Exploit of Online Social Networks with Semi-Supervised Learning

Mingzhen Mo and Dingyan Wang and Baichuan Li and Dan Hong and Irwin King, *Senior Member, IEEE*

***Abstract*— With the rapid growth of the Internet, more and more people interact with their friends in online social networks like Facebook[1]. Current online social networks have designed some strategies to protect users' privacy, but they are not stringent enough. Some public information of profile or relationship can be utilized to infer users' private information. Online social networks usually contain little public available information of users (labeled data) but with a large number of hidden ones (unlabeled data). Recently, Semi-Supervised Learning (SSL), which has the advantage of utilizing fewer labeled data to achieve better performance compared to classical Supervised Learning, attracts much attention from the web research community with a massive set of unlabeled data. In our paper, we focus on the privacy issue of online social networks, which is a hot and dynamic research topic. More specifically, we propose a novel SSL framework that can be used to exploit security issues in online social networks. We first introduce the general SSL framework and outline two exploit models with associated strategies within it, e.g., graph-based models and co-training model. Finally, we conduct a series of experiments on real-world data from Facebook and StudiVZ[2] to evaluate the effectiveness of this SSL exploit framework. Experimental results demonstrate that our approaches can accurately infer sensitive information of online users and more effective compared to previous models.**

I. INTRODUCTION

Recently Semi-Supervised Learning (SSL) has become an active research area in the field of machine learning. SSL is a machine learning framework derived from supervised learning and unsupervised learning. SSL contains a set of efficient algorithms, including Expectation Maximization (EM) algorithm [7], co-training method [4], graph-based methods [24], SVM [17], [18], etc. Different from supervised learning only with labeled data and unsupervised learning only with unlabeled data, SSL learns knowledge with a small set of labeled data and a much larger set of unlabeled data. Compared to supervised learning, SSL has the advantage of avoiding high cost in labeling training data and obtaining better performances with a reasonable amount of labeled data empirically. Considering that, SSL can be applied readily into predicting or learning knowledge from the websites which contain massive unlabeled data.

Online social networks, such as Facebook and Twitter[3], are becoming increasingly popular recently. For example,

Mingzhen Mo, Dingyan Wang, Baichuan Li and Irwin King are with the Department of Computer Science and Engineering, The Chinese University of Hong Kong, Shatin, N.T., Hong Kong (email: {mzmo, dywang, bcli, king}@cse.cuhk.edu.hk).

Dan Hong is with the Department of Computer Science and Engineering, The Hong Kong University of Science and Technology, Clearwater Bay, N.T., Hong Kong (email: csdhong@cse.ust.hk).

[1]http://www.facebook.com.

[2]http://www.studivz.net

[3]http://twitter.com/.

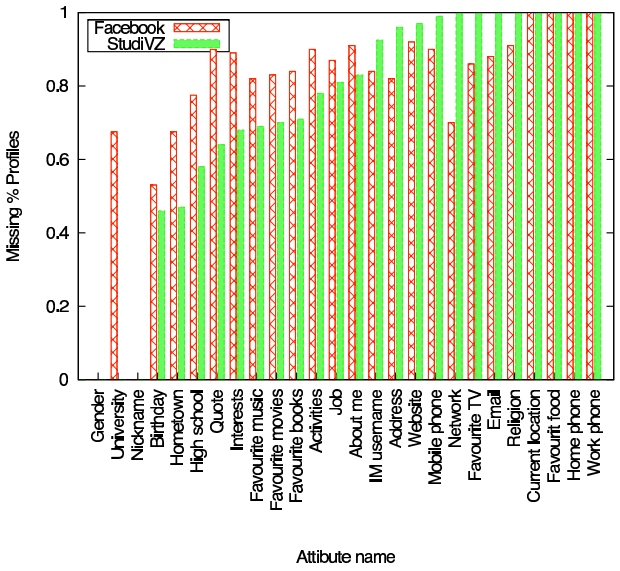

Fig. 1. **Missing % of Users' Profiles on Facebook and StudiVZ**

Facebook currently is utilized by close to 400 million active users and more than 8 billion minutes are spent on Facebook everyday [1]. In these online social networks, people can form social links with others through making friends or joining groups with similar contents. Most of the time, online social networks acquiescently allow people to publish all their profiles. In the meanwhile they also allow people to enable privacy restriction on their profiles. For instance, users can set some or all of their profiles, such as age or university, hidden from strangers.

Currently, one active topic of machine learning application is the privacy exposure problem in online social networks [21], which is to launch exploit models to reveal some private information of users in these websites. Even though a user may set his profile hidden, the friendship and group membership are still visible to the public directly or indirectly. For example, an adversary can find someone's friendship by directly checking his friend list which is available in Facebook. Unfortunately, the public friendship or group information, which online social networks claim to be safe, becomes the potential threat to users' privacy. [8], [9], [21] demonstrate that this information can leak quite a large quantity of sensitive information. Some flaw of these previous models utilizing supervised learning is that they require a lot of labeled data, which increases the exploitative cost. If the exploitation only needs a few labeled data and limited information to reach the same effect, it would be more effective and online social networks should be more aware of this kind of exploitation.

SSL suits well with the scenario that online social net-

works contain little public information and a large number of hidden ones. In our learning model, the public information can be considered as labeled data and that hidden as unlabeled data. According to the statistics shown in Fig. 1, on average 70% users in Facebook have incomplete profiles. It illustrates that labeled data are much fewer than the unlabeled data. Therefore, we propose new exploit models employing SSL. As far as we know, our work is the first one to launch exploitation with SSL effectively in online social networks.

In our SSL exploit framework, we adopt the graph-based and co-training methods of SSL, utilizing personal profile attributes and relationship information of users. The online social network essentially is a connected graph between different people and it includes lots of relationship information. This feature suits graph-based SSL well, which becomes one of the main methods of our SSL exploit framework. As the other method in our framework, the co-training method can apply different classifiers on distinct types of data, like profile attributes and relationship information, and make an agreement between classified results to enhance the learning confidence.

We evaluate our exploit models using real world datasets crawled from Facebook and StudiVZ. In order to evaluate the effectiveness in privacy exposure, we compare our exploit models with previous one applying supervised learning. The experiment results indicate that our SSL exploit framework is more efficient with a small number of labeled data.

Our contributions include the following:

- **Semi-supervised learning framework is firstly used for privacy exposure in online social networks.** As far as we know, our framework is the first one to employ semi-supervised learning as exploit model to expose privacy in online social networks.
- **Semi-supervised learning is superior to supervised learning for privacy exposure in online social networks.** Experiments on the real datasets demonstrate that semi-supervised learning can really achieve a higher accuracy in exposing privacy.
- **Semi-supervised learning aggravates the privacy exposure problem in online social networks.** Applying semi-supervised learning exploit models, users' private information is exposed more seriously than before.

We define this security problem and describe the details of our exploit models on SSL framework in Section II. Section III provides the experimental analysis of our framework and comparison with other models. We introduce related work in Section IV, then presents the conclusion and discusses the future work in Section V.

II. The Problem and SSL Framework

A. Problem Definition

***Definition** 1 (**Social Network**):* We define a social network as a graph $G(V, E)$, where V denotes the set of vertices (users) and E denotes the set of edges (relations) among vertices.

For each vertex (user) $v_i, v_i \in V$, a feature vector P_i describes the personal attributes

$$P_i = (p_i^1, p_i^2, \cdots, p_i^{n_f}),\ p_i^j \in \mathbf{R}, j \in \{1, 2, \cdots, n_f\}, \quad (1)$$

where feature p_i^j describing the j-th attribute of vertex v_i, and n_f is the total number of features.

The weight of relationship on an edge, $W_{i,j},\ v_i, v_j \in V$, which measures the similarity of v_i and v_j in several aspects, is a vector

$$W_{i,j} = (w_{i,j}^f, w_{i,j}^g, w_{i,j}^n),\ w_{i,j}^f, w_{i,j}^g, w_{i,j}^n \in \mathbf{R}, \quad (2)$$

where $w_{i,j}^f$ is a weight for friendship, $w_{i,j}^g$ for group membership and $w_{i,j}^n$ for network relationship. Then all $W_{i,j}$ can form a weight matrix W.

***Definition** 2 (**Labeled Data and Unlabeled Data**):* We define the labeled dataset as $V_l = \{v_{i_1}, v_{i_2}, \cdots, v_{i_l}\}$, $i_1, i_2, \cdots, i_l \in \{1, \cdots, l+n\}$, whose corresponding labels set is $\{L_{i_1}, L_{i_2}, \cdots, L_{i_l}\}$, $L_{i_j} \in L$, $j \in \{1, \cdots, l\}$, where $L = \{1, 2, \dots, n_{class}\}$, l is the number of labeled data and n_{class} is the number of classes.

The unlabeled dataset is $V_u = \{v_{u_1}, v_{u_2}, \cdots, v_{u_n}\}$, $u_1, u_2, \cdots, u_n \in \{1, \cdots, l+n\}$, whose corresponding labels set is $\{L_{u_1}, L_{u_2}, \cdots, L_{u_n}\}$, $L_{u_j} \in L, j \in \{1, \cdots, n\}$, where n is the number of unlabeled data.

In the following of the paper, we always mention the concept of the "class" that is a collection of vertices with the same label. In the experiments, we assume the classes are disjointed and guarantee this by data preprocessing.

Take an example in online social networks, the university name of a user is a label, like "Harvard" and "Cambridge", if learning methods want to classify users according their universities. Any two users with the same label are in the same class.

Now the objective is to predict the labels $\{\hat{L}_{u_1}, \hat{L}_{u_2}, \cdots, \hat{L}_{u_n}\}$, $\hat{L}_{u_1}, \hat{L}_{u_2}, \cdots, \hat{L}_{u_n} \in L$ for the corresponding vertices in $V_u = V - V_l$. We hope that predicting result can agree with the true labels $\{L_{u_1}, L_{u_2}, \cdots, L_{u_n}\}$, $L_{u_1}, L_{u_2}, \cdots, L_{u_n} \in L$. Thus, our objective function is as follows.

***Definition** 3 (**Objective Function**):* The objective function that is to be minimized is defined as

$$\sum_{j=l+1}^{l+n} f_{loss}(\hat{L}_{i_j}), \quad (3)$$

with the loss function being defined as

$$f_{loss}(\hat{L}_i) = \begin{cases} 1 & \hat{L}_i \neq L_i \\ 0 & \hat{L}_i = L_i \end{cases}, \quad (4)$$

where L_i is the real label and $\hat{L}_i$ is the predicted one for vertex v_i.

B. SSL Framework

In this section, we explain why we use these SSL algorithms in our exploitation framework. SSL can be considered as the extension of unsupervised and supervised learning. Thus, it uses both labeled data which supervised learning uses and unlabeled data which are used by unsupervised learning. Here labeled data are those data that we know which classes they belong to and unlabeled data are data that we only know some feature information of the data except their classes.

There are two observations which induct us to apply effective SSL models on online social networks.

***Observation** 1:* Graph-based SSL is especially proposed for learning on graph-structure data. That is suitable for online social networks which are expressed as graphs in mathematics.

In graphs, information can be naturally transmitted from labeled data to unlabeled data through edges with variable weight. This smooth characteristic actually is the important assumption of graph-based SSL. In other word, graph-based SSL can maximize its advantage in solving these problems with graph structure.

***Observation** 2:* Co-training SSL applies different classifiers on data with distinct structure types, e.g. statistical type and graph-structure type. This is suitable for classifying on online social networks.

In general, we can divide all the information on online social networks into two views: relationship information and personal profile information. In most cases, these two views of information are conditionally independent, which satisfies the assumption of co-training semi-supervise learning. In this case, co-training can effectively learn knowledge from different views.

According to these two observations, we use two algorithms to construct the exploit models, local and global consistency (LGC) graph-based SSL [23] and co-training SSL [4]. In the co-training SSL, graph-based SSL with harmonic function [24] and supervised learning are used as the two classifiers. The whole framework is shown in Fig. 2.

In the following parts, we explain two exploit models in detail correspondingly.

C. Local and Global Consistency Graph-Based SSL

Local and Global Consistency (LGC) algorithm is a learning method that considers both local consistency and global consistency in a graph G. In this way, it will obtain a more accurate learning result than other local learning methods.

From the algorithm inputs, we obtain a weight matrix of a weighted graph, W, which is a symmetric and semi-definition matrix. Now, let

$$D_{ii} = \sum_{j=1}^{l+n} w_{i,j},\ i \in \{1, \cdots, l, l+1, \cdots, l+n\} \tag{5}$$

and D be the $(l+n) \times (l+n)$ diagonal matrix by placing $D_{ii},\ i \in \{1, \cdots, l+n\}$ on the diagonal. Then the unnormalized

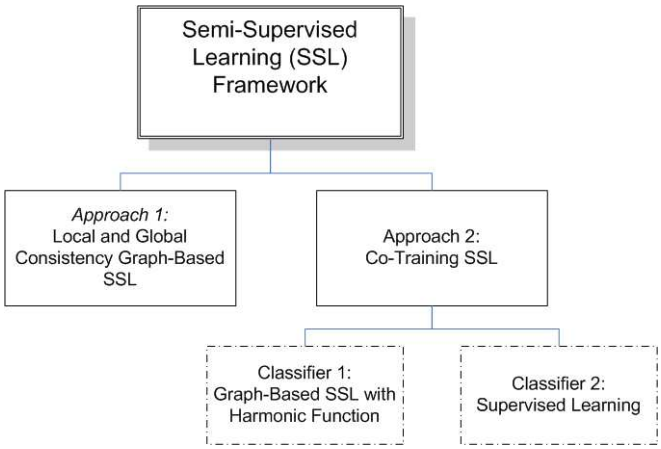

Fig. 2. **Semi-Supervised Learning Framework**

graph Laplacian matrix L^g is defined as

$$L^g = D - W. \tag{6}$$

Local Term. The goal of local consistency is to optimize the following cost to recover the unlabeled information:

$$\min_{\hat{L} \in \mathcal{L}^{l+n}} tr(\hat{L}^\top L^g \hat{L}), \tag{7}$$

where vector $\hat{L} = (\bar{L}_{label},\ \hat{L}_{unlabel})^\top$ is the predicted result, $\bar{L}_{label} = (L_{i_1}, L_{i_2}, \cdots, L_{i_l})$, $\hat{L}_{unlabel} = (\hat{L}_{u_1}, \hat{L}_{u_2}, \cdots, \hat{L}_{u_n})$.

Global Term. In order to predict unlabeled information more accurately, an additional term is considered to keep the consistency in a global view.

Thus, we formalize a new objective function:

$$\min_{\hat{L} \in \mathcal{L}^{l+n}} tr\{\hat{L}^\top L^g \hat{L} + \mu(\hat{L} - \bar{L})^\top (\hat{L} - \bar{L})\}, \tag{8}$$

where vector $\bar{L} = (\bar{L}_{label},\ \bar{L}_{unlabel})^\top$ is the real label information, $\bar{L}_{unlabel} = (L_{u_1}, L_{u_2}, \cdots, L_{u_n})$, $\mu > 0$.

Then define the initial value $\hat{L}_0 = [\bar{L}_{label}\ \underbrace{0, 0, \cdots, 0}_{n}]^\top$. With this initialization, following shows the algorithm.

Algorithm 1 Local and Global Consistency Graph-Based Semi-Supervised Learning

1: Graph construction, obtaining W.
2: Compute the matrix $S = D^{-1/2} W D^{-1/2}$.
3: Initialize $\hat{L}_0$.
4: **while** $\hat{L}_t$ does not convert **do**
5: $\quad \hat{L}_{t+1} = \alpha S \hat{L}_t + (1-\alpha)\hat{L}_0,\ \alpha \in (0,1)$.
6: **end while**

Besides this iterative approach, a directly calculating method was proved to be an equivalent but more efficient approach. [23] has proved that the result of this formula

$$\hat{L}^* = (I - \alpha S)^{-1} \hat{L}_0, \tag{9}$$

is equal to the limit of $\hat{L}_t, t \in \mathbf{N}^+,\ \alpha \in (0,1)$. For the consideration of efficiency, Eq. (9) is a much better choice and implemented in the experiments.

D. Co-Training SSL

Co-training SSL is a procedure that more than one classifier classify different features of the data independently and make an agreement in each step of iterative learning. In contrast with supervised learning, its advantage is that fewer labeled data are used and the similar, or even better, performance is obtained in learning knowledge.

In social network data, we naturally have two views. One is personal information view, and the other is relationship information view. These two views contain different types of data. In this condition, co-training will have great effect in learning result after the recommendation and agreement between classifiers in each iterative step.

We have two classifiers, f^r for relationship information view and f^p personal information view. f^r is a graph-based classifier which can effectively deal with relational information and f^p is a statistics-based classifier which is good at utilizing data with the nature of sets by statistic model. The co-training SSL algorithm is shown in Algorithm 2.

Algorithm 2 Co-Training Semi-Supervised Learning

Input:

Training sample $L = \{(v_{i_1}, L_{i_1}),(v_{i_2}, L_{i_2})\cdots,(v_{i_l}, L_{i_l})\}$.

1: **while** unlabeled data are not used up **do**
2: Training a relationship-information-view classifier f^r, and a personal-information-view classifier f^p.
3: Classify the remaining unlabeled data with f^r and f^p separately.
4: Combine f^r's top k_r most-confident predictions $(v, f^r(v))$ and f^p's top k_p most-confident predictions $(v, f^r(p))$.
5: Add the top k most-confident combination result to L, and remove those from the unlabeled data.
6: **end while**

Though co-training is a wrapper method that does not matter what algorithms of the two classifiers f^r and f^p are, we want to show some details of them in order to emphasize that the classifiers are very suitable for this application.

We classify the relational information by graph-based SSL with harmonic function and deal with the personal information by supervised learning model (Fig. 2). Here we focus on the relational information classifier f^r, and the data translation between co-training SSL and classifiers.

Graph-Based SSL Classifier. Harmonic function algorithm is a basic graph-based SSL method. Though it mainly focuses on the local consistency of a graph, it is simple, efficient and satisfying in some ways.

Here we show how this classifier works. Similar to LGC graph-based SSL, we can also calculate L^g from input data. The next step is to normalize L^g by ordering vertices in the way that labeled data are listed first and then the unlabeled ones. Thus, the new L^g can be partitioned into four submatrices

$$L^g = \begin{bmatrix} L^g_{ll} & L^g_{ln} \\ L^g_{nl} & L^g_{nn} \end{bmatrix}. \tag{10}$$

Finally, solving the constrained optimization problem with Lagrange multiplies with matrix algebra, the final label probability $Pr_{unlabel}$ for the unlabeled data can be calculated by

$$Pr_u = -{L^g_{nn}}^{-1} L^g_{nl} \bar{L}_{label}, \tag{11}$$

where vector $\bar{L}_{label} = (L_{i_1}, L_{i_2}, \cdots, L_{i_l})^\top$, $i_1, i_2, \cdots, i_l \in \{1, \cdots, l+n\}$, and the scalar $Pr_u(i,j)$ in matrix Pr_u means the probability of vertex v_i belonging to the j-th class.

Data Translation. Data between co-training SSL and the classifier graph-based SSL can not be used directly and needs some translations. In the direct output of graph-based classifier, it shows the probability, $Pr_u(i,j)$, of every vertex v_i, $i \in \{1, 2, \cdots, n\}$ belonging to every label L_j, $j \in \{1, 2, \cdots, n_{class}\}$. First, we need to make a decision that which label a vertex v_i should belong to by this way

$$\hat{L}_i = \arg\max_j Pr_u(i,j),\ j \in \{1, 2, \cdots, n_{class}\}. \tag{12}$$

Second, we use these values of probability as the confident information for recommendation in co-training SSL. By this confident information, co-training can enhance the learning result in the procedure of agreement of the recommended data.

The algorithm of the graph-based classifier with data translation in post-processing is shown in Algorithm 3.

Algorithm 3 Graph-Based Semi-Supervised Learning Classifier

1: Construct graph, and obtain weighted matrix W.
2: Compute diagonal matrix D.
3: Compute diagonal matrix L^g.
4: Normalize L^g.
5: Solve the final label result by Eq. (11).
6: Translate Pr_u into class labels (Eq. (12)) and confident values.

III. EXPERIMENTS

In the experiments, we employ both two SSL models and a supervised learning model as comparison. In order to evaluate the learning models, two real-world datasets are applied and the accuracy is considered as the common evaluation criterion. The results of them illustrate that SSL models assuredly expose users' private information and achieve higher accuracy than the supervised learning model does.

Experiments Objective. The task of experiments is to expose which university a user comes from. In this case, university name is set as the private attribute (label) and other attributes are treated as features for training. If a vertex (user) belongs to some class, it means he or she currently is or once was an undergraduate student from the corresponding university. For simplicity's sake, graduate schools are not considered.

Methods Contrast. Two SSL methods explained in Section II and one supervised learning method k-Nearest Neighbor (*kNN*) algorithm as comparison are applied in experiments.

TABLE I

STATISTICS OF FACEBOOK AND STUDIVZ DATASETS

Dataset	Facebook	StudiVZ
Vertices	$5,000$	$1,423$
Edges	$31,442$	$7,769$
Groups	61	406
Networks	78	0
Classes	3	6

Dataset Selection. In order to evaluate the effectiveness of various methods, two different datasets are crawled to check the accuracy of learning private attribute's value. One is Facebook and the other is StudiVZ, a German social network website.

In the following sections, we firstly give the description of these datasets, then we describe data preprocessing and the process of experiments. Finally we make comparison and analysis according to experiment results.

A. Dataset Description

The datasets in our experiment are crawled from two real online social network websites: Facebook and StudiVZ. Table I gives detail statistics of these two datasets.

Facebook Dataset. It has sufficient number of vertices and all kinds of relational information, thus it is similar to the situation of real world. Based on crawled data, we build a graph which contains $5,000$ vertices and $31,442$ edges. Each vertex represents one user and it contains $n_f = 26$ features such as nickname, gender, birthday, high school, university, favorite books, favorite movies, home town, etc. An undirected edge between two vertices means these two users are friends in Facebook. For this dataset, $n_{class} = 3$ labels of universities are used here. Data distribution is shown in Table II.

TABLE II

DATA DISTRIBUTION OF FACEBOOK DATASET

Univ.	CUHK	HKUST	(Others)
Class Size	68	532	$4,400$

StudiVZ Dataset. Comparing with Facebook dataset, StudiVZ dataset has fewer missing values in personal profile and more group information. Moreover, it doubles the number of classes and brings challenges to the learning models. Also, we construct a graph which is made up of $1,423$ vertices and $7,769$ edges from crawled data. $n_{class} = 6$ university names are used as class (label) names. Table III gives the number of users in each class.

TABLE III

DATA DISTRIBUTION OF STUDIVZ DATASET

Univ.	LMU Muenchen	Uni Wien
Class Size	128	79
Univ.	**Uni Frankfurt am Main**	**TU Wien**
Class Size	74	70
Univ.	**Uni Bayreuth**	**(Others)**
Class Size	98	974

B. Data Preprocessing

A series of data preprocessing [19] such as feature selection, data cleaning and data translation are conducted before running algorithms.

*1) **Feature Selection**:* There are $n_f = 26$ features for each user, however, not all of them are needed. In fact, some features such as nickname provide little information for classification. Besides, most people fill only a few of these features, for instance, very few people provide information for work phone and current location. Thus, according to the statistic result for 26 features (see Fig. 1), we select top three features for which most people provide information. After excluding nickname, we finally choose gender, birthday and home town as basic profile information of each user (vertex) for classifying.

For relational information, it also needs to select the helpful data. The original group number of Facebook and StudiVZ data is 371 and $14,400$. Among these groups, most of them are made up with only a small number of people. Thus, lots of small groups are removed and finally 61 and 406 groups left respectively. Networks are processed similar to groups. Apart from that, some networks whose names explicitly reveal universities' names, such as "CUHK", "HKUST" and "LMU Muenchen", are removed manually.

*2) **Data Translation**:* Since home town is just a string and it is a bad way to calculate two users' home town similarity through comparing two strings, we translate home town to its longitude and latitude values through Google maps API[4].

Missing Value. Although top three features on which most users fill information are selected, the number of missing value is still very large and noise information, like birthday with value "$\langle 1/1/0001 \rangle$", exists widespreadly in datasets. For age, missing data are filled with average value of existed data and noise data are treated as missing ones. For gender, 0.5 is used to represent missing value (1 represents male and 0 represents female). For hometown, missing data are filled respectively with average value of longitude and latitude of his friends. Thus, a user's basic information could be expressed by using a vector which contains its age, gender, hometown's longitude and latitude.

Similarity. The value of every attribute in users' profile is scaled into $[0, 1]$ and the cosine similarity between any two profile vectors is calculated. If both of them fail to provide at least 50% information, we set the cosine similarity with mean value.

Another kind of similarity is obtained from relational information, i.e., friendship, group and network membership. Two users' friendship similarity is computed through 1 divided by the shortest hop(s) between them. For example, if two users are friends (linked directly), the hop between them is 1 and the similarity is also 1; if two users are not directly linked but they both link to another user, the shortest hops between them is 2 and thus we set their friendship similarity as $\frac{1}{2}$. Furthermore, a similarity value of group (network)

[4]http://code.google.com/apis/maps/.

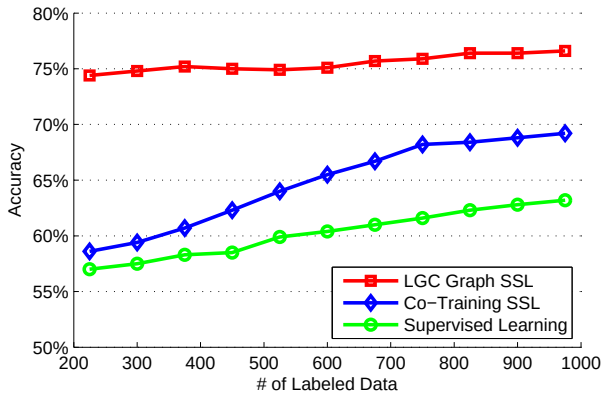

Fig. 3. **Experiment Result on Facebook Dataset with** $5,000$ **Users**

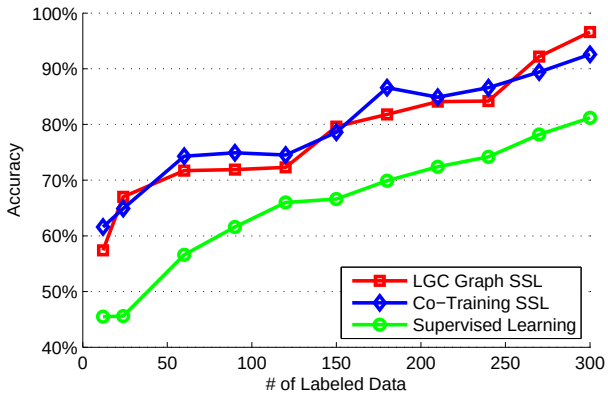

Fig. 4. **Experiment Result on StudiVZ Dataset with** $1,423$ **Users**

membership is calculated by

$$w_{i,j}^{g} = \frac{1}{2}\left(\frac{|G_i \bigcap G_j|}{|G_i| + \varepsilon} + \frac{|G_i \bigcap G_j|}{|G_j| + \varepsilon}\right), \tag{13}$$

where $|G_i|$ means the number of groups user i joined and ε is a small positive constant to avoid dividing by zero.

At last, we calculate the similarity of personal and relational information using the following cosine measure,

$$S(x, y) = \frac{x^{\top} A y}{N(x, y)}, \tag{14}$$

where A is the identity matrix I and $N(x, y)$ is the product of the L_2 norms of vector x and y. This method is widely used and can express the similar information in most cases.

C. Experiment Process

1) ***Labeled Data Selection****:* Labeled data are selected randomly with two constrains below:

- There must be labeled data for each class;
- The number of labeled data in all classes are similar.

2) ***Decision of Label Value****:* From the algorithms' results, we obtain $\langle Pr_u(i,1), Pr_u(i,2), \cdots, Pr_u(i,n_{class})\rangle$ which means the probabilities of an unlabeled data v_i belonging to every label, respectively. And the final label value $\hat{L}_i$ can be attained by Eq. (12).

D. Experiment Result

Table IV and V give the results of experiments, from which various algorithms' performance can be evaluated. Figure 3 and Figure 4 describe the accuracy of private information with Facebook and StudiVZ datasets respectively. What's more, the results of supervised learning are provided for comparison.

1) ***Facebook****:* Figure 3 illustrates various algorithms' performance on Facebook dataset. First of all, it is clear to see in most cases the results of all SSL methods are obviously superior to supervised learning. In specific, LGC graph-based and co-training SSL methods perform better than supervised learning, and LGC graph-based method obtains the best learning accuracy among all three. Second, even there is only a few of labeled data, LGC graph-based method can still make good predictions. For example, the accuracy of LGC graph-based SSL is 74.40% when only 225 (4.5%) data are labeled, while both co-traning SSL and supervised learning method's accuracies are less than 60%. The last point is that the performance of co-traning SSL is not as good as LGC graph-based SSL. We conjecture it is due to the missing of many users' profile information which lead to misclassification of classifier f^p.

In contrast with [22], our SSL models have the distinct advantage. In that paper, experiments on Facebook dataset with 50% labeled data are shown and the accuracy of all the experiments is between 50.0% and 72.5%. However, dealing with 3 classes, only $4.50\% \sim 19.50\%$ labeled data are chosen in our case. Moreover, the accuracy is improved to 76.60% (better than 72.5% in [22]) when less than 20% data are labeled. It demonstrates the advantage of SSL is that only a few labeled data is needed in learning.

2) ***StudiVZ****:* Figure 4 gives similar results, that is, both LGC graph-based SSL and co-training SSL outperform supervised learning. Specifically, the performance of LGC graph-based SSL and co-training SSL are nearly the same. When the labeled data size is above 150, they both obtain a rate of 80% of correctly predicting university names. They can even achieve an accuracy of 90% when labeled data size increases to 270. In contrast, the supervised learning algorithm merely obtains the rate of 80% when given 300 labeled data.

Comparing with Facebook dataset, the learning results on StudiVZ dataset are more accurate because there is more effective group information in StudiVZ dataset and the profile information is more completed. On one side, more sufficient relational information, like group membership, help LGC graph-based learning method a lot. On the other side, the completeness of information balances both classifiers in co-training method, and less incorrectness is accumulated or amplified.

3) ***Summary****:* From the above results, an evident conclusion can be made that in general SSL is far superior to supervised learning for privacy exposure in online social

TABLE IV
ACCURACY OF LEARNING ON FACEBOOK DATA WITH $5,000$ USERS

Number of Labeled Data	% of Labeled Data	LGC Graph SSL	Co-Training SSL	Supervised Learning
225	4.50%	**74.40**%	58.60%	57.00%
300	6.00%	**74.80**%	59.40%	57.50%
375	7.50%	**75.20**%	60.70%	58.30%
450	9.00%	**75.00**%	62.30%	58.50%
525	10.50%	**74.90**%	64.00%	59.90%
600	12.00%	**75.10**%	65.50%	60.40%
675	13.50%	**75.70**%	66.70%	61.00%
750	15.00%	**75.90**%	68.20%	61.60%
825	16.50%	**76.40**%	68.40%	62.30%
900	18.00%	**76.40**%	68.80%	62.80%
975	19.50%	**76.60**%	69.20%	63.20%

TABLE V
ACCURACY OF LEARNING ON STUDIVZ DATA WITH $1,423$ USERS

Number of Labeled Data	% of Labeled Data	LGC Graph SSL	Co-Training SSL	Supervised Learning
12	0.84%	57.40%	**61.60**%	45.50%
24	1.69%	**67.00**%	64.90%	45.60%
60	4.22%	71.70%	**74.30**%	56.60%
90	6.32%	71.90%	**74.90**%	61.60%
120	8.43%	72.30%	**74.50**%	66.00%
150	10.54%	**79.60**%	78.60%	66.60%
180	12.65%	81.80%	**86.60**%	69.90%
210	14.76%	84.10%	**84.90**%	72.40%
240	16.87%	84.20%	**86.60**%	74.20%
270	18.97%	**92.20**%	89.40%	78.20%
300	21.08%	**96.60**%	92.60%	81.20%

networks when only a small number of labeled data exist.

IV. RELATED WORK

Since the online social networks became popular, there has been a growing interest in the security of users' privacy under the current privacy protection. Employing machine learning methods become a popular approach to expose privacy in online social networks. In previous works, the exploiting models only include unsupervised learning models and supervised learning ones.

For exploit models with unsupervised learning in online social networks, they usually employ classical clustering methods to cluster or group objects based on the similarity of attributes or structural information, like *K*-means clustering [12]. Recently, Neville and Jensen [14] use a spectral clustering method based on the node links in the data to discover groups and then classify the nodes with these groups. Airoldi et al. [2] propose a novel mixed-membership clustering of relational data which also can be utilized to disclose privacy in social networks.

As for exploit models with supervised learning methods, they utilize various attributes, including privacy object attributes and linked user profiles [8] to learn the knowledge. Besides, He et al. [9] predict private attributes using Bayesian network with friendship links. A more comprehensive review about collective classification can be found in Sen's work [16]. Most recently Zheleva and Getoor [21] propose a novel model using group-based supervised classification with group membership information apart from friend links.

However, supervised learning has a flaw that it needs a large size of labeled data. Generally, with sufficient labeled data and correct assumptions, supervised learning methods have a higher accuracy than unsupervised learning, which can be thought of as a case that a 'teacher' helps a 'student' in learning. Nevertheless, labeled data are often very time consuming and expensive to obtain, as they require the efforts of human. Especially in online social networks like Facebook.

SSL can be divided into several typical kinds of models, including generative model [7], co-training model [4], graph-based model [24], SVM [17], [18], etc. Based on our application background, we mainly introduce co-training and graph-based methods into our exploit model.

- **Co-training method.** Co-training method [4], [13] assumes that data features can be split into two conditional independent views. And then two initial separate classifiers can be trained with the labeled data, on the two views respectively. Each classifier can classify the

unlabeled data and teaches the other classifier with the most confident unlabeled samples. Nigam and Ghani [15] demonstrate that co-training out-performs generative models and EM when the conditional independence assumption of the two views holds. There are some works about applications using co-training, such as information extraction from text using co-training and co-EM [6], [11].

- **Graph-based method.** Graph-based SSL methods define a graph whose nodes are the labeled and unlabeled samples in the dataset and whose weighted edges represent the similarity between two samples. Graph-based methods are nonparametric, discriminative in nature. Blum and Chawla [3] pose SSL as a graph mincut problem. In the binary labels case, positive labels are described as sources and negative labels as sinks. The goal is to find out a minimum set of edges whose removal blocks all flow from the sources to the sinks. Another graph-based SSL algorithm proposed in [24] is the harmonic function that is a function which has the same values as given labels on the labeled data and satisfied the weighted average property on the unlabeled data. Based on the original harmonic function method, [23] proposes the Local and Global Consistency graph-based method which improves harmonic function method. Experiments of the previous works show that graph-based semi-supervised methods can have a good performance if we can construct suitable graphs. So graph construction is discussed in [5], [20], [10], including *kNN*, ϵ*NN*, etc. Relying on the characteristic of online social networks, our SSL exploit models modify and employ graph-based semi-supervised methods in order to have a better performance in real datasets.

V. CONCLUSION

In contrast to supervised learning, SSL predicts sensitive information in online social network more accurately, by combining a few labeled data and a large number of unlabeled data. In our SSL exploit framework, we use local and global consistency graph-based SSL and co-training SSL to expose the private information in online social networks. From the result, we find it is possible to learn hidden users' attributes based on relational information and profile similarity among users. As a result, the users' security is never secure and SSL framework makes the network security problem more serious. Thus, there is a need of protection based on different users' privacy setting. With this protection, the learning accuracy of SSL will decline and users' privacy will be protected.

ACKNOWLEDGMENT

The work described in this paper is supported by grants from the Research Grant Council of the Hong Kong Special Administrative Region, China (Project No.: CUHK $4128/08E$ and Project No.: 619308).

REFERENCES

[1] http://www.facebook.com/press/info.php?statistics.

[2] E. Airoldi, D. Blei, S. Fienberg, and E. Xing. Mixed membership stochastic blockmodels. *The Journal of Machine Learning Research*, 9:1981–2014, 2008.

[3] A. Blum and S. Chawla. Learning from labeled and unlabeled data using graph mincuts. In *Proceedings of the Eighteenth International Conference on Machine Learning table of contents*, pages 19–26. Citeseer, 2001.

[4] A. Blum and T. Mitchell. Combining labeled and unlabeled data with co-training. In *Proceedings of the eleventh annual conference on Computational learning theory*, page 100. ACM, 1998.

[5] M. Carreira-Perpinan and R. Zemel. Proximity graphs for clustering and manifold learning. *Advances in neural information processing systems*, 17, 2005.

[6] M. Collins and Y. Singer. Unsupervised models for named entity classifi cation. In *Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*, pages 189–196, 1999.

[7] A. Dempster, N. Laird, D. Rubin, et al. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(1):1–38, 1977.

[8] L. Getoor and B. Taskar. *Introduction to statistical relational learning*. The MIT Press, 2007.

[9] J. He, W. Chu, and Z. Liu. Inferring privacy information from social networks. *Lecture Notes in Computer Science*, 3975:154, 2006.

[10] M. Hein, J. Audibert, and U. Von Luxburg. Graph Laplacians and their convergence on random neighborhood graphs. *Journal of Machine Learning Research*, 8:1325–1368, 2007.

[11] R. Jones. *Learning to Extract Entities from Labeled and Unlabeled Text*. PhD thesis, University of Utah, 2005.

[12] J. Macqueen. Some methods for classifi cation and analysis of multivariant Observations. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, page 281. University of California Press, 1967.

[13] T. Mitchell. The role of unlabeled data in supervised learning. In *In Proceedings of the Sixth International Colloquium on Cognitive Science*, 1999.

[14] J. Neville and D. Jensen. Leveraging relational autocorrelation with latent group models. In *Proceedings of the 4th international workshop on Multi-relational mining*, page 55. ACM, 2005.

[15] K. Nigam and R. Ghani. Analyzing the effectiveness and applicability of co-training. In *Proceedings of the ninth international conference on Information and knowledge management*, pages 86–93. ACM New York, NY, USA, 2000.

[16] P. Sen, G. Namata, M. Bilgic, L. Getoor, B. Gallagher, and T. Eliassi-Rad. Collective classifi cation in network data. *AI Magazine*, 29(3):93–106, 2008.

[17] V. Vapnik. Structure of statistical learning theory. *Computational Learning and Probabilistic Reasoning*, page 3, 1996.

[18] Z. Xu, M. Cluster, G. Saarbrucken, R. Jin, J. Zhu, E. Zurich, S. Zurich, I. King, M. Lyu, and Z. Yang. Adaptive Regularization for Transductive Support Vector Machine. In *Proceedings of Advances in Neural Information Processing System 22 (NIPS2009)*, pages 2125–2133, 2009.

[19] Z. Xu, R. Jin, J. Ye, M. Lyu, and I. King. Non-monotonic feature selection. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 1145–1152. ACM, 2009.

[20] X. Zhang and W. Lee. Hyperparameter learning for graph based semi-supervised learning algorithms. *Advances in Neural Information Processing Systems*, 19:1585, 2007.

[21] E. Zheleva and L. Getoor. To join or not to join: the illusion of privacy in social networks with mixed public and private user profi les. In *Proceedings of the 18th international conference on World wide web*, pages 531–540. ACM, NY, USA, 2009.

[22] E. Zheleva and L. Getoor. To join or not to join: The illusion of privacy in social networks with mixed public and private user profi les. In *18th International World Wide Web Conference*, pages 531–531, April 2009.

[23] D. Zhou, O. Bousquet, T. Lal, J. Weston, and B. Scholkopf. Learning with local and global consistency. In *Advances in Neural Information Processing Systems 16: Proceedings of the 2003 Conference*, pages 595–602. The MIT Press, 2004.

[24] X. Zhu, Z. Ghahramani, and J. Lafferty. Semi-supervised learning using Gaussian fi elds and harmonic functions. In *The 20th Internatioanl Conference on Machine Learning (ICML)*, 2003.