A Hybrid Model for Linking Multiple Social Identities across Heterogeneous Online Social Networks

Athanasios Kokkos[1], Theodoros Tzouramanis[1], Yannis Manolopoulos[2]

[1] Department of Information and Communication Systems Engineering,
University of the Aegean – Samos, Greece
{ath.kokkos, ttzouram}@aegean.gr

[2] Department of Informatics,
Aristotle University of Thessaloniki, Greece
manolopo@csd.auth.gr

Abstract. Automated online profiling consists of the accurate identification and linking of multiple online identities across heterogeneous online social networks that correspond to the same entity in the physical world. The paper proposes a hybrid profile correlation model which relies on a diversity of techniques from different application domains, such as record linkage and data integration, image and text similarity, and machine learning. It involves distance-based comparison methods and the exploitation of information produced by a social network identification process for use as external knowledge towards searches on other social networks; thus, the remaining identification tasks for the same individual are optimized. The experimental study shows that, even with limited resources, the proposed method collects and combines accurate information effectively from different online sources in a fully-automated way. The mined knowledge then becomes a powerful toolkit to carry out social engineering and other attacks, or for profit and decision-making data mining purposes.

Keywords. Online identities linkage, online social network profiles data similarity and linkage model, profile matching formulas, machine learning.

1 Introduction

The social web keeps generating swathes of publicly available personal data despite the documented implications of voluntarily exposing personal data on online social networks (OSNs). The methodologies to circumvent traditional privacy preserving countermeasures cover the use of either text analysis for author identification [1], re-identification algorithms on graph-structured data [2], or diverse linkages of the users' footprints in different online data sources [3], and so on.

This study focuses on the concept of online user profiling [4] that can be defined as the cross correlation of publicly available personal information for the successful identification and linking of online social profiles across heterogeneous social networking services that correspond to the same individual in the real world. The same problem in a wider setting is known as record linkage, entity resolution, profile matching, etc. [5].

A hybrid model is proposed that combines data analysis techniques (i.e. from record linkage to text mining) to select information from online data sources in an unsupervised fashion and to weave together the different online social identities of an individual. The

model can be adopted in diverse data mining application domains, from security evaluation tools that focus on privacy risks assessments (e.g. physical identification, de-anonymization attacks, password recovery attacks), to data warehousing technologies for building repositories of OSN data from multiple heterogeneous online sources.

Section 2 develops around the proposed methodology for comparing and linking the users' social identities across heterogeneous OSNs. Section 3 describes the implementation and experimental performance efficiency of a prototype system to identify and link together the different profiles of targeted individuals. Section 4 analyses the advantages and limitations of the related work, and argues in favour of the proposed model. Section 5 summarizes and makes suggestions for future research.

2 The Proposed Model

The proposed model collects publicly available personal information from various OSNs. This personal information is inferred by exploiting their weaknesses e.g. using the email querying functionality offered by some networks, if the targeted individual (from now on called "the target") has been registered with an email address known to the attacker, and by exploiting hidden personal information extracted by applying machine learning techniques on openly available information on the OSNs e.g. the gender inference of a user on Twitter. Its effectiveness increases with every piece of accurate personal information collected during runtime, which, after its collection, is used dynamically to strengthen the search for the target from one OSN to the next. Should the profile of a targeted OSN user be uniquely identified, the proposed method takes into

Algorithm 1: MainRoutine (), Part I
Input: A list $O[1..m]$ of m OSNs to search for a target.
Output: The social identity $Q.v$ of a target Q in the physical world, defined as the union of all the OSNs profiles that belong to Q.

Let $U[1..t]$ be a list of t profiles of users of an OSN. Every such profile is a vector v of personal attributes;
Let $matchedInOSN[1..m]$ be a list of m flags (initially all FALSE) indicating if a matching profile has been found in the m OSNs.
Let $\Theta_{overall}$ be a predefined profile matching threshold (e.g., = 0.60);
$maxW$ = 0; $MatchedProfile$ = NULL; $matched$ = FALSE;

1: FOR i = 1 TO m DO
2: The i^{th} OSN is accessed and the names & types of the personal attributes that the OSN publishes for every user are dynamically recorded.This way, an aggregated vector v of personal attributes in the form $v = (v_1,\ldots,v_d)$ is constructed after accessing all the OSNs;
3: The attacker determines the weight coefficient for every personal attribute in a vector λ of weights in the form $\lambda = (\lambda_1,\ldots,\lambda_d)$;
4: The attacker determines which personal attributes are mutually compatible, by providing values τo a compatibility table $ACT(1..d,1..d]$;
5: For a target Q, the attacker provides the attributes' values of the personal profile $Q.v$ that are known from external sources;

Algorithm 2;

Algorithm 1. The main routine of the proposed model (Part I).

account this verified personal information to expand the set of valid data already available for that individual and uses it in combination with all the remainder of the known personal attribute values of that individual to search for her/his profile in other OSNs. The description of the model follows.

2.1 The Main Routine

The *pseudo*-code of the main routine of the proposed OSN identity identification and correlation model is presented in Algorithmic blocks 1 and 2, with the first initiating the process and the second being the main part of the process.

The algorithm in Lines 1-2 records the personal attributes published online in every OSN under consideration producing a vector v of personal attributes in the form $v = (v_1, v_2, \ldots, v_d)$. For example, if the first OSN provides the attributes (*username*, *firstname*, *lastname*, *work*, *education*, *bio/cv*, *hometown*, *country*, *image*, *friend*) and the second OSN provides the attributes (*username*, *fullname*, *gender*, *organization*, *summary*, *location*, *email*, *follows*, *followedby*), then the vector v has the form v = (*username*, *firstname*, *lastname*, *fullname*, *gender*, *work*, *organization*, *education*, *bio/cv*, *summary*, *hometown*, *country*, *location*, *email*, *image*, *friend*). If an OSN provides the '*follows'* and '*followedby'* attributes (e.g. Twitter and Instagram) only the users that belong in both these sets will be considered as friends. This 'friends' selection can be approached from a different perspective.

	username	firstname	lastname	fullname	gender	work	organization	education	bio/cv	summary	hometown	country	location	email	image	friend
username	√	+	+	√												
firstname	√	√		√					√							
lastname	√		√	√					√							
fullname	√	+	+	√					√							
gender					√											
work						√	√		√	√						
organization						√	√		√	√						
education								√	√	√						
bio/cv		+	+	+		+	+	+	√	√	+	+	+			
summary						+	+	+	√	√	+	+	+			
hometown									√	√	√		√			
country									√	√		√	√			
location									√	√	+	+	√			
email														√		
image															√	
friend																√

Fig. 1. A sample of compatible personal attributes for the matching process[1].

[1] The existence of two or more '+' indicators in a horizontal line means that the attributes on the corresponding columns need to be combined to be compared against the attribute on the horizontal line.

username	firstname	lastname	Fullname	work	...	gender	...	country	...
	Federico García	Lorca	Federico García Lorca	poet	...	M	...	Spain	...
	García		García Lorca			Male			

Fig. 2. A sample of personal information of an individual that may be available to the attacker.

In Line 3 the attacker evaluates the weight coefficient of every personal attribute in the vector *v*, to specify the attribute's importance in the profile identification process in relation to the other attributes. The attacker needs to provide the values to a vector λ of weights of the form $\lambda = (\lambda_1, ..., \lambda_d)$, where $\forall\ i \in [1, d]$: $\lambda_i \leq 1$, $\lambda_1+...+\lambda_d = 1$, and with the exception that $\lambda_{email} = \lambda_{image} =$ NULL. Practical experimentation can also fine tune these weights (e.g., $\lambda_{fullname} = 0.15$, $\lambda_{summary} = 0.01$). In Line 4 the attacker evaluates the attributes' compatibility by filling a three-value table as in Figure 1: the last name of a target can be found in the *fullname* attribute; the full name of an individual can be constructed as the combination of the values of the attributes *firstname* and *lastname*, etc.

In Line 5 the attacker provides the values of the personal attributes of the target that are known from external sources (Figure 2). Every personal attribute in the profile vector *v* of a target's digital identity is a list of values (implementation as non-normalized database relational table or as XML document).

Algorithm 2 implements the main process. For every OSN under consideration, it examines (Lines 8-10) whether it is possible to uniquely identify the target using the email functionality provided by some OSNs. If this test succeeds, the targeted profile has

Algorithm 2: MainRoutine (), Part II

```
6:    FOR i = 1 TO m DO
7:       IF matchedInOSN[i] = FALSE THEN
8:          U[1] = IdentificationViaEmail(Q.v_email);
9:          IF U[1]<> NULL THEN
10:            Q.v = Q.v merge U[1]; matchedInOSN[i] = TRUE; matched = TRUE;
11:         ELSE
12:            Search the i^th OSN for retrieving t profiles U[1..t], one
                  of which may potentially correspond to Q;
13:            FOR j = 1 TO t DO
14:               IF IdentificationViaAnImage(Q.v_image, U[j]).image) THEN
15:                  maxW = Θ_overall; MatchedProfile = j; BREAK;
16:               ELSE
17:                  S=IdentificationViaTheProfileAttributes(Q.v,U[j],ACT),
                        where S is a vector of personal attributes simi-
                        larity scores in the form S = (s_1,…,s_d), in which
                        ∀ k ∈ [1, d]: s_k ≤ 1, and s_email = s_image = NULL;
18:                  Calculate W = λ_1*s_1+…+λ_d*s_d, where W is the OSN-specific
                        normalised overall attributes similarity score bet-
                        ween Q.v and the profile U[j].It is noted that W≤1;
19:                  IF W > maxW THEN maxW = W; MatchedProfile = j;
20:            IF maxW >= Θ_overall THEN
21:               Q.v = Q.v merge U[MatchedProfile]; matchedInOSN[i] = TRUE;
                     matched = TRUE;
22:      IF (i = m AND matched) THEN i = 1; matched = FALSE;
23:   RETURN Q.v;
```

Algorithm 2. The main routine of the proposed model (Part II).

been uncovered and all information published on that OSN is copied and enriches the vector *v* of personal information known to the attacker. Otherwise, in Line 12 of the algorithm the OSN is queried using its own search functionality (e.g., through its API), using prior knowledge obtained (e.g. using the values of the first and last names and the home town, or using any other combination of attributes). This search may select *t* profiles from the OSN, one of which may potentially correspond to the target. Therefore, these *t* profiles are considered for further examination.

Using a high image-similarity threshold Θ_{image} (e.g. $\geq$0.90), an image comparison function in Lines 14-15 (implemented on the basis of any known image similarity measurement [6]) can conclude whether any of the previously known to the attacker profile images for the target coincides with a profile image in any of these *t* profiles in the OSN under consideration. In the negative, in Line 17 a profile-comparison function – see Algorithm 3 – compares the prior knowledge obtained with the related information about every single one of the *t* profiles. Then in Line 18 an overall weighted similarity score *W* (normalized with regard to the attributes available on the OSN) between every selected profile from the OSN and the personal information known to the attacker is calculated; if the highest score *W* exceeds the corresponding predefined similarity threshold $\Theta_{overall}$, then in Line 21 this profile is recognized as the corresponding profile of the target in the OSN under consideration.

The process of Lines 6-21 is repeated for every OSN. Where the identification process has not yet yielded positive results (i.e., if `matchedInOSN[i]=FALSE`), this process is repeated using the enriched information previously gathered from other OSNs. The process terminates if Line 22 indicates that no additional new information can be identified.

2.2 Identification *via* the OSN Profile Attributes

The identification of a profile in an OSN *via* the values of its profile attributes (Line 17 of Algorithm 2) is based on a record linkage method that computes the similarities between the known values of an individual's personal attributes and the corresponding attributes on a potential matched profile in the OSN. The *pseudo*-code of this profile linkage method is illustrated in Algorithm 3.

If the attacker knows the target's gender, the AttributeComparison function (Line 4 of the algorithm) checks whether the gender has been provided by the examined user in the OSN under consideration. If not, a text mining technique method for gender inference is used [7]. If this approach confidently concludes on the value of the OSN user's gender, then this value will be compared against the gender, and a corresponding similarity score of 1 or 0 will be provided. Should at least one friend of the target be known to the attacker, the AttributeComparison function for the *friend* attribute takes the form of Algorithm 4, discussed in the sequel.

Since fine-tuned and extensively tested similarity comparison methods in matching problems can perform poorly in new and different matching problems [8], the comparison performance of compatible text attributes with different representations was tested. In case of a name, the text similarity comparison function may need to overcome *lexical heterogeneities* between any two compared strings, as with 'Federico García Lorca' and 'García Lorca'. This study developed and tested several string comparison functions, such as the similarity functions Jaro-Winkler [9] and Jaccard [10, 11]. Two versions were developed for the Jaccard function: a character-based similarity function and a to-

Algorithm 3: IdentificationViaTheProfileAttributes
Input: The digital identity of a target Q, defined as a vector $Q.v$ of personal attributes, some of which are known to the attacker.
The profile of a given user U in an OSN as a vector $U.v$ of d attributes in the form $U.v = (v_1, \ldots, v_d)$, the values for some attributes of which are available in the OSN.
An attributes' compatibility table $ACT(1..d, 1..d]$.
Output: The vector S of similarity scores between the corresponding d attributes of $Q.v$ and $U.v$, in the form $S = (s_1, \ldots, s_d)$, where $\forall\ i \in [1,d]$: $s_i \leq 1$, and $s_{email} = s_{image} =$ NULL.

Let $\Theta_{attribute}$ be a predefined threshold to get rid of any random noise-based attribute similarity (e.g., $\Theta_{attribute} = 0.20$)
1: FOR $i = 1$ TO d DO
2: IF ($Q.v_i$ <> $Q.v_{email}$) and ($Q.v_i$ <> $Q.v_{image}$) THEN
3: $S.s_i$ = AttributeComparison($Q.v_i$, U, $ACT[i, 1..d]$);
4: IF $S.s_i < \Theta_{attribute}$ THEN $S.s_i = 0$;
5: RETURN S;

Algorithm 3. Identification process of a targeted profile in an OSN *via* its profile attributes.

ken-based one; the first to compare strings using a character-based similarity metric; the second using a word-based similarity metric.

As illustrated in Figure 3, both the character-based and the token-based similarity metrics produce the same results if the strings under comparison are identical (a score of 1.0 indicates a perfect match). In this case the character-based metrics perform more comparisons than the token-based metrics; in the case of the comparison of two strings distinguished by few different characters (e.g. two quite similar usernames), the potentially low similarity scores produced by the Jaccard token-based function make it unsuitable. The optimistically high scores, which both the Jaro-Winkler and Jaccard character-based metrics produced, are not always realistic. As the Hybrid method performed more accurately in most of the preliminary tests carried out, it was selected for the implementation of the attributes-comparison function in Line 3 of Algorithm 3.

With respect to the attribute's lexical heterogeneity, every personal attribute in the profile vector of a target's digital identity that is already known to the attacker might have more than one single value, e.g. for the *city* attribute's values, 'Paris, France' and

Source of information	**username**	**firstname**	**city**	**country**	**organization**
The attacker's knowledge:	consba	Constantin	Linz	Austria	Johannes Kepler University
An OSN:	consbakery	Constantin	-	Linz, Austria	University of Southern California, Johannes Kepler University

(a)

Comparison function	**username**	**firstname**	**city**	**country**	**organization**
Jaro-Winkler (character-based):	0.78	1.0	0.0	0.41	0.61
Jaccard (character-based):	0.56	1.0	0.0	0.64	0.76
Jaccard (token-based):	0.0	1.0	0.0	0.33	0.43
Hybrid (2-grams & tokens):	0.55	1.0	0.0	0.50	0.50

(b)

Fig. 3. (a) Two vectors of personal attributes, (b) The results of the comparison of the compatible attributes of these vectors with four different text similarity comparison functions.

'Paris'. This means that the attributes-comparison function (Line 3 of Algorithm 3) needs to check for a match of every possible alternative known value for a personal attribute and to select the highest possible similarity score.

Besides lexical heterogeneity, a *structural heterogeneity* needs to be dealt with since most of the OSNs represent user profiles differently and use different database schemas. If an OSN provides values to the attribute *fullname* instead of the attributes *firstname* and *lastname* that might be known to the attacker, the attributes-comparison function needs to take into account the predetermined combinations of compatible attributes (Line 4 of Algorithm 1).

2.3 The Friends-Comparison Function

The *pseudo*-code of the function that examines whether the profile of any friend of a target matches the profile of an online friend of an examined user in an OSN is illustrated in Algorithm 4 (for simplicity, the friends of the friends of the target are not considered for examination by the algorithm). The function finally calculates and returns in Line 6 the ratio of the matched friends between the target and the examined user in the OSN.

Algorithm 4: AttributeComparison for the '*friend*' attribute.
Input: The list $Q.v_{friend}[1..m]$ of profiles of the m friends of a target Q.
The list $U.friend[1..h]$ of profiles of the h online friends of a given user U in an OSN.
The attributes' compatibility table $ACT(1..d, 1..d]$.
Output: The ratio of Q's friends matching with the U's friends.

Let *matched* = 0 be the number of matched friends;
1: FOR i = 1 TO m DO
2: FOR j = 1 TO h DO
3: S=IdentificationViaTheProfileAttributes($Q.v_{friend}[i]$,$U.friend[j]$,ACT)
where S is a vector of personal attributes similarity scores in the form $S = (s_1,\ldots,s_d)$, in which $s_{email} = s_{image} = s_{friend}$ = NULL;
4: Calculate $W = \lambda_1 * s_1 + \ldots + \lambda_d * s_d$, where W is the overall attributes similarity score between the profiles $Q.v_{friend}[i]$ and $U.friend[j]$. It is noted that $W \leq 1$;
5: IF W >= $\Theta_{overall}$ THEN *matched*++;
6: RETURN *matched*/m;

Algorithm 4. The friends-comparison function.

3 Experimental Study

3.1 The Model's Preparation Phase

It is assumed that an attacker who aims to construct the digital profile of a group of researchers appearing as authors in articles indexed by the DBLP service compiles a digital dossier by searching the OSNs: Facebook, Twitter, LinkedIn, Google+ and MySpace. A web crawler is developed to access these five OSNs and construct the vector $Q.v$ of personal attributes found online in these data sources for every target Q.

Then a SAX parser is developed to extract the first and last name as well as the publications and the co-authors of every researcher in the DBLP website, by using its XML-based API. The set of co-authors for every researcher is the primary source for finding real-life friends of the researcher. The parser also extracts all the web links that point to external digital libraries, such as the SpringerLink, the IEEE's Xplore, the ACM Digital Library (DL), the Elsevier's ScienceDirect, etc., and adds them to the list of external URLs for browsing to uncover more personal information about the targets. A crawling of the above major digital libraries is performed to gather additional identifiable personal information such as the city and country of residence, institutional affiliation/place of work, email, telephone number, postal address, postcode, etc. The more recent publications stored in the DBLP are considered first, since they might provide more accurate personal information (the SpringerLink and the IEEE's Xplore at the time activated processes to prevent web bots from crawling).

In the absence of gender information, the Baby Name Guesser service[2] is queried and the gender attribute is obtained *via* probabilistic estimation. If this response is accompanied by a high degree of confidence (e.g., 'John'/'Joanna'), it is considered that the researcher's gender is known to the attacker. The Geonames service[3] is queried to perform a cleaning process (e.g., by correcting misspellings) and a verification of the names of cities, countries and locations. The target group produced is a subset of 4,324 researchers for which as much identifiable personal information as possible was gathered *via* external web sources.

The weight coefficients of the importance of the available personal attributes were empirically selected (e.g., $\lambda_{\text{firstname}} = 0.15$, $\lambda_{\text{lastname}} = 0.15$, $\lambda_{\text{gender}} = 0.05$, $\lambda_{\text{education}} = 0.05$, $\lambda_{hometown} = 0.075$, $\lambda_{\text{country}} = 0.075$, *etc.*) by performing some preliminary tests and the values for the three thresholds appearing in the identification process were manually set as follows: $\Theta_{overall} = 0.60$, $\Theta_{image} = 0.90$ and $\Theta_{attribute} = 0.20$.

3.2 The Model's Execution Phase

The next step is to carry out a search, one at a time, for these 4,324 individuals on the five OSNs to uncover all the publicly available information from their profiles. An XML parser, a JSON parser, an HTML parser and an aggregator module were developed for crawling and collecting this accessible information from the OSNs APIs or *via* 'screen-scraping'. The aggregator module performs additional data warehousing functionalities, such as data cleaning, data transformation, data integration, data mining (for gender inference), etc. For every target, the final goal of the proposed model is to identify and correlate at most one social identity from every OSN that possibly belongs to this individual. In the performance evaluation phase, the *Accuracy* metric for measuring the effectiveness of the proposed model in every examined OSN is defined as:

$$Accuracy = \frac{\begin{array}{c}number\ of\ targets\ for\ which\ the\ model\ correctly\ identified\ their\ OSN\ profile \\ or\ correctly\ identified\ that\ no\ such\ OSN\ profile\ exists\end{array}}{total\ number\ of\ targets\ in\ the\ dataset}$$

[2] http://www.gpeters.com/names/

[3] http://www.geonames.org/

in which the model is considered to correctly identify an OSN profile or to correctly point to no existing OSN profile if this is also validated by manual inspection.

Figure 4 shows the accuracy of the proposed profile identification model in every examined OSN for the selected group of 4,324. The accuracy of the profile matching algorithm on Facebook reaches 0.71, and consists of an almost equal percentage of true positives (TP) and true negatives (TN). The cause of the unexpected error ratio of 0.29 is the rather low overall profile matching threshold $\Theta_{overall} = 0.60$, which was selected for such a high number of user profiles of Facebook (which exceeded 1.71 billion as of June 30, 2016[4]), with many users unavoidably sharing the same first, last, or their full name, and other personal attributes, producing a rather high percentage of 28.1% of false positives (FP). This result, however, leads to the expectation that, with an exclusively customized-for-Facebook tuning of the attributes weight coefficients, the model's performance can be markedly improved.

OSN:	Facebook	LinkedIn	Google+	Twitter	MySpace
Accuracy:	0.710	0.900	0.889	0.926	0.956

Fig. 4. The accuracy of the proposed profile identification model for the selected set of OSNs.

On LinkedIn the accuracy level of the model is 0.90, consisting of the majority (i.e., about 72%) of TP, as expected, due to the nature of the targeted group of individuals which was selected on the basis of their occupation. Here, most of the targets take care of ensuring that their profile fields are accurate and comprehensive, which contributes to the profile identification process. In the case of Google+ the model's accuracy rate is 0.889, consisting however in the majority (i.e., about 65%) of TN. The number of FN in Google+ (which is 3.1%) is higher than in Facebook (0.9%) and LinkedIn (0.3%), which can be explained by the higher number of user profiles in Google+ with partial or missing personal information.

On Twitter and MySpace the model also achieves very high accuracy ratings. In these OSNs the number of TP is much smaller than the number of TN, which is explained by the nature of the targeted group. Also, about 30% of the number of TP in MySpace has been successfully encountered by a unique identification *via* a known email address. Notably, in most of these cases the values for the remainder of the personal attributes (emails excepted) in the verified MySpace profiles would not correlate correctly with the profiles of the targets due to missing personal data or deliberate misinformation in the profiles. The percentage % of the targets with identified (by our model) online presence in one, two, three and four of the OSNs under consideration is recorded in Figure 5. As expected, most of the researchers in the DBLP dataset, and who have an online presence in OSNs, maintain personal profiles mainly in LinkedIn and/or in Facebook.

	Presence in one OSN	Presence in two OSNs	Presence in three OSNs	Presence in four OSNs	Presence in five OSNs
Percentage of individuals in the dataset:	44.080%	9.968%	0.902%	0.046%	0%

Fig. 5. The identified presence of the targets in the five OSNs under consideration.

[4] https://newsroom.fb.com/company-info/

4 Related Work

The term record linkage [12] refers to the task of identifying tuples that represent the same entity in one or more, possibly heterogeneous, data sources. In recent years, this concept shifted to matching users' profiles across different OSNs, whereby the identification process detects and weaves together the multiple online social identities of the same entity. Different methodologies have been developed to establish whether a user profile in an OSN belongs to a targeted physical entity: by using the user's email address in [13]; the user's pseudonyms in [3]; the username in [14]; the <*username*, *name*, *location*> attributes in [2]; the Google search service together with the <*occupation*, *education*> attributes in [15]; the <*firstname*, *lastname*, *email*> attributes together with three friends in [16]; the <*instance-messenger-identifier*, *personal website url*, *name*, *hometown*, *birthday*, *university*, *high school*, *gender*, *email*, *friends*> attributes in [17]; machine learning techniques on several personal attributes and the friends' list in [18, 19]; the user's social behavior across time and the close to the user social network structure in [20], *etc*.

The linking task is made difficult by the high degree of heterogeneity of the information available. The methodologies with a limited degree of effectiveness are the simplified approach in [15], relying solely on Google search results on the basis of a predefined set of known personal attributes of the targets to uniquely identify their OSN profiles, and the approaches that rely on the similarity of one or of a small subset of personal attributes (e.g. [3, 14]). While the approach proposed in [13] is effective in a few individual cases, the OSNs which offer the desired *friends-finder* functionality using their known email addresses are not many because this feature threatens user privacy. Closer to the method proposed in this paper are the matching algorithm approach in [17] which takes into account a static predefined set of 10 attributes for the OSN profile identification process, and the work based on supervised learning on several profile attributes and the friends' list in [18]. The main advantages of our model are that: firstly, the set of personal attributes to be utilized in the identification process is practically unlimited (since every possible personal attribute on an OSN can provide valuable data input for the matching process in other OSNs); secondly, the selection of these attributes and the identification operation provided by the model are fully-automated tasks (not manual or supervised tasks for the attacker); and, thirdly, the amount of personal data available to the attacker increases during the identification process, which means that an examined OSN may be accessed several times during this process, every time with an increased pool of prior knowledge, increasing the likelihood of positive results.

Figure 6 summarizes the performance achievements of previous work, setting out the best reported performance ratings, including those achievable only under significantly restricted conditions. The proposed model appears at the bottom indicating that it outperforms most of the earlier work in OSNs profile identification and matching and, to the best of the authors' knowledge, it appears to be the first to address the problem by combining a number of different methodologies, such as machine learning techniques, a variety of linkage methods and, very importantly, by exploiting the verified knowledge produced during the runtime of the identification process in an unsupervised fashion.

OSNs' profiles linkage model	Accuracy	Precision[5]	Recall[6]
Narayanan & Shmatikov [2]	0.308	–	–
Irani et al. [3]	0.600	–	–
Balduzzi et al. [13]	0.049	–	–
Wang et al. [14]	–	0.862	0.685
Vosecky et al. [17]	0.930	–	–
Peled et al. [18]	0.959	–	–
Zhang et al. [19]	–	0.860	0.867
Liu et al. [20]	–	0,968	0.908
Wondracek et al. [21]	0.577	–	–
Goga et al. [22]	–	0.950	0.290
Human inspection [22]	–	0.960	0.400
This paper	**0.956**	**0.904**	**0.985**

Fig. 6. The best provided Accuracy, Precision and Recall by several OSNs profiles linkage models.

5 Conclusions and Future Research

The model proposed for identifying and linking the multiple online social identities of the same physical entity across OSN services combines methodologies to collect, infer and integrate accurate personal information from heterogeneous OSN sources to build a warehouse of digital footprints that can be used in several application domains. This hybrid architecture is built upon four different methods for OSNs profiles matching, and operates using a limited amount of prior knowledge about the target. Every piece of accurate information from one OSN is exploited in other OSNs for the remaining matching tasks, dynamically increasing the model's efficiency. Additionally, the model is operational without any modification in any OSN[7] and in any language.

The empirical performance evaluation of the proposed framework with a dataset of 4,324 individuals indicated that it can successfully retrieve and link together the social identities of the targets across multiple OSN services, regardless of their different database schemas and of lexical and structural heterogeneity. It also indicates that this model outperforms most of the earlier work.

Scope for further exploration includes developing increasingly sensitive modules for measuring the similarity between OSN profiles attributes of textual, date, image, or any other specialized data type. For example, the traditional syntactic-based text similarity metrics might not be able to capture a valuable similarity between two attribute values that are semantically related [23] while lexicographically different (e.g. "MS Corporation" and "Microsoft Inc."). Besides, the effectiveness of our method could be increased by fine-tuning several operational parameters (such as the attributes weight coefficients and the similarity thresholds). An extension of the proposed model could aim to establish links with any information that may be of value for the purpose of targeting individuals from any existing online footprint that can be uncovered and from any trustworthy source.

---

[5] *Precision* is defined as *TP* / (*TP* + *FP*) and represents the ratio of correct user profiles identifications in an OSN.

[6] *Recall* is defined as *TP* / (*TP* + *FN*) and represents the ratio of correct user profiles identifications to the total number of existing user profiles to be identified in an OSN.

[7] refer to e.g. the following list of over 200 OSNs at the time of writing: https://en.wikipedia.org/wiki/List_of_social_networking_websites

References

1. Chaski, C. E.: Empirical evaluations of language-based author identification techniques. *Forensic Linguistics*, 8 (2001) 1-65.
2. Narayanan, A., and Shmatikov, V.: De-anonymizing social networks. *Proceedings 30th IEEE Symposium on Security & Privacy* (2009) 173-187.
3. Irani, D., et al.: Large online social footprints - an emerging threat. *Proceedings IEEE International Conference on Computational Science & Engineering* - CSE, 3 (2009) 271-276.
4. Erlandsson, F., Boldt, M., and Johnson, H.: Privacy threats related to user profiling in OSNs. *Proceedings IEEE International Conference on Social Computing* (2012) 838-842.
5. Christen, P.: *Data matching: concepts and techniques for record linkage, entity resolution, and duplicate detection*. Springer Science & Business Media (2012).
6. Flickner, M., et al.: Query by image and video content: The QBIC system. *IEEE Computer*, 28(9), (1995) 23-32.
7. Kokkos, A., and Tzouramanis, T.: A robust gender inference model for online social networks and its application to LinkedIn and Twitter. *First Monday*, 19(9) (2014).
8. Bilenko, M., Mooney, R., Cohen, W., Ravikumar, P., and Fienberg, S.: Adaptive name matching in information integration. *IEEE Intelligent Systems*, 18(5), (2003) 16-23.
9. Winkler, W.E.: String comparator metrics and enhanced decision rules in the Fellegi-Sunter model of record linkage. *Proceedings of the Section on Survey Research Methods*, American Statistical Association (1990) 354-359.
10. Jaccard P.: Lois de distribution florale. *Bulletin de la Socíeté Vaudoise des Sciences Naturelles* 38 (1902) 67-130.
11. Jaccard P.: The distribution of the flora in the alpine zone. *New Phytologist* 11(2),(1912) 37-50.
12. Winker E.W.: Overview of record linkage and current research directions. *Statistical Research Division U.S. Census Bureau* (2006).
13. Balduzzi, M., et al.: Abusing social networks for automated user profiling. *Proceedings International Workshop on Recent Advances in Intrusion Detection* (2010) 422-441.
14. Wang, Y., Liu, T., Tan, Q., Shi, J., and Guo, L.: Identifying Users across Different Sites using Usernames. *Procedia Computer Science*, 80 (2016) 376-385.
15. Bilge, L., Strufe, T., Balzarotti, D., and Kirda, E.: All your contacts are belong to us: automated identity theft attacks on social networks. *Proceedings 18th ACM International Conference on WWW* (2009) 551-560.
16. Zhou, C., Chen, H., and Yu, T.: Learning a probabilistic semantic model from heterogeneous social networks for relationship identification. *Proceedings 20th IEEE International Conference on Tools with Artificial Intelligence* 1 (2008) 343-350.
17. Vosecky, J., Hong, D., and Shen, V.Y.: User identification across multiple OSNs. *Proceedings 1st IEEE International Conference on Networked Digital Technologies* (2009) 360-365.
18. Peled, O., Fire, M., Rokach, L., and Elovici, Y.: Matching Entities across Online Social Networks. *Neurocomputing* 210, (2016) 91-106.
19. Zhang, Y., Tang, J., Yang, Z., Pei, J., and Yu, P.S.: COSNET: connecting heterogeneous social networks with local and global consistency. *Proceedings 21st ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* - KDD (2015) 1485-1494.
20. Liu, S., Wang, S., Zhu, F., Zhang, J., and Krishnan, R.: HYDRA: Large-scale social identity linkage *via* heterogeneous behavior modeling. *Proceedings ACM International Conference on Management of Data* - SIGMOD (2014) 51-62.
21. Wondracek, G., Holz, T., Kirda, E., and Kruegel, C.: A practical attack to de-anonymize social network users. *Proceedings IEEE Symposium on Security & Privacy* (2010) 223-238.
22. Goga, O., Loiseau, P., Sommer, R., Teixeira, R., and Gummadi, K.P.: On the reliability of profile matching across large online social networks. *Proceedings 21st ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* - KDD (2015) 1799-1808.
23. Egozi, O., Markovitch, S., and Gabrilovich, E.: Concept-based information retrieval using explicit semantic analysis. *ACM Transactions on Information Systems*, 29(2) (2011) article 8.