

Friends don't Lie - Inferring Personality Traits from Social Network Structure

Jacopo Staiano
University of Trento
Trento, Italy
staiano@disi.unitn.it

Fabio Pianesi
Fondazione Bruno Kessler
Trento, Italy
pianesi@fbk.eu

Bruno Lepri
MIT Media Lab
Cambridge, Ma, USA
lepri@fbk.eu

Nicu Sebe
University of Trento
Trento, Italy
sebe@disi.unitn.it

Nadav Aharony
MIT Media Lab
Cambridge, Ma, USA
nadav@mit.edu

Alex Pentland
MIT Media Lab
Cambridge, Ma, USA
sandy@media.mit.edu

ABSTRACT

In this work, we investigate the relationships between social network structure and personality; we assess the performances of different subsets of structural network features, and in particular those concerned with ego-networks, in predicting the Big-5 personality traits. In addition to traditional survey-based data, this work focuses on social networks derived from real-life data gathered through smartphones. Besides showing that the latter are superior to the former for the task at hand, our results provide a fine-grained analysis of the contribution the various feature sets are able to provide to personality classification, along with an assessment of the relative merits of the various networks exploited.

Author Keywords

Personality Recognition, Network Structures, Mobile Sensing, Social Computing.

ACM Classification Keywords

H.m Information Systems: Miscellaneous.; J.4 Computer Applications: Social and Behavioral Sciences.; I.5 Information Systems: Pattern Recognition.

General Terms

Experimentation, Algorithms, Theory.

INTRODUCTION

The rapid global growth of mobile phone usage has reinforced the need to study the psychological and social implications of this technology. Moreover, recent developments in mobile technologies and the advent of smartphones have sensibly broadened the scope of social sciences' studies: researchers can now exploit data collected by means of such

devices, corroborating or even replacing survey-based samplings. Smartphones allow for unobtrusive and cost-effective access to previously inaccessible sources of data related to daily social behavior [40, 28].

Nowadays, these devices are able to sense a wealth of behavioral data: *i*) location, *ii*) other devices in physical proximity through Bluetooth (BT) scanning, *iii*) communication data, including both metadata (logs of who, when, and duration) of phone calls and text messages (SMS) as well as their actual contents, *iv*) scheduled events, *v*) operational status, *vi*) movement patterns, *vii*) usage information, etc.

Recent works have started using smartphone data to automatically infer users' personality traits on the basis of continuously collected data [8, 9, 37]. Chittarajan et al. [8, 9] showed that smartphone usage features (which we will refer to as "actor-based" features from now on, in contrast with "network-based" features) such as the number of calls made or received, their average duration, the total duration of out/in-going calls, the number of missed calls, the number of unique BT IDs seen, Internet usage, and so on, could be predictive of personality traits. Oliveira et al. [37] investigated also the role played by a limited set of nine structural characteristics of the social networks derived from the rich contextual information available in mobile phone data (call logs). On the other hand, by exploiting survey data, works in the tradition of social psychology and network studies (e.g., Kalish and Robins [24]) have proven the existence of important relationships between individual characteristics and the properties of the networks they are part of and, notably, of the so-called ego-networks.

One important individual characteristic that is expected to influence network size and composition is personality. In Social Psychology it is assumed that people's behavior can be explained to some extent in terms of underlying personality traits, which are seen as enduring dispositions that are relatively stable over time [10]. Talks about personality often refer to several dimensions: we are used to talk about an individual as being (non-)open-minded, (dis-)organized, too much/little focused on him/herself, etc. Several existing theories have formalized these informal ways of approaching personalities by means of multifactorial models, whereby an individual's personality is described through a number of fundamental dimensions known as traits, derived through

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to publish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

UbiComp '12, Sep 5-Sep 8, 2012, Pittsburgh, USA.

Copyright 2012 ACM 978-1-4503-1224-0/12/09...\$10.00.

factorial studies. A well known example of a multifactorial model is the Big Five [23] which owes its name to the five traits it takes as constitutive of people's personality: Agreeableness, Conscientiousness, Extraversion, Neuroticism, and Openness.

Kalish and Robins [24] experimentally examined the effect of individual personality differences on their immediate network environment focusing on ego networks, which consist of a focal node ("ego") and the nodes to whom ego is directly connected to (the so-called "alters") plus the ties, if any, among the alters. Their findings showed that psychological predispositions can explain significant portions of the variance of egocentric network characteristics. In line with [24], we investigate the hypothesis that individuals' psychological predispositions tend to shape their immediate network environment. In our work, however, we do not exclusively rely on self-reported data, but prominently exploit real behavioral data, collected by means of smartphones, this way taking full advantage of the power of such technology.

Targeting the automatic recognition of Big Five personality traits, our work extends and merges the lines of research followed by Oliveira et al. [37] and Kalish and Robins [24] by: *i*) exploiting both survey and mobile data and comparing the results obtained thereof; *ii*) focusing on several classes of structural network properties (centrality measures, small world and efficiency measures, triadic structures and transitivity measures) and their relationship to personality traits; and *iii*) comparing the results to those obtained from individual activity (actor-based) data.

Our results show that *i*) personality classification from structural network properties compares in a very favorable manner with (and is often superior to) classification by means of individual activity data; *ii*) mobile phones-based behavioral data can be superior to survey ones for the purposes of personality classification; and *iii*) particular feature set/network type combinations promise to perform better with given personality traits.

RELATED WORKS

In this section we review key works closely related to ours, from two distinct fields: *i*) social psychology and *ii*) social and ubiquitous computing.

Previous Works in Social Psychology

Traditionally, network theorists devote much of their attention to network structure and how the behavior of individuals depends on their location in the network; for instance, individuals occupying central positions and having denser or wider reaching networks may gain faster access to information and assistance [3]. Recently, a special interest in the interaction between personality traits and network positions has emerged: personality traits that predispose people to socialize, such as Extraversion or Openness to experience, might foster and accelerate tie formation in social networks while others like Neuroticism might restrain individuals from creating ties. Mehra et al. [35], for example, found that high self-monitors, i.e. people who are concerned about how they are perceived by others, occupied more central positions in the friendship network of a high-tech company. While these authors applied very specific personality

traits, others [24, 26, 41, 43] have addressed more comprehensive instruments such as the five factor model [19]. For example, previous studies demonstrated a positive correlation between Extraversion and ego-network size (e.g., [43]). However, Extraversion tends to decline with age [11] and, after controlling for age, Roberts et al. [41] found no effect of Extraversion on the size of an individual's social network. Klein et al. [26], instead, found that people who were low in Neuroticism tended to have high degree centrality scores in the advice and friendship networks. Unfortunately, their analysis reports only in-degree centrality and hence it does not allow a complete investigation of relationships between the local network structures and the personality traits of the ego. In order to overcome the limitations of this work, Kalish and Robins [24] presented a new method of examining personal networks of strong and weak ties through a census of nine triads of different types (e.g., WWW, SNS, SSS, where W means "weak tie", S means "strong tie", and N means "no tie"). Their results suggest that people who see themselves vulnerable to external forces tend to inhabit closed networks of weak connections. Conversely, people who seek to maintain their strong tie partners apart tend to be individualists, to believe that they control the events in their lives, and to have higher levels of Neuroticism. Finally, people with strong network closure and "weak" structural holes (where "structural holes" refers to the absence of ties between parts of the network [6]), tend to be more extraverted and less individualistic.

Previous Works in Social and Ubiquitous Computing

A common characteristic of the works reviewed in the previous section is their being based on information collected by means of surveys (e.g., self-reported social relations). Recently, however, researchers in social and ubiquitous computing have started exploring the wealth of behavioral data made available by smartphones, wearable sensors (e.g., sociometric badges [36]), Facebook [39] and Twitter [18, 38]. Exploiting sociometric badges, Olguin et al. [36] found that Extraversion and Neuroticism were positively correlated with *degree*, *closeness*, *betweenness*, and *eigenvector* centrality measures. Moreover, they found a negative correlation between Conscientiousness and *betweenness* centrality. Gloor et al. [17] found a positive correlation between Openness and Agreeableness on the one hand, and *degree* and *betweenness* centrality on the other. Using Facebook data, Golbeck et al. [18] found a positive correlation between the number of friends (taken as a measure of degree centrality) and Extraversion, and a negative correlation between ego-network *density* and Openness and Extraversion. More recently, Quercia et al. [39] argued that Extraversion is a predictor (albeit weak) for the number of social contacts.

Based on a large dataset consisting of recordings of real-life smartphone usage and personality surveys, Chittaranjan et al. [8, 9] exploited actor-based features (e.g. number and duration of calls, BT hits, etc.) in order to automatically classify personality traits. Their results showed that these features could be predictive of the Big Five personality traits. Moreover, the analysis of these features revealed some interesting trends: extroverts were more likely to receive calls and to spend more time on them, while features pertaining

to outgoing calls were found to be not predictive of the Big Five traits. Oliveira et al. [37] extracted 474 variables from Call Data Records (CDRs), at different time scales, and derived from them the users’ social networks; from the latter, they extracted nine structural network features (e.g. *degree*, *efficiency*, etc.). For three personality traits (Extraversion, Agreeableness, and Openness), they obtained significant improvements in classification performance when using some of these structural network characteristics. Inspired by Oliveira et al. [37], our work extends the number and types of global and local social network structural properties to include centrality, small world and efficiency measures, triadic structures and transitivity measures.

DATASET

For our work we exploited a dataset capturing eight complete weeks in the lives of 53 subjects living in a married graduate student residency of a major US university, collected between March and May 2010. Each participant was equipped with an Android-based cell phone incorporating a sensing software explicitly designed for collecting mobile data. Such software runs in a passive manner, and does not interfere with the normal usage of the phone [1].

The data collected consisted of: *i*) call logs, from which we built a Call network whereby participants act as nodes and the numbers of calls between two nodes as edge weights, according to the method used in Eagle et al. [13]; *ii*) proximity data, obtained by scanning near-by phones and other Bluetooth (BT) devices every five minutes, which allowed us to build a BT proximity network with, again, participants as nodes and the counts of social interactions derived from BT data as edge weights; *iii*) data from a survey administered to participants, which provided self-reported information about personality (Big Five) and relationships among subjects. Concerning the latter, the participants were required to assess their closeness to each other on a “0 (no close at all) to 10 (very close)” scale. This information was used to build the Survey network using the obtained scores as edge weights.

More specifically, social interactions were derived from Bluetooth proximity detection data in a manner similar to those in previous reality mining studies [14, 33]. The *Funf* phone sensing platform¹ was used to detect Bluetooth devices in the user’s proximity. The Bluetooth scan was performed periodically, every five minutes, in order to keep from draining the battery while achieving a high enough resolution for social interaction. With this approach, the BT log of a given smartphone would contain the list of devices in its proximity, sampled every 5 minutes. Knowing the BT identifiers of each smartphone in the study, we could thus infer when 2 participants’ phones were in proximity.

The number of subjects varies from one network to another, due to several factors. For instance, some subjects were isolated in the Call network: this could derive from the fact that they had only called people not participating in the data collection (but our data did not include such external calls); or, the call logging might have suffered malfunctions. Thus, these subjects were discarded from the Call network, under the assumption that their empty ego-network structure

¹<http://funf.org>

would introduce undesired noise for the purpose of personality classification. Beside the two basic behavioral networks (Call and BT) and the one based on survey data, we formed a complex behavioral network by combining Call and BT networks in such a way that its node set was the intersection of BT and Call networks’ node sets and its edge weights were a linear combination (the sum of the normalized edge weights) of BT and Call networks’ weights. All our four networks are undirected; they are quantitatively summarized in Table 1.

	Number of nodes	Number of edges
Call net	44	77
BT net	50	823
Call∩BT net	42	609
Survey net	53	590

Table 1: Quantitative summary of the four networks under analysis.

Big Five personality traits were measured by asking subjects to use 1-5 point scales to answer the online version of the 44 questions Big Five questionnaire developed by John et al. [23].

Trait	Description
Agreeableness	sociable, assertive, playful
Conscientiousness	self-disciplined, organized
Extraversion	friendly, cooperative
Neuroticism	calm, unemotional
Openness	creative, intellectual, insightful

Table 2: Big 5 personality traits explained.

The Big Five questionnaire owes its name to the five traits, explained in Table 2, that it takes as constitutive of people’s personality.

The scores of the five traits were computed by summing the (inverted when needed) raw scores of the items (i.e. questions) pertaining to each trait. The results (average, standard deviation, median, minimum and maximum values) are reproduced in Table 3. We performed a Lilliefors’ goodness-

	Mean	St.Dev.	Median	Min	Max
Agre.	34.25	5.03	34	21	45
Cons.	32.49	5.5	34	20	42
Extr.	26.15	6.78	25	12	39
Neur.	22.32	5.85	23	9	34
Open.	33	6.87	34	11	45

Table 3: Statistics for the Big 5 personality traits.

of-fit test of composite normality on each trait’s distribution. All traits are normally distributed ($p < 0.05$).

EXTRACTION OF NETWORK CHARACTERISTICS

Drawing on previous works, we derived a set of network characteristics describing our networks. The features reported in Table 5 have been extracted from both weighted and unweighted networks when applicable.

In the following subsections we describe in detail and justify the features extracted. Previous findings on the relationship between personality traits and structural network properties are reported in Table 4.

	Degree Cent.		Closeness Cent.		Betweenness Cent.		Eigenvector Cent.	Transitivity	WWW triads		SSS triads		SWS triads
	+	-	+	-	+	-	+	+	-	+	-	-	
Agre.	[12, 17, 26]				[12, 17]			[34]					
Cons.	[25]	[46]		[46]		[36, 46]		[34]					
Extr.	[18, 25, 46]		[46]		[46]		[46]		[24]	[24]			
Neur.	[46]	[25, 26]	[46]		[46]		[46]		[24]		[24]	[24]	[24]
Open.	[17]	[25, 26]			[17]								

Table 4: Previously found relations between network measures and personality (+/- indicate positive/negative correlations).

Centrality measures	Degree/Closeness/Betweenness Centrality [16] Eigenvector/Information Centrality [2, 31]
Efficiency measures	Nodal/Local Efficiency [30] Mean Nodal/Local Efficiency*
Transitivity and triadic measures	Global/Local Transitivity* [44] Mean Local Transitivity* Triads {1, 3, 11, 16}* [12] Triads {WWW, SSS, WNW, WSW, SNS, SNW, SWS, SWW, SSW}* [24]

Table 5: Extracted network features (* indicates computation performed on the ego-net).

Centrality Measures

In the literature there is ample, though not always converging, evidence of a relationship between centrality measures and Big Five traits. For instance, according to Kanfer and Tanaka [25] all the Big Five personality traits, with the exception of Agreeableness, correlate closely with *degree*, and more precisely with *in-degree*; moreover, agreeable persons tend to occupy central positions and report more interacting with others while outgoing (extraverted) and secure (low Neuroticism) subjects had more people reporting interacting with them. Klein et al. [26] found negative correlation between *in-degree* centrality from Neuroticism and Openness, and a positive effect of Agreeableness in friendship networks of work group members. Surprisingly, Extraversion had no effect on friendship centrality.

According to [46], Conscientiousness negatively correlates with *closeness*, *betweenness* and *degree* centrality; Extraversion and Neuroticism (the latter in a less evident manner) positively correlate with *degree*, *closeness*, *betweenness* and *eigenvector* centrality. Olguin et al. [36] obtained evidence for the negative correlation of Conscientiousness and *betweenness* centrality. In a more recent study conducted by Gloor et al. [17], the authors found significant positive correlations between Openness and Agreeableness and *degree* and *betweenness* centrality. Inspired by these previous works, we extracted the three standard measures of centrality proposed by Freeman: *degree*, *betweenness*, and *closeness* centrality [16].

These centrality measures can be divided in two classes: those based on the idea that the centrality of a node in a network is related to how close the node is to the other nodes (e.g. *degree* and *closeness* centrality), and those based on the idea that central nodes stand between others playing the role of intermediary (e.g. *betweenness* centrality).

A different interpretation of centrality is given, for instance,

by delta centrality measures, which take into account the contribution of a node to network cohesiveness, inferred from the observed network variation when the node is deleted. We computed a delta centrality measure recently proposed by Latora and Marchiori [31]: *information* centrality, based on the concept of efficient propagation of information over the network [29, 30]. Another centrality measure we extracted, *eigenvector* centrality [2], accords to each node a centrality score depending both on the number and the quality of its connections: having a large number of connections is still valuable, but a vertex with a smaller number of high-quality contacts may outrank one with a larger number of mediocre contacts.

Small World and Efficiency Measures

Latora and Marchiori’s concept [29, 30] of *efficiency* can be used to characterize how close to a “small world” a given ego-network is.

Small world networks are a particular kind of networks that are highly clustered, like regular lattices, and have short characteristic path lengths, like random graphs [45]. The *efficiency* E of a graph G containing N nodes is defined as:

$$E(G) = \frac{1}{N(N-1)} \sum_{i \neq j \in G} \frac{1}{d_{ij}} \quad (1)$$

where d_{ij} is the shortest path length between two nodes i and j in graph G . The use of efficiency measures for the performance evaluation of structural network features is justified by the hypothesis [32] that the rate at which information flows within an ego-network is influenced to some degree by the personality of the ego.

For each node $i \in G$, *local* efficiency is defined as [30]:

$$E_{loc} = \frac{1}{N} \sum_{i \in G} \frac{E(G_i)}{E(G_i^{ideal})} \quad (2)$$

Here, the normalization factor $E(G_i^{ideal})$ represents the efficiency of the ideal case G_i^{ideal} in which i ’s ego-network G_i has all the $k_i(k_i-1)/2$ possible edges, where k_i is the number of edges incident with i . E_{loc} is an average of the *local* efficiency and plays a role similar to transitivity, which will be treated in the next section. Since $i \in G_i$, the *local* efficiency E_{loc} tells how efficient the communication is between i ’s neighbours when i is removed; in other words, *local* efficiency gives a measure of the response, in terms of efficiency, of i ’s ego-network when i is removed.

Conversely, *nodal* efficiency is defined as the inverse of the harmonic mean of path length, hence for a given node $i \in G$

it is calculated as:

$$E_i^{\text{nodal}} = \frac{1}{N-1} \sum_{j \in G} \frac{1}{L_{ij}} \quad (3)$$

We computed *nodal* and *local* efficiency for each node in the networks, along with the *mean nodal* and *mean local* efficiency of each ego-network. All were extracted both on the weighted and unweighted networks, accounting for a total on eight efficiency measures computed.

Transitivity Measures

In [46], extraversion was found to negatively correlate with *local transitivity*; McCarty and Green [34] found that agreeable and conscientious persons tend to have well-connected networks. To account for a possible contribution of this notion to personality prediction, we computed the following three transitivity features: *i) global* transitivity of the ego-networks, *ii) local* transitivity, and *iii) mean local* transitivity. *Global* transitivity gives an indication of clustering properties at the level of the entire ego-network. It is based on triples' counts, where a triple is a set of three nodes connected by either two (open triple) or three (closed triple) ties. The *global* transitivity of a given graph G is then defined as the ratio between the number of closed triples in G and the total number of triples. For each ego-network, we computed this measure, which gives an indication of the clustering in a network, and is often referred to as clustering coefficient. The *local* transitivity of a node, in turn, measures how close its neighbors are to forming a clique (i.e. a complete graph) and the graph to a small-world network [45]. For a node i , *local* transitivity is defined as the proportion of ties between the nodes in i 's neighborhood (i 's ego-net) to the number of ties that could possibly exist between them. Finally, we computed i 's *mean local* transitivity as the mean of the local transitivity values of i 's adjacent nodes.

Triadic Measures

In [12] each triad is described by a string of four elements: the number of mutual (complete) dyads within the triad; the number of asymmetric dyads within the triad; the number of null (empty) dyads within the triad; and, finally, a configuration code for the triads which are not uniquely distinguishable by the first three elements. In the case of directed graphs, every triad may thus occupy one of the 16 possible distinct states. Conversely, in the case of undirected networks, as in our experiments, the triadic census reduces to the following four states: *i) Triad_1*, the empty triad; *ii) Triad_3*, the ratio of triads where two nodes are connected; *iii) Triad_11*, the ratio of triads where a given vertex i is connected to the node j and the node z and there is no edge between the latter two; *iv) Triad_16*, the ratio of triads representing the complete graph, namely i is connected with j and z , and j and z are also connected.

Recently, Kalish and Robins [24] argued that triad proportions can provide more accurate and informative depictions of the egocentric networks than global measures. They also argued that those ego-network properties are significantly associated with the ego's personality traits. In details, when *ego* is connected to two alters, *alter1* and *alter2*, the triad that depicts the relationship between these three actors is

denoted by a three letter combination. The first letter indicates the strength of tie between *ego* and *alter1* (S or W, for Strong or Weak tie), the second letter the strength of the tie between alters (S, W, or N, for Strong, Weak, or No tie) and the third letter the strength of the tie between *alter2* and *ego* (S or W). Given the symmetry of triads, *alter1* and *alter2* are interchangeable and SNW and WNS are thus same triad; hence, a total of nine possible triads can occur in egocentric networks: SSS, SWS, SNS, WWW, WSW, WNW, SSW, SWW, SNW. As for [12], the census is not just a count of the different triad types but rather the proportions of each type against the total number of possible triads given the number of alters in the network: in this way, egocentric networks of different sizes can be compared.

Among the nine triads defined by Kalish and Robins, SSS and WWW represent strong and weak tie network closure, respectively, while WNW, SNS, and SNW represent three different types of structural holes. In particular, WNW represents weak structural holes as implied by Granovetter [21]; SNS represents strong structural holes as permitted by Burt [6]; and SNW represents a mixed structural hole between a strong and weak tie. This third structural hole is permitted by Burt [6] but is also implied by Granovetter [21].

Exploiting this typology of triads, Kalish and Robins suggested that Neuroticism is positively associated with the proportion of WWW triads and negatively associated with the proportion of SSS and SWS triads. Conversely, they found Extraversion to be negatively associated with the proportion of WWW triads and positively associated with the proportion of SSS triads.

In our case, the definition of Strong and Weak ties was established as follows: following [24], from the weighted adjacency matrix, we used as a threshold the 59th percentile of the edge weights array cumulative distribution; then, edges with a weight higher or equal than that threshold were considered as S (Strong) while edges weighting less than the threshold were marked as W (Weak).

AUTOMATIC PREDICTION OF PERSONALITY TRAITS

We turn now to investigating the predictive power of the different features sets discussed above by comparing the results obtained on a personality classification task. To this end, personality traits scores were quantized into two classes (Low/High), using the median values reported in Table 3. Classification was performed by means of Random Forests ensemble classifiers [4]. We chose Random Forests because they satisfy the max-margin property, they do not require parameter tuning, and, importantly, they are feature-space agnostic, i.e. they do not require the specification of a feature-space, as in Support Vector Machines (SVMs) do through the kernels. Moreover, Random Forests are one of the most accurate learning algorithms available [7]. We ran the same experiments described below also by using SVMs with linear and RBF kernels and obtained less stable and accurate results.

The five sets of features introduced above were exploited and compared: *i) centrality* measures, *ii) efficiency* measures, *iii) Davis & Leinhardt's triad census* [12], *iv) Kalish & Robins' triad census* [24], and *v) transitivity* measures. To them, we added three more sets of features, consisting of: *vi) central-*

ity and efficiency features together – i.e. the union of *i*) and *ii*); *vii*) all the triadic measures – i.e the union of *iii*) and *iv*); and *viii*) all the features assessing local connectivity – i.e. the union of *v*) and *vi*). The resulting 8 sets of features were computed on the Survey, Call, BT and on the compound $Call \cap BT$ networks described above.

Classification results were validated by embedding bootstrap [27] in a Leave-One-Out strategy as follows: first, a new dataset *D* was generated by leaving subject *i* out of the original dataset; then, for 100 iterations, a new training set was created by randomly sampling *D* (with replacement) and used to train a classifier, the latter being tested on the left out instance *i*. As a baseline, we chose the classifier that always outcomes the majority class (e.g. in case of perfect balance, the baseline’s accuracy is 50%). The obtained mean accuracy values are reported in Tables 6-10; each table addresses one of the Big Five trait, with columns distinguishing the results according to network type and marginals indicated in italics. As can be seen, in all cases the performances are well above those of the baseline.

Accuracy figures were converted into global ranks and

	Survey	BT	Call	Call \cap BT	
<i>baseline</i>	50.94	58	56.8	52.3	
Centrality measures	65.12	73.59	68.82	62.21	67.57
Efficiency measures	67.14	67.14	72.99	61.88	67.34
Centrality + Efficiency measures	64.09	71.59	69.87	63.75	67.34
Transitivity measures	59.86	73	58.18	66.23	64.36
Kalish & Robins’ triads [24]	65.56	70.2	70.34	72.16	69.37
Davis & Leinhardt’s triads [12]	60.24	70.24	61.15	67.49	64.71
All triads	64.45	71.62	69.97	69.28	68.71
All triads + Transitivity measures	64.56	71.19	68.98	67.9	68.08
	63.88	71.07	67.54	66.36	

Table 6: Accuracies on Agreeableness, and marginals.

	Survey	BT	Call	Call \cap BT	
<i>baseline</i>	54.7	52	56.8	59.5	
Centrality measures	67.07	72.25	63.83	76.97	69.88
Efficiency measures	65.26	67.02	58.82	66.62	64.53
Centrality + Efficiency measures	67.41	69.87	63.41	74.43	67.93
Transitivity measures	67.02	62.08	62.42	65.66	64.34
Kalish & Robins’ triads [24]	65.84	65.85	58.88	64.12	63.37
Davis & Leinhardt’s triads [12]	67.12	66.05	56.84	69.14	64.89
All triads	65.94	67.12	57.35	64.81	64
All triads + Transitivity measures	65.52	67.03	56.59	64.98	63.72
	66.06	67.16	59.52	68.34	

Table 7: Accuracies on Conscientiousness, and marginals.

an all-encompassing analysis of variance on ranks was ran with design Trait(5)*Network-Type(4)*Feature-Set(8). All the various main and interaction effects turned out to be significant ($p < .05$). Pairwise comparisons (with Bonferroni adjustment for multiple comparison and overall $\alpha = 0.05$) on marginal means for Network Type reveals the following ordering of accuracy values: BT (68.56) > Call \cap BT (66.88) > Survey (65.54) > Call (62.86). Concerning the

	Survey	BT	Call	Call \cap BT	
<i>baseline</i>	54.7	60	54.5	54.8	
Centrality measures	66.67	73.08	59.45	68.2	67.02
Efficiency measures	62.53	70.96	58.92	69.62	65.5
Centrality + Efficiency measures	66.27	71.86	61.82	67.7	67.03
Transitivity measures	61.76	79.74	51.55	66.95	65.3
Kalish & Robins’ triads [24]	64.7	70.3	59.85	69.01	66.01
Davis & Leinhardt’s triads [12]	58.45	77.61	51.63	64.87	63.36
All triads	63.47	73.78	59.42	68.53	66.38
All triads + Transitivity measures	64.56	74.89	58.7	67.98	66.69
	63.55	74.03	57.67	67.86	

Table 8: Accuracies on Extraversion, and marginals.

	Survey	BT	Call	Call \cap BT	
<i>baseline</i>	52.8	60	59.1	54.8	
Centrality measures	62.87	60.54	73.74	63.82	64.99
Efficiency measures	64.99	60.6	66.82	63.21	63.86
Centrality + Efficiency measures	62.91	58.53	72.99	63.39	64.21
Transitivity measures	59.25	64.8	57.03	64.02	61.26
Kalish & Robins’ triads [24]	66.67	61.63	59.62	69.75	64.38
Davis & Leinhardt’s triads [12]	59.37	60.19	62.3	61.59	60.76
All triads	64.84	59.59	60.91	66.58	62.92
All triads + Transitivity measures	64.17	59.96	61.51	65.35	62.7
	63.13	60.73	64.37	64.71	

Table 9: Accuracies on Neuroticism, and marginals.

	Survey	BT	Call	Call \cap BT	
<i>baseline</i>	50.9	54	56.8	52.4	
Centrality measures	70.71	65.56	68.39	65.13	67.57
Efficiency measures	71.2	69.44	66.84	66.31	68.63
Centrality + Efficiency measures	70	69.79	68.31	64.02	68.22
Transitivity measures	73.52	70.44	63.37	77.05	71.13
Kalish & Robins’ triads [24]	69.66	68.47	63.21	61.26	65.98
Davis & Leinhardt’s triads [12]	70.82	70.32	63.82	75.52	70.1
All triads	70.69	70.32	64.03	62.85	67.3
All triads + Transitivity measures	72.01	71.71	63.53	64.97	68.39
	71.07	69.51	65.19	67.14	

Table 10: Accuracies on Openness, and marginals.

feature sets, the same procedure revealed that both Centrality (67.40) and Centrality+Efficiency (66.97) outperform Davis & Leinhardt’s triads (64.74) and Transitivity (65.20). Finally, the marginal accuracy values for the Big Five traits could be ordered as follows: Openness (68.23) = Agreeableness (67.21) > Extraversion (65.77) = Conscientiousness (65.27) > Neuroticism (63.24). In summary, at the global level BT is the most, and Call the least, efficient Network for classification purposes; Centrality features (be they alone or in conjunction with Efficiency one) outperform Davis & Leinhardt’s triads and Transitivity while Agreeableness and Openness are the traits that are best recognized. This global picture misses many interesting details, to which we now turn by discussing the results of analyses of variance, one per trait, that were run with Feature-Set and Network-Type

as factors, in a 8*4 design. The results are summarized in Table 11. According to Table 11, the network type has a

	Agre.	Cons.	Extr.	Neur.	Open.
Network-Type	11.422***	17.113***	44.254***	4.082**	7.199***
Feature-Set		3.633**			2.124*
Network-Type * Feature-Set	1.699*		1.412!	2.269**	1.529*

Table 11: ANOVA results. F values and their significance values – !: $p < .1$; *: $p < .05$; **: $p < .01$; *: $p < .001$.**

significant influence on classification results for all traits. A detailed analysis of the sources of these effects (same procedure as above for pairwise comparisons) yields the following patterns:

- *Agreeableness* – BT > Call = Call∩BT > Survey (see Table 6);
- *Conscientiousness* – BT = Call∩BT = Survey > Call (see Table 7);
- *Extraversion* – BT > Call∩BT > Survey > Call (see Table 8);
- *Neuroticism* – no clear ordering, though BT is significantly worse than Call and Call∩BT;
- *Openness* – no clear ordering, though Call is worse than BT and Survey.

In summary, the neat ordering among network type that we detected at the global level is substantially confirmed at the level of the single trait: in all but one case (Neuroticism), BT is the best performing network and Call is the worst one.

Turning to the Feature-Set effects of Table 11, they are significant only with Conscientiousness and Openness. With the former, the effect is due to the better accuracies of Centrality features with respect to the other feature sets; with Openness, the joint analysis of the Feature Set main effect and of the Network Type*Feature Set interaction reveals that the (otherwise quite low) performances of Call∩BT significantly increase when Davis & Leinhardt’s triads or Transitivity are used. The remaining two interaction effects concern Agreeableness and Neuroticism: the former can, at least in part, be attributed to a performance drop of Call with Davis & Leinhardt’s triads and Transitivity, see Table 6; the second interaction effect can be traced back to the accuracy increase obtained when Centrality and Centrality+Efficiency are computed from Call. We also discuss the interaction effect for Extraversion: though only marginally significant ($p < .1$), it is worth commenting because it highlights opposing patterns between BT and Call networks, with the former sensibly increasing its performance with Davis & Leinhardt’s triads and Transitivity and the latter decreasing when the same feature sets are exploited.

We see, therefore, that the pattern highlighted above when discussing the effects of Feature Set at the global level, stems from specific interactions among Network Type, Trait, and Feature Set. Moving from coarser to finer grained considerations, the survey network never outperforms the other network types (though it provides very good results with Openness), suggesting that, despite the many problems that might affect them (e.g., sparseness and incompleteness in the case of the Call network), behavioral data are in a better position than survey data for automatic personality pre-

diction purposes. The second point concerns the relationships between network types and feature sets: in general, Call’s performances tend to decrease with the various types of transitivity and triadic features; BT performance, in turn, are more stable (and higher) across features sets and personality traits. The results from the trait-specific ANOVAs allow refining these general associations: Centrality computed on Call yields high performances with Neuroticism; Davis & Leinhardt’s triads and Transitivity computed on BT improve the classification accuracy with Extraversion, and they do the same for Openness when computed on Call∩BT. The improved results for Extraversion with Davis & Lein-

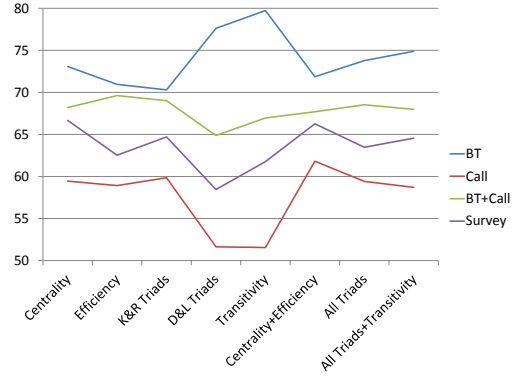


Figure 1: Feature sets’ performance on Extraversion.

hardt’s triads can be further analyzed in the lights of the correlations between the relevant features and the Extraversion trait in the BT network (henceforth, we discuss only correlation with significance level $p < .01$). In particular, we found a positive association (.281) for Triad_16, the (ratio of) triads representing a complete graph, and negative associations (-.345, -.283 and -.237, respectively) for Triad_1, the empty triad, Triad_3, the triads with two connected nodes, and Triad_11, the triads with two edges. In other words, in our BT network, extraverts tend to have more complete triads and less incomplete or empty ones, than introverts. As one may put it, extraverts seems likely to keep their close partners together, perhaps by actively seeking to introduce them to one another at the social gatherings captured by the BT network. Kalish and Robins’ triads, which consider the strength of the ties, seem to be slightly less informative for Extraversion classification, according to our results; and, correspondingly, we find only a couple of significant correlations in this case: with WNW (-.228) and with SNW (-.235) triads. In these respects, our BT network does not replicate Kalish and Robins [24]’ findings that extraverts have proportionally more SSS and SWS triads and fewer WWW triads, a difference that can reasonably be due to the different types of networks these data are drawn from (survey data in [24], BT in the present discussion). Finally, all features in our transitivity set significantly correlate with Extraversion: local transitivity (.301), mean local transitivity (.282) and global ego-network transitivity (.285); in correspondence, classification performance on the BT network gets up to 80%. It should be noticed that our correlation figures contrast with those obtained by Wherli [46] where

Extraversion was found to negatively correlate with local transitivity. However, our results seem supported by those of Hallinan and Kubitschek [22] who, examining the relationship between tolerance for intransitivity and friendliness, found that friendly students have a lower tolerance for intransitive triads and tend to remove them over time. Finally, no significant correlations could be found in the Call network between the Extraversion trait and any of the features composing the Davis & Leinhardt’s triads and the Transitivity sets, possibly explaining the drop in accuracy discussed above.

Turning to Neuroticism, the association with centrality measures in the call network that our classification results reveal can be traced back to the level of correlations to degree centrality (.257), a datum that is in line with findings in [46]. By indicating a more substantial (though not necessarily linear) relationship between centrality features and Neuroticism, our classification results call for further investigation of the underlying phenomena.

Our conclusions concerning Agreeableness are similar to those for Neuroticism. In the literature, this trait has not been investigated much by means of network-level measures. On our side, we could only find a significant positive association in the Call network between Agreeableness and local efficiency (.246), which measures the mean efficiency internal to an ego-network, an index related to small world formation. Correspondingly, the Call network accuracy gets up to 73% with the Efficiency feature set. As it seems, therefore, more agreeable people have some tendency towards forming small worlds than less agreeable ones; again, this is a datum that, together with its import to the explanation of our classification results, needs further investigation. The literature does not offer much to discuss, and compare with, concerning the elusive trait of Openness. Given this lack of information, the level and type of recognition accuracy that we obtained is even more remarkable: at the global level, Openness is one of two best recognized trait, with average accuracy 68.23; it seems capable of taking specific advantage of features, such as Davis & Leinhardt’s triads and Transitivity, which measure the level of connectedness of the egonet; it is also the trait where information from the surveys performs better. Definitely, more work is needed.

DISCUSSION AND COMPARISON WITH PREVIOUS WORKS

A number of recent works have used mobile phones data in order to automatically infer and predict personality traits ([8, 9, 37]). In particular, [8] exploited actor-based features (e.g., the number of calls made or received, their average duration, the total duration of out/in-going calls, the number of missed calls, the number of contacts associated with missed calls, the number of unique BT IDs seen, Internet usage, and so on). In this work, we have focused on the usage of network-level features, arguing that they are more informative for the task at hand than actor-based ones. In order to contribute to the assessment of the relative merits of these two approaches, we computed actor-based features from our networks and compared the results obtained through them to those discussed above. Because of the different nature of our data, we could not fully replicate Chittaranjan et al.’s study as we only had available the following activity level features:

number of outgoing calls, number of incoming calls, number of calls from a unique subject, number of calls directed to a given subject, number of unique subjects in proximity (through BT), max time a subject was seen in proximity, total time seen in proximity. The obtained average accuracies are reported in Table 12.

While consistently performing above the baseline, actor-based features seem to perform worse than network-based ones with any traits, with the possible exception of Conscientiousness, as a comparison between Table 12 and Tables 6-10 above shows. An approach more similar to ours is re-

	Agre.	Cons.	Extr.	Neur.	Open.
Call features only	71.2	65.4	72.9	66.1	69.2
BT features only	67.1	67.7	62.9	63.2	67.1
All (Call + BT) features	69.1	65.7	68.8	65.2	69.8
<i>baseline</i>	50.9	54.7	54.7	52.8	50.9

Table 12: Accuracies obtained using actor-based features.

ported in [37]. In this work, the authors used 9 network-based features including the number and the weight (measured by the number of reciprocal phone calls) of contacts (degree of the nodes), the number and social distance between relevant contacts, etc. The authors reported significant improvements ($p < .05$) in the classification performance for 3 traits, Extraversion, Agreeableness, and Openness, when the network-level features were included among predictors. Although at a first sight in line with the general trends we obtained concerning those three traits, more direct comparisons are made difficult by the limited amount of information about the way the network-level features were computed and used in [37].

More generally, with respect to the quoted studies, the present work has addressed a larger number of aspects relevant to the usage of behavioral data from mobile phones for the task of automatically predicting personality traits. In the first place, we emphasized the importance of egonets’ structural characteristics, as those that more clearly associate with personality traits variations. In the second place, we systematically investigated the predictive power of those structural properties with networks arising from different behaviors (call logs and BT proximity), and compared the obtained results to those with survey data, which still are the most common source of data in the social sciences. The results we obtained provide encouraging evidence that behavioral data are indeed better suited to our task. Moreover, the detailed analyses conducted on the relationships among the feature set exploited, the different networks and the Big Five traits have allowed us to reveal their relative merits for the task of personality prediction. Among the relevant findings, we list the superiority of triadic and transitivity features for Extraversion prediction on BT data; the importance of centrality and efficiency features from Call data for Neuroticism and Agreeableness; the overall greater richness of the information provided by BT data with respect to Call data.

PRACTICAL IMPLICATIONS AND LIMITATIONS

The automatic determination of personality from mobile phone data can provide a new and interesting framework for

mobile and, more in general, pervasive computing. As observed in [9], the ability of inferring and predicting personality and other psychological variables through contextual data collected by mobile phones could be used in various ways in the context of mobile applications.

In the first place, previous works have shown that personality is linked to user interface preferences [5]. The personality of a user might also determine the kind of functions he/she is disposed to use on the phone, as in the case of recommendation systems that attempt to match the preferences and personalities of their users [20].

Another important practical implication of our research program is the use of the automatic understanding of personality from mobile phone data for the design of more effective strategies of mobile persuasion. Given their pervasiveness, mobile phones are becoming the most powerful channel for persuasion, more influential than TV, radio, print, or the Internet [15]. At the same time, some studies (e.g., [42]) have convincingly shown that psychological variables affect if and how people are amenable to persuasion as well as the choice of the best means to bring it about; as a consequence, automatically inferred personality traits can be used to build more effective change-inducing systems.

Turning to the limitations of the present study, we list the following ones: the relatively small size of the sample; the fact that it comes from a population living in the same environment (our subjects were all married graduate students living in a campus facility of a major US university); the non-availability of behavioral data concerning the interaction with people not participating in the data collection, a fact that is common to many other studies of this type and that has been also pointed out by [39]. The first two problems are at least partially attenuated by the large variability of the sample in terms of provenance and cultural background, which can be expected to correspond to a wide palette of interaction behaviors that efficaciously counterbalance the effects of sample's small size and of living-place homogeneity.

CONCLUSIONS

This paper aimed to contribute to advance the state of the art in the automatic analysis of people personality by deepening and extending previous works along several dimensions: a) the usage of network-level features (and in particular those addressing the properties of egonets) and the comparison of the results obtained with those attained through actor-based features; b) the comparison of the results from two different types of mobile phone data with those based on more traditional surveys; c) the systematic analysis of the relative strengths and weaknesses of the exploited feature sets across network types and the Big Five traits. Despite the limitations of this study discussed above, we believe that our results have provided compelling evidence that mobile phones-based behavioral data can be superior to survey ones for the purposes of personality classification and that egonet-based features can improve performance over actor-based ones. Moreover, we have provided many new insights concerning the feature set/network type combinations that promise to perform better with given personality traits.

ACKNOWLEDGEMENTS

The work of Jacopo Staiano and Nicu Sebe was partially supported by the FIRB S-PATTERNS project. Bruno Lepri's research is funded by PERSI project within the Marie Curie COFUND – FP7. We gratefully acknowledge Wei Pan's and Scott Greenwald's support for data preparation.

REFERENCES

1. Aharony, N., Pan, W., Ip, C., Khayal, I., and Pentland, A.: Social fMRI: Investigating and shaping social mechanisms in the real world. *Pervasive Mob. Comput.* 7(6), 643-659 (2011)
2. Bonacich, P.: Power and Centrality: A Family of Measures. *American Journal of Sociology*, 92, 1170-1182 (1987)
3. Borgatti, S.P. and Foster, P.: The network paradigm in organizational research: A review and typology. *Journal of Management*. 29(6), 991-1013 (2003)
4. Breiman, L.: Random Forests. *Machine Learning*, 45(1), (2001)
5. Brinkman, W.P. and Fine, N.: Towards customized emotional design: an explorative study of user personality and user interface skin preferences. *EACE* (2005)
6. Burt, R.S.: *Structural holes: the social structure of competition*. Harvard University Press (1992)
7. Caruana, R., Karampatziakis, N., and Yessenalina, A.: An empirical evaluation of supervised learning in high dimensions. *ICML* (2008)
8. Chittaranjan, G., Blom, J., and Gatica-Perez, D.: Who's who with Big-Five: Analyzing and Classifying Personality Traits with Smartphones. *ISWC* (2011)
9. Chittaranjan, G., Blom, J., and Gatica-Perez, D.: Mining Large-Scale Smartphone Data for Personality Studies. *Personal and Ubiquitous Computing*, in press
10. Costa, P.T. and McCrae, R.R.: Revised NEO Personality Inventory (NEO-PI-R) and NEO Five-Factor Inventory (NEO-FFI) manual. *Psychological Assessment Resources* (1992)
11. Costa, P.T., McCrae, R.R., and Siegler, I.C.: Continuity and change over the adult life cycle: Personality and personality disorders. *Personality and psychopathology*, 129-153 (1999)
12. Davis, J.A. and Leinhardt, S.: The Structure of Positive Interpersonal Relations in Small Groups. In: *Sociological Theories in Progress*, 2, 218-251 (1972)
13. Eagle, N., Pentland, A., and Lazer, D.: Inferring Social Network Structure using Mobile Phone Data. *PNAS* (2007)
14. Eagle, N. and Pentland, A.: Reality mining: sensing complex social systems. *Personal and Ubiquitous Computing*, 10(4), 255–268 (2006).

15. Fogg, B.J. and Eckles, D.: The Behavior Chain for Online Participation: How Successful Web Services Structure Persuasion. *Persuasive*, 199-209 (2007)
16. Freeman, L.: Centrality in social networks: Conceptual clarification. *Social Networks*, 1(3), 215-239 (1979)
17. Gloor, P.A., Fischbach, K., Fuehres, H., Olguin, D., Pentland, A., and Putzke, J.: Towards "Honest Signals" of Creativity - The Association between Actors' Personality Characteristics, Body Movements and Embeddedness in Social Networks. *COINs* (2010)
18. Golbeck, J., Robles, C., Edmondson, M., and Turner, K.: Predicting Personality from Twitter. *SocialCom/PASSAT* (2011)
19. Goldberg, L. R.: An alternative "Description of personality": The Big-Five factor structure. *Journal of Personality and Social Psychology*, 59, 1216-1229 (1990)
20. Graham, L., Gosling, S.: Can the ambiance of a place be determined by the user profiles of the people who visit it? *ICWSM* (2011)
21. Granovetter, M.: The Strength of Weak Ties. *American Journal of Sociology*, 78(6), 1360-1380 (1973)
22. Hallinan, M.T. and Kubitschek, W.: The effects of individual and structural characteristics on intransitivity in social networks. *Social Psychology Quarterly*, (2), 81-92 (1988)
23. John, O.P. and Srivastava, S.: The Big Five trait taxonomy: History, measurement, and theoretical perspectives. In: *Handbook of personality: Theory and research*, 102-138 (1999)
24. Kalish, Y. and Robins, G.L.: Psychological predispositions and network structure: The relationship between individual predispositions, structural holes and network closure. *Social Networks* (2006)
25. Kanfer, A. and Tanaka, J.S.: Unraveling the Web of Personality Judgments: The Influence of Social Networks on Personality Assessment. *Journal of Personality*, 61(4) 711-738 (1993)
26. Klein, K.J., Lim, B.C., Saltz, J.L., and Mayer, D.M.: How do they get there? An examination of the antecedents of network centrality in team networks. *Academy of Management Journal*, 47, 952-963 (2004)
27. Kohavi, R.: A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection. *IJCAI* 1137-1143 (1995)
28. Lane, N.D., Miluzzo, E., Lu, H, Peebles, D., Choudhury, T., and Campbell, A.T: A Survey of Mobile Phone Sensing. *IEEE Communications Magazine* (2010)
29. Latora, V. and Marchiori, M.: Efficient Behavior of Small-World Networks. *Phys. Rev. Lett.* 87(19) (2001)
30. Latora, V. and Marchiori, M.: Economic Small-World Behavior in Weighted Networks. *Eur. Phys. Journ. B Condensed Matter* 32(2) (2003)
31. Latora, V. and Marchiori, M.: A measure of centrality based on the network efficiency. *New J. Phys.* 9, 188 (2007)
32. Lu, Y.E., Roberts, S., Lio, P., Dunbar, R., and Crowcroft, J.: Size matters: variation in personal network size, personality and effect on information transmission. *ICCSE* (2009)
33. Madan, A., Cebrian, M., Lazer, D., and Pentland, A.: Social sensing for epidemiological behavior change. *Proc. ACM UbiComp 2010*, 291-300 (2010)
34. McCarty, C. and Green, H.D.: Personality and personal networks. *Sunbelt XXV* (2005)
35. Mehra, A., Kilduff, M., and Brass, D.J.: The Social Networks of High and Low Self-Monitors: Implications for Workplace Performance. *Administrative Science Quarterly*, 46(1) 121-146 (2001)
36. Olguin, D., Gloor, P.A., and Pentland, A.: Capturing Individual and Group Behavior with Wearable Sensors. *AAAI Spring Symposium on Human Behavior Modeling* (2009)
37. Oliveira, R., Karatzoglou, A., Concejero, P., Armenta, A., and Oliver, N.: Towards a psychographic user model from mobile phone usage. *ACM CHI, WIP*, 2191-2196, (2011)
38. Quercia, D., Kosinski, M., Stillwell, D., and Crowcroft, J.: Our Twitter Profiles, Our Selves: Predicting Personality with Twitter. *SocialCom/PASSAT* (2011)
39. Quercia, D., Lambiotte, R., Stillwell, D., Kosinski, M., and Crowcroft, J.: The personality of popular facebook users. *CSCW* (2012)
40. Raento, M., Oulasvirta, A., and Eagle, N.: Smartphones: An emerging tool for social scientists. *Sociological Methods and Research*, 37(2), 426-454 (2009)
41. Roberts, S.G.B., Wilson, R., Fedurek, P., and Dunbar, R.I.M.: Individual differences and personal social network size and structure. *Personality and Individual Differences*, 44(4), 954-964 (2008)
42. Halko, S. and Kientz, J.A.: Personality and Persuasive Technology: An Exploratory Study on Health-Promoting Mobile Applications. *Persuasive*, 150-161 (2008)
43. Swickert, R.J., Hittner, J.B., Harris, J.L., and Herring, J.A.: Relationships among Internet use, personality, and social support. *Computers in Human Behavior*, 18(4), 437-451 (2002)
44. Wasserman, S. and Faust, K.: *Social Network Analysis: Methods and Applications*. Cambridge Press (1994)
45. Watts, D.J. and Strogatz, S.: Collective dynamics of 'small-world' networks. *Nature* 393(6684), 440-442 (1998)
46. Wehrli, S.: Personality on Social Network Sites: An Application of the Five Factor Model. *ETH Zurich Sociology Working Papers* 7 (2008)