Community-based Identity Validation on Online Social Networks

Leila Bahri, Barbara Carminati, and Elena Ferrari
DiSTA, Università degli Studi dell'Insubria, Varese, Italy
{leila.bahri, barbara.carminati, elena.ferrari}@uninsubria.it

***Abstract*—Identity management in online social networks (OSNs) is a challenging, yet important requirement for effective privacy protection and trust management. Literature offers several proposals addressing issues related to identity breaches and/or identity related attacks on OSNs, but only a few aim at giving means to judge users' reliability in terms of trustworthiness of their claimed identities. In this paper, we propose an identity validation process that relies on OSN community feedback to assign to OSN users identity trustworthiness levels. For this purpose, we define a community-based supervised learning process to detect the set of attributes in a user profile for which it is expected to see a correlation among their values (e.g., job and salary). Once these correlated attribute sets are identified, the profile of a target user is judged by a selected group of raters to estimate her identity trustworthiness level. We demonstrate the effectiveness of our proposal through experimentation under two different scenarios and using real data.The experiments' results under the two scenarios demonstrate the effectiveness and meaningfulness of our proposal.**

I. Introduction

Socializing with others remains one of the crucial needs of humankind by which they maintain social and cultural continuity [1]. In the context of socializing, identification of the others becomes crucial in that it helps create a sense of trust among people and establish strong relationships between them [2]. In fact, the more we know about the others, the safer we are in dealing with them and the higher our trust in the environment is [2][3]. Indeed, as discussed in [3], one of the main requirements for trust to occur is that involved participants should be sure about the identity of each other.

As OSNs provide an alternative environment for people to socialize, having a means to reliably identify the person behind the screen becomes a strong requirement for a safer and trustworthier OSN [4]. That is most probably why, registering to an OSN, in almost all the existing commercialized solutions, requires users to create a profile that exposes personal information with a varying degree of detail. However, when creating a profile, most, if not all, of the provided information is not verifiable. A new joiner, for example, can claim to be a doctor, but the OSN has no one single way of validating the veracity of this information. On the other hand, the OSN cannot decline access to this new user just because this information cannot be validated. Beyond that, some might argue that getting to validate such information might hinder the preservation of users' privacy over the Internet.

Towards addressing these issues, we propose in this paper an identity trustworthiness measure which is expected to reflect the extent to which a claimed identity is indeed reflecting a truthful person behind it. Capitalizing on the social dimension of identity, we investigate the possibility of computing these trustworthiness levels using explicit feedback from the community by requesting them to judge the coherence of the claimed identity based on the coherence among its corresponding profile's attribute values.

As we show in Section VI, the literature has mainly focused on detecting identity related frauds and attacks [9][10][11]. Only a few works aim at giving users themselves means to estimate whether to trust a peer or not based on community feedback, but they generally do so based on historical transactions between users [12][13].

In this paper, we demonstrate that we can design a system which effectively measures trustworthiness of users profiles based only on users' feedback and without the need for any previous transaction between them. We achieve this through a system, which first exploits a community-based supervised learning to identify *coherence relations* among OSN profile attributes. It then uses the defined relations to make the OSN community rate their perceived trustworthiness of a target profile.

The rest of this paper is organized as follows. Section II specifies the goal of the system and discusses the rationale behind it. Sections III and IV detail the two steps of our suggested system. Section V presents the performed experiments and discusses the results. Section VI discusses related work. Section VII draws conclusions.

II. Community-based Identity Validation

We estimate identity trustworthiness on the basis of the coherence within attributes' values on a target profile. This is mainly motivated from sociological works which discuss different theories to explicate the identity formation process and its related conflictual issues. Most of these works tend to agree that the final formed identity converges to satisfy both an inner consistency and a sense of coherence based on homogeneity with regard to some "socially accepted assimilative models" [22]. This is also

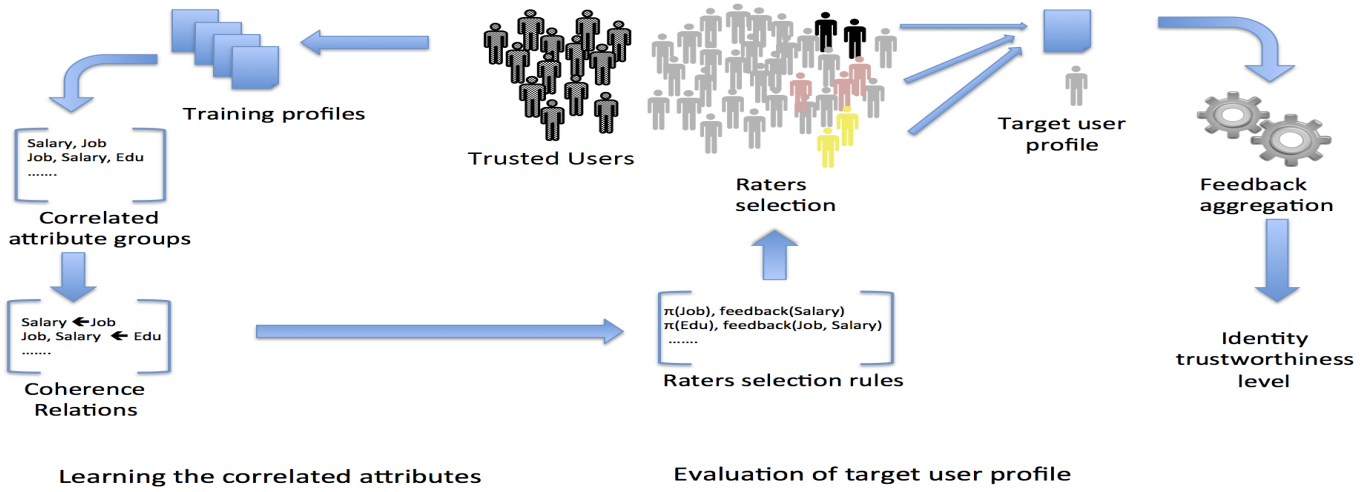

Fig. 1: Community-based identity validation

present in Erik Erikson's theory of identity formation, where he concludes, as mentioned in [23], that "the final identity, [...] includes all significant identifications, but it also alters them in order to make a unique and reasonably coherent whole of them." As such, it is expected that profiles reflecting real identities contain and maintain coherent pieces of information. For example, a profile advertising a person as a University professor, a Ph.D. holder, and earning a salary of about 3K euros per month is coherent as per these three values from some society's perspective.

Having said that, a first step of our proposed process concerns learning the groups of profile attributes for which it is expected to have correlated values. Once these groups are identified, the system evaluates the profile of a target user by asking the community to evaluate the homogeneity between values in the user's profile for the identified correlated attributes. The gathered feedback are then aggregated to estimate an identity trustworthiness level for the target user. For example, if the system learns that job, education, and salary are correlated, then part of the identity trustworthiness level for the target user is estimated based on community feedback on his/her job, education and salary values altogether.

Therefore, the proposed approach implies using the community over two phases (see Figure 1): (1) supervised learning about correlated attribute groups over a set of training profiles, and (2) evaluating values for these attributes on a target user profile.

The first phase is run as bootstrapping of the system over a predefined training profiles' set and a selected group of participants.[1]

Learning correlated attribute groups. In general, not all attribute values are expected to be coherent. In fact, profiles on OSNs mostly represent multiple aspects of identities. A profile, for example, could be advertising both the professional and the personal characteristics of a person which are not necessarily related to each other. A person with a specific job, for instance, can be a parent, single, or in a relationship whatsoever with equal probabilities as job is not determinant of personal relationships. Moreover, some attributes of a profile are obviously not related to each others (e.g., gender and race, or age and gender). Consequently, capturing the coherence of a profile requires first finding those groups of profile attributes whose values are expected to be correlated. We refer to these groups as *correlated attribute groups*. To find out these groups, we exploit community-based supervised learning. As it will be explained in Section III.A, the learning is performed using a group of OSN community participants, referred to as *trusted users*, who are well informed so as to maximize giving reliable feedback for a reliable learning. We assume that the selection of such users is performed by some mechanism, which we do not address in this paper.

Evaluation of a target user profile. In the second phase, the profile of a target user is evaluated by the larger OSN community (which we refer to as *raters*). In general, we expect more meaningful evaluation from *raters* who share the same ground as per some of the values they are asked to provide their opinion on. We propose then to select raters based on the values of correlated attributes of the target profile that has to be judged. For example, assuming that job and salary are correlated, it would be meaningless to ask a student to rate the coherence of the values set *(Job=Plumber, Salary=3K$/month)*, as a student is most probably, if not surely, not knowledgeable about the salary ranges of plumbers. Instead, it would be possible for a rater sharing the same job or working in the same domain to give an informed opinion about that. The

[1]To answer the highly changing dynamics of OSNs with profiles being added, deleted, or modified, this phase is periodically executed by extending the training set over profiles of new OSN joiners.

opposite, however, is not quite straightforward to claim as people sharing the same range for salary value might not be expected to give an informed opinion on the job corresponding to it. This suggests that there should be a relation between the elements of correlated attribute groups specifying which one(s) are determinant to the others.This relation should drive the raters selection for profile evaluation. Therefore, our system needs not only to be aware of the correlated attribute groups, but it also has to learn the direction of relation between them. We call these relations *coherence relations* and we learn them as detailed in Section III.B. Based on these relations, the system selects raters to be involved in the evaluation of correlated attributes over a given target profile. Once all pieces of evaluation are gathered from the selected raters, the system aggregates their values and obtains the estimated *identity trustworthiness level* for the target user.

III. Learning of Correlated Attribute Groups and Coherence Relations

Our problem has some similarities with association mining, where detecting relations among ordered items is done by means of counting the number of occurrences of a group of items across all available historical transactions to infer the strength of the relation between them [6]. Similarly in our scenario, we are interested in finding the set of correlated attributes and the dependency relations between them; however, the occurrence measure cannot be automatically applied within our environment. This is mainly due to the possible wide range of values each attribute can take, as attributes on OSN profiles are in general non-categorical (i.e., users can insert free text in their profile attributes). If we simply count the frequency of occurrence of values, the measure will be very sparse and not informative. Moreover, we will be losing semantics especially in an OSN environment where semiotic patterns are diverse and mostly informal.[2] To overcome this, we measure the strength of relations among attributes by relying on trusted users, who are asked to express a judgment on the coherence they see among their values. Our goal is to learn relations among attributes in a profile rather than relations among their values; we need to learn the strength of the correlation between 'Salary Range' and 'Job' attributes not between their values. This is what we detail in what follows.

A. Correlated Attribute Groups

Towards learning correlated attribute groups, we consider all possible combinations over the profile schema as candidate ones. Trusted users are asked to provide feedback on coherence among values of these candidate groups over the profiles of a training set. This process implies the exposure of profile information, and hence particular attention to users' privacy is required. As a matter of fact, our system shall consider managing the identities of profiles without making the trusted users and the raters able to re-identify the user they are evaluating. This is ensured, in a first instance, by discarding the quasi-identifier attributes from all the reasoning of the system.[3]

[2]Many forms of informal languages are being adopted in digital socializing environments, such as chat-language, abbreviations, etc.

'Definitely yes', 'Definitely no'	1
'Most probably', 'Less likely'	0.5
'Not meaningful to judge', 'I do not know'	0

TABLE I: Feedback types and corresponding values for the learning phase

The formal definition of candidate attribute groups is given below.

Definition 3.1: Candidate Attribute Group. Let $\mathbf{S}$ be the profile schema adopted in the OSN. Let $\mathbf{QI} \subset \mathbf{S}$ be the set of quasi-identifier attributes in $\mathbf{S}$. The set of candidate attribute groups of order m over $\mathbf{S}$ ($|\mathbf{S}| > m > 1$), denoted as CA_m, contains all possible combinations of size m of attributes in $\mathbf{S}\backslash\mathbf{QI}$. We denote as $\mathbf{CA}$ the set of the candidate attribute groups for any size m over $\mathbf{S}$, $|\mathbf{S}| > m > 1$.

Trusted users' feedback is simply a discrete answer to a question on selected values of attributes, which belong to a candidate attribute group, from the profiles in the training set. To make it clearer, let us consider Joanna to be a member of the social network whose profile is in the training set and which contains values *Black African* for race and *Kenya* for country of origin. The question on the profile values of Joanna over {'Race', 'Country of Origin'} is: *'Do you think a real person (not faking an identity) can hold as race* Black African *and as country of origin* Kenya *altogether?'*. Six answers are possible with each of them being related to a discrete value, as specified in Table I. YES and NO answers result, in this phase, in equal feedback scores because both answers imply that users capture in the group of attributes something that makes them able to make a clear judgment.

To evaluate the correlation within a candidate group, we introduce the following definitions. Hereafter, given a user u, we denote the values of attributes in $\mathbf{Y} \subseteq \mathbf{S}$ for user u as a vector $\mathbf{Q}_u^Y$.

Definition 3.2: Feedback on Candidate Attribute Groups. Let $\mathcal{TU}$ be the set of available trusted users in the OSN, let u be a user in the OSN, with $u \notin \mathcal{TU}$. Let $\mathbf{Y} \in \mathbf{CA}$ be a candidate attribute group, and $\mathbf{Q}_u^Y$ be the values for attributes $\mathbf{Y}$ in u's profile. The feedback of $\mathcal{TU}$ on u w.r.t. $\mathbf{Y}$, denoted as $f_{\mathcal{TU}}(\mathbf{Y}_u) \in [0, 1]$, is computed as:

$$f_{\mathcal{TU}}(\mathbf{Y}_u) = \frac{1}{|\mathcal{TU}|} * \sum_{j \in \mathcal{TU}} f_j(\mathbf{Q}_u^Y)$$

where $f_j(\mathbf{Q}_u^Y)$ is the feedback expressed by the trusted user j on $\mathbf{Q}_u^Y$.

[3]We consider quasi-identifiers those attributes for which the values can be used, alone or together with some other external or internal information, to approximate or to determine the identify of their owner [7]. We assume some attributes out of the profile schema are identified as quasi-identifiers and we do not cover in this paper the process of such identification.

The received feedback per candidate group is aggregated to compute its support:

Definition 3.3: Support. Let $\mathcal{TU}$ be the set of trusted users in the OSN, let $\mathcal{T}$ be a set of users in the OSN whose profiles are in the training set, with $\mathcal{T} \cap \mathcal{TU} = \emptyset$. Let $\mathbf{Y} \in \mathbf{CA}$ be a candidate attribute group, the support for $\mathbf{Y}$ is computed as:

$$supp(\mathbf{Y}) = \frac{1}{|\mathcal{T}|} * \sum_{i \in \mathcal{T}} f_{\mathcal{TU}}(\mathbf{Y}_i)$$

Once all the supports are computed, the groups having a support high enough to justify a correlation between their elements are considered *correlated attribute groups.*

Definition 3.4: Correlated Attribute Group. Let $\mathbf{CA}$ be the candidate attribute groups defined over a profile scheme $\mathbf{S}$. Let $sh \in [0, 1]$ be a threshold value. The Correlated Attribute Groups are defined as $\mathbf{CAG} = \{\mathbf{Y} \in \mathbf{CA} | supp(\mathbf{Y}) \geq sh\}$.

The value of the threshold sh is determined dynamically based on all the computed supports and their distribution. In a preliminary step, and for simplicity, we set it to the mean of all computed supports.

Example 1. Consider the candidate group, {Job, Education}. Suppose we have 3 profiles in the training set and 2 trusted users. This results in 3 feedback questions each corresponding to the values of the candidate group's attributes in a training profile, and in 6 feedback rates (1 per question per trusted user). The support of this group will be the average of these 6 feedback rates. Suppose this support is equal to 0.8. Assuming $sh = 0.5$, then this candidate group is a correlated attribute one.

B. Coherence Relations

Always in analogy with association mining, the detection of coherence relations within correlated attribute groups is done by computing their corresponding confidence. In association mining, given two items R and L, the confidence $Conf(R \implies L)$ is the ratio between the support of the two items (i.e., the occurrences where both items appear) and the support of the single item R, that is, $Conf(R \implies L) = \frac{supp(R \cup L)}{supp(R)}$. If the resulting value is greater than a threshold, then $R \implies L$ can be considered a meaningful rule [6]. As the support of a correlated attribute group is given, in our case, by the average of trusted users' feedback, this recalled confidence definition cannot be directly applied. In designing a new confidence definition, we assumed that the best way to decide if $R \implies L$ is a coherence relation is to consider feedback of trusted users that are informed on R (i.e., have values for attributes in R similar to those of training profiles that they are judging) against the feedback of all other trusted users. Therefore, we introduce the following definition:

Definition 3.5: Conditional Support. Let $\mathcal{TU}$ be the set of trusted users in the OSN, let $\mathcal{T}$ be a set of users in the OSN whose profiles are in the training set, with $\mathcal{T} \cap \mathcal{TU} = \emptyset$. Let $\mathbf{Y} \in \mathbf{CAG}$ be a correlated attribute group. The conditional support for $\mathbf{Y}$, given $\mathbf{B} \in \mathbf{Y}$, denoted as $supp(\mathbf{Y}|\mathbf{B})$, is computed as:
$supp(\mathbf{Y}|\mathbf{B}) = \frac{1}{|\mathcal{T}|} * \sum_{i \in \mathcal{T}} f_X(\mathbf{Y}_i)$, where $X \subseteq \mathcal{TU}$ such that $X = \{x \in \mathcal{TU} | \forall b \in \mathbf{B}, \mathbf{Q}^b_x \in [\mathbf{Q}^b_i - \epsilon, \mathbf{Q}^b_i + \epsilon]$, if b is a numerical attribute; $\mathbf{Q}^b_x = \mathbf{Q}^b_i$, otherwise$\}$, where ϵ is a small tolerance value.

Based on the above definition, we define the following:

Definition 3.6: Confidence. Let $\mathcal{TU}$ be the set of trusted users in the OSN, let $\mathcal{T}$ be a set of users in the OSN whose profiles are in the training set, with $\mathcal{T} \cap \mathcal{TU} = \emptyset$. Let $\mathbf{Y} \in \mathbf{CAG}$ be a correlated attributes group, and let $\mathbf{L} \subset \mathbf{Y}$. We compute the confidence of $\mathbf{L}$ over$\mathbf{Y}$ as:

$$Conf(\mathbf{L}, \mathbf{Y}) = \frac{supp(\mathbf{Y}|\mathbf{L})}{supp(\mathbf{Y})}.$$

Definition 3.7: Coherence Relation. Let $\mathbf{Y} \in \mathbf{CAG}$ be a correlated attributes group, and let $\mathbf{L} \subset \mathbf{Y}$ be a set of attributes such that $\mathbf{R}$= $\mathbf{Y}$ - $\mathbf{L}$. We define p =($\mathbf{L} \implies \mathbf{R}$) as a coherence relation over $\mathbf{Y}$, if $Conf(\mathbf{L}, \mathbf{Y}) > ch$, where ch is a given threshold.

Example 2. Considering the correlated group from Example 1, and supposing that Conf(Job, {Job, Education}) = 1.1, Conf(Education, {Job, Education}) = 1.8, and ch = 1.2, then p =(**Education** $\implies$ **Job**) is a coherence relation over {Job, Education}.

IV. Estimation of Identity Trustworthiness Level

The second phase concerns evaluating the profile of a target user, by a group of selected raters, to estimate its Identity Trustworthiness Level (ITL). Raters selection is driven by the assumption that more meaningful feedback is expected from raters sharing the same ground/values as per the attributes they are asked about. For this purpose, we exploit coherence relations to define a set of conditions for raters selection:

Definition 4.1: Raters Selection. Let $\mathbf{Y} \in \mathbf{CAG}$ be a correlated attribute group, and let $\mathcal{P}$ be the set of coherence relations defined over $\mathbf{Y}$. Let $\mathcal{R}$ be a set of available raters in the OSN, and let u be a user in the OSN whose profile is going under evaluation, with $u \notin \mathcal{R}$. The raters selected for $\mathbf{Y}$ and u, based on the coherence relations in $\mathcal{P}$, are computed as follows:
$RS^Y_u = \{r \in \mathcal{R} | \forall p \in \mathcal{P}, \forall s \in p.\mathbf{L}, Q^s_r \in [Q^s_u - \epsilon, Q^s_u + \epsilon]$, if s is a numerical attribute; $Q^s_r = Q^s_u$, otherwise$\}$.
where ϵ is a small tolerance value, and $p.\mathbf{L}$ denotes the subset $L \in Y$ in the coherence relation p.

Example 3. Let us recall from Example 2, p = (**Education** $\implies$ **Job**). Assume our user Joanna has on her profile: Education = 'Fine arts graduate' and Job = 'Fitness trainer'. Rater selection implies choosing raters who are 'Fine arts graduate' to evaluate Joanna on this group.

ITL is defined as the aggregation of feedback received by selected raters. This feedback is collected as answers

'Definitely yes'	1	'Most probably'	0.5
'Definitely no'	-1	'Less likely'	-0.5
'I do not know'		0	

TABLE II: Feedback types and corresponding values for evaluation phase

to the question provided in Section III.A. However, given that the aim here is to rate the profiles and not to learn relationships between attributes, the values corresponding to the questions are different in that negative and positive answers are judgmental at this level (see Table II).

Definition 4.2: Evaluation of a user profile w.r.t. a correlated attribute group. Let u be the user in the OSN to be evaluated. Let $\mathbf{Y} \in \boldsymbol{CAG}$ be a correlated attribute group. Let RS_u^Y be the corresponding selected raters (see Definition 4.1). Let $\mathbf{Q}_u^Y$ be the values of $\mathbf{Y}$ on the profile of u, and let $f_j(\mathbf{Q}_u^{\mathbf{Y}})$ be the feedback of rater j on $\mathbf{Q}_u^Y$. The evaluation of u w.r.t. $\mathbf{Y}$ is defined as:

$$E_Y(\mathbf{Q}_u^Y) = \frac{1}{|RS_Y|} * \sum_{\mathrm{J} \in RS_u^Y} f_j(\mathbf{Q}_u^{\mathbf{Y}})$$

Based on the above definition, the *Identity Trustworthiness Level* for a target user u is computed as follows:

Definition 4.3: Identity Trustworthiness Level. Let $u \in \mathcal{U}$ be a user in the OSN. Let $\mathcal{Y} \subseteq \boldsymbol{CAG}$ be the set of correlated attribute groups such that $\mathbf{Q}_u^y \neq \emptyset \ \forall y \in \mathcal{Y}$. The identity Trustworthiness Level for u is:

$$\boldsymbol{ITL}_u = \frac{1}{|\boldsymbol{CAG}|} * \sum_{Y \in \mathcal{Y}} E_Y(\mathbf{Q}_u^Y);$$

Example 4. For simplicity, assume we only have the correlated attribute group and the coherence relation identified in Examples 1 and 2 respectively. Consider user Joanna as provided in Examples 3 and assume the system selects two raters who provided feedback corresponding to the values: {0, 0.5}. ITL_{joanna} will consequently be equal to 0.25.

V. Experiments and Discussions

We test the utility of our suggested method, hereafter referred to as CB, under two different experimental environments. On the first hand, we test the effectiveness of CB for learning and detecting correlated attribute groups, by comparing it to a competing alternative on a census dataset (Section V.A). On the other hand, we test the effectiveness and the efficiency of our CB method in rating users' profiles within a real OSN environment (Section V.B).

A. Effectiveness of CB Learning

An obvious alternative to our method in learning correlated attribute groups is opting for machine based association mining techniques. For this reason, we have compared CB learning with association mining learning. The goal of this experiment is to find whether CB learning can capture correlated attributes which cannot be detected by simple machine learning from the data (hereafter, we refer to the machine based learning by MB). In running this experiment, we took in consideration that the performance of MB learning is optimized on categorical and sanitized datasets. For this purpose, we run the experiment on a census dataset.

Attribute	**Description**
Age	Age
Work-class	Work Class
Education	Education Level
Educ-num	Number of years spent at school
Marital-status	Marital Status
Occupation	Job
Social-role	Social Role
Race	Race
Sex	Gender
hrsperweek	Number of hours worked per week
Country	Country of origin

TABLE III: Adults dataset adopted profile schema

1) The Dataset: The dataset is from the US Census Bureau, made available under the name of Adults dataset.[4] It contains 45,222 census descriptive and anonymized records capturing 14 attributes. The dataset comes distributed into 2/3 of its records as training data and 1/3 as validation data. We have considered the 2/3 training records to make our training dataset. This dataset fits most of the requirements for the objective of this experiment. First, it is representative of a user profile and it is rich enough in terms of the attributes it covers. Second, it contains a large number of records. Finally, its values are categorical and well sanitized which makes it favorable for running an association mining algorithm. However, for few attributes, the dataset goes into very fine grained levels of detail as per their values, thing which is not expected on an OSN profile. For example, the education level attribute can take one of sixteen values each referring to the exact schooling year. In order to make this more representative of an OSN user profile, we have over grouped some of these values into one to result in ten possible values only out of the original sixteen. The aim is to make the values in the dataset better understood by the participants in the experiment. In addition to this, 3 attributes have been discarded. These are capital loss, capital gain, and gained-salary.[5] Table III lists all the considered attributes.

2) Experimental Settings and Design: We run CB learning and MB learning on the training dataset. More specifically, we evaluated the correlation between all possible and non-trivial candidate groups, i.e., **CA**, of size 2 over attributes in Table III.[6] For each of these methods, we set the support threshold, *sh*, for correlated attribute groups' identification (see Definition 3.4) to the average of all the computed supports.

MB learning. Given that our training dataset is cate-

[4] http://archive.ics.uci.edu/ml/machine-learning-databases/adult

[5] Capital loss and capital gain are not expected as part of a social profile that one shares on an OSN. The gained-salary attribute is represented in the dataset as a binary salary-class field only.

[6] For proof of concept and for simplicity in results' presentation and discussion, we limit the experiment to CAs of size 2 only. Trivial combinations are ones such as, {country, education}, {race, education}, {country, gender}, etc.

gorical and all its values are sanitized, we have opted for a version of the Apriori algorithm [8] to make our MB learning. This algorithm computes the support of a given CA by counting the number of mutual occurrences of its couple-values. We have slightly altered the algorithm to compute the support for only the 34 CAs we have.

CB learning. We made available to the public an online survey in both English and Italian languages. A survey question has the same format as the feedback question in Section III.A. 38 participants, mainly from Morocco and Italy with little representation of other nationalities such as Iran, Lebanon, Turkey and Tunisia, took the survey. The majority (85%) of our participants are University students or fresh graduates who have recently joined the job market as engineers or salespersons (age is in the range [20, 30]). The remaining 15% of our participants are older professionals within the age range of 35 to 55 and with professions such as, medical doctor, government officer, and University professor. To have a reliable feedback, all our participants have been presented with a thorough explanation about the experiment and its aim, and most of them have accepted to participate by providing careful attention and focus to the survey questions. Each one of the participants answered at least 55 feedback questions corresponding to our 34 candidate groups (CAs). The support of each of the 34 candidate groups has been computed as suggested by the equations in Definitions 3.2 and 3.3.

3) Achieved Results:

Utility in learning correlations. Table IV shows a summary of the MB and CB evaluations of our 34 candidate groups. The table shows only candidate groups evaluated as correlated attributes either by MB or CB. The value 'insig' in the supports column in the table means that the candidate group received an insignificant support using the method to which the sub-column refers (i.e., the support by that method was smaller than the threshold *sh*).

As a first observation, the first 19 candidate groups passed the support test using MB, but failed to prove meaningful correlation using CB. Hereafter, we refer to these groups as CAG_{MB}. Second, we see that only three candidate groups, i.e., the second group of rows in the table, are evaluated as correlated by both MB and CB. This disparity in results between MB and CB can be mainly explained by two elements each specific to one of these methods. On the one hand, MB method captures repeated/common trends in the dataset which are inherent to the statistical and distribution trends of the censused population (i.e., the Adults dataset). This means that the detected CAG_{MB} would be expected to change if the dataset changes. On the other hand, CB method captures logical and social connections or definitions of the participants with regard to what they perceive as a socially conforming combination. This point brings us to stress on the fact that our method is community dependent and justifies its reliance on human feedback instead of relying on statistical or association mining techniques. This can be better understood when examining some of the combinations in CAG_{MB}, such as the group {Gender, Education}. The proved correlation in this combination by MB reflects nothing but that people censused in the Adults dataset tend to have one gender dominance within some educational categories of the data. This group would not be expected to pass the support test using CB because there shall be no relationship between gender and achieved education, unless some strong stereotypes undermine.

	Supports	
Candidate Group	**MB**	**CB**
educ-num, gender	0.36	insig
hrsperweek, gender	0.66	insig
educ-num, race	0.34	insig
hrsperweek, race	0.36	insig
gender, race	0.44	insig
educ-num, social-role	0.29	insig
hrsperweek, social-role	0.30	insig
gender, social-role	0.38	insig
educ-num, marital-status	0.27	insig
hrsperweek, marital-status	0.26	insig
gender, marital-status	0.36	insig
gender, education	0.25	insig
educ-num, work-class	0.29	insig
hrsperweek, work-class	0.30	insig
gender, work-class	0.37	insig
race, work-class	0.21	insig
educ-num, age	0.28	insig
race, age	0.21	insig
gender, age	0.37	insig
hrsperweek, age	0.35	0.56
social-role, marital-status	0.21	0.56
educ-num, education	0.37	0.52
education, hrsperweek	insig	0.66
age, marital-status	insig	0.58
education, occupation	insig	0.59
occupation, hrsperweek	insig	0.67
occupation, educ-num	insig	0.63
occupation, work-class	insig	0.63
country, race	insig	0.56
work-class, educ-num	insig	0.57

TABLE IV: Candidate groups considered as correlated attributes either by MB or by CB

In contrast to that, the trends of CB are not related to population but to people's common judgment on what makes sense for them as valid attribute values combinations. Results in Table IV confirm our theory/assumptions and prove the existence of correlations between some attributes which cannot be captured by pure learning from or statistical analysis of the data. Indeed, CB method has identified eight of such correlations that MB considered insignificant, i.e., the last group of rows in Table IV, to which we refer hereafter as CAG_{CB}.

Value of learned correlations. To better understand and validate the value of our CB learning, we compare the usefulness of CAG_{CB} to the one of CAG_{MB} in the estimation of records' trustworthiness. For this purpose, we have considered the remaining 1/3 of the records in the Adults dataset and we have randomly scrambled 50%

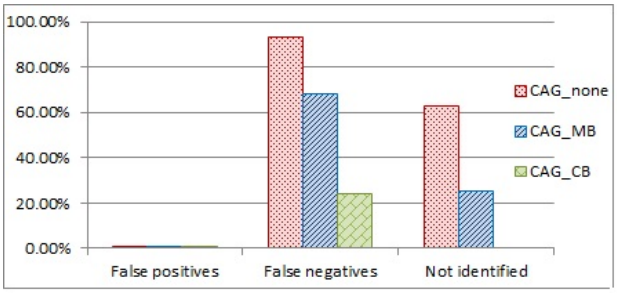

Fig. 2: False positives, false negatives, and not identified profiles under CAG_{CB}, CAG_{MB}, and CAG_{none}

Attribute	Multi-Value
Gender	No
Religious Views	No
Work Place	Yes
Work Location	Yes
Country	No
Education (Ed. Major)	No
Sports	Yes
Pref. Music	Yes
Pref. Movies	Yes
Pref. Books	Yes
Likes (liked pages)	Yes
Groups (joined groups)	Yes

TABLE V: Adopted Profile Schema - OSN dataset

of them to simulate fake profiles.[7] The scrambled records were labeled as fake (F), whereas the original ones as real (R). The union of these real and fake records makes our validation dataset.

We made available a new survey through which we asked participants to rate records from the validation dataset based on CAG_{CB}, CAG_{MB}, and CAG_{none} (this refers to the groups which did not pass the support neither by CB or MB) independently. We compute the resulting ITL of every record under each of the three scenarios. The record is estimated real if ITL is positive. It is estimated fake if ITL is negative. Some profiles could not be identified by the raters as fake or real (i.e., their ITL approximates 0). We refer to these by *NI* - Not-Identified category. Figure 2 summarizes the obtained results; it shows the percentages of false positives, false negatives, and NI category achieved under each of the three scenarios.

We can clearly see on Figure 2 that the reliability of rating records based on CAG_{CB} is the highest compared to CAG_{MB} and to CAG_{none} for all the cases. For example, participants incorrectly rated only 24 % of fake profiles (false negatives) using CAG_{CB} against 68% using CAG_{MB}. By using CAG_{CB}, we also obtain the lowest value for not identified profiles, which means that raters could give an opinion, positive or negative, on almost all the questions they received on CAG_{CB}.

To sum-up, this first experiment proves both utility and value of our method in learning correlated groups which cannot be determined by machine based techniques and which are more significant in reliably rating profiles. Our method successfully passes the test against machine learning under best conditions for this latter; i.e. a categorical sanitized dataset. Such conditions are not expected under a real OSN environment in which users insert free text using different semantics and extensively varied typing patterns.

B. Performance within Real Environment

In this second experiment, we study the feasibility and effectiveness of our CB method within an OSN environment. We choose Al Akhawayn University (AUI)[8] as our study group. In addition, we have opted for Facebook as one of the major, most popular, and widely used social networks. The choice of one community in this experiment is mainly aimed to maximize chances for knowledgeable feedback.

1) The Dataset: With their consensus, we collected the Facebook profile data of 70 alumni students from the cohorts of 2011 and 2012 (see Table V for the collected profile attributes).[9] Our 70 profiles have 64% females and 36% males with approximated average age of 23.[10] These 70 profiles made our training dataset for the learning part of the model.

Some attributes have one dominant value in our training profiles set, such as the attribute Religious Views for which 99% of our training profiles had an equal value, and hence they have not been considered in the definition of candidate groups.[11] We obtained 14 possible combinations of size 2 over the experiment's profile schema to be our candidate groups (CA).

2) Experimental Settings and Design:

Trusted Users. We had 35 participants for the *Trusted Users* group[12] in our model. These are all current students of AUI attending different sections of the same course. 49% of them are females and 51% are males with an average age of 19.5. They all took an online survey in which they answered between 34 and 50 feedback questions on our 14 candidate groups.

Feedback Questions. We set feedback questions as specified under Section III.A. For those attributes accepting multi-values, such as Preferred Music, their different

[7]Fake profiles are made by setting their attributes with values randomly selected from the available records and by ensuring that a fake profile does not have two values taken from the same original record.

[8]AUI is a Moroccan University operating under the American model for education. It offers all living facilities within its campus and its students live in there as a community: www.aui.ma

[9]The second column specifies whether the attribute accepts multiple values or not.

[10]The average age corresponds to the censused mean age of AUI graduating students in 2011 cohort.

[11]Considering one-value-dominant attribute in our candidate groups seemed meaningless as we cannot learn correlations given lack of diversity in values.

[12]The 35 participants are considered trusted because they come from the same community, they understood the objective of the experiment, and they engaged to take the survey with due care and attention to its questions.

Candidate Group (CA)	Support
Gender, Movies	0.72
Ed. Major, Groups	0.67
Gender, Sports	0.65
Groups, Likes	0.61
Pref. Music, Pref. Movies	0.57
Pref. Movies, Pref. Books	0.44
Ed. Major, Likes	0.43
Sports, Likes	0.36
Pref. Music, Gender	0.31
Ed. Major, Sports	0.31
Pref. Music, Pref. Books	0.28
Gender, Pref. Books	0.23
Ed. Major, Books	0.20
Ed. Major, Movies	0.11

TABLE VI: Achieved support per candidate group

values have been considered one at a time when computing the support of groups containing them. For example, if a profile has two values for Preferred Music, *music1* and *music2*, and *female* as value for Gender attribute, then the support of the candidate group {Gender, Preferred Music} was computed considering feedback on [*female, music1*] and [*female, music2*] as two independent combinations.

Evaluation. We have considered the 35 trusted users as the set of target users to be evaluated. In order to make a sound testing dataset, and since these profiles correspond to real identities (i.e. most probably they are all coherent), we created 20 fake profiles out of the 35 real ones. This was done by assigning to each attribute in a fake profile a randomly selected value from a real one, ensuring that one fake profile will never have two attribute values taken from the same real profile. We labeled the testing profiles as RP for the real 35 and as FP for the 20 fake ones. We got 13 raters[13] to evaluate the 55 profiles in our testing dataset.

3) Achieved Results:

Defined Correlated Attribute Groups. Table VI presents the supports received for all the 14 candidate correlated attribute groups (CAs) considered in the experiment. Considering a support threshold of *sh = 0.30*, the 10 first groups in Table VI are the correlated attribute groups (CAG) considered in the experiment.

Defined Coherence Relations. In order to detect the coherence relations in the considered CAGs, the confidence measures were computed. Figure 3 shows these values presenting a comparison between the two possible coherence relations out of each one of them. For example, for the CAG = {'Gender', 'Pref. Movies'}, the two possible coherence relations are ['Gender' $\Longrightarrow$ 'Pref. Movies'] and ['Pref. Movies' $\Longrightarrow$ 'Gender'].

A first remark from Figure 3 is that in all CAGs the two confidence measures are not equal. This implies that a direction in the coherence relation can be always detected. More importantly, this difference is more relevant in some CAGs than in others. In fact, on Figure 3, we can point to 4 specific CAGs for which the difference between the two coherence relations is considerable (i.e., this difference exceeds 0.2 for the 4 of them). These are {'Ed. Major', 'Groups'}, {'Ed. Major', 'Likes'}, {'Movies', 'Gender'}, and {'Ed. Major', 'Sports'}. What is flashing in these 4 CAGs is that 3 of them contain the attribute 'Ed. Major' and that this latter is in the right side of the dominant coherence relation over the 3 of them. We discuss this further under Section V.C.

Evaluation of Profiles. In order to evaluate the utility of raters selection, we performed profile evaluation with and without it.

Scenario 1 - Raters Selection Free. We dropped raters selection and we processed the rates provided by our 13 raters on each of the 55 testing profiles.

Scenario 2 - Raters Selection Applied. We apply rater selection with regard to the 4 most important coherence relations out of the ones mentioned in Figure 3. These are: ['Groups' $\Longrightarrow$ 'Ed. Major'], ['Likes' $\Longrightarrow$ 'Ed. Major'], ['Pref. Movies' $\Longrightarrow$ 'Gender'], and ['Sports' $\Longrightarrow$ 'Ed. Major']. Our 13 raters have been categorized by the determinant attribute (i.e., the left attribute) in each of the 4 considered coherence relations.

Under each of these two scenarios, an estimated ITL is computed for every profile in the testing dataset. Based on the computed ITL, the profile is classified as positively rated (positive ITL), negatively rated (negative ITL), or neutrally rated (ITL approximates 0). Recalling that all testing profiles are preliminary labeled as real (RP) or as fake (FP), we categorize our achieved estimations under 5 groups: 1. the FPs negatively rated (FP-F), 2. the RPs positively rated (RP-R), 3. the RPs negatively rated (RP-F), 4. the FPs positively rated (FP-R), and 5. the RPs and FPs neutrally rated (NR). Figure 4 provides the number of profiles (as percentages) under each of these five categories both when raters selection is applied and when it is not.

The first thing we learn from Figure 4 is the accuracy of our method in correctly rating profiles. For instance, more than 60% of fake profiles and more than 95% of real one are correctly estimated when raters selection was not applied. In addition to that, we clearly notice how raters selection improves the scores under all the five categories. Most importantly, raters selection considerably increased the number of profiles in the FP-F category and decreased the FP-R one. In the first case, raters selection correctly rated 85% of fake profiles against 65% only in the other scenario. In the second case, raters selection minimized the error in incorrectly estimating fake profiles by 5%. Moreover, raters selection minimized the NR category by over 7%.

C. Extended Discussions

Our two initial experiments, each in a different context and from a different perspective, showed how our method draws from and harnesses the wisdom of the community to reliably estimate the trustworthiness of OSN claimed

[13]These 13 raters come from the group of whom we collected the 70 profiles for the training dataset.

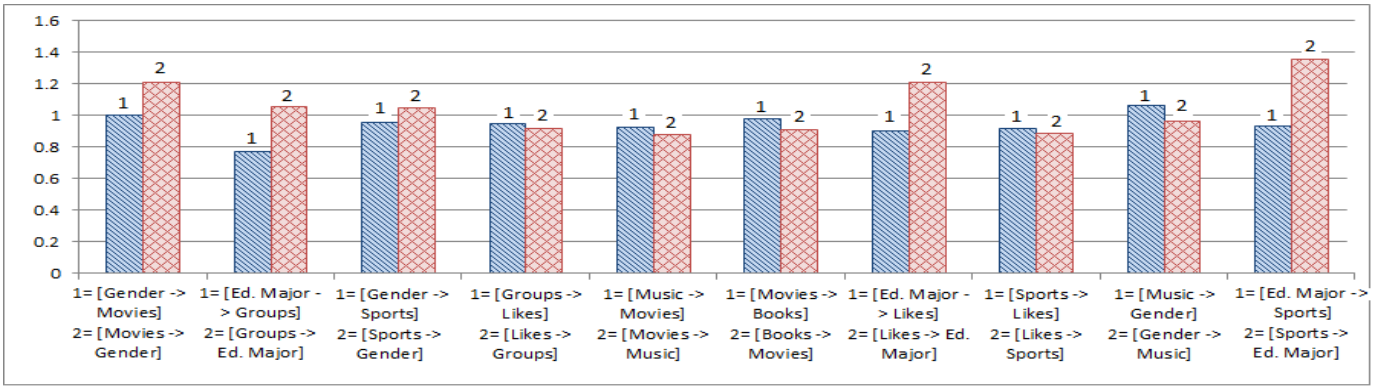

Fig. 3: Coherence Relations for the 10 defined CAGs

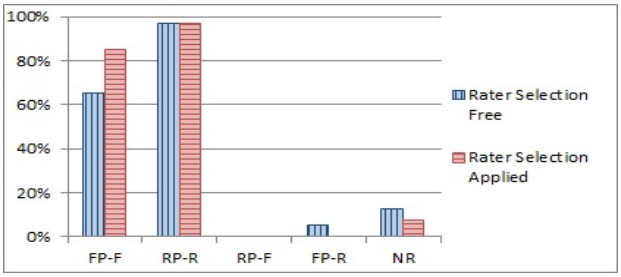

Fig. 4: Evaluation of profiles with rater selection applied vs. rater selection free

identities. In the first experiment, we learned that there are correlations between some profile attributes which cannot be detected by pure machine learning only, even when the training dataset is optimized for this latter. Relying on community-sourcing detects these correlations which we proved have higher effectiveness in reliably rating profiles. Overall, the results of the first experiment justify our method and approve its reliance on community feedback to achieve its objective. Still confirming the same, the second experiment proved that our method can efficiently and effectively stretch to the specificity of an OSN arena.

In parallel to confirming our method, the experiments provided other prominent lessons opening interesting opportunities for future extensions of the work. For example, the coherence relations detected in the second experiment might be interpreted as specific to the experiment's community. In fact, the observation made earlier with regard to the Ed.Major attribute which exists in almost all the strong coherence relations tells us that correlated attribute groups and their coherence relations are bounded within social communities. This is an inherent result given that social norms and social configurations are commonly known to be community and social-groups specific (each community has its own culture and identity which differs from the other). This point brings us to understand that our method is community dependent and shall consider constructing its learning phase within identified communities and not over all the OSN community as one single entity.

VI. Related Work

In general, the works aiming at confirming the identity of a user can be grouped under two categories. The first category are those aiming at giving users a mean to judge the reliability of their online peers mostly by relying on previous transactions or on existing relationships between them [12][13]. A prominent example of these are the work on collaborative filtering for people to people recommendation [12]. Another example is given in [13] where they elicit the opinion of a user's friends on an OSN on her claimed identity attributes on some other system. Our work is similar to these in the sense that we both rely on user feedback, but the core difference is that our work is not based on interactions between users or on any other kind of relationship between them.

The second category relies on machine processing/mining of the user's data and/or on their behavioral traces. Under this category, we find works basing identification on the analysis of biometrics [15] or different types of fingerprints, such as typing patterns [16], chatting patterns [14], etc. Other works suggest methods for inferring some profile attributes from social behavioral traces, such as in [20], where authors infer identities of users from the history of their likes. Others make use of mining techniques to identify users' unique patterns, such as in [17]. These work, however, have not been designed with the explicit aim of ensuring reliable identification of users' online identities.

In a different approach, some works investigated the possibility of determining the user's identity before accepting her/his registration to the service, such as in [5] where authors proposed an identity validation following a theoretical game model. In this work, the identity is prevalidated before accepting the new member, by measuring benefits of accepting her/him against risks she/he will introduce. Although such an approach can be useful in some scenarios, we believe it does not fit with one of a general purpose OSN where censure is not expected.

Some deployed methods rely on hard identification mechanisms via confirming the address by sending code-

embedded post-cards for example, or via requesting the payment of a symbolic 1$ using a valid credit card [couchsurfing.com], or even through requesting scanned copies of identity cards issued by the state they belong to [airbnb.com] to gain an identity verified tag. While the efficiency of these methods is not to be discussed, their feasibility and appropriateness to an open, general purpose OSN is on the edge.

Finally, one of the works which is very close to our method in the sense of considering coherence within profile information to infer identity trustworthiness is the one in [18]. However, the authors focus on online business identity and suggest a theoretical only approach as an extension to the TOGAF framework [19]. They rely on trusted authorities opinion on the validity of information provided in a business profile to judge its veracity. In contrast, we are using the community to provide opinion on the validity of users' profiles on an OSN, but through a practically proved method.

VII. Conclusion and Future Work

In this paper, we have proposed a new approach for estimating identity trustworthiness levels for target profiles on OSNs using community feedback. We base this estimation on a study of the coherence of target profiles w.r.t. correlated attribute groups. The initial experiments performed prove the meaningfulness of our suggested method and its effectiveness in correctly rating target profiles and justify its reliance on human feedback. We plan to extend this work on different dimensions. First, we plan to run the method over larger scopes within OSN data to fine-tune it for sub-communities, learning stopping conditions, and other system parameters. We also plan to investigate incentive mechanisms to improve the engagement of the community in the process and to make OSN users complete their profiles (non complete profiles will get low values for ITL).

Second, we consider enriching our approach by adding trust levels of raters as weights for their provided rates, addressing the colluding friends phenomena, and also ordering the correlated groups by their importance and strength in increasing the reliability of ITLs. In addition, we plan to better address the privacy related issues beyond exclusion of quasi-identifiers to ensure privacy guarantees and safety properties for all the stakeholders of the system. Moreover, we plan to consider a design for the suggested method over a decentralized architecture for OSNs.

VIII. Acknowledgment

This work is partially supported by the iSocial EU Marie Curie ITN project (FP7-PEOPLE-2012-ITN).

References

[1] Vandell, Deborah. L, *Parents, peer groups, and other socializing influences.* Development Psychology, Vol 36(6). 2000.

[2] Pascal. S, Joachim. G, *Organizational Virtualness.* Proc. of the VoNet - Workshop. 1998.

[3] C. L. Corritoire, B. Kracher, S. Wiedenbeck, *On-line trust: concepts, evolving themes, a model.* International Journal of Human-Computer Studies, Vol. 58, no. 6, pp. 737 - 758. 2003.

[4] H. Nissennbaum, *Securing trust online: Wisdom or Oxymoron.* BUL Rev, Vol. 81. 2001.

[5] Squicciarini. A. C, Griffin. C, Sundareswaran. S, *Towards a Game Theoretical Model for Identity Validation in Social Network Sites.* IEEE International Conference on Privacy, Security, Risk, and Trust. 2011.

[6] Michael J. A. Berry, Gordon S. Linoff, *Data Mining Techniques.* Copyright 1997 by John Wiley and Sons. ISBN 0-471-47064-3. 2007.

[7] S. de Capitani di Vimercati, S. Foresti, *Quasi-Identifier.* Encyclopedia of Cryptography and Security: SpringerReference. [Online].

[8] R. Agrawal, R. Srikant, *Fast Algorithms for Mining Association Rules.* IBM Almaden Research Center. Proc. of the 20th VLDB Conference Santiago, Chile. 1994.

[9] B. Viswanath, M. Mondal, A. Clement, P. Druschel, K O. Gummadi, A. Mislove, A. Post, *Exploring the design space of social network-based Sybil defenses.* IEEE Fourth International Conference on Communication Systems and Networks (COSNETS). 2012.

[10] G. Guette, B. Ducourthial, *On the Sybil attack detection in VANET.* IEEE International Conference on Mobile Adhoc and Sensor Systems (MASS). 2007.

[11] L. Jin, H. Takabi, J B.D. Joshi, *Towards Active Detection of Identity Clone Attacks on Online Social Networks.* Proc. of the first ACM conference on Data and Application Security and Privacy (CODAPSY). 2011.

[12] X. Cai, M. Bain, A. Krzywicki, W. Wobcke, Y. S. kim, P. Compton, A. Mahidadia, *Collaborative Filtering for People to People Recommendation in Social Networks.* Advances in Artificial Intelligence: Lecture Notes in Computer SCience Vol, 6464. SpringerLink. 2011.

[13] M. Sirivianos, K. Kim, J. W. Gan, Yang, *Assessing the veracity of identity assertions via OSNs.* IEEE 4th International Conference on Communication Systems and Networks (COMSNETS). 2012.

[14] Roffo, Giorgio, Segalin, Cristina, Vinciarelli, Alessandro, Murino, Vittorio, al. *Reading between the turns: Statistical modeling for identity recognition and verification in chats.* IEEE 10th International Conference on Advanced Video and Signal Based Surveillance (AVSS). 2013.

[15] L. Dehache, L. Souici-Meslati, *A multibiometric system for identity verification based on fingerprints and signatures.* IEEE International Conference on Complex Systems (ICCS). 2012.

[16] P. Chairunnanda, D. R. Cheriton, N. Pham, U. Hengartner, *Privacy: Gone with the Typing! identifying Web Users by Their Typing Patterns.* IEEE 3rd International Conference on Social Computing. 2011.

[17] A. S. Bozkir, S. G. Mazman, E. A. Sezer, *identification of User Patterns in Social Networks by Data Mining Techniques: Facebook Case.* Technological Convergence and Social Networks in Information Management, Vol, 96. 2010.

[18] Y. Yang, E. Lewis, J. Newmarch, *Profile-based digital identity management - a better way to combat fraud.* IEEE International Symposium on Technology and Society (ISTAS). 2010.

[19] TOGAF, *TOGAF version 9.1.* [Online] available at: www.togaf.org. 2013.

[20] M. Kosinski, D. Stillwell, T. Graepel, *Private traits and attributes are predictable from digital records of human behavior.* Proc. of the National Academy of Sciences of the USA. [online]. 2012.

[21] S. Vijayarani, A. Tamilarasi, M. Sampoorna, *Analysis of privacy preserving k-anonymity methods and techniques.* IEEE International Conference on Communication and Computational Intelligence. 2010.

[22] Elli P. Schachter, *Identity configurations: A new perspective on identity formation in contemporary society.* Journal of Personality 72 (1), 167-200. 2004.

[23] SJ. Schwartz, K. Luyckx, VL. Vignoles, *Handbook of Identity Theory and Research, Volume 1.* Springer. 2011.