NATIONAL RESEARCH UNIVERSITY
HIGHER SCHOOL OF ECONOMICS

*Kirill Maslinsky, Sergey Koltsov, Olessia Koltsova*

# CHANGES IN THE TOPICAL STRUCTURE OF RUSSIAN-LANGUAGE LIVEJOURNAL: THE IMPACT OF ELECTIONS 2011

BASIC RESEARCH PROGRAM

WORKING PAPERS

SERIES: SOCIOLOGY
WP BRP 14/SOC/2013

This Working Paper is an output of a research project implemented as part of the Basic Research Program at the National Research University Higher School of Economics (HSE). Any opinions or claims contained in this Working Paper do not necessarily reflect the views of HSE.

*Kirill Maslinsky*[1], *Sergey Koltsov*[2], *Olessia Koltsova*[3]

# CHANGES IN THE TOPICAL STRUCTURE OF RUSSIAN-LANGUAGE LIVEJOURNAL: THE IMPACT OF ELECTIONS 2011

This study investigates the topical structure of the Russian-language blog-publishing service LiveJournal and the change in it that occurred in the course of the public activity after the State Duma elections in December 2011 as compared to a previous “control” period (November 27 – December 27 and August 15 – September 15 respectively). The data for both periods have been automatically obtained from 2000 top-rated blogs on the basis of ratings published by LiveJournal. Unsupervised topic modelling of the sampled posts was done using Latent Dirichlet Allocation algorithm. In December 2011 we found considerable growth in weights of all the topics closely associated with the discussion of voting results and protests, accompanied by a more moderate decrease in the majority of other social topics. the number of users who started posting texts that may be conventionally qualified as political according to LDA in December 2011, considerably outnumbers the number of those who ceased posting political items, which may indicate the existence of a blogger mobilization process in political topics.

JEL Classification: Z.
Keywords: Internet media, blogs, political mobilization, Russia, topic modeling, LDA.

[1] Researcher at the Laboratory for Internet Studies, National Research University Higher School of Economics, Saint-Petersburg, Russia.

[2] Senior researcher at the Laboratory for Internet Studies, National Research University Higher School of Economics, Saint-Petersburg, Russia

[3] Head of the Laboratory for Internet Studies, National Research University Higher School of Economics, Saint-Petersburg, Russia

# 1 Introduction

Relatively recent events, such as the December 2011 political protests in Russia, the May 15th Movement in Spain and, most notably, the Arabic spring have demonstrated that new, Internet and mobile phone based media may play a crucial role in the life of all types of societies. Therefore, social scientists have become increasingly interested in the role of Internet media, social networking services and blogs in various social and political processes, such as the formation of alternative political agendas, expressions of public opinion, and direct political mobilization. In Russia, the clearest episode of mobilization that may be considered in the context of Internet media and their role, was the social protests caused by the State Duma elections in December 2011. This protest, apart from attracting huge public attention itself, has also contributed to the dissemination of the belief in the large role of the Internet for political agenda setting and mobilization. Our research questions inspired by these events may be formulated as follows: does user-generated Internet content really reflect the changes in the social and political domain? Does it form a politically sensitive agenda as much as is claimed or does it limit itself to recreational and consumer activity? Does it contain any signs of direct mobilization? Our major hypothesis was that election-connected topics in December 2011 would replace some other topics or reduce their number, as compared to a "quiet" period, which in our research is exemplified by August-September 2011.

As is known from other studies [e.g. Etling et al 2010], political content is not distributed evenly across blogs and social networks; it is concentrated on certain blog platforms and is shifted towards their most popular users. Thus, the LJ top that is, the most popular and the most followed bloggers at LiveJournal were chosen as representative testing grounds for the study of social discussions on the Russian-language Internet. LiveJournal.com was chosen, because historically it became the leading service where Russian-language blogs first evolved [Gorny 2004], and later on, it became an important public platform for all kinds of civic activism [Lonkila 2008; Alexanyan, Koltsova 2009]. In 2006, when the Russian-language segment of LiveJournal was placed under the control of the Russian company SUP, some authors denounced this as a breakdown of LiveJournal as an important platform for social and political discussions in the Russian language Internet (Runet) [Parkhomenko, Tait 2008]. However, an ambitious quantitative analysis of Russian blogosphere has recently proven that the greatest number of political blogs still reside there [Etling et al. 2010], although it is now fully controlled by SUP.

Within LiveJournal, this research focuses on top users (according to LiveJournal rating) because the aggregate of top blogs may be considered as an independent phenomenon distinguished by specific visibility in the Internet's public space, by their accelerated availability in top search results for blogs and by the number of followers.

The idea of this research was to estimate, on the basis of representative quantitative data, the changes that occurred in the topical spectrum of LiveJournal posts in the course of the public activity that followed the State Duma elections in December 2011 as compared to a previous "quiet" period. The current paper focuses on two principal dimensions of the blogosphere's topical spectrum:

*Degree of involvement*: how much the problem of elections and associated remonstrative activity is discussed in the blogosphere (percentage of all the discussions in the blogosphere, in our case on LJ); which percentage of bloggers touch upon this subject?

*Structural changes*: whether the social and political issues associated with the elections change the general topical structure of the blogosphere, e.g. whether these problems affect the distribution of other discussions, topically unrelated to the elections (in other words, whether the elections replace other topics, in the discussion space or not).

To reach well-grounded conclusions on the topics of the discussions on LiveJournal require the collection and topical classification of a substantial corpora of posts. When using the classical methods of content analysis, the scope of quantitative investigation in this field is limited by high encoding costs. This research uses a fully-automated method of text analysis by means of probabilistic topic modelling, which allows the processing of much larger samples and gets results significantly faster. Unfortunately, the cost of this formalization of text content is a degradation of accuracy when compared to classical content analysis. However, if the interpretation of results takes into account the nature of formalizations used, the data obtained via the method of automated analysis allow us to make, although narrow, reasonable conclusions in relation to broader samples of texts.

The rest of this article is arranged as follows. Section 2 describes the probabilistic topic modelling algorithm used — Latent Dirichlet Allocation. Section 3 describes the selection principles, collection procedures and automated processing of texts, as well as descriptive statistics of the analyzed data. Section 4 describes the results: the specification of the topics found and the statistics of changes in topical structure of LJ in December 2011 compared to August-September the same year. Section 5 summarizes the results of this research and discusses further possible areas of research.

## 2 Methodology

Topic modelling is an approach to text analysis aimed at building a statistical model that would describe a given collection of documents as a mixture of a finite number of “topics” in such a way that in each document several topics are represented in different proportions. The notion of *topic* in such an approach is simplified and defined algorithmically, as a result of the statistical calculation of words co-occurring in the texts.

A topic modelling algorithm named Latent Dirichlet Allocation (LDA) was first proposed in 2002 to extract groups of semantically interconnected words in a text collection [Blei et al. 2003]. Nowadays, LDA with modifications is the most widely used method for extensive text collections in various genres: from research articles or literary databases to blogs or microblogs [Blei 2012, Blei, Lafferty 2009, Ramage et al. 2010].

LDA is a generative probabilistic model that represents each document as a mixture of latent variables called “topics” where each topic is defined by a distribution of a set of words. Being a generative model means that the algorithm assumes the documents to be generated according to a model, in this case by first deciding on the length of the document; then, by deciding on the mixture of topics to be represented in this document, and finally, by picking words from those topics. The topics may be imagined as “bags” of words with different probabilities of being in these bags. However, the algorithm does not perform this process; instead it tries to backtrack from the real documents to find the set of topics that are likely to have generated those documents – that is, it performs an inverse process.

For documents to be generated, the generative probability of each word from a document is represented as a product of the two functions: the distribution of probabilities of documents in the space of topics and the distribution of the probabilities of words in the space of topics. Mathematically the method is expressed as follows [Daud et al 2009]. Notation:

$\Phi_z$ multinominal distribution of words specific to a topic *z* with parameter $\beta$;

$\Theta_d$ multinominal distribution of a document *d* with parameter $\alpha$ in the space of topics;

$w_{di}$ the $i^{\text{th}}$ word token in document *d*;

$z_{di}$ topics assigned to the $i^{\text{th}}$ word token in document *d*;

$\alpha$ Dirichlet distribution associated with topic *z*;

$\beta$ Dirichlet distribution associated with word $w_{di}$.

LDA probability model is described by the expression:

$$p(w \mid d, \theta, \Phi) = \sum_{z=1}^{T} p(w \mid z, \Phi_z) p(z \mid d, \theta_d) ,$$

where:

- *$p(z|d,\Theta_d)$* is a function of distribution of topic probabilities in the space of documents (the probability of a topic in a document *d*). This function depends on parameter *α*; each *α* corresponds to one topic only. A multitude of topics means the necessity for a multitude of *α* parameters;
- *$p(w|z, \Phi_z)$* is a function of distribution of word probabilities in the space of topics. This function depends on parameter *β* (one parameter for each topic).

The right part of the equation contains the prior information about distributions, the left part contains the posterior information.

The calculation according to this methodology proceeds as the evaluation of parameters *α* and *β* for distributions of all words and of all documents in the space of topics. Thus, the output of the algorithm consists of two matrices: one describes the distribution of the probabilities of all words over all topics, and the other, the distribution of the probabilities of all documents over all topics.

To visualize the idea of a document as a topic mixture one may consider a step in the LDA inference process whereby the words of a single document are split between several topics so that each word token is assigned to one of the latent topics. Fig. 1 displays an example of our text collection represented as a mixture of inferred topics.

**Fig. 1. Blog post as a topic mixture**

*"What's the* ***news****[98]? What's* ***happening****[1] in the country? What* ***should****[90] be done? Who is to blame[90]? Who's gay? [90]?*"

| Topic # | Weight in the document | Top words |
|---|---|---|
| Topic 1 | 0.2 | police / policeman / official / detain / to report / militia / be situated / **happen /** several / tell / young / drug / near |
| Topic 90 | 0.6 | understand / **should** / because / go / at all / it is necessary / also / something / think / why about / some / at the moment |
| Topic 98 | 0.2 | the air / program / channel / journalist / Nemtsov / program / lead / TV channel / conversation / **news /** interview / TV / LJembed / radio / television / Moscow |

*Notes*: Number in square brackets after a word represents the number of the topic to which this word occurrence was assigned. Unnumbered words are stopwords removed in the preprocessing stage. In the original Russian text each lexical entry in a topic is represented by a single word. Since LDA does not disambiguate word meanings, the most likely meanings for the given topic were selected in the English translation. Words present in the list of top terms of the corresponding topic are shown in bold.

From a user's point of view, LDA thus makes it possible to automatically split the vocabulary of the collection into semantically connected groups, i.e. topics. Within a topic, words can be arranged by their weight (probability) in the topic. Usually, for a user to detect the general content of a topic, it is enough to look through several dozens of top words, or in more ambiguous cases through several dozens of top texts, ordered according to their probabilities in the topic. In real text collections, topics are unequal: they may include larger or smaller numbers of words with non-zero probabilities, general or special vocabulary, words with higher or lower frequencies. The total of all the words' probabilities in a given topic, makes it possible to estimate the relative "weight" of this topic within the collection, which may be roughly interpreted as the importance of the topic or the volume of the attention to it paid by the authors of the documents.

One of the unresolved problems of topic modelling is the choice of the number of topics (c.f. similar problems in cluster and factor analysis). This is closely connected to the problem of quality assessment of the method in general, since it may be a basis of comparison between solutions with different numbers of topics. External measures where the quality is tested against collections marked up manually by humans, although traditional for various text analysis methods, face a number of obstacles. First, the aim of LDA is to detect latent topics, and not only those that may be observed by humans, so the external measures may be seen as not always optimal. Second, a specifically Russian language problem is that in Russian no manually marked up collection of significant size and quality exists. Furthermore, no benchmark based on blog texts exists in other languages either. Among internal measures, the most common is perplexity. However, as with most other quality measures it changes monotonically with an increase in the number of topics. To find a point in the perplexity function after which the increase of number of topics may be stopped a jump in it must be detected. To solve this problem we have used the jump theory by Sugar and James [2003], having developed a piece of simple software to implement this algorithm. It has worked very well detecting jumps in the perplexity function on other blog text collections we have used [Koltsova, Maslinsky 2013], however, in this particular collection the perplexity function behaved unusually, repeatedly going up and down. As a result, no jump could be detected. The causes of this phenomenon demand special study. Therefore, we have chosen the same number of topics we used elsewhere (100), to make the data comparable.

The analysis of topical evolution in diachronic text collections (e.g., in articles of Science magazine over the period of 100 years) is one of the classical fields where LDA is applied [Blei, Lafferty 2006]. In this work, we are using a simple method of estimation of diachronic changes based on the results of the application of the standard LDA [Hall et al. 2008]. This method includes the evaluation of the distribution of topics in the whole collection. Then the collection is split into "time slices". In this work, each slice represents a set of posts published within a period of time (a month). For each slice we estimated relative weight and vocabulary of each topic. Comparison of time slices makes it possible to follow the increase and decrease in popularity of topics (i.e. changes in the relative weight of different topics), along with the changes in their vocabulary.

It should be noted that this approach does not imply the concept of appearance or disappearance of topics: all the topics are "cross-cutting", i.e. they appear in all time slices. This characteristic of the model means that if there are any texts with a new topic within the collection, the new vocabulary introduced by those texts will be grouped with compatible vocabulary of the texts from other periods. Another approach where topics are modelled independently for each period and compared manually [Koltsova, Maslinsky 2013] gives similar results, although comparison of periods  by a human complicates the discovery of cross-cutting topics and estimation of the discrepancy between the periods.

The analysis of diachronic changes in the blogosphere's topics by means of LDA implies that the formal simplifications should be accepted: changes in the blogosphere's topics will be changes in the total frequency of automatically distinguished lexical groups and their dispersion across the texts.

## 3 Data and software

This research is based on the two samples of posts published on LiveJournal from August 15 to September 15, 2011 ("quiet period") and from November 27 to December 27, 2011 ("active period")[4]. The sample consists of posts from 1 400 top journals in the rating published by LiveJournal, as of the date of data retrieval (September 15 2011 and December 27 2011)[5]. The rating reflecting the number of the author's followers appeared quite stable: 1 150 out of the journals selected in September, were likewise selected in December. The samples included the posts published within the given period, but no more than the 50 latest posts. Topic modelling was based on posts' body texts only, comments being excluded. The preprocessing of texts

[4] Data acquired through *Koltran Blogminer* program developed by the Laboratory of Internet Studies, High School of Economy (St. Petersburg, Russia).

[5] LiveJournal's rating: http://www.livejournal.com/ratings/users/. Before February 2012, rating was only based on the total number of users who included the journal in their friends list.

included their lemmatization by means of *mystem* program [Segalovich 2003], with automated selection of the most frequent or the first lemmas, the removal of 100 most frequent words in the collection and of the words encountered in less than 5 documents, and the removal of documents containing less than 5 words after the word removal was finished. After the preprocessing, the September sample included 24 198 posts, 1 360 users; and the December sample 27 026 posts, 1 393 users. A 100-topic LDA and further analysis of weight and vocabulary of topics for each period were carried out by means of Stanford Topic Modeling Toolbox [Ramage et al. 2009].

This software was chosen because it is the only one found that offers the option of diachronic analysis. In addition, it has been designed specifically for social science goals, it contains an in-built perplexity assessment option, and the choice between the two main LDA algorithms of assessment of $\alpha$ and $\beta$ parameters: variational EM-algorithm introduced by Blei [2003] and Gibbs sampling [Griffiths, Steyvers 2004]. Since according to some authors [Heinrich 2008; Griffiths, Steyvers 2004] the latter algorithm gives better results, it was chosen for this research.

## 4 Results

### 4.1 Description of Topics

The topics automatically found by LDA were manually encoded on the basis of the top 20 words of each topic and of the top 20 texts most probably associated with the topic. In the process of encoding, each topic was assigned a label (a word characterizing the topic), title and type. This work uses conventional classification of topics by 4 classes, depending on the character of top words and top texts:

1. *Domain* — topics with an expressed lexical core referring to a specific subject area, e. g. *health* (medicine and healt h), *USSR* (Soviet history and everyday life).

2. *Register* — topics with prevalence of top words that mostly characterize style (register) of the text, while not being associated with a particular subject area; e.g. *today* (immediate records, future plans), *should* (judgements, modality).

3. *Language* — topics with foreign lexical core (in our collection — English and Ukrainian).

4. *Boilerplate* — topics formed on the basis of clichés, especially textual interface elements, list-based structures etc, and all uninterpretable topics (statistical noise).

Eventually, 65 topics out of 100 were qualified as domain, 21 as register, 2 as language and 12 as boilerplate. It should be noted that margins between these categories are very diffuse, therefore the assignment of a topic to any single class is rather conventional. Nevertheless, various types of topics differ in terms of the total number of words in a topic and variability of vocabulary in its top when comparing September and December time slices. On average, register topics include more words and their lexical core is more stable diachronically compared to domain topics, which can be explained by the fact that vocabulary of the former is mostly general (see Fig. 2a, 2b).

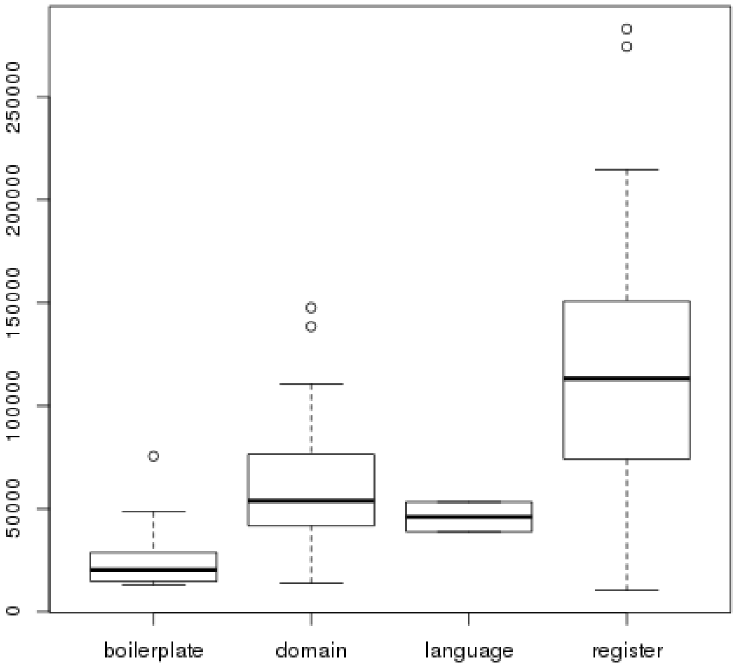

Fig. 2a. Number of word tokens assigned to a topic in a whole sample (by topic type)

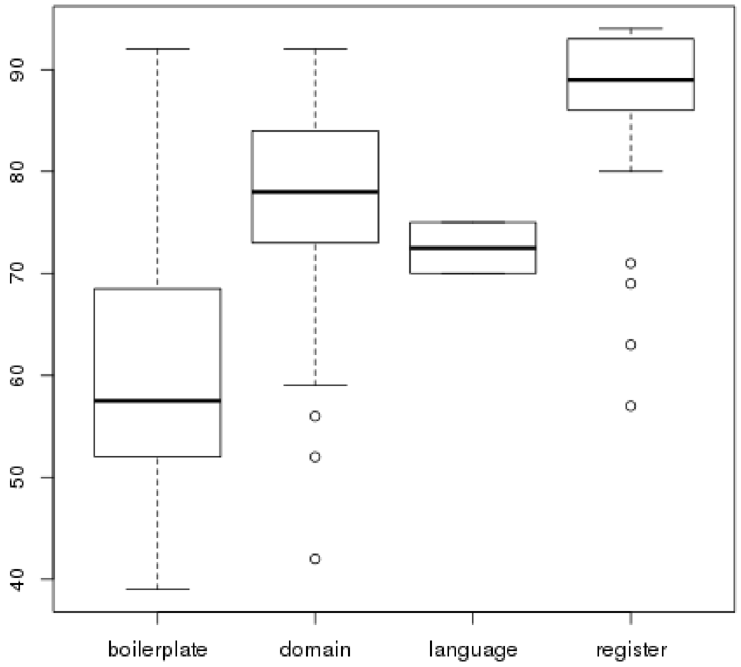

Fig. 2b. Percentage of top words in topic which are shared in both “active” and “quiet” periods (by topic type)

Domain topics are typically stable at the level of around 80 out of 100 top words in the December sample compared to September. For example, in the topic *church* there are 79 common words in top-100 list for September and December, e.g. *church, temple, Russian Orthodox, god, Saint, Christ* etc. Among words specific for top-100 in September there are *Okhlobystin, Chaplin* (political characters who made resonant public claims in September)*, Jew, Jerusalem* etc; in December — *Catholic, Christmas, saviour, shrine/relic* etc. This example shows that, on the basis of the topics found, one cannot regain the “agenda list” that was discussed in the blogosphere, because mentions of specific events and persons fit into the topic which is formed by more frequent and general vocabulary referring to a specific subject area, such as Russian Orthodox church, military operations (topic *Libya*) or legal proceedings (topic *court*). It appears correct to regard the LDA results for blogosphere's texts as helpful for estimating the quantitative trends in the frequency of various subject areas.

## 4.2 Popularity Dynamics of Topics

The simplest way to estimate changes in the topical structure of the blogosphere is to compare the relative weights of topics, i.e. to calculate the proportion of the text collection that is devoted to a topic, for each period. When LDA is applied, each text is distributed between several topics, in different proportions, therefore it is natural to talk about the share of a specific topic in each of the posts, instead of the number of posts devoted to the topic. Therefore, the relative weight of a topic can be calculated by summarizing the topic's proportions for each text and dividing by the total amount of texts. The number obtained can be considered as the total proportion of the topic in the text collection, normalized by the number of posts and the length of a post, because the same topic weight will correspond to a different number of words in a source text depending on its length. For example, in a 5-word post a topic weight of 0.2 corresponds to one word, while in a 500-word post the same weight corresponds to 100 words allocated to the topic. Normalization by text length takes into account the fact that a post is a communicative unit in blogosphere and allows us to estimate the general tendency to mention a topic, irrespective of volubility of the author.

Another way to calculate the relative weight of a topic is to normalize it by the number of words in the collection only, which requires the calculation of the total number of words allocated to the topic in all the considered texts. With this normalization method, topics discussed in longer texts will get more relative weight compared to those discussed in shorter texts.

Changes in the relative weight of topics in the December collection compared to those in September, estimated in proportions of texts and in proportions of words, reveal high correlation with each other ( $r=0{,}94$ ). Our data display no abrupt changes in topic volumes: the range of changes in text proportions is from $-0{,}65$ % to $1{,}92$ % (median $-0{,}08$ %), in word proportions from $-0{,}87$ % to $1{,}86$ % (median $-0{,}06$ %). However, there is a distinct tendency to general and relatively proportional weight decrease for most topics, accompanied by even more distinct increase in several topics (Fig. 3). This picture may be interpreted as a relative decrease of topical diversity in December.

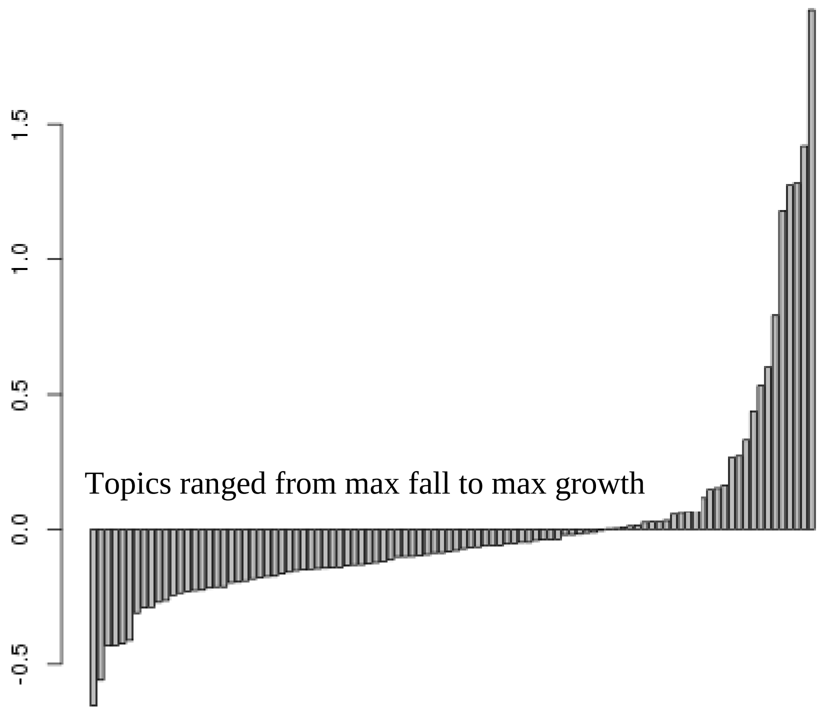

Fig. 3. Changes of topic weights in December compared to September 2011 (in percentage of documents in the collection)

For the list of topics that decreased their relative weight to the maximum extent in December, see Table 1. Part of the changes may be explained by the fact that the topic's core consists of words which are closely associated with specific events and newsmakers that had lost their relevance by December. These topics include *Libya* (military events in Libya and other Arab countries) and *airplane* (aviation vocabulary, associated with the crash of Yak-42 in Yaroslavl). Changes in other topics may be explained by seasonal drop of relevance, e.g. *tourism* and *school* (September 1, the start of the Russian school year). Other changes may be caused by random fluctuations or more complicated conceptual processes that have no obvious explanations at the level of semantic lexical groups.

**Table 1. Topics showing decrease in December 2011**

| № | label | description | in documents | in words | type |
|---|---|---|---|---|---|
| 52 | Libya | revolution in the Middle East | -0.65 | -0.87 | domain |
| 40 | tourism | travel reports | -0.56 | -0.53 | domain |
| 50 | prison | criminal news | -0.43 | -0.62 | domain |
| 89 | today | immediate records | -0.43 | -0.28 | register |
| 92 | airplane | aviation, flight accidents | -0.42 | -0.33 | domain |
| 39 | street | urban environment and architecture | -0.41 | -0.51 | domain |
| 84 | football | football news | -0.31 | -0.28 | domain |
| 82 | school | school system | -0.29 | -0.28 | domain |
| 94 | device | computers and mobile devices | -0.29 | -0.27 | domain |
| 76 | family | family and children | -0.27 | -0.25 | domain |
| 19 | calendar | names of weekdays, event lists | -0.26 | -0.04 | boilerplate |
| 21 | poetry | poetic texts | -0.24 | -0.35 | register |
| 73 | Europe | names of European countries | -0.24 | -0.15 | domain |
| 56 | concert | music and concerts | -0.23 | -0.23 | domain |
| 6 | war | military operations | -0.23 | -0.45 | domain |

For the list of topics with the maximum growth of relative weight in December, see Table 2. It should be noted that the topics prevailing at the top of this list are directly associated with post-electoral protests (*protest, square, revolution*), political parties or voting results (*vote, party, voting station*), legislation (*law*) or state authority/power in general (*power, Putin*). Topic *present* is the only exception associated with the discussion of New Year presents (the considered period ends on December 27), where its relatively low growth in words shows that it was encountered primarily in shorter texts. Topic *voting station* associated with discussions of the electoral process and voting results at specific voting stations, demonstrates, on the contrary, more prominent growth in the number of words, i.e. a tendency to be expressed in longer texts.

Notably, among the vast majority of domain topics, there are at least two register topics that show growth. One of them — *should* (judgements, modality) probably indicates prevalent modality and communicative functions of the majority of texts devoted to political topics that expanded considerably. It may indicate the growth of the mobilizing function of LJ, although this is not a sufficient evidence for such a conclusion. The growth of topic *to share* may mean more active reposting, i.e. actualization of the blogs' function as information transmitting medium.

**Table 2. Topics showing increase in December 2011**

| № | label | description | in documents | in words | Type |
|---|---|---|---|---|---|
| 22 | protest | protests in Moscow | 1.92 | 1.76 | Domain |
| 27 | square | Street protests and arrests | 1.42 | 1.16 | Domain |
| 68 | revolution | protests | 1.28 | 1.58 | Domain |
| 55 | present | presents for the New Year and other occasions | 1.27 | 0.77 | Domain |
| 33 | vote | voting results | 1.18 | 1.45 | Domain |
| 3 | party | political parties | 0.79 | 0.91 | Domain |
| 72 | should | judgements, modality | 0.60 | 0.71 | Register |
| 66 | power | general political vocabulary, state | 0.53 | 0.82 | Domain |
| 90 | Putin | mentions of Putin and Medvedev | 0.44 | 0.35 | Domain |
| 65 | to share | reposts: «share» at social networking sites | 0.33 | 0.18 | Register |
| 57 | Kim Jong-il | death of Kim Jong-il | 0.27 | 0.15 | Domain |
| 83 | law | legislation, Central Election Commission | 0.27 | 0.41 | Domain |
| 60 | channel | mass media and interviews | 0.16 | 0.02 | Domain |
| 34 | posted | reposts: «posted via» | 0.15 | 0.14 | Boilerplate |
| 69 | rating | reposts, blogs | 0.15 | -0.00 | Register |
| 25 | cinema | cinema and theatre | 0.12 | 0.03 | Domain |
| 38 | voting station | vote count, scrutinizers | 0.09 | 0.46 | Domain |

Topics with the smallest change in relative weight in December indicate that LiveJournal's content is stable, irrespective of seasonal or political circumstances (Table 3). The top of this list is formed by topics associated with the daily routine of an Internet user (*money* — salaries, payments; *communication* — mobile communication and Internet) or with his/her hobbies (*photography*). Besides, this list is much richer in formal and stylistic (register) topics than the previous ones, because these two categories are mostly formed by general vocabulary or by formal structures, where dramatic fluctuations in frequency are most unlikely.

**Table 3. Stable topics**

| label | description | in documents | in words | type |
|---|---|---|---|---|
| money | salaries, payments, sums of money | 0.00 | -0.07 | domain |
| communication | mobile communication and Internet | 0.00 | 0.13 | domain |
| photography | photography and photographers | -0.00 | 0.10 | domain |
| false | noise | 0.01 | 0.13 | boilerplate |
| Ukraine | Ukrainian news | -0.01 | -0.07 | domain |
| newspaper | journalism: newspapers, articles | 0.01 | 0.09 | domain |
| court | legal proceedings | -0.01 | 0.00 | domain |
| button | names of interface elements | 0.02 | 0.10 | boilerplate |
| image | links to images | -0.02 | -0.03 | boilerplate |
| occurrence | description of occurrences, narratives | -0.02 | 0.09 | register |
| restaurant | food and drinks: coffee, wine, beer, vodka | -0.02 | 0.02 | domain |
| percentage | percentage assessments | 0.03 | -0.18 | register |
| to see | general vocabulary: verbs and conjunctions | 0.03 | 0.06 | register |
| clip | leisure records: links to interesting materials | 0.03 | -0.02 | register |
| home | apartment and household | 0.03 | -0.00 | domain |
| contest | public events: contests | -0.04 | 0.07 | domain |
| mop | noise: various texts and vocabulary | -0.04 | -0.02 | boilerplate |
| head | news about officials | -0.04 | -0.02 | domain |
| smile | general colloquial vocabulary | -0.04 | -0.09 | register |
| animals | names of animals | -0.05 | -0.08 | domain |

### 4.3 Distribution of Topics by Authors

Another possible estimation of the blogosphere's topic structure is an analysis of topic distribution by authors. Considerable changes in that kind of distribution may be interpreted as structural changes in the blogosphere's conceptual field. In this section we analyze only the posts written by 1 122 authors who have at least one post both in the September and December parts of the sample.

The principal question we are interested in is possible changes in the structure of political discussions on LiveJournal, therefore out of 100 topics allocated by LDA we selected 11 topics associated, in some way or another, with the discussions of domestic political events and actions of the authorities: *court*, *party*, *protest*, *square*, *voting station*, *prison*, *power*, *revolution*, *law*, *administration*, *Putin*. In order to identify posts considerably related to domestic political events, we used an empirically established cut-off criterion: where the total proportion of “domestic political” topics was larger than 0.3, the text was qualified as “political”, otherwise it was qualified as “non-political”. It should be noted that this criterion cannot be used to estimate the author's interest in domestic political affairs, which is not only due to limitations of LDA, but also due to the vagueness of the “domestic policy” concept, from which we excluded topics associated with a wider range of social issues, such as *education* or *society*. However, this approach may be considered as a way to roughly estimate the number of posts with non-random share of words associated with the domestic political events.

The statistics for the total number of posts is given in Table 4. Against the background of the general growth of the average number of posts, the average number of political posts also increased, while the average number of non-political posts decreased. When interpreting the growth of the number of political posts it should be taken into account that a considerable number of domestic political topics are closely associated with the news that evolved in December. Therefore, in the September part of the collection, one may only expect to see texts that were allocated to these topics mostly due to peripheral vocabulary or random coincidence.

**Table 4. Statistics for the number of political and nonpolitical posts***

| Author's posts | September | | December | |
|---|---|---|---|---|
| | average | median | average | median |
| Total | 19.42 | 14 | 20.14 | 15 |
| Political | 1.62 | 0 | 3.74 | 1 |
| Nonpolitical | 17.80 | 13 | 16.46 | 12 |

* $\chi^2$ significance is close to 0.

On the other hand, our data contain no positive evidence that the growth of the total number of posts in December was due to an increase in posts containing political topics. In general, in terms of posting frequency, bloggers' behaviour has been stable: the correlation between the number of a blogger's posts in September and December is quite high (r=0.80). Moderate positive correlation (r=0.40) is observed between the change in the total number of posts and the change in the number of political posts by the same author. However there is a much more expressed correlation between the change in the total number of posts and the change in the number of nonpolitical posts (r=0.88). It is highly probable that the general growth in posts number in December was due to the seasonal growth in blogger activity.

The distribution of political posts in the whole selection of users also demonstrates certain dynamics in December compared to September. For example, in September, 683 users (60.9%) had no political posts, while in December there were only 511 (45.5%) users of that kind. At the same time, 228 (20%) users with no political posts in September produced them in December. On the contrary, there were only 56 (5%) users who ceased producing political posts in December. In general, these data reveal the involvement of a greater number of bloggers in the discussion of domestic political issues in December.

## 5 Conclusion

We conclude that the results obtained by means of the automated lexical analysis based on LDA analysis indicate some structural changes in the topics of LiveJournal's top bloggers in December 2011. First, there is considerable growth in the weighting of all the topics closely associated with the discussion of voting results and protests, accompanied by a more moderate decrease in the majority of other topics. Therefore, we observe the focalization of blog topics on the problems associated with the elections and further civic engagement. Our data also contain indirect evidence of the actualization of the retransmitting function of blogs in December in an elevated frequency of reposts.

In December, one can also observe an increase in the number of users with posts where the vocabulary of domestic political topics gets considerable weight. The The number of users who who started posting texts that may be conventionally qualified as political according to LDA considerably outnumbers the number of those who ceased posting political items. This may indicate the existence of a blogger mobilization process in political topics. At the same time, the hypothesis that the blogosphere's general activity (measured by the number of posts) has increased due to the growth in political posts gets no confirmation in our data. However, due to methodological limitations of the criterion used for the classification of posts, these data should be considered very preliminary and a subject for further verification.

At this stage of the research there are not enough data to estimate the measure of structural changes revealed in the topics of LiveJournal. Further diachronic research is required in order to estimate the rhythms of topical changes and the limits of normal fluctuations in the spectrum of relevant topics that are typical of LiveJournal. Comparative material of that kind would make it possible to distinguish seasonal fluctuations on LiveJournal from significant topical changes provoked by external social processes, in order to get closer to the answer to the question about the role of the Russian blogosphere, and LiveJournal in particular, in these processes.

## References

[Alexanyan, Koltsova 2009] *K. Alexanyan, O. Koltsova.* Blogging in Russia is not Russian blogging.In: Adreinne Russel, Nabil Echchaibi (eds) **International Blogging**: **Identity, Politics, and Networked Publics. Peter Lang 2009.**

[Blei 2012] *D. Blei*. Probabilistic topic models. Communications of the ACM, 55 (4): 77–84, 2012.

[Blei, Lafferty 2006] *D. Blei and J. Lafferty*. Dynamic topic models. In Proceedings of the 23rd International Conference on Machine Learning, 2006.

[Blei, Lafferty 2009] *D. Blei and J. Lafferty*. Topic Models. In A. Srivastava and M. Sahami, editors, Text Mining: Classification, Clustering, and Applications. Chapman & Hall/CRC Data Mining and Knowledge Discovery Series, 2009.

[Blei et al. 2003] Blei, David M.; Ng, Andrew Y.; Jordan, Michael I; Lafferty, John. Latent Dirichlet allo- cation. Journal of Machine Learning Research 3: pp. 993–1022. doi:10.1162, 2003.

[Buntine, Jakulin 2006] *W. Buntine and A. Jakulin*. Discrete component analysis. In Subspace, Latent Structure and Feature Selection. Springer, 2006.

[Etling et al. 2010] *Etling B., Alexanyan, K., Kelly, J., Faris, R., Palfrey, J. and Gasser, U.* “Public Discourse in
the Russian Blogosphere: Mapping RuNet Politics and Mobilization”. Berkman Center Research Publication No. 2010-11. October 19, 2010.
http://cyber.law.harvard.edu/publications/2010/Public_Discourse_Russian_Blogosphere (accessed 31 July 2012).

[Daud et al 2009] Daud, A., Li, J., Zhou, L., Muhammad, F. Knowledge discovery through directed probabilistic topic models: a survey. Front. Comput. Sci. China. DOI 10.1007/s11704-009-0062-y

[Gorny 2004] *Gorny E.* Russian LiveJournal: National specifics in the Development of a Virtual Community. Version 1.0 of 13 May 2004. Russian-cyberspace.org http://www.ruhr-uni-bochum.de/russ-cyb/library/texts/en/gorny_rlj.pdf. (Accessed 05.04.2012.)

[Griffiths, Steyvers 2004] *Griffiths, T.L. and Steyvers, M.*Finding scientific topics. Proceedings of the National Academy of Sciences of the United States of America, 101. Suppl. 1. 5228-—5235.[Hall et al. 2008] *Hall, D. and Jurafsky, D. and Manning, C.D.*, Studying the history of ideas using topic models // Proceedings of the Conference on Empirical Methods in Natural Language Processing. 363–371. 2008.

[Heinrich 2008] Heinrich G. Parameter Estimation for Text Analysis. Technical report, Version 2.

[Lonkila 2008] *Lonkila, Markku* (2008): The Internet and Anti-military Activism in Russia, Europe-Asia Studies, 60:7, 1125—1149.

[Parkhomenko, Tait 2008] *Parkhomenko Y., Tait A.* Blog Talk // Index on Censorship. February 2008 37: 174-178, doi:10.1080/03064220701882822. http://ioc.sagepub.com/content/37/1/174.citation (Accessed 08.06.2012.)

[Ramage et al. 2009] *Ramage D., Rosen E., Chuang J., Manning C.D., McFarland D.A.* Topic Modeling for the Social Sciences. NIPS 2009 Workshop on Applications for Topic Models (Accessed 19.04.2012.)

[Ramage et al. 2010] *Ramage D., Dumais S., Liebling D.* Characterising Microblogs with Topic Models. ICWSM 2010. http://www.stanford.edu/ dramage/papers/twitter-icwsm10.pdf (Accessed 19.04.2012.)

[Segalovich 2003] *Segalovich, I.* A fast morphological algorithm with unknown word guessing induced by a dictionary for a web search engine. // Proceedings of MLMTA. 2003.

http://download.yandex.ru/company/iseg-las-vegas.pdf (Accessed 08.06.2012.)

[Sugar, James 2003] Sugar C., James G. Finding the Number of Clusters in a Data Set: An Information Theoretic Approach. Journal of the American Statistical Association, 98:750–763.

[Кольцова, Маслинский 2013] Чем дышит блогосфера? К методологии анализа больших текстовых данных для социологических задач. *Социология: методология, методы и математическое моделирование* (in print).

Kirill Maslinsky
National Research University Higher School of Economics (St.-Petersburg, Russia). Internet Studies Laboratory. Researcher;
E-mail: kmaslinsky@hse.spb.ru

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