Article Title Page

[Article title] Communities of co-commenting in the Russian LiveJournal and their topical coherence

March 2015

Author Details *(please list these in the order they should appear in the published article)*

Author 1 Name: Olessia Koltsova
Department: Laboratory for Internet Studies
University/Institution: National Research University Higher School of Economics
Town/City: St.Petersburg
State (US only):
Country: Russia

Author 2 Name: Sergei Koltcov
Department: Laboratory for Internet Studies
University/Institution: National Research University Higher School of Economics
Town/City: St.Petersburg
State (US only):
Country: Russia

Author 3 Name: Sergey Nikolenko
Department: Laboratory for Internet Studies
University/Institution: National Research University Higher School of Economics
Town/City: St.Petersburg
State (US only):
Country: Russia

NOTE: affiliations should appear as the following: Department (if applicable); Institution; City; State (US only); Country.
No further information or detail should be included

Corresponding author: [Name] **Sergei Koltcov**
Corresponding Author's Email: kol-sergei@yandex.ru

☐ *Please check this box if you do not wish your email address to be published*

Acknowledgments (if applicable):
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.

Biographical Details (if applicable):

[Author 1 bio] Olessia Koltsova is the Director of Laboratory for Internet Studies (LINIS) at the National Research University Higher School of Economics (HSE), and the associate professor at the Faculty of Sociology. Prior to LINIS, she was the Dean of the Faculty of Sociology at HSE. She holds the PhD in sociology and publishes extensively in the fields of Internet research and media studies. She is also the author of *News Media and Power in Russia* (Routledge, 2006).

[Author 2 bio] Sergei Koltcov is the IT-director of the Laboratory for Internet Studies (LINIS) at the National Research University Higher School of Economics (HSE), responsible for data storage and processing and for mathematical modelling. He is also an Associate Professor at the department of mathematics at HSE. Holding a PhD in Physics, he has an extensive experience in mathematical modelling and software development for various fields in applied Physics. His scientific software has been acquired by academic and applied research organizations from more than 20 countries. Prior to joining LINIS, he also worked in a number of commercial IT companies and ran his own IT start-up; and before that he worked at the Institute for Analytical Instrumentation of the Russian Academy of Sciences and with a number of European research institutions.

[Author 3 bio] Sergey Nikolenko is a Senior Researcher at the Laboratory for Internet Studies (LINIS) at the National Research University Higher School of Economics (HSE); he specializes on machine learning and Bayesian inference. He is also a researcher at the Steklov Institute of Mathematics at St. Petersburg. His research interests include machine learning, data mining (including practical data mining, e.g., recommender systems), networking algorithms and protocols, and theoretical cryptography.

Structured Abstract:

Purpose (mandatory)

The paper addresses the problem of what drives the formation of latent discussion communities, if any, in the blogosphere: topical composition of posts or their authorship? The aim is thus to contribute to the knowledge about structure of co-commenting.

Design/methodology/approach (mandatory)

The research is based on a 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, where posts are vertices. The Louvain algorithm is used to detect communities of co-commenting. Cosine similarity and topic modeling based on latent Dirichlet allocation are applied to study topical coherence within these communities.

Findings (mandatory)

Bloggers unite into moderately manifest communities by commenting upon roughly the same sets of posts. The graph of co-commenting is sparse and connected by a minority of active non-top commenters. Communities are centered mainly around blog authors as opinion leaders and, to a lesser extent, around a shared topic or topics.

Research limitations/implications (if applicable)

The research has to be replicated on other datasets with more thorough hand-coding to ensure the reliability of results and to reveal average proportions of topic-centered communities.

Practical implications (if applicable)

Knowledge about factors around which co-commenting communities emerge, in particular clustered opinion leaders that often attract such communities, can be used by policy makers in marketing and/or political campaigning when individual leadership is not enough or not applicable.

Originality/value (mandatory)

The research contributes to the social studies of online communities. It is the first study of communities based on co-commenting that combines examination of the content of commented posts and their topics.

Keywords: Communities of co-commenting, online discussion, blogs, Russian blogosphere, SNA, topic modeling.

Article Classification: Research paper

For internal production use only

Running Heads:

Structured Abstract:

Purpose (mandatory)

The paper addresses the problem of what drives the formation of latent discussion communities, if any, in the blogosphere: topical composition of posts or their authorship? The aim is thus to contribute to the knowledge about structure of co-commenting.

Design/methodology/approach (mandatory)

The research is based on a 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, where posts are vertices. The Louvain algorithm is used to detect communities of co-commenting. Cosine similarity and topic modeling based on latent Dirichlet allocation are applied to study topical coherence within these communities.

Findings (mandatory)

Bloggers unite into moderately manifest communities by commenting upon roughly the same sets of posts. The graph of co-commenting is sparse and connected by a minority of active non-top commenters. Communities are centered mainly around blog authors as opinion leaders and, to a lesser extent, around a shared topic or topics.

Research limitations/implications (if applicable)

The research has to be replicated on other datasets with more thorough hand-coding to ensure the reliability of results and to reveal average proportions of topic-centered communities.

Practical implications (if applicable)

Knowledge about factors around which co-commenting communities emerge, in particular clustered opinion leaders that often attract such communities, can be used by policy makers in marketing and/or political campaigning when individual leadership is not enough or not applicable.

Originality/value (mandatory)

The research contributes to the social studies of online communities. It is the first study of communities based on co-commenting that combines examination of the content of commented posts and their topics.

Keywords: Communities of co-commenting, blogs, Russian blogosphere, SNA, topic modeling.

Article Classification: Research paper

Introduction

Blogs and later social networks have proven to be vitally important in the social and political life of contemporary societies. Abundant existing evidence proves their effects on policy (see overview in Drezner and Farrell 2008), electoral preferences (Koltsova and Shcherbak 2014) or electoral results (diGrazia et al. 2013), and broader political views and behavior (Parmelee and Bichard 2012). Sometimes they even have been crucial for political regime changes (Howard et al 2011; Lotan et al 2011). In certain situations, individual bloggers or blogs play key roles in the political process (Enikolopov et al 2012), but inherently the blogosphere is a collective phenomenon, a locus of shared content co-production and dissemination embedded in a complex relational structure of mutual reposts, comments, followings, friendships, and other types of links. To understand social and political processes evolving in the blogosphere, and to forecast their potential effects on the "offline world", it is not enough to know the agendas and generally blog content; a profound knowledge of the entire structure is needed. In other words, it is important to know not only what is being talked about, but also who talks to whom, why and whether some talking groups are larger and potentially more influential or more attractive for influence than others.

In blogs, one of the most important components of their relational structure is commenting configuration, because it is the comments where discussions on blog content evolve. Intensive commenting is an indicator of public interest to a post, its topic, and/or its author. If a group of posts united by a topic, an author, or by some other factors is commented upon by a certain set of bloggers, it may indicate that this certain topic, author, or some third factor attracts interest with a certain audience that can be singled out, described, and targeted. While blog content has by now received a relatively large attention, the structure of blog commenting is still under-researched (see review further below). We know very little on how and why people comment, whether they do it randomly, or there is some preferential attachment, what the grounds of their preferences are, and whether users form clusters of dense commenting of certain issues or of certain blogs.

If such clusters 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 commenting clusters 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. Forecasting based on social media is rapidly developing (Schoen et al 2013), and knowledge of the commenting structure may help construct predictive models that will reveal hot topics or emergent opinion leaders at early stages of their development.

In this work, we seek to contribute to the knowledge of the commenting structure by answering the following question: what drives the emergence of commenting clusters 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 networking sites. The latter, focusing on establishment and maintaining of connections, may at the same time incorporate the former, as chains of date-stamped entries in reverse chronological order (Kaplan and Haenlein 2009: 63). Judging by the growing number of users, one might assume that blogs are indeed losing (see, e.g., Alexa statistics on the leading websites' audience and traffic[1]), 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[2]). 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.

As mentioned above, in blogs, unlike forums, discussions have no place to develop other than in threads of comments to individual blog posts 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 clusters – 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 clusters (called here 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 clusters, that is, "denser" fragments of networks connecting posts that have common commenters. The goal of our present work is to test whether such clusters are indeed present, and if so, whether they form around authors or around topics.

[1] http://www.alexa.com/topsites

[2] http://blogs.yandex.ru/services/

Since relational structures are best represented with graphs, most relevant studies widely use graph theory, a corresponding branche of mathematics, and/or social network analysis (SNA), a corresponding branch of social science that employs graph theory as the primary methodological approach. A graph in mathematics is a structure consisting of a set of objects called vertices and a relation between them that shows which pairs of vertices are connected with edges. In SNA, graphs, vertices and edges are usually referred to as networks, nodes, and arcs respectively. For example, an SNA graph may show classmates as nodes with friendships between them as arcs, or bloggers as nodes and mutual commenting as arcs between them. Graph theory is, among other things, used to find dense clusters in large networks: a multitude of algorithms have been developed to locate such clusters.

Dense clusters of nodes in graphs are often called communities. The term has been borrowed from social sciences and, indeed, when vertices represent humans, dense clusters may indicate the existence of actual human communities in the sociological sense of the word: either self-nominated groupings based on self-identification of participants or human groupings whose participants are linked more intensely among each other than with outsiders, whether the participants know it or not. However, in graph theory community is a purely formal notion used to denote graph clusters of various nature, and community detection is the corresponding branch of graph theory.

In community detection, there are two main types of definitions of a community: local and global (Fortunato 2010). For large networks, *global* definitions are usually applied; they define communities in relation to the entire graph, and not only its immediate environment. More precisely, communities in this approach are defined as subgraphs whose density is significantly higher than would be expected in a random graph of the same size. By this logic, a comment-based community would 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 vertices represent posts and edges denote instances of commenting: two posts share an edge if they have received a comment from the same blogger.

Related work

Most social network analysis on the web, including community detection, has been devoted to links of a different nature than commenting. Earlier studies of web-based social networks usually tried to characterize general patterns in the web graph based on in-text hyperlinks between webpages (Albert et al 1999); some of them addressed the issues of densely connected components (Broder et al 2000) or communities, and have had a modest sociological component. Studies coming from social or political sciences have often been using

visualization algorithms to detect hyperlink communities "by eye". Sometimes this approach has been successful; for instance, Adamic & Glance (2005) demonstrated with this approach the polarized character of the US political blog space, while the Berkman center at Harvard launched a series of studies mapping 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 (Ackland & O'Neil 2011).

Another frequent type of links used in the studies has 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*). *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 merging 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). Kaiser and Bodendorf (2012) used a combination of indicators, such as citation and name mentioning, to manually detect dialogical links between users in discussion threads of online forums, which allowed them to construct and analyze communication networks of those users. In some forums, replies can be directed to a specific previous comment, and networks extracted from such forums in fact come close to comment-based networks (Welser et al 2007).

Surprisingly, 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 on the *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. It was not studied whether communities center around particular persons; the issues of online leadership / influence (Huffaker 2010; Watts & Dodds 2007) or followship / fandom (Cohen 2014) were usually studied in different contexts, not related to networks or communities within them.

Also, 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. This excludes from the analysis the topics of posts written by those bloggers. Just a few studies have approached addressing issues of comments' topicality. Jamali and Rangwala (2009) who studied comment-based networks at Digg have discovered a dependence between commenters and topics of commented posts, not through network analysis, but rather by juxtaposing "hand-coded" topics of posts with IDs of commenters. Each commenter, as they revealed, comments across a wide range of topics. Therefore, the authors extracted 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. In 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. Similarly, Ríos and Muñoz (2012) have detected communities in a network of comments where an edge is created between a pair of users only if their messages are semantically similar. Similarity is measured via comparison of topical compositions of each pair of messages, with topics being obtained through topic modeling and their weights being compared through a modification of the cosine similarity measure. After that, a modified label propagation algorithm of community detection reveals overlapping topic-based commenting communities and allows easy hand labeling of topics and communities. However, both this and Qamra's approaches filter out non-topic-based communities *a priori* and thus also do not address the question of how much hyperlink / commenting communities tend to form around topics or other factors.

Probably the closest to our research is the source in literature that has studied whether topological communities tend to contain one or multiple topics, albeit this has been done in relation to scientific papers, not blogs (Ding 2011). Ding first breaks his dataset into co-authorship communities and then detects five topics in each of them via standard topic modeling. Independently of that, he also first applies the author-topic model to the whole dataset and then mines communities inside subsets of authors related to each

detected topic. He finds multiple topics in each community and multiple communities in each topic, which is determined by the nature of both community detection and topic modeling algorithms. Ding learns if the topics are really different by comparing them with Pearson correlation coefficient (which is not quite correct since word probabilities in topics are not independent variables but comprise a multinomial distribution), but he does not test if communities are really manifest at all (e.g. with modularity). This research thus offers some limited evidence to the hypothesis that scholars cluster into groups to co-author papers on multiple topics, but does not provide a methodology that could be directly borrowed for our comment-related task.

Data and methods

The dataset was 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. At the time of data collection, Russian LJ maintained a publicly available list of Russian language accounts rated by three different methodologies. We used LJ's 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 in general 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 were about 2 million accounts in the Russian language LJ in total by the time of the research, 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, to some extent, arbitrary.

Another conventional threshold is the 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 settled on a one week period. One week is a good candidate since most posts get the majority of their comments over the first 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 related to specific 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 19039 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 of community quality ranging from 0 to 1 or from -1 to 1, is the most widely used quality function in community detection; it is optimized in a number of popular algorithms. This measure was originally introduced in (Newman & Girvan 2004) where it was defined as the fraction of within-community edges in the network less the expected value of the same quantity in a network with the same community divisions but random connections between the vertices (Newman & Girvan 2004); modularity then had many subsequent extensions. The Louvain algorithm uses an extension of modularity for weighted graphs ranging from -1 to 1. 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 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). This algorithm has allowed us to reveal groups of posts that have been commented by approximately the same set of bloggers, i.e., that have generated their own actively commenting audience.

To detect topical similarity of texts within and outside communities, we had to rely on automatic methodologies due to the size of the dataset that could not be processed manually. 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)[3]. Then, we used two alternative methodologies: cosine similarity calculation as the main approach and topic modeling with the LDA model as a supplementary approach.

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. The purpose was to detect whether texts commented by the same sets of users are semantically more similar than on average.

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, we summed the weights of all topics in texts belonging to the same comment-based community, getting the relative importance of each topic for each community. Unlike Ding (2011) who looked for topics within communities and for communities within topics, we clustered the whole dataset into 100 topics to see if they overlap with topological communities that, too, were retrieved from the entire dataset: this seems to be a more appropriate approach because inter-cluster divisions are not forced. 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. Moreover, 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

[3] *Yandex MyStem* is freely available at http://company.yandex.ru/technologies/mystem

coders who first worked independently and then agreed on their labels. We thus revealed which groups of posts that possessed their own audiences were also united by common topics, and what those topics were.

More information on methodology can be found in the supplementary material.

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 matches 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. This latter effect might be explained with preferential attachment of new comments to already well-commented posts, as described in an early work of Barabasi and Albert (1999). 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). We thus found that clusters of dense co-commenting are uneven in size and manifestly centered around posts of certain authors.

[TABLE 1 ABOUT HERE]

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 (0.04917), and global average similarity (0.00016). Thus, 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. Anyway, this suggests that posts commented by roughly the same sets of bloggers are united not only by shared authorship, but also to some degree by similar content.

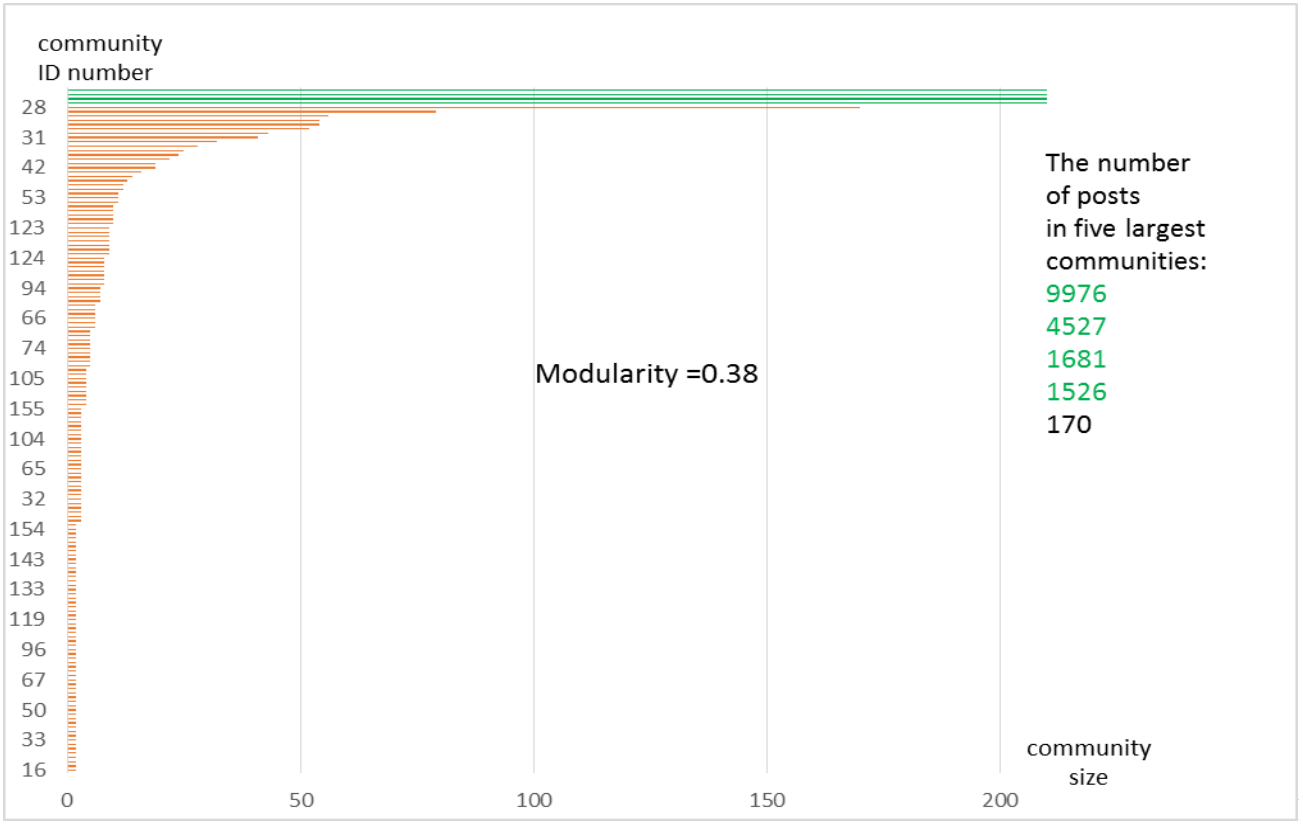

Fig.1. Number of posts in communities: communities 0-158; number range: 2-9976. Louvain algorithm, level 1. Green: bars stretching beyond the picture.
Avg. degree: 237, comments per post: 27, weighted density: 0.012

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 the global cosine similarity average, and the X axis shows 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, even when the data on these similarities are plotted on the logarithmic scale, as suggested in (Raban and Rabin 2009). The skewedness of this distribution lacks an obvious explanation, and this has led us to further explore various properties of the communities.

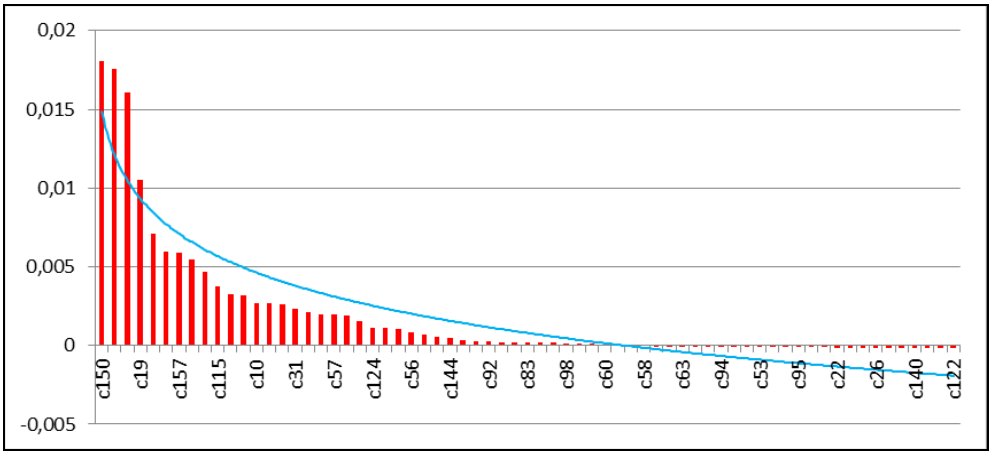

Fig. 2. Distribution of intra-community cosine similarities in comparison with global average (fragment).

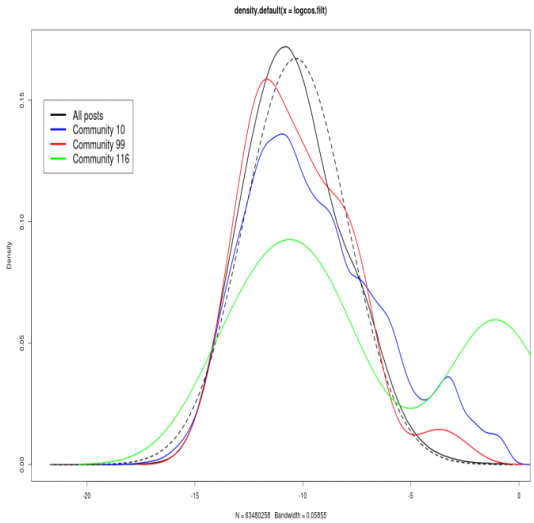

Fig.3. Distributions of logarithms of cosine similarity globally and in some communities.

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 above-average cosine similarity 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 small set of authors, while a relatively large number of disconnected posts by a large number of authors "attaches" to this relatively coherent core. Presumably, it is this core that produces additional peaks in Fig. 3.

To better detect such cores, topic modeling (with 100 topics) with the LDA model trained by the Gibbs sampling algorithm (implemented in the 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 (see e.g. Koltsova and Koltcov 2014). 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 related to a gradual increase in the quality of text preprocessing. The list of topics is given in the supplementary material.

[TABLE 2 ABOUT HERE]

Next, the 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 half of the vertices naturally had the smallest variance, while among other communities different types could be observed (see examples in Fig. 4 a-d). This suggests that while on average posts commented by the same sets of authors are more similar to each other than all posts in the dataset, not all clusters of dense commenting are centered around topically similar posts, but only some of them.

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: communities dominated by a single topic, communities 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, with topics sorted by their weights. Labels of topics tend to match (albeit not perfectly) the 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.

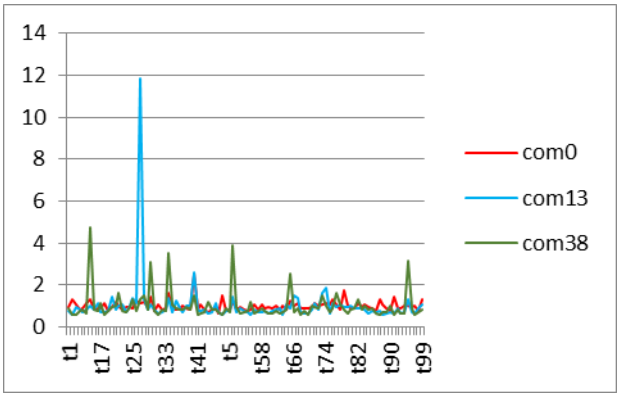

Fig. 4A: distributions of topic weights in three selected communities

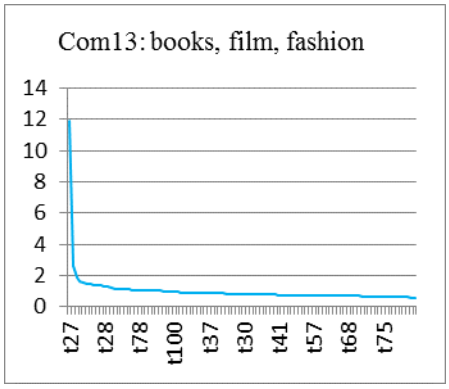

Fig.4b: distribution of topic weights in community 13 (24 posts).

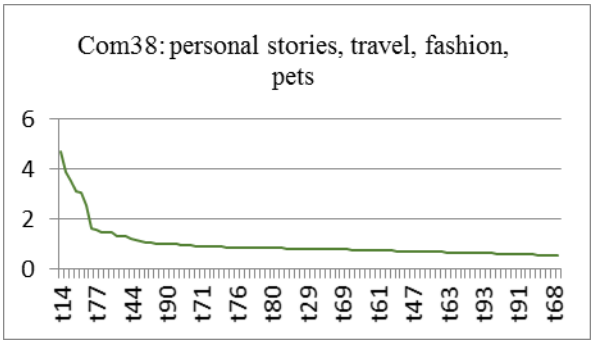

Fig. 4c: distribution of topic weights in community 38 (56 posts)

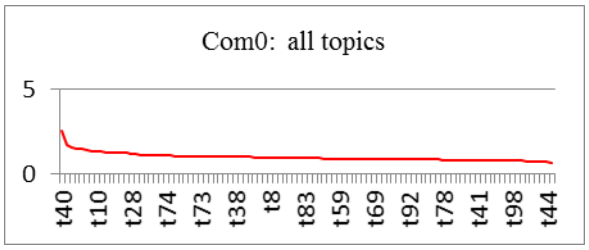

Fig. 4d: distribution of topics' weights in the giant component (Community 0, 9976 posts)

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 topic, but by several 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 calls for further study.

Discussion and practical implications

Our research contributes to the knowledge about the structure of commenting in blogs. It develops earlier research that investigated the properties of friendship communities in LiveJournal (Zakharov 2007) by studying co-commenting communities and their features at this blogging platform. It is the first study that reveals to what degree posts written by popular bloggers may cluster into groups that generate their own actively commenting audiences, and what unites these posts into groups – their content or their authorship.

Thus, our research suggests that people commenting top LJ bloggers tend to unite into moderately manifest communities by (unintentionally) commenting on roughly the same sets of posts. The network of co-commenting is not dense and is connected by a minority of active commenters who tend to be non-top bloggers themselves, thus indicating the predominance of modified fandom commenting in the top LJ. Also, communities strongly tend to emerge around authors of posts, who thus may be treated as opinion leaders of a new type. Traditional fandom may be defined as a form of asymmetrical mediated or parasocial relation, in which a person initiates instances of communication that are not reciprocated and even not known of (Cohen 2014). A classical opinion leader, on the contrary, is the one to whom a person has face-to-face access in his/her small group and by whom he/she may be directly influenced (Lazarsfeld 1950). Blogging transforms both these relations. A leader initiates communication to an indefinite audience, but the feedback is overt and

publicly available, and sometimes reciprocated. Thus an "ordinary" person becomes a public co-contributor to the leader's blog and cannot be completely ignored or not known of. This visible proximity to a leader might be a strong motive for active commenting of top bloggers.

This might also be one of the reasons why communities form around topics of posts to a less visible degree. 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. But some structural reasons for that may be equally plausible.

At a first glance, it could seem natural that bloggers would pick up for commenting those posts that address issues of their interest, and thus tend to comment on multiple posts of the same topic. This would increase the probability of co-commenting the same posts by the same bloggers, even if they do not intend so or even do not know each other. But communication in blogs, at least in LiveJournal, seems to be structured a little differently. LJ is a hybrid of a blogging platform and a social networking site. Besides writing his/her diary, a user can befriend other users and get their messages in an easy to read form akin to a blogroll (friends page). Reciprocation is not necessary, thus celebrities are befriended by much more users than they befriend themselves. This structure causes a tendency of commenting on those bloggers who are befriended and whose posts appear on a user's friends page. This, together with moderate topic-centeredness of co-commenting communities, suggests that perhaps users are inclined not to wander across the LJ space looking for posts on particular topics, but rather to pick up interesting topics from those bloggers who are already in their friends list. That is why an emerging topic of public concern may not necessarily accumulate interrelated comments. It may be the case that such a topic would be commented actively but in disconnected "areas" of LiveJournal and be linked, if at all, through active reposting rather than through co-commenting.

For practical purposes, thus, based on our results, co-commenting communities are useful for studying actively commenting audiences of opinion leaders, as well as intersections and inter-connectedness of those audiences. While detection of individual opinion leaders has been successfully done through various ranking systems, co-commenting communities being often centered around groups of similar authors suggest that leadership may be collective and clustered, and for some goals individual leaders may be not enough. Policy makers and marketing practitioners might search for clusters of bloggers able to generate communities of active co-commenting in order to promote their ideas or goods. If such bloggers are pursued to raise certain issues, they are likely to provoke lavish feedback from 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.

As an example from the current dataset, consider community 52; it reveals three dominant female authors. Upon examination of their blogs and linked websites, it turns out that all three are mutual friends;

two of them explicitly state interest in the issues of women problems, maternity and especially infant feeding, and they are also colleagues and professionals in this sphere, while the third one devotes much of her attention to her three sons. All three formulate their texts in a way that goes beyond the genre of a private diary and corresponds more to the style of a female issues activist's blog. These three bloggers thus form a cluster that attracts an intersecting commenting audience interested in the relevant group of topics. If all the three choose to raise the same issue at the same time, and this issue is relevant to their audience, extremely rich feedback may be produced.

To further develop the practical applicability of these conclusions, it is, of course, useful to learn to what extent they generalize to other blogging platforms, social networking sites, and to other societies. Public blog rankings are actually found in relatively few countries, and not all blogging platforms support the function of friendship. All this may affect the structure of communication. Furthermore, in Russia LiveJournal has historically played a special role both as the first platform for intellectual discussion and later an alternative to highly regulated mainstream media; it was shown that its political content correlated with electoral preferences, but not the electoral results in the 2011-2012 national electoral cycle (Koltsova and Shcherbak 2014). Thus, LJ in Russia may be a unique phenomenon producing unique commenting practices.

Another limitation of this research is absence of data on the content of comments and their dendroid structure. It is known that discussion in comment threads, especially in tree-like ones, may go very far from the initial topic(s) of the post. Mining data on comments' topics, their place in threads, their relation to the number and length of threads and other parameters might shed further light on how co-commenting evolves and why co-commenting clusters emerge.

ACKNOWLEDGEMENTS

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.

REFERENCES

Ackland A., O'Neil M. (2011), "Online collective identity: The case of the environmental movement", *Social Networks*, 33(3), pp. 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*, pp. 36-43.

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

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

Barabási, A.-L., Albert R. (1999), "Emergence of scaling in random networks", *Science*, 286, pp. 509-512.

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, pp.309-320.

Cohen J. (2014), Mediated relationships and social life: current research on fandom, parasocial relations and identification. In: Oliver M.B., Raney A.A. (Eds) *Media and Social Life*, Taylor and Francis.

DiGrazia J., McKelvey K., Bollen J., Rojas F. (2013), "More Tweets, More Votes: Social Media as a Quantitative Indicator of Political Behavior", *PLoS ONE*, 8(11), e79449.

Ding, Y. (2011), "Community Detection: Topological vs. Topical", *Journal of Informetrics*, 5(4), pp. 498-514.

Drezner, D.W., Farrell, H. (2008), "Blogs, politics and power: a special issue of Public Choice", *Public Choice* 134, pp. 1–13.

Enikolopov, R., Petrova M., Sonin K. (2012), 'Do Political Blogs Matter? Corruption in State Controlled Companies, Blog Postings, an DDoS Attacks'. London, Center for Economic Policy Research. http://www.cepr.org/active/publications/discussion_papers/dp.php?dpno=9169 .

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, available at: 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*, available at: 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. (2008), "The power and politics of blogs", *Public Choice*, 134, pp. 15–30.

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

Fortunato, S., Barthelemy, M. (2007), "Resolution limit in community detection", *Proc. Natl. Acad. Sci. USA* 104, pp. 36–41.

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, pp. 645–654.

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

Gruzd A. (2009), "Automated Discovery of Emerging Online Communities Among Blog Readers: A Case Study of a Canadian Real Estate Blog", paper presented at *Internet Research 10.0 - Internet: Critical*, Milwaukee, WI, USA, 7-10 October 2009, available at: http://dalspace.library.dal.ca/bitstream/handle/10222/12831/gruzd_aoir_network_discovery.pdf?sequence=1 (accessed March 17, 2014).

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).

Huffaker, D. (2010), Dimensions of Leadership and Social Influence in Online Communities. *Human Communication Research*, Vol. 36, Issue 4, pp. 593–617.

Jamali S., Rangwala H. (2009), "Digging Digg: Comment Mining, Popularity Prediction, and Social Network Analysis", *Proceedings of the International Conference on Web Information Systems and Mining*, Shanghai, China, pp. 32-38.

Kaiser C., Bodendorf F. (2012) "Mining consumer dialog in online forums", *Internet Research*, Volume 22, Issue 3, pp. 275 – 297.

Kaplan, A.M., Haenlein, M. (2010) Users of the world, unite! The challenges and opportunities of Social Media, *Business Horizons*, 53, pp. 59–68.

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

Koltsova O., Koltcov S. Mapping the Public Agenda with Topic Modeling: The Case of the Russian LiveJournal // *Policy & Internet*. 2013. Vol. 5. No. 2. P. 207-227.

Koltsova O., Shcherbak A. N. 'LiveJournal Libra!': The political blogosphere and voting preferences in Russia in 2011–2012 // *New Media and Society*. 2014. doi: 10.1177/1461444814531875.

Kumar R., Novak J., Tomkins A. (2010), Structure and Evolution of Online Social Networks, in Yu P.S., Han J., Faloustos C. (eds) *Link Mining: Models, Algorithms, and Applications*, Springer, pp. 337-357.

Lazarsfeld, P., Berelson B., and Gaudet H. (1950), *The People's Choice*. New York: Duell, Sloan and Pearce.

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, pp. 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, pp. 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", Paper presented at: Third Annual Workshop on the Web-logging Ecosystem, May 22–26, 2006, Edinburgh, UK, available at: http://staff.science.uva.nl/~gilad/pubs/www2006-blogcomments.pdf (accessed on March 17, 2014).

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

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, pp. 58-67.

Parmelee J.H., Bichard S.L (2012), *Politics and the Twitter Revolution: How Tweets Influence the Relationship between Political Leaders and the Public*. Lanham, MD: Lexington Books.

Raban, D.R., Rabin, E. (2009), "Statistical inference from power law distributed web-based social interactions", *Internet Research*, Volume 19, Issue 3, pp.266 – 278.

Ríos, S.A., Muñoz, R. (2012), "Dark Web portal overlapping community detection based on topic models", In *Proceedings of the ACM SIGKDD Workshop on Intelligence and Security Informatics* (ISI-KDD '12), ACM, New York, NY, USA, article 2.

Schoen H., Gayo-Avello D., Metaxas P.T., Mustafaraj E., Strohmaier M., Gloor P., (2013) "The power of prediction with social media", *Internet Research*, Volume 23, Issue 5, pp.528 – 543.

Segalovich I. (2003), "A fast morphological algorithm with unknown word guessing induced by a dictionary for a web search engine", Proceedings of MLMTA–2003, pp. 273-280.

Watts D.J., Dodds P.S. (2007), Influentials, networks and public opinion formation. *Journal of Consumer Research*, Vol. 34, Issue. 4, pp. 441-458.

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

Welser H., Gleave E., Fisher D., Smith M. (2007) "Visualizing the Signatures of Social Roles in Online Discussion Groups", Journal of Social Structure, Vol. 8, Issue 2. Available at: http://www.cmu.edu/joss/content/articles/volume8/Welser/ (accessed April 7, 2014).

Yano, T. and Smith, N. A. (2010), "What's Worthy of Comment? Content and Comment Volume in Political", *Proceedings of the 4th International AAAI Conference on Weblogs and Social Media*, Melno Park, CA, pp. 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, pp. 550-560.

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.

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

SUPPLEMENTARY MATERIAL

A METHODOLOGICAL NOTE

Russian LiveJournal and its rankings.

LJ daily rates all accounts subscribed for "Cyrillic services". Accounts created from within Russia, Belarus and Ukraine are subscribed automatically, and have to be unsubscribed would the owner wish so. Thus, this segment of LJ includes some proportion of texts in Ukrainian and Belorussian, albeit very small, and its influence on topic modeling is minimal. Originally, LJ rated bloggers by the number of users who have befriended them, but as too many spam accounts penetrated to the top of the ranking, LJ modified this ranking and introduced two others, so by the time of the research the rankings were:

1. Social capital. Based on the number of friends, it weights them by a number of their features: whether they have visited the assessed blog recently, whether they maintain their own blog, whether they comment on other blogs, how many friends they have and others. The recent activity is privileged; the older the activity, the more it is penalized and it gets a zero score if it is older than one year. Thus this ranking filters out bots and spam accounts, inactive accounts, as well as it accounts for recent activity of bloggers, still it is resistant to short-term fluctuations.
2. Social capital 24 hours. Similar to the former, but accounts only for the last 24 hours activities.
3. Number of views. Counts how many times the blog was viewed in the last 24 hours.

The last two rankings are very volatile, while we were interested in detecting long-term LJ leaders whose position in the ranking would not be affected by recent events. None of these rankings is fully transparent, but no other rankings are available. As noted in the paper, constructing networks from random accounts would have brought us to extremely sparse and meaningless networks, therefore we had to use this LJ tool.

Network construction and partition

All texts of posts, post IDs, authors' IDs, commentators' IDs and other data were downloaded to a relational database with the Lab's tool BlogMiner, after which a script was developed that extracted post IDs and respective authors' and commentators' IDs. Another piece of code transformed this into an edge list, where an edge was added between two posts if they were commented by the same commentator. The initial list contained 19,039 vertices and 4,533,077 links. Since vertices could be connected multiple times, edges then were merged and ascribed weights. If a commentator commented on a pair of posts twice, the edge was ascribed the weight of two. If a commentator commented on one post of a pair twice, and on the other only once, the weight ascribed equaled one. If the ID of the commentator coincided with the ID of the author of one of the posts in a pair, no edge was created. This is how self-commenting was filtered out, which constituted a large proportion of would-be edges. After communities were detected, each post was ascribed the number of the community it belongs to. The list of communities and their sizes can be seen the table further below.

Text preprocessing

Text preprocessing was done with the Lab's tool TopicMiner. First, texts were cleared from html tags, single digits and letters and other garbage characters. Second, the texts were lemmatized with MyStem lemmatizer built into TopicMiner. This lemmatizer was developed specially for Russian language by the leading Russian search engine company, Yandex. It is both vocabulary-based and rule-based, and accounts well for the complex Russian suffix system. Usually, for the words found in the vocabulary, one lemma is generated; for the words that are not found, and for polysemic cases, more than one hypothesis is developed, in which case

TopicMiner picks up the first lemma. Since developing more sophisticated rules for lemma choice is a long-term task for the whole Russian linguistic community, this was the only way to perform lemmatization. Next, the collection was cleaned of stop-words, for which absolute word frequencies were calculated. Totally, the text collection contained 187956 unique words and 7489860 word entries. Words occurring more than 5000 times and less than 5 times were deleted, after which 48260 unique words remained. Finally, we filtered out posts that contained no texts (either as a result of preprocessing or initially), which left 17,386 texts for further analysis.

Cosine similarity calculation

Cosine similarity measure is a cosine of the angle between two vectors whose components correspond to the frequencies of words occurring in the texts, but most often absolute frequencies are substituted with weighted ones, usually TF-IDF measure. This was calculated with a specially developed piece of code. After that, another code was used to calculate pairwise cosine similarities, which produced more than 300 million values. Since we did not have sufficient computing and memory resources, values smaller than a certain threshold were not stored. When cosine similarity values were averaged for each community, these unstored values were considered zero.

Topic modeling

There are no clear criteria for the choice of the number of topics in topic modeling, since quality measures, as well as general ideas of what quality of topic modeling might mean, are underdeveloped. One of the approaches is jump theory, a rate distortion theory modification (Sugar and James 2003) was applied in many of our projects, and it generated an optimal number of topics for datasets of similar size in the range between 100 and 130, however, the exact number has been very volatile. Based on this prior experience, we used the number of 100 here. Blei, too, mentions this number for the collection of similar size (Blei 2003). We used α=0.1 and β=0.5 as suggested by (Griffiths and Steyvers 2004). The number of iterations was 2000 which is much larger than it is needed for the algorithm to converge.

Hand coding

Hand coding was used for two purposes: topic labeling and community description. For neither of these purposes pre-defined categories can be developed, thus calculating any index of inter-coder reliability is impossible. Topic labels are needed to help understand what each topic is about, since the algorithm can only break the collection into groups numbered 1 to n, but not tell what those groups mean. On a macro-level, topics may be broken into public affairs, private issues and noise, but this is not very informative. Precise topic composition is dependent on the text collection and, for blogs, on current events in particular. Therefore, two coders were asked to develop short descriptors of one to five words independently of each other, after which they came together to discuss the results and to agree on the single label for each topic. Most often labels coincide, but not precisely enough to measure any coefficients, e.g. “regional elections”, “elections in regions”, “local elections”. Labeling was done by coders after reading 20 most probable words of the topic and 20 most probable texts of the topic.

Describing communities was even more complex since they could be dominated by one, multiple or no topics, as well as one, multiple or no authors of posts. Unlike topic labeling, community labeling was carried out by us for the first time, and no clear understanding of what could be discovered there had been available. Therefore, two coders were asked to account for the described above parameters in the form they chose themselves. After that the descriptions were discussed and agreed on. Coding was based on all texts of the communities; 20 communities with the highest topical variance were coded.

Comment-based communities, sorted by average similarity

Community ID	avg number of words in posts	Number of authors who wrote posts in community	Number of posts in community	Avg similarity
91	4.4	1	5	0.87564400
154	220.0	1	2	0.52173600
145	444.0	1	2	0.28459800
142	878.0	1	3	0.18308300
116	62.0	2	27	0.11325900
51	1277.0	1	3	0.08059800
106	67.8	3	8	0.03576150
86	156.9	5	8	0.02547550
102	165.8	1	4	0.02501340
120	59.5	1	2	0.02128210
150	320.7	2	9	0.01824830
76	264.9	3	8	0.01771450
66	166.2	1	6	0.01623360
19	691.7	1	3	0.01070800
39	129.6	5	20	0.00725582
54	181.2	3	10	0.00615473
157	262.4	2	14	0.00602156
129	351.0	1	3	0.00562149
52	176.2	8	43	0.00484460
115	42.9	1	11	0.00393607
42	1123.9	3	18	0.00342572
7	111.0	14	48	0.00336444
10	83.1	262	1135	0.00287383
99	124.7	32	67	0.00286688
13	162.8	11	52	0.00280106
31	15.0	2	29	0.00246305
11	134.1	34	51	0.00230275
87	58.5	1	2	0.00212892
57	150.0	3	6	0.00210644
32	202.7	2	3	0.00209044
68	364.9	1	9	0.00170151
124	438.5	2	8	0.00125936
59	185.0	1	2	0.00125207
28	321.1	75	161	0.00119606
56	124.9	1	7	0.00100560
101	130.0	4	6	0.00088898
69	167.7	2	3	0.00069568
144	403.3	8	30	0.00063784
1	193.1	573	4208	0.00052220
81	194.8	1	5	0.00044663

92	378.6	5	7	0.00042394
4	369.6	457	1624	0.00035689
8	327.3	2	3	0.00035580
83	244.9	2	7	0.00034336
0	238.7	1353	9226	0.00034037
98	243.3	2	3	0.00025538
44	11.9	1	8	0.00025242
126	520.0	1	2	0.00025226
60	152.0	1	2	0.00022645
89	271.7	1	3	0.00022066
93	349.3	12	26	0.00017299
58	24.5	1	2	0.00017041
24	***137.4***	***2***	***5***	***0.00015698***
88	245.8	2	4	0.00015258
63	1049.0	1	2	0.00012056
155	246.0	2	3	0.00011312
62	83.5	1	2	0.00010951
94	1019.9	5	7	0.00010336
27	146.9	3	7	0.00009196
143	129.5	1	2	0.00007706
53	83.6	8	10	0.00005902
132	219.3	2	12	0.00005146
2	278.3	1	3	0.00005061
95	585.8	2	5	0.00004793
12	337.7	3	10	0.00004461
151	412.0	1	2	0.00003643
22	271.7	5	9	0.00002709
103	376.5	1	2	0.00002161
104	476.3	2	3	0.00001798
26	340.0	1	2	0.00001515
82	57.0	1	3	0.00001231
38	304.0	5	24	0.00000788
140	203.7	2	6	0.00000778
97	114.0	1	2	0.00000662
41	354.0	1	2	0.00000622
122	529.7	1	3	0.00000536
77	106.9	5	43	0.00000349
73	633.1	2	8	0.00000139
158	84.5	3	10	0.00000074
3	213.5	6	11	0.00000000
5	213.0	1	3	0.00000000
6	49.0	2	2	0.00000000
9	7.0	1	2	0.00000000
14	0.0	1	3	0.00000000
15	0.0	1	1	0.00000000
16	2.0	1	1	0.00000000

18	341.3	1	6	0.00000000
20	121.0	1	2	0.00000000
21	80.5	1	2	0.00000000
23	165.2	1	5	0.00000000
25	81.0	1	1	0.00000000
29	10.5	1	2	0.00000000
30	16.5	1	2	0.00000000
33	1491.5	1	2	0.00000000
34	513.9	1	8	0.00000000
35	196.5	1	2	0.00000000
36	507.5	2	2	0.00000000
37	208.7	1	3	0.00000000
40	442.7	1	3	0.00000000
43	65.0	1	1	0.00000000
45	1354.5	1	2	0.00000000
46	494.6	1	5	0.00000000
47	7.0	1	1	0.00000000
48	17.0	1	1	0.00000000
50	33.0	1	2	0.00000000
55	106.0	2	2	0.00000000
61	411.5	2	2	0.00000000
64	114.5	2	6	0.00000000
65	112.0	3	3	0.00000000
67	71.5	2	2	0.00000000
70	76.5	1	2	0.00000000
71	122.7	1	3	0.00000000
72	128.4	1	8	0.00000000
74	95.4	1	5	0.00000000
75	912.3	1	3	0.00000000
78	198.5	1	2	0.00000000
79	71.0	1	2	0.00000000
80	11.0	1	2	0.00000000
84	87.0	1	1	0.00000000
85	245.0	1	1	0.00000000
90	98.0	1	2	0.00000000
96	262.0	2	2	0.00000000
100	225.6	1	5	0.00000000
105	146.0	1	4	0.00000000
107	121.5	2	2	0.00000000
108	38.5	2	2	0.00000000
109	105.3	1	4	0.00000000
110	45.0	2	3	0.00000000
111	863.0	1	9	0.00000000
112	134.0	2	2	0.00000000
113	176.0	1	2	0.00000000
114	403.7	1	6	0.00000000

117	417.8	1	11	0.00000000
118	87.0	1	2	0.00000000
121	244.9	4	8	0.00000000
123	303.0	1	9	0.00000000
125	210.0	2	2	0.00000000
127	19.0	1	1	0.00000000
128	351.5	2	2	0.00000000
130	70.7	1	6	0.00000000
131	47.5	2	2	0.00000000
133	30.0	2	2	0.00000000
134	413.0	2	2	0.00000000
135	116.5	2	2	0.00000000
136	253.5	1	2	0.00000000
137	1278.5	2	2	0.00000000
138	250.5	2	2	0.00000000
139	45.5	1	2	0.00000000
141	36.0	1	1	0.00000000
146	24.5	2	2	0.00000000
147	13.0	1	1	0.00000000
148	134.0	2	2	0.00000000
149	385.7	2	3	0.00000000
152	79.0	3	10	0.00000000
153	2.0	1	1	0.00000000
156	508.0	2	2	0.00000000

Global average similarity corresponds to community 24, marked bold and italic.

Topics revealed and hand-labeled

N	topic	N	topic	N	topic
2	writers	35	general lexicon	68	Ukrainian politics
3	body	36	fairy-tales	69	soccer
4	sex	37	mobile devices	70	Berezovsky's death
5	grain prices	38	Soviet past	71	city transportation
6	medicine as science	39	Xenia Sobchak	72	God
7	Shuvalov in Britain	40	general lexicon	73	LJ jargon
8	international relations: USA, China and Korea	41	uninterpretable	74	shopping
9	Gazprom	42	money	75	church
10	criminal proceedings	43	English words	76	meetings
11	uninterpretable	44	church	77	photo
12	uninterpretable	45	film and movies	78	uninterpretable
13	cars	46	animals and nature	79	problems and solutions
14	travel and tourists	47	word parts	80	trips
15	beach tourism	48	parliamentarians and ministers	81	plagiarism in dissertations

16	toys	49	uninterpretable	82	cooking recipes, food
17	Rock-climbers' death	50	fashion and clothes	83	concerts, musical groups
18	military machinery	51	uninterpretable	84	internet
19	uninterpretable	52	space	85	doctors, healthcare
20	urban setting	53	Caucasus	86	India - history
21	holidays and days off	54	uninterpretable	87	uninterpretable
22	Middle East politics	55	ecology	88	Alexey Navalny
23	uninterpretable	56	industry	89	wars of Russia
24	uninterpretable	57	Korean nuclear program	90	uninterpretable
25	painting	58	Banks	91	Soviet and Russian politics
26	family	59	weapons	92	uninterpretable
27	books	60	village	93	fire in Chechnya
28	family and kids	61	oil and gas	94	education
29	drinking	62	Land code	95	pets and nature
30	uninterpretable	63	USA	96	events
31	laws and courts	64	hotels	97	regional administration
32	uninterpretable	65	nature	98	pirates
33	air flights	66	film and movies	99	love
				100	English internet jargon

References

1. Blei D.M., Ng A.Y., Jordan M.I. (2003) Latent Dirichlet Allocation. *Journal of Machine Learning Research*, vol. 3, pp. 993–1022.
2. Griffiths T.L., Steyvers M. (2004) Finding scientific topics. *Proceedings of the National Academy of Sciences.* USA 101, pp. 5228–5235.
3. Sugar C., James G. (2003) Finding the number of clusters in a data set: An information theoretic approach. *Journal of the American Statistical Association*, 98. Alexandria: ASA, 2003, pp. 750–763.