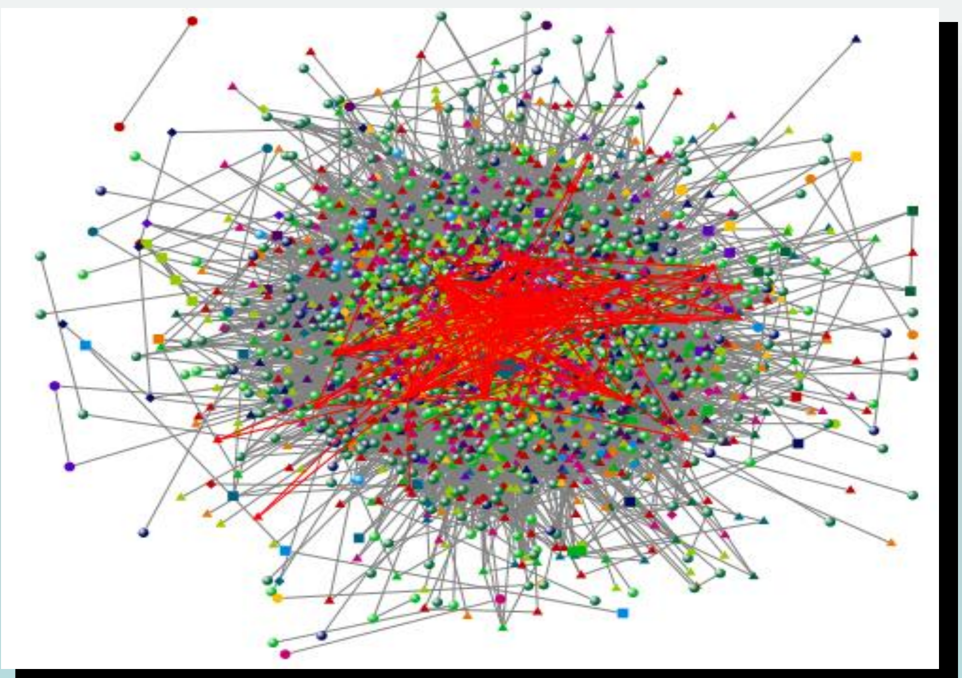
Comment-based communities in the Russian Livejournal and their topical coherence



Olessia Koltsova,
Sergey Koltcov,
Sergey Nikolenko

www.linis.hse.ru

Sunbelt 2013, Hamburg

RESEARCH AGENDA

- Online discussions are socially important.
- In blogs, they develop in comments.
- Comment-based networks may contain denser areas – communities – indicative of some (problematic) social issues.
- However, most research on communities in blogs has been on friendship networks (Lescovec 2008, Zakharov 2007).
- Comment-based network research uses authors, not posts as nodes (Adamic et al 2008, Ali-Hasan & Adamic 2009, Gomez et al 2008).

RESEARCH QUESTIONS

- Do comment-based communities exist?
  - Comment-based community in blogs: exists when a certain (fuzzy) set of posts or bloggers is commented by a certain set of bloggers
- If so, do they form around common topics of the commented posts or around authors of the commented posts?

NETWORK CONSTRUCTION

- For greedy community detection algorithms → bimodal post-commentator network projected to post-post network
- Two posts are considered connected if they have been commented by the same blogger
  - If they have been commented by two different bloggers, they gain two edges in common
  - If they have been commented twice y one blogger, they gain two edges in common
  - Self-commenting is excluded.

RUSSIAN BLOGOSPHERE AND LiveJournal

- Russian blogosphere: about 58 mln blogs, 7-8 mln posts a day(without microblogs).
- Commenting: mostly locked within blog platforms (around 100, 5-6 leading).
- Livejournal (most politicized): 2 (4) mln accounts, 60-70 thousand posts a day.
- Followers-based ratings of bloggers are important in Russia.
- At rating level of 150 thousand LJ produces less than 1 post per blogger per month.

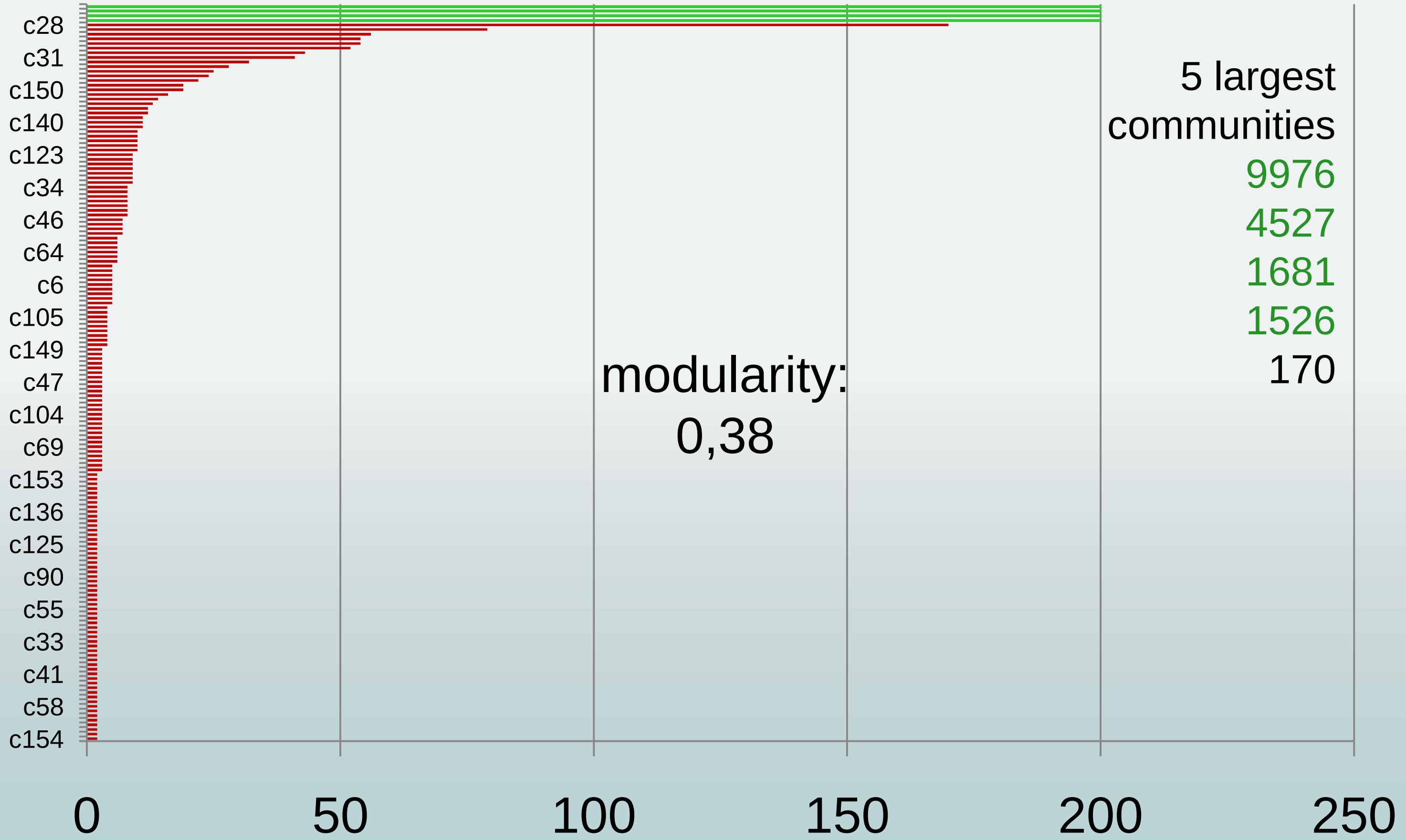
DATA

- Top LJ 2000 bloggers (have 500+ followers, produce avg. 1 post per blogger per day, receive 20 times more comments).
- Time: April 1 – April 7 2013 (reasonable period for an event life-cycle, also: computational complexity limits).
- 24619 posts total,19039 posts with comments,1653 excluded for technical reasons = 17 386 posts in analysis.
- 520 549 comments
- ≈ 4,5 mln edges “post-post”, after self-comments are excluded; 391 posts had no shared commentators.
- 1667 authors, 56217 commentators

METHODS 1

- Data collection: Koltran / LINIS BlogMiner software (full-text & relational structure of LJ).
- Community detection: Louvain algorithm, developers' code.
- Community belonging / authorship correlation: SPSS, nominal measures of association.
- Topic similarity detection: LINIS TopicMiner & C++ codes:
  - Text clearing, cutting & lemmatization;
  - TF/IDF calculation (texts represented as lists of frequencies of words in them);
  - Cosine similarity calculation (each pair of texts compared on the basis of words frequencies in them);
  - Average similarity within comment communities compared to global average similarity.

COMMUNITY STRUCTURE



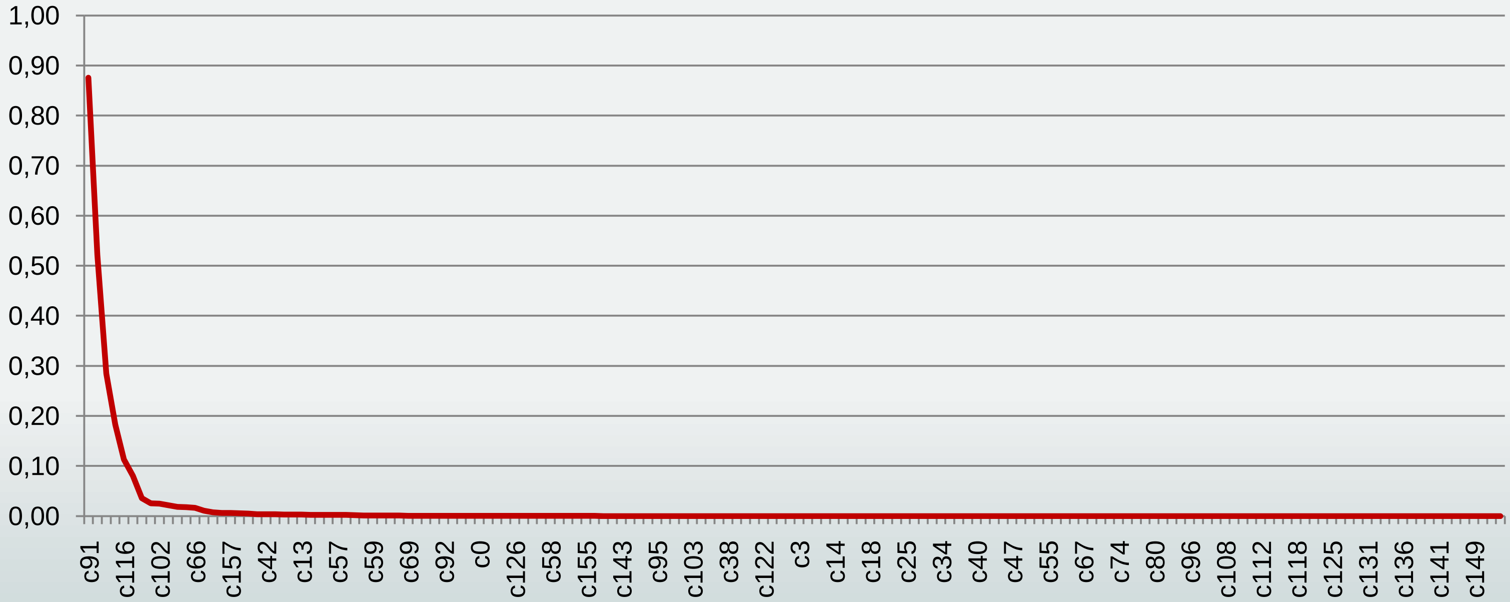
Number of posts in communities: communities 0-158; number range: 2-9976
Louvain, level 1.

AUTHORSHIP

		Value	Asympt.Std. Error	Approx. T	Approx. sig.
Lambda	Symmetric	,209	,003	59,644	,000
	Dependent blogger	,057	,002	26,346	,000
	Dependent community	**,522**	,007	56,832	,000
Goodman & Kruskal Tau	Dependent blogger	,041	,001		,000
	Dependent community	**,510**	,004		,000
Cramer's V		**,466**			,000
Contingency Coefficient		**,985**			,000

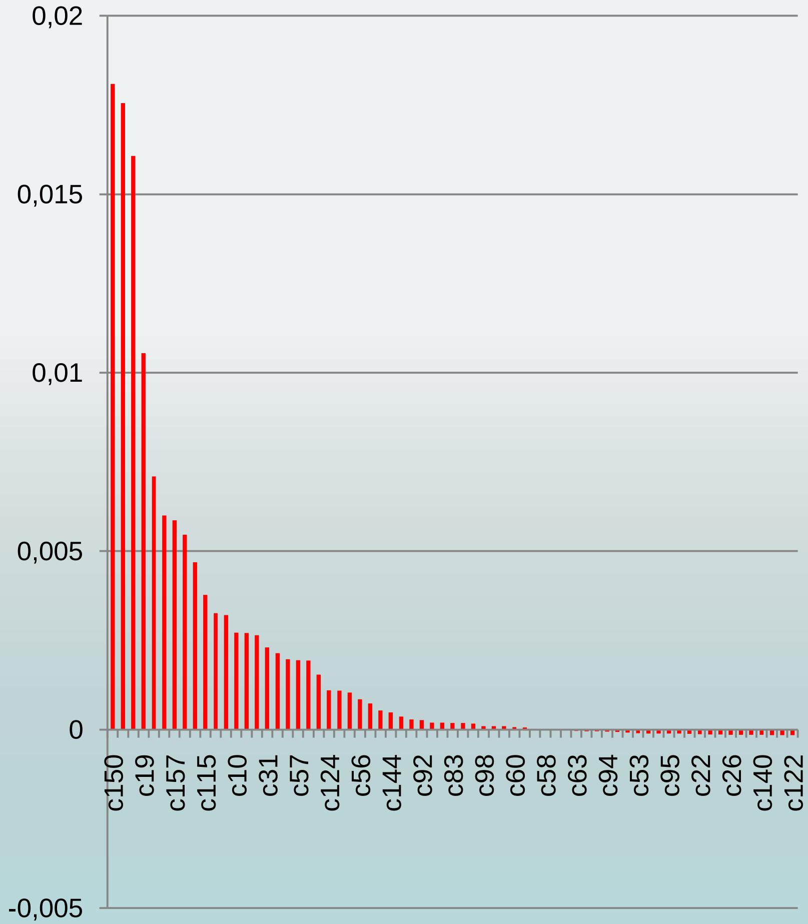
Belonging of a post to a community strongly depends on the post's authorship. I.e. communities tend to form around authors.

TOPICAL SIMILARITY 1



- Global average cosine similarity: 0,00015924;
- Intra-commmunity average cosine similarity: 0,04916513 .
- Distribution of intra-community cosine similarity means (see above) is power-law: there are tighter and looser communities.

TOPICAL SIMILARITY 2



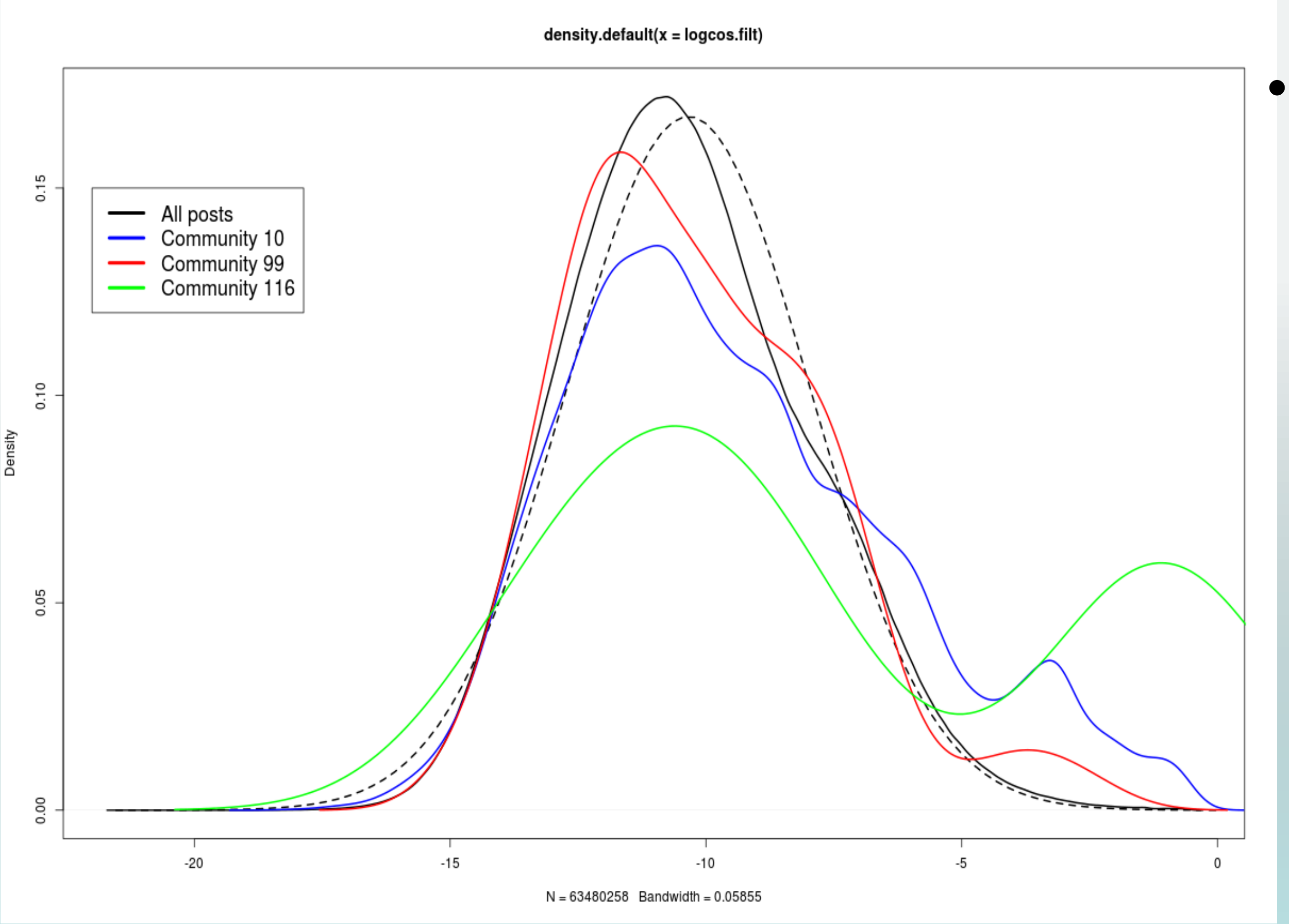
Middle part of intra-community cosine similarity means distribution. X axis: global average cosine similarity

Below average are multiple, but extremely small numbers. I.e. topical similarity in a certain set of communities is manifest.

TOPICS IN COMMUNITIES: EXAMPLES

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

TOPICS IN COMMUNITIES: INDICATORS

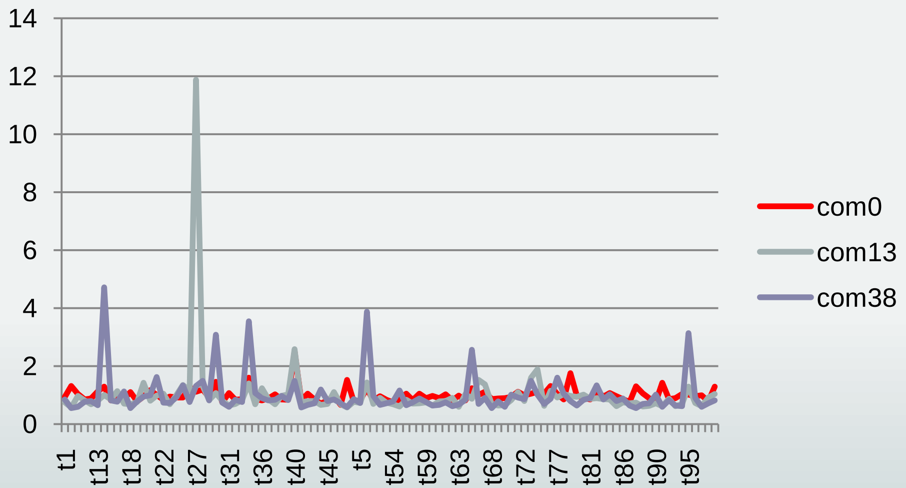


- Distributions of logarithms of cosine distances in communities where dominant topics are clearly present, have additional peaks.

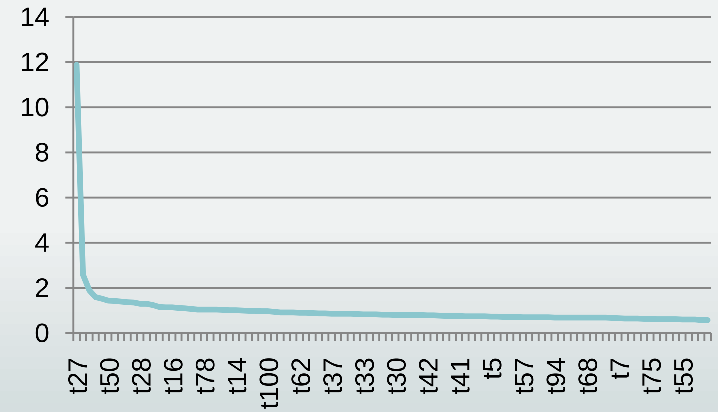
METHODS 2

- LDA Gibbs-sampling 100-topic modeling (software: LINIS TopicMiner)
- Total weight of each topic calculated for each comment-based community
- Normalized
- Topics’ weights variance calculated for each community
- Low variance = multitopic communities

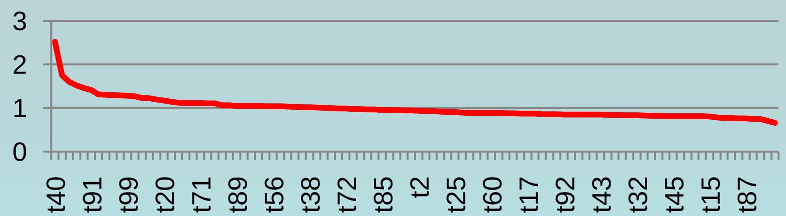
MONO- AND MULTITOPICAL COMMUNITIES



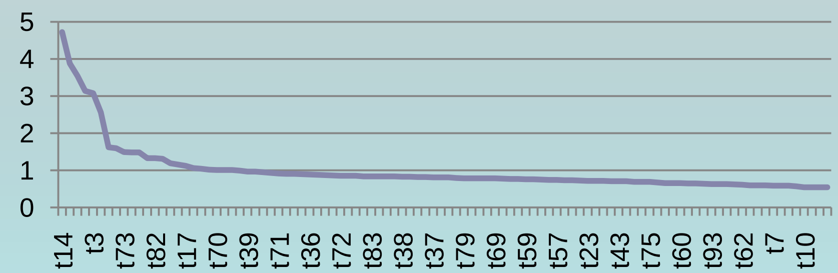
Com13: books, film, fashion



Com0: all topics



Com38: personal stories, travel, fashion, pets



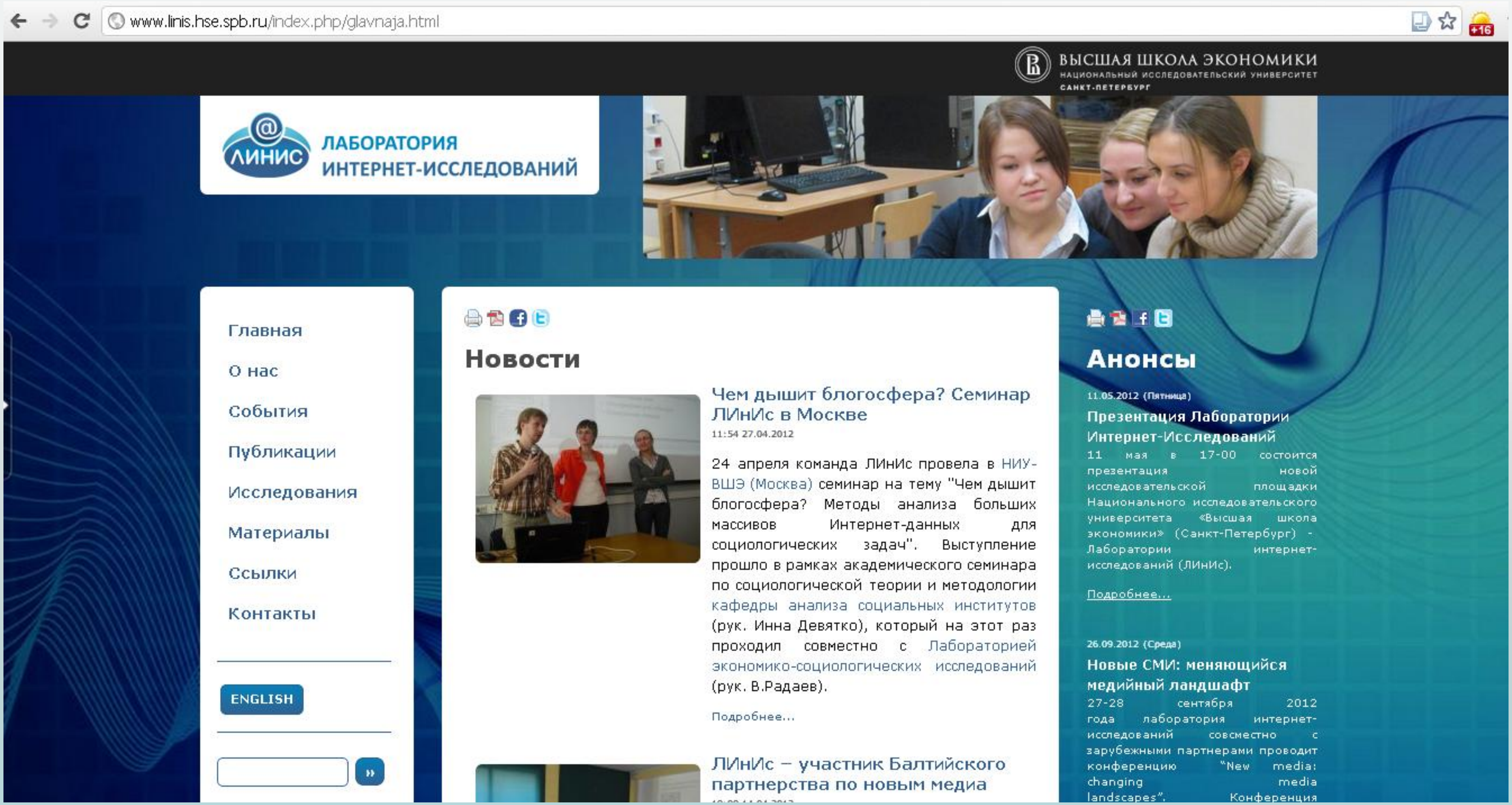
CONCLUSIONS

- Comment-based communities in top LJ exist; community structure moderately manifest.
- Communities are uneven in size.
- Graph is sparse and interconnected by a minority of active commentators.
- Most comments are done by non-top bloggers (fandom commenting)
- Communities strongly tend to emerge around authors of posts.
- Communities have a less manifest tendency to form around topics.
- Some communities are clearly centered around a limited number of topics; they can be detected and described.

FUTURE RESEARCH

- Finalizing LDA results interpretation
- Inclusion of texts of comments into topic modeling.
- Bimodal post-commentator network clustering (inclusion of info about authors of comments).
- Author-commentator network analysis (fandom communities mining).

THANKS!



www.linis.hse.ru

INTERNET STUDIES LAB

HIGHER SCHOOL OF ECONOMICS
NATIONAL RESEARCH UNIVERSITY
SAINT PETERSBURG

ACKNOWLEDGEMENTS

The study is funded by the Fund for Fundamental Research, Higher School of Economics, Russia

CONTRIBUTORS

Olessia Koltsova,

Head of laboratory

Sergey Nikolenko,

mathematician, senior researcher

Sergei Koltcov,

Physicist, techni-cal director

Anastasia Shimorina,

Computer linguist, junior researcher

Ruslan Bahmudov,

Program developer

Yury Rykov,

Sociologist, PhD student, intern

Victoria Seneva,

Sociology MA student, intern