Topology of Thematic Communities in Online Social Networks: A Comparative Study

Valentina Guleva, Danila Vaganov, Daniil Voloshin, and Klavdia Bochenina

ITMO University
guleva@corp.ifmo.ru
vaganov@corp.ifmo.ru
achoched@gmail.com
k.bochenina@gmail.com

Abstract. The network structure of communities in social media significantly affects diffusion processes which implement positive or negative information influence on social media users. Some of the thematic communities in online social networks may provide illegal services or information in them may cause undesired psychological effects; moreover, the topology of such communities and behavior of their members are influenced by a thematic. Nevertheless, recent research does not contain enough detail about the particularities of thematic communities formation, or about the topological properties of underlying friendship networks. To address this gap, in this study we analyze structure of communities of different types, namely, carders, commercial sex workers, substance sellers and users, people with radical political views, and compare them to the 'normal' communities (without a single narrow focus). We discovered that in contrast to ordinary communities which have positive assortativity (as expected for social networks), specific thematical communities are significantly disassortative. Types of anomalous communities also differ not only in content but in structure. The most specific are the communities of radicalized individuals: it was shown that they have the highest connectivity and the larger part of nodes within a friendship graph.

Keywords: network topology, data analysis, online social media, normal communities, anomalous communities, subscribers friendship networks

1 Introduction and Motivation

Online social media play an important role in our daily life, providing official news, presenting a ground for sharing our activities and opinions, and helping in the realization of our personal interests. In this way, one can easily see an enormous informational impact provided by the combination of news sources on each individual, and the contribution of each individual to this flow propagation and its correction, in turn. Consequently, friendship networks reflect the most probable paths of information transmission and their topological properties allows for the estimation of information diffusion effectiveness.

ICCS Camera Ready Version 2018
To cite this paper please use the final published version:
DOI: 10.1007/978-3-319-93698-7_20

Nevertheless, some kinds of informational influence may be undesired or even disruptive providing messages of illegal content, propaganda, or cyberbullying. Since undesired content is often provided by corresponding communities, the problem of their detection is of great interest.

The existing methods of anomaly detection in online social networks do not discover the laws of network formation process, the reasons of the emergence of differences between types of social communities, and topological properties common to them. The study presented contributes to this gap by presenting the analysis of topological differences in various types of communities and thus forms the basis for novel methods of identifying anomalous communities.

A structure of online social networks (OSNs) reflects real-world interactions, on the one hand, and on the other, facilitates the emergence of new virtual links. Formation of a structure is also influenced by the interface and the functionality of particular OSN. The third factor affecting the topology of friendship networks is a thematic of the community as it may cause specific behavioral and friending patterns of the individuals. Depending on these factors, resulting network topologies can vary. For example, Hai-Bo Hu and Xiao-Fan Wang [1] show that assortativity coefficient can be positive or negative depending on the online social network. They show MySpace, Flickr and LiveJournal have positive assortativity, while YouTube, Gnunella, and pussokram have the negative one. At the same time, real social networks were shown to have positive assortativity coefficient.

In this way, one can see social media crucially affects the formation of observed patterns. Twitter and Youtube restrict subscribers in the spectrum of possible interactions via limited functionality, and, consequently, an underlying network does not reflect real-world patterns well. Nevertheless, all of the online social media allows for discovering information flows and studying their informational effects. The most interpretable and comprehensive opportunities are presented by general-purpose online social networks, like Facebook and the largest Russian social network vk.com (VK), which support multilayer interactions between entities of different types.

Due to the prevalence of online social networks and the relative simplicity of OSN data collection via open APIs, nowadays they are widely studied by researchers from different fields. One of them is the analysis and detection of illegal activity, like terrorist communities and dark nets [2, 3]. At the same time, there are techniques for identification of statistical anomalies, which do not correspond to any certain topic and are presented as outliers in the general distribution of a set of systemic characteristic [4]. Online anomalous behavior can be also distinguished between temporal anomalies, local topological outliers, overall outliers or outliers in communities [5]. The vast majority of existing anomaly detection methods are aimed at discovering the suspicious activities of special types mainly based on probabilistic and linguistic methods (especially for anomalies, related to some special topics), or topological differences. Thus, actual structures and formation mechanisms of anomalous communities remain poorly studied.

ICCS Camera Ready Version 2018
To cite this paper please use the final published version:
DOI: 10.1007/978-3-319-93698-7_20

This paper contributes to the topological analysis of user communities in online social media. We consider a friendship graph of community subscribers, which (in some sense) represents the 'backbone' of a given community and determines the aggregated characteristics of information spread. To justify the differences between normal and anomalous communities, and between anomalous communities of different types, we perform the analysis for a wide range of VK communities, from hundreds to several dozens of thousands of subscribers, and compare average characteristics for varying number of users.

The rest of the paper includes a review of literature related to influence maximization and anomaly detection in online social media (see Section 2), the description of data used for analysis (Section 3), the results of a comparative study (Section 4), discussion and conclusion (Section 5).

2 Related Studies

One of the goals of abnormal communities may be an intensive informational-psychological influence on their users. In this way, the important role belongs to the influence maximization and to the formation of propagation-efficient network structures. A local topology of super-spreaders is thought to be related to degree properties, sums of neighboring node degrees, or is conditional on belonging to k-core [6]. In particular, Quax et al. [7] questions an influence of high-degree nodes, which is based on information theory approach along with Markov random fields. Pei et al. [8] note the degree distribution and PageRank do not characterize information spread enough, and these are k-core and sums of neighbors degrees to play an important role. Elsharkawy et al. [6] provide a dynamic simulation of information spreading with the evolution of a k-core, and suggest to consider k-core descendants to estimate a potential effect. They also show an influence of the relation between k-core and cascade size. In addition, a message content affects its contagiosity [9].[1]

Existing methods for detecting the suspicious communities do not discover the principles of network organization or the particularities of the patterns observed. The vast majority of them use machine learning techniques without identifying the factors playing a predominant role. Ratkiewicz et al. present a classification method for political astroturfing detection based on topological features, nevertheless, they do not disclose topological particularities of political communities aimed at informational influence [10]. Varol et al. [11] also build a classifier, based on network and diffusion features, user-based, timing, sentiment, content and language features (total of 487 features). Since they are correlated, authors also accompany the classification process with feature selection procedure.

The description of abnormal communities is met in Bindu research [12] suggesting spammer communities discovering algorithm. Firstly, they suppose spammers to have high local clustering coefficient with other spammers. Then, they

[1] Authors of [9] provide a classification method for rumor detection. In our case, this corresponds to relations between subscribers interests and message content.

ICCS Camera Ready Version 2018
To cite this paper please use the final published version:
DOI: 10.1007/978-3-319-93698-7_20

present an algorithm of spammer clusters detection, showing the clusters discovered have high clustering coefficient (0.15 for the network of 4000 nodes) and diameter of 9. Nevertheless, spammer communities are out of interest in current research since their behavior seems to be artificial and aimed at information spreading.

3 Data

Friendship graphs for different types of communities were analyzed. We have chosen certain thematical communities in social media vk.com (VK), collected their subscribers and then built a friendship graph for the subscribers (fig. 1). Two types of networks were considered: (i) a network of a single community, (ii) a network joining all subscribers for a single category (type) of communities (e.g., a network of all carders).

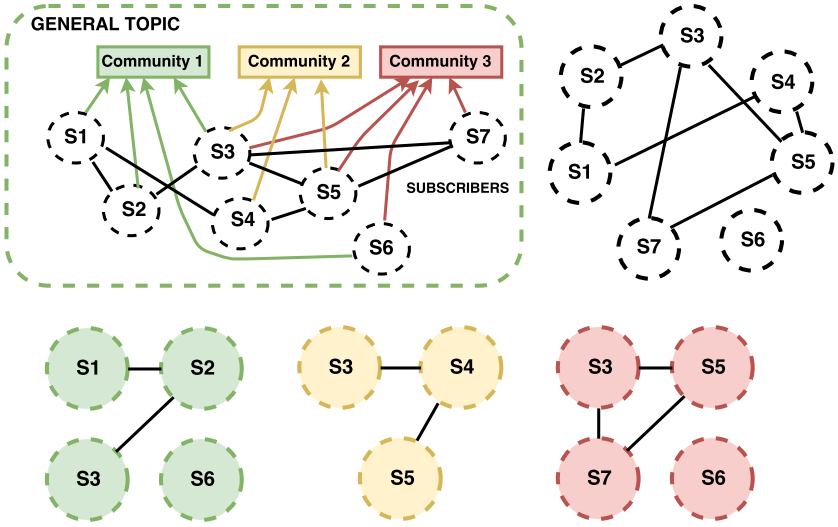


Fig. 1: Subscribers network in social media and resulting friendship networks, based on individual communities and general groups of communities

A set of normal communities ("Work in Saint-Petersburg", "Movie", "Humor", "Ideas for Handicraft and Gifts", etc.) contained about 630 units with maximum subscribers count of 62,949. Also, 630 anomalous communities were collected with maximal subscribers count of 176,527. The anomalous communities contained five following themes, which were marked manually during the stage of data collection: carders — 48 communities, substance users — 9 communities, substance sellers — 75 communities, people with radical political views — 43 communities, commercial sex workers — 210 communities.

ICCS Camera Ready Version 2018
To cite this paper please use the final published version:
DOI: 10.1007/978-3-319-93698-7_20

4 Results

4.1 Network topology of normal and anomalous communities

After data collection, we divide whole range of subscribers count into bins (table 1). As different bins contain the different number of normal and anomalous communities, for the analysis presented further we sampled the equal number of each community type from each bin (maximum possible, e.g. 66 for the third bin) to compare average characteristics for a bin. Sampling procedure for each bin was repeated 100 times to get more consistent results.

Subscribers count	#Normal	#Anomalous
(0, 1000]	227	368
(1000, 2500]	118	91
(2500, 5000]	83	66
(5000, 10000]	84	52
(10000, 15000]	64	26
(15000, 20000]	24	9
(20000, 30000]	18	8
(30000, 40000]	7	3
(40000, 50000]	2	1
(50000, 60000]	1	1
(60000, 200000]	1	4

Table 1: Distribution of a number of normal and anomalous communities by the number of subscribers

The analysis of different bins shows that density tends to decrease with increase in community size. Figure 2 demonstrates that small communities density is ten times greater than the density of the rest of communities. The reason of that may be that the formation of small communities is based on the existing networks of real or virtual friends; also, small communities present an opportunity of more intensive interactions between the participants due to their restricted number. Density in normal communities of medium and large is quite higher than in the anomalous communities and does not change significantly.

The comparison of other topological characteristics also shows significant differences (fig. 2). Anomalous communities have a lower number of links between their participants, which demonstrates a dive with the increase in community size. Clustering coefficient and average degree are also much greater in normal communities. These differences may be related to the fact that the audience of anomalous communities is often composed of strangers with specific interests while normal communities tend to unite people who are familiar with each other. In other words, people in anomalous communities often become friends because of their shared membership while networks of normal communities are mostly formed by previously existing friendship links. The clear illustration of

ICCS Camera Ready Version 2018
To cite this paper please use the final published version:
DOI: 10.1007/978-3-319-93698-7_20

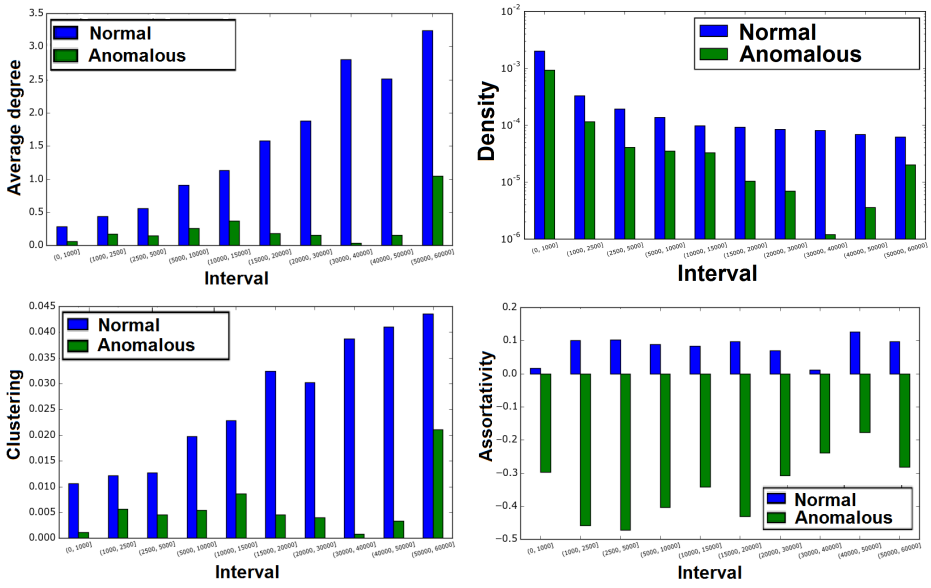


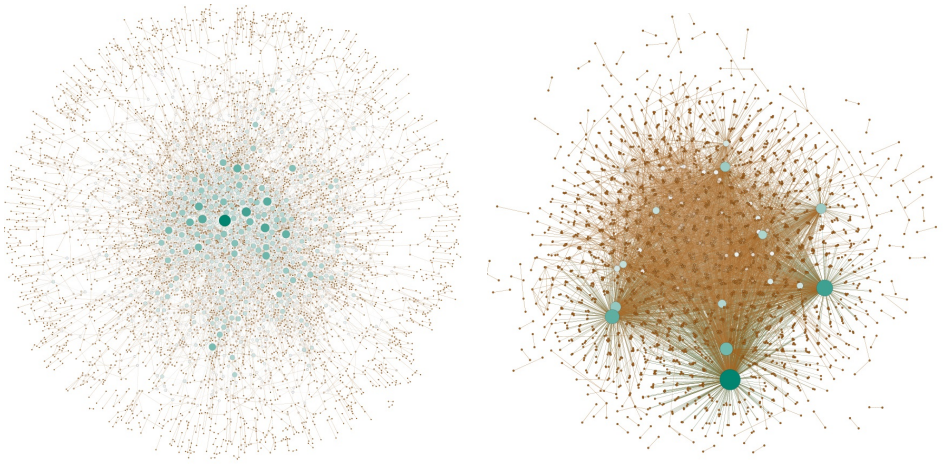
Fig. 2: Comparison of average topological properties for normal and anomalous communities

the difference in the mechanism of topology formation is represented in the plot with degree assortativity coefficients. This plot shows that normal communities are characterized by positive assortativity (as expected for ordinary OSN). In contrast, anomalous communities indicate strong disassortativity. This means that the subscribers with high degree (hubs) are more likely to be connected with low-degree subscribers.

Summarizing, one can see that normal and anomalous communities have drastically different topology. Figure 3 demonstrates the slow transition from high to low node degrees in normal communities, which corresponds to positive assortativity coefficient. This is associated with the necessity of communication prevailing in normal societies. For this case, the community is the space for sharing common interests, and this type of structure exists without special organization, without presenting any special services; therefore activities are mostly provided from bottom to top. The main intention of subscribers, in this case, is to communicate with other similar users. In contrast, anomalous communities often present some special services or content, which determines a specific structure of a network and roles of nodes. The existence of roles, which are impossible for being acted by an arbitrary subscriber, provides a strong distinction between participants, resulting in disassortativity.

Current analysis demonstrates a clear distinction between normal and anomalous communities, which is mostly explained by the degree assortativity charac-

ICCS Camera Ready Version 2018
To cite this paper please use the final published version:
DOI: 10.1007/978-3-319-93698-7_20



(a) Normal network (b) Anomalous network

Fig. 3: Normal and anomalous networks

teristic. These observed patterns allow for wondering how the goals of networks functionality are reflected in observed topological patterns. For this reason, the rest of the section is aimed at comparison of different types of anomalous communities.

4.2 Anomalous communities of different types

Since abnormal communities are aimed at providing different services, we suppose them to be formed by different mechanisms, and, consequently, their structural properties probably demonstrate systematic dependencies on the themes. To check this hypothesis we compare anomalous communities of different types. They are carders, in the context of credit card fraud, people with radical political views, commercial sex workers, substance users, and substance sellers. Figure 4 visualizes corresponding joint networks (connecting the users of all communities from a category) and demonstrates the following particularities:

- Carders friendship networks (fig. 4(a)) are extremely connected, with visually higher density than in communities of other types. One can see two strongly connected modules, corresponding to the union of high-degree nodes and to the union with lower degree nodes. In this way, a k-core, where all hubs (providing informational influence) are located, can be distinguished. Assortativity patterns are visualized similar to the anomalous community in fig. 3(b).
- Radicalized users demonstrate strong connections between different communities of this type. The majority of the hubs are concentrated inside the k-core, nevertheless, there are hubs attracting low-degree nodes to the pe-

ICCS Camera Ready Version 2018
To cite this paper please use the final published version:
DOI: 10.1007/978-3-319-93698-7_20

riphery. A network contains many hubs connecting isolated nodes of degree 1, which results in the negative assortativity.

– Commercial sex workers communities demonstrate more isolated structure (fig. 4(c)). There are many connected components and many separable modules inside the largest connected component. Hubs do not have so many neighbors as in communities of other types, which is reflected by the size and color intensity of green nodes. Modules in the largest connected component are not so interconnected as in radical users communities.
– Substance users network (fig. 4(d)) demonstrates quite high density and connectivity, and bigger hubs than in commercial sex networks. Number of hubs is much less than that in carder and political networks, and they are not concentrated in a core. On the opposite, the network looks more homogeneous. Nevertheless, there are segments weakly connected to giant component and containing own hubs. These modules correspond to separate communities. In this way, one can conclude, that substance users, as well as commercial sex workers, do not demonstrate strong interconnections between communities.
– Substance sellers network is similar to substance users one, for this reason, we do not present them both in figure 4.

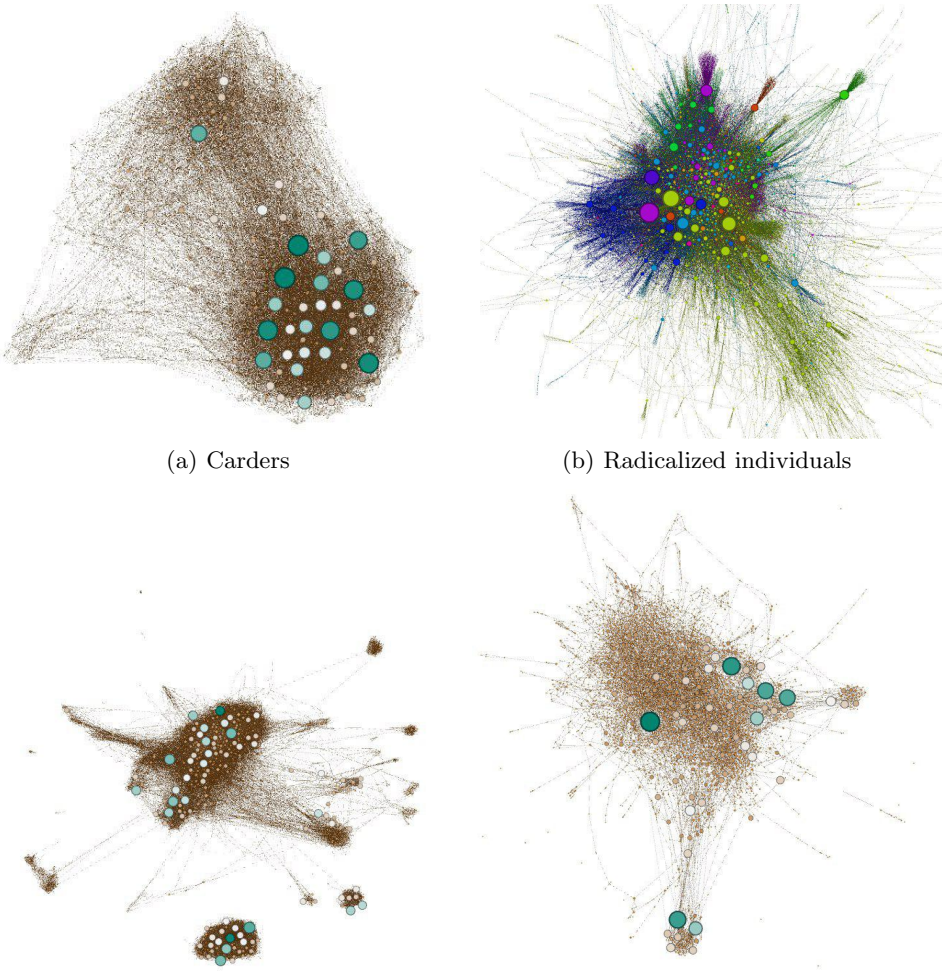
To disclose differences more accurately, the communities have been separated into several bins similarly with Section 4.1. After that, we obtained the distribution demonstrating which sizes of communities are prevailing in different thematic groups (fig. 5). In this way, one can see commercial sex workers group is described by the most representative data, and the community size of [0; 1000] is the most frequent. On the other hand, several community themes tend to be more concentrated in other bins. For example, considered radical politician communities are more common for [1000; 5000] subscribers; at the same time, for the range [10,000; 25,000] substance users are more prevalent.

It should be emphasized, that patterns in community size distribution demonstrated above are not correlated with other topological patterns described below, that means, further results are not reasoned by the imperfection of data used for analysis.

Topological analysis of all groups demonstrates quick dive in density with increase in community size, which is accompanied by a gradual increase in the average shortest path and gradual decrease in assortativity coefficient. One can notice, all types of communities demonstrate similar dynamics reflected by density, average path, and assortativity coefficients since the dynamics is mainly caused by general dependencies between network properties and its size. Clustering coefficient presents no strong dependencies between an increase in community size and network characterization.

Figure 6 shows that community of users with radical political views has four times higher clustering than others communities of more than 10,000 nodes. Their superiority is observed also for communities of 1,000–5,000 nodes. For other community sizes, substance sellers and users have the leadership. The

ICCS Camera Ready Version 2018
To cite this paper please use the final published version:
DOI: 10.1007/978-3-319-93698-7_20



(a) Carders (b) Radicalized individuals

(c) Commercial sex workers (d) Substance users

Fig. 4: Anomalous networks of different types

ICCS Camera Ready Version 2018
To cite this paper please use the final published version:
DOI: 10.1007/978-3-319-93698-7_20

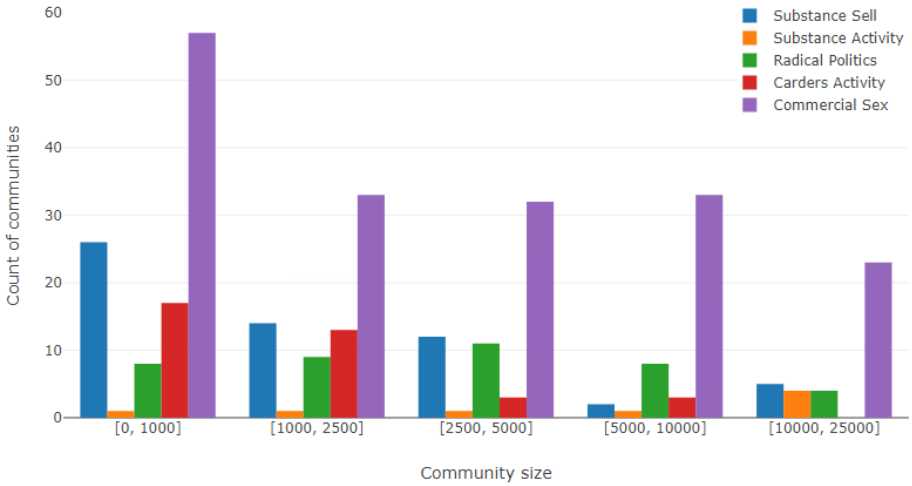


Fig. 5: The distribution of community sizes for different themes

smallest clustering is exhibited by commercial sex workers and carders, which can be due to the undesirability of identity broadcasting.

Density tends to decrease with the increase of community size for all types of abnormal communities. Commercial sex worker communities demonstrate values superiority for communities of 2,500–25,000 subscribers. Their density is several times greater. At the same time, smaller communities do not follow the same pattern. Shortest path naturally increases with the size of societies.

The example of small carders community of 500 nodes (fig. 7) demonstrates the particularities of its organization, which are potentially inefficient for information spreading, but can be quite enough for reaching their local goals.

The vast majority of hubs' friends are of degree 1. Several friends are connected with other hubs. In this way, disassortative mixing arises, since hubs are not connected directly, but they are connected via nodes with lowest degrees. The network contains several connected components. Components without hubs also have the nodes with maximal degrees connected via nodes with minimal degrees. This results in low clustering coefficient and high density as was presented in figure 6.

A plot with degree distribution (fig. 8) demonstrates all types of communities, normal and anomalous, follow a power law with different exponents. Heavy tails demonstrate different behavior due to networks size as well due to their organization. Users with radical political views have higher degrees and more heavy-tailed distribution, which differs with power-law exponent from other types of communities. Normal communities show shorter tails.

The main difference in the degree distributions is in their left part. Here we see straight lines for anomalous communities and rounded line for normal one. That means that normal communities have usually fewer nodes of degree 1, while

ICCS Camera Ready Version 2018
To cite this paper please use the final published version:
DOI: 10.1007/978-3-319-93698-7_20

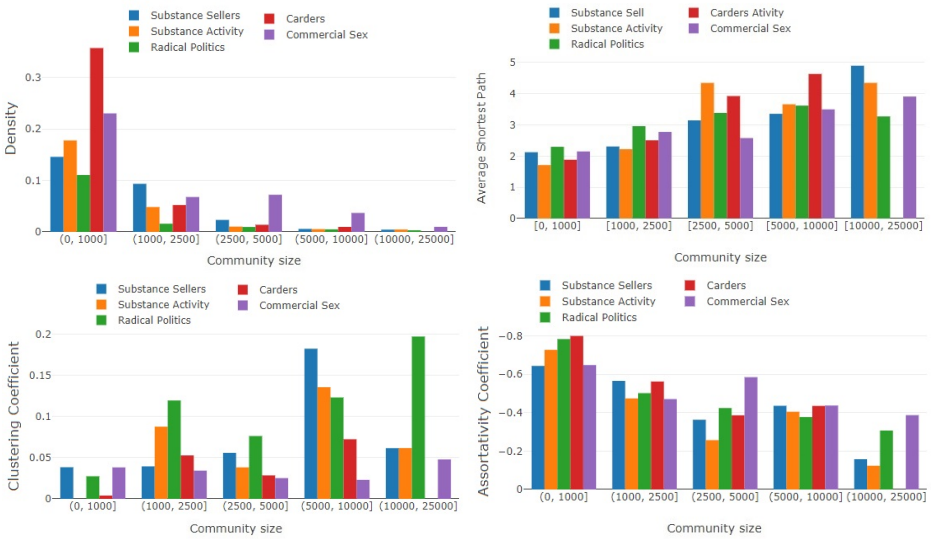


Fig. 6: Comparison of average topological properties for anomalous communities

they have more nodes with degrees 3–10. The relatively small number of users with a medium number of friends within the community and the predominance of hubs and nodes with a single link thus may be considered as distinctive features of anomalous thematic communities.

To liquidate the effect of network size we also analyzed a distribution of the proportion of connected nodes to all subscribers (fig. 9). Firstly, we consider a part of nodes in a friendship network for different size categories. People with radical views are the most "friendly" since the majority of subscribers are in a friendship network. Nevertheless, as it was shown in the previous section, this is due to hubs activity and their interaction with low-degree nodes. Substance users, carders, and commercial sex workers demonstrate less share of nodes in a friendship network. Secondly, we analyzed a cumulative distribution with probability of an arbitrary size network to have a certain part of nodes within a friendship network. One can see the similarity of all community types and sharp difference of political community. Figure 9(b) reflects the fact that users with radical political views have the significant share of nodes in a friendship graph of the thematic community more probably.

Summarizing, the analysis of different types of anomalous communities shows the significant difference of political communities since they have more friends in the subscribers' network, which is associated with lower assortativity coefficient, higher clustering, and strong interconnections between different communities inside the type.

ICCS Camera Ready Version 2018
To cite this paper please use the final published version:
DOI: 10.1007/978-3-319-93698-7_20

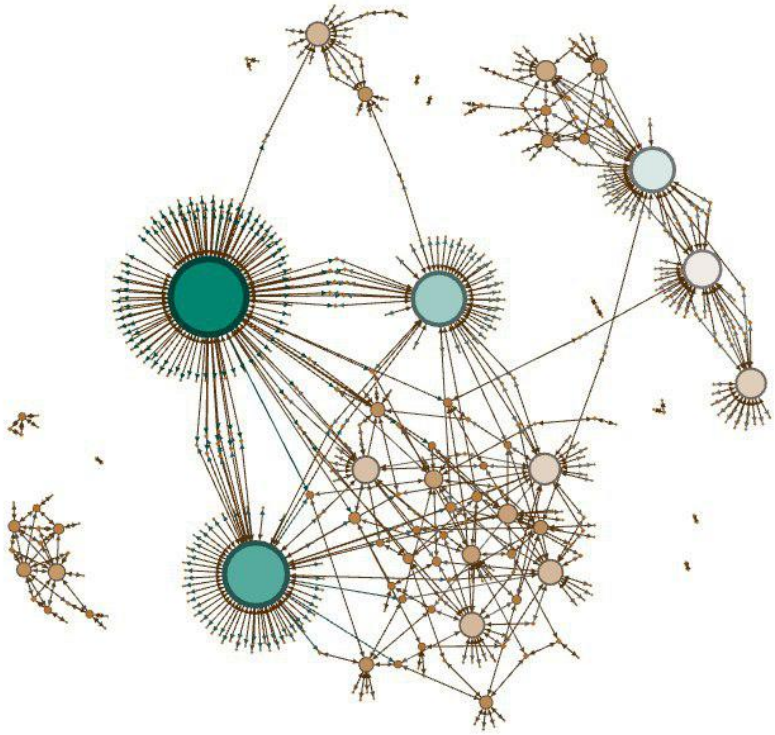


Fig. 7: Carder community of 500 nodes

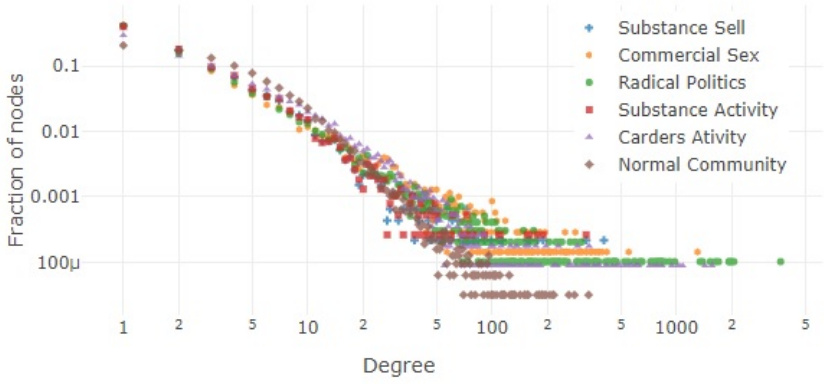
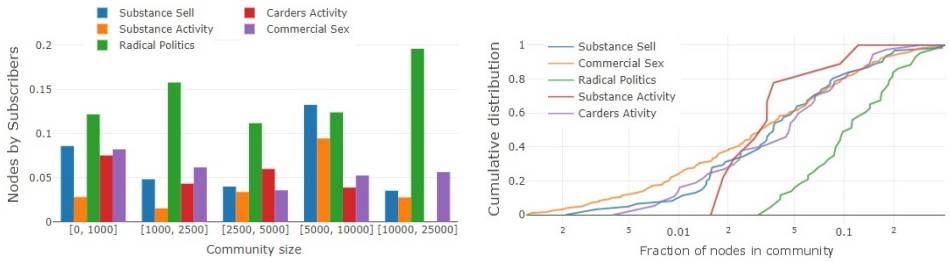


Fig. 8: Degree distribution for biggest connected component

ICCS Camera Ready Version 2018
To cite this paper please use the final published version:
DOI: 10.1007/978-3-319-93698-7_20



(a) Share of nodes in friendship networks

(b) Cumulative distribution

Fig. 9: Relation between the communities size and the corresponding friendship networks

5 Conclusion

Structure of social networks reflects particularities of agents interactions and allows for detection of the "goals" of a system functionality. In particular, it reflects systemic activity and "openness" of considered communities. In this study, we analyzed an extensive dataset from a largest Russian social network `vk.com` consisting of ordinary communities (having publicly acceptable thematic) and abnormal (anomalous) communities of different types.

The comparison showed that normal communities demonstrate positive assortativity coefficient, while negative assortativity is pertinent to abnormal communities. Analysis of different anomalous communities showed the division based on the involvement of subscribers to friendship networks. All considered groups were concentrated around the similar cumulative functions, while users with radical political views showed the highest percentage of users involved in the friendship network. At the same time, political communities demonstrated highest clustering values for groups of 1,000–5,000 and 10,000–25,000. The densest communities were demonstrated by carders (less than 1,000 subscribers) and commercial sex workers (more than 5,000 subscribers). However, for commercial sex workers communities, it was shown that subscribers are not involved in active interaction inside them.

A hypothesis explaining this behavior supposes that groups, providing a kind of services inside social networks, do not require any special broadcast and/or organization. At the same time, radicalized users are aimed at dissemination of their ideas, which implies more publicity. Consequently, their structure is more connected and clustered, and contain more subscribers and users in a friendship network, which should provide a sufficient opportunity for information propagation.

In this study, we showed that the thematic significantly influences the structure. It is even able to emerge opposite patterns of topology formation, as it was demonstrated by degree assortativity. The most interesting part here is that the resulting topology reflects the predominant "use cases" of community and roles

ICCS Camera Ready Version 2018
To cite this paper please use the final published version:
DOI: 10.1007/978-3-319-93698-7_20

of its subscribers within a given context. Uncovering of such hidden roles (as political opinion leaders and their devotees) is a next planned step of our research as well as studying the interplay between user similarity, their context-dependent roles, and characteristics of information spread.

6 Acknowledgements

This research was financially supported by Ministry of Education and Science of the Russian Federation, Agreement #14.578.21.0196 (03.10.2016). Unique Identification RFMEFI57816X0196.

References

1. Hu, H.B., Wang, X.F.: Disassortative mixing in online social networks. EPL (Europhysics Letters) **86**(1) (2009) 18003
2. Rowe, M., Saif, H.: Mining pro-isis radicalisation signals from social media users. In: ICWSM. (2016) 329–338
3. Lau, R.Y., Xia, Y., Ye, Y.: A probabilistic generative model for mining cybercriminal networks from online social media. IEEE Computational intelligence magazine **9**(1) (2014) 31–43
4. Akoglu, L., Tong, H., Koutra, D.: Graph based anomaly detection and description: a survey. Data Mining and Knowledge Discovery **29**(3) (2015) 626–688
5. Savage, D., Zhang, X., Yu, X., Chou, P., Wang, Q.: Anomaly detection in online social networks. Social Networks **39** (2014) 62–70
6. Elsharkawy, S., Hassan, G., Nabhan, T., Roushdy, M.: Effectiveness of the k-core nodes as seeds for influence maximisation in dynamic cascades. International Journal of Computers **2** (2017)
7. Quax, R., Apolloni, A., Sloot, P.M.: The diminishing role of hubs in dynamical processes on complex networks. Journal of The Royal Society Interface **10**(88) (2013) 20130568
8. Pei, S., Muchnik, L., Andrade Jr, J.S., Zheng, Z., Makse, H.A.: Searching for superspreaders of information in real-world social media. Scientific reports **4** (2014) 5547
9. Liu, Y., Jin, X., Shen, H., Cheng, X.: Do rumors diffuse differently from non-rumors? a systematically empirical analysis in sina weibo for rumor identification. In: Pacific-Asia Conference on Knowledge Discovery and Data Mining, Springer (2017) 407–420
10. Ratkiewicz, J., Conover, M., Meiss, M.R., Gonçalves, B., Flammini, A., Menczer, F.: Detecting and tracking political abuse in social media. ICWSM **11** (2011) 297–304
11. Varol, O., Ferrara, E., Menczer, F., Flammini, A.: Early detection of promoted campaigns on social media. EPJ Data Science **6**(1) (2017) 13
12. Bindu, P., Mishra, R., Thilagam, P.S.: Discovering spammer communities in twitter. Journal of Intelligent Information Systems (2018) 1–25

ICCS Camera Ready Version 2018
To cite this paper please use the final published version:
DOI: 10.1007/978-3-319-93698-7_20