Proceedings of the Tenth International AAAI Conference on
Web and Social Media (ICWSM 2016)

Exploring Personal Attributes from Unprotected Interactions

Pritam Gundecha*
IBM Research, Almaden
psgundec@us.ibm.com

Jiliang Tang
Yahoo Labs
jlt@yahoo-inc.com

Xia Hu
Texas A&M University
hu@cse.tamu.edu

Huan Liu
Aizona State University
Huan.Liu@asu.edu

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"). 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

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 and Mistree 2009; Gundecha, Barbier, and Liu 2011) Recent research (Kosinski, Stillwell, and 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 an 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 on such Facebook pages including availability, presence of multilingual text, and extremely sparse user interactions. 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.

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 $\|\mathbf{A}\|_F = \sqrt{\sum_{i=1}^{n}\sum_{j=1}^{m}\mathbf{A}(i,j)^2}$.

*Majority of this work was done when first author was a PhD student at Arizona State University

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 of values of a given attribute. 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, 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. 2013). 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 $\mathbf{U}(i,:)\mathbf{H}\mathbf{V}^\top(j,:)$ 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. 2013), 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, and 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 sentiment analysis (Hu et al. 2013) 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.

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, and 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. 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

$$\frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \Psi(i,j) \|\mathbf{U}(i,:) - \mathbf{U}(j,:)\|_2^2 = Tr(\mathbf{U}^\top \mathcal{L}^u \mathbf{U}), \tag{4}$$

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,

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

Similar to Eq. (4), we can also model the post-post correlations in the matrix form as

$$\frac{1}{2} \sum_{i=1}^{m} \sum_{j=1}^{m} \Phi(i,j) \|\mathbf{V}(i,:) - \mathbf{V}(j,:)\|_2^2 = Tr(\mathbf{V}^\top \mathcal{L}^v \mathbf{V}), \tag{5}$$

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

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

Similar to $\Psi(i,j)$, $\Phi(i,j)$ measures the post-post correlation coefficients between posts v_i and v_j.

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

$$\begin{aligned} \min_{\mathbf{U},\mathbf{H},\mathbf{V} \geq \mathbf{0}} \quad & \|\mathbf{W} \odot (\mathbf{R} - \mathbf{U}\mathbf{H}\mathbf{V}^\top)\|_F^2 + \lambda(\|\mathbf{U}\|_F^2 + \|\mathbf{V}\|_F^2 + \|\mathbf{H}\|_F^2) \\ & + \alpha Tr(\mathbf{U}^\top \mathcal{L}_u \mathbf{U}) + \beta Tr(\mathbf{V}^\top \mathcal{L}_v \mathbf{V}), \end{aligned} \tag{6}$$

Table 1: Statistics of the Facebook Dataset

# 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

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. (6) 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 in Eq. (6) can be obtained through an alternative scheme (Ding et al. 2006). 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.

Experiments

In this section, we conduct experiments to answer whether the proposed framework can predict users' attributes from unprotected interactions.

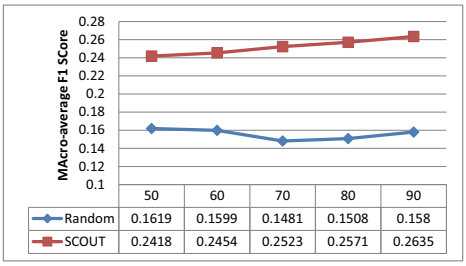
Facebook Datasets

For experiments, we collect a Facebook dataset consisting of users' interactions on the "Basher Kella" Page during the recent events of the Bangladesh protests. 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 to verify efficacy of our framework. Our ground truth dataset contains 2853 Facebook users with five religious affiliations (such as Muslims, Atheist, Buddhist, Hindu, and Christian), 1169 Facebook users with two relationship status values (such as single and not-single), and 2031 Facebook users with three interested-in preference values (such as likes-men, likes-women and likes-both men and women). 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 using commonly adopted *macro-average F1* score.

Performance Evaluation

In this subsection, we conduct experiments to answer - 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

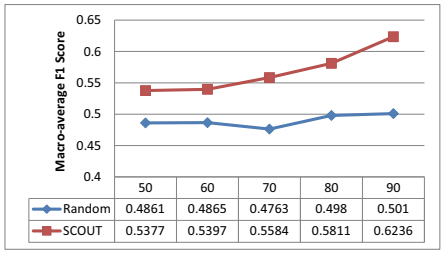


	50	60	70	80	90
Random	0.1619	0.1599	0.1481	0.1508	0.158
SCOUT	0.2418	0.2454	0.2523	0.2571	0.2635

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

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. 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.



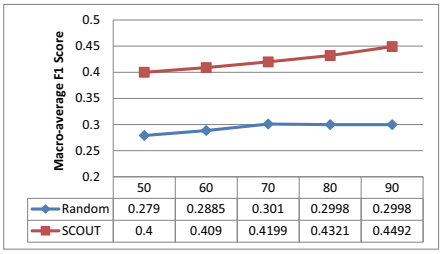
	50	60	70	80	90
Random	0.4861	0.4865	0.4763	0.498	0.501
SCOUT	0.5377	0.5397	0.5584	0.5811	0.6236

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

In conclusion, above results positively suggest that the proposed framework can predict various users' personal attributes from 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



	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.

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.

Acknowledgments

This research is, in part, supported by grants of ARO (#025071) and ONR (N00014-16-1-2257).

References

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

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.

Gundecha, P.; Barbier, G.; and Liu, H. 2011. Exploiting Vulnerability to Secure User Privacy on a Social Networking Site. In *Proceedings of the 17th ACM SIGKDD*.

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.

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*.

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*.

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

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