Accurately Inferring Personality Traits from the Use of Mobile Technology

Aline Carneiro Viana, Adriano Di Luzio, Katia Jaffrès-Runser, Alessandro Mei, Julinda Stefa

Accurately Inferring Personality Traits from the Use of Mobile Technology

Aline Carneiro Viana*, Adriano Di Luzio†, Katia Jaffres-Runser‡, Alessandro Mei†, and Julinda Stefa†
* INFINE Research Team, INRIA Saclay, Île-de-France, France, email: aline.viana@inria.fr
† Computer Science Department, Sapienza University of Rome, Italy, email: {diluzio, mei, stefa}@di.uniroma1.it
‡ IRIT Laboratory, Université de Toulouse, INPT-ENSEEIHT, Toulouse, France, email: kjr@enseeiht.fr

***Abstract*—This paper shows that human personality can be accurately predicted by looking at the data generated by our smartphones. GPS location, calls, battery usage and charging, networking context like bluetooth devices and WiFi access points in proximity, and more give enough information about individual habits, reactions, and idiosyncrasies to make it possible to infer the psychological traits of the user. We demonstrate this by using machine learning techniques on a dataset of 55 volunteers who took a psychological test and allowed continuous collection of data from their smartphones for a time span of up to three years. Openness, Conscientiousness, Extroversion, Agreeableness, and Neuroticism (the so called Big5 personality traits) can be predicted with good accuracy even by using just a handful of features. The possible applications of our findings go from network optimization, to personal advertising, and to the detection of mental instability and social hardship in cities and neighborhoods. We also discuss the ethical concerns of our work, its privacy implications, and ways to tradeoff privacy and benefits.**

I. Introduction

The main factor in human behavior is the individual's personality. In psychology, one of the most commonly used personality model is the Big5 [1], based on five crucial traits: Openness, Conscientiousness, Extroversion, Agreeableness, and Neuroticism. They are relatively stable over time, differ across individuals, and, most importantly, guide our emotions and our reactions to life circumstances. It is so for social and work situations, and even for things as simple as the way we use our smartphone. A person that is curious and open to new experiences will tend to look continuously for new places to visit and thrills to experience. A more conscientious one will instead focus on fewer but constantly fulfilling friends and environments. A neurotic person would constantly check the phone for a response after sending a text instead of using the time in a more productive way.

In this work we investigate on the degree of correlation between the Big5 personality traits and the use of mobile technology. We do so by leveraging a dataset of 55 volunteers who shared their personality traits and allowed that all of the activity on their phone be constantly recorded for research purposes for 3 years. The dataset includes, among other information, also data on the GPS location, battery status, bluetooth devices in the surroundings, connectivity information about WiFi access points, and more. We use this dataset to engineer a set of features that capture three aspects of human behavior: Temporal Mobility (e.g. when the person is at home/work or commuting), Spatial Mobility (e.g. number of most frequent places, maximum distance from home), and the Context of Use (battery charging habits of the user, access points the user constantly sees in her surroundings, and so on). Then, we use the most correlated features to predict the personality of a test-set portion of our dataset through cross validation. More in details, the contribution is the following:

- We extract, through statistical analysis and machine learning techniques, a set of 346 features describing the behavior of the individuals in terms of maximum distance from home, contextual information, and daily habits, and extract the ones that correlate the most to the personality traits of the individuals in the dataset.
- Guided by the fact that certain features are encoded in repeated sequences of numbers (like, for example, distance from home), we use spectral analysis to characterize the time series corresponding to individuals movements [2] (see Section IV).
- We apply Discrete Fourier Transform (DFT) to compute the spectral representation of finite-length sequences representing common and repetitive habits of individuals. This also allows us to capture individual behavioral characteristics that, for example, can describe the exploration tendency (*e.g.,* the tendency to visit frequently new places at certain periods of time) associated to specific traits of personality (see Section IV-A).
- The Temporal, Mobility, and Contextual features engineered are then selected through a greedy heuristic and used to predict the personality traits of the individuals through cross-validation.
- We compare our greedy selection and prediction results with those of Full-Knowledge Benchmark: a strategy that has full knowledge on the traits distribution on the individuals and draws from the corresponding distribution. The results show that the features that we engineer always outperforms the Full-Knowledge Benchmark. The difference in terms of F-score for all Big5 traits is remarkable: 0.23 for Openness, 0.27 for Conscientiousness, 0.31 for Extroversion, 0.24 for Agreeableness, and 0.33 for Neuroticism (see Section V-C).
- We present a detailed discussion about privacy issues, ethical considerations, and investigate on the trade off between features, user sensitive data they encompass,

and their prediction potential in Section VI. In particular, we find out that with as few as 3 features, exhaustively selected among all possible subsets of 3, we can still reveal lot of information on an individual personality.

The applications of our methodology to predict the personality traits from the use of mobile devices are certainly multifaceted: From foreseeing mental health issues within a neighborhood or region, to building better network infrastructures or caching strategies according to people's personal interests; from personality-targeted advertising, to the prediction of individual success in a certain area, and so on. Finally, we believe that the work in this paper paves the way to a new methodology that can in the future be extended to other platforms like social networks, credit card usage circuits, smart watches, and technological devices in general.

II. Background and Related Works

The Big5 personality model was introduced in the 1970s by two groups of researchers [1]. It delineates five traits, initially thought to be orthogonal but later on actually observed to be mutually correlated in experiments done with large populations. The traits, often referred to with the OCEAN acronym, are as follows: **Openness (to experiences) (O)** is associated with intelligence, originality, creativity, and intellectual curiosity. **Conscientiousness (C)** describes self-control, planning, and organisational skills. **Extraversion (E)** accounts for assertiveness, positive emotions, and captures the amount of social stimuli that we look for. **Agreeableness (A)** describes empathy, compassion, and altruism. **Neuroticism (N)** is usually associated with the tendency of experiencing negative feelings, anxiety, mood swings, and emotional instability.

The strength of each trait is determined by a value within a spectrum, typically defined on a scale from 10 to 50. Higher scores correspond to higher levels of the trait. Individuals can obtain their own trait values through well-established questionnaires built for this purpose by psychologists. An example is the 50-item IPIP survey test [3] available online at goo.gl/asZUMJ. The traits of a given population are observed to follow a normal distribution.

The Big5 traits have been the focus of many recent works in the area of psychology. Judge *et al.* [4] investigate on their relationship with career success, denoting a positive correlation with Conscientiousness and Openness and a negative correlation with Neuroticism. Wille *et al.* [5] uncover the link between job instability and low Agreeableness. The authors in [6] show how academic performance correlates significantly with Agreeableness, Conscientiousness, and Openness. The work in [7] finds out that Extraversion has a positive influence on online self-disclosure, that in turn influences life satisfaction.

The Big5 have also attracted the interest of many researchers from the computer science community. Chittaranjan *et al.* lay the groundwork for the relation between human personality and smartphone usage [8]. They study whether the installed applications, the calls and texts logs, and the proximity of other bluetooth devices can lead to the prediction of the Big5 traits. De Montjoye *et al.* [9] follow their steps. They focus on a group of US researchers and use a Support Vector Machine classifier to predict the Big5 traits by using mobile phone-based metrics (*i.e.,* call and text logs, bluetooth proximity, and mobility). Among the OCEAN traits, Neuroticism is the only one that they predicted through any feature related to mobility. Mønsted *et al.* [10] investigate on whether the same set of features predicts the personalities of almost 650 students from the university of Copenhagen. Nonetheless, they find no significant-correlation between the features related to mobility and any of the traits. Staiano *et al.* [11] build a social network from the call logs and nearby bluetooth devices to represent the interactions between users. Then, they exploit it to predict the OCEAN traits. Alam *et al.* [12] take a first step at predicting them from the information that users disclose on online social networks. Chorley *et al.* [13] investigate on the connection between personality and location-based social networks (*i.e.,* Foursquare). They reveal how Openness is correlated with check-ins at popular and social venues, while Neuroticism is negatively correlated with the number of venues visited. Finally, Alessandretti *et al.* [14] also study human mobility, sociality, and individual personality. They investigate on the connection between people social networks, the locations that they consistently visit, and Extraversion.

The contributions that we bring with this work are substantially different and richer. First, we have engineered a number of human-related features that better capture and describe individual spatio-temporal habits in terms of spatial mobility, temporal mobility, and context. Moreover, we introduce the use of spectral analysis of mobility and other human habits as features that can be used to predict personality traits. We did our work by starting from the definition of Big5, the intuition of what it means in terms of usage of technology to have a certain value of a given trait. The features we design, therefore, have also in mind the peculiarities of the single traits. In addition, previous works predict the five traits at a 3-level granularity, e.g., low, average, and high in the Openness scale. In this work, instead, we are able to obtain accurate predictions at a 5-level granularity: Very low, low, average, high, and very high. Lastly, we also investigate on the privacy issues that this type of work can raise, on the prediction quality of just a handful of features, and we give empirically based insights on how users can benefit from systems that make use of this type of analysis yet not leaking too much information on their own life or whereabouts.

III. Dataset description

We evaluate the methodology of this work on a spatio-temporal personal dataset collected through an Android mobile phone application. The application collects the data related to the user's digital activities such as available network connectivity (i.e., Bluetooth, APs, cell towers), resources availability (i.e., battery and memory), and visited GPS locations (as detailed in Table I). These activities are logged every 5 minutes. Before installing the Android mobile phone application, the volunteers were asked to complete the 50-

Table I: Description of the dataset measurements by type, which also have an associated anonymized device ID and a timestamp. WGS84 stands for World Geodetic System 1984.

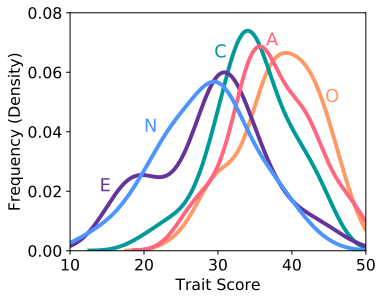
Measurement	Time Series	Description
Location	M	WGS84 Latitude and Longitude
Battery	B	Status (*e.g.* charging, full, discharging) and Level (in %)
WiFi AP	W	SSID, BSSID of the connected AP
APs Nearby	AP	SSID, BSSID of the nearby APs
Bluetooth	BT	MAC address of the nearby BT devices

item IPIP survey for the Big5 [3]. Overall, 99 volunteers completed the survey. Among them, we selected the 55 that appear with more than 500 measurements in the dataset. The resulting data spans 1055 days, from May 2015 to April 2018. The 55 volunteers are distributed among 6 different countries located in 2 different continents. In particular, 41 of them are from the same country. They are students and researchers in the same university. The other 14 come from a heterogeneous group of researchers and PhD students. Each user collected an average of $13,600$ measurements. The time span of the collection is significantly different between users. The actual sampling rate of measurements is also variable. The Android application requests, by default, a new measurement every 5 minutes. Our investigation on the distribution of the time between consecutive measurements per user reveals a sampling interval of 5 minutes for 99% of the measurements. The remaining 1% ($\approx 10,000$ pairs of measurements) has occasional longer gaps. This is due to system background jobs and settings of the specific operating system version. Sometimes the measurements are delayed—e.g., when the mobile device is in its deep sleep phase, or when the operating system kills the application service to free the volatile memory.

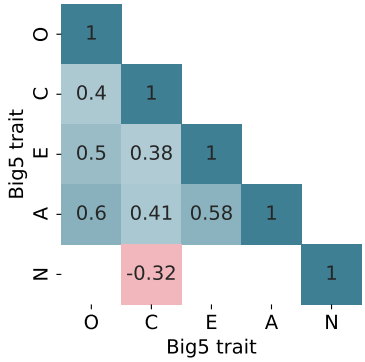
We established the reliability of the answers to the self-administered 50-item IPIP Big5 survey through the Cronbach's α test, a statistical tool designed to estimate a lowerbound on the reliability of a psychometric test. The result confirmed the reliability of the survey: All coefficients are greater than 0.7 [15] with Extroversion being the highest (0.8). In Figure 1, we show the Gaussian Kernel Density estimations of the distributions of the Big5 traits in the dataset, and the Pearson correlation matrix of the trait distributions. The graphs confirm the normal distribution observed before in larger populations and the non-orthogonality of the Big5 personality model [16]. In particular, A exhibits a strong positive correlation with O and E. Instead, N is the only trait that is modelled in a negative sense (i.e., it depicts emotional instability instead of emotional stability). As such it is the only trait that, in general, is known to show a negative correlation with the others. In our dataset, it does so with C.

IV. Feature Engineering

In this section we describe how, starting from the raw dataset of the mobile device usage, we design a carefully crafted set of features that we then use for the personality prediction.



(a) The Gaussian Kernel Density estimations of the distributions of the Big5 traits.



(b) Pearson's correlation matrix from the distributions of the Big5 traits.

Figure 1: Overview of the spatiotemporal personal dataset.

A. Time Series Modeling

We start off with the design of general time series that capture the peculiarities of the device usage. In particular, for each of the measurements shown in Table I we build a representative time series per user u as follows: Mobility M_u gives the geographical position of the user in time; B_u reports on the status and the percentage level of the device battery; W_u tracks the WiFi access points to which the user connects; AP_u tracks the WiFi access points in the surroundings other than those in W_u; BT_u reports on the names of the bluetooth networks available.

These general time series are further manipulated to extract time series that capture the peculiarities and routines of the users in their daily life. In doing so, we focus on three dimensions: Temporal Mobility (typical interval when the user is e.g. at work), Spatial Mobility (e.g. speed of movement, most frequented places), and Context (e.g. if the phone is charging, access points in the surroundings, etc.).

a) Temporal Mobility: This category encompasses the time series that report on *when* the individual is spending time at an important location, a location where she spends a considerable amount of time, or commuting between two important locations. First we focus on identifying the two most important ones: Work and home. Intuitively, they are the ones where a given individual goes often and spends a

Table II: A summary of the advanced time series that we engineer.

Category	Feature	Description
Temporal Mobility	HWC	Captures when the user is at work, at home, or commuting between home and work.
	$OutsideTown$	Captures when the user is outside her home town.
	$NightOutside$	Captures when the user is outside her home town, at night.
	$Abroad$	Captures when the user is in a different country.
	$Geohash^{1,2,3}$	Captures when the user is within the bounds of her top, second, and third most-visited Geohash cells.
	$GeohashNew$	Captures when the user is visiting a new (i.e., previously not visited) Geohash cell.
	$GeohashTen$	Captures the days when the user visited more than 10 different Geohash cells.
Spatial Mobility	$Distance$	Measures the distance traveled by the user between two consecutive measurements.
	$Speed$	Measures the speed at which the user traveled between two consecutive measurements.
	$Moving$	Captures if the user was on the move since the previous measurement and the current one.
	$Displacement$	Measures the distance between the user's home and her current position.
	RoG	Measures the daily radius of gyration of the user.
	$Geohash$	Encodes the Geohash cell of the user, in time.
	$Cluster$	Encodes the DBScan cluster identifier of the current location of the user.
Context	$BatteryStatus$	Keeps track of the status of the battery (e.g., charging, discharging, or full).
	$BatteryLevel$	Keeps track of the charge level of the battery.
	$Charging$	Captures when the user's phone is charging.
	$BurstCharging$	Captures when the user charges her phone for more than 4 hours.
	$HomeCharging$	Captures when the user charges her phone while she is at home.
	$NightCharging$	Captures when the user charges her phone at night.
	$FullCharging$	Captures when the user's phone is plugged in while full.
	$Connected$	Captures when the user is connected to a WiFi network.
	$NearbyAPs$	Counts the number of WiFi access points surrounding the user.
	$NearbyBTs$	Counts the number of Bluetooth devices surrounding the user.
	$Social^{d,b,c,r}$	Measures, day after day, the degree, betweenness centrality, closeness centrality, and page rank of the user in the social network built from WiFi access points.

lot of time [17]. To reach our goal, we first apply a rounding process to the raw GPS coordinates of the time series M_u that tracks the user location in time. We first express the geographic coordinates as decimal fractions; we then truncate them to the fourth decimal digit, reducing their precision to $11\,\mathrm{m}$ at the Earth equator. Then, for each user and for the business days of the weeks (i.e., from Monday to Friday), we determine the two geographical positions that result to be more common daily. We limit the check to two daily intervals: Between 2AM and 6AM (most probably spent at home sleeping) and between 11AM and 5PM (most probably spent at work/university). Intuitively, the resulting locations correspond to home and work. The intervals are chosen so to account for the fact that the participants come from countries with different habits regarding working hours. We also extract, for the same daily intervals, the 5 most common access points seen in the surroundings from AP_u. They are the top 5 BSSIDs of home and work places. Finally, the user is considered to be at home (work) if her distance from the determined home location is less than $100\,\mathrm{m}$ or she has probed any of 5 top BSSIDs for the determined home (work) location. The user is defined to be commuting when she moves from home to work and vice versa.

Then, we determine the geographical region and the country for each participant leveraging on the previously determined work and home locations and on the M_u series. We proceed with building the $NightOutside_u$ series (captures when one spends the night out), the $OutsideTown_u$ series (when outside hometown), and $Abroad_u$ (when out of the country).

In addition, we also analyze the mobility of the dataset users at a coarser-grained geographical resolution—that of a grid. For this, we use the $Geohash$ geocoding system (https://goo.gl/ekkFhH) that tessellates the space with a grid of side $151\,\mathrm{m}$. It assigns to each cell a unique identifier of 7 symbols—its $Geohash$. We then use the Geohash encoding and M_u to detect when an individual: visits for the first time a given region/cell (in the $GeohashNew_u$ series); visits one of her top 3 more frequented regions/cells (in the series $Geohash_u^1$, $Geohash_u^2$ and $Geohash_u^3$); visits more than 10 different $Geohash$ cells within a day (in the $GeohashTen_u$)– days when she moves considerably. The description of the Temporal mobility time series is included in Table II.

b) Spatial Mobility: The second category of advanced time series aims at characterizing the user movement habits in terms of speed, distance from home, and so on. More in details, from the general M_u time series we build the $Distance_u$ time series representing the distance traveled by the user and $Speed_u$ representing the movement speed. $Speed_u$ is used to understand whether the individual is standing still or mobile. Then we compute the $Displacement_u$ series as the distance between the user's current location and home. Intuitively, users that have higher displacement values are those that tend to travel more and more often. The RoG_u time series registers the daily radius of gyration of each user, i.e., the radius of the smallest circle containing all her mobility points.

Finally, we investigate on the density of the mobility traces of each user. To do so, we leverage the density-based spatial clustering algorithm DBScan [18]. For the algorithm tuning we use the Haversine formula to compute the distance between pairs of points and we require any two coordinates in the same

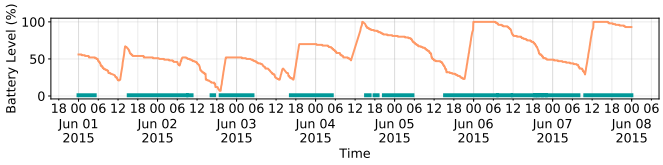


Figure 2: The $BatteryLevel_u$ time series, tracking the charge percentage level of the user's battery during the first 8 days of June 2018. The horizontal line at the bottom marks the instants of time when the user was at home. June 1st was a Monday, June 6 and 7 were respectively Saturday and Sunday.

cluster to be at most $1\,\mathrm{Km}$ apart (i.e., we set the ϵ hyper-parameter of the algorithm to 1). As a result, the algorithm assigns each mobility point in M_u to exactly one cluster. The $Cluster_u$ time series keeps note of the cluster identifiers of the user's positions at all times.

c) Context: This category of time series captures the context of the device usage (see Table II). The general time series B_u keeps track of the charging level and of the status of the battery as reported by the Android operating system (e.g., whether it was charging, discharging, plugged in but not charging, and plugged in while completely charged). Out of it we extract the series $BatteryStatus_u$ and $BatteryLevel_u$ that track separately the status and charge percentage level.

Then, we aim at capturing the user habits in charging the phone—how often, during the night or also the day, and so on. So, we take into consideration whether they usually plug their phones in at night ($NightCharging_u$), how long they keep them on charge while full ($FullCharging_u$), whether they charge them in short bursts of less than 4 consecutive hours ($BurstCharging_u$), and whether charging while at home ($HomeCharging_u$). We couple this information with the Temporal mobility time series HWC_u to detect whether the charge happens at home, work, or elsewhere. Figure 2 shows an example of the $BatteryLevel_u$ time series of an individual in the first 8 days of June 2015. The horizontal line at the bottom of the figure indicates whether the user was at home. From the figure we observe that the user consistently charged her phone at night. Furthermore, the phone was unplugged early in the morning on business days while left unplugged for longer and also while full on the weekends.

We also build series representative of the connectivity level of the individuals: $Connected_u$, $NearbyAPs_u$, and $NearbyBTs_u$ tracking respectively APs to which the user connects, nearby APs, and bluetooth devices. Finally, we leverage the connectivity of the users to model their social behaviours. We build a social network where nodes are the users in our dataset and add an edge each time a couple of users have been connected to the same WiFi access point within a time window of 10 minutes. Intuitively, the edge denotes that the two people have been at the same place roughly at the same time. The series $Social_u^{d,b,c,r}$ capture respectively the node degree, the betweenness centrality, the closeness centrality, and the vertex page rank average daily values in the social network.

B. Time Series Analysis

The advanced time series that we have built so far need to be transformed into prediction features. That is, a set of mathematical and numerical indicators that capture the behavior of the users in the dataset. We focus on the following: i) Studying the times series mean, standard deviation and coefficient of variation after sampling them at a time resolution that is more representative of the human routine than the 5-minute resolution of the measurement (*e.g.,* daily, or weekly); ii) analyzing them from an information theoretic point of view by measuring their entropy, diversity, and irregularity; iii) finally, building metrics that leverage a frequency domain analysis of their Fourier transform.

a) Statistical Analysis: Many of our life's patterns appear at regular time intervals. Some appear daily—the time at which we usually wake up, arrive or leave work. Other appear weekly—pre-week end gathering of friends at a pub. To capture these patterns we subdivide the series in equal-sized intervals of a given frequency (i.e., daily, weekly etc.) that intuitively corresponds to that of the targeted pattern. To avoid distortion or biases we discard weekends and national holidays. Then, we aggregate the different values within each sub-time series to a single measurement. For example, we count the total number of WiFi access points that the user probed around her in a week, compute her average daily speed or her average displacement from home, and so on. Then, we focus on the mean μ, the standard deviation σ and the variation coefficient $\gamma = \frac{\sigma}{\mu}$ of the coarser-grained time series.

b) Entropy, Repetitiveness, Irregularity, and Stationarity: Some of the time series that we engineer assume values from a categorical or discrete domain. Some examples are the $Geohash_u$, the $Cluster_u$, and the $BatteryStatus_u$ series. If properly analyzed, they can give insightful information on the individuals. For example, $Geohash$ tessellation can be used to measure the repetitiveness of visits or the entropy of a sequence of daily or weekly region visits of an individual. So, it can tell on whether the individual tends to explore new, unseen regions. We perform the same analysis on the region clusters given by the DBScan algorithm.

Additionally, we also quantify the irregularity of their $Geohash$ visits, of how they visit the cluster neighborhoods, and of their charging habits. We again extract coarser grained time series from a given one. Then, compute an aggregate value like the information entropy or the repetitiveness from each sub-series. Finally, we compute the average value μ, the standard deviation σ, and the coefficient of variation γ values from the overall distributions. Finally, we investigate whether the statistical descriptors of our time series change over time. That is, their average value, standard deviation, and autocorrelation are constant in time. To do so, we test their stationarity with the Dickey-Fuller test.

c) Frequency Domain Analysis: So far, our analysis of the time series focused on the domain of time. For example, we can know at what time every morning the user leaves home.

However, an important part of our behavior can be described by events that occur with a regular frequency. We leave home every 24 hours, we stay home longer in the morning every 7 days (Saturdays and Sundays), and may be we go to the gym every couple of days. The regularity of our behavior can be unraveled by translating the same data from the time domain to the frequency domain. In this way, we can clearly see how strong are the daily, weekly, and other patterns that occur with a particular frequency, even in a data series full of noise, since we can just look at the magnitude of the corresponding frequencies.

To do so, we use the Discrete Fourier Transform (DFT in short, with its fast implementation FFT) to divide the time series into their frequency components and to study their amplitude and their phase. The DFT requires the time series to be uniformly sampled in time. Our series, instead, are sampled with a non perfectly uniform frequency of roughly 5 minutes (sometimes there are gaps between pairs of consecutive measurements). Therefore, the preparation phase of the series for the DFT has to be done carefully: First, we re-sample the time series so that there is exactly one sample every t minutes. We divide the time in equal-sized intervals of t minutes each. Timeframes with multiple measurements are assigned the average value of its measurements. To fill the gaps, we take the mode (i.e., the most frequent value) of all the other measurements taken during the same day of the week and at the same time [19]. Additionally we derive an up-sampled version with a period of 60 minutes from the uniformly-sampled 5 minutes time series.

From the DFT of each uniformly-sampled time series we extract the frequency that carries the highest energy and the magnitudes at the daily and weekly frequencies. Then, we measure the periodicity of the transform defined as the percentage of overall energy that is accounted for by the three highest-energy frequencies. Finally, we build a pure sine wave with a period corresponding to the most energetic frequency. That is, we put all the other frequencies to zero and invert the DFT, and we measure its euclidean distance from the corresponding original time series, so to see how much of the series can be explained by the most energetic frequency.

Lastly, we also estimate an additional periodogram for each of our series by leveraging Welch's method [20]. To do so, we use a moving window size of 14 days, so that the repeated measurements on the series average out the noise caused by the finite number of measurements in our samples. As before, we analyze the resulting Welch's periodograms to find out the frequencies of highest energy, the periodicity, and the euclidean distance of the pure sine wave with the original time series.

C. Final feature set

Overall we have engineered 346 features from the time series of the dataset. Figure 3 sheds light on how correlated the features are to each trait. In particular, we plot number of features that have a significant Pearson's correlation (2-tailed p-value at the 0.05 significance level) and a significant linear regression model (p-value of the F-statistic less than 0.05) with the Big5 traits. It is worth pointing out that, for each trait, the set of significant features according to the Pearson's correlation is always a subset of the features having a significant regression model to the trait. Therefore, we focus our analysis on the most comprehensive sets obtained from the linear regression models.

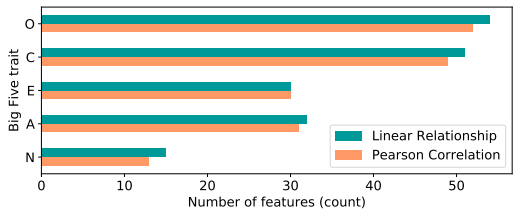


Figure 3: Number of features that have a statistically significant Pearson's correlation as well as linear regression model to the Big5 traits. The sets of features from Pearson's correlation are subsets of those from the linear regression.

Openness and Conscientiousness are the two traits correlated to the largest number of features (respectively 52 and 49). Agreeableness and Extraversion have a significant linear relationship, respectively, with 31 and 30 features. Neuroticism is only correlated with 15 features only. In the reminder of this section we report on the nature of the subsets having significant correlations when taken singularly to each of the traits.

Conscientiousness: Individuals high on C are well organized and highly reliable, and typically stick to routines [21]. According to our correlation analysis, C is the only trait that is characterized mostly by features related to *Spatial Mobility*. Our features having a significant relationship with the C trait capture these indicators: In particular, C is positively correlated with the repetitiveness of the visits to locations captured by the $Cluster$ time series. Additionally, it is negatively correlated with measures capturing dispersal and variability, like the standard deviation σ and the variation coefficient γ smartphone charging times or get-to-home-from-work times. C has also a positive correlation with the user's daily displacement and radius of gyration average values μ. The higher a user scores on C, the more likely she is to travel further from home and the bigger the radius of the circle enclosing the locations that she visits.

Openness: Individuals high on the Openness spectrum are known to be always looking for new thrills and experiences. We observe this to be true also in the type of features that have a relevant correlation with Openness. In fact, the O trait is mostly characterized by features from the *Temporal Mobility category*. In particular, O is negatively correlated with the daily average time individuals spend at home (they tend to be out more often). While, there is a positive correlation with the dispersal and variability (i.e. the σ and γ) of the home-time length of these home periods and with the daily displacement. This indicates irregularities in the routine of individuals high

on Openness. Finally, 3 of the features extracted from the DFT of the $GeohashNew$ time series are positively correlated to O. This analysis confirms the new-thrill searching nature of individuals high on the O spectrum.

Agreeableness: We find the A trait to be mostly correlated with features related to *Temporal Mobility*. In particular, individuals high on the A spectrum tend to arrive at work earlier in the morning. In addition, the time they spend at work and at home is significant and variable. In fact, the correlation with the average values μ and standard deviation values σ of work and home time are positive. These results are consistent with previous result suggesting that Agreeableness and Conscientiousness have a significant correlation with job performance [22].

Extraversion: The E trait is also mostly connected to features in the *Temporal Mobility* category. In particular, extroverted people tend to spend a significant amount of time at home and to have a low variability on the working time. Our analysis shows that they spend longer but consistent working periods. In addition, individuals scoring higher in the E trait have high dispersal in their interactions. That is, a high standard deviation σ of their betweenness in the social network built from WiFi access points. This sheds light on the high diversity of social relationship of E individuals.

Neuroticism: Lastly, N is the trait that is significantly correlated with less features overall, 15 only. It is mostly characterized by the *Context* category. This result suggests that it is harder to capture the emotional instability of an individual by looking at her routine and her mobility patterns. But, individuals high on the N spectrum are prone to social anxiety [23]. Therefore, they might tend to visit less social venues in average [13]. Our results confirm these findings. N is positively correlated with the variation coefficient γ of their betweenness in the social network built from WiFi access points. This shows that highly neurotic people tend to be unstable in their social relationships. N also results negatively correlated with the number of locations visited daily given by the $GeohashTen$ time series. Here, we noticed that N individuals seldom visit 10 or more locations in a day.

V. Predicting the Big5 traits

After investigating how movement and behavior of individuals correlate with traits of personality, this section details our evaluation on whether it is possible to predict human personality from their mobility and their routines.

More in details, to predict the user's Big5 traits (i.e., the dependent variables $y_i \in Y_t$ of user i and Big5 trait t) from her engineered features (i.e., the independent variables X_i of user i), we first standardize the features and the traits: We subtract the mean of their distributions and then we divide by their standard deviation. Then, we build a linear regression model from the ordinary least squares method to estimate the β coefficients of $y_i = X_i^T \beta + \varepsilon_i$, where ε is a mean-zero random error term and ^{T} denotes the transpose, such that the sum R of the squared residuals is minimum.

Creating a linear model with all the 346 features is not reasonable. Linear regression works best if variables (i.e., features in this case) are non-correlated. The more correlated features are fed to the model, the more the measurement errors are captured by linear regression. Thus, it is fundamental to work on a reduced and independent set of variables. For this, in the following, we first define an appropriate quality prediction methodology that we will leverage in the feature subset construction. Next, we describe the method we have engineered to create statistically significant subsets of features with high prediction power.

A. Evaluation methodology

To evaluate the performance of the linear models we proceed as follows. First, we (independently) predict the traits of each individual in our dataset by leave-one-out cross validation. More specifically, for each trait, we estimate the β parameters of n different models (for n corresponding to the number of users) and at each time, a single individual j is left out of the model fitting. Then, we use the estimated β and the X_j features of the excluded user j to predict the score $\hat{y_j}$ on that trait. As a result, we get a vector $\hat{Y_f}$ of n predicted scores, one for each user.

Next, we measure the performance of the model by comparing the original Y_f and the predicted vectors $\hat{Y_f}$ of scores. One way to evaluate the performance of a linear model is to look at its Mean Square Error or at its R^2 score, i.e., at the proportion of variance explained by the model. Nonetheless, these numerical scores do not have any immediate interpretation—a Mean Square Error of 0.3 does not provide much information on the quality of the predictions generated by the model. For these reasons, we choose to evaluate the performance of our models in a more meaningful way.

To do so, we first dichotomize the y and $\hat{y}$ vectors into 5 discrete categories corresponding to very low, low, average, high, and very high scores on that trait categories coming from the Revised NEO Personality Inventory [1] by Costa and McCrae. Then, we measure 5 F_1 scores, one for each category. To account for the unbalanced population in the categories, we compute the overall F_1 score of the prediction by using a weighted average, where the single F_1 scores are weighted by the size of the true population of their category (*i.e.,* their support).

B. Feature Selection

We describe hereafter the method we have used to select a reduced subset of features to be used in the personality prediction of the linear regression model. We test the F_1 score of subsets constructed with a greedy approach based on forward feature selection. For each trait, we do the following operations: We first select the "seed" of the greedy algorithm, i.e., those features described in Section IV-C and Figure 3 that produce a statistically significant linear model when evaluated singularly against that trait. Then, we iteratively build the greedy subset (initially empty) by moving a feature from the seed set to the greedy set at each iteration. The selected feature

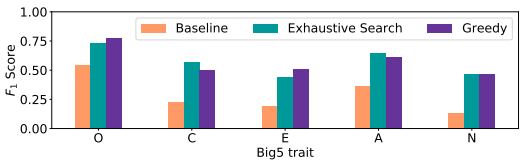


Figure 4: The F_1 scores of the best-performing subsets, by Big5 trait, for the greedy methodology and the exhaustive search of the best three features. In all cases the linear regression model between the subsets of features and the single Big5 trait is statistically significant (p-value of the F-statistic < 0.05).

is the one that, together with the others already in the greedy subset, maximises the weighted F_1 score of a new set of n linear models. In fact, when testing each candidate feature we use the leave-one-out cross validation approach described before. This serves two purposes: To be sure that there is no over-fitting in our feature selection process and that we do not introduce any bias in our model. After each iteration, we remove the selected feature from the seed, we add it to the greedy subset, and we move to the next iteration step. The process ends when the seed is empty or when the greedy set is composed of 25 features.

C. Experimental Evaluation

In the remainder of this section we detail the experimental evaluation of our prediction methodology against each Big5 trait, independently. More in details, Figure 4 compares the F_1 scores of our greedy methodology against their baseline—the weighted-average F_1 score of a classifier that simply predicts the most likely category for each trait—to measure their improvement. Generally speaking, our linear model performs significantly better than the baseline for all the traits: The baseline F_1 scores for the OCEAN traits are respectively 0.541, 0.229, 0.194, 0.364, and 0.131. The improvements of the greedy methodology to the baselines for the OCEAN traits are 0.23, 0.27, 0.31, 0.24, 0.33. In particular, our methodology finds two subset of 14 and 16 features that improve the prediction of Openness (with the highest F_1 score overall, 0.77) and Extraversion. The mean and the standard deviation of the absolute values of the pairwise Pearson's correlations of the features in these sets are respectively 0.26 and 0.20 for Openness, and 0.27 and 0.22 for Extraversion—*i.e.,* they are composed of independent and sufficiently uncorrelated features. In addition, both for Openness and for Extraversion the features selected by our approach come mostly from the Temporal Mobility category. These results match those of our correlation analysis of Section IV-C and confirm our intuition that Openness and Extraversion characterise the amounts and the distributions of the that we spend at our important locations (*e.g.,* home, work, etc.). The prediction also agree with the results of correlation analysis for Neuroticism.

VI. Ethical Considerations and Privacy Issues

This work brings up critical ethical issues. First of all, in the construction of the dataset volunteers have given full consent to the storage of data about their psychological traits, to the collection of personal networking data, and to the use of these data for research purposes in an anonymized form.

Clearly, systematic collection of personal data raises privacy concerns. Users might be willing to reveal some aspects of their personality to get better services, but they might have concerns about sharing with the service provider all the data about their temporal mobility, their spatial mobility, and the context they live in. To better understand whether it is possible to restrict the collection of data and get similar prediction results, we have run an experiment in which we limit to 3 the number of features that can be used in the prediction. This can be easily done by exhaustive search. The results, shown in Figure 4, give some interesting insights. First, for each personality trait three features are enough to get good F_1 scores (actually, sometimes even better than the greedy methodology). Second, and more importantly, most of the personality traits can be predicted by using only two of the three feature categories discussed in Section IV.

For example, according to our experiments, Openness and Extraversion can be predicted well without the use of features of the category of Spatial Mobility. In other words, O and E can be predicted without revealing the position of the user, but only when the user visits some special locations, that can be kept private. Indeed, Openness and Extraversion are significantly correlated, as shown in Figure 1b. Therefore, if a service needs to know just your openness and your extraversion to give you better quality of experience (for example, targeted advertising), then the service does not need to know your position. Conversely, Conscientiousness and Neuroticism can be predicted well without using features of the category of Temporal Mobility. It is indeed interesting to note that C is the only trait, albeit negatively, correlated to N. Lastly, features from the category of Context are in the best three features used to predict all of the 5 traits, which interestingly show the importance of how our personality influences the way we interact with the context in which we live. Agreeableness is the only trait that needs data from all of the feature categories to get the best results using only three features. Arguably, A has to do with the way we behave with the other humans, and it is indeed hard to predict (Figure 4 shows that, though the baseline for A is high, prediction results do not do better than .6 of F_1 score, whereas O can be predicted with 0.77 of F_1 score).

In order to protect personal data and still be able to get the benefits of services based on personality, the linear regression model described in this paper can be computed in a trusted environment in the smartphone. Then, only the traits that are actually needed for the service can be given to the service provider. This information can be sent in a very limited form like, for example, high or low Openness (higher or lower than the median of all users).

Another important contribution of this work is to give strong evidence that sharing networking data, for example about mobility, does not only reveal where you are during the day, it also gives very intimate information about your personality. Users should know that this is indeed possible and the choice of giving personal data to get better services, like, for example, location services, should be taken sensibly and with full awareness. Actually, our experiments show that even information about charging our smartphone gives important clues about four out of five traits (O, C, E, and N), and similarly does information about the time people go to work and come back home.

VII. Conclusions

The goal of this work was to shed light on the link between human personality and mobility, daily routine, and sociability. In particular, we aimed at characterizing and capturing, for the first time, the human-related aspects that make our lives unique: How we commute to and from work, the diversity of the time we spend at home, and whether we regularly charge our smartphones. By studying a set of 55 volunteers from around the world, we discovered that, in fact, our personality has a strong influence on the small details of our everyday life. For example, individuals high in Conscientiousness move in repetitive and predictable ways, and commute back home at the same time every day; someone high in Extraversion has, on average, a higher number of social interactions with the others; Open individuals seek new experiences to try and, as a result, travel further from home and visit, on average, a higher number of location. In addition, we also investigated on the predictability of human personality. More in details, we engineered a set of 346 features that we specifically designed to capture the human-related aspects of our mobility and of our routine. Then, we exploited these features to predict the Big5 traits of the volunteers in our dataset. As a result, we found out how it is possible to predict the human personality, with considerable accuracy, even when using a handful of features related to Temporal and Spatial Mobility and Context (3, in our case). We hope that the consequences of this work will be two-fold: On one side to contribute with a first, ground-breaking step, on understanding how to characterise from a mathematical, statistic, and information theoretical point of view the aspects that make our lives, our mobility, and our routines unique. On the other to deepen our understanding of human personality by evaluating and quantifying, experimentally, how it influences our lives, the decisions we make, and the actions we take.

References

[1] P. T. Costa and R. R. McCrae, "The revised NEO personality inventory (NEO-PI-r)," in *The SAGE Handbook of Personality Theory and Assessment: Volume 2 — Personality Measurement and Testing*. SAGE Publications Ltd, pp. 179–198.

[2] C. Song, Z. Qu, N. Blumm, and A.-L. Barabási, "Limits of predictability in human mobility," *Science*, vol. 327, no. 5968, pp. 1018–1021, 2010.

[3] L. R. Goldberg, "The development of markers for the Big-Five factor structure." *Psychological Assessment*, vol. 4, no. 1, pp. 26–42, 1992.

[4] T. A. Judge, C. A. Higgins, C. J. Thoresen, and M. R. Barrick, "The big five personality traits, general mental ability, and career success across the life span," *Personnel Psychology*, vol. 52, no. 3, pp. 621–652, sep 1999.

[5] B. Wille, F. D. Fruyt, and M. Feys, "Vocational interests and big five traits as predictors of job instability," *Journal of Vocational Behavior*, vol. 76, no. 3, pp. 547–558, jun 2010.

[6] A. E. Poropat, "A meta-analysis of the five-factor model of personality and academic performance." *Psychological Bulletin*, vol. 135, no. 2, pp. 322–338, 2009.

[7] S. S. Wang, ""I Share, Therefore I Am": Personality traits, life satisfaction, and facebook check-ins," *Cyberpsychology, Behavior, and Social Networking*, vol. 16, no. 12, pp. 870–877, dec 2013.

[8] G. Chittaranjan, J. Blom, and D. Gatica-Perez, "Mining large-scale smartphone data for personality studies," *Personal and Ubiquitous Computing*, vol. 17, no. 3, pp. 433–450, dec 2011.

[9] Y.-A. de Montjoye, J. Quoidbach, F. Robic, and A. Pentland, "Predicting personality using novel mobile phone-based metrics," in *Social Computing, Behavioral-Cultural Modeling and Prediction*. Springer Berlin Heidelberg, 2013, pp. 48–55.

[10] B. Mønsted, A. Mollgaard, and J. Mathiesen, "Phone-based metric as a predictor for basic personality traits," *Journal of Research in Personality*, vol. 74, pp. 16–22, jun 2018.

[11] J. Staiano, F. Pianesi, B. Lepri, N. Sebe, N. Aharony, and A. Pentland, "Friends don't lie," in *Proceedings of the 2012 ACM Conference on Ubiquitous Computing - UbiComp '12*. ACM Press, 2012.

[12] F. Alam, E. A. Stepanov, and G. Riccardi, "Personality traits recognition on social network-facebook," *WCPR (ICWSM-13), Cambridge, MA, USA*, 2013.

[13] M. J. Chorley, R. M. Whitaker, and S. M. Allen, "Personality and location-based social networks," *Computers in Human Behavior*, vol. 46, pp. 45–56, may 2015.

[14] L. Alessandretti, P. Sapiezynski, V. Sekara, S. Lehmann, and A. Baronchelli, "Evidence for a conserved quantity in human mobility," *Nature Human Behaviour*, jun 2018.

[15] P. Kline, *The Handbook of Psychological Testing*. Routledge, 1993.

[16] J. M. Digman, "Higher-order factors of the Big Five." *Journal of Personality and Social Psychology*, vol. 73, no. 6, pp. 1246–1256, 1997.

[17] S. Phithakkitnukoon, Z. Smoreda, and P. Olivier, "Socio-geography of human mobility: A study using longitudinal mobile phone data," *PLoS ONE*, vol. 7, no. 6, p. e39253, jun 2012.

[18] A. Ram, S. Jalal, A. S. Jalal, and M. Kumar, "A density based algorithm for discovering density varied clusters in large spatial databases," *International Journal of Computer Applications*, vol. 3, no. 6, pp. 1–4, jun 2010.

[19] E. M. R. Oliveira, A. C. Viana, C. Sarraute, J. Brea, and I. Alvarez-Hamelin, "On the regularity of human mobility," *Pervasive and Mobile Computing*, vol. 33, pp. 73–90, dec 2016.

[20] P. Welch, "The use of fast Fourier transform for the estimation of power spectra: A method based on time averaging over short, modified periodograms," *IEEE Transactions on Audio and Electroacoustics*, vol. 15, no. 2, pp. 70–73, jun 1967.

[21] J. J. Jackson, D. Wood, T. Bogg, K. E. Walton, P. D. Harms, and B. W. Roberts, "What do conscientious people do? development and validation of the behavioral indicators of conscientiousness (BIC)," *Journal of Research in Personality*, vol. 44, no. 4, pp. 501–511, aug 2010.

[22] L. A. Witt, L. A. Burke, M. R. Barrick, and M. K. Mount, "The interactive effects of conscientiousness and agreeableness on job performance." *Journal of Applied Psychology*, vol. 87, no. 1, pp. 164–169, 2002.

[23] J. Newby, V. A. Pitura, A. M. Penney, R. G. Klein, G. L. Flett, and P. L. Hewitt, "Neuroticism and perfectionism as predictors of social anxiety," *Personality and Individual Differences*, vol. 106, pp. 263–267, feb 2017.