# Image and Attribute Based Convolutional Neural Network Inference Attacks in Social Networks

Bo Mei, *Student Member, IEEE,* Yinhao Xiao, *Student Member, IEEE,* Ruinian Li, *Student Member, IEEE,* Hong Li, Xiuzhen Cheng, *Fellow, IEEE,* and Yunchuan Sun, *Member, IEEE*

**Abstract**—In modern society, social networks play an important role for online users. However, one unignorable problem behind the booming of the services is privacy issues. At the same time, neural networks have been swiftly developed in recent years, and are proven to be very effective in inference attacks. This paper proposes a new framework for inference attacks in social networks, which smartly integrates and modifies the existing state-of-the-art Convolutional Neural Network (CNN) models. As a result, the framework can fit wider applicable scenarios for inference attacks no matter whether a user has a legit profile image or not. Moreover, the framework is able to boost the existing high-accuracy CNN for sensitive information prediction. In addition to the framework, the paper also shows the detailed configuration of Fully Connected Neural Networks (FCNNs) for inference attacks. This part is usually missing in the existing studies. Furthermore, traditional machine learning algorithms are implemented to compare the results from the constructed FCNN. Last but not least, this paper also discusses that applying Differential Privacy (DP) can effectively undermine the accuracy of inference attacks in social networks.

**Index Terms**—inference attack, social network, neural network, machine learning.

## 1 Introduction

Nowadays, online social networks are an important part for everyone. In the US, the amount of time that people spend on online social networks is constantly increasing; 30% of all time spent online is now allocated to social network interaction. Social network platforms constantly evolve their tools and options to attract and engage new audiences.

There are billions and billions of messages being sent every day in online social networks. However, one unignorable problem behind the booming of the services is privacy issues. When a person registers a valid account for a social network, a profile has to be created, which makes sure that his family, friends and colleagues are able to identify himself. The profile includes several pieces of information. Some of them are mandatory, and some are optional. Depending on a user's preference, he needs to find a balance between expressing sufficient information of himself and hiding sensitive personal information. However, by deploying the attacks initiated by Convolutional Neural Networks (CNNs), Fully Connected Neural Networks (FCNNs), or other machine learning algorithms, onto a large amount of available data, the hidden sensitive personal information can be inferred and users' privacy can be compromised.

The motivation of this paper lies in that sensitive information attacks have profound impacts for both online users and advertisers. Understanding the attacks can radically help to develop defensive measures to prevent privacy leakage for online users. There are several consequences when users' privacy is exposed. First, sensitive information may facilitate adversaries to recover users' passwords since current password recovery mechanisms usually ask users' sensitive information before sending password recovery links. Second, online user privacy leakage could affect users' offline activities. For example, knowing users' detailed information such as name, birthday, and address could aid to forge credit cards or even identification documents. Third, sensitive information can help advertisers to deliver ads for targeted users. For example, knowing a user's age, gender, and zip code could reveal the user's lifestyle, which dramatically increases the accuracy of ad delivery.

In this paper, an extensive study is conducted to infer sensitive personal information from users' profile images and insensitive attributes. Specifically, Sina Weibo data are used to infer sensitive age information based on available public user profiles. Sina Weibo is the most popular social network in China. Each Weibo profile contains a profile image and many attributes like location, gender, work information, number of followings, number of posts, and so on. Among the attributes, sensitive age information is usually hidden. Thus, this paper focuses on inferring a user's age range based on his available public profile.

Three challenges are involved in the considered age


- *B. Mei is with the Department of Computer Science, The George Washington University, Washington, DC, 20052*
  *E-mail: bomei@gwu.edu*
- *Y. Xiao is with the Department of Computer Science, The George Washington University, Washington, DC, 20052*
  *E-mail: xyh3984@gwu.edu*
- *R. Li is with the Department of Computer Science, The George Washington University, Washington, DC, 20052*
  *E-mail: ruinian@gwu.edu*
- *H. Li (corresponding auther) is with School of Cyber Security, Chinese Academy of Sciences, Beijing 100049, China, and with Beijing Key Laboratory of IoT Information Security Technology, Institute of Information Engineering, Chinese Academy of Sciences, Beijing 100093, China*
  *E-mail: lihong@iie.ac.cn*
- *X. Cheng is with the Department of Computer Science, The George Washington University, Washington, DC, 20052*
  *E-mail: cheng@gwu.edu*
- *Y. Sun is with the School of Business, Beijing Normal University, Beijing 100875, China*
  *E-mail: yunch@bnu.edu.cn*

inference attack. First, a feasible way needs to be discovered to collect a large amount of sensitive information from users. This information is treated as essential ground-truth data. Second, the correlations among all the attributes in Weibo data need to be explored to assist the construction of neural networks. Third, because currently there are no neural networks built on Weibo for inference attacks, the most effective configuration for the proposed framework needs to be established.

There are several contributions in this paper. First, the paper invents a new framework that integrates faster Regional Convolutional Neural Network (R-CNN), traditional CNN, and FCNN. This integration not only broadens the scope of inference attacks in social networks but also boosts the performance for the existing models. Second, the paper modifies the existing CNN age classifier based on images. The proposed framework adds users' attributes in the classifier to make the classifier more powerful and robust. Third, the paper illustrates the configuration of a fully connected age classifier based on attributes. Although there are multiple papers focusing on the similar classifiers, none of them gives detailed explanation of the structures of the built neural networks. Fourth, the paper demonstrates a feasible way about how to collect sensitive users' information in real-world social networks. Fifth, the paper compares the performances among neural networks and different traditional machine learning algorithms for inference attacks in social networks. Sixth, the paper proposes an effective countermeasure by leveraging Differential Privacy (DP) against inference attacks.

Particularly, to achieve high accuracy on the age range prediction, a novel framework is proposed. This framework is an integration and modification of a few existing neural networks. One main integration is the combination of faster R-CNN and CNN age classifier on images. The extracted face region from R-CNN feeds the input of CNN age classifier, which can significantly boost the attack performance. Another main integration is the combination of CNN age classifier and FCNN age classifier. This integration widens the scope of the existing models making the proposed framework be able to predict a user's age range no matter whether the user has a valid profile image or not. One main modification is that users' attributes are injected for the CNN age classifier. As a result, unlike the existing classifiers that are only based on images, both images and attributes are utilized to seek the best prediction accuracy.

On the other hand, this paper also gives the construction details about a fully connected age classifier based on attributes. Both faster R-CNN and CNN age classifiers based on images are well illustrated in the existing research. However, there is no detailed information about how to construct a FCNN age classifier based on attributes. The discussion in this paper facilitates to understand how FCNN works for inference attacks since different FCNN configurations can have dramatic differences in performance.

To better evaluate the ability of inference attacks in social networks, several traditional machine learning algorithms are also applied. Specifically, decision tree, Naïve Bayes, and $k$-Nearest Neighbors ($k$-NN), are selected for our comparison study. These algorithms have been proven to be effective to execute inference attacks in different areas. However, in the area of social networks, few studies have been performed to compare the performances with neural networks.

Last but not least, this paper also proposes an effective defense mechanism which leverages DP to prevent learning-based inference attacks. The evaluation results show that our defense mechanism can not only efficiently hinder the inference attacks mentioned in this paper by lowering the success rates, but also preserve the data usability with best effort by finding a relatively large privacy budget and minimal number of attributes to perturb.

The paper is organized as follows. Section 2 presents the recent related works for inference attacks in social networks. Section 3 details the proposed framework, explains how the framework is applied onto the targeted social network, and deduces the DP defense mechanism. Section 4 evaluates the performance of both the attack framework and the DP defense mechanism. At last, Section 5 concludes the whole paper.

## 2 Related Work

Many studies [1], [2], [3], [4], [5], [6], [7], [8] have been done focusing on inferring hidden user information based on traditional machine learning algorithms. These studies generally fall into two categories, with the first one clustering users into different categories by deploying unsupervised machine learning algorithms [1], [2], [3], [4] and the second one employing traditional machine learning algorithms combined with natural language processing [5], [6], [7], [8].

However, there are two drawbacks from these studies. First, the performances of traditional machine learning algorithms on social networks are generally poor. Due to the complexity of social networks, there are usually more than 10 public attributes, and it is usually challenging to pursue a linear relationship between the public attributes and the target hidden attribute. Second, some studies use natural language processing to boost the performance. Conversely, the added technique is time-consuming and is also unable to increase the performance in essence because of the limitation of traditional machine learning algorithms.

Recently, a few studies [9], [10], [11], [12] have been realized to utilize FCNN based inference attacks in social networks. FCNN is chosen because it is good at seeking complex relationships between input attributes and output attributes. This advantage becomes especially important when it is hard to represent a relationship by a liner expression. However, existing studies have two major shortcomings. First, these studies neither clearly show the configurations of FCNNs nor the detailed characteristics of the studied datasets. It is critical to demonstrate how a FCNN is constructed since different configurations dramatically affect the performance of inference attacks in social networks. Second, none of the studies compare the performances among FCNNs and traditional machine learning algorithms. Since FCNN is basically one type of machine learning algorithms, it is important, from the global perspective, to deploy such kind of comparisons.

More recently, CNNs have been developing rapidly. Computer vision and pattern recognition have several breakthroughs due to the improvement of CNN. Basically,

the CNN for image processing contains two components: the convolutional (conv) layers followed by the fully connected layers. Conv layers are what CNN distinguishes from the other kinds of neural networks. The purpose of these layers is to extract the features of an image. The tasks of the lower conv layers are to extract the features like curves or triangles in an image while those of the higher layers are to extract more sophisticated features like faces, cars, or flowers. After the conv layers, the fully connected layers take over. The purpose of these layers, like all the layers from FCNNs, is to do classification. In this case, the fully connected layers classify the features that are extracted from the conv layers into different categories.

Levi and Hassner [13] provides the details about using CNNs to classify users' ages and genders based on their profile images in social networks. It shows that by learning representations through the use of CNNs, a significant increase in performance can be obtained on automatic age and gender classification. However, the classifier only works when a user has a profile image that contains his clear face. This is not always true as many users have privacy concerns about uploading personal profile photos online.

Ren *et al.* [14] introduces a faster R-CNN. The purpose of the original R-CNN is to solve the problem of object detection. Given a certain image, bounding boxes can be drawn over all of the objects. The process can be split into two general components, the region proposal step and the classification step. However, the original R-CNN contains complex pipelines and runs relatively slow. The faster R-CNN comes to combat the complexity to make the model run more efficiently. Currently, the faster R-CNN has become the gold standard for object detection in images. Thus, faster R-CNN is utilized as the first step of our proposed framework in this paper.

DP is a popular privacy-protecting mechanism that is widely adopted in various realms of research. One of the most-adopted fields is recommendation system. McSherry and Mironov [15] explores the methods to apply DP onto the leading approaches in the Netflix Prize competition without significantly degrading the accuracy. DP is also gradually applied to social networks. Gao *et al.* [16] applies DP to hierarchical random graph structures within social networks to guarantee stronger de-anonymization protection. Both works mentioned above focus on large-scale user identity anonymization protection and may not be feasible to small-scale datasets. By comparison, our method can protect sensitive attributes of scalable social network datasets against learning-based inference attacks.

## 3 Methodology

### 3.1 System Overview

Fig. 1 shows the overview of the proposed system. It contains three major parts, which are the faster R-CNN face detector, the CNN age classifier based on images and attributes, and the FCNN age classifier based on attributes. The faster R-CNN detects the existence of human faces in a user's profile image. If R-CNN detects a legit single face in an image, the CNN age classifier based on images and attributes is then used. If R-CNN can't detect any human face or there are multiple faces in an image, the FCNN age classifier based on attributes is then used. Inside the CNN age classifier based on images and attributes, there are two basic components: the conv layers and the fully connected layers. The extracted single face image from the faster R-CNN is the input of the conv layers. The conv layers work only onto the face image. The output of the conv layers is a vector that contain critical features for age classification. Then, the attributes from the same user's profile like gender, work information, and education information combined with the output from the conv layers become the input of the fully connected layers. After that, the fully connected layers work onto both image features and attributes to classify the user's age.

Since the proposed framework is the integration and modification of the existing models, the way how the models are integrated and the motivations behind the modification are illustrated from section 3.2 to 3.4. Massive ground-truth data have to be crawled to train the neural networks, so section 3.5 shows how to collect data from real-world social networks. Section 3.6 intends to seek the relationships between the user public attributes and the target hidden attribute. In this step, a correlation matrix is involved to understand the characteristics of the available data. Since existing studies don't show the detailed structure of FCNNs for social network attacks, section 3.7 solves this prominent issue by elaborating a feasible and effective FCNN for inference attacks based on public user information. Section 3.8 shows the implementation of some typical traditional machine learning algorithms. Those algorithms are applied to compare the performance of neural networks. Specifically, decision tree, Naïve Bayes, and $k$-NN are used. These algorithms are proven to be effective for information attacks. Section 3.9 describes the cross-validation used in this paper. Cross-validation is one of the impartial ways to analyze the correctly classified percent for machine learning algorithms. Lastly, Section 3.10 provides defense mechanisms with DP.

### 3.2 Integration of Faster R-CNN Face Detector and CNN Age Classifier

Faster R-CNN face detector and CNN age classifier based on images are mature models. Faster R-CNN face detector can achieve ideal results to extract faces in an image accurately and efficiently. CNN age classifier based on images can also achieve great results if images containing single face are the input. One of the main contribution of this paper is that both models are integrated to perform inference attacks.

There are two advantages to deploy this integration. First, R-CNN face detector can filter out the profile images that don't contain human faces. As a result, time is saved to skip the complex CNN classifier on images and instead, execute light-weight FCNN classifier. Second, the integration can boost the accuracy of CNN age classifier on images. Since the original classifier is trained on face images only, it performs best when the input is clear face images. Faster R-CNN face detector guarantees that every image that feeds the CNN classifier contains a clear single face. The reason is that R-CNN face detector has the ability to extract all the face regions in an image and at the same time, eliminate noisy background and other objects. Then, if there is only one legit face, the extracted face region rather than the original image becomes one input of the CNN classifier.

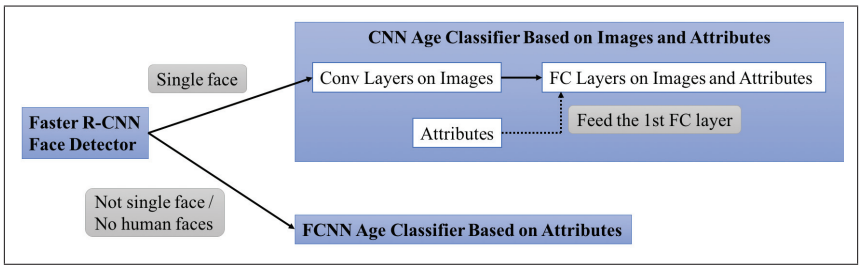



Figure 1: System Overview of Inference Attacks Based on Images and Attributes

### 3.3 Integration of CNN Age Classifier and Users' Attributes

The original CNN age classifier is only based on images. Although the classifier itself already provides accurate prediction, improvements can be done for this mature model. Another main contribution of this paper is that users' attributes are fused into the traditional CNN model. As a result, the CNN age classifier becomes more powerful and robust by taking both images and attributes as inputs.

To inject attributes into the existing classifier, transfer learning is used. Transfer learning is the process of taking a pre-trained model and tuning the model to target a similar problem. As mentioned before, a CNN model contains conv layers and fully connected layers. In this case, all the weights in the conv layers in the existing CNN are frozen. Then, the features extracted from the profile image by conv layers and the same user's attributes are concatenated. The concatenated vector becomes the input of the fully connected layers. When training, only the weights in the fully connected layers are to be determined.

There are two advantages when taking this approach. First, the attributes from a user can complement his face image if the information predicted from the image is not sufficient. Second, the attributes of a user have high potential to increase the accuracy even if his image already provide adequate information. The reason behind the above two advantages is that attributes bring valuable additional information about the users. Instead of depending on just one source – profile images, two sources increase the chance to better predict users' hidden information.

### 3.4 Integration of CNN Age Classifier and FCNN Age Classifier

Users' profile images can provide abundant information toward their hidden attributes. However, it is rather common for users to not upload legit profile images onto social networks. One reason behind this phenomenon is that users try to protect their privacy. Another reason is that most social networks don't verify users' profile images. To make the purposed system target more general scenarios, another main contribution of this paper is the integration of FCNN classifier based only on attributes and CNN classifier based both images and attributes.

There are two advantages for this integration. First, the system can adapt more general situations since not every user has a profile image. Although a user may upload an image, the image may be irrelevant to the user's face. Note that it is not uncommon for users to choose group pictures as profile images. In this situation, those images are still treated as invalid since it's difficult to verify which face in a group image is the user's at the training stage. Second, the integration can increase prediction accuracy. On one hand, if only CNN classifier based on images and attributes is considered, the image may contain adverse information that hinder accurate prediction. For example, if a user's profile image is just a flower, using this photo to train the CNN classifier play an opposite effect. On the other hand, if only FCNN classifier based on attributes is considered, valuable profile images are wasted, and the prediction accuracy can be affected negatively. As a result, the integration of both CNN and FCNN classifiers has significant potential to improve prediction accuracy.

### 3.5 Data Collection

There are two reasons to choose Sina Weibo as the primary site to collect users' data. First, it is the largest online social network in China. By the end of 2015, there are over 222 million subscribers and 100 million daily users. About 100 million messages are posted every day. Second, the data quality in Weibo is very high. Due to Chinese government policy, users in Weibo must register with their true identities. Specifically, the website asks for a valid mobile phone number during the registration process, and Chinese government enforces that all mobile phone numbers must under real names with verified identity. In other words, the user data quality in Weibo can be promised. Baidu Baike is a Chinese-language, collaborative, web-based encyclopedia. It contains accurate personal information for all the public, famous and influential people in China.

Since inference attacks are conducted through neural networks, the data to be trained are essential. Massive ground-truth data have to be retrieved from the Internet. However, the data containing private personal information are difficult to collect. To solve this issue, data from both Weibo and Baidu Baike are crawled. Then, age information extracted from Baidu Baike and profile information retrieved from Weibo are connected. The combination of both websites constructs sufficient ground-truth data that contain sensitive age information. As a result, personal information from 2030 people is collected.

The profile images are collected only through Baidu Baike. The reason is that in order to ensure authoritativeness, Baidu Baike guarantees that each person has a legit and clear profile image. Otherwise, Baidu Baike doesn't contain any profile images for that person. As a result, all the collected images can be treated as ground-truth images, and can reflect people's age correctly.

Aside from the profile images, for the attribute part in both CNN and FCNN, the following attributes are considered: profile location, gender, existence of blog, length of slogan, registration year, existence of work information, existence of education information, number of followings and number of posts. Note that to maintain the universality of the problem, number of followers is not considered since a famous person usually has far more followers than a general user. 18 classes are set for different age ranges, which are from 5–10 years old to 90–95 years old.

### 3.6 Correlation Matrix

Correlation matrix can represent the relationships among the chosen attributes. It shows the dependency between any two attributes. Equation (1) shows the correlation matrix between two variable $x$ and $y$. $x$ is the independent variable. $y$ is the dependent variable. $n$ is the number of data points in the sample. $\bar{x}$ and $\bar{y}$ are the mean of $x$ and $y$, respectively. $s_x$ and $s_y$ are sample standard deviation of $x$ and $y$, respectively. The correlation value $r(x, y)$ is a number between $-1$ to 1. In a positive correlation, as $x$ increases, $y$ increases. In a negative correlation, as $x$ increases, $y$ decreases. If the value is close to 0, it means that $x$ and $y$ are loosely related.

$$r(x,y) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{(n-1)s_x s_y} \tag{1}$$

Fig. 2 is the correlation matrix. If two attributes have higher correlation, the background color of that cell is darker. In Fig. 2, all the public attributes have weak relationships with each other, and all the public attributes have weak relationships with the private age attribute. Since all the attributes are not independent with each other, it is an ideal setting to implement neural networks. The reason is that all the public attributes can effectively aid each other to learn hidden relationships with the targeted private attribute.

### 3.7 Fully Conncected Neural Network Construction

R-CNN face detector and CNN age classifier based on images are clearly demonstrated in [13], [14]. However, existing papers don't clearly show the configuration of FCNN age classifier on attributes. Different configurations can dramatically affect the performance of FCNN, so it is urgent to show the details of FCNN construction.

There are two reasons to select FCNNs. First, by the nature of the problem, the data types for the attributes are simple, and there are just a few known attributes and several targeted classes. FCNN is good at dealing with this kind of settings. Second, FCNN is efficient in computation and can be easily implemented. As a result, it is flexible to adjust the parameters of the network to seek the best possible configuration for the problem.

In a typical FCNN, there are several layers and each layer has several neurons. A neuron is essentially a logistic unit and is represented by logistic activation function. In this paper, sigmoid activation function shown in (2) is used. In (2), $h_\theta(x)$ is the targeted hypothesis that needs to be expressed. $\theta$ and $x$ are undetermined weight vector and known attribute vector, respectively. For the input layer, there are nine units since a user's nine public attributes are measured. For the output layer, there are 18 units because 18 classes are set for different age ranges.

$$h_\theta(x) = \frac{1}{1 + e^{-\theta^T x}} \tag{2}$$

For FCNNs, the key is to find the best number of layers and number of units in each layer to fit the problem properly. Since there are no mandatory rules to follow to define those values, the best way is through trial and error, and the characteristics of the problem. After repeated attempts, the best structure of the neural network is that there are 2 hidden layers and each layer has 13 units. The complete structure of neural network is virtually shown in Fig. 3.

The standard neural network cost function, deduced from logistic regression, is shown in (3). The part before $\frac{\lambda}{2m}$ in (3) is to calculate the error between the training data and the current hypothesis result. The part after $\frac{\lambda}{2m}$ is the regularization part to prevent overfitting. Specifically, $m$ is the number of training data. $K$ is the number of output classes. $(x^{(i)}, y^{(i)})$ represents an entry in the training dataset. $\lambda$ is the regularization parameter to control overfitting. $L$ is the total number of layers in the network. $s_l$ is the number of units in layer $l$, excluding the bias unit.

$$\begin{aligned} J(\Theta) = -\frac{1}{m}(&\sum_{i=1}^{m}\sum_{k=1}^{K} y_k^{(i)} \log\left(h_\Theta(x^{(i)})\right)_k \\ &+ (1 - y_k^{(i)}) \log\left(1 - h_\Theta(x^{(i)})_k\right) \\ &+ \frac{\lambda}{2m}\sum_{l=1}^{L-1}\sum_{i=1}^{s_l}\sum_{j=1}^{s_{l+1}}(\Theta_{ji}^{(l)})^2 \end{aligned} \tag{3}$$

To minimuze $J(\Theta)$, backpropagation algorithm is implemented to compute $\frac{\partial}{\partial \Theta_{jk}^{(l)}} J(\Theta)$. Noted that to achieve best performance, gradient checking and random initialization are also executed.

### 3.8 Traditional Machine Learning Implementation

For this section, several traditional machine learning algorithms are examined to compare with the results from the proposed framework. These algorithms are selected because they have been widely proven to be effective for inference attacks. Particularly, decision tree, Naïve Bayes and $k$-NN are implemented.

The first machine learning algorithm to be applied is decision tree. Decision tree employs top-down and divide-and-conquer strategy. The key is to know which is the best attribute, and the aim is to get the smallest tree. Technically, the selected attribute should have the greatest information gain. Information gain is calculated as the difference between entropy of distribution before the split and entropy of distribution after the split. Here, information is measured in bits. Entropy is defined in (4), and $p$ represents probability.

$$entropy(p_1, p_2, \ldots, p_n) = -\sum_{i=1}^{n}(p_i \log p_i) \tag{4}$$

The second machine learning algorithm is Naïve Bayes. The algorithm assumes that knowing the value of one attribute says nothing about the value of another. Equation (5) shows the Bayes theorem. $Pr[H|E]$ is the probability of

| | Location | Gender | Blog | Slogan | Registration | Work | Education | Followings | Posts | Age |
|---|---|---|---|---|---|---|---|---|---|---|
| Location | 1.00 | -0.12 | -0.24 | 0.10 | -0.39 | -0.14 | -0.28 | -0.30 | -0.20 | 0.11 |
| Gender | -0.12 | 1.00 | -0.19 | -0.19 | -0.18 | 0.20 | 0.13 | 0.13 | 0.15 | 0.33 |
| Blog | -0.24 | -0.19 | 1.00 | 0.32 | 0.51 | 0.21 | 0.22 | 0.29 | 0.22 | -0.12 |
| Slogan | 0.10 | -0.19 | 0.32 | 1.00 | 0.22 | 0.32 | 0.26 | 0.30 | 0.20 | -0.24 |
| Registration | -0.39 | -0.18 | 0.51 | 0.22 | 1.00 | 0.14 | 0.23 | 0.40 | 0.27 | 0.18 |
| Work | 0.14 | 0.20 | 0.21 | 0.32 | 0.14 | 1.00 | 0.41 | 0.23 | 0.14 | -0.11 |
| Education | 0.28 | 0.13 | 0.22 | 0.26 | 0.23 | 0.41 | 1.00 | 0.25 | 0.16 | -0.21 |
| Followings | -0.30 | 0.13 | 0.29 | 0.30 | 0.40 | 0.23 | 0.25 | 1.00 | 0.51 | -0.11 |
| Posts | -0.20 | 0.15 | 0.22 | 0.20 | 0.27 | 0.14 | 0.16 | 0.51 | 1.00 | 0.23 |
| Age | 0.11 | 0.33 | -0.12 | -0.24 | 0.18 | -0.11 | -0.21 | -0.11 | 0.23 | 1.00 |

|r|=1.00
0.50≤|r|<1.00
0.40≤|r|<0.50
0.30≤|r|<0.40
0.20≤|r|<0.30
0.00≤|r|<0.20

Figure 2: Correlation matrix of the training data.

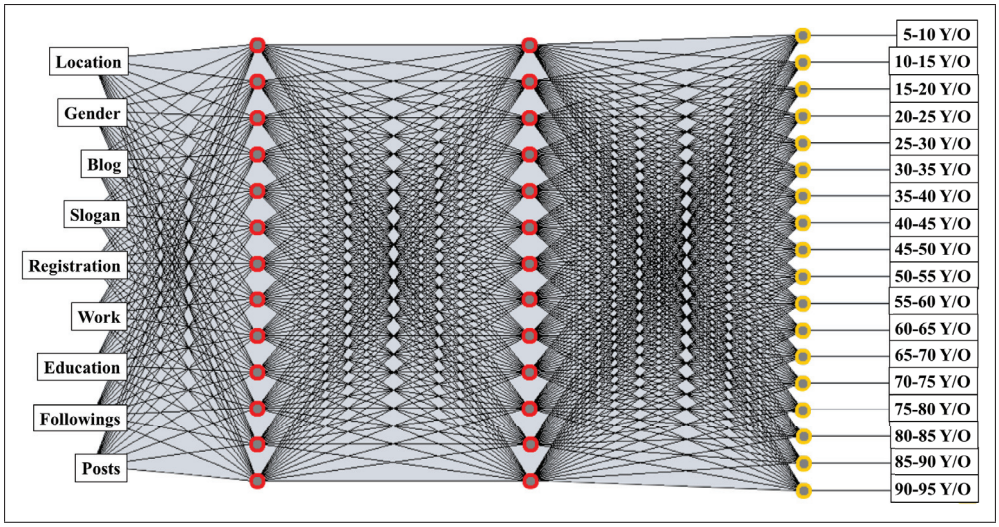



Figure 3: Constructed neural network for Weibo information inference.

event $H$ given evidence $E$, and is also called a posteriori probablity of $H$. $Pr[H]$ is a priori probability of $H$.

$$Pr[H|E] = \frac{(\prod_{i=1}^{n} Pr[E_i|H])Pr[H]}{Pr[E]} \quad (5)$$

The third algorithm is $k$-NN. After many attempts, $k = 20$ best fits the problem. An object is classified by a majority vote of its neighbors, with the object being assigned to the class most common among its $k$ nearest neighbors. A similarity function is used to search the training set. In this paper, Euclidean distance is implemented. The distance $d(p, q)$ is shown in (6), where $p$ and $q$ are two $n$ dimensional points.

$$d(p, q) = \sqrt{\sum_{i=1}^{n}(p_i - q_i)^2} \quad (6)$$

### 3.9 Cross-Validation

Cross-validation is a model validation technique for assessing how the results of a statistical analysis generalizes to an independent dataset. It is mainly used in settings where the goal is prediction, and one wants to estimate how accurately a predictive model will perform in practice. In a prediction problem, a model contains a dataset of known data on which training is run, and a dataset of unknown data against which the model is tested. The goal of cross-validation is to limit problems of overfitting, giving an insight on how the model generalizes to an independent dataset.

In this paper, cross-validation is set as 10 folds. It means that the data is partitioned evenly in 10 portions, and each portion has 203 individuals. Afterwards, 9 portions are used to train, and 1 portion is used to test. This process is iterated 10 times by choosing different portions to train and test to let results truly reflect the fact.

### 3.10 Defense With Differential Privacy

#### 3.10.1 Basic Concepts

So far, DP is the only known privacy-preserving technique that provides strong mathematical foundation and guaranteed protection. The purpose of developing DP is to prevent an adversary from inferring the sensitive information out of the aggregation of databases. For every pair of neighboring databases which differ in only one element, DP hinders an adversary from distinguishing or isolating the value of that particular element based on the query results of this pair of neighboring databases. Before diving into the defense mechanism with DP, basic definitions and theorems of DP used in this paper are illustrated in the following.

***Definition 1 (Neighboring Database).*** Given a database $D_1$, a neighboring database of $D_1$, denoted as $D_2$, has only one element different from $D_1$.

***Definition 2 (Sensitivity).*** Given a positive integer $d$ and a query function $f : \mathcal{D} \rightarrow \mathbb{R}^d$, the sensitivity of the function $f$, denoted as $\Delta f$, is defined as

$$\Delta f = \max_{D_1, D_2} \|f(D_1) - f(D_2)\|_1 \tag{7}$$

where $\|\cdot\|$ is the $\ell_1$ norm. Sensitivity means how sensitive $D_1$ and $D_2$ are based on the query function $f$.

***Definition 3 ($\epsilon$-Differential Privacy).*** A randomized algorithm $\mathcal{A}$ is $\epsilon$-differentially private if for all pairs of neighboring databases $D_1$ and $D_2$, such that

$$Pr[\mathcal{A}(D_1) \in S] \leq e^{\epsilon} \times Pr[\mathcal{A}(D_2) \in S] \tag{8}$$

where $S$ is the collection of all subsets of image of $\mathcal{A}$. $Pr[B]$ represents the probability if $B$ happens. $\epsilon$ is called the privacy budget, which is the most important factor in DP since it directly determines the how indistinguishable $\mathcal{A}(D_1)$ and $\mathcal{A}(D_2)$ are.

The above definitions give all the essentials of DP. Since the goal is to construct the $\epsilon$-differentially private algorithm $\mathcal{A}$ proposed in definition 3, the Laplace mechanism for constructing an $\epsilon$-differentially private algorithm is further presented.

***Theorem 1 (The Laplace Mechanism).*** Given a positive integer $d$ and a query function $f : \mathcal{D} \rightarrow \mathbb{R}^d$, $M(D) = f(D) + Laplace(0, \sigma)^d$ is defined. Given any pair of two neighboring databases $D_1$ and $D_2$, $M$ achieves $\epsilon$-differential privacy if

$$\sigma \geq \max \|f(D_1) - f(D_2)\|_1 / \epsilon \tag{9}$$

where $\sigma = \frac{\Delta f}{\epsilon}$.

Theorem 1 is well proven in [17]. According to the Laplace mechanism, in order to construct an $\epsilon$-differentially private mechanism $M$ based on a query function $f$, the bigger the sensitivity $\Delta f$ is or smaller $\epsilon$ is, the larger noise imposes on $f$.

#### 3.10.2 Applying DP for Defense

The most critical part for constructing a differentially private mechanism is to find out the sensitivities. Since the problem setting is to study how nine Weibo attributes of a person in figure 3 are relevant with that person's age, it is reasonable to define nine query functions, respectively, for the corresponding attributes. Thus, each person is considered as a database since a protection mechanism is proposed for every *individual* user in a social network. In this case, applying a query function on a person returns the value of his one specific attribute. For example, $f_{gender}(u_1)$ retrieves the gender of the user $u_1$, 1 (male) or 0 (female). Under this setting, each pair of two different users is considered as a pair of neighboring databases. Therefore, the sensitivity for a query function $\Delta f_j$ can be further deduced based on definition 2 as

$$\Delta f_j = \max f_j(D) - \min f_j(D) \tag{10}$$

where $j$ represents one of the nine attributes like location or gender, and $D$ is the set of all users.

Hence, nine $\epsilon$-differentially private mechanisms are constructed for nine attributes, respectively. Based on theorem 1, each mechanism $M_j$ is deduced as

$$M_j(D) = f_j(D) + Laplace(\frac{\max f_j(D) - \min f_j(D)}{\epsilon}) \tag{11}$$

## 4 Results and Discussion

### 4.1 Evaluation of the Integrated Framework

Fig. 4 shows the evaluation and comparison among the different components of the proposed framework. It can be seen clearly that FCNN based on attributes has a great ability to predict the age range for a user in the social network. The result is about $8$ times of the result from the random guess. Thus, it proves that the public attributes have great potential to predict users' private attributes by solely utilizing FCNNs. However, since the information contained in the public attributes is still limited, the prediction rate is still under $50\%$.

The result from CNN based on images in Fig. 4 shows the results of inference attack that only utilizes users' profile images. This result is in accord with the result from [13]. However, in this paper, the existing model is modified. Attribute factors are fused in the CNN through transfer learning technique. The process of this technique is elaborated in Section 3.3. As a result, a reasonable boost can be seen from Fig. 4. For R-CNN unadjusted situations, the prediction rate goes from $79.55\%$ up to $85.34\%$, a $7.28\%$ increase. There are two reasons behind the performance boost. First, if the image is lack of ability to predict a user's age range, the added attribute information can compensate the image to provide a better insight of the user. Second, if the image contains sufficient information about a user's age, the added attributes have the ability to reinforce the information and aid the inference attack.

Fig. 4 also shows a comparison between the results before R-CNN adjusted and after R-CNN adjusted. The introduction of R-CNN is to increase the flexibility of the proposed framework. Thus, regardless of the percent of users who have valid profile images in a dataset, the framework can adapt the dataset and maximize prediction accuracy. To evaluate flexibility, the absolute difference before R-CNN adjusted and after R-CNN adjusted under each situation is measured. If the differences are small, it means that not only the framework is flexible to deal with different datasets but also high overall prediction rate can be guaranteed. In Fig. 4, it can be seen that there is not much fluctuation under all the situations, which shows that the prediction ability of the framework is promising and robust. It also can be seen that the framework yields high overall prediction rate at $78.00\%$.

Table 1 shows the results from the constructed FCNN as well as several other machine learning algorithms. There are four indices being considered, which are correctly classified percent, kappa statistic, mean absolute error and root mean squared error. For correctly classified percent, cross-validation is implemented to achieve convincing results. Kappa statistic is to measure inter-rater agreement for qualitative items. Mean absolute error measures the average magnitude of the errors in a set of predictions, without considering their directions. Root mean squared error is

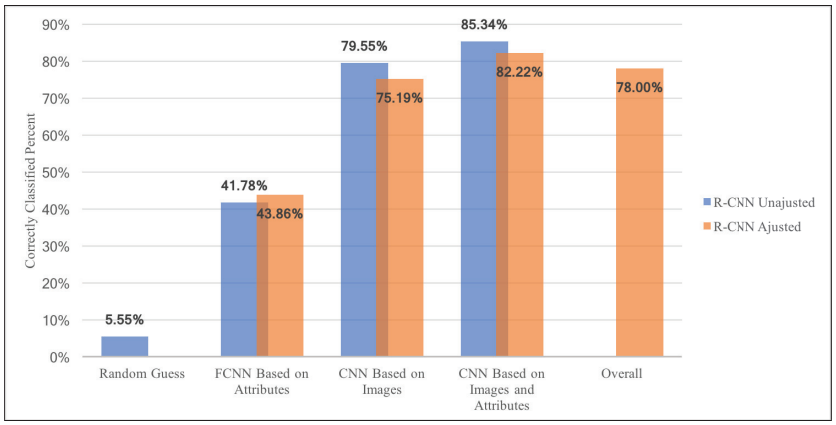



Figure 4: Comparison among different neural networks.

a quadratic scoring rule that also measures the average magnitude of the error.

Correctly classified percent is the most important index to reflect the performance of each algorithm. The percent from each traditional machine learning algorithm can achieve about 6 times of the percent from the random guess. Since the data is crawled directly from the Internet, the distribution of the data is not normalized. $k$-NN has the best ability to handle noisy data when $k$ is set to a relatively large number. Thus, $k$-NN performs the best among the three traditional machine learning algorithms. In addition, from the nature of the problem, intuitively, it is hard to seek the relationships between sensitive age information and the combination of all the insensitive attributes. Fig. 2 confirms the same difficulty since all the available attributes have very limited correlations. From this perspective, the three machine learning algorithms provide relatively reasonable and effective results.

Table 1 also shows that the FCNN is about 8 times of the random guess accuracy. When comparing the result from the FCNN with the result of any traditional machine learning algorithm, it is clear that the FCNN outperforms all three traditional algorithms. There are three reasons about why FCNN performs better. First, FCNN is good at finding complex and subtle relationships between known attributes and unknown attributes. In social network inference attacks, the known attributes are very loosely connected to the targeted hidden attribute. Second, the hypothesis of FCNN is complicated and is not limited to linear expressions. As traditional machine learning algorithms can only pursue relationships expressed by linear expressions, FCNN breaks this restriction and is able to seek non-linear relationships. Third, since the social network data are noisy, FCNN has stronger ability to deal with the noise than any other traditional machine learning algorithm. Because of these three reasons, the constructed FCNN has higher correctly classified percent to infer hidden attributes for social networks.

When taking the results from Fig. 4 and Table 1 as a whole, it is obvious that current social networks have considerable privacy issues. Nowadays social network providers try their best to protect users' private information and give the users right to control their profiles. However, from the results of this paper, the users' private information is still vulnerable. On one hand, simple neural network attacks like FCNN attacks can achieve nearly $50\%$ of the prediction accuracy. On the other hand, complex neural network attacks like the proposed framework in this paper can achieve almost $80\%$ of overall prediction accuracy. Note that along with the development of modern computing, complex neural networks don't need unbearable amount of time to be executed. This situation further puts users' privacy at risk. In a word, current social network privacy is nevertheless susceptible to the inference attacks initiated by modern neural networks.

### 4.2 Evaluation of DP as Defense Mechanism

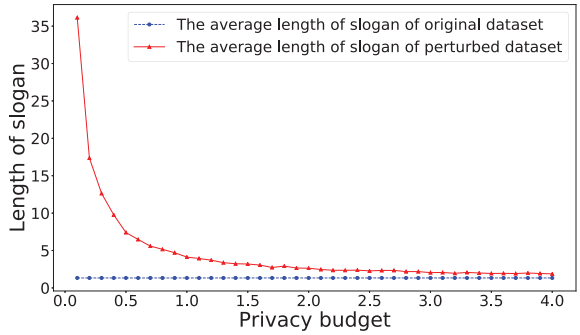
To apply DP as the defense mechanism, the sensitivity for each query function is needed to be calculated. According to (10), the sensitivities of queries of location, gender, blog, slogan, registration, work, education, followings and posts are 35, 1, 1, 7, 7, 1, 1, 3257 and 104544, respectively, based on the collected data. Higher sensitivity means that data are more closely correlated under a query function. Thus, the query for number of posts is the most sensitive, meaning that it is easy for an attacker to infer a user's number of posts based on others' numbers of posts. After figuring out sensitivities, differentially private algorithm for each query function can be further constructed according to Theorem 1. Note that applying Laplace distribution as noises may result in negative values, which are meaningless toward the nature of the attributes. As a result, all the negative outputs are zeroed out to keep the perturbed results physically meaningful.

Prior to probe into the performance of the inference attacks under protection of DP, it is valuable to evaluate the trend of how changing the privacy budget $\epsilon$ poses a difference in the original dataset. Thus, the privacy budget $\epsilon$ is set from $0.1$ to $4.1$ and the changing trends of the average attribute values of the dataset is recorded. Fig. 5 shows the changing trends of the average values of three of the attributes: length of slogans, number of posts and number of followings.

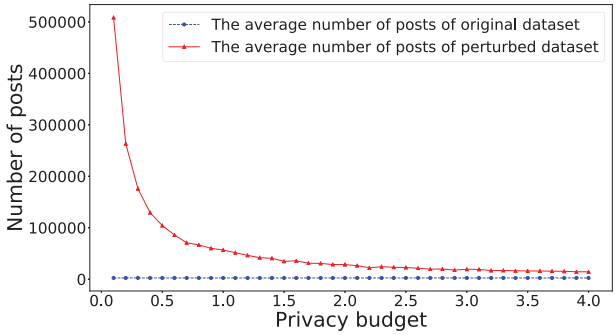
From Fig. 5, it can be seen that the three attributes are very sensitive because the differences of the maximum values and the minimum values are large. If the privacy budget $\epsilon$ is set small, the averages of the perturbed data can deviate far from the averages of the original data, which provides

Table 1: Comparison among different machine learning algorithms

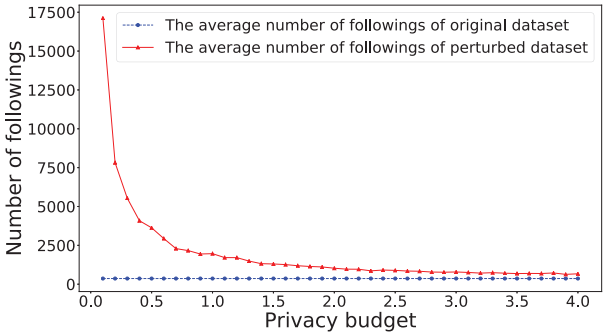
| | Random Guess | Decision Tree | Naïve Bayse | $k$-NN ($k = 20$) | FCNN |
|---|---|---|---|---|---|
| **Correctly Classified Percent** | 5.5556% | 27.9585% | 30.0873% | 34.9454% | 41.7801% |
| **Kappa Statistic** | — | 0.0357 | 0.0636 | 0.0960 | 0.0457 |
| **Mean Absolute Error** | — | 0.0922 | 0.0926 | 0.0928 | 0.0923 |
| **Root Mean Squared Error** | — | 0.2614 | 0.2279 | 0.2207 | 0.2361 |




(a) Changing trend of slogan length




(b) Changing trend of number of posts




(c) Changing trend of number of followings

Figure 5: Changing trends of attribute values against $\epsilon$

strong privacy protection. However, the usability of the data is compromised, making the data useless for many benign applications such as research and recommendation systems. Hence, a balance between the value of $\epsilon$ and the usability of the data needs to be sought. Specifically, the goal is to find out the minimum $\epsilon$ that can effectively lower the possibility of a successful inference against sensitive attribute, the age range, to a reasonable threshold. To conduct this study, the experiments are split into two steps. First, $\epsilon$ for *all* nine attributes are continuously changed, and the cross-validation inference results from four machine learning algorithms including FCNN mentioned in Table 1 are observed. Second, a more fine-grained situation is studied. In this case, DP is iteratively applied to each attribute while the other attributes are fixed. The purpose of the second study is to identify which attribute affects the inference results the most and how perturbing of one specific attribute can hinder the inference.

For the first-step experiment, $\epsilon$ is continuously changed from $4.1$ to $0.1$. The trends of accuracy change for the four machine learning algorithms are shown in Fig. 6. It can be seen that if the privacy budget $\epsilon$ is set to $1$, DP defense mechanism reaches great results. At this point, the accuracy of all the inference attacks are reduced to a value that is no better than random guess. Note that from Fig. 6, the neural network is the most sensitive one to data perturbation, which has a sharp decrease between $\epsilon = 3.0$ and $\epsilon = 2.3$.

For the second-step experiment, each time, DP is applied to only one attribute and the values of the other attributes are fixed. In this case, it is no need to study all four machine learning algorithms, so only the trend of accuracy change of the neural network is examined. Again, $\epsilon$ is changed from $4.1$ to $0.1$, and the trends of accuracy change of the neural network are shown in Fig. 7. It can be seen that length of slogan, gender are the two key attributes, which leads to success of the neural network inference attacks. If $\epsilon$ is set to $1.5$, DP defense mechanism can successfully reduce the accuracy of both length of slogan and gender to a value that is close to random guess.

## 5 Conclusion

This paper proposes a new framework for inference attacks in social networks. The proposed framework smartly integrates and modifies the existing state-of-the-art CNN models. As a result, it can fit wider scenarios for inference attacks no matter whether a user has a legit profile picture or not.

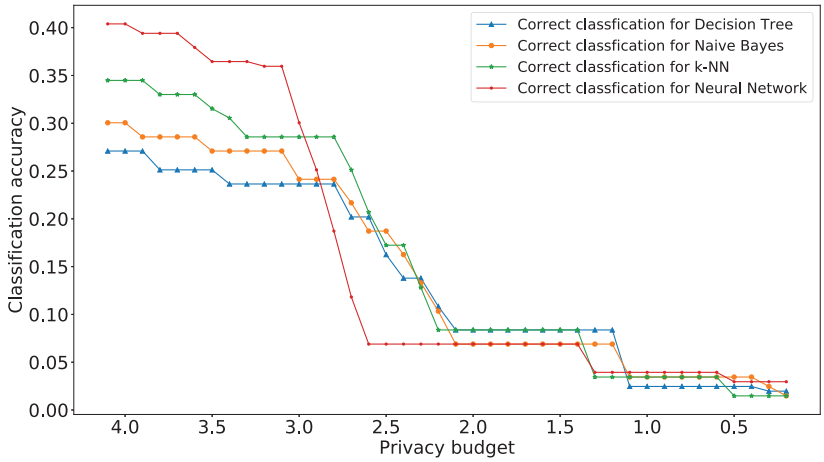



Figure 6: Cross-validation accuracy of machine learning algorithms against $\epsilon$

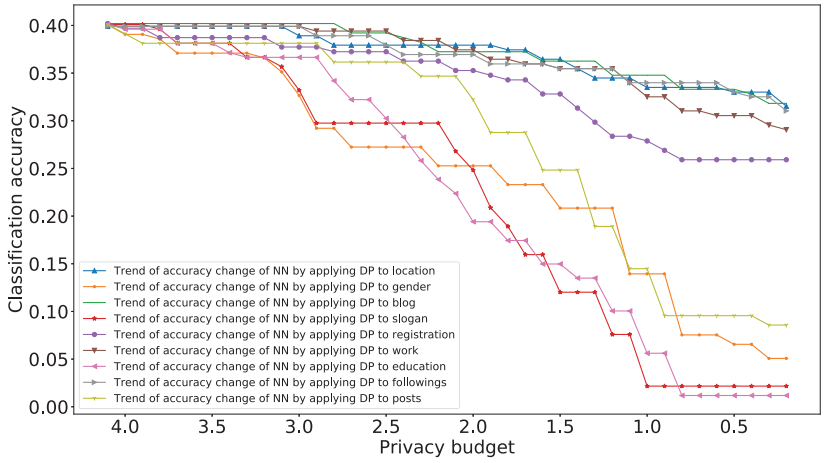



Figure 7: Cross-validation accuracies of Neural Networks against $\epsilon$

The experiments show 78.00% overall prediction accuracy, which is nearly 16 times of the random guess accuracy. The experiments also show that the framework can boost the existing high-accuracy CNN for inference attacks by 7.28%. In addition to the new framework, the paper also provides the detailed configuration of FCNN for inference attacks, which is usually missing in the existing studies. Moreover, traditional machine learning algorithms are used to compare the results from the constructed FCNN. Last but not least, DP is used and evaluated as a defense mechanism. As a whole, the paper convincingly demonstrates that current social networks are vulnerable to the information inference attacks initiated by modern neural networks.

There are two major limitations in this paper. First, only one social network is trained and tested. Second, only one targeted private attribute, the age range, is considered. In the future work, more social networks and more targeted attributes will be explored to further validate the performance of the proposed framework and to show more diverse inference attacks in social networks.

## Acknowledgments

This work was supported by the US National Science Foundation under grant CNS-1704397 and the National Natural Science Foundation of China under grant 61702503.

## References

[1] D. Liben-Nowell and J. Kleinberg, "The link-prediction problem for social networks," *journal of the Association for Information Science and Technology*, vol. 58, no. 7, pp. 1019–1031, 2007.

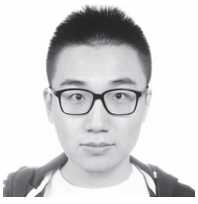
[2] M. Pennacchiotti and A.-M. Popescu, "A machine learning approach to twitter user classification." *Icwsm*, vol. 11, no. 1, pp. 281–288, 2011.

[3] N. Benchettara, R. Kanawati, and C. Rouveirol, "Supervised machine learning applied to link prediction in bipartite social networks," in *Advances in Social Networks Analysis and Mining (ASONAM), 2010 International Conference on*. IEEE, 2010, pp. 326–330.

[4] I. H. Witten, E. Frank, M. A. Hall, and C. J. Pal, *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann, 2016.

[5] W. M. Soon, H. T. Ng, and D. C. Y. Lim, "A machine learning approach to coreference resolution of noun phrases," *Computational linguistics*, vol. 27, no. 4, pp. 521–544, 2001.

[6] A. McCallum, A. Corrada-Emmanuel, and X. Wang, "Topic and role discovery in social networks," *Computer Science Department Faculty Publication Series*, p. 3, 2005.
[7] Y. Matsuo, J. Mori, M. Hamasaki, T. Nishimura, H. Takeda, K. Hasida, and M. Ishizuka, "Polyphonet: an advanced social network extraction system from the web," *Web Semantics: Science, Services and Agents on the World Wide Web*, vol. 5, no. 4, pp. 262–278, 2007.
[8] I. Habernal, T. Ptáček, and J. Steinberger, "Sentiment analysis in czech social media using supervised machine learning," in *Proceedings of the 4th workshop on computational approaches to subjectivity, sentiment and social media analysis*, 2013, pp. 65–74.
[9] R. Michalski, P. Kazienko, and D. Król, "Predicting social network measures using machine learning approach," in *Proceedings of the 2012 International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2012)*. IEEE Computer Society, 2012, pp. 1056–1059.
[10] A. Luna, M. N. del Prado, A. Talavera, and E. S. Holguín, "Power demand forecasting through social network activity and artificial neural networks," in *2016 IEEE ANDESCON*, Oct 2016, pp. 1–4.
[11] Z. Li, D. y. Sun, J. Li, and Z. f. Li, "Social network change detection using a genetic algorithm based back propagation neural network model," in *2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, Aug 2016, pp. 1386–1387.
[12] A. Khadangi and M. H. F. Zarandi, "From type-2 fuzzy rate-based neural networks to social networks' behaviors," in *2016 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, July 2016, pp. 1970–1975.
[13] G. Levi and T. Hassner, "Age and gender classification using convolutional neural networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2015, pp. 34–42.
[14] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," in *Advances in neural information processing systems*, 2015, pp. 91–99.
[15] F. McSherry and I. Mironov, "Differentially private recommender systems: building privacy into the net," in *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2009, pp. 627–636.
[16] T. Gao, F. Li, Y. Chen, and X. Zou, "Preserving local differential privacy in online social networks," in *International Conference on Wireless Algorithms, Systems, and Applications*. Springer, 2017, pp. 393–405.
[17] C. Dwork, A. Roth *et al.*, "The algorithmic foundations of differential privacy," *Foundations and Trends® in Theoretical Computer Science*, p. 32, 2014.



**Bo Mei** is a Ph.D. candidate at the Department of Computer Science in the George Washington University. He got his bachelor degree of Material Science and Engineering in Beijing Institute of Technology, China, in 2010. Then, he granted one master degree of Interdisciplinary Engineering in Purdue University in 2011 and one master degree of Computer Science in the George Washington University in 2013. He began his research focusing on Mobile Computing since 2014, and has conducted extensive study on system applications for IoT devices. Currently, his research spans the broad area of mobile computing, IoT, and social network inference attacks.

**Yinhao Xiao** is a Ph.D. candidate at the Department of Computer Science in the George Washington University. He received his bachelor degree in Information and Computing Science in Guangdong University of Technology, China, in 2012. Later, he attended the George Washington University where he was granted a Master degree in Applied Mathematics and a Master degree in Computer Science by the year of 2014 and 2015, respectively. His research interests cover a broad area of system security (more specifically, mobile security and IoT security) and social network privacy.

**Ruinian Li** is a Ph.D. student at the Department of Computer Science in the George Washington University. He received his bachelor degree in Software Engineering from Nanchang University, China, in 2011. He attended the joint program of Nanchang University and SUNY Polytechnic Institute, where he attained his master degree in computer science from Suny Polytechnic Institute in 2013. His research interests include network security, applied cryptography and privacy preserving computations.

**Hong Li** is an Assistant Professor in the Institute of Information Engineering, Chinese Academy of Sciences. He received his bachelor degree from Xi'an Jiaotong University in 2011, and received his Ph.D. degree from University of Chinese Academy of Sciences in 2017. His current research interests include IoT security and privacy, social network privacy and blockchain.

**Xiuzhen Cheng** received her M.S. and Ph.D. degrees in Computer Science from University of Minnesota Twin Cities in 2000 and 2002 respectively. She is a professor in the Department of Computer Science, the George Washington University. Her current research interests include privacy-aware computing, wireless and mobile security, cyber physical systems, mobile computing, and algorithm design and analysis. She has served on the editorial boards of several technical journals and the technical program committees of various professional conferences. She also has chaired several international conferences. She is a Fellow of the IEEE and a member of ACM.

**Yunchuan Sun** received the Ph.D. degree from the Institute of Computing Technology, Chinese Academy of Science, Beijing, China, in 2009. He is currently an Associate Professor with Beijing Normal University, Beijing. His research interests include big data modeling and analysis, event-linked network, Internet of Things, semantic technologies, knowledge engineering, and information security. He has been the Secretary of the IEEE Communications Society Technical Subcommittee for the Internet of Things since 2013. He has also been an Associate Editor of the Personal and Ubiquitous Computing (Springer) since 2012.