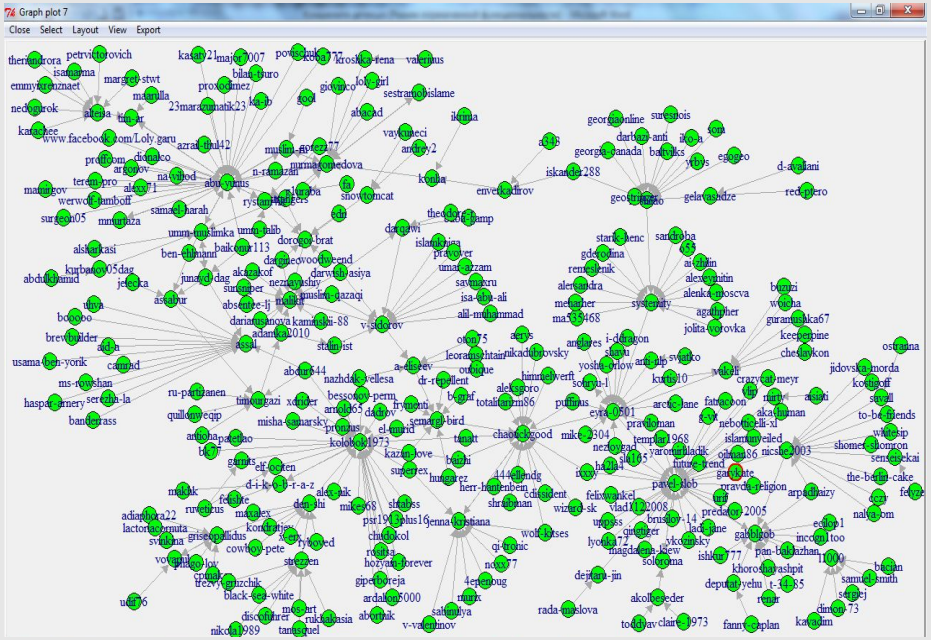
ВНУГ

Интернет-исследования

Руководитель:

Олеся Кольцова

Высшая школа экономики – Санкт-Петербург



О ВНУГе

- ВНУГ вырос из проекта «Учитель-ученики», грант Научного фонда НИУ-ВШЭ 11040006, 2011-2012 гг.
- После получения гранта ЦФИ на 2012-2013 год преобразуется во временную лабораторию интернет-исследований
- Участники:
  - Олеся Кольцова (руководитель)
  - Анастасия Кинчарова (сетевой анализ)
  - Кирилл Маслинский (анализ текстов)
  - Елизавета Терещенко (анализ текстов - стажер)
  - Юлия Павлова (анализ текстов – стажер)
  - Татьяна Ефимова (анализ текстов, администратор)
  - Сергей Кольцов (постановщик задач, математик)
  - Руслан Бахмудов (программист)
  - Виктория Сенева (сетевой анализ - стажер)
  - Алиса Баснарева (анализ текстов – волонтер)

ЗАДАЧИ ЛАБОРАТОРИИ

- выявление спектров мнений в сети по социально значимым темам, изучение структуры и динамики сообществ, характера распространения информации в сети, предикция социальной мобилизации через интернет
- Разработка методов решения этих задач, в т.ч. адаптация матметодов, решение проблем сбора данных, создание баз данных

МЕТОДЫ

- Автоматизированные методы анализа текстов, основанные на подходе bag of words: кластеризация, выявление тем (topic detection, topic modeling), sentiment analysis
- Методы сетевого анализа сетей комментирования

*большие массивы данных

ЗАДАЧИ ВНУГа

- Доработка программного обеспечения Koltran BlogMiner
- Продолжение выявления тематической структуры блогосферы с на основе Латентной Дирихле-аллокации (инструмент Stanford Topic Modelling Toolbox)
- Адаптация методов sentiment analysis для выявления эмоциональной заряженности групп блогов.
- Волонтерский проект: освещение протестов декабря 2011 – тексты и сообщества комментирования

ДАННЫЕ

- Сплошная закачка постов, комментариев и метаданных ЖЖ на основе собственного ПО Koltran Blogminer
- На данный момент: несколько тестовых выборок за август – декабрь из топ-2000 блоггеров.

Спасибо за внимание!

koltsova@hse.spb.ru

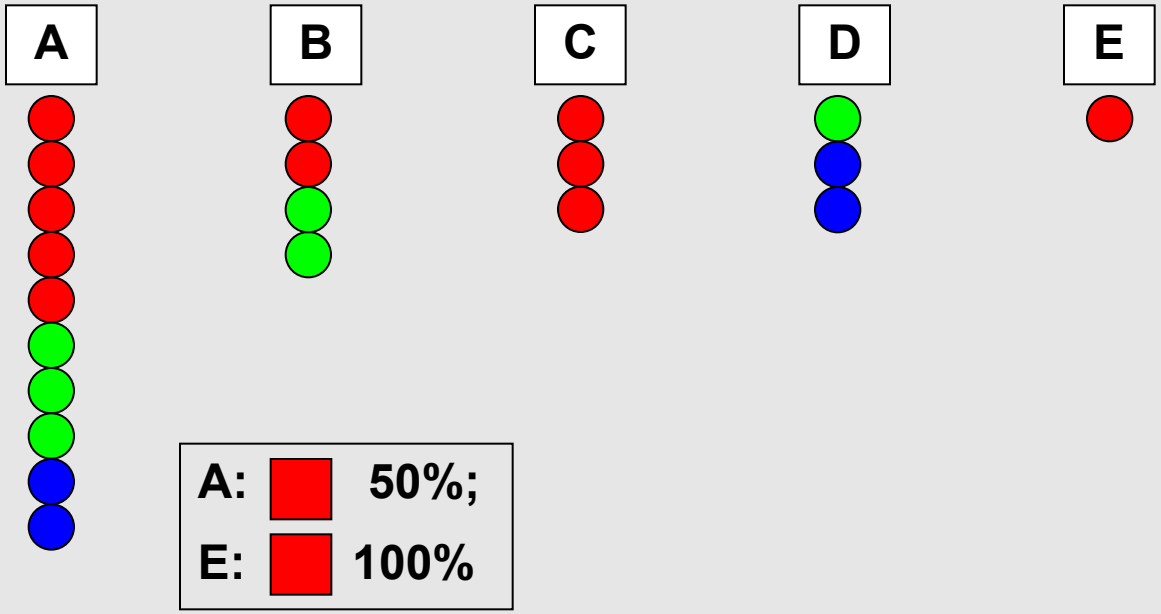
blogruresearch@gmail.com

Дополнительные слайды

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

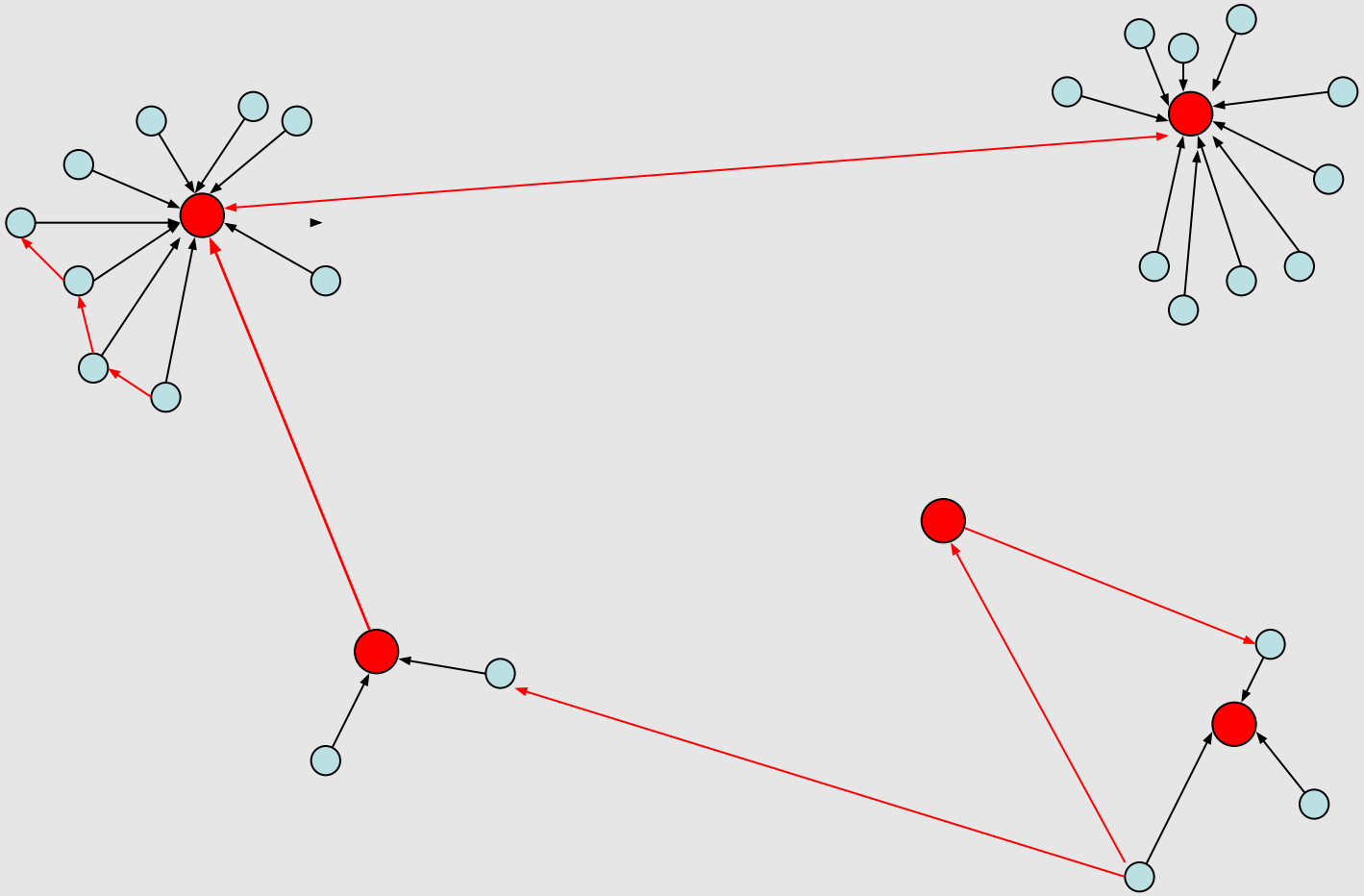
PROBLEM C



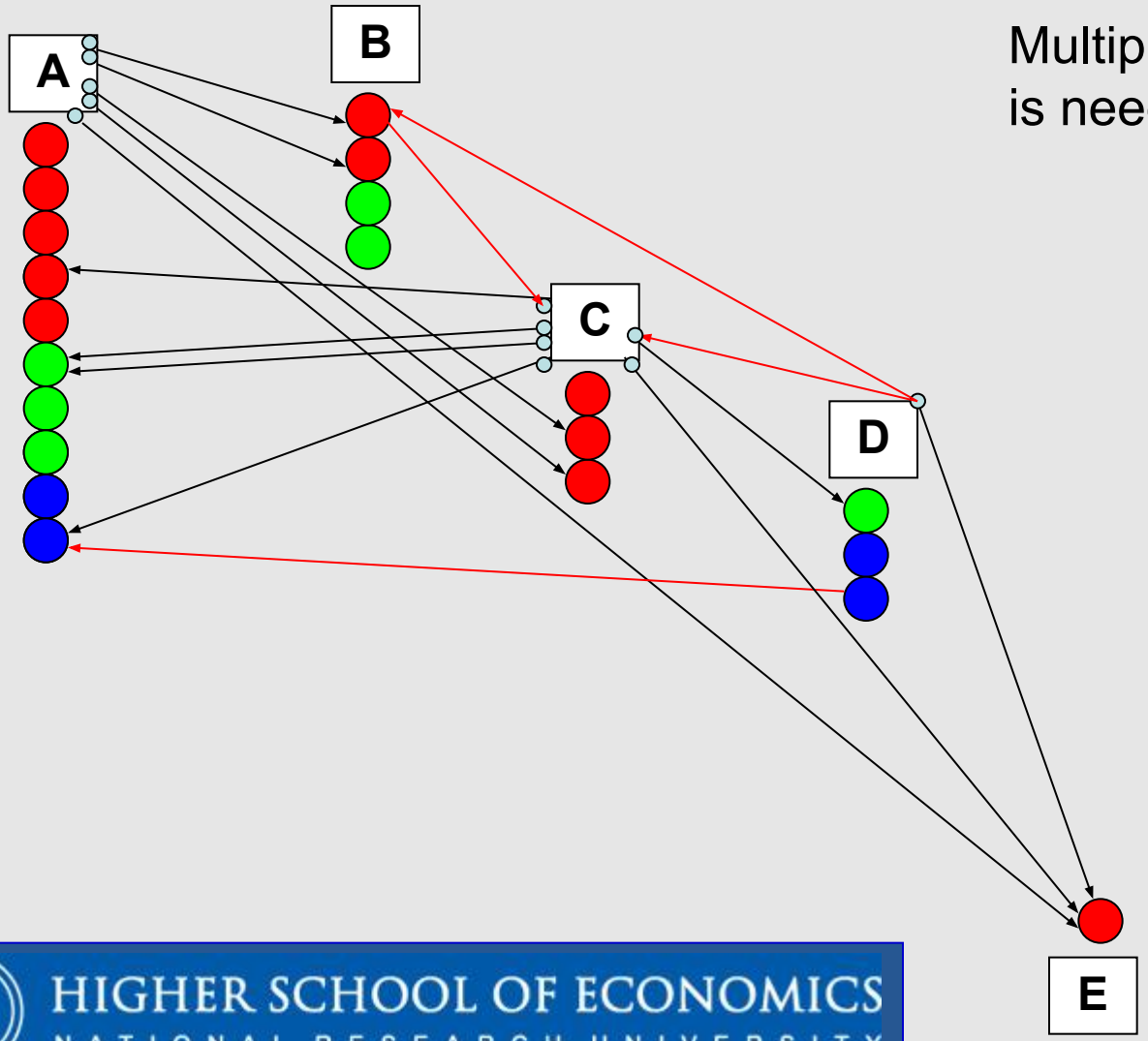
UNIT OF NETWORK ANALYSIS

- 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



Multiplex graph analysis is needed?

PROBLEM OF SUBGROUP / COMMUNITY DETECTION

- Problem 1: choice of definition
  - Traditional (n-cliques / n-clans, k-plexes / k-cores, LS-sets / λ-sets)
  - Definitions based on comparison with random graphs
  - Definitions based on vertex similarity
- Problem 2: choice of algorithms
- Problem 3: choice of software
  - It should work with large datasets
  - It should contain applicable algorythms

HIGHER SCHOOL OF ECONOMICS
NATIONAL RESEARCH UNIVERSITY