

# Your Digital Image: Factors Behind Demographic And Psychometric Predictions From Social Network Profiles

## (Demonstration)

Yoram Bachrach  
Microsoft Research  
yobach@microsoft.com

Thore Graepel  
Microsoft Research  
thoreg@microsoft.com

Pushmeet Kohli  
Microsoft Research  
pkohli@microsoft.com

Michal Kosinski  
University of Cambridge  
mk583@cam.ac.uk

David Stillwell  
University of Cambridge  
ds617@cam.ac.uk

### ABSTRACT

We demonstrate how information gathered from social network profiles can be used to predict personal attributes such as gender and age, religious and political views, intelligence, happiness and personality traits.

Our approach is based on applying machine learning techniques to a large dataset of people who volunteered their Facebook profiles along with their demographic and psychometric test results. We combine various features from the profile, including the numbers or rates of posting status updates, pictures and group memberships, and the specific items liked by individuals. Our system provides insights regarding how the predictions are made, allowing people to understand how they may be perceived by others based on their social network profiles.

### Categories and Subject Descriptors

K.4 [Computers and Society]: Social Issues

### Keywords

Social Networks, Analytics, Machine Learning, Personality Predictions, Demographic Predictions, Privacy

## 1. INTRODUCTION

Social networking sites (SNS), such as Facebook, have become immensely popular in recent years. Such sites record much information about their users, including their opinions, preferences and interactions. Recent work shows that many traits of individuals can be predicted automatically and accurately by using the wealth of information from SNS, including age and gender [9] personality [7, 4, 1], and even ethnicity, religious or political views, intelligence and happiness [6]. Automated approaches typically rely on obtaining a large dataset consisting of social network profiles along with personal traits of individuals, and applying machine learning techniques to build models that predict traits of individuals given the features of their social network profile.

**Appears in:** *Alessio Lomuscio, Paul Scerri, Ana Bazzan, and Michael Huhns (eds.), Proceedings of the 13th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2014), May 5-9, 2014, Paris, France.*  
Copyright © 2014, International Foundation for Autonomous Agents and Multiagent Systems (www.ifaamas.org). All rights reserved.

Some profile features that are used to predict properties of the profile owner are *macro-features* — aggregates which quantify the total activity or behavior of a certain type, such as the total number of photos uploaded to a profile in a month, the number of friends or the average number of words in status updates. Other approaches use *micro-features*, relating to very specific profile properties — whether the user is a member of a certain group, or expressed interest in a particular item (such as a movie, website, or book).

### 1.1 Prediction Approaches

One approach that uses micro-features predicts many user traits, such as sexual orientation, ethnicity, religious and political views, personality, intelligence, happiness, use of addictive substances, parental separation, age, and gender [6]. It uses Facebook Likes, which allow Facebook users to express a positive association (“Like”) with online content, including status updates, photos, or websites.

Given a set  $S$  of objects that can be “Liked”, it represents a user SNS profile as a binary vector, with entries set to 1 if the user Liked the corresponding object (i.e. if there is an association between the user and object) and 0 otherwise. An encoding of the user demographic or psychometric traits is concatenated to the encoding of the SNS profile, so each user is expressed as a vector of equal length. The set of all users is encoded as a matrix where each row is the encoding of a user. After applying singular-value decomposition (SVD) [3], regression methods are used to predict a user’s traits given their SNS profile.

Another approach uses macro-features to predict personality traits [1], expressed using the Five Factor Model [2, 8], a widespread model used by psychologists to represent the key features underlying human personality. This approach predicts personality by multiple linear regression, using features such as the number of Facebook friends, group associations, “Likes”, photos and status updates.

Both the above approaches use the same dataset to train the machine learning model, released by the *myPersonality* project<sup>1</sup>. This is a Facebook application, first deployed in 2007, which lets Facebook users complete a demographics questionnaire and a standard personality questionnaire to obtain feedback regarding their personality. Users may give their consent to record their profile information responses,

<sup>1</sup>available at: <http://mypersonality.org/wiki/>

allowing the myPersonality dataset to correlate profile features and user traits for a very large user population.

In order to improve the prediction accuracy, it is advisable to use all available features. Further, we wish to help users understand how parts of the profile information affect predictions, as this allows them to determine how others may perceive them and what affects these perceptions. We demonstrate an approach that combines both micro and macro features to predict traits of the profile owners, and that helps users to understand what drives these predictions. Our approach also relies on the myPersonality dataset, and allows predicting traits such as demographics and personality, similarly to the methods discussed above [6, 1].

## 2. COMBINED PREDICTION

Our approach uses multiple linear regression for continuous variables, and logistic regression for boolean variables. To process “Liked” items, we use an SVD of the “Liked” item matrix, similarly to the approach discussed in the introduction [6]. However, rather than relying solely on these micro-features, we augment the vector encoding a user with the macro-level features used by personality predicting approaches [4, 1].

We build a model for each predicted trait, such as gender, age or each of the Big Five personality traits, and train the model using the myPersonality dataset. To identify the key features that influence the predictions, our system shows which of the “Liked” items resulted in a change in the predicted value of traits, and allows viewing the predictions based only on the macro features or only on the micro features. This allows users of our prediction engine to better understand how various factors affect our predictions. Our combined prediction approach is flexible, allowing integrating additional explanatory variables easily: in order to use another profile feature to improve the prediction accuracy of all the user traits, one only needs to add these as an additional column to the matrix representing the users.

We have built the system using the myPersonality project, which only contains Facebook profile features. However, we emphasize that the same approach allows integrating data from several sources, including other social networks, or shopping histories or browsing histories.<sup>2</sup> Our system only analyzes the status updates in a shallow manner, at the macro level; it does not examine the textual content of the status updates, but only the total number of these updates. Recent work indicates that linguistic characteristics are predictive of a user’s traits [9]. Our approach can use such information by augmenting the vector encoding with columns representing specific words in the updates.

## 3. CONCLUSION

We demonstrate an approach for aggregating information from SNS profiles to predict user traits. Such predictions have many applications, ranging from tailoring products or recommendations to users, through personalized search engines to better targeted advertising. However, the ability to infer potentially sensitive information about users is also a cause for privacy concerns. This highlights the need for

<sup>2</sup>For example, previous work already shows that personality can be predicted using information about the websites people tend to visit [5]. Such information is contained in a user’s browsing history, and can be integrated into our framework.

proper controls, that would balance enable the important applications of this technology, while protecting user privacy.

Our system allows users to examine the factors influencing the predictions, so users can determine how “Liking” a certain item changes the predictions regarding their intelligence, or how changing the number of friends they have affects the predictions regarding their personality. Clearly, these factors are under the control of the user, and users may modify their behavior on Facebook to be perceived in a positive manner. As people can form judgments on others based on their social media profiles [4], this phenomenon is not new. However, we believe an automated tool can allow people to easily determine how others may perceive them based on their behavior on social networks.

Many questions are open for future research. How can we design systems that would give people more control over the information they reveal? Can such accurate predictions be made based on other publicly available information, such as personal webpages or blogs? How is the availability of tools that allow making such predictions likely to change the way in which people communicate and interact on social media?

## 4. REFERENCES

- [1] Y. Bachrach, M. Kosinski, T. Graepel, P. Kohli, and D. Stillwell. Personality and patterns of Facebook usage. In *Proceedings of the 3rd Annual ACM Web Science Conference*, pages 24–32. ACM, 2012.
- [2] P. Costa Jr and R. McCrae. Neo personality inventory–revised (NEO-PI-R) and NEO five-factor inventory (NEO-FFI) professional manual. *Odessa, FL: Psychological Assessment Resources*, 1992.
- [3] G. Golub and W. Kahan. Calculating the singular values and pseudo-inverse of a matrix. *Journal of the Society for Industrial & Applied Mathematics, Series B: Numerical Analysis*, 2(2):205–224, 1965.
- [4] S. D. Gosling, S. Gaddis, S. Vazire, et al. Personality impressions based on Facebook profiles. In *ICWSM*, 2007.
- [5] M. Kosinski, P. Kohli, D. Stillwell, Y. Bachrach, and T. Graepel. Personality and website choice. In *ACM Web Science Conference*, pages 251–254, 2012.
- [6] 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, 2013.
- [7] C. Ross, E. S. Orr, M. Sisic, J. M. Arseneault, M. G. Simmering, and R. R. Orr. Personality and motivations associated with Facebook use. *Computers in Human Behavior*, 25(2):578–586, 2009.
- [8] M. Russell, D. Karol, I. for Personality, and A. Testing. *The 16PF fifth edition administrator’s manual*. Institute for Personality and Ability Testing Champaign, IL, 1994.
- [9] H. A. Schwartz, J. C. Eichstaedt, M. L. Kern, L. Dziurzynski, S. M. Ramones, M. Agrawal, A. Shah, M. Kosinski, D. Stillwell, M. E. Seligman, et al. Personality, gender, and age in the language of social media: The open-vocabulary approach. *PLoS one*, 8(9):e73791, 2013.