Predicting Personality Using Novel Mobile Phone-Based Metrics

Yves-Alexandre de Montjoye[1,⋆], Jordi Quoidbach[2,*], Florent Robic[3,*], and Alex (Sandy) Pentland[1]

[1] Massachusetts Institute of Technology - The Media Laboratory, Cambridge, MA
[2] Harvard University - Department of Psychology, Cambridge, MA
[3] Ecole Normale Supérieure de Lyon, Lyon, France

Abstract. The present study provides the first evidence that personality can be reliably predicted from standard mobile phone logs. Using a set of novel psychology-informed indicators that can be computed from data available to all carriers, we were able to predict users' personality with a mean accuracy across traits of 42% better than random, reaching up to 61% accuracy on a three-class problem. Given the fast growing number of mobile phone subscription and availability of phone logs to researchers, our new personality indicators open the door to exciting avenues for future research in social sciences. They potentially enable cost-effective, questionnaire-free investigation of personality-related questions at a scale never seen before.

Keywords: Personality prediction, Big Data, Big Five Personality prediction, Carrier's log, CDR.

1 Introduction

How much can one know about your personality just by looking at the way you use your phone? Determining the personality of a mobile phone user simply through standard carriers' log has became a topic of tremendous interest. Mobile cellular subscriptions have hit 6 billion throughout the world [1] and carriers have increasingly made available phone logs to researchers [2] as well as to their commercial partners [3]. If predicted correctly, mobile phones datasets could thus provide a valuable unobtrusive and cost-effective alternative to survey-based measures of personality. For example, marketing and phone companies might seek to access dispositional information about their customers to design customized offers and advertisements [4]. Appraising users dispositions through automatically collected data could also benefit the field of human-computer interface where personality has become an important factor [5]. Finally, finding ways to extract personality and, more broadly, psycho-social variables from country-scale datasets might lead to unprecedented discoveries in social sciences.

The idea of predicting people's personalities from their cellphone stems from recent advances in data collection, machine learning, and computational social

⋆ These authors contributed equally to this work.

A.M. Greenberg, W.G. Kennedy, and N.D. Bos (Eds.): SBP 2013, LNCS 7812, pp. 48–55, 2013.

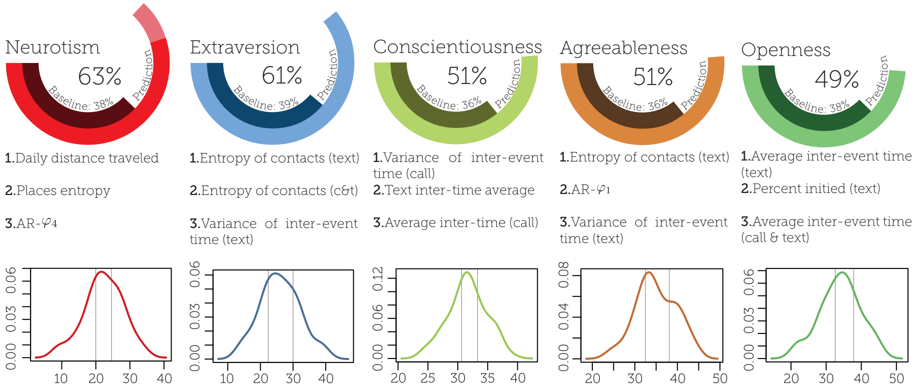

Fig. 1. (A) Accuracy of the prediction with respect to the baseline, **(B)** most useful features to predict personality traits, and **(C)** the distribution of personality traits across our dataset

science showing that it is possible to infer various psychological states and traits from the way people use everyday digital technologies. For example, some researchers have shown that pattern in the use of social media such as Facebook or Twitter can be used to predict users' personalities [6,7,8]. Others have used information about people's usage of various mobile phone applications (e.g., YouTube, Internet, Calendar, Games, etc.) or social network to draw inferences about phone owners' mood and personality traits [9,10,11,12,13]. Although these approaches are interesting, they either require to have access to extensive information about people's entire social network or people to install a specific tracking application on their phone. These constraints greatly undermine the use of such classification methods for large-scale investigations.

The goal of the present research is to show that users' personalities can be reliably inferred from basic information accessible from *all* mobile phones and to *all* service providers. Specifically, we introduce five sets of psychology-informed metrics–Basic phone use, Active user behaviors, Mobility, Regularity, and Diversity–that can be easily extracted from standard phone logs to predict how extroverted, agreeable, conscientious, open to experience, and emotionally stable a user is.

2 Results

Table 1 displays the different indicators and their respective contribution in predicting the big 5. Specifically, 36 out of our indicators were significantly related to personality and were all included in the final SVM classifier. As depicted in Figure 1, the model predicted whether phone users were low, average, or high in neuroticism, extraversion, conscientiousness, agreeableness, and openness with an accuracy of 54%, 61%, 51%, 51%, and 49%, respectively. The baselines being between 36 and 39%, we predict on average 42% better than random. For neuroticism, the predictive power of the model was further increased by including participants' gender as a predictor, increasing the accuracy to 63%. This finding is not surprising given that neuroticism is one of the traits that is most strongly associated with gender, with women having higher means levels than men in most countries world-wide [14].

An investigation of the most important feature to predict each trait revealed interesting associations. Indicators linked to users' mobility (i.e., distance traveled and entropy of places) were useful to predict Neuroticism. The entropy of participants' contacts helped predict both Extraversion and Agreeableness. These findings are inline with past research showing these traits both relate to different aspects of the diversity of one's social network: extraverts tend to seek more friends than introverts, agreeable individuals tend to be selected more as friends by other people [15]. Highly consistent with past research showing that conscientious individuals tend to like organization, precision, and punctuality [16], we found that the best predictor of Conscientiousness was the variance of the time between phone calls. Lastly, the strongest predictor of Openness was the average time between text interactions–a finding that remains be explained be future research.

3 Methodology

3.1 Participants and Procedure

The empirical sections of this work are based on a dataset collected from March 2010 to June 2011 in a major US research university [17]. Each participant was equipped with a Android smartphone running the open sensing framework *Funf* [18]. While the framework is designed to collect a wide range of behavioral data from the user's phone, we voluntarily limit ourself to data available in standard carriers's logs such as phone calls, text messages sent and received, etc. These CDR (Call Data Record) have recently become widely use for computational social science research [2,19,20,21,22]. After removing participants who had less than 300 call or text per year and/or that failed to complete personality measures, our final sample was composed of 69 participants (51% male, Mean age = 30.4, S.D. = 6.1, 1 missing value).

3.2 Metrics

We developed a range of novel indicators allowing us to predict users' personality. To build our list of indicators, we examined theories and research in personality psychology and, more specifically, the literature five factor model of personality, the dominant paradigm in personality research [23]. The five-factor model is a hierarchical organization of personality traits in terms of five basic dimensions: Extraversion (i.e, the tendency to seek stimulation in the company of others, to be outgoing and energetic), Agreeableness (i.e, the tendency to be warm, compassionate, and cooperative), Conscientiousness (i.e., the tendency to show self-discipline, be organized, and aim for achievement), Neuroticism (i.e, the tendency to experience unpleasant emotions easily), and Openness (i.e, the tendency to be intellectually curious, creative, and open to feelings).

From this literature review, we generated novel indicators that can be easily computed from carriers logs and that we believed would meaningfully account for potential differences in personality (see Table 1). These indicators fall under 5 broad categories: Basic phone use (e.g., number of calls, number of texts), Active user behaviors (e.g., number of call initiated, time to answer a text), Location (radius of gyration, number of places from which calls have been made), Regularity, (e.g.,temporal calling routine, call and text inter-time), and Diversity (call entropy, number of interactions by number of contacts ratio). These indicators are detailed hereafter.

Entropy: Is a quantitative measure reflecting how many different categories there are in a given random variable, and simultaneously takes into account how evenly the basic units are distributed among those categories. For example, the entropy of one's contacts is the ratio between one's total number of contacts and the relative frequency at which one interacts with them. $H(a{-}c) = -\sum_c f_c \log f_c$ where c is a contact and f_c the frequency at which a communicates with c. The more one interacts equally often with a large number of contacts the higher the entropy will be. This work considers the entropy of calls, text, calls+text but also the entropy of places one visits.

Inter-event Time: Is the time elapsed between two events. This work then consider both the average and variance of the inter-event time of ones' call, text, call+text. call+text means that an interaction, a call or an text, happened between two users. Therefore, even though two users have the same inter-event time for both call and text, their mean inter-event times for call+text can be very different.

AR Coefficients: We can convert the list of all calls and texts made by a user into a time-series. We discretized time by steps of 6 hours. For example, the time-series X_t contain the number of calls made by a user between 6pm and 12am on Monday followed by the number of calls made by the same user between 12am and 6am on Tuesday and so on. We then train a *auto-regressive* model per user. This model takes the form $X_t = c + \sum_{i=1}^{p} \varphi_i X_{t-i} + \varepsilon_t$ where c is a constant and ε_t are noise terms. The coefficients φ_i can thus be interpreted as the extent to which knowing how many calls a person made in the previous 6 hours, the day before at the same time predicts the number of calls that person will make in the coming 6 hours. We only kept the coefficient that were statistically significant for at least 3 traits: $\varphi_{1,4,8,12,18,24}$. Note that while we see some patterns in the statistically significant coefficients, interpretation of such patterns requires caution given that (1) this analysis has been done post-hoc and (2) our relatively small sample size.

Response Rate and Latency (Text): We consider a text from a user (A) to be a response to a text received from another user (B) if it is sent within an hour after user A received the last text from user B. The response rate is the percentage of texts people respond to. The latency is the median time it takes people to answer a text. Note than by definition, latency will be less or equal to one hour.

Number of Places and Their Entropy: The dataset was collected using the open sensing framework *Funf* which prevent us from directly using cell phone towers. We instead empirically defined places by grouping together the GPS points of a user that are less than 50m apart and by defining their center of mass as the lat-long coordinate of the place. 50m made sense given the sampling resolution of our dataset. Finally, we only kept the places where a user spend more than 15 minutes in a row.

Radius of Gyration: This is the radius of the smallest circle that contains all the places a user have been to on a given day.

Distance Per Day: This is sum of the distance between the consecutive places a user has visited in a given day.

Home and Call Regularity: We look at regularity at which a user is coming back home (home regularity) or receiving/making a call (call regularity) using a neural coding inspired metric [24].

3.3 Personality

As part of a larger questionnaire, participant completed the Big Five Inventory (BFI-44 [25]), a 44-item self-report instrument scored one a 5-point Likert-type scale measuring the Big Five personality traits. The BFI-44 has been widely used in personality research and has been shown to have excellent psychometric properties [25]. As depicted in Figure 1, participants personality scores follow a normal distribution: Neuroticism ($A = 0.3012$, $p = 0.5698$), Openness ($A = 0.2592$, $p = 0.7042$), Extraversion ($A = 0.2884$, $p = 0.6074$), Conscientiousness ($A = 0.4380$, $p = 0.2869$), and Agreeableness ($A = 0.4882$, $p = 0.2162$).

3.4 Class Prediction

Because the relationship between personality traits and numerous behavioral and psychological factors can often be non-linear [26,27], we choose to use SVM over the more traditional GLM as the former automatically model non linear relationships. Consequently, following [28] we classified each user as low, average, or high on each on the five personality dimensions.

We then selected the most relevant features using a greedy method similar to [29]. At each iteration, features are ranked using the squared weight and the worst feature of the set is removed. We stop removing features when removing a subset of worst features of size less than 3 degrades the performance and report the 3 highest ranked features. We then classified using an SVM with a 10-fold cross validation.

4 Discussion

The present study provides the first evidence that personality can be predicted from standard carriers' mobile phone logs. Using a set of novel indicators that we developed based on personality research and that are available to virtually anyone, we were able to predict whether users were low, average or high on each of the big five from 29% to 56% better than random. These levels of accuracy were obtained while we purposefully adopted a restrictive approach only using phone logs.

To our knowledge, these predictions exceed all previous research linking psychological outcomes to mobile phone use. In particular, a previous study that used a combination of information from mobile phone logs *and* people's usage of mobile phone applications such as YouTube, Internet, and Games predicted the personality of their owners with a mean accuracy of 15% [9]. In comparison, the mean accuracy in the present research is almost three times as high (i.e., 42%).

It is interesting to note that Extraversion and Neuroticism were the traits that were best predicted in our study. These two traits

Table 1. Metrics

	N	E	O	C	A
Regularity					
Average inter-event time (call)	●	●	●	●	●
Average inter-event time (text)			●	●	
Average inter-event time (c&t)		●	●	●	
Variance of inter-event time (call)		●		●	
Variance of inter-event time (text)		●		●	●
Variance of inter-event time (c&t)	●	●		●	●
Home regularity		●			●
AR- φ_1			●		
AR- φ_4	●				●
AR- φ_8	●	●	●		
AR- φ_{12}				●	●
AR- φ_{24}		●	●	●	
Number of call regularity				●	●
Diversity					
Entropy of contacts (call)			●	●	●
Entropy of contacts (text)	●	●	●	●	●
Entropy of contacts (c&t)		●			
Contacts to interactions ratio (call)	●	●	●		●
Contacts to interactions ratio (text)	●	●		●	
Contacts to interactions ratio (c&t)	●	●	●		
Number of contacts (call)		●	●		
Number of contacts (text)		●		●	
Number of contacts (c&t)		●	●		
Spatial behavior					
Radius of gyration (daily)	●	●			
Distance traveled (daily)	●	●	●	●	●
Number of places	●	●	●	●	●
Entropy (places)	●	●	●	●	●
Active behavior					
Response rate (call)		●			
Response rate (text)	●		●	●	●
Response latency (text)			●		
Percent during the night (call)	●		●		●
Percent initiated (text)	●		●		
Percent initiated (call)	●	●	●	●	
Percent initiated (c&t)				●	
Basic Phone use					
Number of interactions (text)		●			
Number of interactions (call)	●	●	●		●
Number of interactions (c&t)	●	●		●	●

are the dimensions of personality that are the most directly associated with emotion. In particular, extraversion is a strong predictor of positive emotions and neuroticism is a strong predictor of negative emotion [30]. This raises the hypothesis that our indicators might be picking up on the emotional components associated with these two traits. It would be interesting to investigate whether our indicators can predict emotional variable such as happiness in future studies. In addition, contrasting cellphone-based vs. questionnaire-based measures of personality when predicting various psycho-social outcomes might lead to interesting asymmetries. In line with this idea, recent research in personality shows that ratings of one's personality that are made by oneself and ratings of one's personality that are made by others are both valid but different predictors of behavior. For example, self-ratings predict behaviors like arguing or remaining calmn, whereas other-ratings predict behaviors like humor and socializing [31].

Although more research is needed to validate our model and the robustness of our indicators for use on a large-scale and more diverse population, we believe that our findings open the door to exciting avenues of research in social sciences. Our personality indicators and the ability to predict personality using readily available mobile phone data may enable cost-effective, questionnaire-free investigation of personality-related questions at the scale of entire countries.

Acknowledgments. The authors would like to thanks Nadav Aharony, Wei Pan, Cody Sumter, and Bruno Lepri for sharing data.

References

1. CNET, 2011 ends with almost 6 billion mobile phone subscriptions, http://news.cnet.com/8301-1023_3-57352095-93/2011-ends-with-almost-6-billion-mobile-phone-subscriptions/
2. de Montjoye, Y.-A., Hidalgo, C., Verleysen, M., Blondel, V.: Unique in the Crowd: The privacy bounds of human mobility. Nature Sci. Rep. (2013)
3. CNN, Your phone company is selling your personal data, http://money.cnn.com/2011/11/01/technology/verizon_att_sprint_tmobile_privacy/index.htm
4. de Oliveira, R., et al.: Towards a psychographic user model from mobile phone usage. In: Proceedings of the 2011 Annual Conference Extended Abstracts on Human Factors in Computing Systems. ACM (2011)
5. Arteaga, S.M., Kudeki, M., Woodworth, A.: Combating obesity trends in teenagers through persuasive mobile technology. ACM SIGACCESS Accessibility and Computing 94, 17–25 (2009)
6. Back, M.D., et al.: Facebook profiles reflect actual personality, not self-idealization. Psychological Science 21(3), 372–374 (2010)
7. Counts, S., Stecher, K.: Self-presentation of personality during online profile creation. In: Proc. AAAI Conf. on Weblogs and Social Media (ICWSM) (2009)
8. Stecher, K., Counts, S.: Spontaneous inference of personality traits and effects on memory for online profiles. In: Proc. Int. AAAI Conference on Weblogs and Social Media (ICWSM) (2008)
9. Chittaranjan, G., Blom, J., Gatica-Perez, D.: Mining large-scale smartphone data for personality studies. In: Personal and Ubiquitous Computing (2012)

10. Do, T.M.T., Gatica-Perez, D.: By their apps you shall understand them: mining large-scale patterns of mobile phone usage. In: Proceedings of the 9th International Conference on Mobile and Ubiquitous Multimedia. ACM (2010)
11. Verkasalo, H., et al.: Analysis of users and non-users of smartphone applications. Telematics and Informatics 27(3), 242–255 (2010)
12. Staiano, J., et al.: Friends dont Lie–Inferring Personality Traits from Social Network Structure (2012)
13. Pianesi, F., et al.: Multimodal recognition of personality traits in social interactions. In: Proceedings of the 10th International Conference on Multimodal Interfaces. ACM (2008)
14. Lynn, R., Martin, T.: Gender differences in extraversion, neuroticism, and psychoticism in 37 nations. J. Soc. Psychol. 137(3), 369–373 (1997)
15. Selfhout, M., et al.: Emerging late adolescent friendship networks and Big Five personality traits: A social network approach. J. Pers. 78(2), 509–538 (2010)
16. MacCann, C., Duckworth, A.L., Roberts, R.D.: Empirical identification of the major facets of conscientiousness. Learning and Individual Differences 19(4), 451–458 (2009)
17. MIT Human Dynamics Lab, Reality Commons, `http://realitycommons.media.mit.edu/`
18. Aharony, N., et al.: Social fMRI: Investigating and shaping social mechanisms in the real world. In: Pervasive and Mobile Computing (2011)
19. Onnela, J.P., et al.: Structure and tie strengths in mobile communication networks. Proc. Natl. Acad. Sci. U S A 104, 7332–7336 (2007)
20. Meloni, S., et al.: Modeling human mobility responses to the large-scale spreading of infectious diseases. Nature Scientific Reports 1 (2011)
21. Balcan, D., et al.: Multiscale mobility networks and the spatial spreading of infectious diseases. Proc. Natl. Acad. Sci. USA 106, 21484–21489 (2009)
22. Gonzalez, M., Hidalgo, C., Barabasi, A.: Understanding individual human mobility patterns. Nature 453, 779–782 (2008)
23. McCrae, R.R., John, O.P.: An introduction to the fivefactor model and its applications. Journal of personality 60(2), 175–215 (1992)
24. Williams, M.J., Whitaker, R.M., Allen, S.M.: Measuring Individual Regularity in Human Visiting Patterns. In: ASE International Conference on Social Computing (2012)
25. John, O.P., Srivastava, S.: The Big Five trait taxonomy: History, measurement, and theoretical perspectives. In: Handbook of personality: Theory and Research 2, pp. 102–138 (1999)
26. Benson, M.J., Campbell, J.P.: To be, or not to be, linear: An expanded representation of personality and its relationship to leadership performance. Int. J. Select. Asses. 15(2), 232–249 (2007)
27. Cucina, J.M., Vasilopoulos, N.L.: Nonlinear personality performance relationships and the spurious moderating effects of traitedness. J. Pers. 73(1), 227–260 (2004)
28. MacCallum, R.C., et al.: On the practice of dichotomization of quantitative variables. Psychol. methods 7(1), 19 (2002)
29. Guyon, I., Weston, J., Barnhill, S., Vapnik, V.: Gene selection for cancer classification using support vector machines. Mach. Learn. 46, 389–422 (2002)
30. Gomez, A., Gomez, R.: Personality traits of the behavioural approach and inhibition systems: Associations with processing of emotional stimuli. Pers. Indiv. Differ. 32(8), 1299–1316 (2002)
31. Vazire, S.: Who knows what about a person? The self-other knowledge asymmetry (SOKA) model. J. Pers. Soc. Psychol. 98(2), 281 (2010)