Digital Inequality in Russia through the Use of a Social Network Site: A Cross-Regional Comparison

Author 1, Author 2, Author 3

Abstract. An important role of digital inequality for hindering the development of civil society is being increasingly acknowledged. Simultaneously, differences in availability and the practices of use of social network sites (SNS) may be considered as major manifestations of such digital divide. While SNS are in principle highly convenient spaces for public discussion, lack of access or domination by socially insignificant small talk may indicate underdevelopment of the public sphere. At the same time, agenda differences between regions may signal about local problems. In this study we seek to find out whether regional digital divide exists in such a large country as Russia. We start from a theory of uneven modernization of Russia and use the data from its most popular SNS "VK.com" as a proxy for measuring digital inequality. By analyzing user activity data from a sample of 77,000 users and texts from a carefully selected subsample of 36,000 users we conclude that regional level explains an extremely small share of variance in the overall variation of behavioral user data. A notable exception is attention to the topics of Islam and Ukraine. However, our data reveal that historically geographical penetration of "VK.com" proceeded from the regions considered the most modernized to those considered the most traditional. This finding supports the theory of uneven modernization, but it also shows that digital inequality is subject to change with time.

Keywords: digital inequality, social network site use, online user behavior, topic modeling, Russian regions, VK.com

1 Introduction

An important role of digital inequality for hindering the development of civil society is being increasingly acknowledged. Simultaneously, differences in availability and the practices of use of the Internet and social network sites (SNS) may be considered as major manifestations of such digital divide. While SNS are in principle highly convenient spaces for public discussion, lack of access or domination by socially insignificant small talk may indicate underdevelopment of the public sphere. At the same time, agenda differences between regions may signal about local problems. In this study we seek to find out whether regional digital divide exists in such a large country as Russia. We start from a theory of uneven modernization of Russia and use the data from its most popular SNS "VK.com" as a proxy for measuring digital inequality.

2 Literature review

Related literature covers at least two research fields: (1) digital divide and (2) socio-economic development of Russian regions. The literature on digital divide discusses benefits and social implications of Internet use among different groups of population. Works about differences in development of the Russian regions can form a basis to explain contemporary digital inequality.

2.1 The digital divide

The main focus of the studies on digital divide concerns penetration and accessibility of the Internet among different groups of population. The initial assumption of these studies is that the access to the Internet and information is a valuable resource and an undoubted good expanding social opportunities and life chances of users compared to non-users. The digital divide is a factor of social inequality additional to traditional sources of inequality. The unequal Internet access depends on demographic and socio-oeconomic differences such as gender and age, income and education, race and ethnicity, location and type of settlement. According to the study of Pew Internet & American Life Project, the most accurate predictors of intensity and diversity of Internet use are levels of income and education [Witte, Mannon, 2010].

There are three hypothetical scenarios of social inequality implications under the influence of Internet use [Hargittai, Hsieh, 2013]: a) if Internet access is provided mainly for upper classes of society, social inequality grows ("the rich get richer" model); b) if equal Internet access is provided, inequality remains the same; c) if deprived and marginal groups of population benefit from the Internet use in the first place, inequality is reduced. Finally, the relation between socioeconomic status and access to the Internet is reciprocal, the traditional forms of inequality and digital inequality could strengthen each other [Van Dijk, 2005].

Early research used quite basic indicators of digital divide and limited them to the material dimension of Internet use: availability and quality if computer equipment, Internet access, signal speed and quality, number of places with Internet access, time spent on the Internet. When the Internet penetration in developed countries reached saturation, more refined indicators related to user qualities and behavior came to the forefront: user skills and abilities, use goals, topics of information search, etc. Furthermore, rates of social media content activity were used as an indicator of digital inequality and applied in the study of unevenness of urban space [Indaco, Manovich, 2016]. Thus, nowadays digital divide has evolved into a complex concept that includes at least two levels. The digital divide of the first level is connected with the material characteristics of access to ICT, the second level accounts for the characteristics of use (goals, skills, activity). Furthermore, the uneven access to users' attention is an additional aspect of second level digital divide [DiMaggio et al., 2001].

Approached from the digital divide perspective, two main dimensions of SNS use — contents and online user engagement — could be considered as indicators of digital inequality. Topic variation on the SNS could point at digital inequality because greater attention to socially significant topics in a particular region (such as human

rights, economy, social policy, housing and utilities, urban planning, etc.) indicates a stronger online public sphere and could lead to potential benefits for population, while greater presence of everyday topics (such as sports, celebrities, cars, cooking recipes, etc) might mean lack of SNS-mediated public sphere. Likewise, differences in online user engagement on the SNS could be an indicator of digital inequality because greater online activity and social involvement may lead to a larger amount of social capital and stronger communication power of the population [Castells 2013].

2.2 Socioeconomic development of Russian regions

According to [Auzan, Belyakov, 2011] Russia is a fundamentally segmented society, and one of the most well-known theories explaining unequal development of Russian regions is a so-called "Theory of four Russias" [Zubarevich, 2011]. Zubarevich distinguished four types of Russian settlements, that differ from each other in terms of population and socioeconomic modernization. The "First Russia" is represented by large cities with population over 500,000 and is characterized with high speed of the post-industrial transformation. The majority of Internet users are concentrated in these cities. The "Second Russia" can be found in industrial cities with population between 20,000 and 500,000. Inhabitants of these cities are employed at industrial, often state-owned companies; they keep struggling for economic well-being and are indifferent to the problems of the middle class. The "Third Russia" is conservative and passive population from rural periphery and small towns of most of the regions. They have the lowest level of education and mobility, are employed in the state-owned organizations and agriculture, and are completely focused on their own problems [Zubarevich, 2016]. Finally, Zubarevich singles out North Caucasus and southern Siberia republics and places them into the "Fourth Russia". The population of these regions is said to exercise traditional, pre-industrial lifestyles and often retreat to subsistence farming. The "Four Russias" theory is close to the more general theory of post-materialism [Inglehart, 2012], which argues that the growth of economic well-being causes transition to the so-called post-materialist values including self-expression and greater participation of the individual in public life.

Bodrunova and Litvinenko [2015] used Zubarevich's classification to analyze the fragmentation of communication in the Russian public sphere. The authors examined the contribution of online media to the split of Russian public sphere in the electoral cycle 2011-2012 and found that the split in the online media reproduces social divisions from the "Four Russias" theory. Thus, territorial differences have been related to the differences in media consumption. Authors assigned the main role in this split to the confrontation between the "First" and "Second Russia", which differ in their values, attitudes and behaviors. The "Third" and "Fourth Russia" are much less represented in the online media and the public sphere. However, as this elation has not been tested quantitatively neither in Bodrunova and Litvinenko [2015] nor in any other research on Russia, it needs further investigation.

3 Research Questions

This study seeks to find a meaningful regional variation in SNS use. As the latter is a complex phenomenon we split it into two dimensions regarding contents and formal online user behavior. Online content is characterized and measured through its topical structure. Topic variation as an important feature of the second level digital divide because it reflects uneven distribution of public attention towards different social issues, and uneven practice of use of an SNS as a media outlet in different Russian regions.

RQ1: What are the differences in representation of online content topics across Russian regions?

Formal online user behavior is represented as the aggregate of elementary user actions such as posting messages and comments, giving "likes", making "friends", reposting, etc. All these elementary actions form digital user biography and patterns of socially oriented using of an SNS. Differences in patterns of SNS use are also an important feature of the second level digital divide because they reflect different purposes of SNS use, and uneven user activity and social engagement.

RQ2: What are the differences in online social engagement of the SNS users across Russian regions?

4 Data and methods

In this research we use the data from the most popular Russian SNS "VK.com", a Russian replication of Facebook. At the time of the study "VK.com" had over 350 million registered users. Since the "VK.com" user ID is generated incrementally (the first registered is assigned ID 1, the second is 2, etc.), we were able to generate a random sample by selecting the required number of random numbers from a range of 1 to 350 million and download user information filtering out removed profiles. However, due to the extremely uneven distribution of users across regions, we refused from random sampling, because users from Moscow, accounted for almost a quarter of all registered accounts, and Saint Petersburg accounted for 11%, while some regions were represented by a tiny fraction. This did not fit our main goal of cross-regional comparison. To avoid such bias the upper limit of 1000 users per region was set.

The data was collected by using the "VkMiner" software developed in The Laboratory for Internet Studies. Only publicly available data on "VK.com" users were collected, such as information from the profile and records from the "walls". Some regions were excluded from the analysis because user data from these regions were incorrectly downloaded during the data collection. Thus, the final sample included 7,827,384 entries from the "walls" of 42,459 users from 69 out of 85 regions of the Russian Federation. They represent all "four Russias".

Our task of topic detection was solved with an approach known as topic modeling. We used the most popular topic modeling algorithm — latent Dirichlet allocation

(LDA) with Gibbs sampling, implemented in the TopicMiner[1] software. The output of the LDA consists of the matrix of probabilities of words in topics and the matrix of probabilities of topics in texts [Steyvers & Griffiths, 2006]. In other words, it is assumed that each document belongs to all topics and each word has the potential to generate any topic, but with highly different probabilities. Top words form interpretable sets that can be easily labeled by human analysts.

One of the research questions was the detection of topical profiles of users, but the data was collected from the "walls" of users. A “wall” is the main message board of the user’s profile and the place for public communication with user’s “friends”, followers and other visitors. “Wall” works as a personal media outlet, because posts from the “wall” appear in the news feed of the user’s “friends” and followers. Users can also make posts on each other's walls. So we decided to group the texts not by the "wall" but by the author. The preliminary selection of the texts thus included 5,392,586 entries from the "walls" of 74,303 users including those who posted only on the walls of others.

To reduce the dimensionality of the data and to improve the results of topic modeling we set a time limit for the texts. It reflects the need to take into account the fact that the topical profile of the author changes over time. It can occur as a result of changes in the author's preferences or as a result of changes in the policy of the SNS. In particular, since October 20, 2010 “VK.com” has changed its communication concept: the walls of users have ceased to be a place for communication, the focus of the service has been transferred from users' pages to the news feeds with updates of records, statuses and friends’ photos[2]. At the time a new functionality was introduced that has made it more convenient to receive news from public pages. Since then “VK.com” has become less focused on communication between private users. As the change in the technical features has definitely influenced patterns of social interaction, from the analysis the entries older on October 21, 2010 were excluded.

All texts were preprocessed with the TopicMiner software and passed through the following stages: 1) removing of HTML-tags, 2) tokenization, 3) lemmatization, and 4) stopwords removing. Thereby the dictionary of our text corpus comprises about 220,000 unique words.

The next modification of the original sample was performed due to difficulties of topic modeling of short texts from SNS. Since the topics are formed by words that often occur together within one document and in short texts too few words co-occur, modeling on these texts led to the emergence of uninterpretable topics.

There are several ways to deal with topic modeling of short texts that can be divided into two groups: 1) modification of the source data; 2) modification of the topic modeling algorithms. In the absence of ready-to-use implementations of topic modeling algorithms for short texts, the first approach was chosen. This group of methods is easier to implement, and therefore it is more popular [Weng et al., 2010; Hong, & Davison, 2010]. We, first, merged all texts of the same user, and, second, filtered out all users whose merged texts contained fewer than fifty words. As a result, the final

[1] https://linis.hse.ru/soft-linis
[2] https://vk.com/blog/blog154

sample consisted of texts from 36,396 authors. For a model with 150 topics it allowed us to achieve more than a hundred interpretable topics, which was much better than the result obtained with the initial texts.

Finally, we got rid of unstable topics in our model. The algorithms of topic modeling have a well-known limitation: each time they are run on the same texts and with the same parameters, slightly different topics are obtained. To fight this, we used normalized Kullback-Leibler similarity measure [Koltcov et al., 2014]. Thus we ran our algorithm three times and selected only the topics that appeared in all three runs. The two topics from different runs were considered identical if the similarity between them exceeded the 95% threshold. As a result, we obtained 33 stable topics almost all of which were easily interpreted.

5 Results

5.1 Topic modeling

After determining the stable topics, they were manually labeled. The following topics were identified: (1) Tips for cleaning and washing; (2) Recipes of baking; (3) Strength exercises; (4) Music; (5) Babies; (6) Automatically generated messages from "VK.com" applications; (7) Recipes of desserts; (8) Competitions and prizes; (9) War, the Great Patriotic War; (10) Obscene vocabulary; (11) Movies; (12) Football; (13) Musical events; (14) Patterns of behavior of men and women; (15) Beauty, make-up, manicure; (16) Automatically generated messages from "VK.com" applications; (17) City events; (18) Cars; (19) Russian-Ukrainian relations; (20) Tourism; (21) Christianity; (22) Physical exercises; (23) Recipes of main dishes; (24) Uninterpretable; (25) Automatically generated messages from games; (26) Recipes of main dishes; (27) Salad recipes; (28) Malady fundraising for children; (29) Search for hosts for stray animals; (30) Recipes of desserts; (31) Islam; (32) Gardening; (33) Knitting.

It can be clearly seen that the VK user-generated agenda is dominated by consumption, everyday small talk and private approach to problem-solving. Most of these topics are related to such everyday activities as games, listening to music, cooking, solving everyday problems. People use "VK.com" SNS mainly as a place where they can save interesting information about an unusual culinary recipe, recommendations for body care and garden management, select new films and music. This huge part of user' records is created with the instrumental aim to ensure quick access to potentially useful information. Another significant part of the texts is automatically generated by numerous applications, most often for advertising purposes.

Together with uninterpretable and non-subject topics, these topics centered on private life comprise the bulk of the content of "VK.com" SNS. At the same time, we have identified a small number of public affairs topics that are especially interesting for sociological analysis. Such topics are "Christianity", "Islam", "Ukraine-Russia relations", "City events" and "Malady fundraising for children". These topics touch socially significant and potentially problematic phenomena of the public life.

To account for regional differences, we have aggregated the Russian regions into seven Federal Districts (traditional semi-formal administrative division in Russia) and

calculated the average probability of topics in each of them. As the revealed topic composition has seemed gender-sensitive, we have done the same for male and female users (Fig. 1) As seen from it, the topic preferences of men and women are very different. In their posts users of "VK.com" reproduce the common gender stereotypes. The most typical topics for women are those about cooking recipes, children, needlework and beauty, while men engage in the talk about football, games, cars and politics. This finding can significantly supplement previous studies in this area [Moore, 1922; Bischoping, 1993]. In particular, Bischoping claims that "consistent patterns can be seen in the gender differences for most topic areas, with women holding the majority of conversations about people and relationships and appearances, and men typically holding the majority of conversations about work and money and issues". We confirm some of these findings — women indeed give more attention to talks about their appearance and men are more interested in topics about "issues" (war, politics). But we also identify other patterns of gender behaviour. According to our data, women talk a lot more about cooking and men take the lead in games, including football and computer games.

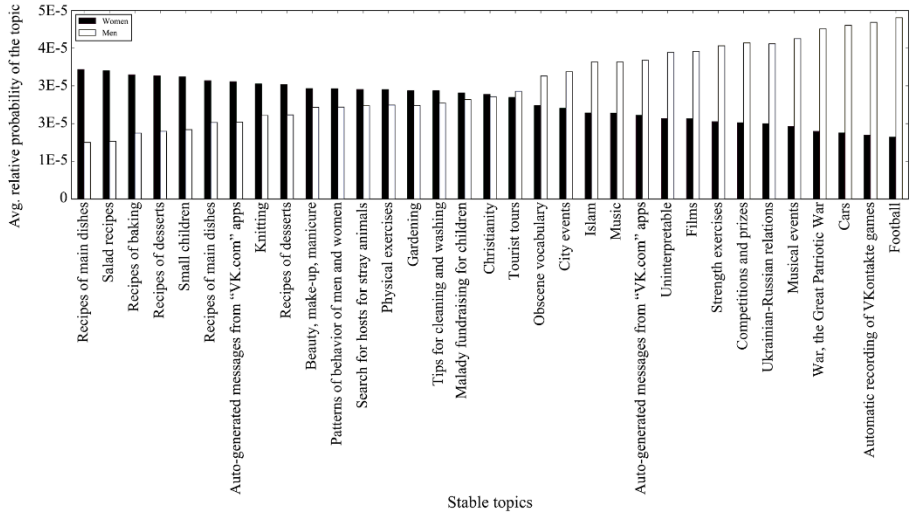

Fig. 1. Topic modeling results: distribution of topics by gender

Regarding regional differences, we have found that the most unevenly distributed topic is "Islam" (Fig. 2). It is much more pronounced in North Caucasian Federal District of Russia, which consists of republics whose population has traditionally practiced Islam — Dagestan, Ingushetia, Chechnya and others.

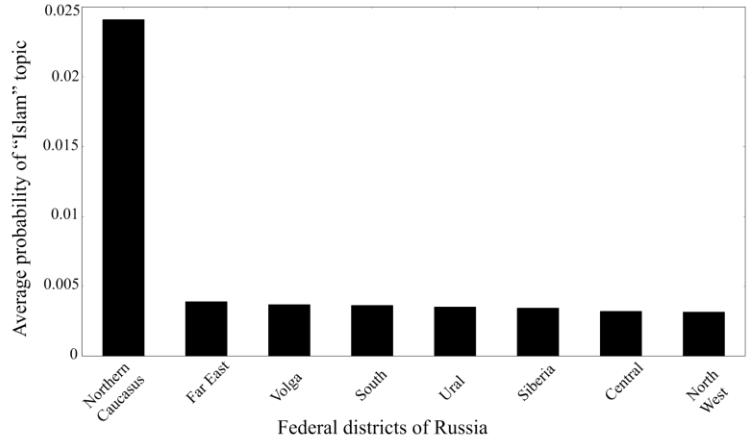

Fig. 2. Topic modeling results: distribution of "Islam" topics by Federal dis-tricts of Russia

Another interesting finding concerns topic called "Ukraine-Russia relations". It is more vivid in the Southern Federal District, that has a common border with Ukraine (Fig. 3). This may indicate the concern of residents of this border district with recent events in Ukraine. In particular, the regime change in Ukraine and subsequent military actions led to the influx of refugees into Russia.

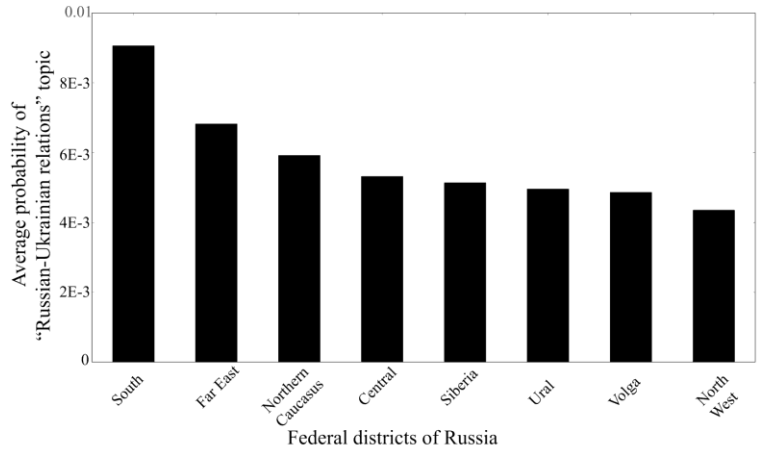

Fig. 3. Topic modeling results: distribution of "Ukraine-Russia relations" topic by Federal districts of Russia

5.2 Formal online user engagement

Distribution of user meta-data and indicators of online user activity are extremely uneven and right-skewed.

Table 1. Descriptive statistics of online user behavior indicators

		Min	Max	Mean	Std.dev	1st Q	Median	3rd Q
User meta-data								
Birth year	Birth year from a user profile	1902	2002	1986	13.18	1982	1989	1995
Friends	Number of user's "friends"	0	8895	45.8	178.44	0	1	35
Followers	Number of user's followers	0	18235	23.5	146.28	0	0	9
Groups	Number of groups joined by a user	0	4681	24.7	89.63	0	0	13
Duration	The duration of VK.com use computed as difference between dates of the first and the last user posts on the "wall"	0	3254	663.2	685.51	32	460	1072
Indicators of public user engagement								
Posts	Total number of posts on a user's "wall"	0	53285	113.04	585.45	0	3	33
Comments	Total number of comments on a user's "wall"	0	15513	11.20	115.05	0	0	0
Likes	Total number of "likes" on a user's "wall"	0	71304	181.49	985.12	0	1	28
In Reposts	Number of posts reposted by the user on the own "wall"	0	52957	68.18	502.26	0	0	3
Out Reposts	Number of posts reposted by other from the a user's "wall"	0	4374	8.60	57.61	0	0	0
Contributors	Number of unique users who contributed into activity on the a user's "wall"	0	13656	35.13	151.62	0	2	17
Sources	Number of unique IDs which posts were reposted by a user on his "wall" (diversity of sources)	0	3601	14.02	60.02	0	0	2
Originality	Share of original (authored by a user himself) posts among total number of posts on a user's "wall"	0	1	0.684	0.374	0,32	0.93	1
Other' Posts Share	Share of posts authored by other users among total number of posts	0	1	0.347	0.394	0	0.133	0.75
Other' Comments Share	Share of comments authored by other users among total number of comments	0	1	0.601	0.319	0,42	0.6	1

Other' Likes Share	Share of "likes" leaved by other users among total number of "likes"	0	1	0.831	0.242	0,76	0.94	1

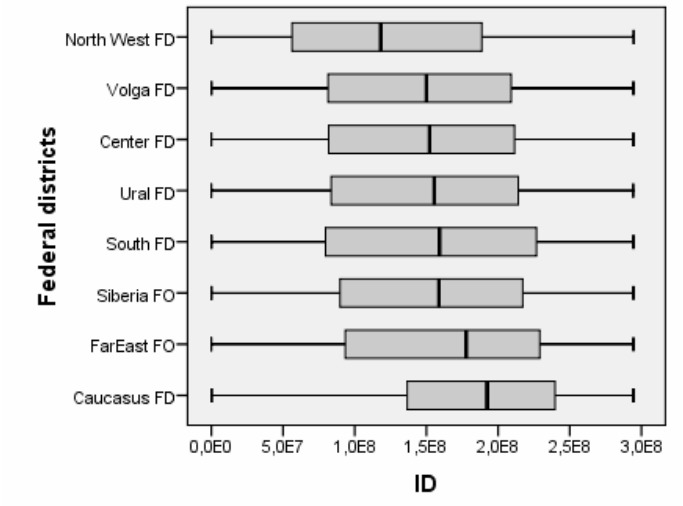

Fig. 4. Boxplots of user ID distribution among Federal districts of Russia.

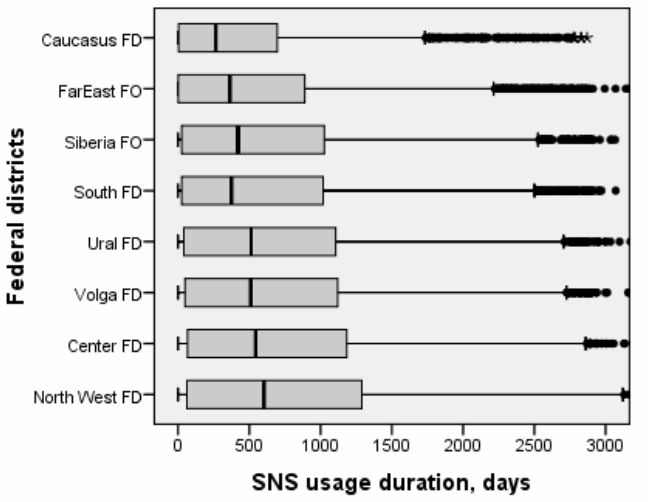

Fig. 5. Boxplots of SNS user duration distribution among Federal districts of Russia.

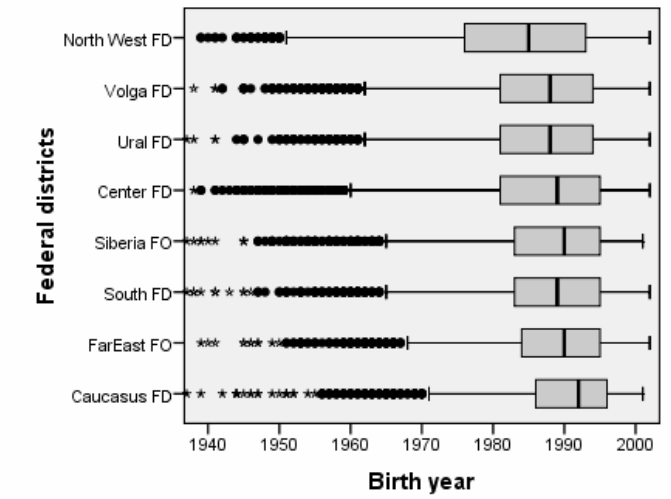

Fig. 6. Boxplots of users birth years distribution among Federal districts of Russia.

Boxplots on Figure 4 indicate that pioneer users of "VK.com" were inhabitants from North West Federal district. Users from the other regions registered later, and the latest were those from the republics of North Caucasus and Far East. These data reflect the process of "VK.com" penetration through the geographical space of Russia after being originated in Saint-Petersburg. The boxplots of SNS usage duration (Fig. 5) also illustrate this process. Boxplots from Figure 6 show the distribution of birth years of users among Russian regions and indicate the age of inhabitants from the North West region is slightly shifted to the older side. Such distribution could be explained with geographical paths and dynamics of SNS penetration and by the innovation diffusion theory [Wejnert, 2003]. According to the the latter technologies first spread among younger population and further are adopted by the older. Older population of the North West region had received an advantage comparing to their peers from other regions because of the earlier start of "VK.com".

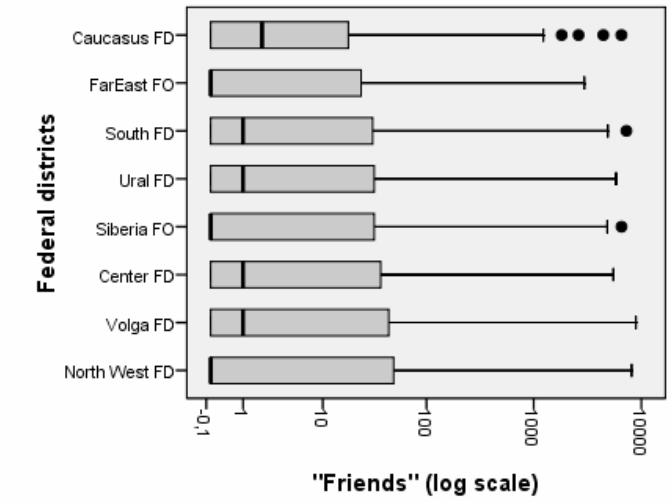

Fig. 7 Boxplots of user SNS "friends" distribution among Federal districts of Russia (logarithmic scale).

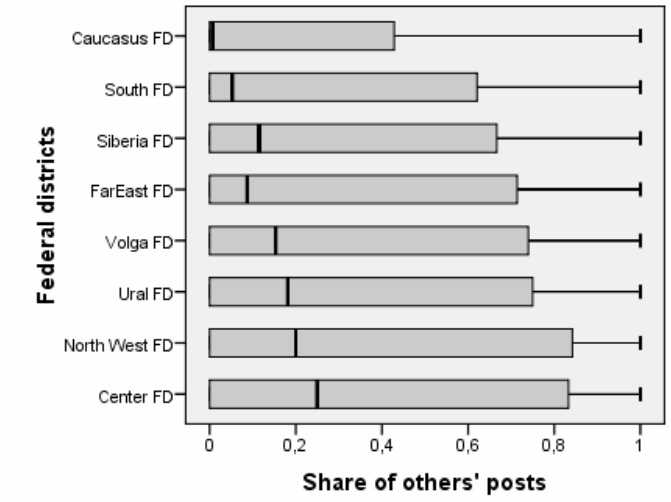

Fig. 8 Boxplots for share of others' posts distribution among Federal districts of Russia.

Boxplots from Figure 7 indicate differences, albeit not very pronounced, in the number of users' SNS "friends" among Russian Federal districts. The rank of each district roughly corresponds to the order of VK's geographical dissemination: users from North West regions tend to have more "friends" compared to all other regions, while users from republics of North Caucasus and Far East tend to have fewer SNS

"friends". Since the number of SNS "friends" could be a component of social network capital these differences may be interpreted as display of digital inequality,
The measure from Figure 8 — share of others' posts on a user's "wall" — indicates the extent of users' online involvement and public engagement with others. Boxplots show significant differences in the distribution of shares among Russian regions: Central and North West regions have the highest scores for online social engagement, while Caucasus Federal district has the lowest.
Other indicators of online social engagement did not indicate any significant or interpretable differences.

6 Conclusion

Obtained results contribute to the theory of the impact of geographic location on online communication and user behavior. First, differences in the overall activity of "VK.com" users across Russian regions are not significant in general. The regional level explains the extremely small share of variance in the overall variation of behavioral user data. This evidence is in favor of the geographical independence of online users engagement. A notable exception is attention to the topics of Islam and Ukraine. Second, distributions of user IDs, user age and duration of active SNS use reflect the natural process of SNS diffusion across Russia and the gradual penetration into various regions of the country. The dynamics and nature of SNS spreading can be explained with the digital divide theory [Deviatko, 2013] and the conception of the unequal modernization of Russia [Zubarevich, 2011]. The delayed penetration of the "VK.com" into the Far East and republics of the North Caucasus is explained by the poorer availability of the Internet and significant cultural differences. The delayed and weak representation of these regions in the online media space has been confirmed in other studies [Bodrunova, Litvinenko, 2015].

7 Aknowledgements

Withdrawn for anonymization purposes

8 References

1. Auzan, A., Belyakov, E.: Economist Aleksander Auzan: «Rossia prevraschaetsa v stranu menedzherov, ohrannikov, migrantov i pensionerov». Komsomolskaya pravda (2011)
2. Bischoping, K.: Gender differences in conversation topics, 1922–1990. Sex Roles. 28, 1–18 (1993)
3. Bodrunova, S. S., Litvinenko, A. A.: Four Russias in Communication: Fragmentation of the Russian Public Sphere in the 2010s. In: Dobek-Ostrowska, B., Glowacki, M. (eds.) Democracy and Media in Central and Eastern Europe 25

Years On. pp. 63-80. Peter Lang, London (2015) DOI:10.3726/978-3-653-04452-2
4. Castells, M.: Communication Power. Oxford University Press, Oxford (2013)
5. Deviatko, I. F.: Digitizing Russia: The Uneven Pace of Progress Toward ICT Equality. In: Ragnedda, M., Muschert, G.W. (eds.) The Digital Divide: The Internet and Social Inequality in International Perspective. pp. 118–133 Routledge, New York (2013)
6. DiMaggio, P., Hargittai, E., Neuman, W. R., Robinson, J. P.: Social Implications of the Internet. Annu. Rev. Sociol. 27, 307–336 (2001)
7. Hargittai, E., Hsieh, Y. P. Digital Inequality. In Dutton, W. H. (ed.) The Oxford Handbook of Internet Studies. Oxford University Press, Oxford (2013) DOI:10.1093/oxfordhb/9780199589074.013.0007
8. Hong, L., Davison, B.D.: Empirical Study of Topic Modeling in Twitter. In: Proceedings of the First Workshop on Social Media Analytics, pp. 80–88. ACM, New York (2010)
9. Indaco, A., Manovich, L.: Social Media Inequality: Definition, Measurements, and Application. Urban Studies and Practices, 1, 11-22 (2016)
10. Inglehart, R.: Modernization and Democracy. In: V. Inozemtsev, and P. Dutkiewicz (eds.) Democracy versus Modernization: A Dilemma for Russia and for the World. pp. 123–144. Routledge, New York (2012)
11. Koltcov, S., Koltsova, O., Nikolenko, S.: Latent dirichlet allocation: stability and applications to studies of user-generated content. In: Proceedings of the 2014 ACM conference on Web science (WebSci '14), pp. 161-165. ACM, New York (2014)
12. Moore, H. T.: Further data concerning sex differences. J. Abnorm. Psychol. 17(2), pp. 210–214 (1922)
13. Steyvers, M., Griffiths, T.: Probabilistic topic models. In Landauer, T., McNamara, D., Dennis, S., Kintsch, W. (eds.) Latent Semantic Analysis: A Road to Meaning. Lawrence Erlbaum (2006)
14. Van Dijk, J. A. M. G.: The Deepening Divide: Inequality in the Information Society. Sage, London (2005)
15. Wejnert, B.: Integrating Models of Diffusion of Innovations: A Conceptual Framework. Annu. Rev. Sociol. 28, 297–326 (2003)
16. Weng, J. et al.: TwitterRank: Finding Topic-sensitive Influential Twitterers. In: Proceedings of the Third ACM International Conference on Web Search and Data Mining, pp. 261–270. ACM, New York (2010)
17. Witte, J. C., Mannon, A. P.: The Internet and Social Inequalities. Routledge, New York (2010)
18. Zubarevich, N.: Perspectiva: Chetire Rossii. Vedomosti, 3014 (2011)
19. Zubarevich, N.; Chetire Rossii i novaya politicheskaya realnost'. Polit.ru (2016) http://polit.ru/article/2016/01/17/four_russians/