PSEUDO-BIMODAL COMMUNITY DETECTION IN TWITTER-BASED NETWORKS

Aleksandr Semenov[1], Igor Zakhlebin[1], Alexander Tolmach[2], Sergey I. Nikolenko[3,4,5]

ICUMT 2016, Lisbon, October 20, 2016

[1] International Laboratory for Applied Network Research, NRU Higher School of Economics, Moscow

[2] Institute of Sociology, Russian Academy of Sciences

[3] Laboratory of Internet Studies, NRU Higher School of Economics, St. Petersburg

[4] Steklov Institute of Mathematics at St. Petersburg

[5] Kazan Federal University, Kazan

Random facts:

- on October 20, 1517, the Portugese Ferdinand Magellan, then Fernão de Magalhães, arrived to Seville where he would later secure a large grant for his voyage of circumnavigation;
- in Russia, October 20 is the Military Communication Officer Day.

TWITTER AND SOCIAL SCIENCES

- Structure, evolution, and topical content of social networks are important for computational social science.
- And Twitter is one of the most important social networks for researchers in social and political studies:
  - Twitter has been instrumental in many political movements; e.g., "Twitter revolutions" include
    - 2009 Moldova civil unrest,
    - 2009–2010 Iranian election protests,
    - 2010–2011 Tunisian revolution,
    - Egyptian revolution of 2011,
    - Euromaidan revolution in Ukraine starting from 2013.
  - there is relatively easy access to the data via Twitter API.
- Existing works mainly deal with one of two topics:
  - either they analyze the tweets themselves, as short texts,
  - or they deal with the network structure of Twitter.
- This work is in the second category...

PREVIOUS WORK

- Our subject: political polarization (people and sources tend to one of the extremes, and it's interesting to see which one).
- Adamic, Glance, *The political blogosphere and the 2004 US election: divided they blog*:
  - an already classical work from before Twitter;
  - shows clear political polarization based on hyperlink patterns;
- Conover et al., *Political polarization on twitter*:
  - studies political polarization on Twitter;
  - uses community detection to show polarization.
- Twitter gives rise to different graphs via different relations:
  - *followers* (social structure),
  - *mentions* (in tweets),
  - *retweets* (shares).

OUR MAIN HYPOTHESIS

- Our main hypothesis: *users are not equal*.
- They are roughly divided in two kinds:
  - «top» users, trendsetters, accounts of politicians, media, other celebrities, and popular bloggers with thousands of followers;
  - «bottom» users, who mainly follow «top» users due to their stance on issues, not social effects.
- These two types of users differ in their behaviour, including following other users.
- So the network becomes *pseudo-bimodal*...

COMMUNITY DETECTION

- We propose an algorithm for pseudo-bimodal community detection:
  - select a set of top users V^{top} (with some threshold k, according to a centrality measure which can be different);
  - remove internal links, making the graph bipartite (bimodal network);
  - project the graph onto one of its node sets with Newman's projection (paths of length 2); the graph becomes unimodal again;
  - run community detection (Louvain method) on the resulting one-mode network; community detection aims to maximize *modularity*

$$Q = \frac{1}{2m} \sum_{v_1, v_2} \delta(c_1, c_2) \left(1 - \frac{k_1 k_2}{2m}\right).$$

DATASETS

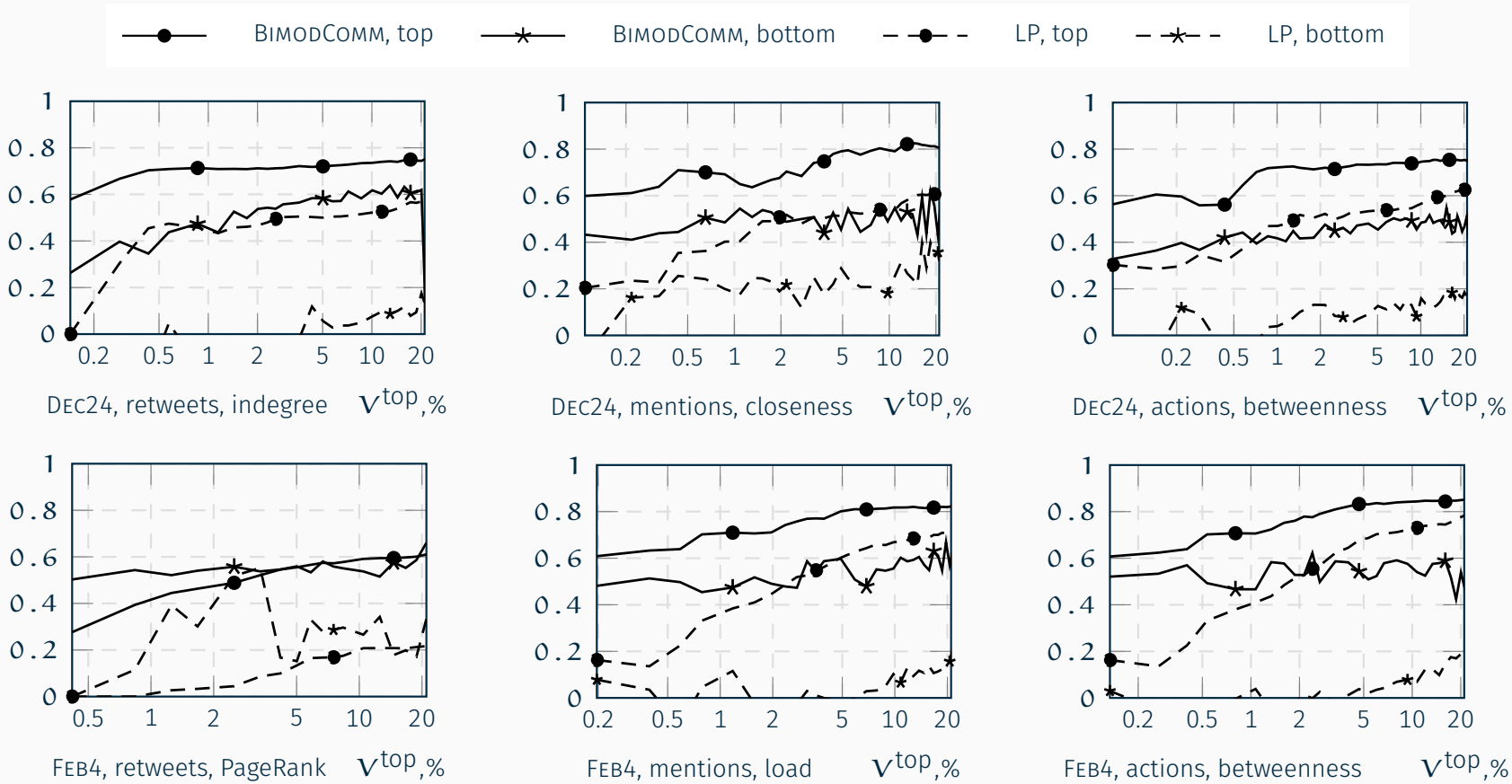
- Datasets about protest movements in Russia:
  - meetings in Moscow on December 24, 2011 (prospekt Sakharova);
  - protest meetings in Russia on February 4, 2012;
  - tweets on the World Economic Forum in Davos, 2012;
  - retweet network collected six weeks prior to the 2010 U.S. midterm elections (Conover et al.).

Dataset	Description	Number of			
		users	retweets	mentions	actions
DEC24	Russian protests on Dec 24th, 2011	3,485	6,529	6,197	12,725
FEB4	Russian protests on Feb 4th, 2012	3,742	1,498	5,893	7,391
WEF	World Economic Forum, Davos, 2012	4,555	1,977	6,372	8,348
CON	U.S. Elections, 2010	22,405	61,156	15,159	77,920

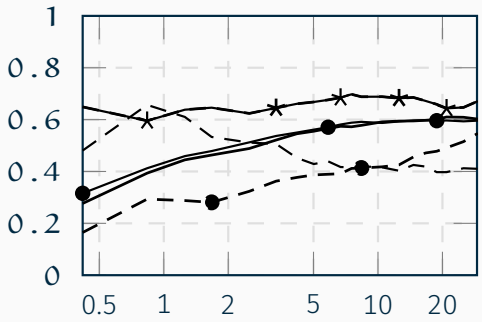
OUR EXPERIMENTS

- Two main experiments:
  - compare our algorithm with semi-supervised label propagation on the original graph;
  - compare different centrality measures for choosing top users:
    - *indegree* (% of nodes with edges incoming to v),
    - *betweenness* (total % of shortest paths between all pairs of vertices going through v),
    - *load* (simply total % of shortest paths through v),
    - *closeness* (sum of inverse shortest path sizes from v to all others),
    - *eigenvector* (for the largest eigenvalue of the adjacency matrix),
    - *PageRank* (chance that a random path will pass through v).
- The objective is to improve modularity in the resulting community structure.

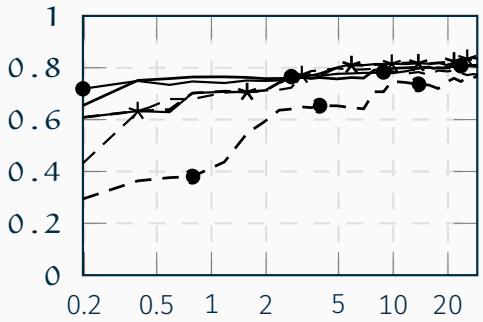
BIMODAL ALGORITHM OUTPERFORMS LABEL PROPAGATION



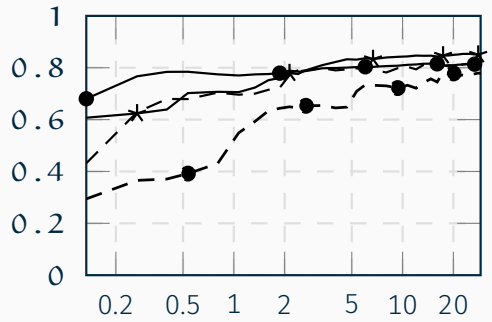
COMPARING CENTRALITY MEASURES



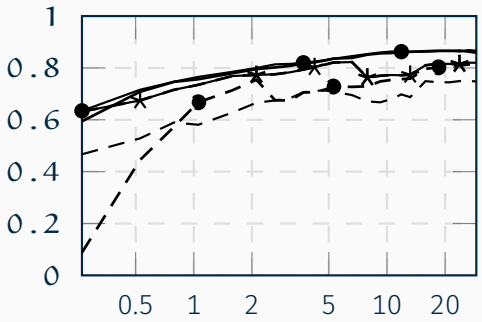
FEB4, retweets $v^{top},\%$



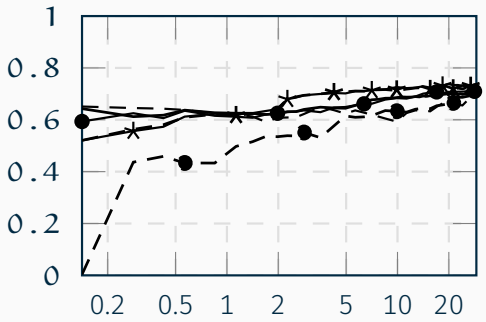
FEB4, mentions $v^{top},\%$



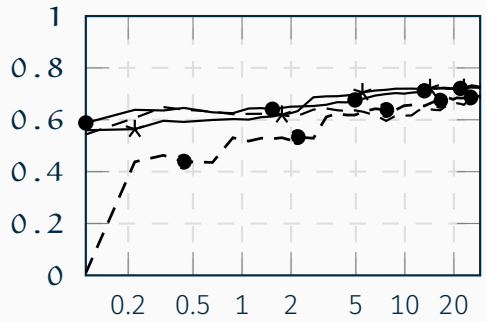
FEB4, actions $v^{top},\%$



WEF, retweets $v^{top},\%$

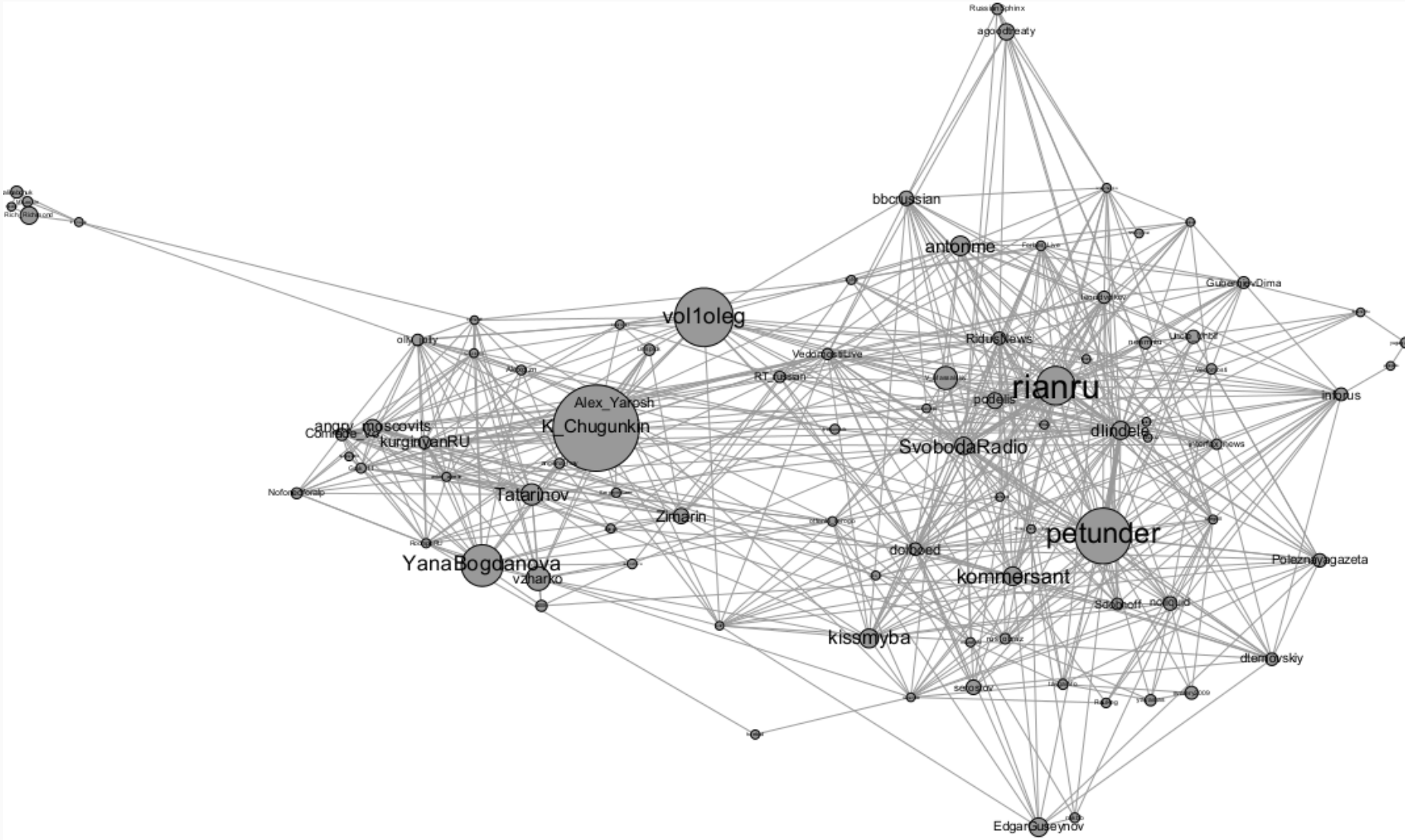


WEF, mentions $v^{top},\%$

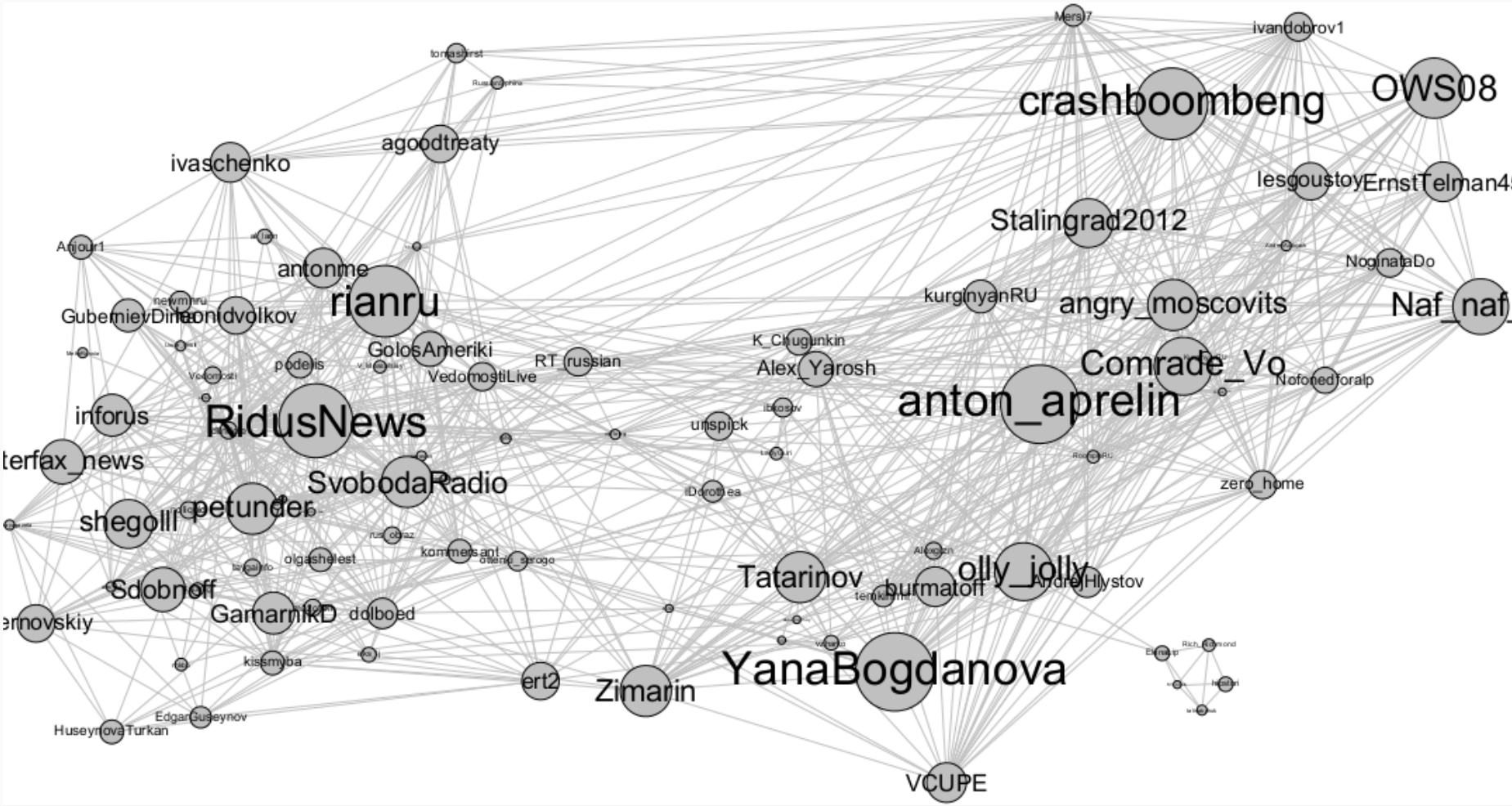


WEF, actions $v^{top},\%$

TOP USER PROJECTION, DEC24, PAGERANK



TOP USER PROJECTION, DEC24, INDEGREE CENTRALITY



THANK YOU!

Thank you for your attention!