Supporting Information

Kosinski et al. 10.1073/pnas.1218772110

SI Text

SI Results. Table S1 presents Likes characterized by the most extreme average levels for each of the numeric variables (e.g. personality traits) or most extreme frequencies of classes (e.g. being a Democrat). Fig. S1 shows the average levels of personality traits and age of the users associated with selected Likes presented on the percentile scale. Fig. S2 presents relative popularity of selected Likes within groups of Democrat, Homosexual, Christian, and African-American users.

Sample. We used data from 58,466 US Facebook users, including their psychodemographic profile and their list of Likes. The data were obtained from the myPersonality application (www.mypersonality.org). Users opted in to provide their data for this study and gave their consent to have their scores and profile information recorded for analysis.

An important limitation of our sample is that some of the predicted variables are from Facebook profile information. Individuals who declare their political and religious views, relationship status, and sexual orientation on their profile may be different from nondeclaring members of those groups; they may associate with distinct Likes, which may lead to an overestimate of prediction accuracies for these groups. Nevertheless, the model was still able to predict privately reported information, such as personality or intelligence quotient questionnaire results, and survey results on addictive substance use.

Political and Religious Views, Sexual Orientation, Relationship Status. Political and religious views were recorded from the respective fields of users' Facebook profiles. Both fields allow users to input text freely (but suggest popular choices). Political views "Democrat," "Democratic," or "Democratic Party" were recoded to "Democrat." "Republican," "GOP," and "Republican Party" were recoded to "Republican"; other entries were ignored. Religious views "Christian," "Catholic," and "Jesus Christ" were recoded to "Christian." "Moslem," "Muslim," "Islam," and "Sunni" were recoded to "Muslim." Sexual orientation was taken from the "Interested in" section of users' Facebook profiles; users who listed being interested in only the opposite gender were labeled as being heterosexual, whereas users who listed only the same gender were labeled as being homosexual. Relationship status was recorded from the "Relationship Status" profile field, where the options were "Single," "It's complicated," "In an open relationship," "In a relationship," "Engaged," and "Married." The latter three options were recoded to "In a relationship."

Ethnicity. Labels for ethnicity were assigned to users by visual inspection of their profile pictures. This procedure has the advantage that the data are not explicitly self-reported and, hence, does not suffer from disclosure bias. However, some users do not include any picture with their profile or use a picture that does not show themselves. To confirm the reliability of the manual classification procedure, a subsample of the data were compared with self-reported ethnic background from a survey, and there was $r = 0.98$ agreement between the two sets of labels.

Substance Use and User's Parents Together at Age Twenty-One Years. Both substance use and whether a user's parents stayed together or split up before the user was 21 y old were measured using self-report survey measures on the myPersonality application. These questions were explicitly labeled as optional. Individuals were asked if they smoked daily, less than daily, or were nonsmokers; less than daily and daily were recoded as "smokers." They were also asked if they drank alcohol by offering the choices "weekly or more often," "less than once a week," or "never"; the first two options were recoded as "drinkers." For drug use, the options were the same as for drinking; the first two options were recoded as "drug users."

Personality. Five-Factor Model (FFM) (1) personality scores ($n = 54,373$) were established using the International Personality Item Pool (IPIP) questionnaire with 20 items (2). This test is widely used in both traditional and online studies and is known to be successful at explaining variability across individuals. FFM encompasses the following traits.

Openness to Experience. Openness to experience ("Openness") is related to imagination, creativity, curiosity, tolerance, political liberalism, and appreciation for culture. People scoring high on Openness like change, appreciate new and unusual ideas, and have a good sense of aesthetics.

Conscientiousness. "Conscientiousness" measures preference for an organized approach to life in contrast to a spontaneous one. Conscientious people are more likely to be well organized, reliable, and consistent. They enjoy planning, seek achievements, and pursue long-term goals. Nonconscientious individuals are generally more easy-going, spontaneous, and creative. They tend to be more tolerant and less bound by rules and plans.

Extraversion. "Extraversion" measures a tendency to seek stimulation in the external world, the company of others, and to express positive emotions. Extraverts tend to be more outgoing, friendly, and socially active. They are usually energetic and talkative; they do not mind being at the center of attention and make new friends more easily. Introverts are more likely to be solitary or reserved and seek environments characterized by lower levels of external stimulation.

Agreeableness. "Agreeableness" relates to a focus on maintaining positive social relations, being friendly, compassionate, and cooperative. Agreeable people tend to trust others and adapt to their needs. Disagreeable people are more focused on themselves, less likely to compromise, and may be less gullible. They also tend to be less bound by social expectations and conventions and more assertive.

Emotional Stability. "Emotional Stability" (reversely referred to as neuroticism) measures the tendency to experience mood swings and emotions such as guilt, anger, anxiety, and depression. Emotionally unstable (neurotic) people are more likely to experience stress and nervousness, whereas emotionally stable people (low neuroticism) tend to be calmer and self-confident.

Intelligence. Intelligence ($n = 1,350$) was measured using Raven's Standard Progressive Matrices (SPM) (3), a multiple choice nonverbal intelligence test drawing on Spearman's theory of general ability. SPM is a proven standard intelligence test used in both research and clinical settings, as well as in high-stake contexts such as in military personnel selection and court cases (4). The SPM test was shortened for the purpose of this study and contained 20 items only. Note that SPM was used only to compare between users of this study, and no comparisons with the general population were made.

Satisfaction with Life. "Satisfaction with Life" (SWL) ($n = 2,340$) was measured using the SWL Scale (5), a widely used, five-item instrument designed to measure global cognitive judgments of satisfaction with one's own life.

Facebook Likes and User-Like Matrix. Facebook Likes allow Facebook users to connect with virtually any object that has an online

presence. Likes are one of the most typical and pervasive forms of digitally recorded behavior. People can Like quotes, Web sites, press articles, products, activities, places they visit (or would like to visit), and content such as pictures, movies, books, and music. Likes span a diverse set of entities, from "Bible" and "Philosophy" through "Bonfires," "BMW," and "cnn.com" to statements such as "I hate myself." People's Likes are shared with their friends and can be used as a way of expressing support, bookmarking, or enhancing online identity, by indicating individual preferences.

We recorded more than 9 million unique objects liked by users, a great majority of which were associated with one or very few users only. For the purpose of building a predictive model, Likes associated with fewer than 20 users, as well as users with fewer than two Likes, were removed from the sample. The remaining 58,466 users and 55,814 unique liked objects were arranged in a sparse matrix (user–Like matrix), the columns of which represent Likes and the rows of which represent users. The entries were set to 1 if there existed an association between a user and a Like and 0 otherwise. The matrix contained roughly 10 million associations between users and Likes. To facilitate the predictive analysis, the dimensionality of the user–Like matrix was reduced using singular-value decomposition (SVD) (6) such that each user is represented by a vector of k component scores. SVD provides a low-rank approximation to the original matrix, and the approximation quality increases with the number k.

Choosing the Number of the SVD Components. To choose the optimum number of SVD components to be used in this study, we examined the cross-validated prediction accuracy as a function of the number of components. Fig. S3, based on Openness and Extraversion, shows that prediction accuracy increases steeply in the beginning but flattens out relatively early (note that the horizontal axis is not linear). Interestingly, including some of the components abruptly increases prediction accuracy for certain traits. For example, including component 3 in the model increases the accuracy of Openness estimates from $r = 0.1$ to $r = 0.4$. Similarly, component 5 sharply increases the accuracy achieved in predicting Extraversion. This suggests that particular components are specifically related to a given attribute. We used the first $k = 100$ SVD components, which explained 28% of the variance in the user–Like matrix (Fig. S4). For sexual orientation, parents' relationship status, and drug consumption, only $k = 30$ top SVD components were used because of the smaller number of users for which this information was available.

Predictions. SVD components were used to build models that predict users' individual traits and preferences. Predictions related to numeric variables, such as age or intelligence, were calculated using a linear regression model based on the users' $k = 100$ SVD components as covariates. Dichotomous variables such as gender, relationship status, and political views were modeled using logistic regression based on the same SVD components. In both cases 10-fold cross-validation was used to assess the out-of-sample prediction accuracy: the sample was randomly split into 10 equally sized subsets of users, and predictions for each subset were calculated based on parameters determined on the remaining users. Prediction accuracy was measured in two ways. For the numeric variables, such as age in years, we report the Pearson product–moment correlation coefficient between the actual and predicted values across users. For the dichotomous variables such as gender, we report the area under the receiver-operating characteristic (ROC) curve (AUC) coefficient, which can be interpreted as the probability of correctly classifying two randomly selected objects: one of each class (e.g., male and female).

AUC. AUC relates to the ROC curve, which is a plot of true-positive rate (or sensitivity) versus false-positive rate (or 1 specificity) for detection or classification tasks. Positive cases are those classified by the model to belong to a target class (e.g., "male" or "Democrat"). Thus, true positive cases are the cases that were correctly classified by the model as belonging to a target class, whereas false-positive cases were classified incorrectly as belonging to a target class. The true-positive rate is the ratio of the number of true positives to the number of all cases in the target class, whereas the false-positive rate is the ratio of the number of false positives to the number of all cases in the background class. The logistic regression model used in this study to predict dichotomous outcomes assigns a probability of belonging to a target class to each of the users. To avoid having to select a single threshold for assigning users to a given target category, an ROC curve can be used to analyze the entire spectrum of possible thresholds. An example of an ROC curve is presented in Fig. S5. In general, ROC curves for random (or null) models should be close to diagonal, because the probability of seeing a true positive is not greater than the probability of seeing a false positive. The more an ROC curve bulges to the upper left, however, the higher the accuracy of the model, because higher true-positive rates are achieved for a given number of false positives. The AUC is simply the area below the ROC curve, and it is equal to the probability that a classifier will rank a randomly chosen positive instance higher than a randomly chosen negative one.

1. Costa PT, McCrae RR (1992) *Revised NEO Personality Inventory (NEO-PI-R) and NEO Five-Factor Inventory (NEO-FFI) Manual* (Psychological Assessment Resources, Odessa, FL).
2. Goldberg LR, et al. (2006) The international personality item pool and the future of public-domain personality measures. *J Res Pers* 40(1):84–96.
3. Raven JC (2000) The Raven's progressive matrices: change and stability over culture and time. *Cognit Psychol* 41(1):1–48.
4. Lubinski D (2004) Introduction to the special section on cognitive abilities: 100 years after Spearman's (1904) "'General intelligence,' objectively determined and measured". *J Pers Soc Psychol* 86(1):96–111.
5. Diener E, Emmons RA, Larsen RJ, Griffin S (1985) The satisfaction with life scale. *J Pers Assess* 49(1):71–75.
6. Golub GH, Kahan W (1965) Calculating the singular values and pseudo-inverse of a matrix. *J Soc Ind Appl Math* 2(2):205–224.

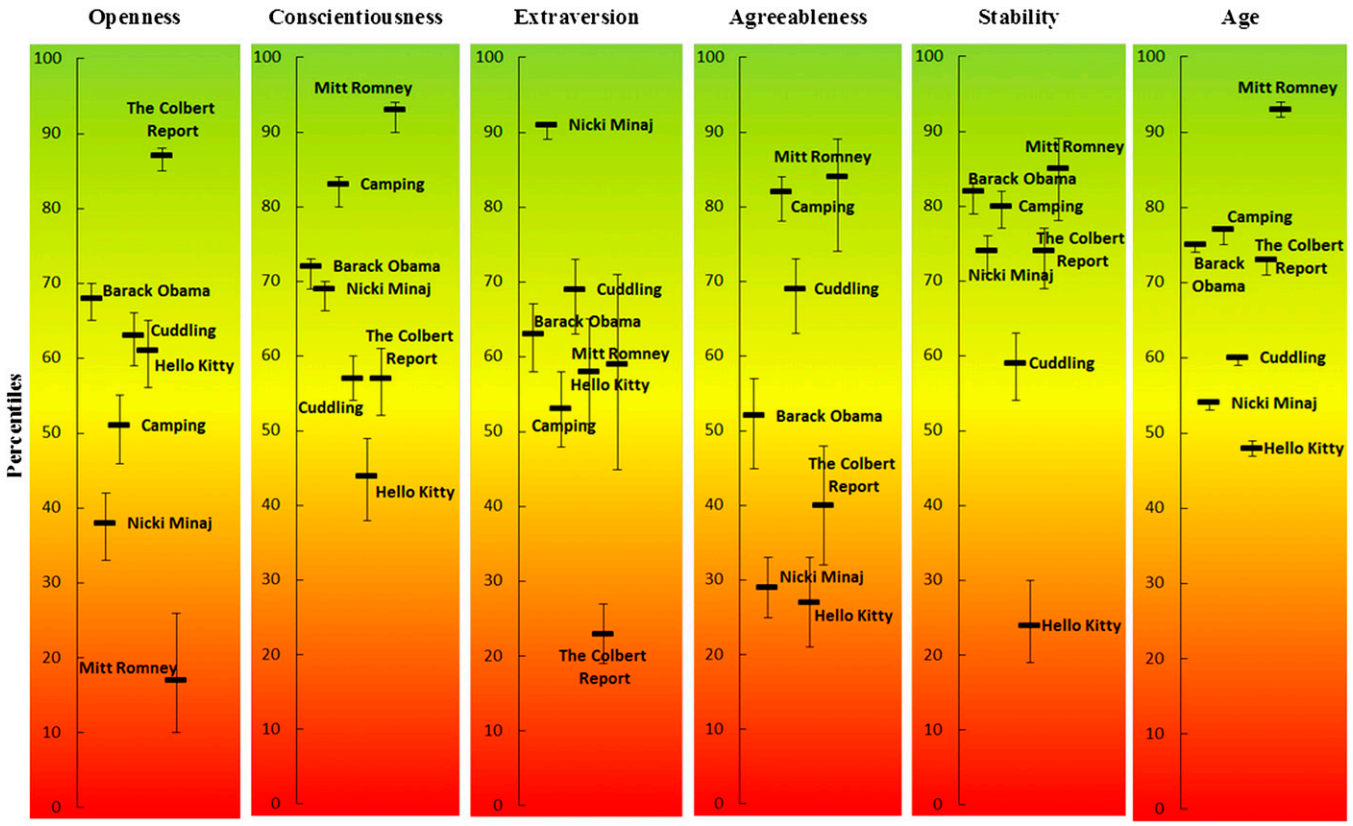

Fig. S1. Average levels of five personality traits and age of the users associated with selected Likes presented on the percentile scale. For example, the average extraversion of users associated with "The Colbert Report" was relatively low: it was lower only for 23% of other Likes in the sample. Error bars signify 95% confidence intervals of the mean.

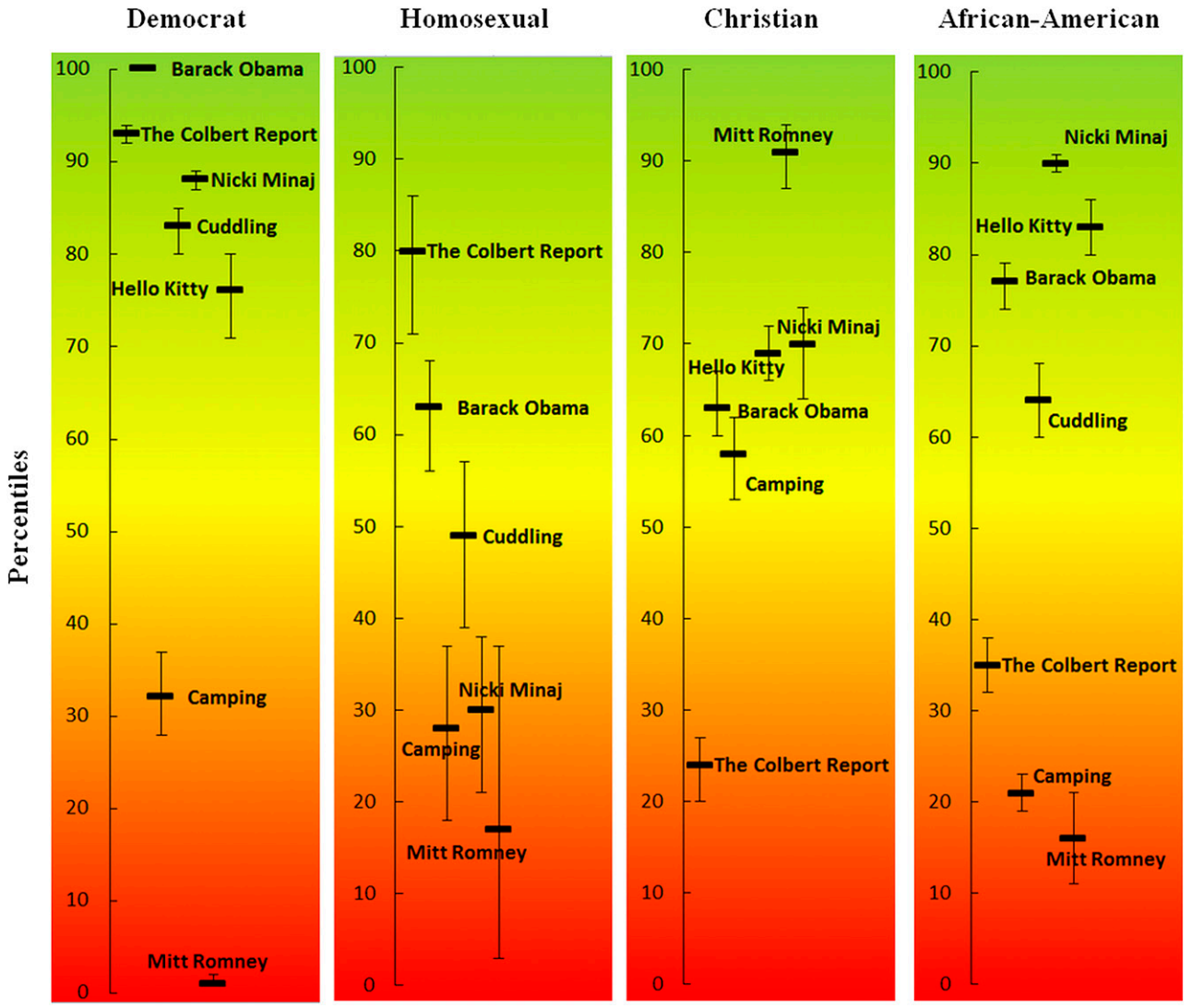

Fig. S2. Relative popularity of selected Likes within groups of Democrat, Homosexual, Christian, and African American users. Because Likes differed greatly in popularity (e.g., "Barack Obama" was nearly four times more popular than "Mitt Romney"), we calculated relative popularity by dividing the frequencies of associations with a given Like within the studied groups by the respective frequency in the entire sample. Relative popularity was transformed into a percentile scale. Error bars signify 95% confidence intervals of the population proportion. For example, The Colbert Report is relatively popular within Democrats and Homosexual groups (93th and 80th percentile respectively) but rather unpopular among Christians and African Americans (24th and 35th percentile, respectively).

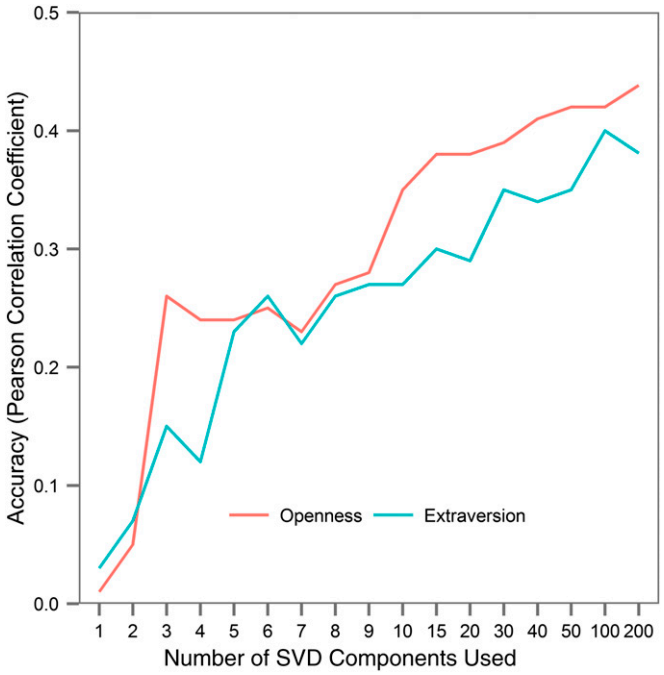

Fig. S3. Prediction accuracy in terms of Pearson correlation coefficient as a function of the number of SVD components used for the two personality traits Openness and Extraversion.

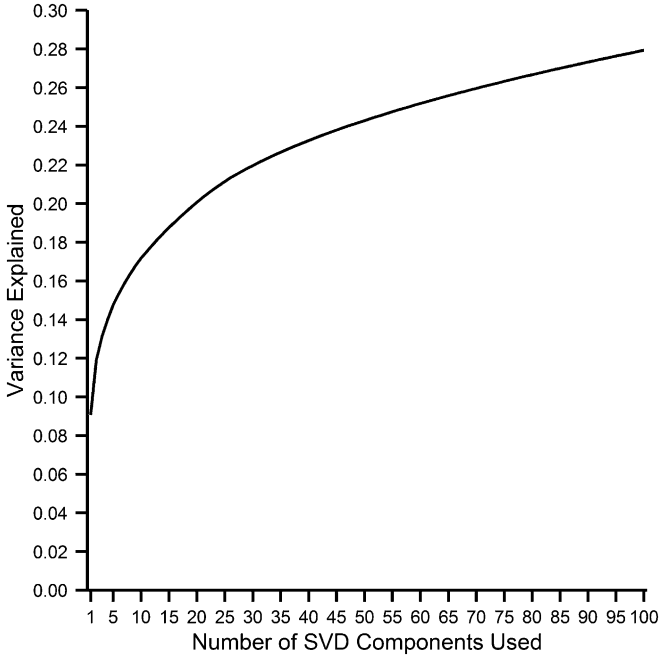

Fig. S4. Fraction of variance explained as a function of the number of SVD components for the dimensionality reduction of the user–Like matrix.

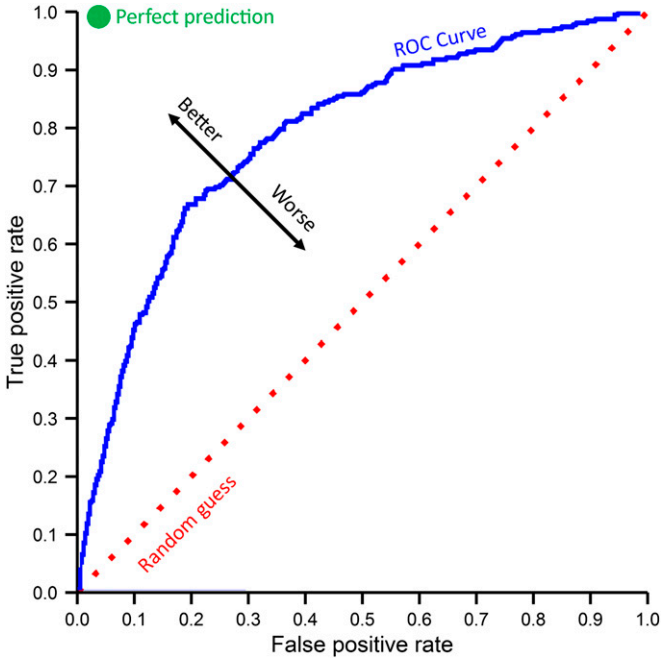

Fig. S5. Example of an ROC curve, detecting users associated with the Facebook Page associated with the FailBlog.org website. The AUC for this plot is 0.79.

Other Supporting Information Files

Table S1 (PDF)