

Analysis of open data of a social network in order to identify deviant communities

Rostislav Mikherskii
Physico-technical Institute
V.I. Vernadsky Crimean Federal University
Simferopol, Russia
mrm03@mail.ru

Dmitry Kuznetsov
Physico-technical Institute
V.I. Vernadsky Crimean Federal University
Simferopol, Russia
dimabrayankuznetsov@mail.ru

Abstract—The system of analysis of open data of the social network Vkontakte is developed and programmatically implemented. Two ways of identification of deviant communities are proposed. The first way is by the number of community subscribers blocked by the social network for violating the rules. The second way, by the presence of common subscribers between the studied community, and the community about which it is precisely known that it is deviant. It is experimentally established that the second method of identification of deviant communities gives the best result.

Keywords—big data, open data, social network

I. INTRODUCTION

Analysis of open data from social networks is a significant area in the field of big data processing. In particular, an important task for both law enforcement agencies and social network administrators is to identify communities of these networks that disseminate socially dangerous content. Many works that were written recently have been devoted to discussion of this problem. The work [1] is devoted to the development of a method for assessing the degree of connectedness of user profiles of social networks based on open data. The degree of connectedness of user profiles is understood as the probability of meeting profile owners in real life. In [2], a review of methods that detect the demographic attributes of a user from their profile and messages is made. In [3,4], forms of deviant behavior of users of the Russian-language segment of the Internet are examined in detail. In particular, in [4] it was shown that the main reason for deviant behavior in social networks is virtuality and anonymity. In [5], according to foreign sources, a review of the main methods of analysis of social networks in relation to the task of identifying suspicious and criminal communities is carried out.

To study social networks in terms of social relationships, the Social Network Analysis (SNA) method is often used. The SNA method is described in detail in [6–8]. In this method, the objects of research are the nodes, and the relationships characterizing the relationship between them. Nodes can be communities, users of social networks, etc. The connections between these nodes can be money transfers, communication, friendship, etc. This method has been successfully used to study the organization structure of the Al-Qaida terrorist network [9], to study the network of terrorist organizations operating in India [10], to analyze the topological structure of criminal networks, in particular the network of methamphetamine traffic [11]. These research studies are mainly motivated by the need to find effective methods to undermine criminal or terrorist organizations.

Anorexia-oriented online communities have been studied in [12–14]. A wide range of issues was studied in these works, including the construction and management of member identities, the processes of social recognition, the

emergence of group norms, and the use of linguistic markers. Similar studies have been conducted for groups promoting suicidal behavior [15].

In [16–20], the categorization of pornographic content and the frequency of its use were studied.

In [21], the authors also focused on consumption networks for adult content, which is present in many online social networks and on the Internet as a whole. The authors of this work investigated how such communities interact with the entire social network. They found that few small and closely related communities are responsible for much of the production of content. Produced content is distributed through the rest of the network mainly directly or through bridge communities, reaching at least 450 times more users. In this work, a demographic analysis of the networks of producers and consumers of adult content was also carried out. It has been shown that it is possible to easily identify several key users in order to radically eradicate the process of distribution of pornographic content.

The issue of community polarization in social networks was studied in detail in [22]. It proposed a new polarization metric based on the analysis of the boundary of a pair of (potentially polarized) communities, which better reflects the concepts of antagonism and polarization.

Cyber aggression, as a form of deviant behavior in the Internet environment, was studied in detail in [23–28]. This socio-psychological phenomenon has many forms, the main of which are trolling, cybermobbing and astroturfing.

As can be seen from the above review of published scientific papers, the search for deviant communities is an important task both for scientists involved in researching such communities and for law enforcement agencies. Unfortunately, most often, the identification of deviant communities is carried out manually, often only by user complaints.

The aim of this work was to develop a methodology for identifying deviant communities in the social network Vkontakte in automatic mode. To achieve this goal, two options have been proposed to search for such communities.

II. RESULTS

In the first version, the following algorithm for searching for such communities is proposed and programmatically implemented. For the studied community, the number l of subscribers blocked by the social network for breaking the rules, as well as the total number L of subscribers of this community, is determined. The coefficient $k = \frac{l}{L}$ is found. It

is assumed that if the coefficient k is greater than some critical value k_d , then the community under study is deviant.

The software implementation of the above algorithm was implemented in the Python programming language. During

the implementation of this program, 50,704 communities of the Vkontakte social network were randomly selected. In order to shorten an influence of statistical error, only communities with 100 or more total subscribers were selected from the general list. Due to system and API limitation, communities with a few members were considered. A coefficient was calculated for each of these communities. Further, all communities were sorted in descending order of magnitude of this coefficient. Table 1 presents the first 20 communities from the list.

TABLE I. COMMUNITIES WITH A HIGH PERCENTAGE OF BLOCKED SUBSCRIBERS

№	Community identification number	Number of subscribers in the community	Number of blocked subscribers	Percentage of blocked subscribers from the total number of community subscribers, $k \cdot 100\%$
1	172017411	104	101	97.1154
2	171896750	122	114	93.4426
3	41398959	107	98	91.5888
4	125043269	1017	904	88.8889
5	19613748	960	852	88.75
6	176328754	226	193	85.3982
7	148023353	495	419	84.6465
8	188941498	530	438	82.6415
9	23811356	1116	921	82.5269
10	164252296	152	123	80.9211
11	150230769	198	157	79.2929
12	130381011	200	157	78.5
13	154988787	410	317	77.3171
14	155397881	847	654	77.2137
15	149830913	107	81	75.7009
16	170030633	577	428	74.1768
17	***	174	129	74.1379
18	164288533	153	113	73.8562
19	143657800	424	312	73.5849
20	157513161	420	309	73.5714

In order to prevent propaganda of deviant communities, in this table and further in table 2, the identification number of all such communities is replaced by the symbols “***”.

As can be seen from this list, there is only one deviant community in it (community under No. 17). This community was classified as deviant due to the presence of pornographic material in it.

Thus, the hypothesis that the percentage of blocked users in deviant communities is greater than in non-deviant communities has not been experimentally confirmed. Furthermore, it's clear that some communities are abandoned and they can contain a lot of banned users because of lack of moderation and new subscribers. Other communities can be related to advertising or temporary events. But they are still not deviant despite the fact that social network Vkontakte has special rules that restrict the creation of such communities as communities with an inappropriate content.

The second option for searching for deviant communities is based on the following algorithm: One community is found for which it is known for certain that it is deviant. For this community, a list of subscribers is defined. Each of these subscribers defines the communities to which it is subscribed. For each of the communities in this list, the number of subscribers who are also subscribers of the studied

deviant community is determined. It is assumed that a sufficiently large number of communities from this list will also be deviant. This algorithm was programmatically implemented using the Python programming language.

To test the performance of this program, the deviant community “Mom Anarchy” was chosen with an identification number of 177615404. This community is engaged in popularizing the ideas of anarchism and has 32097 subscribers. The data processing time was 18 hours. Followers of this community are also subscribed to 940512 other communities. All of them were sorted in descending order by the number of users who are also subscribed to the Mama Anarchy community. Table 2 presents the first 20 communities from this list.

TABLE II. COMMUNITIES WHOSE SUBSCRIBERS ARE ALSO SUBSCRIBERS OF THE MOM ANARCHY COMMUNITY

№	Community identification number	Number of subscribers in the community	The number of subscribers who are also subscribers of the Mama Anarchy community
1	***	5539982	15035
2	91050183	9356399	12924
3	***	707327	12712
4	159146575	1162785	12521
5	***	563784	11987
6	***	4403183	11644
7	***	2768306	11317
8	***	2508543	11246
9	57846937	11275065	11224*
10	***	2684988	11154
11	***	2586853	11145
12	150550417	937052	10916
13	149094324	2076903	10832*
14	30316056	1809325	10451
15	66678575	4976245	10299
16	12353330	3555825	10167
17	154168174	1264550	10145
18	173556111	641480	10005
19	***	3802683	9576
20	133180305	3116645	9540

As can be seen from this table, out of 20 communities of the presented list, 9 are deviant. The main reasons that these communities are attributed to deviants are: propoganda of violence, criticism of the existing constitutional system, and the use of profanity. This results show us that algorithm results should not be considered as final predictions but as an assumption. Still results must be managed by special person to make a conclusion about community content. The main aim of this algorithm for now is to narrow the search for deviant communities.

Furthermore, this algorithm allows to consider communities with bigger amount of subscribers. However, it should be mentioned that API has a strong impact on algorithms productivity. Therefore such systems have a portability limitations. Nevertheless, the core idea of this system is to show dependencies between blocked users amount and community content.

The scientific novelty of this work lies in the proposed algorithm, which helps to identify deviant communities. Despite the fact that current algorithm can only help us to

make suggestion about community content, there could be ways to improve it by using extra algorithms and tools, such as image recognition tools and text analyzer. Therefore, holistic recognition system could be developed to make more accurate predictions about deviant communities in social networks with open API.

III. CONCLUSION

Thus, the second way of identifying deviant communities is much more effective than the first. This technique for identifying deviant communities in automatic mode can be applied not only on the social network Vkontakte but also in other social networks. We also note that the second method can be applied not only to search for deviant communities, but also when searching for communities related to the studied community, for example, in marketing research. In the case of such studies, it is possible to determine the interests of community users and, accordingly, build a policy to attract new users to this community.

Another possible use of this method is to conduct an advertising campaign of a certain community. In this case, as the studied community, you can choose the community whose advertising you want to conduct. Define a list of communities associated with this community and place advertising messages in these communities.

It should also be noted that to search for deviant communities, it may be useful to use machine learning methods, such as, for example, artificial immune systems [29-31] or convolutional neural networks [32-41]. However, even when using machine learning, the method of identifying deviant communities by the presence of common subscribers between the studied community, and the community about which it is known for certain that it is deviant will not lose its relevance. This is primarily due to the fact that this method has a high degree of transparency in interpreting the results obtained, in contrast to machine learning methods, which are often a black box, the results of which are often incomprehensible.

Thus, in this study, a new method is proposed that allows you to quickly, cheaply and efficiently search for deviant communities.

ACKNOWLEDGMENT

In conclusion, we would like to thank Marina Vsevolodovna Glumova, Director of the Physico-technical Institute of the V. I. Vernadsky Crimean Federal University, and Victor Vasilyevich Milyukov, head of the Department of computer engineering and modeling of the Physico-technical Institute of the V. I. Vernadsky Crimean Federal University, for their assistance in organizing research.

REFERENCES

- [1] V. Kataeva, I. Pantyukhin and I. Yurin, "Methods for assessing the degree of connectivity of social network user profiles based on open data," *Open education*, vol. 21, no. 6, pp. 14-22, 2017.
- [2] A. Gomzin and S. Kuznetsov, "Methods for constructing socio-demographic profiles of Internet users," *Proceedings of the ISP RAS*, vol. 27, no. 4, pp. 129-142, 2015.
- [3] A. Baklantseva, "Transformation of social norms and deviations in the Russian-language Internet," *News of universities in the North Caucasus region. Social sciences*, vol. 3, pp. 21-25, 2014.
- [4] D. Cherenkov, "Deviant behavior in social networks: causes, forms, consequence," *Nauka-Rastudent.Ru*, vol. 7, no. 19, pp. 29, 2015.
- [5] M. Basarab, I. Ivanov, A. Kolesnikov and V. Matveev, "Detection of illegal activities in cyberspace based on the analysis of social networks: algorithms, methods and tools (review)," *Cybersecurity issues*, vol. 4, no. 17, pp. 11-19, 2016.
- [6] L.C. Freeman, "The development of social network analysis: A study in the sociology of science," *Social Networks*, vol. 27, no. 4, pp. 377-384, 2005.
- [7] A. Hopkins, "Graph theory, social networks and counter terrorism," *Univ. of Massachusetts Dartmouth*, pp. 22, 2010.
- [8] L.C. Freeman, "Centrality in social networks conceptual clarification," *Social networks*, vol. 1, no. 3, pp. 215-239, 1979.
- [9] V. Krebs, "Mapping networks of terrorist cells," *Connections*, vol. 24, no. 3, pp. 43-52, 2002.
- [10] P. Choudhary and U. Singh, "A survey on social network analysis for counter-terrorism," *Int. Journal of Computer Applications*, vol. 112, no. 9, pp. 24-29, 2015.
- [11] J. Xu, H. Chen, "The topology of dark networks," *Communications of the ACM*, vol. 51, no. 10, pp. 58-65, 2008.
- [12] J. Gavin, K. Rodham and H. Poyer, "The presentation of "pro-anorexia" in online group interactions," *Qualitative Health Research*, vol. 18, no. 3, pp. 325-333, 2008.
- [13] J.D.S. Ramos, A.D.F. PereiraNeto and M. Bagrichevsky, "Pro-anorexia cultural identity: characteristics of a lifestyle in a virtual community," *Interface (Botucatu)*, vol. 15, no. 37, pp. 447-460, 2011.
- [14] N. Boero and C.J. Pascoe, "Pro-anorexia communities and online interaction: Bringing the pro-ana body online," *Body & Society*, vol. 18, no. 2, pp. 27-57, 2012.
- [15] S.M. Haas, M.E. Irr, N.A. Jennings and L.M. Wagner, "Communicating thin: A grounded model of Online Negative Enabling Support Groups in the pro-anorexia movement," *New Media & Society*, vol. 13, no. 1, pp. 40-57, 2010.
- [16] M. Schuhmacher, C. Zirn and J. Volker, "Exploring youporn categories, tags, and nicknames for pleasant recommendations," *Workshop on Search and Exploration of X-Rated Information. ACM*, pp. 27-28, 2013.
- [17] G. Tyson, Y. Elkhatib, N. Sastry and S. Uhlig, "Are People Really Social in Porn 2.0?" *Proceedings of the 9th International AAAI Conference on Web and Social Media (ICWSM)*, pp. 436-444, 2015.
- [18] G.M. Hald and A. Stulhofer, "What types of pornography do people use and do they cluster? Assessing types and categories of pornography consumption in a large-scale online sample," *Journal of Sex Research*, pp. 1-11, 2015.
- [19] G.M. Hald, N.N. Malamuth, T. Lange, "Pornography and sexist attitudes among heterosexuals," *Journal of Communication*, vol. 63, no. 4, pp. 638-660, 2013.
- [20] G.M. Hald, "Gender differences in pornography consumption among young heterosexual danish adults," *Archives of sexual behavior*, vol. 35, no. 5, pp. 577-585, 2006.
- [21] M. Coletto, L.M. Aiello, C. Lucchese and F. Silvestri, "On the Behaviour of Deviant Communities in Online Social Networks," *Proceedings of the 10th International AAAI Conference on Web and Social Media (ICWSM)*, pp. 72-81, 2016.
- [22] P.H.C. Guerra, Jr.W. Meira, C. Cardie, R. Kleinberg, "A measure of polarization on social media networks based on community boundaries," *Proceedings of the 7th International AAAI Conference on Web and Social Media (ICWSM)*, pp. 215-224, 2013.
- [23] J.S. Chibbaro, "School counselors and the cyberbully: interventions and implications," *Journal of Professional School Counseling*, vol. 11, no. 1, pp. 65-68, 2007.
- [24] R. Gable, J. Snakenborg and R. Van Acker, "Cyberbullying: Prevention and Intervention to Protect Our Children and Youth," *Preventing School Failure*, vol. 55, no. 2, pp. 88-95, 2011.
- [25] W. Heirman and M. Walrave, "Cyberbullying: Predicting Victimization and Perpetration," *Children & Society*, vol. 25, pp. 59-72, 2011.
- [26] J.S. Donath, "Identity and Deception in the Virtual Community," *Communities in Cyberspace*, London: Routledge, pp. 26, 1999.
- [27] N.E. Willard, "From Cyberbullying and Cyberthreats: Responding to the Challenge of Online Social Aggression, Threats, and Distress," *Champaign, IL: Research Press*, pp. 303, 2007.
- [28] R.A. Vnebrachnykh, "Trolling as a form of social aggression in virtual communities," *Bulletin of the Udmurt University. Philosophy. Sociology. Psychology. Pedagogy*, vol. 1, pp. 48-51, 2012.

- [29] R. Mikherskii, "Application of an artificial immune system for visual pattern recognition," *Computer Optics*, vol. 42, no. 1, pp. 113-117, 2018. DOI: 10.18287/2412-6179-2018-42-1-113-117.
- [30] G. Luh, "Face recognition based on artificial immune networks and principal component analysis with single training image per person," *Immune Computation*, vol. 2, no. 1, pp. 21-34, 2014.
- [31] D. Dasgupta, S. Yu and F. Nino, "Recent advances in artificial immune systems: Models and applications," *Applied Soft Computing*, vol. 11, no. 2, pp. 1574-1587, 2011. DOI: 10.1016/j.asoc.2010.08.024.
- [32] Y. Li, X. Zhang and D. Chen, "CSRNet: Dilated convolutional neural networks for understanding the highly congested scenes," *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pp. 1091-1100, 2018.
- [33] M. Kalayeh and M. Shah, "Training Faster by Separating Modes of Variation in Batch-Normalized Models," *IEEE Transactions on Pattern Analysis & Machine Intelligence*, vol. 42, no. 6, pp. 1483-1500, 2020. DOI: 10.1109/TPAMI.2019.2895781R.
- [34] A. Farrugia and C. Guillemot, "Light Field Super-Resolution Using a Low-Rank Prior and Deep Convolutional Neural Networks," *IEEE Transactions on Pattern Analysis & Machine Intelligence*, vol. 42, no. 05, pp.1162-1175, 2020. DOI: 10.1109/TPAMI.2019.2893666.
- [35] Z. Zhou, J. Shin, L. Zhang, S. Gurudu, M. Gotway and J. Liang, "Fine-tuning convolutional neural networks for biomedical image analysis: Actively and incrementally," *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pp. 7340-7349, 2017.
- [36] C. Lian, M. Liu, J. Zhang and D. Shen, "Hierarchical Fully Convolutional Network for Joint Atrophy Localization and Alzheimer's Disease Diagnosis Using Structural MRI," *IEEE Transactions on Pattern Analysis & Machine Intelligence*, vol. 42, no. 04, pp. 880-893, 2020. DOI: 10.1109/TPAMI.2019.2895781.
- [37] A. Bulat and G. Tzimiropoulos, "Hierarchical Binary CNNs for Landmark Localization with Limited Resources," *IEEE Transactions on Pattern Analysis & Machine Intelligence*, vol. 42, no. 02, pp. 343-356, 2020. DOI: 10.1109/TPAMI.2018.2866051.
- [38] V.A. Sindagi and V.M. Patel, "A survey of recent advances in CNN-based single image crowd counting and density estimation," *Pattern Recognit. Lett.*, vol. 107, pp. 3-16, 2017.
- [39] K. He, G. Gkioxari, P. Dollar and R. Girshick, "Mask R-CNN," *IEEE Transactions on Pattern Analysis & Machine Intelligence*, vol. 42, no. 02, pp. 386-397, 2020. DOI: 10.1109/TPAMI.2018.2844175.
- [40] S. Lin, R. Ji, C. Chen, D. Tao and J. Luo, "Holistic CNN Compression via Low-Rank Decomposition with Knowledge Transfer," *IEEE Transactions on Pattern Analysis & Machine Intelligence*, vol. 41, no. 12, pp. 2889-2905, 2019. DOI: 10.1109/TPAMI.2018.2873305.
- [41] I. Rocco, R. Arandjelovic and J. Sivic, "Convolutional Neural Network Architecture for Geometric Matching," *IEEE Transactions on Pattern Analysis & Machine Intelligence*, vol. 41, no. 11, pp. 2553-2567, 2019. DOI: 10.1109/TPAMI.2018.2865351.