НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ

Network Science
Project 1: Analyzing VK Friends Graph.

Alexander Plentsov

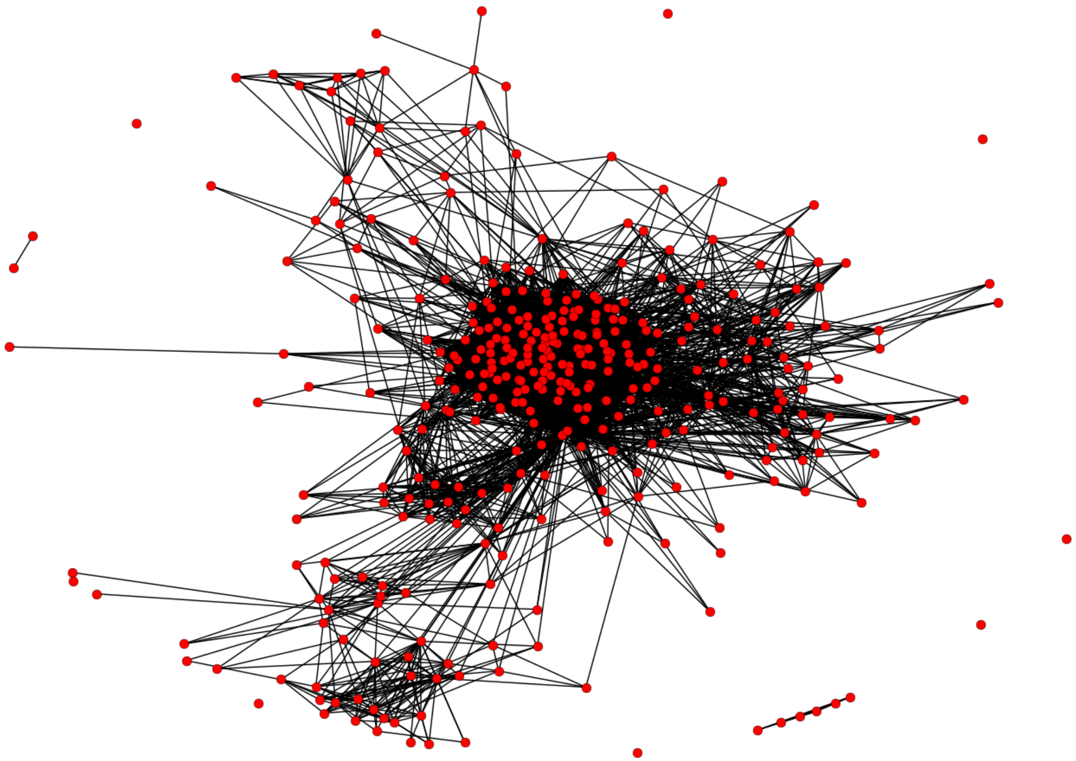
HSE Data Science, ISSA Group

Network Summary

- Network of my friends from the social network VK.com
- Network was extracted via VK.com Open API
- Each node represent one person, edge – friendship
- Each node was augmented by following attributes:
  - Sex
  - Current City
  - First Name
  - Last Name
- A total of 364 friends was extracted
- NetworkX was used for further analysis

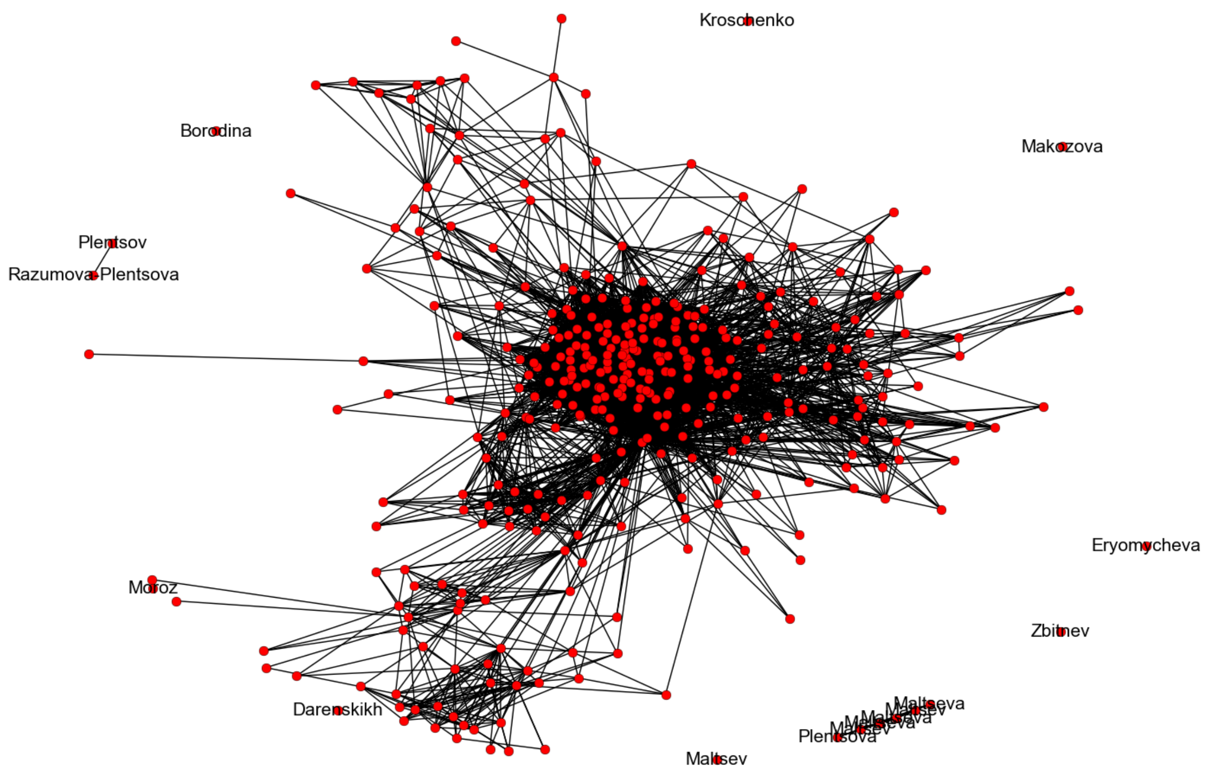
Network Summary

As we can see the network was not connected



Network Summary

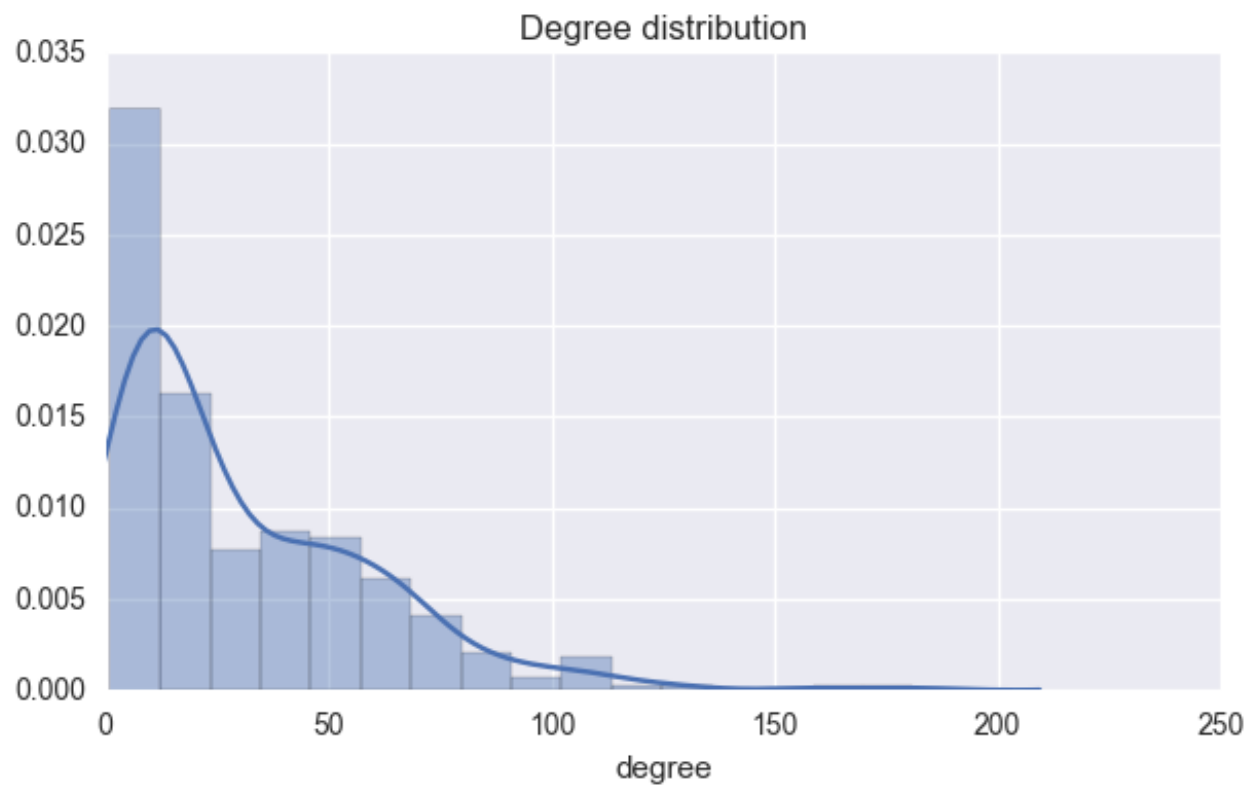
Small connected components contained relatives and childhood friends



In order to facilitate analysis, further research used only GCC

Network properties [Gigantic Connected Component]:

- **Number of nodes: 348**
- **Number of edges: 5435**
- **Average node degree: 31.2**
- **Diameter: 6**
- **Radius: 3**
- **Average path length: 2.4**
- **Clustering coefficient : 0.43**



Structural Analysis

Analysis of centralities highlighted most important nodes

- **Degree Centrality**

Roman Kharitonov
Natalya Kirsanova
Nikolay Tesla
Alexandra Kramchenkova
Darya Morozova

These are simply people who directly knows more people in my friends list.

- **Betweenness Centrality**

Roman Kharitonov
Alexandra Kramchenkova
Natalya Kirsanova
Nikolay Tesla
Vlad Kharitonov

Nodes with largest portion of shortest paths going through them. In the context of my friends list this means that if I'm not available, these people are the most probable to connect two unacquainted persons.

Analysis of centralities highlighted most important nodes

- **Closeness Centrality**

Roman Kharitonov
Natalya Kirsanova
Nikolay Tesla
Alexandra Kramchenkova
Kirill Ponomarev

They represent people who can reach other people in my network with minimum number of intermediates.

- **PageRank**

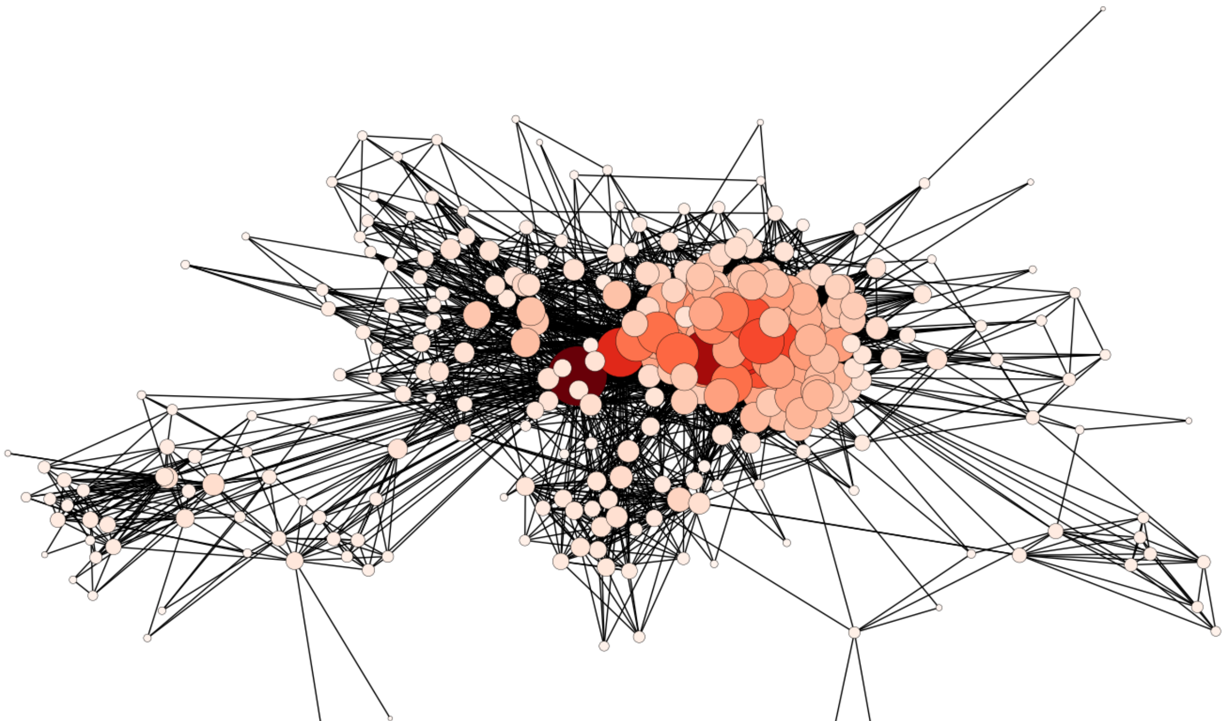
Roman Kharitonov
Alexandra Kramchenkova
Natalya Kirsanova
Nikolay Tesla
Vlad Kharitonov

Imagine that person starts with one of my friends at random and clicks on one of my common friends with this person, and then again and again. Sometimes he gets bored and starts with random person again. People with highest pagerank are those who are most likely to be met during such excursions.This top exactly coincides with top of betweenness centrality.

Structural Analysis

Analysis of centralities highlighted most important nodes

As all these metrics try to emphasize the importance of some nodes in the network, it is not surpising that there is a major overlap in top nodes.

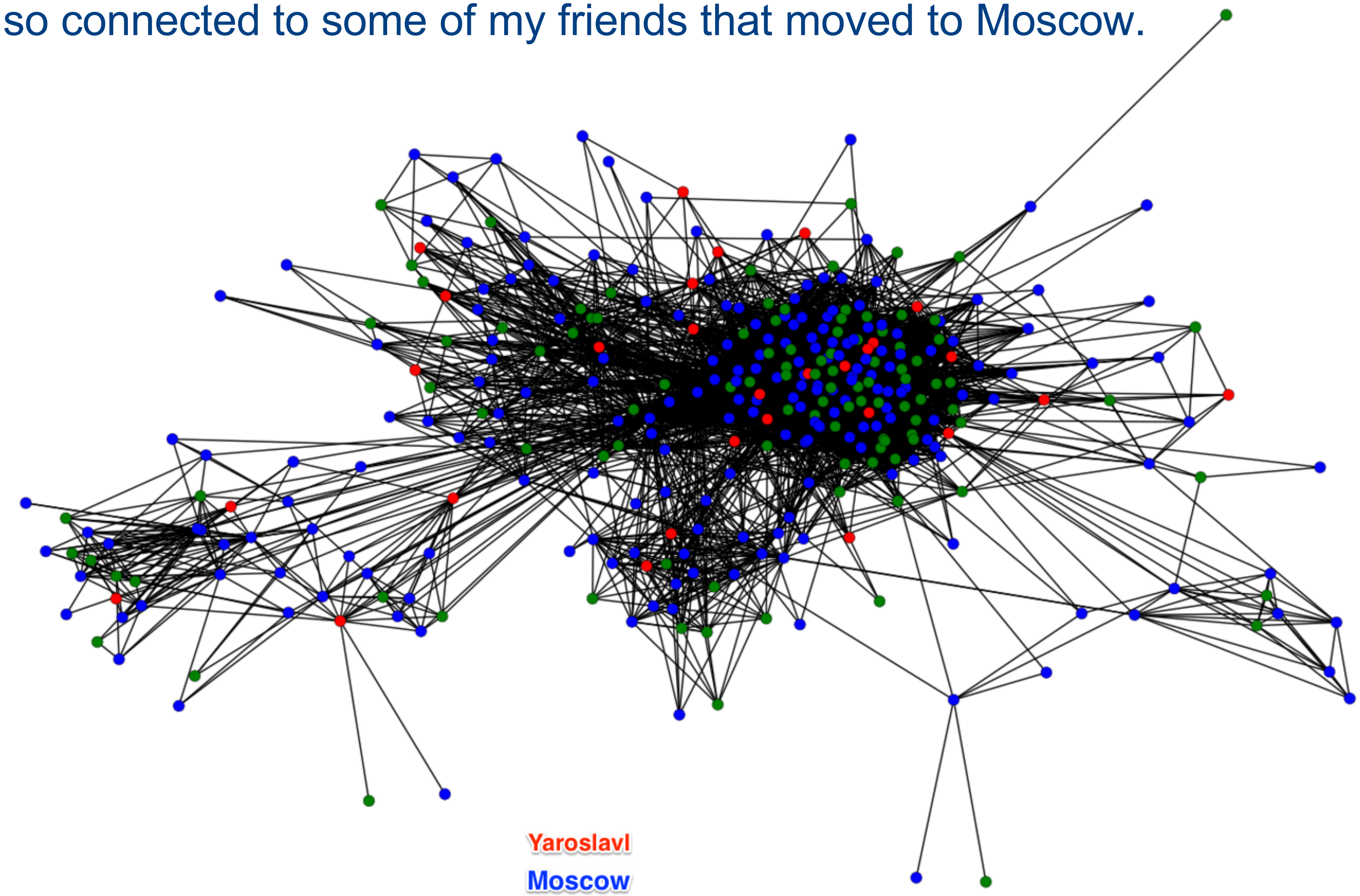


This visualization of the network emphasizes nodes with large degree centrality. It is obvious that variance in centralities is very high. It seems that there is some "core" of the network with large centrality values (closely connected friends, probably from the university) and a lot of nodes with centrality much smaller.

Structural Analysis

There is slight assortative mixing by city in the network

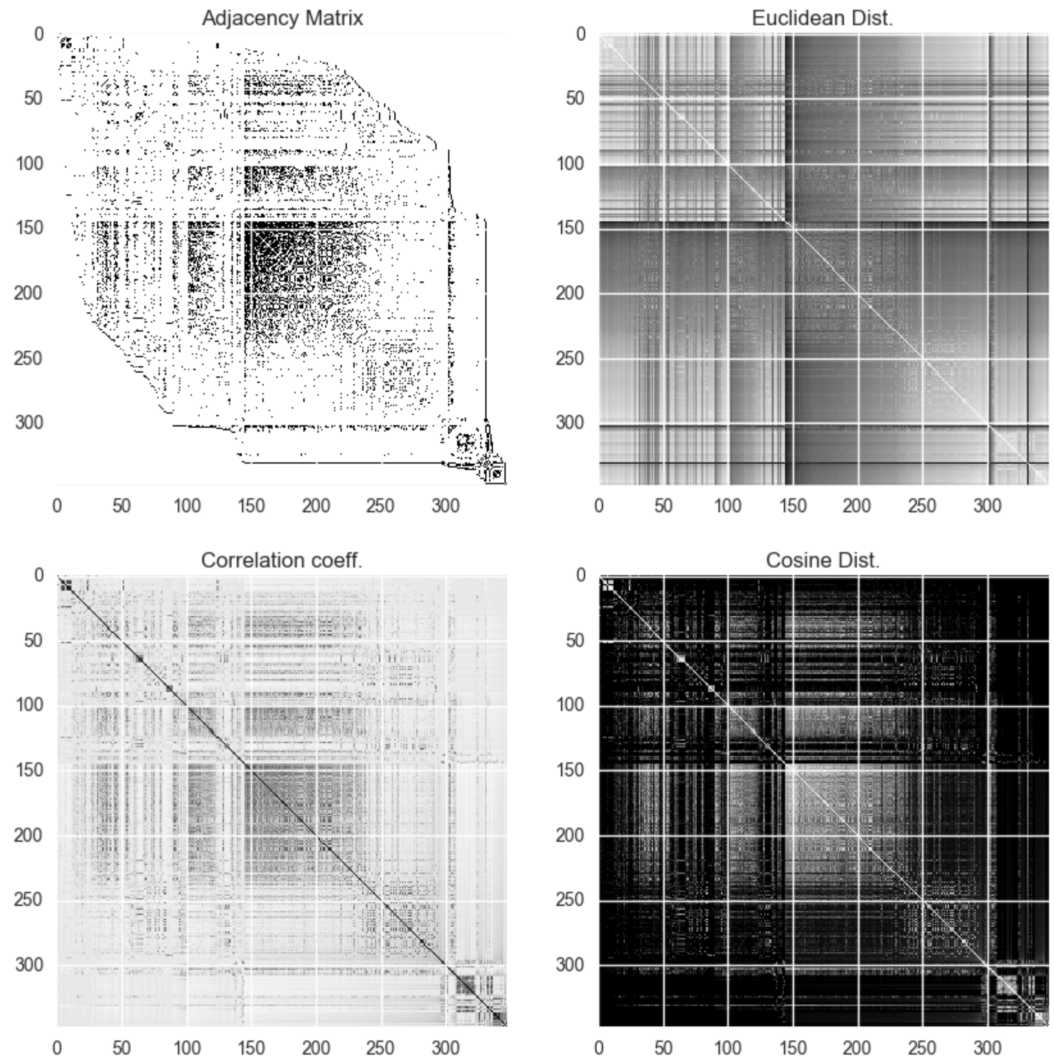
Assortativity coefficient over current city attribute is negligible: 0.07.
It could be explained by the fact that I have not so many VK friends in Yaroslavl and most of them are also connected to some of my friends that moved to Moscow.



Structural Analysis

Assortative mixing by degree is also not huge

Degree Assortativity coefficient = 0.08



Structural Analysis

Random graph modelling is not an option

- Next we try to model network using some random model
- Base models to consider:

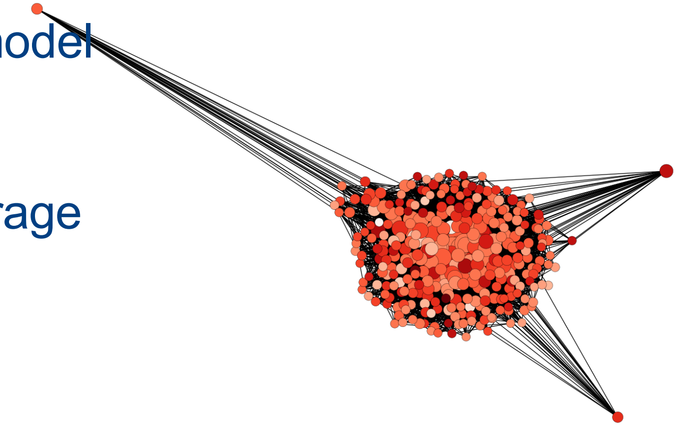
  - Random Graph

  - Barabasi-Albert preferential attachment growth model

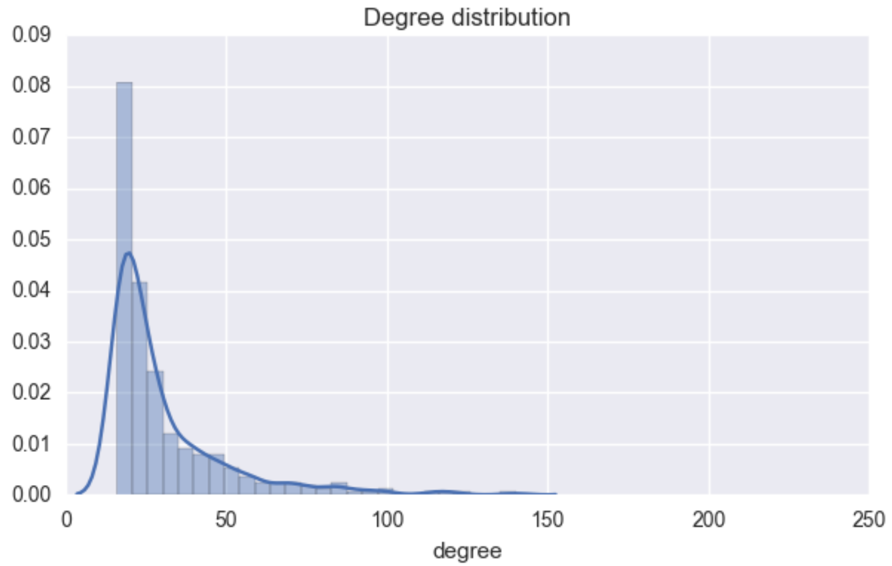
  - Watts-Strogatz model
- Random graph model was rejected immediately due to obvious existence of hubs in the network [max degree in the network is 181]

Barabasi-Albert model primarily fails with transitivity

- First model that was fitted is Barabasi-Albert dynamic model with preferential attachment



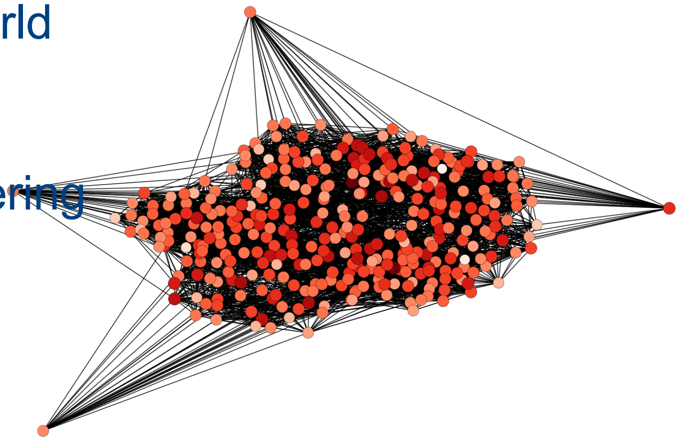
- Parameter *m* was optimized in order to have target average node degree. *m = 16 was chosen.*
- **Obtained properties:**
  - \# Nodes = 348
  - Diameter = 3
  - Radius = 2
  - Average path length = 2
  - Clustering coefficient = 0.16
  - Avg Node degree = 30.53

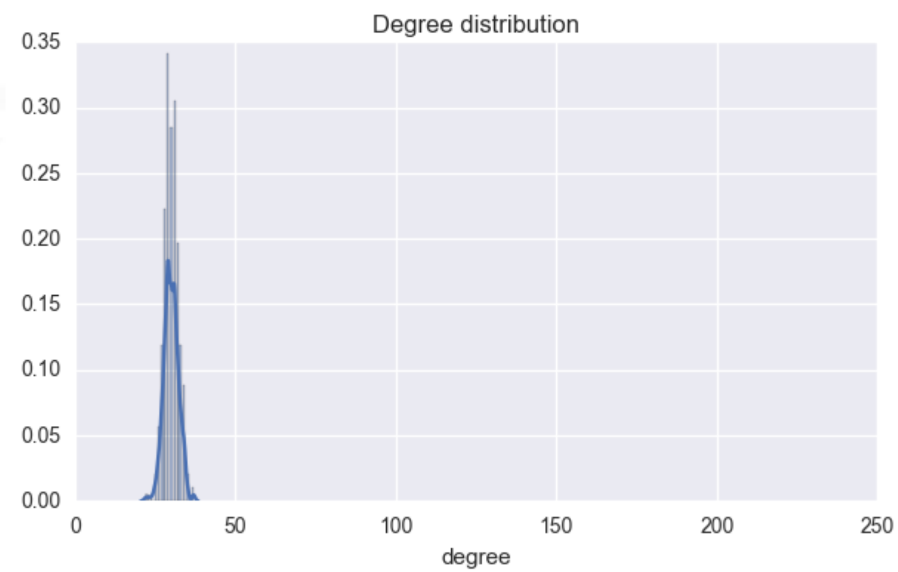


Structural Analysis

Watts-Strogatz model fails to have correct diameter and degree distribution

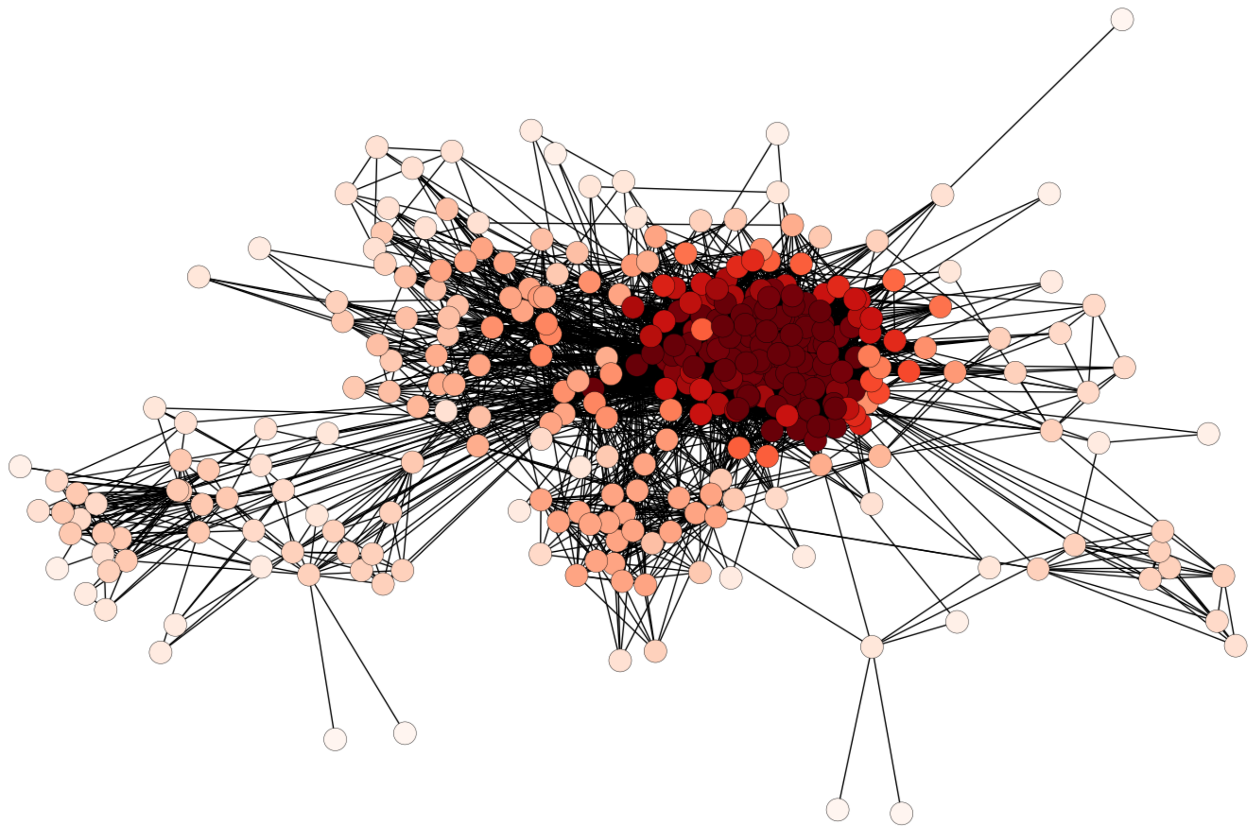
- Second model that was fitted is Watts-Strogatz small world model
- Parameter *p* was optimized in order to have target clustering coefficient. *p = 0.18 was chosen.*
- **Obtained properties:**
  - # Nodes = 348
  - Diameter = 3
  - Radius = 3
  - Average path length = 2.24
  - Clustering coefficient = 0.42
  - Avg Node degree = 30





Network was decomposed into its K-cores

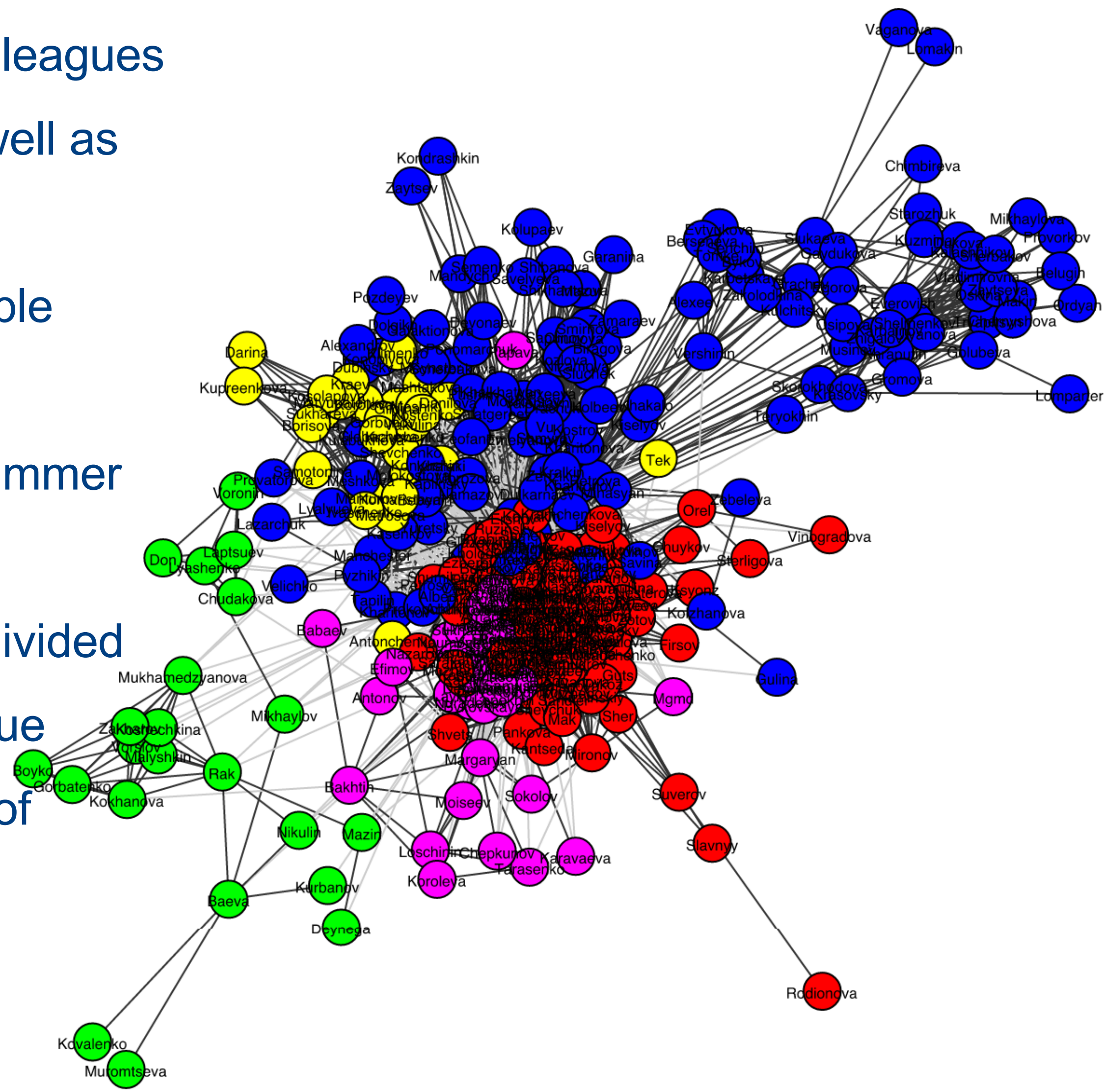
k-core is the largest subgraph such that each vertex is connected to at least k others in subset



Community Detection

Greedy modularity optimization have provided meaningful results

- Here **green** community includes my colleagues from both current and previous job as well as guys from Master's program.
- **Pink** community aggregates some people from HSE [my year and the next one]
- **Yellow** one includes guys from HSE summer camp for scholars.
- It seems like **red** and **blue** ones were divided arbitrarily. Moreover separate part of blue community to the right consists mostly of Yaroslavl guys.



Community Detection

Edge betweenness output is even greater interpretable

- Here **green** community includes my colleagues from current job
- **Grey** – from previous job
- **Dark green** – guys from audit school I've attended 3 years ago
- **Yellow** one includes guys from HSE Master's program.
- **Red** – most of my friends from economics faculty and/or dormitory
- **Blue** – Yaroslavl friends
- As can be seen there are several other small communities. It is explained by the choice of n = 10.