A Semi-supervised Approach to Measuring User Privacy in Online Social Networks

A Semi-supervised Approach to Measuring User Privacy in Online Social Networks

Ruggero G. Pensa and Gianpiero Di Blasi

Department of Computer Science
University of Torino, Italy
ruggero.pensa@unito.it

Abstract. During our digital social life, we share terabytes of information that can potentially reveal private facts and personality traits to unexpected strangers. Despite the research efforts aiming at providing efficient solutions for the anonymization of huge databases (including networked data), in online social networks the most powerful privacy protection is in the hands of the users. However, most users are not aware of the risks derived by the indiscriminate disclosure of their personal data. With the aim of fostering their awareness on private data leakage risk, some measures have been proposed that quantify the privacy risk of each user. However, these measures do not capture the objective risk of users since they assume that all user's direct social connections are close (thus trustworthy) friends. Since this assumption is too strong, in this paper we propose an alternative approach: each user decides which friends are allowed to see each profile item/post and our privacy score is defined accordingly. We show that it can be easily computed with minimal user intervention by leveraging an active learning approach. Finally, we validate our measure on a set of real Facebook users.

Keywords: privacy metrics, active learning, online social networks

1 Introduction

Online social networks are among the main traffic sources in the Internet. At the end of 2014, they attracted more than 31% of the worldwide internet traffic towards the Web. Facebook, the most famous social networking platform, drives alone 25% of the whole traffic. As a comparison, Google search engine represents just over 37% of the global traffic[1]. More than two billions people are estimated to be registered in at least one of the most popular social media platforms (Facebook hits the goal of one billion users in 2012). Overall, the number of active "social" accounts are more than two billions. The famous "six degrees of separation" theory has been far exceed in Facebook, where an average degree of 3.57 has been recently observed[2]. Consequently, social network users are constantly exposed to privacy leakage risks. Although most users do not disclose

[1] Source: http://www.alexa.com/

[2] https://research.facebook.com/blog/three-and-a-half-degrees-of-separation/

very sensitive facts (private life events, diseases, political ideas, sexual preferences, and so on), they are simply not aware of the risks due to the disclosure of less sensitive information, such as GPS tags, photos taken during a vacation period, page likes, or comments on news. As an example, the research project myPersonality [14] carried out at the University of Cambridge has shown that, by leveraging Facebook user's activity (such as "Likes" to posts or fan pages) it is possible to "guess" some very private traits of the user's personality. According to another study, it is even possible to infer some user characteristics from the attributes of users who are part of the same communities [18]. As a consequence, privacy has become a primary concern among social network analysts and Web/data scientists. Also, in recent years, many companies are realizing the necessity to consider privacy at every stage of their business. In practice, they have been turning to the principle of *Privacy by Design* [5] by integrating privacy requirements into their business model.

Despite the huge research efforts aiming at providing efficient solutions to the anonymization of huge databases (including networked data) [3, 25], in online social networks the most powerful privacy protection is in the hands of the users: they, and only they, decide what to publish and to whom. Even though social networking sites (such as Facebook), notify their users about the risks of disclosing private information, most people are not aware of the dangers due to the indiscriminate disclosure of their personal data when they surf the net. Some social media provide advanced tools for controlling the privacy settings of the user's profile [24]. However, yet a large part of Facebook content is shared with the default privacy settings and exposed to more users than expected [17]. According to Facebook CTO Bret Taylor, even though most people have modified their privacy settings[3], in 2012, still "13 million users [in the United States] said they had never set, or didn't know about, Facebook's privacy tools[4]".

Some studies try to foster risk perception and awareness by "measuring" users' profile privacy according to their privacy settings [16, 23]. These metrics usually require a *separation-based* policy configuration: in other terms, the users decide "how distant" a published item may spread in the network. Typical separation-based privacy policies for profile item/post visibility include: visible to no one, visible to friends, visible to friends of friends, public. However, this policy fails when the number of user friends becomes large. According to a well-known anthropological theory, in fact, the maximum number of people with whom one can maintain stable social (and cybersocial) relationships (known as Dunbar's number) is around 150 [10, 20], but the average number of user friends in Facebook is more than double [5]. This means that many social links are weak (offline and online interactions with them are sporadic), and a user who sets the

[3] `http://www.zdnet.com/article/facebook-cto-most-people-have-modified-their-privacy-settings/`

[4] `http://www.consumerreports.org/cro/magazine/2012/06/facebook-your-privacy/index.htm`

[5] `http://www.pewresearch.org/fact-tank/2014/02/03/6-new-facts-about-facebook/`

privacy level of an item to "visible to friends" probably is not willing to make that item visible to *all* her friends.

To address this limitation, in this paper we propose a *circle-based* formulation of the privacy score proposed by Liu and Terzi [16]. We assume that a user may set the visibility of each action and profile item separately for each other user in her friend list. For instance, a user u may decide to allow the access to all photo albums to friends f_1 and f_2, but not to friend f_3. In our score, the sensitivity and visibility of profile item i published by user u are computed according to the set of u's friends that are allowed to access the information provided by i. Since the expression of explicit allow/deny policy for each friend and each item may require huge labeling efforts, we also propose an active learning labeling approach to limit the number of manual operations. We show experimentally that i) our circle-based definition of privacy score better capture the real privacy leakage risk and ii) the active learning approach provides accurate results in terms of both predicted privacy settings and final privacy score.

The remainder of the paper is organized as follows: we briefly review the related literature in Section 2; the overview and the theoretical details of our score are presented in Section 3; the active learning approach is presented in Section 4; Section 5 provides the report of our experimental validation; finally, we draw some conclusions in Section 6.

2 Related work

Most research efforts in social network privacy are devoted to the identification and formalization of privacy breaches and to the anonymization of networked data [25]. All these works focus on how to share social networks owned by companies or organizations masking the identities or the sensitive connections of the individuals involved. However, increasing attention is being paid to the privacy risk of users caused by their information-sharing activities (e.g., posts, likes, shares). In fact, since disclosing information on the web is a voluntary activity, a common opinion is that users should care about their privacy during their interaction with other social network users. Thus, another branch of research has focused on investigating strategies and tools to enhance the users' privacy awareness and help them act more safely during their day-to-day social network activity. In [6] the authors present an online game, called Friend Inspector, that allows Facebook users to check their knowledge of the visibility of their shared personal items and provides recommendations on how to improve privacy settings. Instead, Fang and LeFevre [11] propose a social networking privacy wizard based on active learning. The wizard iteratively asks the user to allow or deny the visibility of profile items to selected friends and assign privileges to the rest of the user's friends using a classifier. [4] presents a tool to detect unintended information loss in online social networks by quantifying the privacy risk attributed to friend relationships in Facebook. The authors show that a majority of users' personal attributes can be inferred from social circles. In [22] the authors present a privacy protection tool that measures the inference probability of sensitive at-

tributes from friendship links. In addition, they suggest self-sanitization actions to regulate the amount of leakage. [12], instead, introduces a machine learning technique to monitor users' privacy settings and recommend reasonable privacy options. Other approaches to privacy control in social networks investigate the problem of the risk perception. In [1, 2], for instance, the authors propose to provide users with a measure of how much it might be risky to have interactions with them, in terms of disclosure of private information. They use an active learning approach to estimate user risk from few required user interactions.

The privacy measure we propose in this paper is closely related to the work of Liu and Terzi [16]. They propose a framework to compute a privacy score measuring the users' potential risk caused by their participation in the network. This score takes into account the sensitivity and the visibility of the disclosed information and leverages the item response theory as theoretical basis for the mathematical formulation of the score. Another privacy measure has been proposed in [23] where the authors introduce a privacy index to measure the user privacy exposure in a social network. This index, however, strongly relies on predefined sensitivity values for users' items. Furthermore, in both proposals, the privacy measures are computed by leveraging separation-based privacy policies. Differently from the above mentioned papers, our proposal considers circle-based policy settings that better suits the real user visibility preferences.

3 A circle-based definition of privacy score

In this section we introduce our circle-based privacy score aiming at supporting the users participating in a social network in assessing their own privacy leakage risk. Most social networking platforms (such as Facebook or Google+), provide an adequate flexibility in configuring privacy of profile items and user's actions. Moreover, they offer some advanced facilities, such as the possibility of grouping friends into special lists or social circles. But privacy is not just a matter of users' preferences; it also relies on the context in which an individual is immersed: the position within the network (very central users are more exposed than marginal users), her or his own attitude on disclosing very private facts, and so on. Hence, we propose a privacy score that takes all these aspects into account and fits the real user expectations about the visibility of profile items.

Before entering the technical details of our approach, we briefly introduce some basic mathematical notation required to formalize the problem.

3.1 Preliminaries and notation

Here we introduce the mathematical notation we will adopt in the rest of our paper. We consider a set of n users $\mathcal{U} = \{u_1, \ldots, u_n\}$ corresponding to the individuals participating in a social network. Each user is characterized by a set of m properties or profile items $\mathcal{P} = \{p_1, \ldots, p_m\}$, corresponding, for instance, to personal information such as gender, age, political views, religion,

workplace, birthplace and so on. Hence, each user u_i is described by a vector $\boldsymbol{p}^i =< p_{i1}, \ldots, p_{im} >$.

Users are part of a social network. Without loss of generality, we assume that the link between two users is always reciprocal (if there is a link from u_j to u_j then there is also a link from u_j to u_i). Hence, the social network here is represented as an undirected graph $G(V, E)$, where V is a set of n vertices $\{v_1, \ldots, v_n\}$ such that each vertex $v_i \in V$ is the counterpart of user $u_i \in \mathcal{U}$ and E is a set of edges $E = \{(v_i, v_k)\}$. Given a pair of users $(u_i, u_k) \in \mathcal{U}$, $(v_i, v_k) \in E$ iif users u_i and u_k are connected (e.g., by a friendship link).

For any given vertex $v_i \in V$ we define the neighborhood $\mathcal{N}(v_i)$ as the set of vertices v_k directly connected to the vertex v_i, i.e., $\mathcal{N}(v_i) = \{v_k \in V \mid (v_i, v_k) \in E\}$. Conversationally speaking, $\mathcal{N}(v_i)$ is the set of friends (also known as *friend-list*) of user u_i, hence we use $\mathcal{N}(v_i)$ or $\mathcal{N}(u_i)$ interchangeably. Given a user u_i and her friend-list $\mathcal{N}(u_i)$, we also define the *ego network* centered on user u_i as the graph $G_i(V_i, E_i)$, where $V_i = \mathcal{N}(v_i) \cup \{v_i\}$ and $E_i = \{(v_k, v_l) \in E \mid v_k, v_l \in V_i\}$.

Finally, for any user u_i we introduce a *privacy policy matrix* $\boldsymbol{M}_i \in \{0, 1\}^{n_i \times m}$ (with $n_i = |\mathcal{N}(u_i)|$) defined as follows: for any element m^i_{kj} of $\boldsymbol{M}_i$, $m^i_{kj} = 1$ iif profile item $p_j \in \mathcal{P}$ is visible to user $u_k \in \mathcal{N}(u_i)$ (0 otherwise, i.e., iif user u_k is not allowed to access profile item p_j).

It is worth noting that our framework can be easily extended to the case of directed social networks (such as Twitter): in this case, the privacy policies are defined only on inbound links.

3.2 Privacy score

Our measure is inspired by the privacy score defined by Liu and Terzi [16]. It measures the user's potential risk caused by his or her participation in the network. A $n \times m$ response matrix $\boldsymbol{R}$ is associated to the set of n users $\mathcal{U}$ and the set of m profile properties $\mathcal{P}$. In [16], each element r_{ij} of $\boldsymbol{R}$ contains a privacy level that determines the willingness of user u_i to disclose information associated with property p_j. In the binomial case $r_{ij} \in \{0, 1\}$: $r_{ij} = 1$ (resp. $r_{ij} = 0$) means that user u_i has made the information associated with profile item p_j publicly available (resp. private). In the multinomial case, entries in $\boldsymbol{R}$ take any non-negative integer values in $\{0, 1, \ldots, \ell\}$, where $r_{ij} = h$ (with $h \in \{0, 1, \ldots, \ell\}$) means that user u_i discloses information related to item p_j to users that are at most h links away in the social network G (e.g., if $r_{ij} = 0$ user u_i wants to keep p_j private, if $r_{ij} = 1$ user u_i is willing to make p_j available to all friends, if $r_{ij} = 2$ user u_i is willing to make p_j available to the friends of her or his friends, and so on). For this reason, we call this policy *separation-based*. However, in this work, we adopt a different meaning for the entries r_{ij} of $\boldsymbol{R}$: in our framework r_{ij} is directly proportional to the number of friends to whom u_i is willing to disclose the information of profile property p_j. Hence, we can compute $\boldsymbol{R}$ according to the *circle-based* privacy policies defined by matrices $\boldsymbol{M}_i$'s using this formula:

$$r_{ij} = \left\lfloor \ell \cdot \frac{1}{|\mathcal{N}(u_i)|} \sum_{k=1}^{|\mathcal{N}(u_i)|} m^i_{kj} \right\rfloor \tag{1}$$

where $\mathcal{N}(u_i)$ is the set of friends of user u_i, m^i_{kj} denotes the visibility of user u_i's profile item p_j for friend u_k, and $\lfloor \cdot \rfloor$ is the floor function. As a consequence, $r_{ij} = \ell$ iif $\forall u_k \in \mathcal{N}(u_i)$, $m^i_{kj} = 1$. Our definition is conceptually different from the original one, since the latter does not take into account the possibility of disclosing personal items to just a part of friends.

In the following, we use $\boldsymbol{R}^S$ when we refer to the response matrix computed with the original separation-based policy approach defined in [16]. We use $\boldsymbol{R}^C$ when we refer to our circle-based definition of response matrix.

Using the response matrix it is possible to compute the two main components of the privacy function: the sensitivity β_{jh} of a profile item p_j for a given privacy level h, and the visibility V_{ijh} of a profile item p_j due to u_i for a given level h. The sensitivity of a profile item p_j depends on the item itself (attribute "sexual preferences" is usually considered more sensitive than "age"). The visibility, instead, captures to what extent information about profile item p_j of user u_i spreads in the network. For the computation details of β_{jh} and V_{ijh} we invite the reader to refer to [16], where a mathematical model based on item response theory (a well known theory in psychometrics) is used to compute sensitivity and visibility. Intuitively, sensitivity β_j is such that the more users adopt at least privacy level h for privacy item p_j, the less sensitive p_j is w.r.t. level h. Instead, visibility V_{ijh} is higher when the sensitivity of profile items is low and when users have the tendency to disclose lots of their profile items. Moreover, it depends on the position of user u_i within the network and can be computed by exploiting any information propagation models [13].

The privacy score $\phi_p(u_i, p_j)$ for any user u_i and profile property p_j is computed as follows:

$$\phi_p(u_i, p_j) = \sum_{h=0}^{\ell} \beta_{jh} \cdot V_{ijh}. \tag{2}$$

and the overall privacy score $\phi_p(u_i)$ for any user u_i is given by

$$\phi_p(u_i) = \sum_{j=1}^{m} \phi_p(u_i, p_j). \tag{3}$$

From Equation 2 and 3 it is clear that users that have the tendency to disclose sensitive profile properties to a wide public are more prone to privacy leakage. Intuitively, $\phi_p(u_i) = 0$ means that, in each element of the summation, either $\beta_{jh} = 0$ (the profile item p_j is not sensitive at all), or $V_{ijh} = 0$ (the profile item p_j is kept private). On the contrary, the privacy score is maximum when a user discloses to all her or his friends ($V_{ijh} = 1$) all sensitive information ($\beta_{jh} = 1$).

In this paper, we use ϕ^S_p when we refer to the score computed using the original separation-based response matrix $\boldsymbol{R}^S$; we use ϕ^C_p when we refer to the privacy score leveraging our circle-based definition of response matrix $\boldsymbol{R}^C$.

Table 1. Example of Input Dataset for the Classification Task

Friend ID	Age	Gender	Hometown	Community	No. of friends	C_{work}	C_{photos}	$C_{politics}$
102030	"21-30"	Male	Rome	C10	"501-700"	allow	allow	deny
203040	"31-40"	Female	Madrid	C5	"201-300"	allow	deny	deny
304050	"15-19"	Female	Paris	C7	"101-200"	allow	deny	deny
405060	"41-50"	Female	Berlin	C5	"701-1000"	allow	deny	deny
506070	"51-60"	Male	Rome	C10	"501-700"	allow	allow	deny
607080	"21-30"	Female	Rome	C10	"301-500"	?	?	?
708090	"41-50"	Male	Madrid	C5	"301-500"	?	?	?

4 Semi-supervised privacy policy definition

Our definition of privacy score requires the availability of visibility preferences for all user friends. However, setting them correctly is often an annoying and frustrating task and many users may prefer to adopt simple but extreme strategies such as "visible-to-all" (exposing themselves to the highest risk), or "hidden-to-all" (wasting the positive social and economic potential of social networking websites). In this section we present a semi-supervised approach to minimize the user's intervention while computing the circle-based privacy policy matrices $\boldsymbol{M}_i$. The classification model should be as accurate as possible in predicting those privacy preferences not explicitly set by the users. Moreover, the model should be easily updatable when the user sets more privacy preferences or adds new users. Our choice is to use a Naive Bayes classifier [19], which is simple and converge quickly even with few training data. Moreover, it can be easily embedded in an active learning framework using, for instance, uncertainty sampling [9] thus minimizing the intervention of the user in the model training phase.

For any given user $u_i \in \mathcal{U}$ and any given profile item $p_j \in \mathcal{P}$ we define a classification problem in which we have a set of $|\mathcal{N}(u_i)|$ instances $D = \{d_1, \ldots, d_{|\mathcal{N}(u_i)|}\}$ corresponding to all friends of u_i. Each instance d_k is characterized by a set of p attributes $\{A_1, \ldots, A_p\}$ with discrete values and m class variables $\{C_1, \ldots, C_m\}$ that take values in the domain $\{allow, deny\}$: $C_j = allow$ (resp. $C_j = deny$) means that friend u_k is allowed (resp. is not allowed) to access the information of profile item p_j of user u_i. The values of attributes $\{A_1, \ldots, A_p\}$ are partly derived from the profile vector $\boldsymbol{p}^k =< p_{k1}, \ldots, p_{km} >$ of users u_k, partly from the ego network $G_i(V_i, E_i)$ of user u_i (see Section 3.1). For instance, they may contain information such as the workplace and home-town of u_k, or the communities in G_i u_k belong to. Table 1 is an example of possible small dataset for a generic user consisting of five training instances and two test instances with three profile-based attributes, two network-based attributes and three class variables.

The Naive Bayes classification task can be regarded as estimating the class posterior probabilities given a test example d_k, i.e., $Pr(C_j = allow|d_k)$ and $Pr(C_j = deny|d_k)$. The class with the highest probability is assigned to the example d_k. Given a test example d_k, the observed attribute values are given by the vector $\boldsymbol{d}^k = \{a_1^k, \ldots, \ldots, a_p^k\}$, where a_s^k is a possible value of A_s, $s = 1, \ldots, p$. The prediction is the class c ($c \in \{allow, deny\}$) such that $Pr(C_j = c|A_1 = a_1^k, \ldots, A_p = a_p^k)$ is maximal. By Bayes' theorem, the above quantity can be

expressed as

$$Pr(C_j = c|A_1 = a_1^k, \ldots, A_p = a_p^k) =$$
$$= \frac{Pr(A_1 = a_1^k, \ldots, A_p = a_p^k|C_j = c)Pr(C_j = c)}{Pr(A_1 = a_1^k, \ldots, A_p = a_p^k)} =$$
$$= \frac{Pr(A_1 = a_1^k, \ldots, A_p = a_p^k|C_j = c)Pr(C_j = c)}{\sum_{c_x} Pr(A_1 = a_1^k, \ldots, A_p = a_p^k|C_j = c_x)Pr(C_j = c_x)} \quad (4)$$

where, $Pr(C_j = c)$ is the class prior probability of c, which can be estimated from the training data. If we assume that conditional independence holds, i.e., all attributes are conditionally independent given the class $C_j = c$, then

$$Pr(A_1 = a_1^k, \ldots, A_p = a_p^k|C_j = c) = \prod_{s=1}^{p} Pr(A_s = a_s^k|C_j = c) \quad (5)$$

and, finally

$$Pr(C_j = c|A_1 = a_1^k, \ldots, A_p = a_p^k) =$$
$$= \frac{Pr(C_j = c)\prod_{s=1}^{p} Pr(A_s = a_s^k|C_j = c)}{\sum_{c_x} Pr(C_j = c_x)\prod_{s=1}^{p} Pr(A_s = a_s^k|C_j = c_x)} \quad (6)$$

Thus, given a test instance d_k, its most probable class is given by:

$$c = \arg\max_{c_x} \left\{ Pr(C_j = c_x) \prod_{s=1}^{p} Pr(A_s = a_s^k|C_j = c_x) \right\} \quad (7)$$

where the prior probabilities $Pr(C_j = c_x)$ and the conditional probabilities $Pr(A_s = a_s^k|C_j = c_x)$ are estimated from the training data.

To predict all C_j's accurately without requesting too much labeling work to u_i, we adopt an *active learning* approach named *uncertainty sampling* [15] based on the *maximum entropy* principle [9]. In an active learning settings the learning algorithm is able to interactively ask the user for the desired/correct labels of unlabeled data instances. A way to reduce the amount of labeling queries to the users is to sample only those data instances whose predicted class is most uncertain. Different measures of uncertainty have been proposed in the literature, e.g., least confidence [8], smallest margin [21] and maximum entropy [9], but for binary classification tasks they are equivalent. Hence, we decide to adopt the maximum entropy principle. According to this principle, the most uncertain data instance d_u is given by:

$$d_u = \arg\max_{d_k} \left\{ -\sum_{c_x} Pr(C_j = c_x|d_k) \log Pr(C_j = c_x|d_k) \right\} \quad (8)$$

Since probabilities $Pr(C_j = c_x|d_k)$ are exactly those computed by the Naive Bayes classifier to take its decision, this principle can be easily adapted to our classification task.

Once all friends' labels are predicted, each entry of the policy matrix $\boldsymbol{M}_i$ can be updated as follows:

$$\forall u_k \in \mathcal{N}(u_i),\ m^i_{kj} = \begin{cases} 1, & \text{if } C_j = allow \text{ for } u_k \\ 0, & \text{if } C_j = deny \text{ for } u_k. \end{cases} \tag{9}$$

The entries of $\boldsymbol{M}_i$ are then used to compute the response matrix $\boldsymbol{R}^C$ as described in Section 3. Note that the original separation-based definition of privacy score can not take advantage of this active learning strategy.

5 Experimental results

In this section we report and discuss the results of an online experiment that we conducted on real Facebook users. The main objectives of our experiment are: i) to study the relationship between the separation-based privacy policies and our circle-based policy definition; ii) to analyze the relationship between the separation-based privacy score ϕ^S_p defined in [16] and our circle-based score ϕ^C_p; iii) to assess the performances of our active learning approach in terms of classification accuracy and privacy score robustness.

The section is organized as follows: first, we describe the data and how we gathered them; then we provide the details of our experimental settings; finally we report the results and discuss them.

5.1 Dataset

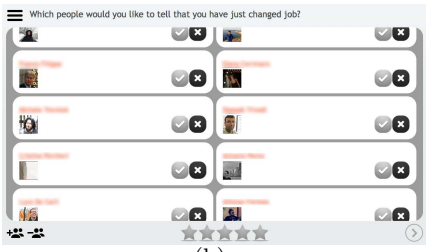
Our online experiments were conducted in two phases. In the first phase we promoted the web page of the experiment[6] where people could voluntarily grant us access to some data related to their own Facebook profile and friends' network. We were not able to access any other information rather than what we asked the permission for, i.e.: email (needed to contact the users for the second phase of our experiment), public profile, friend list, gender, age, work, education, hometown, current location and pagelikes. The participants were perfectly aware about the data we asked for and the purpose of our experiment. In this first phase, data were gathered through a Facebook application developed in Java JDK 8, using Version 1.0 of Facebook Graph API. From March to April 2015, we collected the data of 185 volunteers, principally from Europe, Asia and Americas. The social network consisting of all participants plus their friends is an undirected graph with 75,193 nodes and 1,377,672 edges.

During the second phase, all the remaining participants were contacted for the interactive part of our experiment. First, the participants had to indicate to which level (0=no one, 1=close friends, 2=friends except acquaintances, 3=all friends, 4=friends of friends, 5=everyone on Facebook) they were willing to allow the access to five personal profile topics. The topics were proposed in form of

[6] `http://kdd.di.unito.it/privacyawareness/`

Q1	Which people would you like to tell that you have just changed job?
Q2	If your relationship status changed, which friends would you like to tell?
Q3	After a nice holiday, which friends would you share your photos with?
Q4	With whom would you like to share a comment on current affairs/politics?
Q5	With whom would you like to share your mood or something personal that happened to you?

(a)



(b)

Fig. 1. The five questions (a) and the graphical interface (b) of our online survey

direct questions (see Figure 1(a)) with different levels of sensitivity. We used the answers to fill the response matrix $\boldsymbol{R}^S$. Then, to each participant, we proposed a list of 60 randomly chosen friends and 6 randomly chosen friends of friends (when available). The participants had to indicate to which people they were willing to allow the access to the same five topics. For this phase, we developed a Java JDK 8 mobile-friendly web application leveraging Version 2.0 of Facebook Graph API. Figure 1(b) provides a screenshot of our online survey. We used the answers on friends to fill the response matrix $\boldsymbol{R}^C$. From May 2015 to February 2016, 74 out of 185 participants answered all questions of two surveys. Hence, in our experiments, we consider the network data provided by all 185 participants and the survey data related to the 74 participants who completed the questionnaire. All the data have been anonymized to preserve volunteers' privacy. The entries in the two resulting 74×5 matrices $\boldsymbol{R}^S$ and $\boldsymbol{R}^C$ take values in $\{0, \ldots, 5\}$.

5.2 Separation-based vs. circle-based policies

As a preliminary analysis, we measure how the perception of topic sensitivity changes when the two policies (separation-based and circle-based) are presented to the participants. To this purpose we compare the two response matrix $\boldsymbol{R}^S$ and $\boldsymbol{R}^C$ in several ways. First, we measure the Pearson's correlation coefficient between the two matrices. Given two series of n values $X = x_1 \ldots, x_n$ and $Y = y_i, \ldots, y_n$, the Pearson's coefficient is computed as:

$$\rho(X, Y) = \frac{\sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n} (x_i - \overline{x})^2} \sqrt{\sum_{i=1}^{n} (y_i - \overline{y})^2}} \tag{10}$$

where $\overline{x} = \sum_{i=1}^{n} x_i/n$ and $\overline{y} = \sum_{i=1}^{n} y_i/n$. It basically captures the correlation between the two series of values and ranges between -1 (for inversely correlated sets of values) and $+1$ (for the maximum positive correlation). In our experiment, $n = 74 \cdot 5$. We obtain a moderate positive correlation ($\rho(\boldsymbol{R}^S, \boldsymbol{R}^C) = 0.4632$), that indicates a substantial difference between the two policies. Then, for each question Q_j, we measure the average difference between each entry of the two matrices as $\sum_i (r_{ij}^d - r_{ij}^b)/n$. All the average differences are positive, i.e., the

Table 2. Policy differences in visibility

Measure	**Q1**	**Q2**	**Q3**	**Q4**	**Q5**
A	2	2	4	9	1
B	0	0	4	9	1
C	20	5	19	21	4
D	0	0	4	9	1

given separation-based policies are less restrictive than circle-based ones. In particular, we measure an average difference of 0.54 for Q_1, 0.43 for Q_2, 0.32 for Q_3, 0.35 for Q_4 and 0.15 for Q_5. Moreover, we measure the overall sensitivity of each topic as $\beta_j = \sum_h^\ell \beta_{jh}$ (see Section 3.2) in the two cases. As can be seen in Figure 2(a), all sensitivity values increase when the circle-based policy is adopted. The improved sensitivity perception is confirmed when we look at the users' policies more deeply. In particular, for each question Q_j, we count:

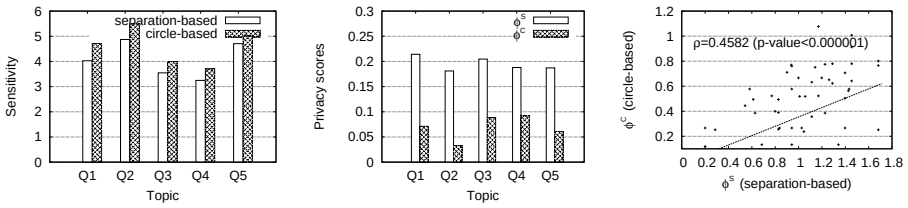
- the number **A** of participants that, in the separation-based test, have made Q_j at least visible to friends of their friends ($r_{ij}^S \geq 4$), but have denied the access to Q_j to some of the friends of their friends in the circle-based test;
- the number **B** of users that have granted the access to some of the friends of their friends in the circle-based test while $r_{ij}^S < 4$ in the separation-based test;
- the number **C** of participants that, in the separation-based test, have made Q_j visible at least to all friends ($r_{ij}^S \geq 4$), but have denied the access to Q_j to some of their friends in the circle-based test $r_{ij}^C < 5$;
- the number **D** of participants that, in the circle-based test, have made Q_j visible to all friends ($r_{ij}^C = 5$), but have denied the access to Q_j to some of their friends in the separation-based test $r_{ij}^S < 3$.

The results in Table 2 indicate that the major differences are on questions Q_3 and Q_4, that are the less sensitive according to Figure 2(a). However, then passing from a separation-based policy to a circle-based one, many users have reviewed their choices in a more restrictive way for question Q_1 and Q_2 as well.

Finally, we also compute the privacy scores $\phi_p^S(u_i, p_j)$ and $\phi_p^C(u_i, p_j)$ for each question Q_j and each participant u_i. The average score values are given in Figure 2(b). Interestingly, although the circle-based policy increases the perception of topic sensitivity, the related privacy scores are sensibly smaller than those computed within the separation-based hypothesis, i.e., the participants have a safer behavior w.r.t. the visibility of the topics. For the sake of completeness, we perform a correlation analysis between the values of $\phi_p^S(u_i)$ and $\phi_p^C(u_i)$ in Figure 2(c). The value of the Pearson's ρ coefficient (0.4582) shows moderate positive correlation between the two series of scores.

5.3 Assessment of the active learning approach

To measure the performances of the active learning approach, we generate 74×5 datasets (one for each pair of users and questions) that we use to train and



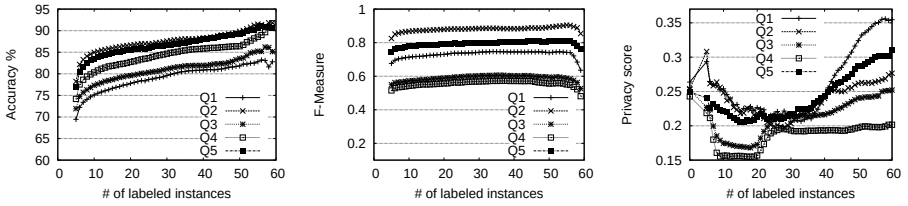
(a) Sensitivity (b) Privacy score (c) Correlation analysis

Fig. 2. Comparative results (separation-based approach vs. circle-based approach)

test the Naive Bayes classifier. These datasets contain, for each friend u_k of a user u_i, the following attributes: *gender* and *age* of u_k, *countryman* (true, if u_k and u_i were born in the same place, *fellow_citizen* (true, if u_k and u_i live in the same place), *coworker* (true, if u_k and u_i work or have worked in the same place), *schoolmate* (true, if u_k and u_i are or have studied in the same school/college/university), and the *Jaccard similarity of page likes* of u_i and u_k. All attribute values are derived from the information extracted by the Facebook profiles, when available. Additionally, we also consider the *list of communities* u_k is part of. To this purpose, we execute a community detection algorithm on the so called "ego-minus-ego" networks (the subgraph induced by the vertex set $\mathcal{N}(u_i) \setminus \{u_i\}$) of all 74 users. We use *DEMON* [7], a local-first approach based on a label propagation algorithm that is able to discover overlapping communities. The algorithm requires two parameters as input: the minimum accepted size for a community ($minCommunitySize$) and a parameter ϵ that determines the minimum overlap two communities should have in order to be merged. In our experiments, we set $minCommunitySize = 3$ (to discard very small communities) and $\epsilon = 0.5$ (to admit an average overlap degree). Finally, each friend has a class variable that takes values in the set $\{allow, deny\}$.

We conduct the experiment as follows. To simulate the active learning framework, for each user and question, i) we start with just five (randomly chosen) labeled friends with which we train the Naive Bayes classifier described in Section 4; ii) we test the classifier on the remaining 55 friends and iii) choose the friend whose prediction is the most uncertain, following the maximum entropy criterion (see Equation 8 in Section 4); iv) we assign to this friend the same label declared by the participant and v) we re-train the classifier on $5 + 1$ instances (friends); vi) finally, we test the new classifier on the remaining 54 instances. We repeat iteratively the last four steps until there are no test instances left.

At the end of each prediction step, we measure the following performance parameters: i) the *Accuracy* of the predictions; ii) the *F-Measure* of the predictions, computed as $F\text{-}Measure = 2 \cdot (precision \cdot recall)/(precision + recall)$ where precision and recall are computed by considering the *deny* class as the positive one; iii) the privacy score (Equation 3) computed by considering both given and predicted $\{allow, deny\}$ labels for all 74 users and applying Equation 9 to calculate matrices $\boldsymbol{M}_i$ and Equation 1 to compute the response matrix $\boldsymbol{R}^C$).



(a) Average Accuracy (b) Average F-Measure (c) Average ϕ_p^C

Fig. 3. Prediction Accuracy vs. Privacy function: average results

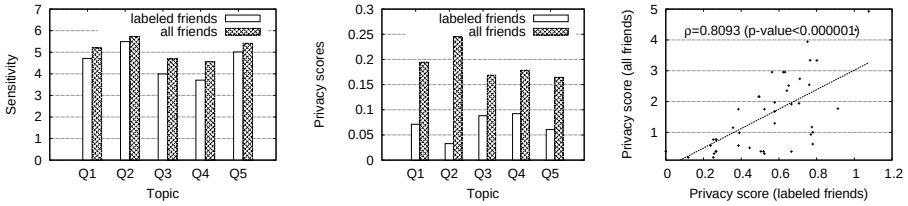
The values of the three parameters are averaged on all 74 users and 30 runs. In each run, the first five labeled friends are chosen randomly. The initial value of the privacy function (when no labels are given) is computed by assigning random labels to all 60 friends.

The results are provided in Figure 3. The values of the three parameters are reported for each question separately. As a general observation, the accuracy of the prediction increases significantly with the number of labeled friends (see Figure 3(a)). The growth of the F-Measure is less sharp, instead (Figure 3(b)). We recall that both measures are computed on the test instances only. The small drop of Accuracy and F-Measure in the last steps can be explained by the fact that misclassification errors of few test instances (less than 5 samples) are more likely to happen. Interestingly, predictions are more accurate for the two most sensitive questions (Q2 and Q5). As for the privacy scores (Figure 3(c)), they start to decrease when few friends (5 to 15) are labeled, then they start to grow almost monotonically and the differences among them are more emphasized. This behavior can be partially explained by noting that, as the amount of labeled friends increases, the sensitivity perceived by the users gets closer to the realistic sensitivity of the five topics.

5.4 Reliability of the predictions

We also study the robustness of the approach by extending the prediction to all participants' friends. Since we do not have the correct labels for friends who do not belong to the list proposed to the participants, we can only measure the privacy scores computed on the basis of the predicted set of labels. We compare these measures with those computed by just considering the labeled friends.

To do that, we first compare the sensitivity values in the two cases (see Figure 4(a)). All questions are subject to an increase of their sensitivity, but when looking at the average privacy scores (Figure 4(b)) we note that all scores are higher than those computed when considering only labeled friends. This means that the visibility of the topics is high. Hence, we perform a correlation analysis in order to check whether the behavior of scores is coherent in the two cases and measure the Pearson's ρ coefficient on the two series of privacy score values. We obtain a Pearson's coefficient of $\rho = 0.8093$ with a p-value



(a) Sensitivity (b) Privacy scores (c) Correlation analysis

Fig. 4. Privacy scores computed with labeled friends only Vs. privacy scores computed on all friends

$p < 0.000001$ (see Figure 4(c)) denoting high positive correlation. This result confirm that: i) the experiments on the limited set of 60 friends per user are significant enough and that, ii) the approach is reliable even for users with a realistic number of friends and few given labels. Note that the overall number of friends of the participants spans between 120 and 1558 (with an average of 435).

6 Conclusions

With the final goal of fostering users' privacy awareness in the Web, we have proposed a privacy score based on an active learning approach to provide the users of online social networks with a measure of their privacy leakage. We have validated experimentally our metrics on an original dataset obtained through an online survey on real Facebook users. The experiments have shown the effectiveness and the reliability of our approach. In particular, we have shown that state-of-the-art metrics are based on a distorted perception of sensitivity of published items. Based on these results, we believe that our framework can be easily plugged into any domain-specific or general-purpose social networking platforms. Furthermore, it may inspire the design of privacy-preserving social networking components for *Privacy by Design* compliant software [5].

Acknowledgments The work presented in this paper has been co-funded by Fondazione CRT (grant number 2015-1638). The authors wish to thank all the volunteers who participated in the survey.

References

1. Akcora, C.G., Carminati, B., Ferrari, E.: Privacy in social networks: How risky is your social graph? In: Proc. of ICDE 2012. pp. 9–19 (2012)
2. Akcora, C.G., Carminati, B., Ferrari, E.: Risks of friendships on social networks. In: Proc. of ICDM 2012. pp. 810–815 (2012)
3. Backstrom, L., Dwork, C., Kleinberg, J.M.: Wherefore art thou R3579X?: anonymized social networks, hidden patterns, and structural steganography. Commun. ACM 54(12), 133–141 (2011)

4. Becker, J., Chen, H.: Measuring privacy risk in online social networks. In: Proc. of Web 2.0 Security and Privacy (W2SP) 2009 (2009)
5. Cavoukian, A.: Privacy by design [leading edge]. IEEE Technol. Soc. Mag. 31(4), 18–19 (2012)
6. Cetto, A., Netter, M., Pernul, G., Richthammer, C., Riesner, M., Roth, C., Sänger, J.: Friend inspector: A serious game to enhance privacy awareness in social networks. In: Proc. of IDGEI 2014 (2014)
7. Coscia, M., Rossetti, G., Giannotti, F., Pedreschi, D.: Uncovering hierarchical and overlapping communities with a local-first approach. TKDD 9(1), 6:1–6:27 (2014)
8. Culotta, A., McCallum, A.: Reducing labeling effort for structured prediction tasks. In: Proc. of AAAI 2005. pp. 746–751 (2005)
9. Dagan, I., Engelson, S.P.: Committee-based sampling for training probabilistic classifiers. In: Proc. of ICML 1995. pp. 150–157 (1995)
10. Dunbar, R.I.M.: Do online social media cut through the constraints that limit the size of offline social networks? Royal Society Open Science 3(1) (2016)
11. Fang, L., LeFevre, K.: Privacy wizards for social networking sites. In: Proc. of WWW 2010 (2010)
12. Ghazinour, K., Matwin, S., Sokolova, M.: Monitoring and recommending privacy settings in social networks. In: Proc. of 2013 EDBT/ICDT Workshop. pp. 164–168 (2013)
13. Kempe, D., Kleinberg, J.M., Tardos, É.: Maximizing the spread of influence through a social network. In: Proc. of SIGKDD 2003. pp. 137–146 (2003)
14. Kosinski, M., Stillwell, D., Graepel, T.: Private traits and attributes are predictable from digital records of human behavior. PNAS 110(15), 5802–5805 (2013)
15. Lewis, D.D., Gale, W.A.: A sequential algorithm for training text classifiers. In: Proc. of SIGIR 1994. pp. 3–12 (1994)
16. Liu, K., Terzi, E.: A framework for computing the privacy scores of users in online social networks. TKDD 5(1), 6 (2010)
17. Liu, Y., Gummadi, P.K., Krishnamurthy, B., Mislove, A.: Analyzing facebook privacy settings: user expectations vs. reality. In: Proc. of SIGCOMM IMC '11. pp. 61–70 (2011)
18. Mislove, A., Viswanath, B., Gummadi, P.K., Druschel, P.: You are who you know: inferring user profiles in online social networks. In: Proc. of WSDM 2010. pp. 251–260 (2010)
19. Mitchell, T.M.: Machine learning. McGraw-Hill (1997)
20. Roberts, S.G.B., Dunbar, R.I.M., Pollet, T.V., Kuppens, T.: Exploring variation in active network size: Constraints and ego characteristics. Social Networks 31(2), 138–146 (2009)
21. Scheffer, T., Decomain, C., Wrobel, S.: Active hidden markov models for information extraction. In: Proc. of IDA 2001. pp. 309–318 (2001)
22. Talukder, N., Ouzzani, M., Elmagarmid, A.K., Elmeleegy, H., Yakout, M.: Privometer: Privacy protection in social networks. In: Proc. of M3SN'10. pp. 266–269 (2010)
23. Wang, Y., Nepali, R.K., Nikolai, J.: Social network privacy measurement and simulation. In: Proce. of ICNC 2014. pp. 802–806 (2014)
24. Wu, L., Majedi, M., Ghazinour, K., Barker, K.: Analysis of social networking privacy policies. In: Proc. of 2010 EDBT/ICDT Workshops (2010)
25. Zheleva, E., Getoor, L.: Privacy in social networks: A survey. In: Social Network Data Analytics, pp. 277–306 (2011)