The Demographics of Social Media Users in the Russian-Language Internet

Sergey Vinogradov, Vera Danilova, Alexander Troussov and Sergey Maruev
International laboratory for mathematical modelling of social networks,
RANEPA,
Moscow, Russia
e-mails: derbosebar@gmail.com, maolve@gmail.com, troussov@gmail.com, Maruev@ranepa.ru

***Abstract*—Our study focuses on the demography of the largest European social network VK and the representativeness of VK population sample with respect to the real-world state demography. The relationships between the variables, such as region code, settlement type, age and gender are explored. A special-purpose tool has been developed for ethnic group labeling purposes, which performs the classification given the user forename, patronymic and/or surname and ensures 99.2% accuracy. The analysis of the considered variables is helpful in finding a solution to the cold start problem in recommender systems.**

Keywords- internet sociology; internet demography; internet surveys; social network analysis.

I. INTRODUCTION

Scientists and politicians are concerned about the accuracy of traditional sociology's methods (questionnaires). Traditional polls are no longer a gold standard for the evaluation of politicians activity, elections outcome prediction and studies of voters' preferences. Predictions failed for the US 2016 elections, events in Israel, U.K. and Greece last year, Scottish independence referendum and US congress elections of 2014 [2]. As for the Internet surveys, the initial confidence in the accuracy of their methods was replaced by doubts about their consistency, mainly because (i) Internet demography differs from the real-world one, and (ii) demographies of different Internet segments vary significantly from one another, e.g., as of 2008, the Facebook audience was dominated by young white people from middle-class families, while MySpace was represented mainly by Afro-Americans [4]. Taking into account the discrepancies between the Internet demography and the real-world demography allows us to perform a more efficient study of potential customers, voters, patients and others on the Web and to apply the results to real-world situations.

Our study considers 239.044.903 user profiles of the largest European online social networking service VKontakte (VK) [28], whose average daily audience equals to 64.525.950. The study includes the following stages: (i) data collection via VK Open Application Programming Interface (API) [29]; (ii) filtering; (iii) user location identification using the Federal Information Address System (FIAS); (iv) analyzing the relationship between the variables: region code, settlement type, age and gender; (v) developing the ethnic group classifier for further use of the corresponding variable.

Current results include 1) the analysis of the age distribution of VK users across regions and settlement types, and 2) the analysis of the gender distribution of VK users across regions and settlement types. A tool for ethnic group labeling of target users has been developed and validated, so that the ethnic group feature can be further processed. An important application of our study of Social Network (SN) users demographics is to improve user profiling and mitigate the problem of the cold start in recommender systems (how to recommend a movie to see or a product to buy to new users with no history of activities).

The paper is organized as follows. Section II overviews the state of the art in the domain. Section III considers data gathering, filtering and description. In Section IV, the analysis of the obtained data is covered. Section V outlines the tool for ethnic group labeling and Section VI concludes the paper.

II. RELATED WORK

Among papers dealing with the study of Facebook and other English-speaking SN population are the works by 1) J. Chang et al. [5] on the study of ethnic groups represented in the SN, 2) P. Corbett [6] on Facebook demographic groups, 3) T. Strayhorn [23] on gender differences in the use of Facebook and Myspace by first-year college students.

Twitter data has been widely used for quantitative measurement and prediction of real-world events including prediction of the price level in the stock market [3], sentiment analysis of brands [21], prediction of movie revenues [1], earthquake prediction [20], and prediction of election outcomes [14] [27]. A large number of these predictions turned out to be inaccurate, which leads to the following questions: (i) Is Twitter population representative enough? (ii) In case it is, in which segments? Since Twitter-based predictions are considered to allow for a better understanding of real-world events and phenomena, Twitter sample should be representative to a certain extent: e.g., young people tweet far more often than older people, therefore, the corresponding part of the population is better represented. In any case, SN data analysis can be fruitful only if the SN demography data is approached scientifically.

One of the first papers on Twitter demography published by A. Mislove et al. [13] asked the following questions: (i) Who are Twitter users? (ii) To which extent the Twitter profile collection is representative of the whole population of the country? The authors undertook the first steps to explore

 ISBN: 978-1-61208-563-0

the Twitter dataset that covers over 1% of the US population. A comparative study of the Twitter population and the real-world one was conducted based on 3 criteria: location, gender, race. It was shown that the SN population is non-uniformly distributed: densely populated regions are represented redundantly, while underpopulated regions representation is scarce. Moreover, the male population prevails on Twitter and results in a non-random sample.

Twitter has been selected as an example of social platform, because, as opposed to the users of other SNs, 91% of Twitter users made their profile and communication history visible to non-registered users. At the moment, Twitter data use is the only possibility to perform a free large-scale study of particular features of the communication and information exchange. The investigation of the Twitter platform started quite a long time ago and is being considered promising, however, there remain questions on the sufficiency of the representation of demographic groups in the sample. To compare Twitter data to the US Census 2000 data, users were mapped to their location (districts).

In [22], it was found out that most UK Twitter users prefer to indicate hobbies instead of their job. The authors explain it by the fact that the creative sector on Twitter is represented by an excessive number of users, as compared to the Census 2011 data. The tool for the automatic extraction of data on user age, developed by the authors, allowed to find out that, according to proportions, Twitter is dominated by young population as opposed to the Census 2011 data. However, in view of the predicted values, there is a significant number of older twitters. The study has shown that there is a way to extract the data on age and occupation of Twitter users from metadata with varying accuracy that depends particularly on professional groups.

III. DATA

A. Collection

For the purposes of our study, 239 044 903 profiles from VK social network are considered. The data is accessed via VK Open API and collected using an in-house Python-based crawler. The gathered data is stored as JavaScript Object Notation (JSON) in a Hadoop cluster (HDFS) [11]. The processing is done using Apache Spark framework for in-memory cluster computing together with Hive data warehouse software [12].

B. Filtering

The following account types are excluded to sort out reliable data on the population of the considered social network: groups, deactivated accounts, fake accounts, and profiles, where no current city and/or no date of birth are indicated.

C. Description

As a result of data collection and filtering, two arrays are built. The first array in Comma-Separated Values (CSV) format contains 29.053 lines with the following information on VK population: location name, number of VK users, number of women among VK users, number of men among VK users, region code, location type (town, settlement or other). The second array in CSV format contains 1.048.576 lines with the following information on VK population: location name, age, number of VK users of this age, region code, location type.

The data was being collected during December 2015 for 83 regions, including the Moscow Region. The city of Moscow will be considered independently for comparison purposes in our future work. The information on user location is extracted from the ”Current city” field in the general information section of user personal profiles. In the present work, we perform a comparison with the real demography (data from the official Federal State Statistics Service (Rosstat) statistics as of January 2015).

Each of the resulting nodes (non-fake profiles) is described by certain attributes, such as age, gender, settlement type, etc., that differ in format and dimensions. The majority of these attributes can be used for opinion mining, decision making in recommender systems, etc.. The main questions are: (i) how to evaluate the quality of these attributes? (ii) how to efficiently use them? To answer these questions, it is crucial to use the topology of the network where the given nodes live and communicate. We build a model, which, however imperfect, provides a high-resolution picture of the behavior of a large number of nodes (people). To evaluate the quality of attributes, the preferential connectivity coefficient is estimated: if nodes having a certain attribute connect more often, this attribute is considered useful. Attributes can be effectively used to study the structure of massive networks as it is shown in [24], where a method is introduced that automatically creates a network of overlapping clusters, from the largest to the smallest, up to a threshold of term frequencies that is used to detect the similarity of interests.

D. Filtering

The following account types are excluded to sort out reliable data on the population of the considered social network: groups, deactivated accounts, fake accounts, and profiles, where no current city and/or no date of birth are indicated.

IV. DATA ANALYSIS

The analysis is based on two types of the collected data: 1) gender distribution of users across regions and settlement types (49.334.562 users) and 2) age distribution of users across regions and settlement types (41.236.001 users). The information on the location of users is aligned with the FIAS data. The parameters of the above samples have the following differences. The first sample (gender distribution) includes the total number of users, the number of men, the number of women, the number of the region according to the Constitution and the settlement type corresponding to each of the settlements, where 83 regions and 35 settlement types are covered. The sample does not include the city of Moscow, Republic of Crimea and Sevastopol. The sample includes the city of Baikonur - one of the territories serviced by the Administration of secure facilities under the Ministry of Internal Affairs (No. 94 according to the Constitution and

 ISBN: 978-1-61208-563-0

No. 99 in our database). The second sample includes such parameters as age, the number of users of a given age, the number of the region according to the Constitution and the settlement type corresponding to each of the settlements, where 83 regions and 45 settlement types are covered. This sample does not include the city of Baikonur, the Republic of Crimea and Sevastopol.

The distribution of VK users among regions is presented in Figures 1 and 2. The plot is divided into two parts for visual clarity. The values are sorted from the smallest to the largest. The first plot is built in arithmetic scale (absolute values of the number of users), the second plot is built in logarithmic scale (the difference between the minimum and the maximum values in this area is over 8 millions). The fewest number of users (44) has been registered in the region defined by the number 99 in our database, which corresponds to Baikonur, a city leased from Kazakhstan until 2050. VK population under 10.000 has been registered in the following constituent regions: 33 - Vladimir Oblast (2148), 79 - Jewish Autonomous Oblast (2553), 6 - Republic of Ingushetia (6858).

The values change by 1-2 orders of magnitude in the range from 12.348 (87 - Chukotka Autonomous Okrug) to 262.389 (89 - Yamalo-Nenets Autonomous Okrug).

Figure 2 shows the demography of VK for the other 42 regions. The values change by one order of magnitude in the interval from 275.411 (40 - Kaluga Oblast) to 2.740.984 (66 - Sverdlovsk Oblast). The maximum value (8.377.856) is observed for Saint Petersburg (78).

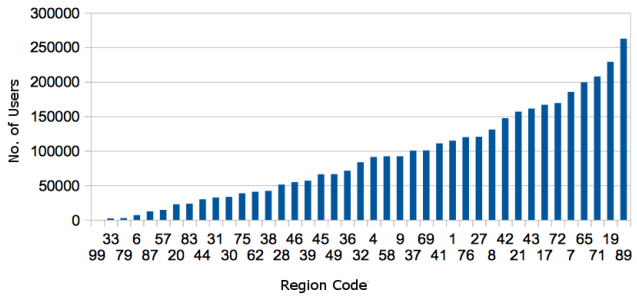

Figure 1. Distribution of the number of VK users among the constituent regions of the Russian Federation.

The real demography data has been taken from the Rosstat website [18]. The most recent statistics on the number of residents per region that can be accessed via the online Rosstat service reflects the situation as of January 1 of 2015, which allows us to perform the comparison with the data collected by our research group in December 2015. The results of the comparison of VK users distribution and the real Russian population distribution among the constituent regions are presented in Figures 3 and 4. These figures are parts of one plot. The first part describes data for the first 42 regions and the second part describes the remaining 40 regions, for a total of 82. The comparison is made for 82 regions, because Rosstat does not provide statistics for leased territories (Baikonur city). Moreover, the number of users in Baikonur is low (44). The comparison shows that, for most regions, the number of social network users is lower than the number of residents. For 30% of regions the number of registered users is the lowest (less than 10% of residents), as it is shown in Table I.

Furthermore, a relatively low number of VK users (under 20%) as compared to the number of residents is observed in the following constituent regions, as shown in Table II:

TABLE I. PERCENTAGE OF REGISTERED USERS PER REGION (1)

Region No.	Region Name	Percentage, %
33	Vladimir Oblast	0.15
6	Republic of Ingushetia	1.48
79	Jewish Autonomous Oblast	1.52
20	Chechen Republic	1.65
38	Irkutsk Oblast	1.74
57	Oryol Oblast	1.88
31	Belgorod Oblast	2.09
36	Voronezh Oblast	3.05
30	Astrakhan Oblast	3.23
75	Chita Oblast	3.54
62	Ryazan Oblast	3.59
44	Kostroma Oblast	4.56
72	Tyumen Oblast	4.72
46	Kursk Oblast	4.88
42	Kemerovo Oblast	5.40
39	Kaliningrad Oblast	5.86
28	Amur Oblast	6.31
32	Bryansk Oblast	6.77
58	Penza Oblast	6.79
45	Kurgan Oblast	7.59
69	Tver Oblast	7.65
27	Khabarovsk Krai	8.99
76	Yaroslavl Oblast	9.42
37	Ivanovo Oblast	9.67
24	Krasnoyarsk Krai	9.94

TABLE II. PERCENTAGE OF REGISTERED USERS PER REGION (2)

Region No.	Region Name	Percentage, %
43	Kirov Oblast	12.36
50	Moscow Oblast	12.56
21	Chuvash Republic	12.66
71	Tula Oblast	13.72
16	Tatarstan Republic	15.75
2	Bashkortostan Republic	16.28
5	Republic of Dagestan	17.05
9	Karachay-Cherkess Republic	19.64

 ISBN: 978-1-61208-563-0

The number of users in Moscow Oblast is quite low, according to the gathered data, probably because Moscow Oblast residents work and spend most of their time in the city of Moscow, so they prefer to specify ”Moscow” as their current city. The situation will become clearer when we analyze the VK population in the city of Moscow.

In the following regions, the majority of residents have VK profiles: 3 - Republic of Buryatia (89.82%), 10 - Republic of Karelia (83.78%), 18 - Udmurt Republic (86.64%), 48 - Lipetsk Oblast (77.02%), 54 - Novosibirsk Oblast (75.83%).

The above considered regions, where we observe the excess of real demography statistics values, are situated in the Northwestern Federal District. Vologda Oblast is a highly industrialized region, where one of the largest metallurgical plants of the country (Severstal) is located. The metallurgical industry is followed by chemical, food (center of butter industry), timber and machine building industries. Also, salt production, glass making and textile industries are well developed.

Figure 2 shows the demography of VK for the other 42 regions. The values change by one order of magnitude in the interval from 275.411 (40 - Kaluga Oblast) to 2.740.984 (66 - Sverdlovsk Oblast). The maximum value (8.377.856) is observed for Saint Petersburg (78).

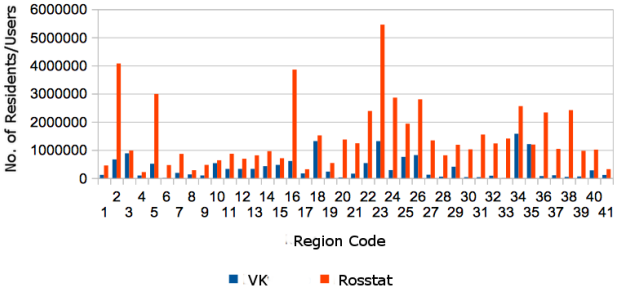

Figure 2. Comparison of the distribution of VK population through regions to the official Rosstat statistics as of January 1st, 2015 (1).

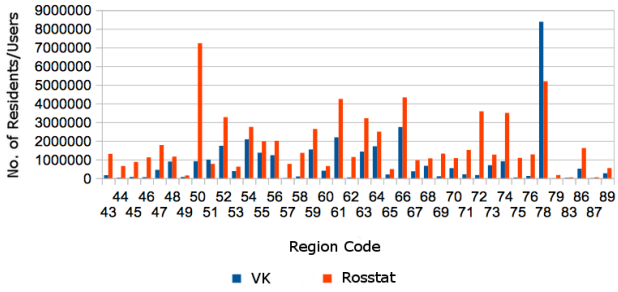

Figure 3. Comparison of the distribution of VK population through regions to the official Rosstat statistics as of January 1st, 2015 (2).

Murmansk Oblast is the home of one of the largest Russian ports; therefore, the fishing industry is well developed here. These regions afford good job opportunities, they are always in need of qualified human resources. Having these circumstances in mind, the excess in the number of users compared to that of real residents can be partially explained by the possible recent movement of people corresponding to the indicated percentage of VK users to their new workplaces, which is why they have not yet been officially registered. As for Saint Petersburg, these users can be among those, who reside in this constituent region without a temporary residence permit nor an official workplace. Moreover, residents of Leningrad Oblast may have indicated ”Saint Petersburg” in their current city field, because most of them work and spend most of their time there. Also, users may have selected Saint Peterburg, because they plan to move there in the nearest future and/or they do it on purpose to attract attention from other users. The distribution of VK users according to the settlement (locality) type is presented in Figures 5, 6, 7 and 8. Figure 5 shows the extent to which urban population prevails over that of other settlement types (46.803.588 citizens against 2.530.974 (5.13%) in rural settlements, Cossack villages, mountain villages and other).

In Figure 6, the VK population in a range of settlements is presented, where cities and settlement types with population under one thousand users are excluded for visual clarity. The standard notation for settlement types is transliterated (sl - sloboda, kp - health resort settlement, gorodok - community area, st and zd/d st - train station, aul - mountain village, u - ulus (administrative division of the Sakha Republic), np - inhabited locality (settlement), kh - khutor (a type of rural locality in Eastern Europe), st-tsa - Cossack village, rp - workers’ settlement, p - settlement, pgt - small town, d - small village, s - big village).

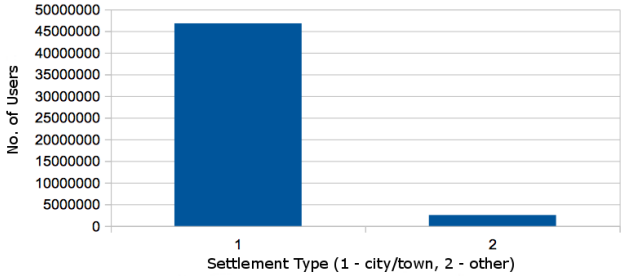

Figure 4. VK users distribution according to settlement type.

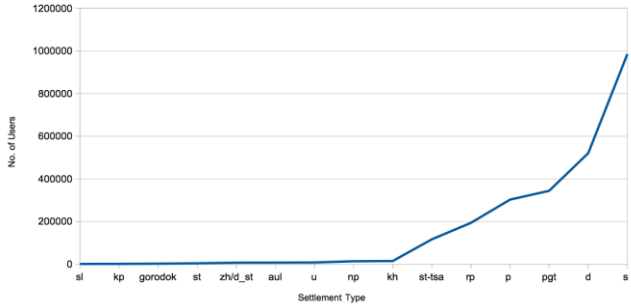

Figure 5. VK social network population distribution across different settlement types.

Population values for settlement types between sloboda and khutor change smoothly from 1427 to 15113. In the khutor - big village, the interval population values change drastically from 15113 to 985069.

 ISBN: 978-1-61208-563-0

Figures 7 and 8 show VK population values for all of the considered settlement types in logarithmic scale, organized from the smallest to the largest. The standard notation of settlement types is used (the list of settlement types can be found in [19]. The notation is as follows: s/a - rural administration, zh/d op - railway station, sh - highway, zh/d post - railway post, zh/d rzd - railway junction, s/o - rural okrug, r-n - district, dp - suburban settlement, s/s - rural council, p/o - post office, p/st - settlement near a railway station. The lowest population has been registered for the following settlement types: s/a - 1 user, zh/d op - 3 users, zh/d post - 4 users, sh - 4 users. In the s/a - p/o interval the population grows by 3 orders of magnitude to 628 users (p/o).

Furthermore, in the p/st - bv range it changes by 3 orders of magnitude from 675 to 985069 users. So, the leading settlement types as to the number of VK users are (in descending order) city, big village, small village, small town, settlement, workers' settlement, Cossack village.

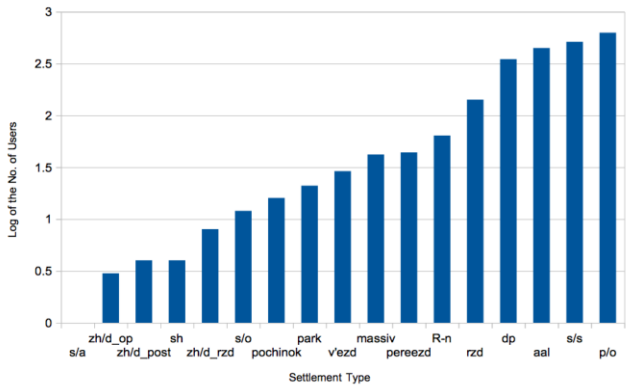

Figure 6. Quantitative distribution of VK social network users according to the settlement type (1).

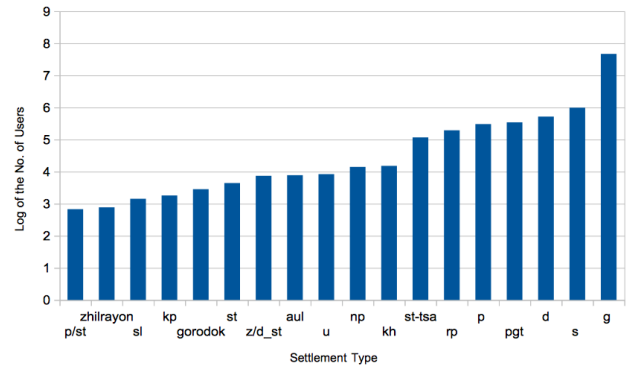

Figure 7. Quantitative distribution of VK social network users according to the settlement type (2).

A. Gender

Gender distribution of VK users across regions is depicted in Figures 9 and 10 (percentage ratio). The number of men is marked with blue color, the number of women - with orange color. According to the obtained data, in 19 out of 82 regions most users indicated gender in their profiles (Table III).

TABLE III. PERCENTAGE OF THE USERS THAT INDICATED GENDER IN THEIR PROFILE

Region No.	Region Name	Percentage, %
3	Republic of Buryatia	75.34
48	Lipetsk Oblast	76.05
64	Saratov Oblast	76.13
56	Orenburg Oblast	76.28
18	Udmurt Oblast	76.89
68	Tambov Oblast	77.54
51	Murmansk Oblast	77.83
15	North Osetia Republic	78.14
74	Chelyabinsk Oblast	78.31
63	Samara Oblast	79.35
24	Krasnoyarsk Oblast	85.45
35	Vologda Oblast	85.91
22	Altai Krai	86.05
42	Kemerovo Oblast	86.95
47	Leningrad Oblast	89.60
23	Krasnodar Krai	93.94
61	Rostov Oblast	95.78
50	Moscow Oblast	97.98
52	Nizhniy Novgorod Oblast	98.70

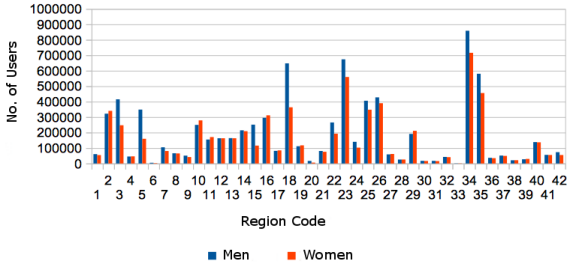

Figure 8. Gender distribution of VK social network users across regions (1).

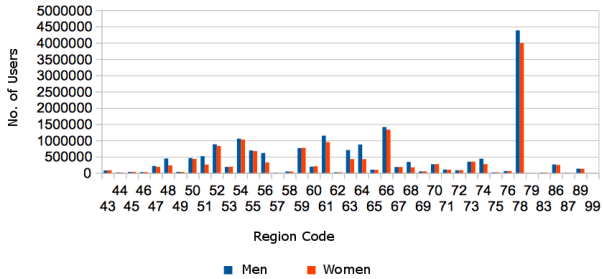

Figure 9. Gender distribution of VK social network users across regions (2).

In 40 of the considered regions, over 50% of users are men. Region 99 (Baikonur city) cannot be taken into account in the gender analysis, because only 44 users are registered there as of December 2015 and 30 of them are men. The lowest number of male users has been observed in Yamalo-Nenets Autonomous Okrug 83 (46.34%). Also, this value is lower in the following constituent regions: 6 - Republic of Ingushetia (68.14%), 5 - Republic of Dagestan (68.35%), 20 - Chechen Republic (74.03%). In 58 regions there are less than 50% of female users. The highest number of female users (53.65%) resides in Yamalo-Nenets Autonomous Okrug (83). In the following constituent regions women constitute less than 30% of users: 15 - North Osetia Republic (24.62%), 64 - Saratov Oblast (24.80%), 20 - Chechen Republic (25.29%), 51 - Murmansk Oblast (25.98%), 48 - Lipetsk Oblast (26.23%), 68 - Tambov Oblast (26.43%), 56 - Orenburg Oblast (26.46%), 18 - Udmurt Oblast (27.64%), 3 - Republic of Buryatia (28.12%), 74 - Chelyabinsk Oblast (29.79%), 63 - Samara Oblast (29.88%). According to the processed data on VK demography, the number of men in this social network is 3.970.836 higher than the number of women.

B. *Age*

The following account types are excluded to sort out reliable data on the population of the considered social network: groups, deactivated accounts, fake accounts, and profiles, where no current city and/or no date of birth are indicated.

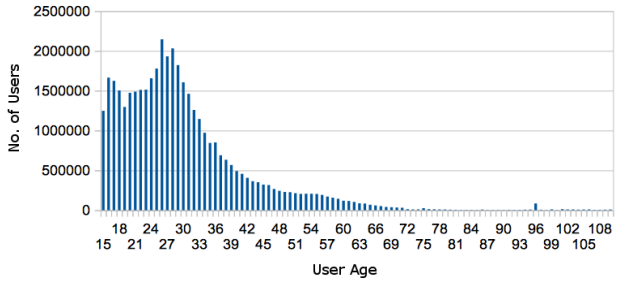

Figure 10. Age distribution of VK users across regions.

Age distribution of VK social network users across regions is shown in Figure 11. Most users are from 15 to 33 years old. The peaks are achieved at age 26 (2.145.219) and 28 (2.031.442). In what follows, the higher the age value, the lower the number of users. However, there is an unusual maximum at the age of 96. Most probably, users chose this age at random.

A comparison with Rosstat data has been conducted (see Figures 12 and 13).

It is important to note that the VK sample includes only those users whose profiles contain age data. The sum of all users that show (or fake) their age constitutes 41.236.001, while the sum of all Russian Federation residents with the age attribute recorded in Rosstat database is 121.861.482. The comparison shows that the main maximums stay close for both samples: the majority of users/residents are of ages 26-28. According to Rosstat calculations, the number of residents of age 27 is the highest (2.607.083). Also, there are over 2 million residents with the age values 23-42, 44 and 51-60 in Russian Federation. Only two age values exceed 2 million users (26 and 28 years old) and the age corresponding to the most VK users is 26 (2.145.219 users).

TABLE IV. EXCESS IN THE NUMBER OF VK USERS COMPARED TO THE OFFICIAL NUMBER OF RESIDENTS, ACCORDING TO ROSSTAT

Age	Excess, No. of People	Percentage, %
16	311.207	23%
17	307.609	23.4%
18	113.946	8.2%
96	67.320	406.87%
99	4.479	83.42%

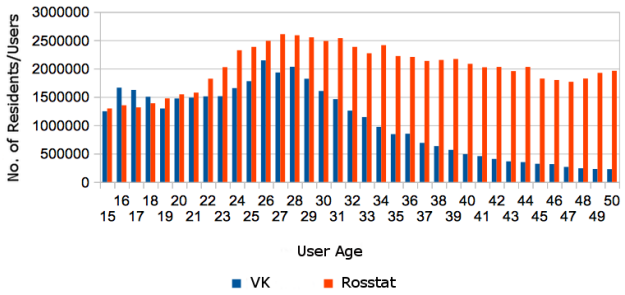

Figure 11. Comparison of VK and Rosstat data on age distribution (1).

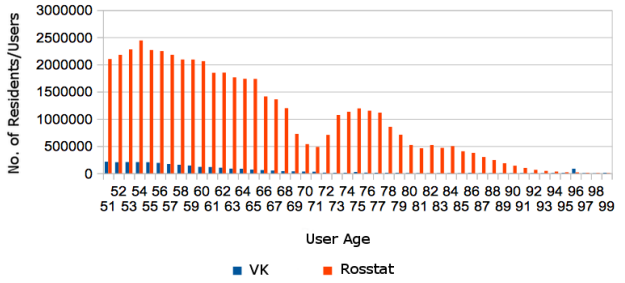

Figure 12. Comparison of VK and Rosstat data on age distribution (2).

Age distribution of rural and urban population, as well as the comparison of VK and official Rosstat statistics is presented in Figures 14, 15, 16 and 17.

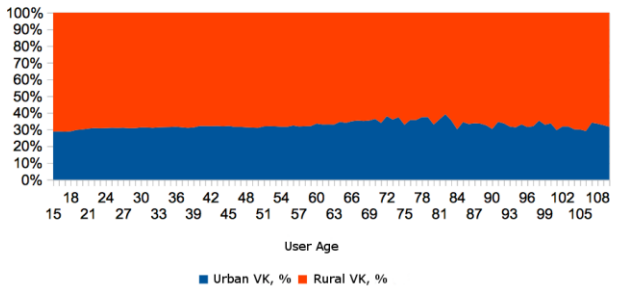

Figure 13. Age percentage distribution of urban and rural VK population.

According to the processed data for the second sample, the amount of urban VK population is significantly lower than that of rural population (Figure 14). In general, urban population equals to 30.81% and rural population - to 69.19% of the total population (41.236.001 users). Values are in the range from 28.67% (17 years) to 39.21% (82 years), the average value is 32.51%. It turns out that the least number of users of age 17 live in towns/cities and over 40% of users at the age of 82 are urban citizens as well. Surprisingly, the same ratio is observed for all of the considered ages (15-110). Since we have been expecting the opposite results, there will be another iteration of data collection process to build a new age distribution across different settlement types. A detailed examination of the sample gives us an idea on the origins of these results. Only 316 out of 1114 towns and cities of the Russian Federation (data as of January 1st, 2015[31]) form part of our sample. The complete list of towns and cities has been taken from [16]. When comparing the lists, we have found out that our automatically created list of town/city names includes wrong entries that probably refer to some areas/objects (district, street, metro station, etc.) inside a town/city or suburban area, such as ”Roshcha”, ”Malaya”, ”Yuzhnaya”, ”Yuzhny”, ”Chekhov-1”, ”Chekhov-4”, ”Chekhov-5”, ”Chekhov-6”, ”Chekhov-7”, ”Druzhba”, ”Kirova”, ”Krutaya”, ”Krutoy”, ”Lenina”, ”Nizhniy”, ”Nizhnyaya”, which do not correspond to any of the entries in the official list.

Figure 15 shows the age distribution of rural and urban population of the Russian Federation, according to the official Rosstat information, based on the data on 121.861.482 residents, which is 3 times higher than the considered number of VK users. The urban population of the Russian Federation is 2.95 higher than that of rural population as of January 1st, 2015. Rural population constitutes only 25.30% of the total population.

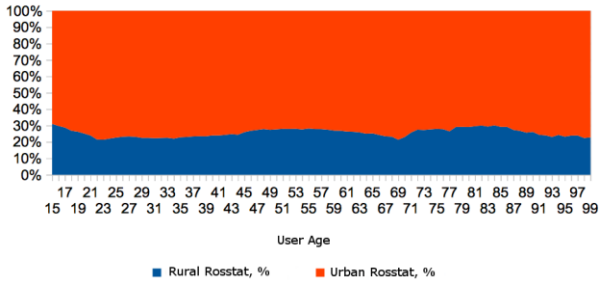

Figure 14. Age percentage distribution of urban and rural population, according to Rosstat.

Figure 16 shows age distribution of the Russian Federation rural population, according to VK API data as compared to Rosstat calculations. The number of users between 15 and 37 years residing in the rural area is higher than that of the registered rural residents in this age interval (almost 2 times higher in the 15-31 range). In what follows, as the age value increases, the number of users smoothly decreases in the range from 37 to 99 years, except the previously seen unusual peak at 96 years: the number of rural residents at the age of 96 (57.424) is significantly higher than that of the officially registered Russian Federation residents at this age (3.952).

Figure 17 depicts the comparative age distribution of the Russian Federation urban population, according to VK API data as compared to Rosstat calculations. As it has been noticed before, VK urban population sample is not representative enough, which is why another iteration of data collection and analysis is needed. On the basis of the processed data, we conclude that urban population of ages between 15 and 34 prevails (from 300.000). School and university students turn out to be the most active part of VK users: 909.412 users at the age between 15 and 25 years (in total) against 666.719 at the age between 26 and 99 years (in total). In the following, as the age grows, the VK urban population smoothly decreases (34-99 years), except the maximum at the age of 96, where the obtained value (26.424) is higher than the official statistics (12.594) as in the case of rural population.

While comparing VK and Rosstat plots, we have noticed that the largest number of users/residents is observed in the interval between 25 and 30 years in both cases: over 500.000 VK users between 24 and 30 years and around 2.000.000 officially registered residents between 27 and 29.

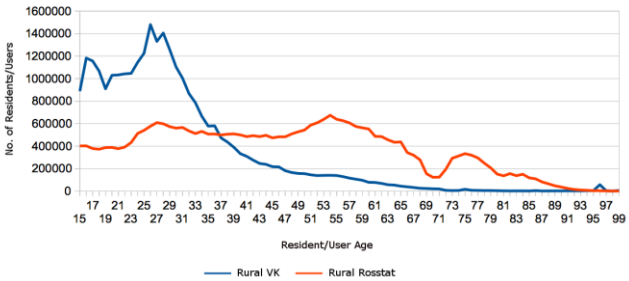

Figure 15. Comparative age distribution of rural population (VK and official Rosstat statistics).

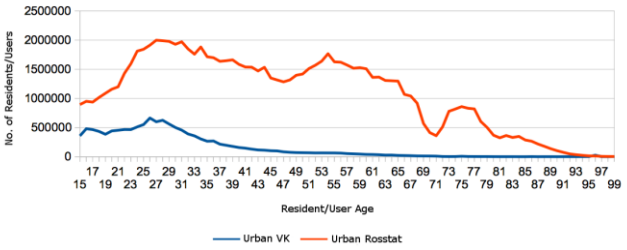

Figure 16. Comparative age distribution of urban population (VK and Rosstat).

V. ETHNIC GROUP LABELING

In the previous section, we considered VK demography variables, such as age, gender, region and settlement type, their relationships, and a comparison with the real-world demography has been conducted (Rosstat statistics). Another variable that is of key importance for sociological analysis (e.g., user behavior prediction) is the ethnic group. Due to the fact that this parameter is not indicated in SN profiles, our task is to devise an algorithm for the automatic

 ISBN: 978-1-61208-563-0

classification of users according to their ethnic group. We plan to compare the results with the official data on the ethnic composition of the Russian Federation [30].

Our tool for ethnic group labeling takes forename, patronymic, if any, and surname data as input and outputs the ethnic affiliation. It has been successfully tested on the lists of eminent people (e.g., Time's magazine list of 100 most important people of the 20th Century [26], famous Georgians [7], Ossetians [17], List of People's Artists of Azerbaijan [32], etc.). The average accuracy of the system is 99.2%. The core of the algorithm is a neural network trained on representative samples. The existing similar-purpose systems of the prior art are Onolytics [15], E-tech [8], EthnicSeer [10], Ethnea [9] and TextMap [25].

These systems use predefined hierarchies/group lists, however, no hierarchy/list construction standards are mentioned. Except for Wikipedia, no training resources are referenced in case of machine learning-based systems. Moreover, the hierarchies represent a mix of ethnic and religious (e.g., "Muslim" is listed as an ethnic group) groups. Russian ethnic diversity is not covered by these hierarchies/lists, therefore, a special-purpose structure has been developed. The details on the classification algorithm and comparison to the similar-purpose systems will be given in an upcoming paper.

VI. Conclusions And Future Work

This paper presents the first steps toward understanding the attributes of social media users. It includes the analysis of VK social network population sample and the comparison to Rosstat data. The attributes taken into account include region code, settlement type, age and gender. VK covers the following regions the best (in descending order): Saint Petersburg, Murmansk Oblast, Vologda Oblast, Republic of Buryatia, Republic of Karelia, Udmurt Republic, Lipetsk Oblast, Novosibirsk Oblast. According to the first sample that takes into account the gender feature, the urban VK population is 18 times larger than the rural one. The number of men in VK is 3.970.836 higher than the number of women. The majority of VK users indicated their gender in 19 out of 82 regions, including Murmansk Oblast (77.83%), Vologda Oblast (85.91%), Republic of Buryatia (75.34%), Udmurt Oblast (76.89%) and Lipetsk Oblast (76.05%) that are among the best covered by VK. According to the second sample including users that specified their age, most VK users are between 15 and 33 years old. The number of users who specified their age in VK is 3 times lower than the official number of Russian Federation residents. The number of users of age 16-18, 96 and 99 exceeds the official number of residents with these age values. 81.159 VK users introduced the age in the interval from 100 to 110 years. For these age values there are no official statistics. Also, in the age-oriented VK sample, the rural population is two times higher than the urban one.

The ethnic group feature will also be considered upon additional testing of the corresponding tool. Our current tasks include comparing the ethnic composition of VK with the official data on the Russian Federation ethnic composition and assessing the name feature quality (whether it is useful for the analysis of user groups behavior and preferences, and, consequently, whether profiles can be improved using kins and/or friends profiles data). Also, we are developing an approach to effectively assess the quality of other account attributes. It can be helpful in mitigating the cold start problem in recommender systems.

References

[1] S. Asur and B. Huberman, "Predicing the future with social media," 2010, URL: http://arxiv.org/abs/1003.5699 [retrieved: September, 2016].

[2] "Bloomberg: European Edition," URL: http://www.bloombergview.com/quicktake/perils-of-polling [retrieved: December, 2016].

[3] J. Bollen, H. Mao, and X.-J. Zeng, "Twitter mood predicts the stock market," in Proceedings of ICWSM, 2010, pp. 1-8.

[4] D. Boyd, "Taken out of context: American teen sociality in networked publics," PhD thesis, University of California, Berkeley, 2008, URL: http://oskicat.berkeley.edu/record=b18339028~S1 [retrieved: November, 2015].

[5] J. Chang, I. Rosenn, L. Backstrom, and C. Marlow, "ePluribus: Ethnicity on Social Networks," Artificial Intelligence, 2010, pp. 18-25.

[6] P. Corbett, "Facebook demographics and statistics report 2010," 2010, URL: http://www.istrategylabs.com/2010/01/facebook-demographics-and-statistics-report-010-145-growth- in-1-year [retrieved: June, 2016].

[7] "Famous Georgians," http://www.countries.ru/?pid=1950 [retrieved: November, 2016].

[8] "E-tech," URL:http://www.ethnictechnologies.com [retrieved: September, 2016].

[9] "Ethnea," URL:http://abel.lis.illinois.edu/cgi-bin/ethnea/search.py?Fname=kim&Lname=jung [retrieved: September, 2016].

[10] "EthnicSeer," URL:http://singularity.ist.psu.edu/ethnicity?name=kim+jung&commit=Analyze [retrieved: September, 2016].

[11] "HDFS Architecture Guide," URL: https://hadoop.apache.org/docs/r1.2. 1/hdfs design.html) [retrieved: June, 2016].

[12] "Hive on Spark: Getting Started - Apache Hive - Apache Software Foundation", URL:https://cwiki.apache.org/confluence/display/Hive/Hive+on+Spark%3A+Getting+Started [retrieved: June, 2016].

[13] A. Mislove, S. L. Jorgensen, Y-Y. Ahn, J-P. Onnela, and J. N. Rosenquist, "Understanding the demographics of Twitter users," in Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media, AAAI Press, 2011, pp. 554-557.

[14] B. O'Connor, R. Balasubramanyan, B. Routledge, and N. Smith, "From tweets to polls: Linking text sentiment to public opinion time series," in Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media, 2010, pp. 122-129.

[15] "Onolytics," URL:http://onolytics.com [retrieved: September, 2016]

[16] "Official Russian Cities List," URL: http://hramy.ru/regions/city abc. htm [retrieved: December, 2016].

[17] "Ossetians," URL: http://ossetians.com [retrieved: December, 2016].

[18] "Rosstat Federal State Statistics Service: Demography", URL: http://www.gks.ru/wps/wcm/connect/rosstat

main/rosstat/ru/statistics/ population/demography/ [retrieved: June, 2016].

[19] "Rosreestr. The Federal Service for State Registration, Cadastre and Cartography," URL: https://rosreestr.ru/upload/documenty/doc Pril 1 k P48.PDF [retrieved: June, 2016].

[20] T. Sakaki, M. Okazaki, and Y. Matsuo, "Earthquake shakes Twitter users: realtime event detection by social sensors," presented at WWW 2010, Raleigh, NC USA, 2010, pp. 851-860.

[21] M. Scarfi, "Social Media and the Big Data Explosion," in Forbes, 2012, URL:http://forbes.com/sites/onmarketing/2012/06/28/social-media-and-the-big-data-explosion/ [retrieved: September, 2016].

[22] L. Sloan, J. Morgan, P. Burnap, and M. Williams, "Who Tweets? Deriving the Demographic Characteristics of Age, Occupation and Social Class from Twitter User Meta-Data," PLoS ONE 10(3): e0115545, 2015, pp. 1-20, doi:10.1371/journal.pone.0115545.

[23] T. Strayhorn, "Sex differences in use of facebook and myspace among first-year college students," Stud. Affairs 10(2), 2009, URL: https: //www.studentaffairs.com/Customer-Content/www/CMS/files/Journal/Sex-Differences-in-Use-of-Facebook-and-MySpace.pdf[retrieved: June, 2016].

[24] A. Troussov, S. Maruev, S. Vinogradov, and M. Zhizhin, "Spreading Activation Connectivity Based Approach to Network Clustering," in Graph Theoretic Approaches for Analyzing Large-Scale Social Networks, N. Meghanathan (ed) IGI Global, 2017 (in print).

[25] "TextMap," URL:http://www.textmap.com/ethnicity/ September, 2016].

[26] "Time Magazine: 100 Most Important the 20th Century," URL: http://www.ranker.com/list/time-magazine-100-most-important-people-of-the-20th-century/theomanlenz [retrieved: December, 2016].

[27] A. Tumasjan, T. O. Sprenger, P. G. Sandner, and I. M. Welpe, "Predicting elections with Twitter: what 140 characters reveal about political sentiment," in Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media, 2010, pp. 178-185.

[28] "VK," URL: http://vk.com [retrieved: June, 2016].

[29] "VK Open API," URL: https://vk.com/dev/openapi [retrieved: June, 2016].

[30] "Wikipedia: Ethnic groups in Russia", URL: https://ru.wikipedia.org/ wiki/Ethnic groups in Russia [retrieved: June, 2016].

[31] "Wikipedia: List of cities and towns in Russia," URL:https://ru. wikipedia.org/wiki/List of cities and towns in Russia [retrieved: June, 2016].

[32] "Wikipedia: List of People's Artists of Azerbaijan," URL: https: //en.wikipedia.org/wiki/List of People%27s Artists of Azerbaijan [retrieved: December, 2016].

 ISBN: 978-1-61208-563-0