2017 IEEE First Ukraine Conference on Electrical and Computer Engineering (UKRCON)

Clustering of Users in Social Networks by Their Activity

Viktor Hozhyi
Applied Mathematics Dept.,
Faculty of Applied Mathematics,
NTUU "I. Sikorsky KPI"
Kyiv, Ukraine
hozhyi.victor@ukr.net

Bart Lamiroy
Université de Lorraine – Loria (UMR 7503),
Campus Scientifique – BP 239,
54506 Vandoeuvre-lès-Nancy, France
Bart.Lamiroy@loria.fr

Abstract — In this paper we present some preliminary experimental results related to searching correlations between users' psychological traits and the properties of their social network graph. We explore simple clustering techniques on standard available data extracted from mainstream social media platforms. Our experiments on approximately 1000 users show that we are capable of identifying 9 groups of users by their activity on social networks

Keywords — psychology; social graph; social network; clustering; psychological traits

I. INTRODUCTION

The main drive behind this work is an attempt to classifying users of social networks in function of factual usage statistics and investigate whether these groups can be brought into correspondence with general psychological traits defining their members.

In this work, we will classify users by their perceived psychological profile, extracted from their day-to-day use of social networks. This profile is composed of traits like consciousness, agreeableness, extraversion and neuroticism. It is possible to define user psychological traits without other data than just their social graph. However, this work could be also interesting for classifying users by other criteria.

Quite a lot of research in this area already exists. MasterCard Consumer Research [1] in 2013 defined five personas: Open Sharers, Simply Interactors, Solely Shoppers, Passive Users and Proactive Protectors. AllOutDigital blogger Ranjan in 2012 obtained eight types of Internet users [2]: Centers, Navigational browser, Networkers, Expressers, Regular Web Person, Schoolers and Gamers. Enterprise Social Network "Zyncro" defined seven types of Internet users in 2013 [3]: the Hyperconnected, the Geek, the Digital, the New digital, the Senior, the "Too late for me" and the "Against the world". "LoadMeNa.com" in 2016, similar to Master Card, identified five types [4]: Searcher, Shopper, Downloader, Gamer and Socializer. However, all these researches used just activity classifying, meanwhile we need to classify users by psychological traits. In addition, they didn't present any function of user statistics, so we don't know how to classify users automatically.

The rest of this paper is organized as follows: first, we will define the concept of psychological traits, how they have been handled in the literature and how some of them relate to specific behavior that can be observed using social networks. In Section III we look into the means of defining specific descriptors from a social network graph that relate to these defined traits. Section IV then addresses the more technical issues of collecting and analyzing these data, as to automatically obtain clusters of similar use cases.

II. DEFINING PSYCHOLOGICAL TRAITS

Defining psychological user profiles is essential for social interaction – each person can interact with specific set of people, high on specific psychological traits or have compatible temperament. That is why it is sometimes very important to know temperaments of people to have better communication or better social better social interaction (on work or somewhere else)."Four temperaments" [5] were first defined by Hippocrates in about 400 BC: sanguine (optimistic, active and social), choleric (short-tempered, fast or irritable), melancholic (analytical, wise and quiet), or phlegmatic (relaxed and peaceful). However,, those temperaments are not quite fully representative of human psychology anymore, nor are they in line with contemporary psychological state-of-the-art.

In 1884, Sir Francis Galton [6] investigated psychological personality traits, that later were more investigated by Gordon Allport and S. Odbert through putting Galton's hypothesis into practice by extracting 4,504 traits in 1936. In [6], The Big Five personality traits, also known as the five factor model (FFM), is a model based on common language descriptors of personality. These descriptors are grouped together using a statistical technique called factor analysis (*i.e.* this model is not based on experiments). This theory suggests five broad dimensions used by some psychologists to describe the human personality and psyche. Those five factors are openness to experience, conscientiousness, extraversion, agreeableness, and neuroticism, often listed under the acronyms OCEAN or CANOE.

Openness to experience: (inventive/curious vs. consistent/cautious). Appreciation for art, emotion, adventure, unusual ideas, curiosity, and variety of experience. Openness reflects the degree of intellectual curiosity, creativity and a preference for novelty and variety a person has. It is also described as the extent to which a person is imaginative or

978-1-5090-3006-4/17/$31.00 ©2017 IEEE

independent, and depicts a personal preference for a variety of activities over a strict routine. High openness can be recognized as unpredictability or lack of focus. Individuals with high openness are said to pursue self-actualization specifically by seeking out intense, euphoric experiences, such as skydiving, living abroad, gambling, etc. Conversely, those with low openness seek to gain fulfillment through perseverance, and are characterized as pragmatic and data-driven

Conscientiousness: (efficient/organized vs. easy-going/careless). A tendency to be organized and dependable, show self-discipline, act dutifully, aim for achievement, and prefer planned rather than spontaneous behavior. High conscientiousness is often perceived as stubborn and obsessive. Low conscientiousness is flexible and spontaneous, but could be sloppy and unreliable.

Extraversion: (outgoing/energetic vs. solitary/reserved). Energy, positive emotions, surgency, assertiveness, sociability and the tendency to seek stimulation in the company of others, and talkativeness. High extraversion is often perceived as attention-seeking, and domineering. Low extraversion (introverts) causes a reserved, reflective personality, which can be perceived as aloof or self-absorbed.

Agreeableness: (friendly/compassionate vs. analytical/detached). Rather compassionate and cooperative than suspicious and antagonistic towards others. It is also a measure of one's trusting and helpful nature, and whether a person is generally well-tempered or not. High agreeableness is often seen as naive or submissive. Low agreeableness personalities are often competitive or challenging people, which can be seen as argumentative or untrustworthy.

Neuroticism: (sensitive/nervous vs. secure/confident). The tendency to experience unpleasant emotions easily, such as anger, anxiety, depression, and vulnerability. Neuroticism also refers to the degree of emotional stability and impulse control. A high on this trait need for stability manifests as a stable and calm personality, but can be seen as uninspiring and unconcerned. A low on neuroticism need for stability causes a reactive and excitable personality, often very dynamic individuals, but they can be perceived as unstable or insecure.

Narcissism (from a free Wikipedia [7]) is the pursuit of gratification from vanity or egotistic admiration of one's own attributes.

III. PSYCHOLOGICAL TRAITS EXTRACTING FROM A SINGLE USER PROFILE

The problem of extracting psychological traits from a user profile in social network has been investigated for many years by different researchers: Ph.D. Susan Krauss Whitbourne [8], Amy Morin [9], Ph.D. Gwendolyn Seidman [10], Agata Blaszczak-Boxe [11] and many others.

Below is a shortened, synthetic review of these works that allow us to relate key features of a user profile with their psychological portrait:

- As in [8], extroverts, in general, have more friends and they spend more time on social networks, but according to [9] there is also a risk that person with big amount of friends on social network has low self-esteem.
- Introverts watch what their friends are doing, and regret what they do share about themselves, as figured out in [8].
- Extroverts post more likely about social activities and their everyday lives [10].
- Extroverts and neurotics both upload significant numbers of photos to their Facebook pages, but extroverts tend to change their profile cover photos, while neurotics tend to upload more photos per album [11]
- Extroverts upload photos and update their status more often than introverts [9]
- Open people are likely to post updates about their intellectual interests and fill out their personal profiles most thoroughly [10]
- High on conscientiousness generally staying off Facebook [8]
- Conscientiousness individuals were more likely to post updates about their children [10]
- Conscientious people in the study uploaded more videos and created more "self-generated" photo albums and organize it more carefully [11]
- Neurotic people post mostly photos [9]
- Agreeable people tended to attract more comments and "likes" to their posts [11]
- Agreeable people are tagged in other people's photos most often [9]
- Narcissists were more likely to post about their achievements, about diet and exercise [10]
- Narcissists receive the greatest number of likes and comments, particularly if they posted about achievements [10]

As a conclusion, we have 5 characteristics of extroversion, 1 of openness, 3 of conscientiousness, 1 of neuroticism, 2 of agreeableness and 2 of narcissism. Therefore, we have characteristics of 5 main traits from the Big Five Model and for one additional trait – narcissism.

IV. PROCESSING THE DATA

A. *Collecting the data*

We want to download the relevant usage statistics mentioned in the previous section of about 1000 social network users: count of friends / photos / profile photos / tagged photos / albums / videos, average count of likes and comments on posts.

We can use a few popular social networks for collecting relevant data by querying information of big amount of user's profiles.

- Facebook. According to Facebook API [12], from Facebook we can query almost all information of a user, like count of albums, friends, photos, likes (of pages) and

comments. However, according to privacy rules, we cannot query any information about user friends except their count. That is why we cannot explore the full social graph of a user in Facebook.

- VK. According to [13], VK Open API has fewer restrictions and we can observe all friends of a user, not just count of opened photographs.

We therefore decided to use the VK social network to extract the required data. The security settings of the social network (one cannot query user information repeatedly, without delay) constrained us to use a 8 sec between each query. As a result, retrieving the data of approximately 1000 users requires a day of time.

B. *Normalizing and cleaning data*

After the initial phase, we have information of users, represented as set of vectors. However, some contained empty vectors, representing deleted or inactive users. We do not need those vectors, and they were removed them from our set.

In addition, each dimension of our description vectors has different meanings and obviously has different distribution of values. Therefore, we decided to normalize those values by their expected value, because it gives us similar distribution (with expected values equal 1), not just similar max value equal to 1 in case of normalizing by max value.

Finally, there is a small amount of very big values per some dimensions (for example, count of tagged photos, as shown in Fig. 1).

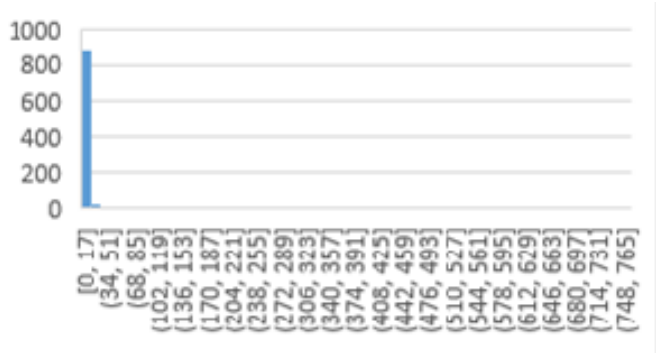

Fig. 1. Distribution of values of photos tagged count.

Therefore, we have removed from our database all outlier profiles, containing very strong standard deviation values.

TABLE I. Standard deviation for each dimension before cleaning

Dimension	Standard deviation
friends count	3.806119
profile_photos_count	8.715536
photos_tagged_count	26.8862
photos_count	5.097114
albums_count	5.200779
videos_count	5.772548
avg_likes_count	7.301909
avg_comm_count	575.6317

For defining which vectors of data are too big, we have computed the standard deviation for each dimension. As we can see on Table I – we have very big standard deviation on average count of comments, so we should remove some vectors with biggest value on this dimension. After these removals standard deviation decreased to 23.6.

However, same operation on count of tagged photos do not give us same result. With each removal, the standard deviation decrease is very slow. We therefore decided not to touch them.

C. *Clustering*

Actually, we do not know how many clusters we have, that is why we used hierarchical clustering. We tried different methods (like 'single', 'complete', 'average' etc.) and metrics ('Euclidean', 'Minkovski', 'Mahalanobis' etc.), that are available in science library for Python [14] [15], and the best distribution of clusters we have got with Ward method in Euclidean metric (see the result on the Fig. 2).

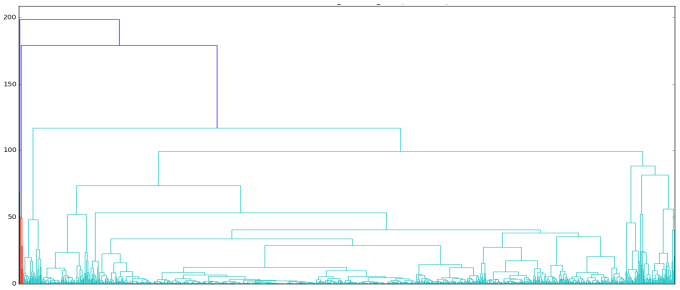

Fig. 2. Hierarchical clustering dendrogram.

As we decided to use the Big Five model for defining user traits, we have 5 different traits, on which user can be low or high, it is 2^5 = 32 combinations, so we decided to analyze hierarchical clusters from number of clusters equal to 32. By removing statistically insignificant clusters (less than 6 members)), we get 18 classes of users. For visualizing those clusters, we have normalized their centers (dividing by max value per dimension) and used a Radar chart type, as shown in Fig. 3.

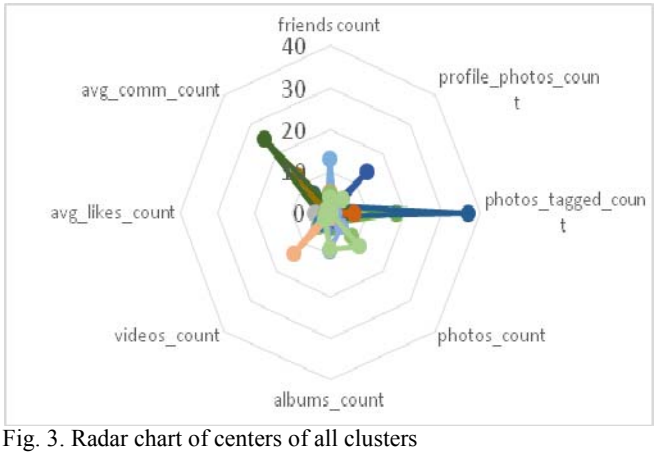

Fig. 3. Radar chart of centers of all clusters

TABLE II. Normalized by expected value centers of clusters (read as average values, except column "Cluster size")

№	Friends count	Profile photos count	Photos tagged count	Photos count	Albums count	Videos count	Avg likes count	Avg comm count	Cluster size
1	1.66500479	2.207084469	36.74772037	1.310347061	1.931518383	0.824589922	1.852074918	6.576912913	7
2	0.715313473	0.883855586	6.482871074	0.961191302	1.039169941	0.53289588	0.561866256	1.233411556	32
3	1.026664342	2.052201348	17.69610195	3.070745727	3.692689484	1.141484247	0.636907243	1.812672376	19
4	0.490448341	0.119422684	0.050711331	0.269293525	0.206501719	0.300960619	0.558165649	0.175335947	384
5	5.030435979	0.194141689	0.098150963	0.493917677	0.182711198	0.698533523	0.830116431	0.908652444	16
6	12.73514885	1.89373297	0.23556231	0.639075231	0.791748527	0.814684937	3.624433189	0.964742101	12
7	1.539810597	0.286103542	3.559608241	0.453808352	0.182711198	0.128764792	3.237862638	24.99355112	6
8	0.940856758	2.634667784	0.15704154	0.503878894	0.281094151	0.588013187	1.44750301	12.39024443	13
9	0.621245795	0.20185731	0.371771809	0.397210939	0.365422397	0.443980803	1.484868974	5.693489373	49
10	4.557293447	1.951634877	0.039260385	8.047649384	0.959233792	4.57470974	1.508249128	0.538460705	8
11	1.362547729	0.177111717	0.15704154	1.468765249	1.248526522	13.84004849	0.992699827	0.336537944	12
12	0.722867288	0.464453802	0.019630193	0.825314456	0.589658868	3.609269257	0.836166041	0.434322216	88
13	3.630516934	4.547002725	0.412234043	11.14638117	8.724459724	1.998794827	0.722034157	1.177882793	8
14	0.622941235	0.515682068	0.133652375	1.544196786	2.464657442	0.459593894	0.624188761	0.269230355	94
15	1.668583078	0.790190736	1.196506972	4.390088306	9.048554587	1.135712546	0.493547064	0.824357702	21
16	1.275770553	0.21358882	0.053191489	0.713909002	0.288801572	0.709923756	4.217459533	0.030396975	62
17	1.084892101	13.89645777	0.687056738	2.836015709	1.233300589	0.892609304	0.737767371	0.722504896	16
18	0.80576141	4.025146909	0.052977869	1.364076001	1.144696665	0.688679817	0.712991278	0.175161916	83

However, there is could be similar centers not by values, but by form of radar, so we decided to view on each separate cluster and compare them to find similar. On Fig. 4, you can see each separate scaled radar view of cluster centers, and as you can see, some of them are very similar (for better visibility charts already are sorted by similarity). We decided to merge those clusters: {1, 2, 3}, {7, 8, 9}, {13, 14, 15}, {5, 6}, {11, 12}, {17, 18}.

D. Results

We have chosen a Big Five Model for characterizing users. Unfortunately, we have not enough data and knowledge about openness to experience, so we have investigated just four dimensions of this model.

Therefore, we defined nine classes of users by their activity:

1. Those who are often tagged on photographs by others are clusters #1-3. According to the theory part we can assume that they are

- Extraverts, except cluster #2 (they are low on extroversion). Cluster #1 has bigger than average amount of friends, #2 has smaller than average, #3 – average; about photos count - #3 has 3 times more photos than general, #1 – average amount, #2 – less than average. As we know from the theory sections – high on extraversion have more friends and upload more photos. #1 and #3 are high on conscientiousness, #2 – general. As mentioned earlier, high on this trait have more self-generated photo albums and upload more videos, and clusters #1 and #3 have more than average amount of albums (they actually have average amount of videos, but we think, that amount of albums is more important for this trait).
- General on neuroticism. High on neuroticism posts mostly photos. Yes, clusters #1-3 have big amount of photos, but they also have not small amount of videos.
- They seem to be agreeable people, but only cluster #1 has higher than average amount of likes under their posts. Another two clusters have less than average amount of likes. Nevertheless, we still think that people from clusters #2-3 are agreeable.

2. Clusters #7-#9, people with big amount of comments under their posts.

#8-9 – introverts, #7 – could be high on extraversion, but not enough amount of photos; Low on conscientiousness (not organized albums), except cluster #9 – they are general on this trait; #8-9 - low on neuroticism (smaller then general amount of photos), #7 – possibly high on this trait (almost no videos on pages, e.g. mostly photos)

#7 – agreeable, #8-9 – could be narcissists. Cluster #7 should be set of true agreeable people, because there is big number of comments, likes and they are often tagged on others' photos. #8 and #9 are tagged on others photos much more rarely then general, that is why they could be rather categorized as narcissists.

3. #Clusters #13-15 – big amount of photos:

#13 and #15 are extraverts (big amount of friends, photos and profile photos), #14 rather introverts (opposite to previous two, see Table II). All of them are conscientious people (big amount of photo albums), they are general on neuroticism (there is no just photos on their pages) and they are not so agreeable people (mostly smaller than average amount of likes and comments)

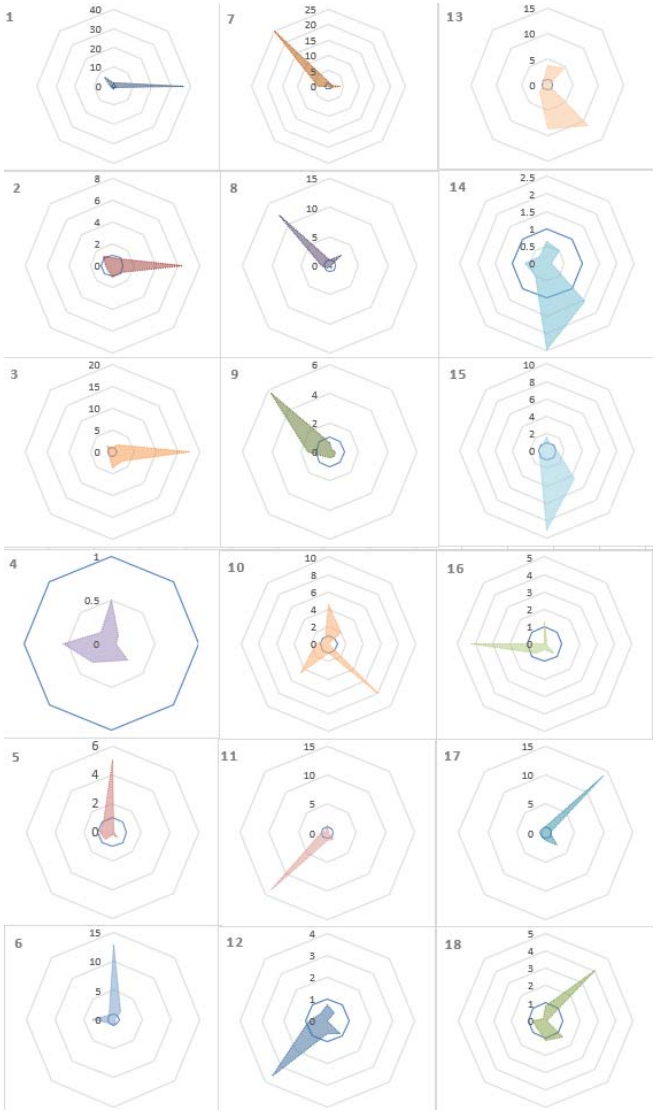

Fig. 4. Scaled radar charts of centers of all separate clusters

4. Cluster #4, the biggest one by number of points:

Introverts (less than average amount of everything); They should be conscientious people, because of low activity; low on neuroticism (small amount of photos); low on agreeableness (small amount of likes and comments)

5. Cluster #10 – big amount of photos, videos and friends:

Extraverts (big amount of friends, photos and profile photos); Low on conscientiousness (small amount of albums comparing to big amount of photos); low on neuroticism (there is also big amount of videos); low on agreeableness (very small amount of tagged photos)

6. Cluster #16 – users with big amount of likes:

Undefined on extraversion (average amount of photos and friends); Low on conscientious (very small amount of albums); General on neuroticism (general amount of photos and videos); Undefined on agreeableness (very small amount of comments but big amount of likes)

7. Clusters #5-6 As we can see – the only value of those pages is big amount of friends. The difference just in number of friends. Biggest number of friends has cluster #6, it's a set of people oriented on just adding to their friend list every man in the world, so it is not informative for us. Cluster #5 has smaller amount of friends comparing to #6:

undefined on extraversion (big amount of friends, but small amount of photos); #5 – low on conscientiousness (much smaller amount of albums comparing to amount of photos), #6 – general; general on neuroticism (similar amount of photos and videos, so we can't say that they are posting mostly photos); #6 could be high on agreeableness (big amount of likes), #5 - general

8. Clusters #11-12 – oriented on videos count (up to 5 times more videos then general):

General on extraversion (general amount of friends and photos); general on conscientiousness (similar amount of photos and albums); low on neuroticism (big amount of videos); general on agreeableness

9. Clusters #17-18 – big amount of profile photos:

Possibly high on extraversion (general count of friends but big amount of profile photos); #17 – low on conscientiousness (smaller amount of albums comparing to amount of photos), #18 – general on this trait; general on neuroticism (because they also have almost average amount of videos); #17 - general on agreeableness, #18 – low on this trait (smaller than average amount of likes and small amount of comments)

VALIDATION

We have conducted a micro-survey, where 10 volunteers passed a test for defining their five personality traits. The result is a vector for each person containing five values expressing a high, low or undefined score on some specific trait (for example, [-1, -1, 0, 1, 1]). Also we`ve included their profile in the set of data for users grouping in this paper, so we know to what cluster they belong. Knowing this and their real personality traits, we have check is this grouping is correct.

Validating of groups by activity has no reason because it is obviously that some profile more oriented for photos, and another is oriented on big number of friends. Validating of grouping by personality traits we made checking the predefined value of trait with real. It shows next results: 27.5% of traits are match with real, 40% - there is small deviation, and 32.5% of traits are wrong. We therefore assume our validation is inconclusive and needs further enhancement.

CONCLUSIONS

In this work, we analyzed methodologies for defining personal psychological characteristics. For classifying users by psychological traits we chose the Big Five Model and hierarchical clustering with Ward distance. We extracted and downloaded 1000 user profiles. Their analysis shows that we can identify 9 groups of users by activity and 15 groups by psychological characteristics.

Future work will consist in better analysis by extending the range of data used in this paper (more user profiles and other information from the profile like posts' themes, define closest friends; try to define theme of photos and so on). In addition, 10 people for validation is too small, therefore we should extend validation to a larger panel of people for their real traits for calibration of model. Besides, it is unsure how to validate the obtained data. Using this classification, we aim to find dependencies between user psychological characteristics and their friends' psychological characteristics.

This work highlights (although not proves in any way) how far very simple data analysis can go and infer very sensitive information from rather freely available information. This work obviously lacks experimental backing but it is a first step into creating awareness to the fact of how social networks can reveal "hidden" information of their users.

REFERENCES

[1] M. Grimes, "What's Your Digital DNA? MasterCard Study Reveals Five Global Online Personas," Mastercard, 2 October 2013. [Online]. Available: http://newsroom.mastercard.com/press-releases/whats-your-digital-dna-mastercard-study-reveals-five-global-online-personas/.

[2] Ranjan, "Top 8 Types Of Internet Users In The World," All out digital, 18 November 2012. [Online]. Available: http://www.alloutdigital.com/2012/ 11/top-8-types-of-internet-users-in-the-world/.

[3] A. Asuero, "[INFOGRAPHIC] 7 types of internet users you encounter when using an Enterprise Social Network," Zyncro, 17 May 2013. [Online]. Available: http://en.blog.zyncro.com/2013/05/17/infographic-7-types-of-internet-users-you-encounter-when-using-an-enterprise-social-network/

[4] G. Espino, "Did you know that there are 5 types of Internet Users?," LoadMeNa.com, 4 May 2016. [Online]. Available: http://www.slideshare.net/GenesisEspino/did-you-know-that-there-are-5-types-of-internet-users.

[5] "Four temperaments," Wikipedia, 21 November 2002. [Online]. Available: https://en.wikipedia.org/wiki/Four_temperaments.

[6] "Big Five personality traits," Wikipedia, 8 Mar 2014. [Online]. Available: https://en.wikipedia.org/wiki/Big_Five_personality_traits. [Accessed 21 Nov 2016].

[7] "Narcissism," Wikipedia, 8 February 2017. [Online]. Available: https://en.wikipedia.org/wiki/Narcissism

[8] S. K. W. Ph.D., "What is Your Facebook Personality?," Psychology Today, 20 Dec 2011. [Online]. Available: https://www.psychologytoday.com/blog/fulfillment-any-age/201112/what-is-your-facebook-personality. [Accessed 20 November 2016].

[9] A. Morin, "What Your Facebook Use Reveals About Your Personality And Your Self-Esteem," Forbes, 31 October 2014. [Online]. Available: http://www.forbes.com/sites/amymorin/2014/10/31/what-your-facebook-use-reveals-about-your-personality-and-your-self-esteem/#38b5c6ab5bda. [Accessed 16 November 2016].

[10] G. S. Ph.D., "What Can You Learn About People from Facebook?," Psychology Today, 02 Jul 2015. [Online]. Available: https://www.psychologytoday.com/blog/close-encounters/201507/what-can-you-learn-about-people-facebook. [Accessed 20 Nov 2016].

[11] A. Blaszczak-Boxe, "What Your Facebook Photos Say About Your Personality," Live science, 5 Aug 2014. [Online]. Available: http://www.livescience.com/47191-facebook-photos-reveal-personality-traits.html. [Accessed 20 Nov 2016].

[12] facebook-team, "Graph-API references," Facebook, 2016. [Online]. Available: https://developers.facebook.com/docs/graph-api/reference.

[13] "Open API," VK, 2017. [Online]. Available: https://vk.com/dev/openapi.

[14] "scipy.cluster.hierarchy.linkage," SciPy.org, 19 September 2016. [Online]. Available: https://docs.scipy.org/doc/scipy-0.18.1/reference/generated/scipy.cluster.hierarchy.linkage.html#scipy.cluster.hierarchy.linkage.

[15] "scipy.spatial.distance.pdist," SciPy.org, 11 May 2014. [Online]. Available: https://docs.scipy.org/doc/scipy-0.14.0/reference/generated/scipy.spatial.distance.pdist.html.