Schools are segregated by educational outcomes in the digital space
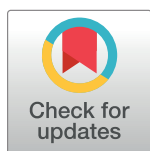
Ivan Smirnov [ORCID] *

Institute of Education, National Research University Higher School of Economics, Moscow, Russia

* ibsmirnov@hse.ru

Abstract

The Internet provides students with a unique opportunity to connect and maintain social ties with peers from other schools, irrespective of how far they are from each other. However, little is known about the real structure of such online relationships. In this paper, we investigate the structure of interschool friendship on a popular social networking site. We use data from 36, 951 students from 590 schools of a large European city. We find that the probability of a friendship tie between students from neighboring schools is high and that it decreases with the distance between schools following the power law. We also find that students are more likely to be connected if the educational outcomes of their schools are similar. We show that this fact is not a consequence of residential segregation. While high- and low-performing schools are evenly distributed across the city, this is not the case for the digital space, where schools turn out to be segregated by educational outcomes. There is no significant correlation between the educational outcomes of a school and its geographical neighbors; however, there is a strong correlation between the educational outcomes of a school and its digital neighbors. These results challenge the common assumption that the Internet is a borderless space, and may have important implications for the understanding of educational inequality in the digital age.

Introduction

The Internet creates unique opportunities for people to connect with each other. It may, therefore, be significantly beneficial for its users because social ties are known to play a significant role in human well-being including life-satisfaction [1], health [2, 3], and professional development [4, 5]. There is growing evidence that these findings apply not only to offline social ties but to online friendship as well [6, 7]. This role of the internet may be particularly important for underprivileged groups of people such as students from low-performing schools who lack resources in their immediate environment. Connections with students from high-performing schools might potentially influence their university aspirations [8], improve educational outcomes [9], and promote positive behavioral change [10].

People from underprivileged backgrounds tend not to benefit as much as their peers from the Internet (a phenomenon usually referred to as digital inequality [11]). While well-educated people often use the Internet for medical or juridical advice, job seeking or education, their

less educated peers use it predominantly for entertainment [12–14]. The use of social media by students is known to be differentiated in a similar way depending on their academic performance. High-performing students use it for information seeking while low-performing students for chatting and entertainment [15, 16]. It may be expected that online social ties would also depend on academic achievements and that students might be segregated by the educational outcomes in the digital space. At a general level, segregation is the degree to which several groups of people are separated from each other [17]. In this paper, we investigate whether students from high- and low-performing schools are separated (i.e. not connected via online friendship) in the digital space.

We use data from 36, 951 15-year-old students from 590 schools of Saint Petersburg, Russia, registered on a popular social networking site VK (http://vk.com) (see Methods for details about the sample). VK is the Russian analog of Facebook and the largest European social networking site. It is ubiquitous among young Russians: more than 90% of 18-24-year-olds use it regularly [18]. The information in users' public profiles includes their age and the schools they are studying in. This information is available via the open application programming interface (API) of VK. We use the VK API to download information about all students who indicate that they study in one of Saint Petersburg's schools and who were born in 2001 (i.e. that students were 15 years old at the time of data collection).

Similar to other social networking sites, users might become "friends" on VK if they mutually confirm this status. We use information about such online friendships to construct a weighted network of schools (Fig 1), where two schools are connected if there is at least one friendship tie between their students (see Methods for details), and the weight corresponds to the number of such ties. For each school, the information about its geographical coordinates along with the performance of its graduates on the unified state examination (USE) is available (see Methods). The USE scores serve as a proxy for schools' educational outcomes.

Residential segregation by income is believed to be an important source of variation in schools' educational outcomes in some countries [19–21]. It means that low-performing schools are concentrated in less affluent neighborhoods and the educational outcomes of a school could be effectively predicted from the socioeconomic status of its district [22]. The situation might be different in Saint Petersburg thanks to the egalitarian nature of the Russian educational system inherited from the Soviet period. To account for potential effects of residential segregation, we collect data from 11, 034 apartments from the largest Russian real estate site CIAN (http://cian.ru) and use average apartment price as a proxy of neighborhood affluence. We then check whether schools' educational outcomes are correlated with the affluence of their neighborhood.

We measure geographical segregation of schools as a correlation between the educational outcomes of a school and those of its closest geographical neighbors. We then compare this segregation with that in the digital space. In this case, instead of the closest geographical neighbors, we examine the educational outcomes of schools' closest digital neighbors. We assume that the distance between two schools in the digital space is inversely proportional to the number of online friendship ties between them.

The probability of an online friendship between two people is known to be strongly dependent on the geographical distance between them [23–26]. It is, therefore, important to ensure that any observed effect for the digital network of schools is not solely driven by the geographical constraints. To achieve this, we use a random graph model that preserves geographical constraints—namely, the probability of a friendship tie between two schools given the geographical distance between them. We then compare the results obtained for such random networks with the observed results for the real network.
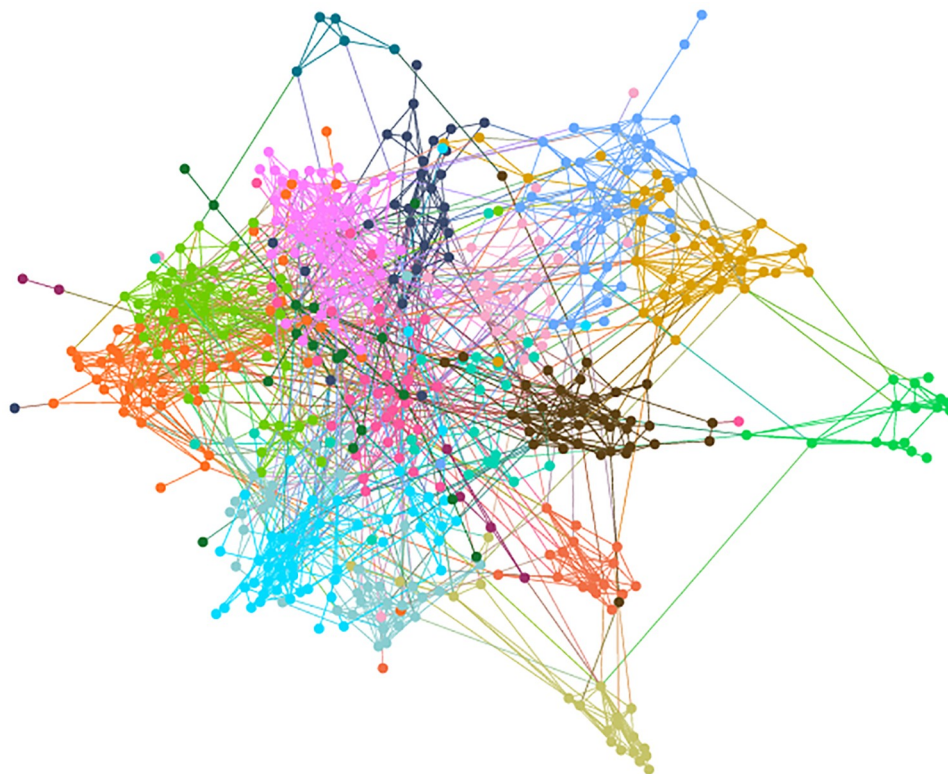
Fig 1. The school network. Nodes represent schools. Different colors correspond to administrative districts of Saint Petersburg. Two schools are connected if there is a friendship tie between their students. For visual clarity, only strong connections (at least three friendship ties) are shown.

https://doi.org/10.1371/journal.pone.0217142.g001

Results

Distance and online relationships

We find that geographical distance plays an important role in the formation of an interschool friendship. The probability of a friendship tie between two close schools is high (0.75) but it declines rapidly with distance following the power law (Fig 2). The best fit is provided by the exponent −0.62 (Fig 2 inset), which is similar to the previously observed results [26].

Geographical segregation

We find that the educational outcomes of schools do not depend on their distance from the city center (Pearson correlation coefficient between USE scores of schools and their distance from the center is 0.018, $P = 0.65$). The distance from the center may be, however, a poor proxy for neighborhood affluence. Hence, we additionally collect information about apartment prices across the city. We use the average apartment price in the area where schools are located as a proxy for their neighborhood affluence. We then compute the correlation between schools' USE scores and neighborhood affluence, $S_n(R)$ (see Methods). The exact value depends on R (see S1 Fig), and the maximum value is $S_n = 0.12$ ($P = 0.007$), indicating a weak correlation between educational outcomes and neighborhood affluence. Finally, we compute a correlation between USE scores of schools and average USE score of their N closest geographical neighbors, $S_g(N)$ (see Methods). We find no correlation $S_g(N) = 0.01$ ($P = 0.73$) for $N = 20$ (Fig 3a); this result holds true for all values of N (S2 Fig).
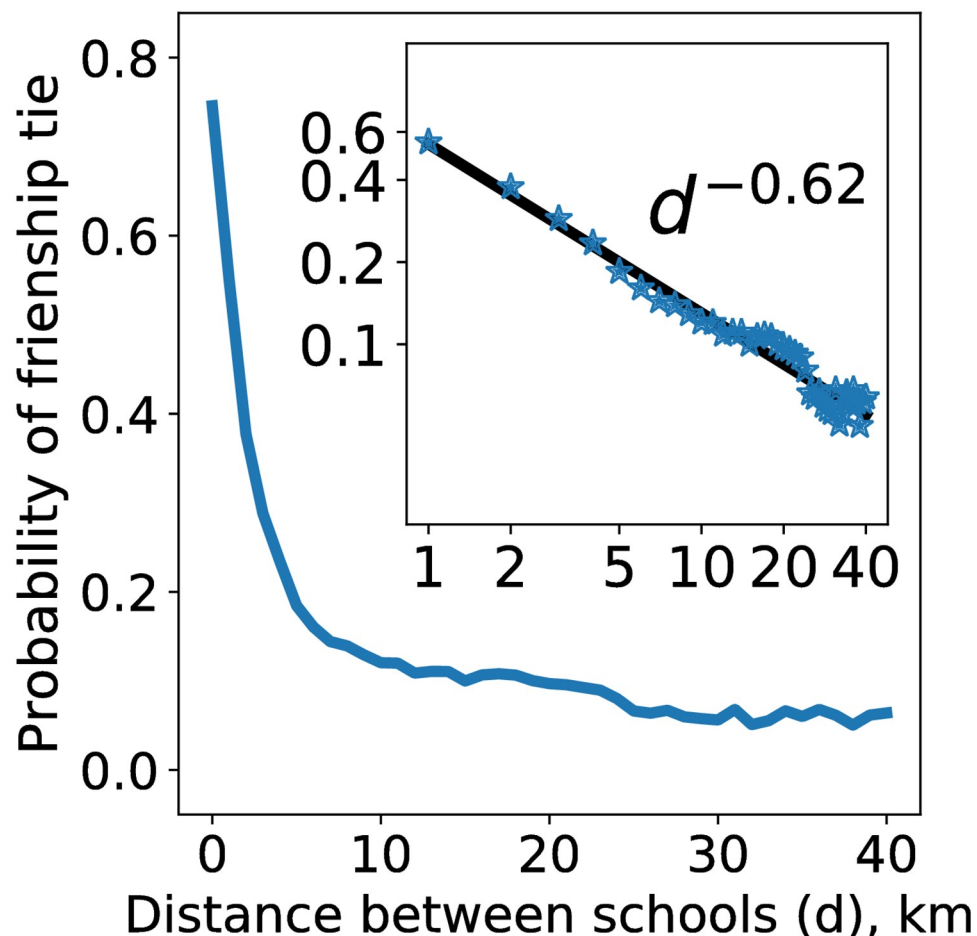
Fig 2. Probability of a friendship tie between two schools as a function of distance between these schools. For close schools, the probability is 0.75 and it then declines with distance following the power law (inset).

We, therefore, find that there is only a weak if any relationship between educational outcomes of a school and its location in physical space. However, as we show in the next section, this result does not apply for the school location in the digital space.

Digital segregation

We find that there is a relatively strong correlation between the educational outcomes of schools and their N closest digital neighbors (see Methods). $S_d(N) = 0.47$ $(P < 10^{-33})$ for $N = 20$ (Fig 3b). The correlation is significant for all N (S2 Fig).

To rule out the role of geographical constraints in the observed digital segregation, we use a random graph model that preserves relationships between distance and probability of a friendship tie from the observed network (i.e. we create a tie between two schools with a probability from distribution represented in Fig 2 that depends on the distance between schools). We compute $S_d^{rand}(1)$ for generated random networks and compare it with $S_d(1)$. After 10,000 simulations we obtain $\langle S_d^{rand}(1) \rangle = -0.0005$ and $SD(S_d^{rand}(1)) = 0.04$. The maximum value $\max(S_d^{rand}(1)) = 0.14 < S_d(1)$. This result makes the observed digital segregation significant with $P < 10^{-4}$.
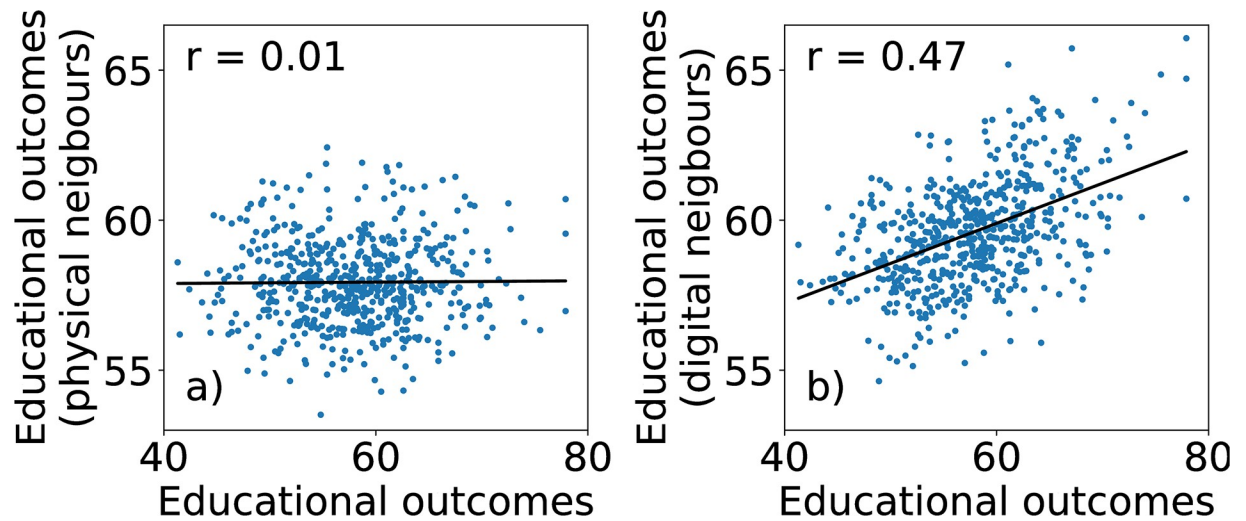
Fig 3. Correlation between educational outcomes of schools and their 20 closest geographical (a) and digital neighbors (b). While there is no correlation for physical neighbors, there is a relatively strong correlation for digital neighbors. These results hold true regardless of the number of neighbors used in the analysis.

https://doi.org/10.1371/journal.pone.0217142.g003

We also find that high-performing schools not only tend to be connected with each other but also have more connections on average than low-performing schools. The correlation between the degree centrality of schools in the network and their educational outcomes is 0.49 (S3 Fig). This correlation might be partially explained by the presence of high performing selective schools that attract students from all over the city (see S1 Text for details).

One of the strongest predictors of academic achievements is the socioeconomic status of students [27]. This is true not only on the individual level but also on the school level, i.e. the socioeconomic composition of the student body is the strong predictor of school's educational outcomes. For Russian schools, 34%–41% of the variance in average USE scores is explained by the socioeconomic composition of the student body, the same amount that is explained by school's material and human resources [28]. It is, therefore, noteworthy, that the degree centrality is such a strong predictor of educational outcomes. Note that this is a simple network property and that it does not contain any information about schools or students themselves.

We show, therefore, that the educational outcomes of a school are closely related to its location in the digital space. More central schools tend to be high performing. We also show that schools with similar academic performance tend to be connected in the digital space. We demonstrate that these results cannot be explained by schools' locations in the physical space.

Discussion

Both for research and policy-making purposes, it is crucial to understand the context in which schools operate. This requirement traditionally means collecting information about school resources and the socioeconomic status of its students. Today, students spend much of their time online [29], and it may be warranted to consider students' online environment on a par with their home environment. In this paper, we focus only on one dimension of such an online environment, namely interschool friendship on a social networking site. We find that school position in an online friendship network could explain as much variation in the educational outcomes of its students as their socioeconomic status, indicating the importance of the digital context. Online inequalities might merely reflect existing socioeconomic inequality or rather complement it. In particular, it is not known if students from different schools who are friends

on VK know each other offline or these connections are only virtual. Future research is required to clarify this relationship.

Social media have become the main source of information for young people. In Russia, VK is referred to as the main source of information about the country and the world by 70.3% of respondents—more than any other information source [30]. It is also considered more trustworthy than traditional media [30]. The news feed of the social network mainly comprises posts shared by online friends. Friends from different schools may, therefore, be an important source of diversity in the information environment of students. In particular, the connections with students from high-performing schools could have a positive impact on students from low-performing schools. However, our results suggest that interschool friendship ties mainly exist between schools with similar educational outcomes. Intriguingly, this digital separation cannot be explained by the geographical location of schools. This result means that the digital environment not only fails to remove segregation but rather might amplify it.

Methods

Data collection

According to the open data government portal (http://data.gov.spb.ru), there are 638 high schools in Saint Petersburg. This number excludes specific types of schools such as boarding schools, cadet schools, and educational centers. We use open VK API to find these schools in the VK database. We find VK IDs for 628 of the schools. We exclude school №1 from the sample because it has an unreasonable number of users (more than 1000 per cohort). We also exclude two pairs of schools with identical names. We then use data from the web portal "Schools of Saint Petersburg" (http://www.shkola-spb.ru) to obtain the average performance of schools' graduates at the Unified State Examination. This is a mandatory state examination that all school graduates should pass in Russia. This information was available for 590 schools from our sample.

We then perform requests to VK API to obtain the lists of all users who were born in 2001 and indicate that they are studying in one of the schools from our sample. To exclude users who provided false information about their school, we remove profiles with no friends from the same school, as previously recommended [31]. We also exclude students who indicate several schools in their profiles. Finally, we download the lists of all VK friends for users from our sample. All collected data is publicly available. The VK team confirmed to us that we can use its API in this way for research purposes.

We also use data from the largest Russian real estate site CIAN to collect information about the prices of all 2-room apartments in Saint Petersburg listed on the site. For each apartment, its price per square meter was calculated. CIAN team approved the use of this data for research purposes.

Network of schools

We define a 36,951 × 36,951 adjacency matrix F that represents the friendship network of students (i.e. $F_{i,j} = 1$ if students i and j are friends on VK and $F_{i,j} = 0$ otherwise). We assume that student i studies in school $s(i)$, and construct a weighted network of schools by counting the number of all friendship ties between two schools. This network is represented by 590 × 590 matrix A where

$$A_{k,l} = \sum_{\{i,j \mid s(i)=k, s(j)=l\}} F_{i,j}.$$

One potential disadvantage of this definition is that two schools could be considered as closely connected when only one student from the first school has a lot of friends from the other. We therefore also use an alternative way to define the weight of the school tie. In this case, instead of friendship ties, we count the number of students from one school that have friends from another (i.e. we define $\tilde{A}_{k,l} = |\{i|s(i) = k$ and $\exists j : F_{i,j} = 1, s(j) = l\}|$). We could then construct a symmetric matrix $\hat{A}_{k,l} = \min(\tilde{A}_{k,l}, \tilde{A}_{l,k})$. This alternative metric leads to the same results, and therefore we opted for the first more straightforward approach.

Segregation measures

If U_i is the average performance on the Unified State Examination of graduates from school i, we could then define segregation based on the affluence of school neighborhoods in the following way:

$$S_n(R) = \text{corr}(U_i, \sum_{\{j|d(i,j)<R\}} P_j / |\{j|d(i,j) < R\}|),$$

where P_j is the price of apartment j in rubles per square meter and $d(i, j)$ is the distance between school i and apartment j.

We denote geographical neighbors of school i by $N_g(i)$. $N_g(i) = (s_{i,1}, \ldots, s_{i,590})$ is an ordered list of all schools such as $\tilde{d}(i, s_{i,k}) <= \tilde{d}(i, s_{i,k+1})$, where $\tilde{d}$ is the geographical distance between schools. We then denote the list of k-closest geographical neighbors by $N_g^k(i) = (s_{i,1}, ..., s_{i,k})$. We define the k-closest digital neighbors $N_d^k(i)$ by replacing geographical distance with the digital distance that is equal to $1/A_{i,j}$.

We then define geographical and digital segregations by academic performance in the following manner:

$$S_g(k) = \text{corr}(U_i, \sum_{\{j|j \in N_g^k(i)\}} U_j / k)$$

$$S_d(k) = \text{corr}(U_i, \sum_{\{j|j \in N_d^k(i)\}} U_j / k)$$

Note that in the case of digital segregation, there could be several schools with exactly the same distance from a certain school. In this case, $N_d^k(i)$ is not uniquely defined. In our computations, we randomly select with equal probabilities one of the possible $N_d^k(i)$.

Ethical considerations

The data was collected as part of the "Digital Trace" project that was approved by the Institutional Review Board of the National Research University Higher School of Economics. Note that the units of our analysis are schools rather than individuals. Public information about friendship ties between users who indicated their high schools on VK was used to construct a friendship network between schools. Neither names of users nor other personal information available from VK were analyzed or collected as part of this research.

Supporting information

S1 Text.
(PDF)

S1 Fig. Segregation $S_n(R)$ as a function of the radius R that defines school neighborhood.
(TIF)

S2 Fig. Digital $S_d(N)$ and geographical $S_g(N)$ segregations as functions of the number of neighbors N used in the analysis.
(TIF)

S3 Fig. Correlation between educational outcomes of schools and their degree centrality.
(PDF)

Acknowledgments

Author Contributions

Conceptualization: Ivan Smirnov.

Formal analysis: Ivan Smirnov.

Writing – original draft: Ivan Smirnov.

References

1. Diener E, Suh EM, Lucas RE, Smith HL. Subjective well-being: Three decades of progress. Psychological bulletin. 1999; 125(2):276.

2. Holt-Lunstad J, Smith TB, Layton JB. Social relationships and mortality risk: a meta-analytic review. PLoS medicine. 2010; 7(7):e1000316. https://doi.org/10.1371/journal.pmed.1000316 PMID: 20668659

3. Kawachi I, Berkman LF. Social ties and mental health. Journal of Urban health. 2001; 78(3):458–467. https://doi.org/10.1093/jurban/78.3.458 PMID: 11564849

4. Podolny JM, Baron JN. Resources and relationships: Social networks and mobility in the workplace. American sociological review. 1997; 62(5):673–693. https://doi.org/10.2307/2657354

5. Ng TW, Eby LT, Sorensen KL, Feldman DC. Predictors of objective and subjective career success: A meta-analysis. Personnel psychology. 2005; 58(2):367–408. https://doi.org/10.1111/j.1744-6570.2005.00515.x

6. Hobbs WR, Burke M, Christakis NA, Fowler JH. Online social integration is associated with reduced mortality risk. Proceedings of the National Academy of Sciences. 2016; 113(46):12980–12984. https://doi.org/10.1073/pnas.1605554113

7. Manago AM, Taylor T, Greenfield PM. Me and my 400 friends: The anatomy of college students' Facebook networks, their communication patterns, and well-being. Developmental psychology. 2012; 48(2):369 PMID: 22288367

8. Cohen J. Peer influence on college aspirations with initial aspirations controlled. American Sociological Review. 1983; 48(5):728–734. https://doi.org/10.2307/2094931

9. Lomi A, Snijders TA, Steglich CE, Torló VJ. Why are some more peer than others? Evidence from a longitudinal study of social networks and individual academic performance. Social Science Research. 2011; 40(6):1506–1520. https://doi.org/10.1016/j.ssresearch.2011.06.010 PMID: 25641999

10. Maxwell KA. Friends: The role of peer influence across adolescent risk behaviors. Journal of Youth and adolescence. 2002; 31(4):267–277. https://doi.org/10.1023/A:1015493316865

11. DiMaggio P, Hargittai E, Celeste C, Shafer S. From unequal access to differentiated use: A literature review and agenda for research on digital inequality. Social inequality. 2004; p. 355–400.

12. Pearce KE, Rice RE. Somewhat separate and unequal: digital divides, social networking sites, and capital-enhancing activities. Social Media+ Society. 2017; 3(2):2056305117716272.

13. Büchi M, Just N, Latzer M. Modeling the second-level digital divide: A five-country study of social differences in Internet use. New Media & Society. 2016; 18(11):2703–2722. https://doi.org/10.1177/1461444815604154

14. Van Deursen AJ, Van Dijk JA. The digital divide shifts to differences in usage. New media & society. 2014; 16(3):507–526. https://doi.org/10.1177/1461444813487959

15. Junco R. Too much face and not enough books: The relationship between multiple indices of Facebook use and academic performance. Computers in human behavior. 2012; 28(1):187–198. https://doi.org/10.1016/j.chb.2011.08.026

16. Smirnov I. Predicting PISA Scores from Students' Digital Traces. International Conference on Web on Social Media. 2018; p. 360–364.

17. Allen R, Vignoles A. What should an index of school segregation measure? Oxford Review of Education. 2007; 33(5):643–668. https://doi.org/10.1080/03054980701366306

18. Public Opinion Foundation. Online practices of Russians: social networks; 2016. http://fom.ru/SMI-i-internet/12495.

19. Flores CA. Residential segregation and the geography of opportunites: a spatial analysis of heterogeneity and spillovers in education. The University of Texas at Austin; 2008.

20. Gordon I, Monastriotis V. Urban size, spatial segregation and educational outcomes. London: London School of Economics; 2003.

21. Owens A. Income segregation between school districts and inequality in students' achievement. Sociology of Education. 2018; 91(1):1–27. https://doi.org/10.1177/0038040717741180

22. Reardon SF, Kalogrides D, Shores K. The Geography of Racial/Ethnic Test Score Gaps. CEPA Working Paper No 16-10. 2017;.

23. Takhteyev Y, Gruzd A, Wellman B. Geography of Twitter networks. Social networks. 2012; 34(1):73–81. https://doi.org/10.1016/j.socnet.2011.05.006

24. Shin WY, Singh BC, Cho J, Everett AM. A new understanding of friendships in space: Complex networks meet Twitter. Journal of Information Science. 2015; 41(6):751–764. https://doi.org/10.1177/0165551515600136

25. Lengyel B, Varga A, Ságvári B, Jakobi Á, Kertész J. Geographies of an online social network. PloS one. 2015; 10(9):e0137248. https://doi.org/10.1371/journal.pone.0137248 PMID: 26359668

26. Grabowicz PA, Ramasco JJ, Gonçalves B, Eguíluz VM. Entangling mobility and interactions in social media. PloS one. 2014; 9(3):e92196. https://doi.org/10.1371/journal.pone.0092196 PMID: 24651657

27. Sirin SR. Socioeconomic status and academic achievement: A meta-analytic review of research. Review of educational research. 2005; 75(3):417–453. https://doi.org/10.3102/00346543075003417

28. Yasterbov G, Bessudnov A, Pinskaya M, Kosaretsky S. Contextualizing Academic Performance in Russian Schools: School Characteristics, the Composition of Student Body and Local Deprivation. Higher School of Economics Research Paper. 2014;.

29. Koroleva D. Always online: Using mobile technology and social media at home and at school by modern teenagers. Educational Studies. 2016; 1:205–224.

30. Kasamara, Valeria and Sorokina, Anna. Russian Students' Values; 2017. http://tass.ru/obschestvo/4255020.

31. Smirnov I, Sivak E, Kozmina Y. In Search of Lost Profiles: The Reliability of VKontakte Data and its Importance in Educational Research. Educational Studies. 2016; 4:106–122.