DEVELOPING METHODS FOR SEMANTIC & NETWROK ANALYSIS OF BLOGS: DISCOURSES OF ISLAM IN THE RUSSIAN BLOGOSPHERE

Olessia Koltsova
Lidia Pivovarova
Anastasia Kincharova
Tatiana Yefimova
Elisaveta Tereschenko
Yulia Pavlova

HIGHER SCHOOL OF ECONOMICS
NATIONAL RESEARCH UNIVERSITY

PROJECT GOALS

- to extract a population and samples of all texts of a given topic (Islam) (with hyperlinks)
- to divide the corpus of texts into semantically similar clusters
- to detect hyperlink-based subgraphs (cohesive subgroups) in the given corpus
- to juxtapose semantic clusters and cohesive subgroups

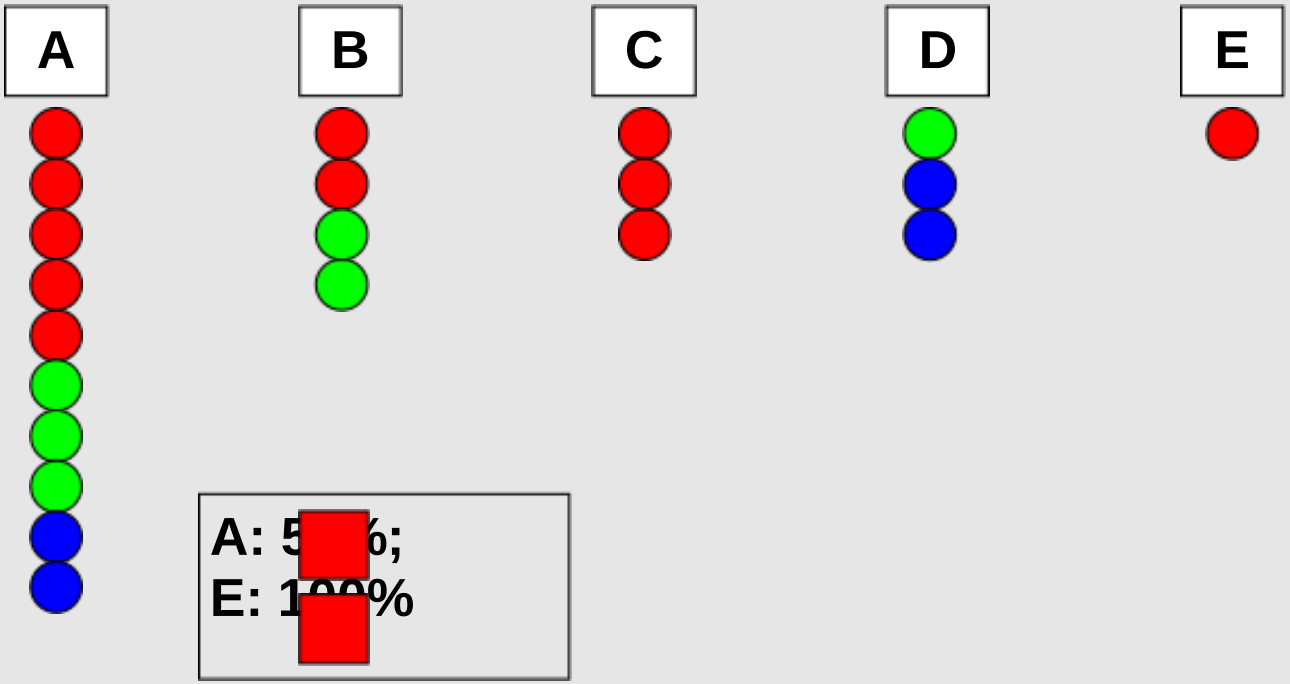
DATA

- Two one-month collections of Russian blog texts & metadata from Yandex archive
- Russian blogosphere graph file from Yandex
- One-month collection contains:
  - ca. 30 mln “proper”posts
  - ca. 210 mln “proper” & microblog posts
  - over bln comments
- *Forming a population of topic-relevant texts is a special task involving human coding & expertise*

UNIT OF SEMANTIC ANALYSIS

- Entire blogs are multi-topical and can not be clusterized except by fuzzy clustering
  - Problem A: still much noise
- Single posts are usually uni-topical and can be divided into strict clusters with low noise
  - Problem B: juxtaposing with SNA results
- Populations of topic-relevant posts from each blog can be units to be fuzzily clusterized with low noise
  - Problem C: blogs with more posts will have lower coefficients of belonging to clusters than single-post blogs

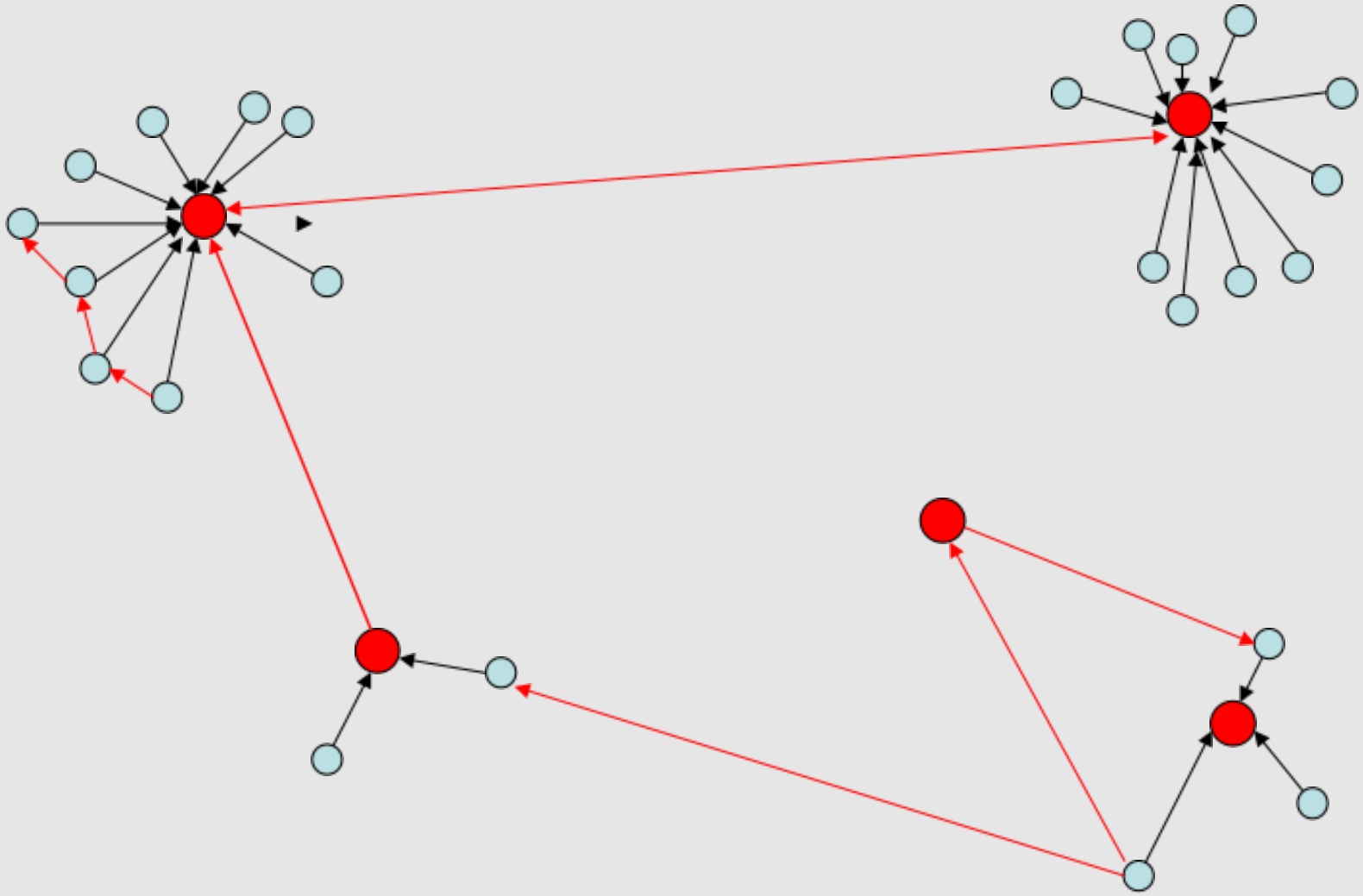
PROBLEM C



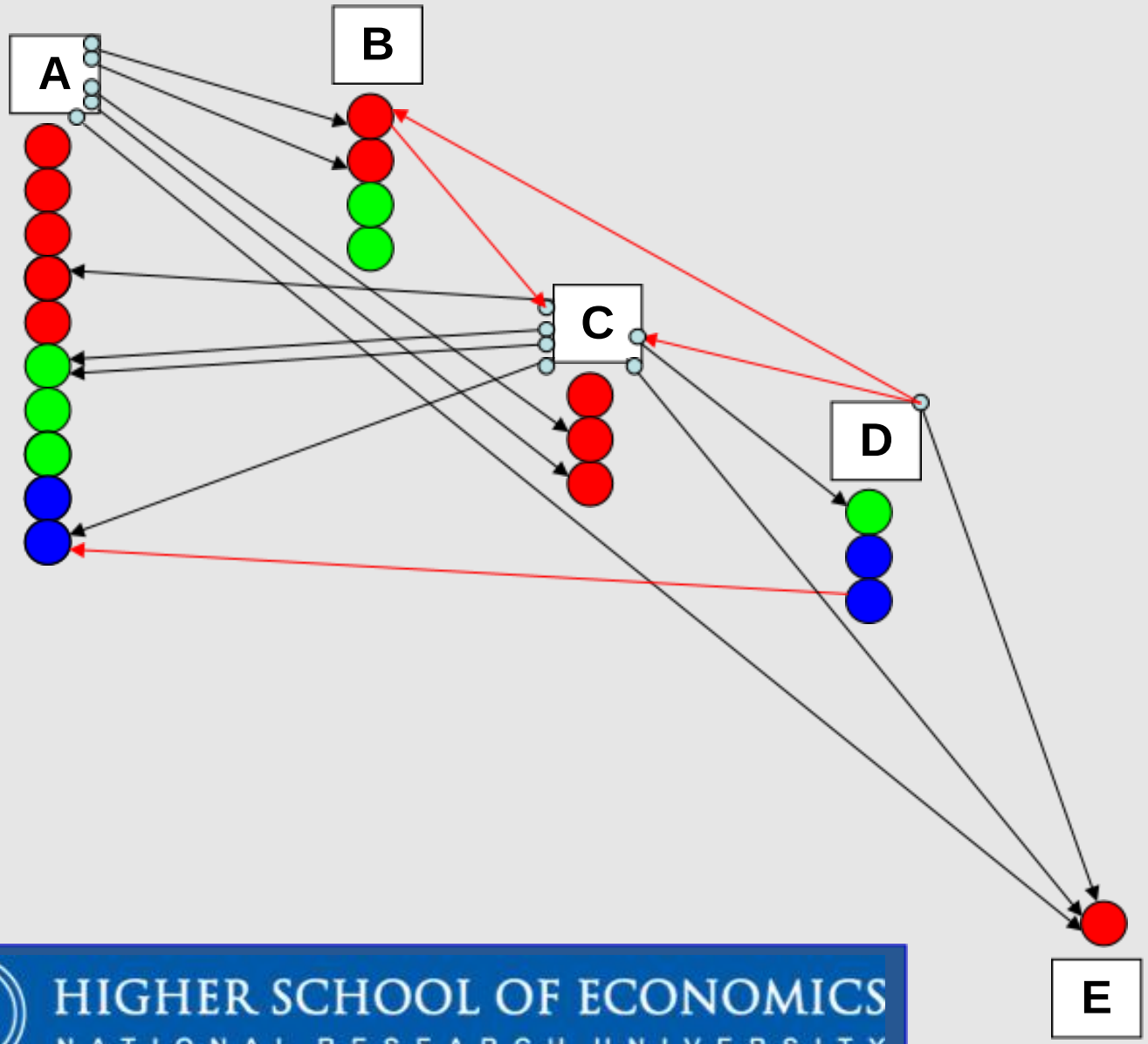
UNIT OF NETWORK ANALYSIS

- Entire blogs: network is easily interpreted
  - Problem 1.1: uncomparable with semantic clusters of posts
  - Problem 1.2: structure of intext and friending links in the Russian blogosphere (fusion of blogplatforms and social network platforms; platform dependence)
- Posts: data comparable
  - Problem 2.1: too few links between posts
  - Problem 2.2: too many links to non-blog resources
- Posts and comments: detects real conversational networks
  - Problem 3.1: star-like loosely connected subgraphs with unhomogeneous nodes and ties

PROBLEM 3.1.



SOLUTION & NEW PROBLEMS



PROBLEM OF SUBGROUP DETECTION

- Problem 1: choice of metrics
  - Cliques do not apply
  - N-cliques - ?
  - N-clans - ?
  - K-plexes and cores seem to apply if weighted by the size of the subgroup
- Problem 2: choice of software
  - It should work with large datasets
  - It should fulfill subgroup detection and clustering

Thank you for listening
about our problems