2012 IEEE 28th International Conference on Data Engineering

Privacy in Social Networks: How Risky is Your Social Graph?

Cuneyt Gurcan Akcora, Barbara Carminati, Elena Ferrari

DICOM, Università degli Studi dell'Insubria
Via Mazzini 5, Varese, Italy
{cuneyt.akcora, elena.ferrari, barbara.carminati}@uninsubria.it

***Abstract*—Several efforts have been made for more privacy aware Online Social Networks (OSNs) to protect personal data against various privacy threats. However, despite the relevance of these proposals, we believe there is still the lack of a conceptual model on top of which privacy tools have to be designed. Central to this model should be the concept of *risk*. Therefore, in this paper, we propose a risk measure for OSNs. The aim is to associate a risk level with social network users in order to provide other users with a measure of how much it might be risky, in terms of disclosure of private information, to have interactions with them. We compute risk levels based on similarity and benefit measures, by also taking into account the user risk attitudes. In particular, we adopt an active learning approach for risk estimation, where user risk attitude is learned from few required user interactions. The risk estimation process discussed in this paper has been developed into a Facebook application and tested on real data. The experiments show the effectiveness of our proposal.**

I. Introduction

Having an account in an Online Social Network (OSN) opens a path to opportunities but it also brings about certain risks; social network users can be bullied, their pictures can be stolen or their status posts can reach unwanted audiences. Even when profiles do not list any information, social graphs can be analyzed to infer personal information. Despite all these, social networks have hundreds of millions of users, because users think that positives outweigh negatives and they maintain their online presence. However, this user attitude is not totally care-free; individual cases of privacy breaches and their consequences have been widely discussed in social media and privacy risks have only grown with time as social networks have grown in size exponentially [1].

Several research efforts have been carried out to alleviate these problems, with the results of some tools helping users to be more privacy-aware. Notable examples are relationship-based access control mechanisms [2], tools on support of privacy preference specification [3], [4], as well as more expressive privacy settings recently adopted by commercial OSNs, like Facebook. However, despite the relevance of these proposals, we believe there is still the lack of a conceptual model on top of which privacy tools have to be designed. Central to this model, that should be the basis for any conscious user decision about profile information disclosure, should be the concept of *risk* one might be exposed when interacting with other OSN users.

The introduction of this concept, new in OSNs, is mainly motivated by the consideration that a large amount of friendships in social networks are just virtual. Indeed, people connected in an OSN very often have never met before establishing a relationship (or a long period has passed before they keep in touch again in an OSN). However, several events reported in the media testify that users establish virtual relationships with never-met people without considering the consequences these new relationships might have for their privacy and, some time, personal safety. Indeed, a new relationship in an OSN always implies the release of some personal information. In general, this release is controlled by user privacy settings. Unfortunately, empirical studies show that social network users are not used to specifying privacy settings, and very often they do not change the default privacy settings that are very permissive [5], [6]. As a consequence, the creation of new relationships without specifying the appropriate privacy settings might expose a user to the risk of unconscious release of personal data. This is due to the fact that in almost all OSNs users can reference resources of other users in their social graph; and it is generally not possible or very difficult for a user to control the resources published by another user. This uncontrolled information flow highlights the fact that creating a new relationship might expose one to some privacy risk. In addition, as said before, new relationships with never-met users might have most serious consequences, since social networks have drawn the attention of sexual predators and child molesters, just to mention the worst.

Despite the possible serious consequences of the establishment of new relationships, current situation is that social network users are poorly informed about their friends, and with whom their friends keep company. Hugely popular social networks still think that a friend is someone with whom a user can share everything, and furthermore, that the user can even share most of his/her profile information with friends of friends. Recommended privacy settings are in accordance with this friend notion. For instance, on Facebook, friends of friends are by default allowed to see most parts of a user profile.

To cope with these issues, in this paper we investigate a risk measure to help users in judging a stranger (i.e., a never-met user) with whom they might establish a relationship so as to be informed of how much it might potentially be risky, in terms of disclosure of private information, to have interactions with him/her. In defining the proposed measure we made several design choices, inspired by real cases. The first is that risk evaluation is impacted by several different dimensions, whose importance may greatly vary from user to user. Indeed, different from other scenarios where user risk evaluation can be made by a unique trusted entity (e.g., a bank,

1084-4627/12 $26.00 © 2012 IEEE
DOI 10.1109/ICDE.2012.99

an insurance company) by processing profile information and users' past activities, we believe that in OSNs user risk attitude should also be taken into consideration. As such, we adopt a solution closer to the social computing paradigm, where rather than a trust entity judging social network users, we prefer the judgments given by community members. This paradigm has resulted to be a winner choice in other communities, where knowledge deriving from the community is used to take decisions, such as recommendation systems, collaborative filtering, and so on. To this end, in this paper we try to assess the risk of a stranger through the eyes of the user, hence the risk score is user-dependent and subjective. Thus, our risk measure captures the personal judgment that a user, hereafter called owner, expresses on a stranger Y with respect to his/her risk. However, to help owners in forming their own opinion about risk, we provide them with objective information about strangers. More precisely, inspired by homophily and heterophily theories, we provide information on *similarity* between owner and the stranger as well as the *benefits* owner might have in terms of new information from strangers profile he/she might be authorized to access. We believe that, by using this information, users are able to evaluate the trade-off between similarities and benefits so as to estimate the risk of a stranger.

Other design choices for our risk measure are motivated by social network statistics, which show that new relationships are mainly established among contacts of direct friends (i.e., second level friends). As such, we focus on second hop connections which act as strangers in the rest of the paper. Statistics also show that this set of users can be very large (as an example, according to Facebook statistics each user has 16,900 strangers on the average). This necessitates a prediction technique, since asking an owner to insert risk judgments for each stranger is unfeasible. To this purpose, in this paper we adopt an active learning process for risk estimation, where owner risk attitude is learned from few required interactions. According to the proposed learning algorithm, once the classifier is built with the training data (i.e., the owner risk labels), the system can predict risk labels of all those strangers that were not included in the training data without any user intervention.

The risk estimation process discussed in this paper has been developed into a Facebook application and tested on real Facebook data. Our experiments show that with limited user involvement, we can get accurate risk estimation. In particular, as experiments in Section IV show, we are able to correctly predict risk labels with 83.38% of accuracy. Furthermore, in Section IV we discuss how user decisions in assigning risk levels are affected by the characteristics of other social network users.

To the best of our knowledge, we are the first to provide a framework for computing OSN users' privacy risks which takes into account a rich set of dimensions, ranging from structural properties of the social graph and profile similarity, to benefits arising from user interactions and subjective risk perception.

The remainder paper is organized as follows. We discuss our risk measure in Section II, whereas Section III details the three phases of the risk learning process. In Section IV we discuss the performance of risk computation. Section V covers related work. Finally, Section VI concludes the paper.

II. How to measure risk

In designing the proposed risk measure we take into account several design principles that have been inspired from social network theory. In the following, we discuss these principles and our assumptions, whereas in Section III we present our approach to estimate risk levels according to them.

Strangers. As mentioned in the introduction, social network users tend to interact and create new relationships with those that are close to them in the social network graph [7], [8]. As an example of this trend, experiments in [9] show that 80% of new Facebook relationships are created between users and contacts of their friends. For this reason, we limit risk estimation to second-level contacts. More precisely, given a social network user, hereafter $owner$, we compute risk levels for those users that are connected to a friend of owner's friends. In the following, we will refer to these second-level contacts as $strangers$.

Subjective risk perception. Risk attitude has been found to be very subjective [10], [11]. As such, we do not believe that it is possible to automatically estimate the risk just on the basis of the characteristics of a user social graph and/or strangers' profiles. In contrast, we believe that we need to take into account also users' risk attitude. To cope with this issue, during the first phase of our risk estimation process we ask for owner feedback. More precisely, we ask users to provide risk judgments for few selected strangers from the social graph. Then, from the collected risk judgments, the process learns how to estimate risk levels of the remaining strangers.

Risk judgment. Since we want to estimate how much it might be risky to interact with a stranger, the risk notion we want to model is related to social interactions. We believe that to receive an educated risk estimation for a given stranger, we should provide the owner with meaningful information that can help him/her to make a correct decision. To determine what kind of information should be provided to owners, we take into account two key tendencies that, according to social network theory, regulate social interactions.

Homophily. This is the tendency of people to interact with those that are similar to them [7]. In social networks, users who are similar along certain attributes, such as gender, education and hometown, tend to create new friendship links. Furthermore, as previously discussed, users create new friendship links with others who are closer to them on the social graph.

To take homophily into account, we provide owner with information on how much the stranger, for which the risk judgment is required, is similar to him/her. This information is provided by means of a 'similarity measure'. It is worth

noting that literature offers several similarity measures [12], in this paper we adopt a similarity function that considers both network similarity (e.g., information on owner and strangers mutual friends), as well as profile similarity (e.g., how many common/similar profile attributes they have).

Benefits. In contrast to homophily, heterophily is the tendency to interact with people that are not similar to you [13]. In social networks, this tendency is mainly motivated by the fact that social interactions (e.g., creation of new relationships with a stranger) might give owner some benefits in terms of acquaintance of new information (e.g., strangers profile and social graph).

An illustrious example of heterophily can be observed on the social network Twitter. Twitter users do not need approval from others before they can 'follow' them and view their status posts. This allows users to follow people whenever they can get benefits, and, in many cases, followed users are distinctly different from the followers in their profile and social graph. We wish to keep also this tendency into account, and therefore we provide owner with information on which benefits he/she might get in starting social interactions with the stranger.

To have a measure of benefits, the owner assigns an importance coefficient θ_i for each benefit item i, whereas the benefit an owner o gets from a stranger s is defined as:

$$B(o,s) \;=\; \frac{1}{|M|} \times \sum_{i \in M} (\theta_i \times V_s(i,o))$$

where M is the set of benefit items on the profile of s (e.g., photos, friends, wall), θ_i is the importance of being able to see item i, and $V_s(i,o) = 1$, if item i of the profile of s is visible to the owner, $V_s(i,o) = 0$, otherwise.

III. Risk Learning Process

Our risk learning process relies on owner feedback to take into account the subjective nature of risk. Feedback is gathered as owner feeling/estimation about how much it could be risky to start social interactions with a given stranger. Since the owner might not know the stranger, risk estimation is based both on the similarity the stranger has with the owner and the benefits he/she might provide to the owner (cfr. the principles discussed in the previous section). Collected risk judgments, hereafter *risk labels*, are then used to learn owner risk attitude and predict risk estimation for those strangers for which the owner does not provide a direct judgment.

The proposed risk learning process is based on supervised learning techniques [14], where, by analyzing training instances (i.e., few labeled instances) it is possible to generate a classifier to predict labels for unlabeled instances. In designing the supervised learning process to be adopted for risk estimation we have considered two main issues. The first is the potentially big number of strangers an owner might have. As an example, in our experiments the average count of strangers per owner is 3,661.

Another important factor is the dynamic nature of the owner's social graph, in that stranger connections might change very fast. These changes are due to acquisition of new strangers (i.e., some friends create new contacts) as well as new connections between strangers themselves, which might impact their similarity measures with the owner. This means that in our scenario it is not efficient to adopt a pre-defined and fixed training set. Rather, it is preferable to select the training set on the fly so that changes in the social graph are immediately reflected, and the number of required data to be labeled (i.e., strangers) can be reduced as the accuracy of predictions improves. Due to these requirements, from the many existing supervised learning models, we choose the active learning paradigm [15], where the learning phase is computed in several rounds. During each round an annotator (i.e., the owner) is interactively queried for labels (i.e., risk judgments) of selected unlabeled data samples (i.e., unlabeled strangers).

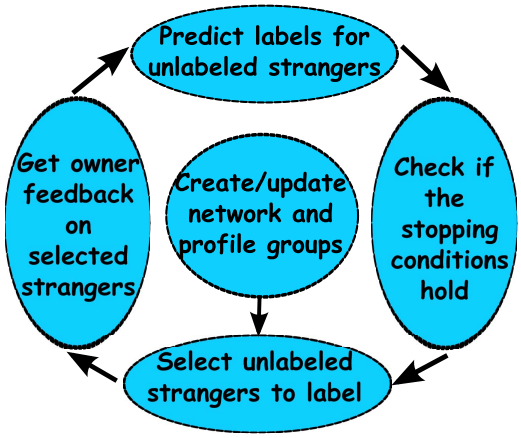

Fig. 1. Risk learning process

Using active learning in our context implies that initially all strangers belong to the set of unlabeled strangers. In each round, the risk learning process asks the owner to give risk labels to a selection of the most informative strangers[1] among those not labeled, which are then moved into the labeled strangers set. Then, collected labels are used by a classifier to predict labels for unlabeled strangers. Rounds of labeling and prediction continue until a good level of prediction is met, where the required level is given as input by the owner to take into account different risk attitude. The learning process is summarized in Figure 1.

To adopt active learning in our scenario, we need to cope with several issues. (1) In order to really get the owner risk attitude, the learning process has to pose an appropriate question to the owner, so as to get a meaningful feedback with respect to his/her risk perception. Possible risk label values have to be carefully designed too. (2) Particular attention has to be paid to the strategy for the most informative strangers selection, which should be tailored to the social network scenario. (3) A good classifier should be chosen to predict risk

[1]Most informative strangers are selected according to a clustering-based approach explained in Section III-B.

labels for unlabeled strangers. (4) Finally, we need to design stopping criteria, which ensure the required accuracy.

All the above issues are discussed in the rest of the section.

A. Risk Labels

In order to explain owners that our risk estimation is based on similarity and benefits, we query the owner for the risk level of a stranger with the following question:

"You and *stranger_name* are $x/100$ similar and he/she provides you $y/100$ benefits in terms of information you are allowed to see now on his/her Facebook profile. Do you think it might be risky to establish a relationship with *stranger_name*? Please respond by considering how much you are similar to *stranger_name* and that, after you become friends of him/her, benefits might increase as you might be allowed to see more resources in addition to his/her profile, e.g., his/her posts, photos, if privacy settings allow you."

Rather than allowing owners to select any value in [0,1], we give them only three options for risk labels, namely *very risky*=3 , *risky*=2, and *not risky*=1. We believe that these three values are more easily understandable by average OSN users.

B. Stranger sampling

The literature offers several methods for sampling selection (see e.g., [15] for a survey). Among these, we are interested in pool-based criteria according to which data to be labeled in each round are greedily selected from the pool of unlabeled strangers.

Our idea is to cluster strangers based on their network connections and profiles, so as to generate a set of pools. Thus, the set of strangers is divided into a set of disjoint pools $\mathcal{P}_{st}$ from which a number of strangers are randomly selected at each round i to be labeled by the owner. Therefore, at each round i each pool $P \in \mathcal{P}_{st}$ has associated the corresponding set of labeled strangers L^i_{st} as well as the gathered labels L^i, whereas U^i_{st} denotes the set of unlabeled strangers. Then, based on labels in L^i the classifier predicts labels for unlabeled strangers in U^i_{st}. The process is iterated until the required accuracy in labeling prediction is met in the corresponding pool P.

Pools are defined by clustering strangers according to two dimensions. The first takes into account how much a stranger is connected to the owner in the network. Therefore, we create a first-level grouping based on strangers' network similarity with the owner as follows:

***Definition** 1 (Network similarity groups):* Let S_o be the set of strangers for a given owner o. Let NS be a function returning the network similarity between o and $s \in S_o$, $NS(o,s) \in$ [0,1]. The Network Similarity Groups for o (NSG) consists of α disjoint sets of strangers $nsg_1, \ldots, nsg_\alpha$, such that for each $x \in \{1, \ldots, \alpha\}$, $nsg_x = \{s_i \in S_o \mid \frac{x-1}{\alpha} \leq NS(o,s_i) < \frac{x}{\alpha}$ and $0 < x \leq \alpha\}$.[2]

[2]We defer the discussion on α value to the experimental results section.

To estimate the network similarity between owner o and a stranger s, we adopt the measure NS() defined in [9]. Unlike existing similarity measures [12] which only consider mutual friends of the owner and a stranger, the measure works by also considering the connections among mutual friends. If the stranger is connected to a dense community around the owner, the measure returns a higher similarity value.

Additional knowledge that can help in defining stranger pools is the one contained in OSN user profiles. Indeed, these can be used to further refine network similarity groups, so as to cluster similar strangers together, which we expect to improve label prediction accuracy. Therefore, the first-level groups of strangers are based on network similarity with the owner, whereas the second-level groups of strangers within the same network similarity group are determined according to their mutual profile similarity.

In selecting the algorithm for profile-based clustering [16] that better fits our scenario, we have to take into account that (1) social network profiles mainly contain categorical data (such as hometown, gender and education); (2) it might be required to perform profile-based clustering several times, since we need to generate separate profile clusters for each network similarity group. Due to these requirements, we selected the Squeezer algorithm [17]. This is a clustering algorithm for categorical data that makes only one pass on the data set, and this helps us to deal with computational complexity issues when there are thousands of strangers in a network similarity group. A further benefit of Squeezer is that it allows us to customize clustering by associating different weights with each profile item (e.g., age, hometown, educations). As it will be discussed in the section reporting experiments, these weights help us in catching the relevance of some profile items over the others while grouping strangers. In order to apply Squeezer to our scenario, we have adapted it as follows. Given a network similarity group nsg, the algorithm starts with the first stranger as a profile similarity cluster psg and iterates over each stranger $s \in nsg$. An overall similarity of stranger s with profile similarity cluster psg is computed by weighted averaging all profile attribute similarities. More precisely, to find the similarity of a stranger to a psg, the algorithm checks how many strangers from psg have the same value as s for each profile attribute. Stranger s is put into the most similar cluster. If no similar cluster can be found, that is, the similarity is below a given threshold, the stranger s itself creates a new cluster. We have adapted the similarity measure defined in [17] for relational tuples, to profile attributes as follows.

***Definition** 2 (Profile Similarity):* Let S_o be the set of strangers for a given owner o. Let nsg be a network similarity group generated on S_o. Let c be a subset of strangers in nsg (i.e., a cluster) and s be a stranger such that $s \notin c$. Let PA be the set of profile attributes, we denote with $s.pa_i$ the value of the i-th attribute in s profile, where $i \in |PA|$. The profile

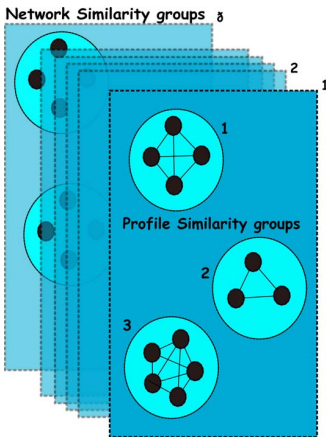

Fig. 2. Network and profile based pools

similarity of c and s is given by:

$$Sim(s,c) = \sum_{i \in |PA|} w_i \left(\frac{Sup(s.pa_i)}{\sum_{x \in VAL_{pa_i}(c)} Sup(x)} \right)$$

where w_i is the weight associated with the i-th profile attribute, $VAL_{pa_i}(c)$ is the set of values for attribute pa_i in the profiles of strangers in c, and $Sup(x) = |\{s_i \in c | s_i.pa_i = x\}|$ returns the support of attribute x.

By using the above profile similarity definition in Squeezer, we define the second-level grouping obtaining thus the set of network and profile based stranger pools $\mathcal{P}_{st}$, defined as follows.

***Definition** 3 (Network and profile based pools):* Let S_o be the set of strangers for a given owner o. Let α and β be two parameters. Let NSG be the set of α network similarity groups generated according to Definition 1. For each $nsg \in NSG$, let $Squeezer(nsg, \beta)$ be the set of profile-based clusters formed on nsg by algorithm Squeezer using the similarity measure in Definition 2 and β as similarity threshold to create a new cluster. The set of disjoined pools $\mathcal{P}_{st}$ on S_o is given by $\bigcup_{nsg \in NSG} Squeezer(nsg, \beta)$.

Figure 2 graphically depicts the resulting network and profile based pools.

Up to this point, we explained how strangers can be grouped so that sampling can be done. Next we will detail the prediction process which is performed over each distinct pool, and the related termination condition. Therefore, in what follows we explain the process for a given network and profile based pool $P \in \mathcal{P}_{st}$.

C. Classifier

In searching a classifier for our scenario, we focused on those tailored for network data. In this field, literature offers several classifiers, among these we selected the graph based approach by Zhu et al. [18]. This is a graph based classifier that works well with few labeled samples, and this complies with our need to reduce labeling efforts.

According to Zhu's approach, both labeled and unlabeled strangers are represented as nodes in a graph, where each pair of nodes is connected by a weighted edge. Weights represent the similarity according to the Euclidean similarity metric. The authors note that different similarity metrics can be used in different settings. In our scenario data are categorical, so we use edge weights based on strangers' profile similarity. To compute this similarity, we adopt the profile similarity function PS() defined in [9] which takes as input two profiles. For each attribute, if values are identical on both profiles the attribute similarity is set to 1. If they are non-identical, a non-zero value is computed by considering the frequency of the item values in the data set (i.e., the profiles in the considered pool). We refer the reader to [9] for more details.

After weights have been computed, the classifier predicts similar labels for similar neighbors on the graph, by exploiting the random walk strategy presented in [18].

In our privacy context, erroneous predictions can have two types of consequences; a stranger's label can be predicted lower or higher than what the owner would give. Higher label prediction poses no immediate threat to privacy; it only calls for more vigilance. On the other hand, lower prediction can have the system assume that the owner is safe when there is a real privacy threat.

D. Termination

In general, the risk learning process aims at soliciting owner labels till a good accuracy is reached. Accuracy can be estimated by comparing the predicted labels against the labels given by the owner. If the accuracy is low, we can ask owner to continue labeling. In the case of a good accuracy value, we can stop and spare owner effort in labeling. Since accuracy validation requires owner effort (i.e., owner labeling), we need to stop the process even if not all predicted labels have been validated against owner labels.

Several measures have been proposed to understand when to stop labeling when all predictions cannot be validated [19]. Among these, we chose the classification change principle. Classification change looks into predicted labels of strangers in consecutive rounds. If predicted labels do not change in these rounds, we conclude that further owner labeling will not change them as well and stop owner labeling. In what follows, we give more details on the adopted accuracy and stabilization measures.

Accuracy. The basic idea to estimate accuracy is to compute the root mean square error between labels predicted by the classifier and labels given by the owner for the same set of strangers. This error can be iteratively computed. More precisely, in each round r_i, the learning process asks the owner to give risk labels for some strangers s for which labels have been already predicted in previous round r_{i-1}. The error is computed as the root mean square between the predicted and owner-defined labels.

***Definition** 4 (Root Mean Square Error):* Let $P \in \mathcal{P}_{st}$ be a network and profile based pool, let U^i_{st} and L^i_{st} be the subsets of P denoting, respectively, the set of unlabeled strangers and the set of strangers for which the owner has given a risk label at round r_i. Moreover, let L^i be the set of collected owner labels for pool P till round r_i, and $\hat{L}^i_{st}$ be the set of predicted labels for strangers in U^i_{st}. Let S be a set of strangers selected from those that have been labeled at round r_{i+1} (i.e., from $L^{i+1}_{st} \setminus L^i_{st}$), where, for each $s \in S$, $\hat{L}(s)^i$ and $L(s)^{i+1}$ denote the label for s predicted at round r_i and the label given by owner at round r_{i+1}, respectively. Root mean square error of predictions for the set S is then:

$$RMSE_S = \sqrt{\frac{\sum_{s \in S}(L(s)^{i+1} - \hat{L}(s)^i)^2}{|S|}}$$

Since risk labels have values in the range [1,3] (i.e., very risky=3, risky=2, not risky=1), the root mean square error can have a value in [0,2], where 0 implies that all the predictions were correct.

Stabilization. For employing classification change in our unvalidated predictions, we solicit owner to give a confidence value $c \in$ [0,100], which is then used as a tolerance value for the classification change, as the following definition clarifies.

***Definition** 5 (Classification Change):* Let $P \in \mathcal{P}_{st}$ be a network and profile based pool, and let $\hat{L}(s)^i$ and $\hat{L}(s)^{i+1}$ be labels for stranger $s \in P$ predicted at round r_i and r_{i+1}, respectively. Let c be the confidence value selected by owner o. We say that P is stabilized according to confidence c if, for all $s \in P$, the following condition does not hold:

$$\hat{L}(s)^{i+1} - \hat{L}(s)^i \geq \frac{(L_{max} - L_{min}) \times (100 - c)}{100}$$

where L_{max} and L_{min} are the lower and upper bound of the labels range.

Note that, if the owner wants to manually label all strangers, the confidence value can be set as $c = 100$.

Stopping Condition. On the one hand, accuracy helps in validating label predictions, but it requires owner effort. On the other hand, stabilization in predicted labels does not guarantee accuracy in labeling; it merely informs us that further labeling by the owner will not change predicted labels. To keep into account both these aspects, we adopt a stopping condition that combines accuracy and stabilization. That is, we stop the risk learning process when risk labels are predicted with a good accuracy (i.e., RMSE between owner given and predicted labels has to be less than 0.5) and for at least n rounds there should be no classification changes with a confidence c selected by the owner, where n is a system-defined parameter.

IV. Experiments on Facebook data

In this section, we discuss the effectiveness of the proposed risk measure and related learning process. At this aim, we conducted several experiments to validate that our proposed risk measure indeed captures the risk perception of social network users and that the learning process correctly predicts strangers' risk labels. Before giving the details of experiments, we introduce the dataset that has been used.

A. Facebook Dataset

To perform the experiments on Facebook real data, we developed a Facebook application (called Sight[3]). The aim of this application is to: 1) gather strangers' information, so as to generate the network and profile based pools, 2) ask owner (i.e., the social network user that launched the application) risk labels only for selected strangers, on the basis of similarities and benefits. To perform this last task we developed a Google Chrome extension. Due to Facebook API limitations we cannot retrieve the complete social graph at once. Rather, we listen owner profile to see friends' interactions (e.g., tagging, posting) and, once a friend of friend is found, we query Facebook for its mutual friends/profile information. According to our experiments, the time period to learn a big portion of the social graph (4,000 strangers) can take up to 1 week. However, the user can start label and learn about the risk since the first day. In our experiments, in 2 months we were able to discover around 30,000 strangers.

Through Sight, owner can specify the benefit coefficients for each single item (e.g., profile attributes, wall, photos), as well as the confidence he/she wishes to reach. We list received benefits on the extension interface, along with the similarity value of the selected strangers, and ask the owner our query from Section III-A. A snapshot from the extension interface is shown in Figure 3.

During the test, Sight application has been used by 47 Facebook users, where 32 were males and 15 were females, all aged in the range [18-35]. 17 of the users were from Turkey, 5 from Italy, 9 from USA, 1 from India, and 7 from Poland.[4] We collected thus a total of 172,091 stranger profiles, and 4,013 owner-defined risk labels. On the average, owners have 3,661 strangers and gave 86 risk labels.

B. Parameters setting

The first couple of parameters are used to customize the creation of network and profile based pools. These are α and β. We recall that the first determines the number of network similarity groups computed using network similarity function NS(). As this function returns value in [0,1], for this first set of experiments, for simplicity, we have set α=10, that should represent 10% of strangers for each network similarity group under the assumption that strangers follow a uniform distribution w.r.t. their network similarity with owner.

[3]http://www.facebook.com/developers/apps.php?app_id=103656703030138

[4]The statistics have been computed on those available user profiles.

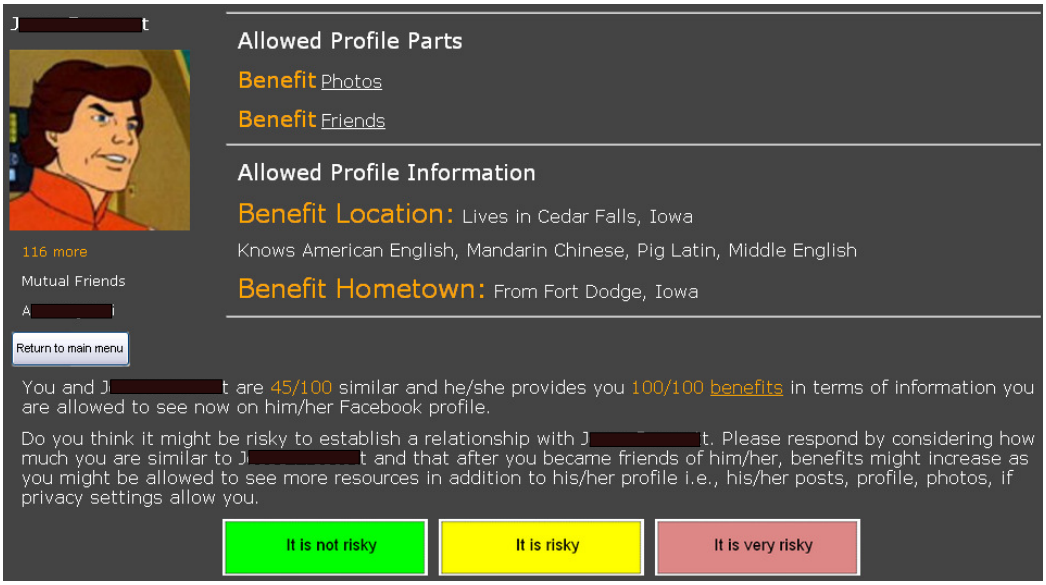

Fig. 3. Snapshot from the extension

However, Figure 4 reports that stranger distribution in network similarity groups is not uniform. More precisely, in the x-axis network similarity groups are ordered based on increasing network similarity values. Thus, for example strangers in the 3rd group have higher network similarity with owner than strangers in the 2nd network similarity group. Figure 4 highlights that most of the strangers are weakly connected with owners. Moreover, we found that no stranger has network similarity values greater than 0.6, so in the following we do not report experiments for these network similarity groups.

In contrast, β represents the threshold for the creation of a new profile based cluster, so it constraints the number of profile based clusters generated by Squeezer over each network similarity group. Thus, increasing β could result in too many profile based clusters each of which with few strangers, which implies too many distinct learning processes to be executed. To avoid this situation, we select β=0.4.

A further parameter is related to the number of strangers that are required to be labeled by the owner in each round, which in these experiments have been set to 3, to keep minimum the owner effort. Finally, a cluster is considered stabilized when classification change does not happen in 2 rounds of labeling (i.e., parameter n in Section III-D).

C. Risk learning process

In order to show the effectiveness of the proposed risk learning process we run several experiments.

Risk Label Prediction. As explained in Section III-D, the risk learning process stops when both the accuracy and stabilization conditions hold. We recall that if both these hold it means that risk labels are predicted with a good accuracy (i.e., RMSE between owner and predicted labels is less than 0.5) and for at least 2 rounds there is no classification changes with a confidence c selected by the owner.

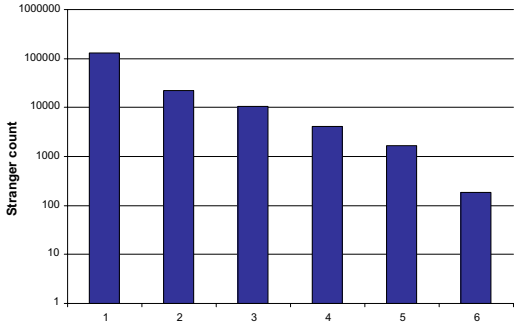

Fig. 4. Stranger count for network similarity groups

At the average, the confidence selected by our 47 users is about 80% (i.e., 78,39). With these confidence values, experiments highlight that labels prediction stabilize in about 3 rounds (i.e., 3,29 is the average).

Whilst these results confirm that we have good accuracy (an RMSE error less than 0.5), we are also interested in those predicted labels that exactly match what the owner would have given (i.e., $\hat{L}(s)^{i+1} = L(s)^i$). To have this estimation, during the accuracy evaluation, we count the number of predicted labels that match the owner labels. The promising result is that 83,36% of predicted labels exactly match the owner labels.

Stranger sampling. To validate the strategy adopted for pool generation, we perform experiments to compare network and profile based pools (NPP) against pools generated by only considering network similarity (NSP).

The impact of profile similarity grouping can be seen in Figures 5 and 6, showing better results for NPP in terms of

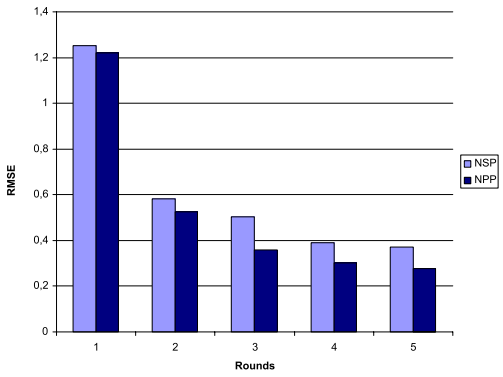

Fig. 5. Error rate by rounds for NPP and NSP pools.

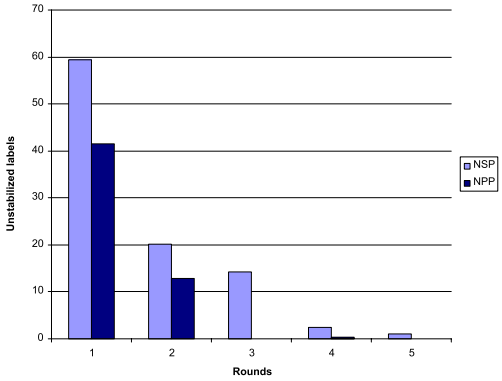

Fig. 6. Average number of unstabilized labels for NPP and NSP pools.

root mean square error and stabilization per round.

D. Risk Measure

As discussed in Section II, risk labels are based on similarity and benefits. In the following experimental setting, we will evaluate how these two factors affect the owner's risk judgment.

Network and Profile Similarity. We first consider network similarity. In creating network similarity groups, we wanted to capture how owner judgment is affected by network connections. Our data shows that some strangers can have more than 40 mutual friends with an owner. Hence, with increasing network similarity we assume that the possibility of an acquaintance between owner and stranger increases. We expect that this increasing possibility is reflected in lower assigned risky labels. Indeed, Figure 7 shows that with increasing network similarity, the percentage of very risky labels in network similarity groups consistently decreases.

Along with network similarity, we evaluated the importance of stranger profile similarity in owner labeling. To this purpose, we look for patterns in profile attributes of strangers that are assigned by owner to the same risk labels, so to determine how much each stranger profile attribute affects owner decision. As an example, if risk labels given by owner show a pattern with gender (e.g., all males are labeled as very risky), we can

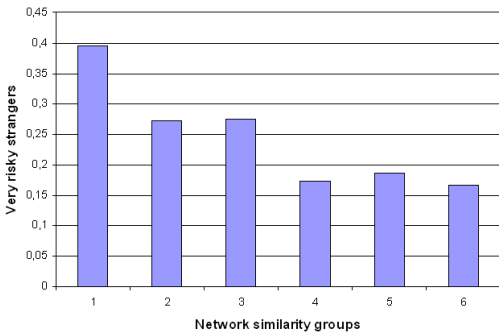

Fig. 7. Percentage of very risky strangers in network similarity groups

infer that gender is an important attribute in owner decision. In information theory, information gain ratio [20] is used to capture the importance of a variable in decision trees. More precisely, a high information gain ratio for a profile attribute implies a reduction in entropy[5] caused by attribute value. A higher gain ratio item carries more information in understanding the rationale behind owner labeling (e.g., all males are risky).

In our setting, we use three profile attributes for clustering with Squeezer algorithm. These are gender, last_name and locale, where this latter denotes Facebook web interface language of the owner. For example, English speaking owners from USA have EN_US locale, while owners from Great Britain have EN_GB locale.

***Definition** 6 (Attribute importance):* Given a profile attribute $pa_i \in PA$, we estimate its importance $\mathcal{I}_{pa_i}$ by normalizing the information gain ratio as follows:

$$\mathcal{I}_{pa_i} = \frac{IGR(pa_i)}{\sum_{pa_j \in PA} IGR(pa_j)}$$

where $IGR(pa_i)$ denotes information gain ratio of attribute pa_i.

In Table I, we show profile attribute importance for our owners in an ordered list, where $\mathcal{I}_1$ refers to the most important profile attribute (highest information gain ratio), and $\mathcal{I}_3$ refers to the least important attribute. For example, in the table, values in $\mathcal{I}_i$-th column and gender row correspond to the number of owners for which gender is the i-th most important item. As highlighted in the table, gender has the biggest average weight for owners, and for 34 owners it is the most important item ($\mathcal{I}_1$). Gender is followed by locale, and last_name. Although last_name has low average weight, it is more important than locale for two owners.

Benefits. The discussion on similarity aimed at finding similarity patterns in owners' risk judgments. Now we will

[5]Here we refer to the entropy of label distribution in three risk classes (i.e., very risky, risky, not risky).

TABLE I
PROFILE ATTRIBUTES IMPORTANCE

	$\mathcal{I}_1$	$\mathcal{I}_2$	$\mathcal{I}_3$	Avg Imp.
gender	34	12	1	0.6231
locale	13	33	1	0.3226
last_name	0	2	45	0.0542

TABLE III
OWNER GIVEN θ WEIGHTS

	Average θ weight
hometown	0.155
friend	0.149
photo	0.147
location	0.143
education	0.1393
wall	0.1328
work	0.1321

look at what benefit patterns can be found in owners' risk judgment. More precisely, we adopt for benefit items the same importance definition from information gain ratio (Definition 6).

Whereas in similarity we have categorical item values such as gender:male, in benefits we work with visibility values such as photos:1. Here item values 0 and 1 correspond to visibility:false and visibility:true, respectively on stranger profiles.

In line with importance of items in similarity, we found importance of benefits in owners' risk judgment. In Table II we show the average importance as well as the order of items' importance. As reported, photos are considered the most important benefit item for 21 users, followed by education and work. Home wall has the least average importance, but for 4 owners it is the most important benefit.

According to our benefit measure, we asked owners to give a θ weight to each benefit item. In Table III those θ weights are shown. When comparing these weights to the ones we found in Table II, we see that home wall benefit is given a low importance in both tables and some other weights are correlated. The relevant issue this table highlights is that for some benefit items it is better to use system suggested weights.

Another aspect of benefits is related to stranger profile attributes, as some patterns can be found with respect to the benefits they provide to owners. On social networks, gender of a stranger has been known to affect the visibility of profile items. In [11], Fogel et al. show that females have stricter privacy settings than males. As a result, the benefits they provide as strangers are lower than those of male users. To understand how provided benefits change with gender, we computed visibility of profile items based on gender. In Table IV, female strangers are shown to have lower visibility values for profile items. However, in photos male and female strangers do not show a big difference and visibility is almost the same for both genders. In Fogel's study, social network users were asked this question: *Do you include a picture of yourself on your profile?*. This *picture of yourself* can be the profile picture, or a picture in photo albums. In our experiment, the photos are photo albums on Facebook, excluding a profile picture. Fogel reports that the answer to the asked question was yes for 90.4% of males, and 81.6% of females. In our experiment, we found the visibility of photos to 88% and 87% for males and females, respectively. However, in our experiments we could not check if the visible photos on Facebook were self-pictures. Visibility percentages can be lower when non personal photos are excluded.

Similar to male/female difference, benefits show a pattern with locale values. In Table V we show visibility of profile items for strangers from 7 locales (TR:Turkey, DE:Germany, US:USA, IT:Italy, GB:Britain,ES:Spain, PL:Poland). Work has the lowest visibility among items.

It is interesting to note that although home walls are considered very private and hidden, home was one of the lowest importance items in both Tables III and II. It seems that social network users do not share their walls with strangers, but they are also not interested in other strangers' walls. Photos have very high visibility among all locales while friends list visibility ranges from 41% to 72%. In Table V we can also see some correlation between locales. Visibility of profile items do not change greatly between IT and ES locales and the difference between percentages is around 5%.

V. RELATED WORK

In the literature, risk models [21], [22], [23] have been proposed to put privacy in a context so that social network users would be empowered to understand privacy risks. Privacy risks have been evaluated from access control [24], [25] and information leakage [26] perspectives. Lately, analyzing the social graph of a user [26], [27], [28] has been found successful in revealing personal information even when the user does not enter personal data to the social network.

Our work is related to these approaches because we also consider the information leakage to friends of friends. However, we do not a priori assume that all revealed information poses a privacy risk, and we do not consider all network users to be privacy threats. In contrast, in our approach we try to closely monitor users' risk attitude and find out when interactions with other social network users are desirable. In this sense, our view is closer to the mindset of social network users.

Other related work are those proposing enhanced privacy setting tools [3], [4]. In these proposals, feedback from a user is used to suggest the best privacy settings. However, these privacy tools are not based on the concept of risk, and therefore their functionality is similar to access control models.

Similar to our work, the risk of maintaining an online presence has been studied in [29], from which we borrowed our visibility measure for items. However, risk is studied in [29] from a different point of view. Indeed, in [29] a privacy score is computed for social network users, which measures the user's potential privacy risk due to his/her online information sharing behaviors. Such measure depends on the sensitivity of the information being revealed and the visibility the revealed information gets in the network. In contrast, in

TABLE II
MINED IMPORTANCE OF BENEFITS

	$\mathcal{I}_1$	$\mathcal{I}_2$	$\mathcal{I}_3$	$\mathcal{I}_4$	$\mathcal{I}_5$	$\mathcal{I}_6$	$\mathcal{I}_7$	Avg Imp.
photo	21	8	6	4	3	0	5	0.27
education	11	9	4	3	10	4	6	0.143
work	8	7	9	7	5	7	4	0.140
friends	2	10	7	6	6	8	8	0.13
hometown	0	7	9	11	6	9	5	0.11
location	1	4	8	9	11	8	6	0.092
wall	4	2	4	7	6	11	13	0.091

TABLE IV
ITEM VISIBILITY FOR DIFFERENT GENDERS

	wall	photo	friend	loc.	edu.	work	hometown
male	25 %	88 %	56 %	42 %	35 %	20 %	41%
female	16 %	87 %	47 %	32 %	28 %	12 %	30%

TABLE V
VISIBILITY OF PROFILE ITEMS FOR DIFFERENT LOCALE STRANGERS

	wall	photo	friend	loc.	edu.	work	hometown
TR	20%	84%	41%	36%	31%	15%	32%
DE	20%	77%	46%	34%	17%	17%	34%
US	17%	89%	52%	42%	34%	18%	37%
IT	27%	92%	68%	32%	38%	14%	41%
GB	12%	91%	46%	38%	25%	17%	32%
ES	22%	87%	63%	37%	28%	13%	37%
PL	31%	95%	72%	33%	23%	13%	31%

our work we are interested in finding risk scores of friends of friends of the owner by considering their similarity with the owners and potential owner benefits.

VI. CONCLUSIONS

In this paper, we have proposed a measure for estimating the risk, in terms of disclosure of personal information, of interacting with OSN users. Risk levels are computed for 2 hop connections by taking into account both similarity and benefit metrics as well as user risk attitude. They are computed through an active learning process, where owner risk attitude is learnt from few required interactions. The initial experiments we have performed show the effectiveness of our proposal. We plan to extend this work along several directions. First, we plan to extend our tests to a greater data set and to data sets coming from different social networks. Moreover, we plan to develop techniques to mine from the data most of the values for the parameters on which our learning process relies. Finally, we envision a variety of applications for our risk labels that we would like to explore in the future, such as privacy settings/friendships suggestion or label-based access control.

ACKNOWLEDGMENT

The research presented in this paper was partially funded by a Google Research Award.

REFERENCES

[1] D. Boyd and N. Ellison, "Social network sites: Definition, history, and scholarship," *Journal of Computer-Mediated Communication*, vol. 13, no. 1, pp. 210–230, 2008.

[2] B. Carminati and E. Ferrari, "Privacy-aware access control in social networks: Issues and solutions," in *Privacy and Anonymity in Information Management Systems*, ser. Advanced Information and Knowledge Processing, X. Wu, J. Nin, and J. Herranz, Eds. Springer London, 2010, pp. 181–195.

[3] L. Fang and K. LeFevre, "Privacy wizards for social networking sites," in *Proceedings of the 19th International Conference on World Wide Web, WWW 2010, Raleigh, North Carolina, USA, April 26-30*, 2010, pp. 351–360.

[4] H. Lipford, A. Besmer, and J. Watson, "Understanding privacy settings in facebook with an audience view," in *Proceedings of the 1st Conference on Usability, Psychology, and Security*. USENIX Association, 2008, pp. 1–8.

[5] A. Joinson, "Looking at, looking up or keeping up with people?: motives and use of facebook," in *Proceeding of the twenty-sixth annual SIGCHI conference on Human factors in computing systems*. ACM, 2008, pp. 1027–1036.

[6] D. M. Boyd and E. Hargittai, "Facebook privacy settings: Who cares?" *First Monday*, vol. 15, no. 8, p. 23, 2010.

[7] M. McPherson, L. Smith-Lovin, and J. Cook, "Birds of a feather: Homophily in social networks," *Annual review of sociology*, vol. 27, pp. 415–444, 2001.

[8] G. Kossinets and D. Watts, "Empirical analysis of an evolving social network," *Science*, vol. 311, no. 5757, p. 88, 2006.

[9] C. Akcora, B. Carminati, and E. Ferrari, "Network and profile based measures for user similarities on social networks," in *Information Reuse and Integration (IRI), 2011 IEEE International Conference on*. IEEE, 2011.

[10] A. Acquisti and R. Gross, "Imagined communities: Awareness, information sharing, and privacy on the Facebook," in *Privacy Enhancing Technologies*. Springer, 2006, pp. 36–58.

[11] J. Fogel and E. Nehmad, "Internet social network communities: Risk taking, trust, and privacy concerns," *Computers in Human Behavior*, vol. 25, no. 1, pp. 153–160, 2009.

[12] E. Spertus, M. Sahami, and O. Buyukkokten, "Evaluating similarity measures: a large-scale study in the orkut social network," in *Proceedings of the Eleventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Chicago, Illinois, USA, August 21-24*. ACM, 2005, pp. 678–684.

[13] E. Rogers, *Diffusion of innovations*. Free Pr, 1995.

[14] S. Kotsiantis, I. Zaharakis, and P. Pintelas, "Supervised machine learning: A review of classification techniques," *Emerging artificial intelligence applications in computer engineering: real word AI systems with applications in eHealth, HCI, information retrieval and pervasive technologies*, vol. 160, p. 3, 2007.

[15] B. Settles, "Active learning literature survey," *Machine Learning*, vol. 15, no. 2, pp. 201–221, 1994.

[16] G. Gan, C. Ma, and J. Wu, "Data clustering: theory, algorithms, and applications," *ASASIAM Series on Statistics and Applied Probability*, vol. 20, pp. 219–230, 2007.

[17] Z. He, X. Xu, and S. Deng, "Squeezer: an efficient algorithm for clustering categorical data," *Journal of Computer Science and Technology*, vol. 17, no. 5, pp. 611–624, 2002.

[18] X. Zhu, Z. Ghahramani, and J. Lafferty, "Semi-supervised learning using gaussian fields and harmonic functions," in *Machine learning-international workshop then conference-*, vol. 20, no. 2, 2003, p. 912.

[19] J. Zhu, H. Wang, and E. Hovy, "Multi-criteria-based strategy to stop active learning for data annotation," in *Proceedings of the 22nd International Conference on Computational Linguistics-Volume 1*. Association for Computational Linguistics, 2008, pp. 1129–1136.

[20] D. MacKay, *Information theory, inference, and learning algorithms*, D. MacKay, Ed. Cambridge Univ Pr, 2003.

[21] B. Carminati, E. Ferrari, S. Morasca, and D. Taibi, "A probability-based approach to modeling the risk of unauthorized propagation of information in on-line social networks," in *Proceedings of the first ACM conference on Data and application security and privacy*. ACM, 2011, pp. 51–62.

[22] T. Wang, M. Srivatsa, D. Agrawal, and L. Liu, "Modeling data flow in socio-information networks: A risk estimation approach," *SACMAT 2011, 16th ACM Symposium on Access Control Models and Technologies, Innsbruck, Austria, June 15-17, Proceedings*, 2011.

[23] A. Dubrawski, P. Sarkar, and L. Chen, "Trade-offs between agility and reliability of predictions in dynamic social networks used to model risk of microbial contamination of food," in *Social Network Analysis and Mining, 2009. ASONAM'09. International Conference on Advances in*. IEEE, 2009, pp. 125–130.

[24] B. Carminati, E. Ferrari, and A. Perego, "Enforcing access control in web-based social networks," *ACM Transactions on Information and System Security*, vol. 13, no. 1, pp. 1–38, 2009.

[25] K. Gollu, S. Saroiu, and A. Wolman, "A social networking-based access control scheme for personal content," in *Proceedings of the 21st ACM Symposium on Operating Systems Principles (SOSP07)-Work-in-Progress Session*, 2007.

[26] B. Krishnamurthy and C. Wills, "On the leakage of personally identifiable information via online social networks," *ACM, Computer Communication Review*, vol. 40, no. 1, pp. 112–117, 2010.

[27] J. Bonneau, J. Anderson, and G. Danezis, "Prying data out of a social network," in *First International Conference on Advances in Social Networks Analysis and Mining*. Citeseer, 2009.

[28] A. Mislove, B. Viswanath, K. Gummadi, and P. Druschel, "You are who you know: inferring user profiles in online social networks," in *Proceedings of the Third International Conference on Web Search and Web Data Mining, WSDM 2010, New York, NY, USA, February 4-6*. ACM, 2010, pp. 251–260.

[29] K. Liu and E. Terzi, "A framework for computing the privacy scores of users in online social networks," in *2009 Ninth IEEE International Conference on Data Mining*. IEEE, 2009, pp. 288–297.