Assessing the Reliability of Facebook User Profiling

Thomas Theodoridis, Symeon Papadopoulos, and Yiannis Kompatsiaris
Information Technologies Institute (ITI)
Centre for Research and Technology Hellas (CERTH)
Thessaloniki, Greece
{tomastheod,papadop,ikom}@iti.gr

ABSTRACT

User profiling is an essential component of most modern online services offered upon user registration. Profiling typically involves the tracking and processing of users' online traces (e.g., page views/clicks) with the goal of inferring attributes of interest for them. The primary motivation behind profiling is to improve the effectiveness of advertising by targeting users with appropriately selected ads based on their profile attributes, e.g., interests, demographics, etc. Yet, there has been an increasing number of cases, where the advertising content users are exposed to is either irrelevant or not possible to explain based on their online activities. More disturbingly, automatically inferred user attributes are often used to make real-world decisions (e.g., job candidate selection) without the knowledge of users. We argue that many of these errors are inherent in the underlying user profiling process. To this end, we attempt to quantify the extent of such errors, focusing on a dataset of Facebook users and their likes, and conclude that profiling-based targeting is highly unreliable for a sizeable subset of users.

1. INTRODUCTION

Our research focuses on the problem of user profiling in the context of Online Social Networks (OSN), such as Facebook. Such platforms offer their users simple mechanisms, such as *likes*, to express their endorsement or affiliation to news stories, topics, and groups. Aggregating inputs by many users has been recently shown [2] to enable the training of machine learning models to predict, sometimes quite accurately, a variety of user attributes, ranging from demographic information (gender, age) to information of sensitive nature, such as sexual orientation and political opinion. Performing such predictions results into the construction of *inferred* user profiles that are then typically used for ad targeting and other kinds of personalized services (e.g., news recommendation).

Although the motivation behind user profiling practices, such as the aforementioned ones, may appear to be legit at first sight, there has been a growing concern among OSN users about the use of their observed data to perform such profiling in ways that are not transparent, especially with respect to their implications. Such concerns have been exacerbated by recent evidence suggesting that sensitive pieces of information inferred from OSN profiles are systematically used in relation to discrimination practices, e.g., job candidate selection [1], or loan underwriting and pricing [3]. Despite such concerns, user profiling practices are still widely used, and even when used solely for the purpose of offering improved online services, e.g., through personalized recommendations, there are often cases where their results are surprising or even upsetting.

One of the reasons behind users' concerns about profiling is the considerable number of erroneous inferences made on the basis of observed data. To this end, we present an experimental study on a fully anonymized Facebook dataset[1] with the goal of quantifying the sensitivity of automatic inferences and accordingly the extent of erroneous decisions involved in user profiling, and of mitigating the risk of erroneous user classification.

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage, and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s). Copyright is held by the author/owner(s).
WWW 2015 Companion, May 18–22, 2015, Florence, Italy.
ACM 978-1-4503-3473-0/15/05.
http://dx.doi.org/10.1145/2740908.2742728.

2. METHODOLOGY OUTLINE

We build upon the user profiling approach proposed by Kosinski et al. [2]. Given a set of n users and the likes for each one of them (to a total of m Facebook pages), a very sparse user-like matrix L $(n \times m)$ is first created where L_{ij} is set to 1 if user i likes page j and 0 otherwise. In the original approach, Singular Value Decomposition (SVD) is first applied on the matrix to select a small set of k SVD components to represent each user as a k-dimensional vector. Then, given a set of known traits/attributes for a number of users, e.g., "gender", "sexual orientation", "political views", "religion", etc., predictive models are built that can subsequently be used to infer these attributes for a set of unknown users based on their like history. Similar to [2], we used logistic regression for classification[2].

In our study, we added a feature selection step $L' = fs(L)$ $(L' : n \times m', m' \ll m)$, which we found beneficial for the classification accuracy. In this step, we remove features (liked pages) that are selected by only few users. For this, we take into account the average number of likes in each set and recreate the user-like matrix by filling in likes at random. We then count the number of likes for each page and use the 95 percentile as our threshold. Pages in the original matrix with fewer likes than the threshold are removed.

[1] `http://mypersonality.org/wiki/doku.php`
[2] In [2], both classification and regression are considered. Here, we focus on classification without loss of generality.

Table 1: Dataset overview

L#	labels	users (n)	balance	m	m'
L1	gay/straight	2,412	50/50	218,490	15,609
L2	single/married	7,732	50/50	511,775	29,389
L3	liberal/conserv.	4,106	55/45	296,298	18,658
L4	christian/muslim	1,196	74/26	134,120	10,333

A conventional means of reporting the accuracy of classification models is to carry out N-fold cross-validation (with N typically set to 10), i.e. split the set in N equally sized parts, use the $N-1$ for training and the remaining for testing, repeat N times (each time using a different part for testing), and report the average classification accuracy over the N folds in terms of the area under the receiver-operating characteristic curve (AUC), which is equivalent to the probability of correctly classifying two randomly selected users one from each class (e.g., male and female). Although cross-validation provides an overall estimate of the classifier performance, it treats the inferences for the whole set of data points (users) as being equally reliable.

Given the sensitive nature of the problem at hand, we devised an evaluation approach that attempts to capture the risk involved in misclassifying particular data points of the collection. To this end, we independently sample from the available training set B *bags*, each covering α% of the training set, and we use them to train B classification models for the same target label. We then apply the ensemble of all B model outputs on the test set to derive the final predictions using majority vote. Comparing the individual model predictions with the ensemble one, we then quantify the extent to which the prediction for user x is reliable.

$$S_x = \frac{|\sum_i^B (m_i(x) = +1) - \sum_i^B (m_i(x) = -1)|}{B} \quad (1)$$

where S_x stands for the reliability score for the prediction about user x, $m_i(x)$ is the prediction of model m_i for x, and -1, $+1$ denote the two possible labels for the classification at hand. In the case that all models agree on the decision, the score takes a value of 1, while in the case that the decision is made on the basis of just one vote in favour of the majority (B should be always an odd number), the score takes a value of $\frac{1}{B}$. Hence, analyzing the distribution of these scores for all test users, we could reason about the reliability of the performed inferences for different groups of users.

3. EXPERIMENTS & KEY RESULTS

We performed our experiments on a subset of the *myPersonality* dataset [2], focusing on four target label pairs of sensitive nature: a) `gay` vs `straight`, b) `single` vs `married`, c) `liberal` vs `conservative`, and d) `christian` vs `muslim`. For each of those, we performed the feature selection process described in Section 2, and removed those users that had no features associated with them. The resulting dataset statistics per label pair are presented in Table 1. Note that n and m' denote the number of users and liked pages respectively that remained after the feature selection process. We used the top $k = 100$ SVD components for building the classification models. For each of the label pairs, we reserve 10% of the associated users as test samples, and keep the 90% to use for performing the training of models as described in Section 2. We used $B = 25$ bags to create an equal number of models and tested different values of α between 20% and 100% (the latter is equivalent to not using bags).

Table 2: Results. AUC stands for the overall Area Under Curve, while AUC_{HC} and AUC_{LC} stand for the AUC scores for high- and low-confidence classifications respectively. A high-confidence classification for user x occurs when $S_x = 1$, while a low-confidence one occurs when $S_x \leq 0.5$.

L#	α	AUC	U_{HC} (%)	AUC_{HC}	U_{LC} (%)	AUC_{LC}
L1	80	83.0	70.5	90.6	11.8	56.6
	40	82.9	46.5	94.6	22.9	60.6
	20	82.3	28.7	97.6	35.1	65.1
L2	80	69.8	72.6	75.0	11.1	53.4
	40	69.8	52.2	78.5	20.6	52.1
	20	70.1	33.1	80.1	31.2	55.5
L3	80	77.3	77.3	82.9	8.6	53.8
	40	77.7	53.6	88.8	19.3	57.1
	20	77.7	31.0	94.0	32.2	59.3
L4	80	85.1	64.5	94.7	14.7	53.9
	40	84.1	40.9	97.2	23.6	58.8
	20	84.4	27.4	95.7	29.3	67.1

Table 2 presents some of the obtained results. The reported prediction accuracies are consistent with the ones of [2], pointing for instance to the fact that the classification between `gay` and `straight` can be performed more accurately compared to the one between `single` and `married`.

However, a noteworthy observation is that even for labels that can be predicted with high accuracy, there is a sizeable percent of users that are classified to one of the two with very low confidence. For instance, in the case of the `gay/straight` label and for $\alpha = 80$%, there are 11.8% of the test users that are classified to one of the two labels with a reliability score below 0.5. As expected, for those users the classification accuracy drops considerably (56.6% compared to the overall accuracy of 83%). This is even more alarming, given the fact that in the case of $\alpha = 80$%, the 25 classification models were built using many common training samples (since each of those is an 80% random subset of the same set). For lower values of α, the percentage of users who are classified with low confidence is even higher, e.g., 32.2% for labels `liberal/conservative` and $\alpha = 20$%.

Hence, given the fact that such highly unreliable classifications are often used for targeting users, and sometimes have serious real-world consequences (e.g., not being selected for a job), one should be really cautious against inferred user profiles, and could raise serious ethical concerns with respect to the overall practice of mining user profiles from observed data. Yet, being able to quantify the reliability of the performed inferences based on the methodology of Section 2, could be at least used as a measure to mitigate the risk of erroneous profiling (by deciding to not profile at all those users for which the S scores are low).

Acknowledgments: We thank Michal Kosinksi and David Stillwell for giving us access to the myPersonality dataset. This work is supported by the USEMP FP7 project, partially funded by the EC under contract number 611596.

4. REFERENCES

[1] A. Acquisti and C. M. Fong. An Experiment in Hiring Discrimination Via Online Social Networks. *Social Science Research Network Working Paper Series*, Apr. 2012.

[2] M. Kosinski, D. Stillwell, and T. Graepel. Private traits and attributes are predictable from digital records of human behavior. *Proceedings of the National Academy of Sciences*, 110(15):5802–5805, Apr. 2013.

[3] A. S. Raman, J. L. Barloon, and D. M. Welch. Social media: Emerging fair lending issues. *The Review of Banking and Financial Services*, 28(7), July 2012.