NATIONAL RESEARCH UNIVERSITY
HIGHER SCHOOL OF ECONOMICS

Olessia Y. Koltsova, Sergei N. Koltcov,
Sergey I. Nikolenko

COMMENT-BASED DISCUSSION COMMUNITIES IN THE RUSSIAN LIVEJOURNAL AND THEIR TOPICAL COHERENCE

BASIC RESEARCH PROGRAM

WORKING PAPERS

SERIES: SOCIOLOGY
WP BRP 33/SOC/2013

Olessia Y. Koltsova[1], Sergei N. Koltcov[2], Sergey I. Nikolenko[3]

COMMENT-BASED DISCUSSION COMMUNITIES IN THE RUSSIAN LIVEJOURNAL AND THEIR TOPICAL COHERENCE[4]

We study the structure of online discussions in order to uncover latent communities of socially important debate. Our research reveals that discussion communities defined by mutual commenting in the Russian language blogosphere are centered mainly around blog authors as opinion leaders and, to a lesser extent, around a shared topic or topics. We have derived these conclusions from the dataset of 17386 full text posts written by top 2000 LiveJournal bloggers and over 520,000 comments that result in about 4.5 million edges in the network of co-commenting.

JEL classification: Z19.

Keywords: Comment-based communities, online discussion, blogs, Russia.

[1] National Research University Higher School of Economics, Director of the Laboratory of Internet Studies, PhD in Sociology. E-mail: ekoltsova@hse.ru

[2] National Research University Higher School of Economics, Laboratory of Internet Studies, Deputy Director, PhD in Mathematics and Physics. E-mail: kol-sergei@yandex.ru

[3] National Research University Higher School of Economics, Laboratory of Internet Studies, Senior Researcher, PhD in Mathematics and Physics. E-mail: snikolenko@gmail.com

[4] This research is supported by the Basic Research Program of the National Research University Higher School of Economics, 2013. The authors are grateful to Anastasia Shimorina for initial dataset preparation and to Eduard Ponarin for his methodological advice.

Introduction

Online discussions and networking have proven to be vitally important in the social and political life of contemporary societies. Sometimes they even have been crucial for political regime changes (Howard et al 2011; Lotan et al 2011). The structure of online discussions and communities that presumably arise around them presents an important and relatively new research problem for social scientists. If such discussions arise around specific topics, hot topics may be revealed by studying them. They may correspond to social problems or political issues important for the online public and therefore may lead to social conflicts or mobilization, so knowledge about such discussions may be used by policy makers. If discussion communities center around prominent personalities, such as popular bloggers, rather than around topics, this knowledge, apart from being used by policy makers, can also prove useful for marketing or political campaigning.

In this work, we focus on the following question: what drives the emergence of discussion communities in Russian-language blogs: their topical composition or their authorship? It has been widely alleged that blogs have ceded their leading positions in political influence to social networks. Judging by the growing number of users, one might indeed assume so (see e.g. Alexa statistics on the leading websites audience and traffic http://www.alexa.com/topsites), but in a broader socio-political sense it is hardly the case, at least in the Russian-speaking sector of the web. First, blogs, unlike much of the content of social networks, are available publicly and thus have a greater chance for stronger social impact; for instance, blogs are often further disseminated by regular media (Farrell & Drezner 2008). Second, Russian blogs, unlike social network accounts, are publicly rated with a number of indices which makes finding popular items in them much easier both for journalists and for ordinary seekers of alternative points of view (see Yandex rating of blog services at http://blogs.yandex.ru/services/). Since most public interest discussions in Russia have been housed by the LiveJournal blogging service (LJ) (Etling et al 2010), and since it is mostly LJ blogs that inhabit the public "top" of the Russian blogosphere ratings, this service has become the platform of choice for our research.

In blogs, unlike forums, discussions have no place to develop other than in threads of comments to individual diaries that are, unlike forums, not labeled in terms of topics. Thus, author-based discussions may develop in multiple threads of comments for the same post or for different

posts of the same author, while topic-based discussions may involve posts scattered around multiple authors. Therefore, discussion communities – that is, groups of people discussing something with each other – may be latent even when they exist. Our assumption has been that bloggers do unite into comment-based communities – by which we mean they may be divided into groups tending to comment approximately the same sets of posts, either posts on the same topic or of the same blogger, and to develop discussions around those sets of posts. Thus, these posts may be also said to form comment-based communities, where communities are understood as "denser" fragments of networks connecting posts that have common commenters. The goal of our present work is to test whether indeed such communities are present, and if so, whether they form around authors or around topics.

What is a community

The sociological notion of a community has been strongly challenged with the advent of the Internet in many respects; for a detailed discussion, see, e.g., (Smith & Kollock 1999; Wellman et al 2002; Gruzd et al 2011). In relation to our goals, the use of the concept of a community in social sciences may be divided into two major classes, not entirely mutually exclusive. The first class refers to communities as self-nominated groupings based on self-identification of participants (examples of such communities on the web include Facebook groups and LiveJournal's community blogs); this is characteristic of both traditional sociology stemming from Tönnies and the recent surge of literature on "virtual communities" (e.g. Rheingold 2000; Porter 2006). The other view, which is more characteristic of the social network analysis (SNA) tradition, describes communities as systems of empirically detected connections / interactions of a certain kind; see, e.g., (Smith et al. 2010). Members of such communities may be unaware of which communities they belong to; moreover, these members may not be human at all and may be represented by organizations, countries, texts, purchases, and other social objects. Several types of social connections, technically embodied in hyperlinks, may be found in blogs and online social networks: declared links (friends, followers, subscribers, members of self-nominated groups), comments, and in-text hyperlinks. Some links are not, or are not always technically expressed; for instance, borrowed content does not always have a clear and unified indication of being borrowed and is not always linked to its source,

which makes it significantly harder to study. For the purposes of revealing communities in blogs bonded by real activities, the closest matching type of a link is a *comment*; online comments are technical embodiments of a speech act addressed to a counterpart. Self-nominated bonds are static and therefore often inactive; furthermore, they usually do not include "negative" ties with opponents, while discussions may occur among users holding different views. When patterns of mutual commenting are represented in a graph, all methods of social network analysis and graph theory may be applied to them afterwards.

In graph theory, two main types of definitions of a community may be discerned: local and global (Fortunato 2010). *Local* definitions define a community in relation to a pre-determined threshold (e.g., density threshold) or to the closest environment (e.g., minimal ratio of internal and external total degree which shows whether vertices within a community have more links inside or outside the community). Local definitions include such notions as n-cliques, n-clans, n-clubs, K-plexes, k-cores , LS-sets, lambda-sets and others (see, e.g., Hanneman 2005); these notions are well-known to SNA sociologists. However, they are suboptimal for large sparse networks like the Internet, although this approach has been used for small-scale online communities (Chin & Chignell 2006).

In large networks, *global* definitions are usually applied; they define communities in relation to the whole graph. Most generally, communities in such approach are defined as subgraphs whose density is significantly higher than would be expected in a random graph. More precisely, a community is usually defined as the result of a specific community detection algorithm. Random graphs used in such comparisons are modeled to have the same number of vertices and edges as the real networks in question, but the edges are assigned to the vertices randomly. Since community detection algorithms are usually quite complex, the most transparent criteria for the choice between them are their quality and/or scalability confirmed in experiments, which in our case has led us to almost the only possible option: the Louvain algorithm (Blondel et al. 2008) that has been very efficiently implemented by its authors (source code available at https://sites.google.com/site/findcommunities/).

Thus, in this study we define a community as a segment of a network where the quantity of edges is significantly larger than should be in a random graph of the same size. Comment-based communities arise in the blogosphere or in an online social network when the same set of posts is

commented by approximately the same set of users. A comment-based community here is a community detected in a graph whose edges denote instances of commenting: two posts share an edge if they have received a comment from the same blogger.

Related work

Most network analysis on the web, including community detection, has been devoted to links of a different nature than commenting. A large portion of this research has been mathematical in nature and has been devoted to developing community detection algorithms in graphs, where the web data is only used for testing; a comprehensive survey of such literature is beyond the scope of this paper, and we recommend, e.g., (Fortunato 2010). We also leave out studies where community detection algorithms are in fact used as mathematical approaches to the purposes of clustering, e.g for retrieving clusters of online consumers and respective sets of purchasing interests.

Earlier studies of web-based social networks usually tried to characterize general patterns of the web graph based on in-text hyperlinks between webpages (Albert et al 1999), and some of them addressed the issues of densely connected components (Broder et al 2000) or communities, and have had a modest sociological component. Studies from the social or political science have been often using visualization algorithms to detect hyperlink communities "by eye"; sometimes this approach has been successful, e.g., Adamic & Glance (2005) thus demonstrated the polarized character of the US political blog space, while the Berkman center at Harvard launched a series of studies mapping the Iranian, Arabic, and Russian blogospheres and identifying different political groups in them (Kelly & Etling 2008, Etling et al 2009, Etling et al 2010). Later a new type of studies evolved that employed community detection algorithms for sociological studies of web data; one example of such a study correlates types of ecological concerns with hyperlink-based communities of websites of ecological social movements (Ackland & O'Neil 2011).

Another frequent type of links used in the studies have been, so to say, person-to-person links, that is, self-declared links between personal accounts. Such links may be either undirected (e.g., "friendships" in Facebook or "connections" in LinkedIn) or directed (e.g., "followers" in Twitter and Academia or "friending" in LiveJournal). Here LiveJournal friendship-based communities have already attracted some attention (Zakharov 2007; Lescovec et al 2008). The latter

paper revealed that the best separated communities happen to be of the size around 100 nodes in a wide range of different networks: hyperlink webgraphs, co-citation networks, online friendship networks, and some others. Larger communities are less discernable and more integrated into the largest component of the network that has no obvious underlying geometry. One of the rare examples of large-scale dynamic network studies (Kumar et al 2010) has explored evolution and merge of friendship communities in Flickr and Yahoo! 360 social networking platforms. Another type of person-to-person link studied has been in-text mentionings of persons, including those occurring in texts of commenters (Gruzd 2009).

However, while many important discussions in blogs develop in comments, not only comment-based communities but even comments in general have received relatively little attention from researchers. Among a few relevant non-community studies of comments we can list the following: Yano and Smith (2010) proposed to predict the volume of comments to political blogs with topic modeling; Mishne and Glance (2006), having outlined some general characteristics of weblog comments, also offered a method of detecting discussions, i.e., disputative sequences of comments in threads treating the task as a text classification problem; Ali-Hasan and Adamic (2009) studied comment links along with blog-roll links and citations in blogs and found significant overlap between them, but the communities they detected were not comment-based.

One of the first studies of comment-based communities (Chin & Chignel 2006) made a valuable attempt to merge graph-based notions of a community and social science concept of community as a self-nominated grouping. However, the scope of that research was very limited and the final method, based on local measures, quite unclear. The research which is most relevant to this work is a large-scale study of comment-based networks in Slashdot news website (Gomez et al 2008). In line with Lescovec et al (2008), they have found out that, due to the network's sparseness, communities are multiple and small, with a single giant component quickly absorbing middle-size communities in the process of hierarchical graph clustering.

To the best of our knowledge, all studies of comment-based communities, including those mentioned above, have had authors / bloggers as nodes in the networks. Most often, two commenters were considered connected if they have commented each other. Some other ways of constructing comment-based networks include: commenters who get a common edge if they have commented (a) the same post or (b) the same blogger, or, on the contrary, (c) post authors who get a

common edge if they have received a comment from the same commenter. When community detection methods are applied to such networks, one can obtain: communities of people commenting either (a) the same posts or (b) the same bloggers or (c) communities of people commented by the same commenters. The version (c) is the furthest from our goals; in the case of (b) it is hard to determine the degree to which communities tend to center around authors and to isolate other possible factors that may have a stronger effect on community formation. The version (a) lets us find whether clusters of commenters are centered around subsets of posts, but no information about posts' content is available, since the network does not contain post IDs. Jamali and Rangwala (2009) who studied comment-based networks at Digg have discovered a dependence between commenters and topics of commented posts, but not through network analysis, but by juxtaposing "hand-coded" topics of posts with IDs of commenters. Each commenter, as they revealed, commented across a wide range of topics. Therefore, the authors obtained the topical compositions of sets of posts commented by each blogger, but did not learn (and did not intend to learn) if bloggers could be united into communities of commenting based on topics of the commented sets of posts or on other parameters. By contrast, Qamra et al. (2006) have proposed an algorithm that finds sets of posts simultaneously united by a shared topic, extensive mutual hyperlinking, and proximity in publishing time. It thus reveals "hot topics" that are actively discussed in temporary communities of interested bloggers. However, this approach filters out non-topic-based communities *a priori* and thus also does not address the question of how much the hyperlink communities tend to form around topics or other factors.

Data and methods

The data were retrieved from the Russian language LiveJournal website via its API into an MS SQL database with the Koltran BlogMiner downloading software developed by the authors. Russian LJ maintains a publicly available list of Russian language accounts rated by three different criteria. We used the so-called "social capital" rating list which, although it is not explicitly stated by its developers, uses the general idea of Pierre Bourdieu as well as the general idea behind PageRank. Its methodology is not fully available and represents a commercial secret but, generally, it counts people who have befriended a given blogger favoring those who really read it on a

permanent basis. It also uses a number of penalizing coefficients whose purpose is to fight various methods for artificial boosting of the social capital (since social capital can be monetized, various forms of blog optimization similar to search engine optimization for web sites have arisen). As a result, the top of the rating list contains accounts that are highly active, read, and commented, and bots rarely can get to the top. Our downloading experiments have shown that the number of posts per blogger and especially the number of comments per post drop very fast as we move down the rating list: there are 2 million accounts in the Russian language LJ in total, but already at the level of places around 50,000 in the rating list comments are too few to construct a network, and bots are quite apparent. The first 2000 bloggers (approx. 0.1% of all accounts) usually attract 20 times more comments than they write posts: this is quite sufficient for a meaningful graph, although the threshold is, of course, conventional.

Another conventional threshold is time limit: how many days, weeks, or months should we include into our network analysis? Ideally, it would be desirable to conduct a series of experiments with a moving window in time and with a varying width of the window, to detect which period produces best-discernable communities on the most permanent basis. However, we did not have sufficient computational resources for this, so we hypothesized that a one week period would be a plausible candidate since we observed that most posts get the majority of their comments in a few days. Longer periods might add more permanent blogger-based links, but would hardly be suitable to detect topic-based communities, especially if topics are events. Going months back would have also added links to posts of bloggers that are really no longer read by a given commenter and thus produce false communities. Choosing a week for analysis, we have also ensured that it had no major events, like national elections or large-scale disasters that might have skewed the topical or community distribution.

The data used in this research includes all posts by top-2000 bloggers for one week between April 1 and 7, 2013, as well as the relational structure of their comments (who commented which post and how many times). After clearing and excluding uncommented posts the resulting graph contains 17386 vertices (i.e., posts, written by 1667 authors) and around 4.5 million edges derived from approx. 520,000 comments left by about 56,000 commenters. Two posts get connected by an edge every time they have been commented by the same blogger, which is actually a unimodal

projection of a bimodal post-commenter network; we have used this projection because there are virtually no publicly available community detection algorithms for large weighted bipartite graphs.

Among those available, the Louvain algorithm is not only the most scalable, but also has the best quality in comparison with other modularity-optimizing algorithms, according to the tests performed by the developers (Blondel et al 2008). Modularity, a measure ranging from 0 to 1, is the most widely used quality function in community detection that is optimized in a number of popular algorithms. Having had many subsequent extensions, this measure was originally introduced in (Newman & Girvan 2004) where it was defined as the fraction of the within-community edges in the network minus the expected value of the same quantity in a network with the same community divisions but random connections between the vertices (Newman & Girvan 2004: 7). The Louvain algorithm uses the extension of modularity for weighted graphs. This algorithm scales well because, having done the initial partition, it then treats the revealed communities as single vertices and merges them in two more phases. However, as we saw, in line with observations of Lescovec et al (2008) and Gomez et al. (2008), the second and third phases tend to blend the middle-sized communities into one giant component, leaving the smallest and the least interesting communities intact. Therefore we used the results of the first phase (level).

To detect topical similarity of texts within and outside communities, we used the classical bag-of-words approach: texts were considered thematically similar if they shared a large amount of words, and each text was treated as a multiset of words, discarding their sequence. We represented each text as a vector whose components corresponded to the frequencies of words occurring in it; we used weighted frequencies known as tf-idf (term frequency – inverse document frequency) measure; see e.g. (Manning et al. 2008). Prior to calculating them, each text was cleared of html tags and other special symbols and then lemmatized with the Yandex Mystem lemmatizer (Segalovich 2003); the software is freely available at http://company.yandex.ru/technologies/mystem. Then, we used two alternative methodologies: cosine similarity calculation and topic modeling with the LDA model.

Cosine similarity is a measure widely used to calculate the proximity between texts for text clustering and for information retrieval. The cosine of the angle between a pair of vectors representing those texts is assumed to measure the similarity between them; for details, see, e.g., (Manning et al. 2008). Using cosine similarity, we computed average distances between all texts

within comment-based communities and the global average distance. One disadvantage of the cosine similarity measure is that it tends to assign zero similarities to most pairs (namely, to every pair of documents that have no shared terms), which is why we also used topic modeling. This approach views topics as latent variables, akin to factors, whose distribution over words and texts is simultaneously modeled. The output of the algorithm we used (Griffiths, Steyvers 2004) includes two matrices: the term-topic matrix and the topic-document matrix, where cells contain probabilities of "words in topics" and of "topics in documents", respectively. Thus, each text is represented as a probability distribution over topics, and each probability can be considered as a "weight of importance" of a particular topic in this text.

After we obtained the topical composition of each text, weights of all topics in texts belonging to the same comment-based community were totaled, getting the relative importance of each topic for each community. We hypothesized that while some communities might be equally distributed across all topics, in others importance of only a few topics would peak, therefore intra-community variance of topics' weights in each community was calculated and normalized to the range (0-100). This gave us a possibility to treat communities with low variance as topic-independent, and communities with high variance as mono-topical or at least topic-centered. Besides that, topic modeling provides a possibility to judge not only about topical similarity, but also about the content of topics by considering the most probable words / texts in them. This was done by hand coding of topics and of communities with the highest topical variance by two coders who first worked independently and then agreed on their labels.

Results

Community detection has revealed a moderately manifested but clearly evident community structure with modularity $Q = 0.38$ and a highly skewed distribution of community sizes, the largest community comprising more than half of the vertices (9976 out of 17386) (Figure 1). This corresponds to the findings of (Gomez et al 2008) and (Lescovec et al 2008) addressed above. A large number of small communities (85) are isolated pairs and triads of little interest; this is the result of the highly skewed distribution of comments per post and especially per commenter. Around one third of commenters have left only one comment, thus not participating in the comment-based

network at all, while most of it has been formed by less than a thousand commenters who generated from a hundred to more than a thousand comments each, and two thirds of all comments together. However, about 70 middle-size communities are potentially interesting. Analysis of dependence of posts' belonging to a community on their authorship has revealed a strong positive correlation (Table 1).

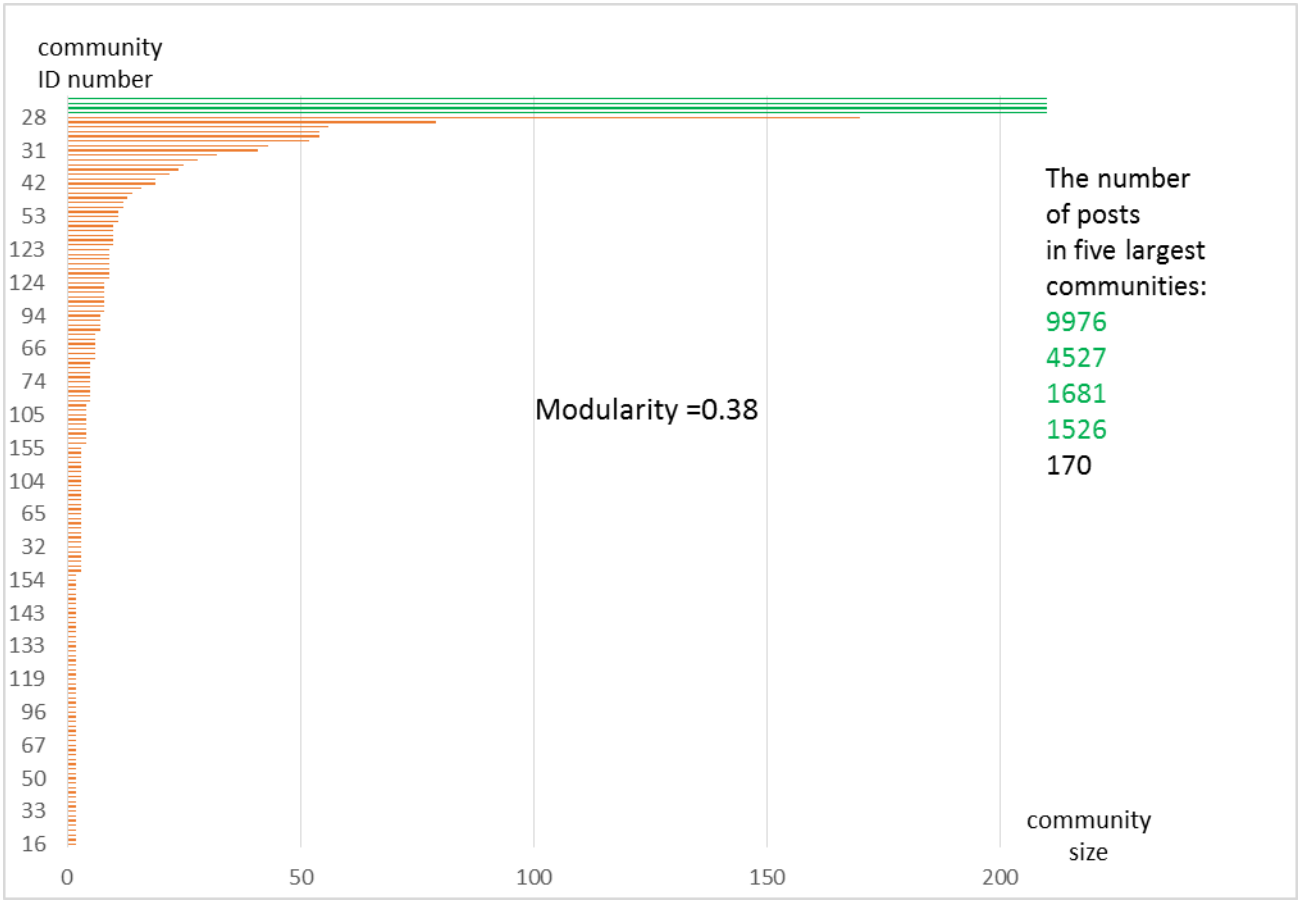

FIG.1. NUMBER OF POSTS IN COMMUNITIES: COMMUNITIES 0-158; NUMBER RANGE: 2-9976. LOUVAIN ALGORITHM, LEVEL 1. GREEN: BARS STRETCHING BEYOND THE PICTURE.

Next, we calculated all cosine similarities between each pair of texts and obtained the following averages: average similarity within each community, average intra-community similarity (491,7E-4) and global average similarity (1,6E-4) As can be seen, similarity between two texts assigned to the same community is on average two orders of magnitude higher than the global average. This difference is statistically significant as determined by one-way ANOVA; however, it is known that ANOVA produces very large values of F-test with large samples (in our case, the total

number of cosine distances within all the communities is more than 53 million), and thus may assign statistical significance to very small differences.

At the same time, the distribution of intra-community cosine similarity means is highly skewed, with a minority of communities being highly above the global average and a vast majority only slightly above or even slightly below the global average. The middle part of this distribution is shown on Figure 2, where 0 on Y axis is global cosine similarity average, X axis contains communities sorted by their average cosine similarities. Average intra-community similarities do not correlate with community size, number of bloggers who authored the community's posts, or average post length. The skewedness of the distribution and the absence of an obvious explanation for it has led us to further exploration of various communities' properties.

TABLE 1. DEPENDENCE OF POSTS' BELONGING TO A COMMUNITY ON THEIR AUTHORSHIP.

		Value
Lambda	Symmetric	.209***
	Dependent blogger	**.057***
	Dependent community	.522***
Goodman & Kruskal Tau	Dependent blogger	**.041***
	Dependent community	**.510***
Cramer's V		**.466***
Contingency coefficient		**.985***

Note: The symbol *** denotes 2-tailed statistical significance of less than 0.001.

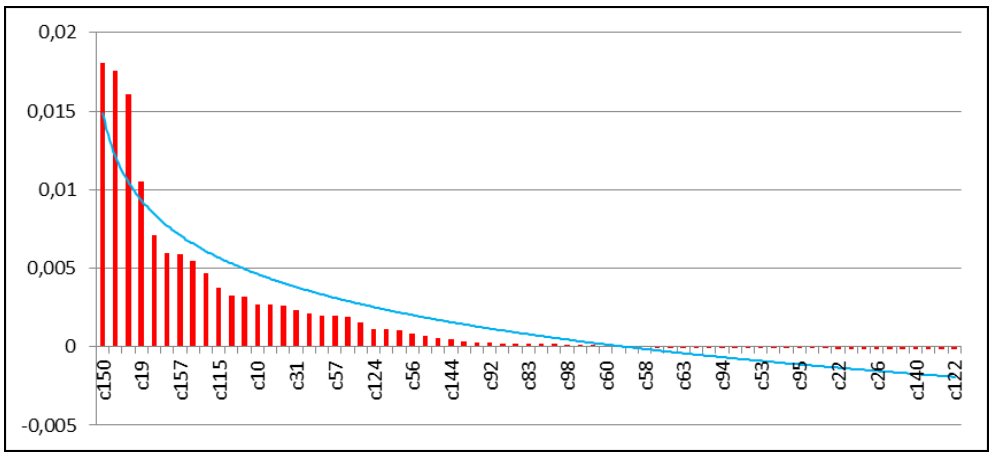

FIG. 2. DISTRIBUTION OF INTRA-COMMUNITY COSINE SIMILARITIES IN COMPARISON WITH GLOBAL AVERAGE (FRAGMENT).

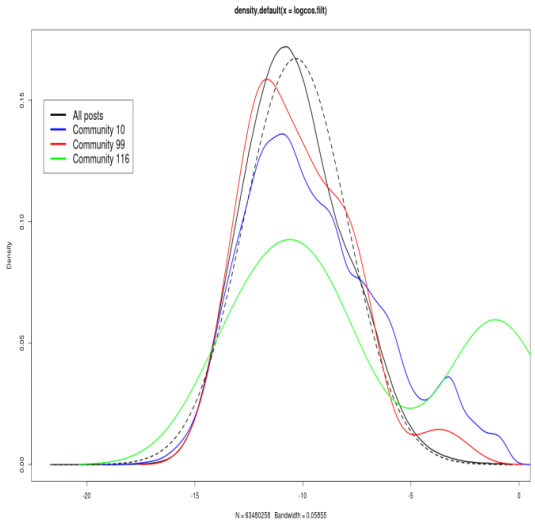

FIG.3. DISTRIBUTIONS OF LOGARITHMS OF COSINE SIMILARITY GLOBALLY AND IN SOME COMMUNITIES.

The distribution of logarithms of cosine similarity (Fig. 3) shows that while globally they clearly follow a bell-shaped distribution (black line), some communities that stand high above the global cosine similarity average produce additional peaks shifted closer to the higher values of cosine similarity (X axis). Selective analysis of communities (see examples in Table 2) shows that those with cosine similarity above the average tend to be (albeit not always are) dominated by a set of posts covering a roughly similar set of issues and written by the same author or by a very limited

set of authors, while a relatively large number of disconnected posts by a large number of authors "sticks" to this relatively coherent core. Presumably, it is this core that produces additional peaks in Fig. 3.

For better detection of such cores, topic modeling (N of topics =100), LDA with Gibbs sampling algorithm (authors' LINIS TopicMiner software) was then performed on the dataset. Hand-coding of topics revealed no substantial difference in the topic composition of the dataset, as compared to datasets covering other periods that had been studied by the laboratory in previous projects. Topics were approximately evenly divided between public affairs, including some event-driven topics, and private, recreational and consumption issues. Number of uninterpretable topics did not exceed 20%, which is lower than before and is mostly connected with gradual increase of quality of text preprocessing.

TABLE 2. EXAMPLES OF HAND-CODING OF COMMUNITIES.

Comm ID	Num of authors in comm	Num of posts in comm	Rank by avg cos sim	Description
c154	1	2	2	author: sontucio, one post is a cut version of another
c86	5	8	10	culture and privacy
c150	2	9	13	author: bragin_sasha - on politics in Ulianovsk region
c39	5	20	17	dominant author: lumbricus, where she went and what pictures she took
c52	8	43	21	15 natashav, 7 orange_sky_bird, 14 pelageya, most are women; dominant topics: maternity, pregnancy, women's problems; other private issues are present
c7	14	48	24	29 posts by hope1972, dominant topic: popstars and films; others also have a mixture of other issues.
c10	262	1135	25	Post/author distr. - power law, short posts (mean 83 words against global mean 375), private messages dominate

Next, total weight of each topic for each comment-based community and the normalized topic weight variance in each community were calculated, as described in the methodological section. The largest community containing more than the half of vertices naturally had the smallest

variance, while among other communities different types could be observed (see examples in Fig. 4 a-d).

Y axis in all figures shows the weight of topics in % of the total topic weight of each community, so large communities scoring high in all topics in absolute numbers are comparable to small communities. Fig. 4a illustrates the diversity of topical "profiles" of communities with three examples: the community dominated by a single topic, another one dominated by a small number of less pronounced topics, and the giant component (community 0) whose topical distribution is close to the global distribution. Figures 4b-d show these examples separately, and topics are sorted by their weights. Labels of topics tend to coincide (although not perfectly) with labels of communities in which their presence is visible through the analysis of topical variance: that is, hand-coding of texts belonging to community 13 assigned it to the topic "books", while independent hand-coding of the topic 27 that dominates the community assigned it the same label.

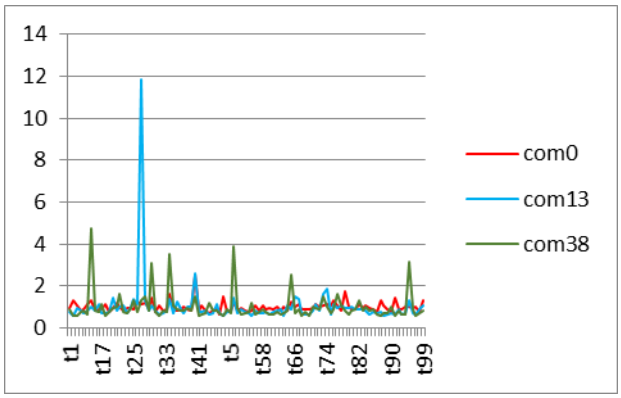

FIG. 4A: DISTRIBUTIONS OF TOPICS' WEIGHTS IN THREE SELECTED COMMUNITIES

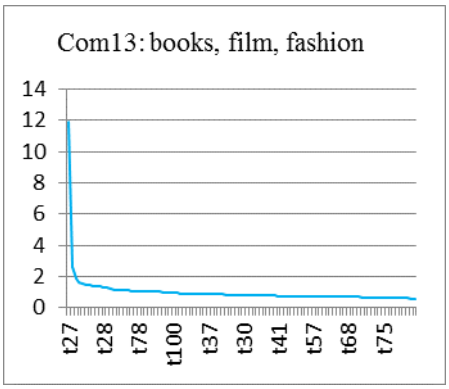

FIG.4B: DISTRIBUTION OF TOPICS' WEIGHTS IN COMMUNITY 13, 24 POSTS.

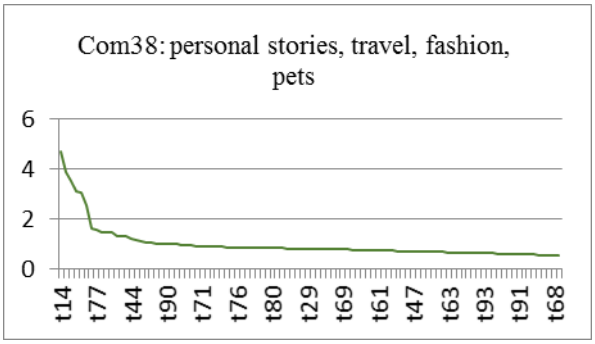

FIG. 4C: DISTRIBUTION OF TOPICS' WEIGHTS IN COMMUNITY 38, 56 POSTS

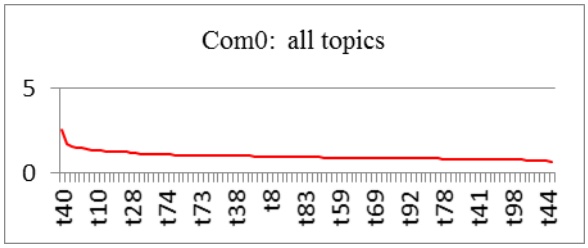

FIG. 4D: DISTRIBUTION OF TOPICS' WEIGHTS IN THE GIANT COMPONENT (COMMUNITY 0), 9976 POSTS

At the moment, we have found no correlation between proximity of texts in a community as determined by cosine similarity and topical variance within a community. A possible explanation might be that communities with high topical variance might be dominated not by a single, but by a number of topics, not necessarily similar to each other. This may push the communities containing completely different texts that simultaneously differ much from the global topical distribution to the top of the "rating" of topical variances. It means that although potentially topic modeling may help find communities dominated by a small number of topics and determine which topics they are, this issue demands further extensive study.

Conclusions and implications for further research

The research contributes to the social studies of online communities being the first study of communities based on mutual commenting while simultaneously examining the commented posts and their topics. It thus links commenting activity of bloggers to what and who is commented and uncovers some principles of formation of comment-based groupings. It confirms its initial hypothesis and finds that people commenting top LJ bloggers tend to unite into moderately manifest communities by (unintentionally) commenting roughly the same sets of posts. The graph of co-commenting is sparse and connected by a minority of active commenters that tend to be non-top bloggers themselves, thus indicating the predominance of fandom commenting in the top LJ. The research also confirms that communities strongly tend to emerge around authors of posts, who thus may be treated as opinion leaders of a new type. However, to a less visible degree communities form around topics of posts. Topical coherence of some communities is presumably connected with the topical coherence of the author (or a small number of authors) dominating these communities; testing this assumption has been an unresolved methodological problem and thus a limitation of this research. A few communities are obviously dominated by a single topic or a small number of related topics, while a large number of communities are not topically coherent at all. This is an important finding about the nature of discussions in blogs, telling us that topic-centered discussions, although they do exist, are not the dominant type.

This poses further questions for research: what is the proportion of topic-based discussion communities? is it sufficient to be worth searching for? if so, what is the best way to differentiate such communities from "noisy" talk of everybody about everything? Topical coherence measured with mean cosine similarity and with topical variance as determined through topical modeling are possible candidates for such methods, however, their match with hand coding has not been perfect so far, and hand-coding experiments themselves have to be carried out more thoroughly to be sufficient for reliable conclusions. This limitation of the research is coupled with another one connected to its relatively small scale: it has yet to be replicated on other datasets (of different periods, time windows and sizes) to ensure repeatability of the issues found. Its conclusions may be immensely enriched if it will be supplemented with an analysis of the composition of commenters.

Thus, at this stage of the research, since it is clear that topic-centered communities do not dominate, it is not advisable, for practical purposes, to try finding hot latent topics in blogs via detecting comment-based communities. Rather, policy makers and marketing practitioners might search for clusters of bloggers able to generate communities of active co-commenting. If such bloggers are made to raise certain issues, they are likely to provoke a lavish feedback of their partially overlapping audiences that can be used for different purposes: from preliminary screening of public reaction to mining new ideas by studying their "wisdom of crowds" to attracting attention to socially important problems.

REFERENCES

Ackland A., O'Neil M. (2011) Online collective identity: The case of the environmental movement // Social Networks, 33(3): 177-190.

Adamic L.A., Glance N. (2005) The political blogosphere and the 2004 US election: divided they blog // Proceedings of the 3rd international workshop on Link discovery. P.36-43.

Albert R., Jeong H., Barabási A.-L. (1999) Diameter of the world wide web. Nature, 401, P.130-131.

Ali-Hasan N., Adamic L.A. (2009) Expressing social relationships on the blog through links and comments // International conference on weblogs and social media. – San Jose, CA, USA.

Blondel V.D., Guillaume J.-L., Lambiotte R., Lefebvre E. (2008) Fast unfolding of communities in large networks // Journal of Statistical Mechanics: Theory and Experiment, P 10008.

Broder A., Kumar R., Maghoul F., Raghavan P., Rajagopalan S., Stata R., Tomkins A., Wiener J. (2000) Graph structure of the web // Computer Networks. 33. P.309-320.

Chin A., Chignell M. (2006) A social hypertext model for finding community in blogs // HYPERTEXT '06: Proceedings of the seventeenth conference on Hypertext and hypermedia. P. 11 - 22. New York, NY, USA, ACM Press.

Etling B., Alexanyan, K., Kelly, J., Faris, R., Palfrey, J. and Gasser, U. (2010). Public Discourse in the Russian Blogosphere: Mapping RuNet Politics and Mobilization // Berkman Center for Internet and Society Research Publication. 2010. http://cyber.law.harvard.edu/publications/2010/Public_Discourse_Russian_Blogosphere (accessed 30 September 2013).

Etling D., Kelly J., Faris R., Palfrey J. (2009) Mapping the Arabic Blogosphere: Politics, Culture and Dissent // Berkman Center for Internet and Society Research Publication No. 2009-06. http://cyber.law.harvard.edu/sites/cyber.law.harvard.edu/files/Mapping_the_Arabic_Blogosphere_0.pdf (accessed December 1, 2013).

Farrell, H., Drezner, D.W. The power and politics of blogs. Public Choice (2008) 134: 15–30. DOI 10.1007/s11127-007-9198-1.

Fortunato S. (2010) Community detection in graphs // Physics Reports. Vol. 486. Issue 3-5. – Elsevier B.V. P. 75-174.

Gomez V., Kaltenbrunner A., Lopez A. (2008) Statistical analysis of the social network and discussion threads in Slashdot // WWW '08: Proceeding of the 17th international conference on World Wide Web. – NY: ACM. P. 645–654.

Griffiths T.L., Steyvers M. (2004) Finding scientific topics // Proceedings of the National Academy of Sciences, 101. P. 5228–5235.

Gruzd A. (2009) Automated Discovery of Emerging Online Communities Among Blog Readers: A Case Study of a Canadian Real Estate Blog // Internet Research 10.0 - Internet: Critical. Milwaukee, WI, USA. 7-10 October 2009.

Gruzd. A., Wellman B., Takhteyev Yu. (2011) Imagining Twitter as an Imagined Community // American Behavioral Scientist. 55(10). P. 12941318.

Hanneman, R. A., Riddle M. (2005) Introduction to social network methods. Riverside, CA: University of California, Riverside.

Hansen D., Shneiderman B., Smith M.A. (2010) Analyzing Social Media Networks with NodeXL: Insights from a Connected World. Morgan Kaufmann.

Howard PN, Duffy A, Freelon D et al. (2011) Opening Closed Regimes: What Was the Role of Social Media During the Arab Spring? *The project on Information Technology and Political Islam (PIPTI)*, Working paper 2011-1. Available at: - http://pitpi.org/index.php/2011/09/11/opening-closed-regimes-what-was-the-role-of-social-media-during-the-arab-spring/ (accessed 9 January 2014).

Jamali S., Rangwala H. (2009) Digging Digg: Comment Mining, Popularity Prediction, and Social Network Analysis // International Conference on Web Information Systems and Mining. – Shanghai, China. P. 32-38.

Kelly J., Etling B.(2008) Mapping Iran's Online Public: Politics and Culture in the Persian Blogosphere // Berkman Center for Internet and Society Research Publication No. 2008-01. http://cyber.law.harvard.edu/sites/cyber.law.harvard.edu/files/Kelly&Etling_Mapping_Irans_Online_Public_2008.pdf (accessed December 1, 2013)

Kumar R., Novak J., Tomkins A. (2010) Structure and Evolution of Online Social Networks // Link Mining: Models, Algorithms, and Applications. P. 337-357.

Leskovec J., Lang K.J., Dasgupta, A., Mahoney M.W. (2008) Statistical properties of community structure in large social and information networks // WWW '08 Proceedings of the 17th international conference on World Wide Web. – Beijing, China: ACM. P. 695-704

Lotan G, Graeff E, Ananny M et al. (2011) The Revolutions Were Tweeted: Information Flows During the 2011 Tunisian and Egyptian Revolutions. *International Journal of Communication* 5:1375–1405.

Manning C.D., Raghavan P., Schütze H. (2008) Introduction to Information Retrieval, Cambridge University Press.

Mishne G., Glance N., (2006) Leave a Reply: An Analysis of Weblog Comments // Third Annual Workshop on the Web-logging Ecosystem.

Newman, M. E. J., Girvan M. (2004) Finding and evaluating community structure in networks // Phys. Rev. E **69**, 026113.

Porter C.E. (2004) A Typology of Virtual Communities: A Multi-Disciplinary Foundation for Future Research // Journal of Computer-Mediated Communication. Volume 10, Issue 1, P. 00.

Qamra A., Tseng B., Chang E.Y. (2006) Mining blog stories using community-based and temporal clustering // CIKM '06 Proceedings of the 15th ACM international conference on Information and knowledge management, New York. P. 58-67.

Rheingold H. (2000) The Virtual Community: Homesteading on the Electronic Frontier. Revised Edition. Cambridge, Mass.: MIT University Press.

Segalovich I. (2003) A fast morphological algorithm with unknown word guessing induced by a dictionary for a web search engine // MLMTA–2003.

Smith M.A., Kollock P. (1999) Communities in Cyberspace. Routledge.

Wellman B., Boase J., Chen W. (2002) The Networked Nature of Community: Online and Offline // IT& Society. Vol. 1, Issue 1, P. 151-165.

Yano, T. and Smith, N. A. (2010) What's Worthy of Comment? Content and Comment Volume in Political Blogs // Proceeding of the 4th International AAAI Conference on Weblogs and Social Media, P. 359-362.

Zakharov P. (2007) Diffusion approach for community discovering within the complex networks: LiveJournal study // Physica A: Statistical Mechanics and its Applications, Vol. 378, Issue 2. P. 550-560.

Olessia Y. Koltsova
National Research University Higher School of Economics
Director of the Laboratory of Internet Research
E-mail: olessia.koltsova@gmail.com
Tel.: +7 (812) 677-94-52

Any opinions or claims contained in this Working Paper do not necessarily reflect the views of HSE.