Exploring Personal Attributes from Unprotected Interactions

Abstract

Research, so far, has shown that many personal attributes, including religious and political affiliations, sexual orientation, relationship status, age, and gender, are predictable providing users' interaction data. To address these privacy concerns, users on a social networking site like Faceboook are usually left with profile settings to mark some of their data invisible. However, users sometimes interact with others using unprotected posts (e.g., posts from a "Faceboook page[1]"). Although the aim of such interactions is to help users to become more social, visibilities of these interactions are beyond their profile settings and publicly accessible to everyone. The focus of this paper is to explore such unprotected interactions so that users' are well aware of these new vulnerabilities and adopt measures to mitigate them further. In particular, we ask - *are users' personal attributes predictable using only the unprotected interactions*? To answer this question, we design a novel problem of predictability of users' personal attributes with unprotected interactions. The extreme sparsity patterns in users' unprotected interactions pose a serious challenge for the proposed problem. Therefore, we first provide a way to mitigate the data sparsity challenge and propose a novel attribute prediction framework using only the unprotected interactions. Experimental results on Faceboook dataset demonstrates that the proposed framework can predict users' personal attributes.

Introduction

Users' interactions are cornerstone to the popularity of social media, improving a plethora of services related to friend connections, job recommendations, news disseminations, restaurant reviews, product advertisements, and event participations. Users' interactions even played a critical role during the recent events of the "Arab Spring", "Bangladesh Protests" (Shahbag), and "Assam Violence" (Assam), to organize and mobilize protesters, give the latest updates, and spread deceptions. Posts containing violent videos, morphed photos and instigating messages during such events triggered a large amount of user interactions and often led to negative consequences for societies. Recent research (Kosinski, Stillwell, & Graepel 2013) shows that many personal attributes, including religious and political affiliations, sexual orientation, relationship status, age and gender, are predictable providing that users' personal data including interactions are available. To address such privacy concerns, users often use their profile settings to mark their personal data, including status updates, lists of friends, videos, photos, and interactions on posts, invisible to others. Hence, in this study we only focus on the data which is *publicly available and beyond individual users profile settings*. We refer to such data as *unprotected* data.

Social media wants their users to be more social and at the same time less concerned about unwarranted access to their personal data. Recent social media advancements are creating new opportunities for meaningful interactions among users, while enabling new profile settings for users to better protect their personal information. New mechanisms such as Facebook page allow users' to interact through posts without requiring them to be friends, while keeping their personal information, including demographic profiles, lists of friends, and interactions with friends private. Users' interactions on these pages are often centrally administered and publicly available for everyone. Based on whether a user can control the visibility of her actions, a post can be categorized into two parts: *protected or unprotected* post. A protected post is a post which can be controlled by a user's individual profile settings, otherwise it is referred as a unprotected post. In this paper, we exclusively focus on unprotected posts, and the users' actions, including liking, commenting and sharing, on unprotected posts are together referred as their unprotected interactions. Given the pervasive availability of unprotected interactions, we ask – *are users' personal attributes predictable using only the unprotected interactions on posts?*

To answer the question, we study the problem of the predictability of users' personal attributes in the context of Facebook pages. There are several challenges regarding the data in Facebook pages. The first challenge is about the *availability* of data. Based on the literature (Mislove *et al.* 2010; Kosinski, Stillwell, & Graepel 2013), users' connections and interactions on all types of posts play a vital role in predicting personal attributes. However, the users' connections and interactions on protected posts can be marked invisible using profile settings. Since our focus is exclusively on unprotected data, we assume user connections and protected posts are not available for further analysis. The second challenge is about

[1] The term "Faceboook page" refers to the page which are commonly dedicated for businesses, brands and organizations to share their stories and connect with people.

the complexity of *multilingual text*. During the recent events across the globe including “Arab Spring”, “Assam riots”, and “Bangladesh Protests”, Facebook users primarily communicated in their local languages such as Arabic, Assamese, and Bengali rather than English alone, which makes text data complex to analyze. In this work, we exclusively focus on clickable interactions such as likes, comments and shares on unprotected posts. Third challenge is about the *extremely sparse* interactions on unprotected posts. Due to the intrinsic design, all the (Facebook) users can perform actions on unprotected (Facebook) posts. This leads to extremely sparse interaction patterns which further exacerbates the difficulty of the proposed prediction problem.

In this paper we systematically investigate how to deal with extremely sparse interactions on unprotected posts; and propose a novel framework to predict users' personal attributes from such interactions. The major contributions of this paper are summarized as follows:

- Identify the novel problem whether users personal attributes are predictable using only the unprotected interactions. To the best of our knowledge, we are the first to address this problem. In the next section, we formally define the proposed problem.
- Propose a novel framework to predict users' personal attributes from unprotected interactions. Framework provides a way to systematically address the sparsity problem of unprotected interactions with the help of popular social theories. We elaborate the framework details in a separate section later.
- Evaluate the proposed framework with real-word data from Facebook pages. Section on experiments exhibit all the details establishing the efficacy of the proposed Framework.

Problem Statement

We first present the notations used in this paper. Let $\mathbf{A} \in \mathbb{R}^{n\times m}$ be the matrix, where n is the number of rows and m is the number of columns. The entry at i-th row and j-th column of $\mathbf{A}$ is denoted as $\mathbf{A}(i,j)$. $\mathbf{A}(i,:)$ and $\mathbf{A}(:,j)$ denote the i-th row and j-th column of $\mathbf{A}$, respectively. $\|\mathbf{A}\|_F$ is the Frobenius norm of $\mathbf{A}$, and $\|\mathbf{A}\|_F = \sqrt{\sum_{i=1}^{n}\sum_{j=1}^{m}\mathbf{A}(i,j)^2}$.

Typically, two types of objects are involved in interactions: users and posts. Let $\mathbf{u} = \{u_1, u_2, \ldots, u_n\}$ be the set of users, and $\mathbf{v} = \{v_1, v_2, \ldots, v_m\}$ be the set of unprotected posts, where n and m are the total numbers of users and unprotected posts, respectively. Depending on the social media site, users' interactions involve different types of clickable actions. For example, Facebook users mainly perform three types of clickable actions on unprotected posts and their associated items including liking, commenting, and sharing. For each allowed action, we can construct the user-post action matrix $\mathbf{R} \in \mathbb{R}^{n\times m}$, where $\mathbf{R}_{ij} = 1$ if i-th user's perform the action to j-th post, otherwise 0. For the simplicity of discussion, we assume that $\mathbf{R}$ contains user-post like actions.

The problem of predicting users' attributes is extensively studied. It assumes that there are N labeled users in $\mathbf{u}$ with $N < n$. We assume that $\mathbf{u}_L = \{u_1, u_2, \ldots, u_N\}$ is a set of labeled users where $\mathbf{u}_L$ is a subset of $\mathbf{u}$. Let $\mathbf{Y}_L \in \mathbb{R}^{N\times K}$ be the label matrix of $\mathbf{u}_L$ where K is the total number values of a given attribute. The vast majority of existing attribute prediction algorithms make use of users' personal data (with no consideration of the fact that they can be protectable) such as their all types of posts (Rao & Yarowsky 2010; Conover *et al.* 2011) or their social networks (Jernigan & Mistree 2009; Mislove *et al.* 2010; Tang & Liu 2009) to obtain a predictor f to predict the attribute of users in $\{\mathbf{u} \setminus \mathbf{u}_L\}$. To seek an answer to the question of whether users' personal attributes are predictable using only the unprotected interactions, we formally investigate the following problem -
Given users' unprotected interactions on posts, and the known attribute labels $\mathbf{Y}_L$, we aim to learn a predictor f to automatically predict the personal attribute for unlabeled users i.e, $\{\mathbf{u} \setminus \mathbf{u}_L\}$.

Framework for Attribute Prediction: SCOUT

A user usually performs like actions with a small proportions of all posts, resulting in a sparse user-post action relationships. One of the key difference between protected and unprotected posts is that only friends can perform interactions on protected posts, whereas all users can perform interactions on unprotected posts. Hence, interaction patterns on unprotected posts are likely to be more sparse than protected posts. Thus, the problem of predicting the personal attributes from such sparse unprotected interactions is more challenging for traditional classification methods including support vector machines (SVM), logistic regression, and naive Bayes. Our proposed framework, SCOUT[2], aims to address the sparse interactions problem by learning a compact representation of users with the help of social theories. This compact representation is later used to build a predictor f to automatically predict the personal attributes.

Learning a Compact Representation

The low-rank matrix factorization-based method is one of the popular way to obtain the compact representation of users (Tang *et al.* 2013a). In this paper, we adopt the well known matrix factorization model (Ding *et al.* 2006) to obtain low rank representation of users. The matrix factorization model seeks a low rank representation $\mathbf{U} \in \mathbb{R}^{n\times d}$ with $d << n$ via solving following optimization problem.

$$\min_{\mathbf{U},\mathbf{H},\mathbf{V}\geq\mathbf{0}} \quad \|(\mathbf{R} - \mathbf{U}\mathbf{H}\mathbf{V}^\top)\|_F^2 + \lambda(\|\mathbf{U}\|_F^2 + \|\mathbf{V}\|_F^2 + \|\mathbf{H}\|_F^2), \tag{1}$$

where $\mathbf{V} \in \mathbb{R}^{m\times d}$ is a low-rank space representation of the set of unprotected posts; and $\mathbf{H} \in \mathbb{R}^{d\times d}$ captures the correlations between the low rank representations of users and unprotected posts such as $\mathbf{R}(i,j) = \mathbf{U}(i,:)\mathbf{H}\mathbf{V}^\top(j,:)$. λ is non-negative and introduced to control the capability of $\mathbf{U}$, $\mathbf{V}$ and $\mathbf{H}$ and avoids model over-fitting. The learnt compact representation may be inaccurate because of the sparsity of $\mathbf{R}$. The number of zero entities in $\mathbf{R}$ is much larger than that of non-zero numbers, which indicates that

[2] Merriam-webster dictionary defines “scout” as to explore an area to obtain information. Philosophically, our proposed framework also scouts for users' attributes from unprotected interactions on posts.

$\mathbf{U}(i,:)\mathbf{H}\mathbf{V}^\top(j,:)$ will fit to be zero. The extreme sparsity of $\mathbf{R}$ will result in the learnt representation $\mathbf{U}$ close to a zero matrix.

One way to mitigate the data sparsity challenge is to give different weights to the observed and missing actions. We introduce a weight matrix $\mathbf{W} \in \mathbb{R}^{n\times m}$ where $\mathbf{W}(i,j)$ is the weight to indicate the importance of $\mathbf{R}(i,j)$ in the factorization process. The new formulation is presented in Eq. (1) as

$$\min_{\mathbf{U},\mathbf{H},\mathbf{V}\geq\mathbf{0}} \quad \|\mathbf{W}\odot(\mathbf{R}-\mathbf{U}\mathbf{H}\mathbf{V}^\top)\|_F^2 + \lambda(\|\mathbf{U}\|_F^2 + \|\mathbf{V}\|_F^2 + \|\mathbf{H}\|_F^2), \tag{2}$$

where $\odot$ is the Hadamard product and $(\mathbf{A}\odot\mathbf{B})(i,j) = \mathbf{A}(i,j)\times\mathbf{B}(i,j)$ for any two matrices $\mathbf{A}$ and $\mathbf{B}$ with the same size. $\mathbf{W}(i,j) = 1$ if $\mathbf{R}(i,j) = 1$. Following the suggestions in (Tang *et al.* 2013b), we set $\mathbf{W}(i,j)$ to a small value close to zero when $\mathbf{R}(i,j) = 0$, which allows negative samples in the learning process. In this work, we set $\mathbf{W}(i,j) = 0.01$ when $\mathbf{R}(i,j) = 0$.

In addition to like actions, users can perform other actions such as sharing and commenting. There are many social theories such as homophily (McPherson, Smith-Lovin, & Cook 2001) and consistency (Abelson 1983) theories developed to explain users' actions. These social theories pave a way for us to model user-user and post-post correlations, which can potentially further mitigate the data sparsity problem.

Modeling Correlations

User-user and post-post correlations in social media are widely used to improve various tasks such as feature selection (Tang & Liu 2012), sentiment analysis (Hu *et al.* 2013; Li, Zhang, & Sindhwani 2009) and recommendation (Lu *et al.* 2010). In this paper, we propose a novel way to compute the user-user and post-post correlations using users' actions on unprotected posts and their associated items such as comments and shared posts. We exploit these correlations to tackle the sparsity problem further and more details are discussed in the following subsections.

Modeling User-User Correlations. Apart from likes, users also perform other actions including commenting, replying and sharing on different types of objects such as posts, shared posts, and comments. This subsection provides a way to include these users' activities by modeling user-user correlations. Homophily (McPherson, Smith-Lovin, & Cook 2001) is one of the important social theories developed to explain users' actions during interactions in the real world. Homophily theory suggests that similar users are likely to perform similar actions. These intuitions motivate us to obtain low-rank space representation of users based on their historical actions during interactions. We define $\Psi(i,j)$ to measure the user-user correlation coefficients between u_i and u_j. There are many ways to measure user-user correlation, such as similarity of users' behavior (Ma *et al.* 2011) and connections in social networks (Lu *et al.* 2010). In this paper, we choose the similarity of users' historical behavior to measure user-user correlations. A user can perform a variety of actions, including liking, commenting, and sharing. Hence, similarity is calculated as a function of the total amount of actions performed by two users together as $\Psi(i,j) = h(l(i,j),\ c(i,j),\ s(i,j))$, where $l(i,j)$, $c(i,j)$ and $s(i,j)$ record the number of likes, comments and shares, respectively, performed by u_i and u_j together. $h(\cdot)$ combines these users' behaviors together, which is defined as a sign function in this paper. $\Psi(i,j) = 1$ if $l(i,j) + c(i,j) + s(i,j) > 0$, 0 otherwise. With $\Psi(i,j)$, we model user-user correlations by minimizing the following term as

$$\min_{\mathbf{U}\geq\mathbf{0}} \quad \sum_{i=1}^{n}\sum_{j=1}^{n} \Psi(i,j)\|\mathbf{U}(i,:) - \mathbf{U}(j,:)\|_2^2 \tag{3}$$

Users close to each other in the low-rank space are more likely to be similar and their distances in the latent space are controlled by their correlation coefficients. For example, $\Psi(i,j)$ controls the latent distance between u_i and u_j. A larger value of $\Psi(i,j)$ indicates that u_i and u_j are more likely to be similar. Thus, we force their latent representation as close as possible, while a smaller value of $\Psi(i,j)$ tells that the distance of their latent representation should be larger.

For a particular user u_i, the terms in Eq. (3) related to her latent representation $\mathbf{U}_i$ are,

$$\min_{\mathbf{U}\geq\mathbf{0}} \quad \sum_{j=1}^{n} \Psi(i,j)\|\mathbf{U}(i,:) - \mathbf{U}(j,:)\|_2^2 \tag{4}$$

We can see that the latent representation of u_i is smoothed with other users, controlled by $\Psi(i,j)$, hence even for long tail users, with a few or even without any actions, we can still get an approximate estimate of their latent representation via user-user correlations, addressing the sparsity problem in Eq. (2). We can rewrite the matrix form of Eq. (3) as

$$\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} \Psi(i,j)\|\mathbf{U}(i,:) - \mathbf{U}(j,:)\|_2^2 = Tr(\mathbf{U}^\top\mathcal{L}^u\mathbf{U}), \tag{5}$$

where $\mathcal{L}^u = \mathbf{D}^u - \mathbf{S}$ is the Laplacian matrix and $\mathbf{D}^u$ is a diagonal matrix with the i-th diagonal element $\mathbf{D}^u(i,i) = \sum_{j=1}^{n}\Psi(j,i)$. The user-user correlation matrix $\mathbf{S}$ is,

$$\mathbf{S} = \begin{pmatrix} \Psi(1,1) & \Psi(1,2) & \cdots & \Psi(1,n) \\ \Psi(2,1) & \Psi(2,2) & \cdots & \Psi(2,n) \\ \vdots & \vdots & \ddots & \vdots \\ \Psi(n,1) & \Psi(n,2) & \cdots & \Psi(n,n) \end{pmatrix}$$

Modeling Post-Post Correlations. Apart from likes, posts also receive other actions including commenting and sharing from users. This subsection provides a way to include these activities on posts. Consistency (Abelson 1983) is one of the important social theories developed to explain users' actions, which suggests that users' actions on similar posts are likely to remain consistent. These intuitions motivate us to obtain low-rank space representation of posts based on historical actions received by them. We define $\Phi(i,j)$ to measure the post-post correlation between v_i and v_j. In this paper, we choose the similarity of actions received by posts to measure post-post correlations. A post can receive a variety of actions, including liking, commenting, and sharing. Hence, similarity is calculated as a function of the total amount of actions received by two posts together as $\Phi(i,j) = g(l(i,j),\ c(i,j),\ s(i,j))$, where $l(i,j)$, $c(i,j)$ and $s(i,j)$ record the number of users who perform likes, comments and shares, respectively, on p_i and p_j together. $g(\cdot)$ combines these users' behaviors together, which is defined as a sign function in this paper. $\Phi(i,j) = 1$ if $l(i,j) + c(i,j) + s(i,j) > 0$, 0 Otherwise. With $\Phi(i,j)$, we model post-post correlations by minimizing the following term as

$$\min_{\mathbf{V} \geq \mathbf{0}} \quad \sum_{i=1}^{m} \sum_{j=1}^{m} \Phi(i,j) \|\mathbf{V}(i,:) - \mathbf{V}(j,:)\|_2^2 \tag{6}$$

Posts close to each other in the low-rank space are more likely to be similar and their distances in the latent space are controlled by their correlation coefficients. For example, $\Phi(i,j)$ controls the latent distance between v_i and v_j. A larger value of $\Phi(i,j)$ indicates that v_i and v_j are more likely to be similar. Thus, we force their latent representations should be as close as possible, while a smaller value of $\Phi(i,j)$ tells that the distance of their latent representation should be larger.

For a particular post v_i, the terms in Eq. (3) related to its latent representation $\mathbf{V}_i$ are,

$$\min_{\mathbf{V} \geq \mathbf{0}} \quad \sum_{j=1}^{m} \Phi(i,j) \|\mathbf{V}(i,:) - \mathbf{V}(j,:)\|_2^2 \tag{7}$$

We can see that the latent representation of v_i is smoothed with other posts, controlled by $\Phi(i,j)$, hence even for long tail posts, with a few or even without any actions, we can still get an approximate estimate of their latent representation via post-post correlations, addressing the sparsity problem in Eq. (2). Similar to Eq. (5), we can rewrite the matrix form of Eq. (6) as

$$\frac{1}{2} \sum_{i=1}^{m} \sum_{j=1}^{m} \Phi(i,j) \|\mathbf{V}(i,:) - \mathbf{V}(j,:)\|_2^2 = Tr(\mathbf{V}^\top \mathcal{L}^v \mathbf{V}), \tag{8}$$

where $\mathcal{L}^v = \mathbf{D}^v - \mathbf{P}$ is the Laplacian matrix and $\mathbf{D}^v$ is a diagonal matrix with the i-th diagonal element $\mathbf{D}^v(i,i) = \sum_{j=1}^{m} \Phi(j,i)$. The post-post correlation matrix $\mathbf{P}$ is

$$\mathbf{P} = \begin{pmatrix} \Phi(1,1) & \Phi(1,2) & \cdots & \Phi(1,n) \\ \Phi(2,1) & \Phi(2,2) & \cdots & \Phi(2,n) \\ \vdots & \vdots & \ddots & \vdots \\ \Phi(n,1) & \Phi(n,2) & \cdots & \Phi(n,n) \end{pmatrix}$$

With the components of modeling user-user and post-post correlations, the proposed algorithm is to solve the following optimization problem first.

$$\begin{aligned} \min_{\mathbf{U},\mathbf{H},\mathbf{V} \geq \mathbf{0}} \quad & \|\mathbf{W} \odot (\mathbf{R} - \mathbf{U}\mathbf{H}\mathbf{V}^\top)\|_F^2 + \lambda(\|\mathbf{U}\|_F^2 + \|\mathbf{V}\|_F^2 + \|\mathbf{H}\|_F^2) \\ & + \alpha Tr(\mathbf{U}^\top \mathcal{L}_u \mathbf{U}) + \beta Tr(\mathbf{V}^\top \mathcal{L}_v \mathbf{V}), \end{aligned} \tag{9}$$

where the first term is used to exploit the available users' like actions on posts, second term captures user-user correlations, and post-post correlations are captured by third term. The parameter α and β is introduced to control the contribution from user-user and post-post correlations, respectively.

The optimization problem in Eq. (9) is a multi-objective with respect to the three variables $\mathbf{U}$, $\mathbf{H}$, and $\mathbf{V}$ together. A local minimum of the objective function $\mathcal{J}$ in Eq. (9) can be obtained through an alternative scheme. By following the derivations in (Ding *et al.* 2006), the optimization problem in Eq. (9) leads to following updating rules for $\mathbf{H}$, $\mathbf{U}$, and $\mathbf{V}$.

$$\mathbf{H}(i,j) \leftarrow \mathbf{H}(i,j) * \sqrt{\frac{[\mathbf{U}^\top (\mathbf{W} \odot \mathbf{W} \odot \mathbf{R})\mathbf{V}]\,(i,j)}{[\mathbf{U}^\top (\mathbf{W} \odot \mathbf{W} \odot \mathbf{U}\mathbf{H}\mathbf{V}^\top)\mathbf{V} + \lambda \mathbf{H}]\,(i,j)}} \tag{10}$$

$$\mathbf{U}(i,j) \leftarrow \mathbf{U}(i,j) * \sqrt{\frac{[(\mathbf{W} \odot \mathbf{W} \odot \mathbf{R})\mathbf{V}\mathbf{H}^\top + \alpha \mathbf{S}\mathbf{U}]\,(i,j)}{[(\mathbf{W} \odot \mathbf{W} \odot \mathbf{U}\mathbf{H}\mathbf{V}^\top)\mathbf{V}\mathbf{H}^\top + \alpha \mathbf{D}_u \mathbf{U} + \lambda \mathbf{U}]\,(i,j)}} \tag{11}$$

$$\mathbf{V}(i,j) \leftarrow \mathbf{V}(i,j) * \sqrt{\frac{[(\mathbf{W} \odot \mathbf{W} \odot \mathbf{R})^\top \mathbf{U}\mathbf{H} + \beta \mathbf{P}\mathbf{V}]\,(i,j)}{[(\mathbf{W} \odot \mathbf{W} \odot \mathbf{U}\mathbf{H}\mathbf{V}^\top)^\top \mathbf{U}\mathbf{H} + \beta \mathbf{D}_v \mathbf{V} + \lambda \mathbf{V}]\,(i,j)}} \tag{12}$$

It can be proven that updating rules in Eq. (10), Eq. (11) and Eq. (12) are guaranteed to converge. Since the proof process is similar to that in (Seung & Lee 2001; Ding *et al.* 2006), to save space, we omit the detailed proof of the convergence of the updating rules in Eq. (10), Eq. (11) and Eq. (12).

After obtaining the low-rank representation of $\mathbf{U}$, we choose the well-known linear SVM as the basic classifier for the attribute prediction task. We train a SVM classifier based on the representation of the labeled users $\mathbf{U}_L$ and their labels $\mathbf{Y}_L$. Note that the proposed framework uses SVM as the basic classifier which only takes discrete values as labels such as gender, interested-in, relationship status, and religious affiliations etc. For a continuous valued attribute such as age, we simply perform discretization for SVM. Actually we can choose regression models as basic models to deal with continued value attributes and we would like to leave it as the future work.

Experiments

In this section, we conduct experiments to answer the following two questions - (1) can the proposed framework predict users' attributes from unprotected interactions? and (2) is it necessary to learn a compact representation? After the introduction of experimental settings, we elaborate results answering above two questions with respect to the proposed framework.

Facebook Datasets

For experiments, we collect a Facebook dataset consisting of users' interactions on the "Basher Kella" Page (Basherkella) during the recent events of the Bangladesh protests (Shahbag). The "Basher Kella" Facebook page represents the influential political organization in Bangladesh which also has the records of supporting violence(Instigation). This Facebook page[3] was founded on March 7, 2013. From March 7, 2013 till April 21, 2013, we collect Facebook users' actions on all the posts published on this page. For each post, we collect all the users who likes, comments and shares it. For each comment on a post, we collect all the users who like, and reply it. For each reply, we collect all the users who likes it. For each share, we collect all the users who like and comment on it. Finally, for each comment on a share, we also collect users who like, and reply it. Also, for each reply on a share comment, we collect all the users who likes it. Table 1 shows the overall statistics of the dataset used for experiments.

In this work, we choose three attributes, i.e., religious affiliation, relationship status and interested-in preference.

[3] Old version of the Basher Kella Facebook page was banned during the recent events of the Bangladesh protests due to its violent content. On March 7, a new page was created which can be accessed using`https://www.facebook.com/newbasherkella`

Table 1: Statistics of the Facebook Dataset

# of days crawl	47
# of users	498,674
# of public posts	9,907
Avg # of Likes per user	15.87
Avg # of Likes per post	580.10
Avg # of Comments per user	1.20
Avg # of Comments per post	44.11
Avg # of Shares per user	2.79
Avg # of Shares per post	139.89

For the religious affiliation attribute, to establish the ground truth for evaluation, we first use the Facebook graph search API results to examine the set of users who set the attribute available to the public and then collect the attributes of these users with their public interaction data to establish a dataset, Facebook-religion, to assess the performance of the proposed framework. The statistics of Facebook-religion is shown in Table 2. We follow the similar process to establish another two ground truth data datasets such as Facebook-relation and Facebook-interest to evaluate the proposed framework in predicting attributes such as relationship status and interested-in preference, respectively. The statistics of Facebook-relation and Facebook-interest are shown in Tables 3 and 4, respectively. Note that all the Facebook users with relation status values as “married”, “engaged” and “in a relationship” are considered not-single, whereas “single”, “widowed”, and “divorced” are considered single.

For each dataset, we choose $x\%$ of the dataset for training and the remaining $(1-x)\%$ as testing. In this work, we vary x as $\{50, 60, 70, 80, 90\}$. For each x, we repeat the experiments 5 times and report the average performance. From the evaluation perspective, precision and recall are equally important for the prediction task. For example, in an Islamic country like Bangladesh, the cost of incorrectly predicting someone as atheist could be disastrous, as it carries connotations of blasphemy(Atheism). However, precision, recall, and F1-score are biased towards one of the labels. Hence, it is unsuitable for our unbalanced evaluation dataset. For this purpose, we use commonly adopted *macro-average F1* score to assess the prediction performance, as it gives equal weight to all the labels. The macro-average F1 score is defined as $macro-F1 = \frac{\sum_{i=1}^{K} F_i}{K}$, where $F_i = \frac{2p_i r_i}{p_i + r_i}$, p_i and r_i refer to the precision and recall values associated with the i-th label, respectively. Note that F-score can not be computed for a baseline which always picks the majority label for the prediction.

Performance Evaluation

In this subsection, we conduct experiments to answer the first question - can the proposed framework predict users' personal attributes from users' unprotected interactions? To answer this question, we investigate the performance of the proposed framework by comparing it with the random performance. For SCOUT, we choose the cross-validation to determine the parameter values and more details about the parameter analysis will be discussed in the following subsection. We empirically set the number of latent dimensions d to 50. The performance results for Facebook-religion, Facebook-relation and Facebook-interest are demonstrated in Figures 1, 2 and 3, respectively.

Table 2: Statistics of Facbook-religion Dataset

Religion	# of Users	Percentages(%)
Muslim	1866	65.40
Atheist	216	7.57
Buddhist	113	3.96
Hindu	463	16.23
Christian	195	6.84
Total	2853	100

Table 3: Statistics for the Facebook-relation Dataset

Values	# of Users	Percentages(%)
Single	760	65.01
Not-single	409	34.99
Total	1169	100

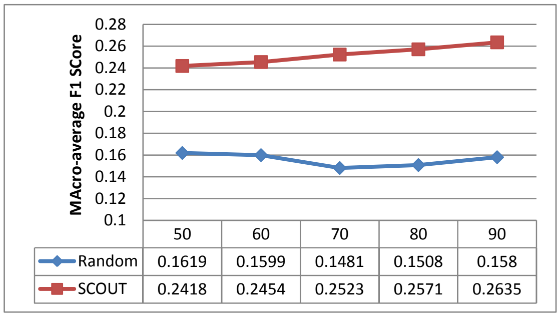
Table 4: The Statistics for Facebook-interest Dataset

Values	# of Users	Percentages(%)
Likes Men	196	9.65
Likes Women	507	24.96
Likes both Men and Women	1328	65.39
Total	2031	100

We have the following observations:

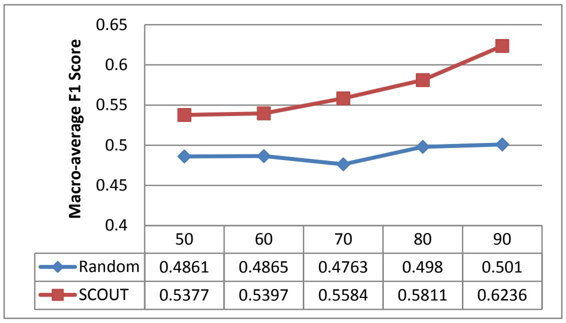
- For all the Figures 1, 2 and 3, the performance of SCOUT increases with the increase of x. This is due to the fact that more training data helps to build a better SVM classifier.
- The proposed framework consistently outperforms the random method. The proposed algorithm gains up to 70.49% and 49.83% relative improvement in Facebook-religion and Facebook-interest, respectively. We conduct a t-test on these results and the evidence from t-test suggests that the improvement is significant. These results support that users' personal attributes are predictable from public interaction data. In the following subsections, we will investigate the contributions from different components to this improvement.

In conclusion, above results suggest a positive answer to the first questions - the proposed framework can predict various users' personal attributes from unprotected interactions.



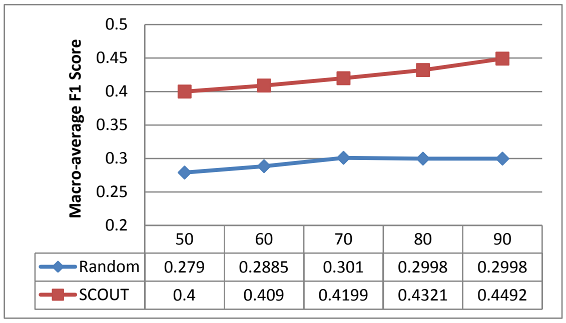
	50	60	70	80	90
Random	0.1619	0.1599	0.1481	0.1508	0.158
SCOUT	0.2418	0.2454	0.2523	0.2571	0.2635

Figure 1: Performance of the Proposed Framework in Predicting Religious Affiliation.



	50	60	70	80	90
Random	0.4861	0.4865	0.4763	0.498	0.501
SCOUT	0.5377	0.5397	0.5584	0.5811	0.6236

Figure 2: Performance of the Proposed Framework in Predicting Relationship Status.



	50	60	70	80	90
Random	0.279	0.2885	0.301	0.2998	0.2998
SCOUT	0.4	0.409	0.4199	0.4321	0.4492

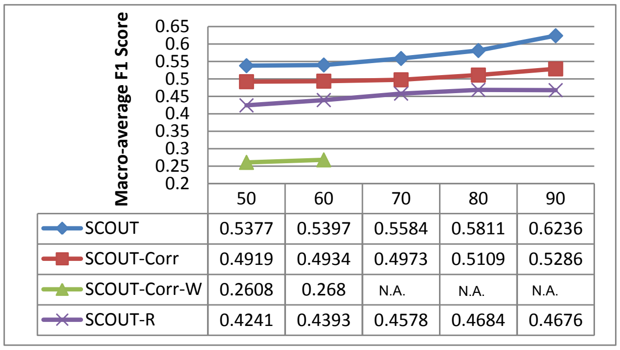
Figure 3: Performance of the Proposed Framework in Predicting Interested-in Preference.

Impact of the Learnt Compact Representation

As mentioned above, the public interaction data representation matrix $\mathbf{R}$ is very sparse and we proposed an algorithm to learn a compact representation $\mathbf{U}$ with the help of social theories. In this subsection, we study the impact of the compact representation on the performance of the proposed framework to answer the second question. In detail, we define the following variants:

- SCOUT-Corr: Eliminate the impact of the user-user and post-post correlations by setting $\alpha = \beta = 0$ in Eq. (9);
- SCOUT-Corr-W: Eliminate the effects from both the correlation and the weight matrix $\mathbf{W}$ by setting $\alpha = \beta = 0$ and $\mathbf{W}$ to be a matrix with all entities equal to 1 in Eq. (9);
- SCOUT-R: Eliminate the impact of the compact representation by learning the SVM classifier with the original matrix $\mathbf{R}$.

We only show results on Facebook-relation in Figure 4 since we have similar observations with other settings. Note that sometimes the classifier gives the majority prediction and we can not compute macro-F1 in this situation; hence we use "N.A." to denote the performance in the table. When eliminating the impact of the user-user and post-post correlations, the performance of SCOUT-Corr reduces, which indicates the importance of incorporating these correlations based on social theories. When eliminating these correlations and the weight matrix $\mathbf{W}$, a traditional matrix factorization algorithm learns the compact representation and the performance of SCOUT-Corr-W reduces dramatically. As mentioned before, the extreme sparsity of $\mathbf{R}$ will lead to the learned compact representation close to zero. When building the classifier based on the sparse matrix $\mathbf{R}$, the performance of SCOUT-R also reduces a lot. We can not learn a good classifier based on the sparse and high-dimensional matrix $\mathbf{R}$, which directly supports the importance of learning a compact representation for uesrs.



	50	60	70	80	90
SCOUT	0.5377	0.5397	0.5584	0.5811	0.6236
SCOUT-Corr	0.4919	0.4934	0.4973	0.5109	0.5286
SCOUT-Corr-W	0.2608	0.268	N.A.	N.A.	N.A.
SCOUT-R	0.4241	0.4393	0.4578	0.4684	0.4676

Figure 4: Impact of the Learnt Compact Representation on the Proposed Framework. Note that sometimes the classifier gives the majority prediction and we can not compute macro-F1 in this situation; hence we use "N.A." to denote the performance in the table.

In conclusion, the learned compact representation can mitigate the sparsity problem of unprotected interactions and plays an important role in the performance improvement of the proposed framework, which correspondingly answers the second question.

Related Work

Psychologists have long been predicting user traits and attributes based on various types of information such as samples of written text (Fast & Funder 2008), answers to psychometric tests (Costa & McCrae 1992), or the appearances of places people inhibit (Gosling *et al.* 2002). Most of these researches are based on an assumption that users have tendencies to inadvertently leave behind cues which correlate with their personal attributes. Recently computer scientists are also exploring users' personal attributes based on cues from the web, such as user's web site browsing logs (Murray & Durrell 2000; Hu *et al.* 2007; De Bock & Van den Poel 2010; Goel, Hofman, & Sirer 2012), contents of personal web sites (Marcus, Machilek, & Schutz 2006), and music collections (Rentfrow & Gosling 2003).

Social media popularity has created several opportunities for users to create data. Massive amount of social media data has attracted attention of privacy researchers to identify, measure and mitigate the risks of predicting personal attributes (Jernigan & Mistree 2009; Mislove *et al.* 2010; Rao & Yarowsky 2010; Conover *et al.* 2011; Quercia *et al.* 2011; Golbeck *et al.* 2011; Li *et al.* 2012; Cohen & Ruths 2013). An inspiring work from Kosinski et.al. (Kosinski, Stillwell, & Graepel 2013) shows that wide variety of people's highly sensitive personal attributes can be automatically and accurately inferred using the variety of Facebook likes. To the best of our knowledge, this work is different from most of previous work as previous work do not distinguish between protected and unprotected data and use them uniformally for attribute prediction tasks. The proposed framework SCOUT focuses exclusively on unprotected interactions.

Future Work

We believe that proposed problem can be explored further along following directions. Some attributes are likely to be correlated. For example, age values are likely to have correlation with relationship statuses; gender values are likely to have correlation with occupations; and religious affiliations are likely to have correlation with political affiliations. These observations can be explored further to see whether attribute correlations can be explored further to achieve better prediction performance. Once aware of such privacy attacks from unprotected interactions, it is interesting to design different ways to protect users' privacy while creating minimal restriction on their social interactions. Finally we would like to extend the proposed framework with regression methods to deal with continuous valued attributes directly.

References

Abelson, R. P. 1983. Whatever Became of Consistency Theory? *Personality and Social Psychology Bulletin*.

Assam. http://en.wikipedia.org/wiki/2012_Assam_violence.

Atheism. http://en.wikipedia.org/wiki/Atheism_and_religion\#Atheism_in_Islam.

Basherkella. https://www.facebook.com/newbasherkella.

Cohen, R., and Ruths, D. 2013. Classifying political orientation on twitter: Its not easy! In *Proceedings of the 7th International Conference on Weblogs and Social Media*.

Conover, M. D.; Gonçalves, B.; Ratkiewicz, J.; Flammini, A.; and Menczer, F. 2011. Predicting the political alignment of twitter users. In *Privacy, security, risk and trust (passat), 2011 ieee third international conference on and 2011 ieee third international conference on social computing (socialcom)*, 192–199. IEEE.

Costa, P. T., and McCrae, R. R. 1992. *Revised neo personality inventory (neo pi-r) and neo five-factor inventory (neo-ffi)*, volume 101. Psychological Assessment Resources Odessa, FL.

De Bock, K., and Van den Poel, D. 2010. Predicting website audience demographics forweb advertising targeting using multi-website clickstream data. *Fundamenta Informaticae* 98(1):49–70.

Ding, C.; Li, T.; Peng, W.; and Park, H. 2006. Orthogonal nonnegative matrix tri-factorizations for clustering. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, 126–135. ACM.

Fast, L. A., and Funder, D. C. 2008. Personality as manifest in word use: correlations with self-report, acquaintance report, and behavior. *Journal of personality and social psychology* 94(2):334.

Goel, S.; Hofman, J. M.; and Sirer, M. I. 2012. Who does What on the Web: A Large-scale Study of Browsing Behavior. In *Proceedings of ICWSM*.

Golbeck, J.; Robles, C.; Edmondson, M.; and Turner, K. 2011. Predicting Personality from Twitter. In *Proceedings of PASSAT and SOCIALCOM*.

Gosling, S. D.; Ko, S. J.; Mannarelli, T.; and Morris, M. E. 2002. A room with a cue: personality judgments based on offices and bedrooms. *Journal of personality and social psychology* 82(3):379.

Hu, J.; Zeng, H.-J.; Li, H.; Niu, C.; and Chen, Z. 2007. Demographic Prediction based on User's Browsing Behavior. In *Proceedings of WWW*.

Hu, X.; Tang, J.; Gao, H.; and Liu, H. 2013. Unsupervised sentiment analysis with emotional signals. In *Proceedings of the 22nd international conference on World Wide Web*, 607–618. International World Wide Web Conferences Steering Committee.

Instigation. http://www.thedailystar.net/beta2/news/net-instigation-in-full-force/.

Jernigan, C., and Mistree, B. F. 2009. Gaydar: Facebook friendships expose sexual orientation. *First Monday* 14(10).

Kosinski, M.; Stillwell, D.; and Graepel, T. 2013. Private Traits and Attributes are Predictable from Digital Records of Human Behavior. *Proceedings of the National Academy of Sciences*.

Li, R.; Wang, S.; Deng, H.; Wang, R.; and Chang, K. C.-C. 2012. Towards social user profiling: unified and discriminative influence model for inferring home locations. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, 1023–1031. ACM.

Li, T.; Zhang, Y.; and Sindhwani, V. 2009. A Non-negative Matrix Tri-factorization approach to Sentiment Classification with Lexical Prior Knowledge. In *Proceedings of the ACL*.

Lu, Y.; Tsaparas, P.; Ntoulas, A.; and Polanyi, L. 2010. Exploiting Social Context for Review Quality Prediction. In *Proceedings of WWW*.

Ma, H.; Zhou, D.; Liu, C.; Lyu, M. R.; and King, I. 2011. Recommender systems with social regularization. In *Proceedings of WSDM*.

Marcus, B.; Machilek, F.; and Schutz, A. 2006. Personality in Cyberspace: Personal Web Sites as Media for Personality Expressions and Impressions. *Journal of Personality and Social Psychology* 90(6):1014–1031.

McPherson, M.; Smith-Lovin, L.; and Cook, J. M. 2001. Birds of a Feather: Homophily in Social Networks. *Annual review of sociology* 415–444.

Mislove, A.; Viswanath, B.; Gummadi, K. P.; and Druschel, P. 2010. You are who you know: inferring user profiles in online social networks. In *Proceedings of the third ACM international conference on Web search and data mining*, 251–260. ACM.

Murray, D., and Durrell, K. 2000. Inferring demographic attributes of anonymous internet users. In *Web Usage Analysis and User Profiling*. Springer. 7–20.

Quercia, D.; Kosinski, M.; Stillwell, D.; and Crowcroft, J. 2011. Our Twitter Profiles, Our Selves: Predicting Personality with Twitter. In *Proceedings of PASSAT and SOCIALCOM*.

Rao, D., and Yarowsky, D. 2010. Detecting latent user properties in social media. In *Proc. of the NIPS MLSN Workshop*.

Rentfrow, P. J., and Gosling, S. D. 2003. The Do Re Mi's of Everyday Life: The Structure and Personality Correlates of Music Preferences. *Journal of personality and social psychology* 84(6):1236–1256.

Seung, D., and Lee, L. 2001. Algorithms for Non-negative Matrix Factorization. In *Proceedings of the NIPS*.

Shahbag. `https://en.wikipedia.org/wiki/2013_Shahbag_protests`.

Tang, L., and Liu, H. 2009. Relational learning via latent social dimensions. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, 817–826. ACM.

Tang, J., and Liu, H. 2012. Feature Selection with Linked Data in Social Media. In *SDM*.

Tang, J.; Gao, H.; Hu, X.; and Liu, H. 2013a. Exploiting homophily effect for trust prediction. In *Proceedings of the sixth ACM international conference on Web search and data mining*, 53–62. ACM.

Tang, J.; Hu, X.; Gao, H.; and Liu, H. 2013b. Exploiting Local and Global Social Context for Recommendation. In *Proceedings of IJCAI*.