NETWORK ANALYSIS.

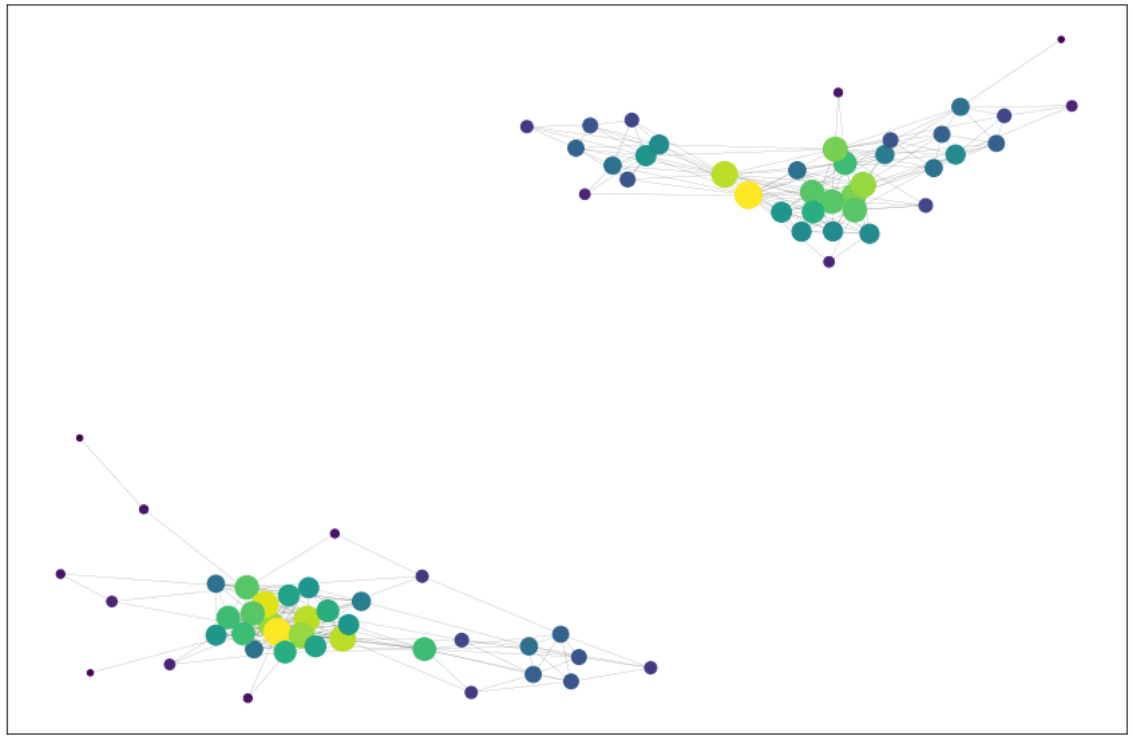
Anosov Roman.
Group : мНОД2020_ИССА

Information about data

- Data is extracted from vk using vk api. Using methods : users.get, friends.get;
- My profile is https://vk.com/anubo2;
- Receiving the information about my friends : first name, last name, sex, city, school, personal information, (attitude to alcohol, smoking; people main; life main) ,university.
  - attributes (alcohol, smoking, people main, life main) was dropped after analysis.

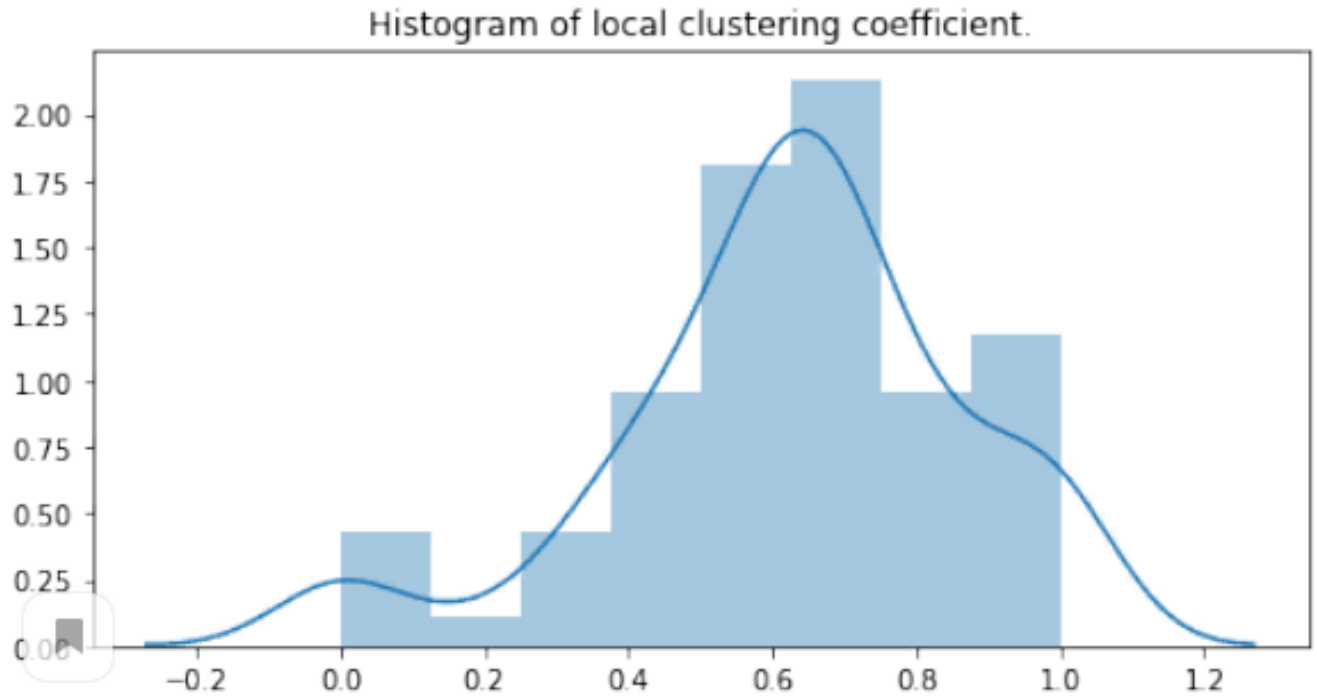
Ego network

- Size of graph 75
- Order of graph 349
- My ego network contain 2 connected components. First component is friends from Kashira. Second component is friends from bachelor (Moscow)
- Size component of friends from Moscow is 37.
- Size component of friends from Kashira is 38
- We will analyze the largest components (friends from Kashira).



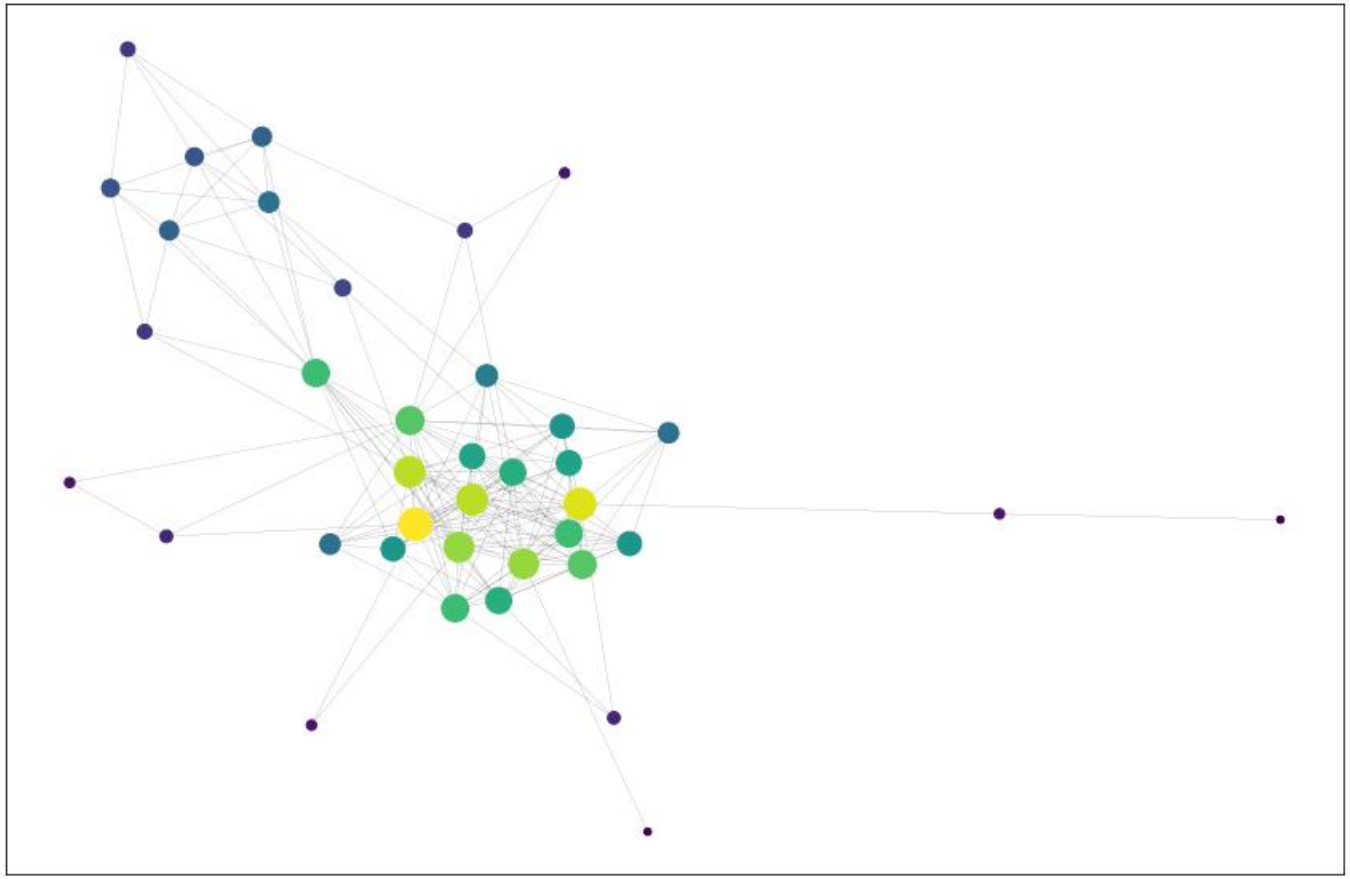
Clustering coefficient

Average clustering coefficient of my network is high. Also, a lot of my friends have high local clustering coefficient. So, my network contain some communities.



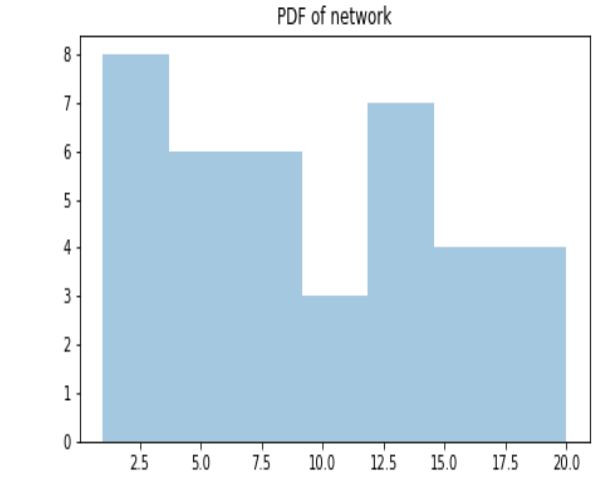
Largest componetnt

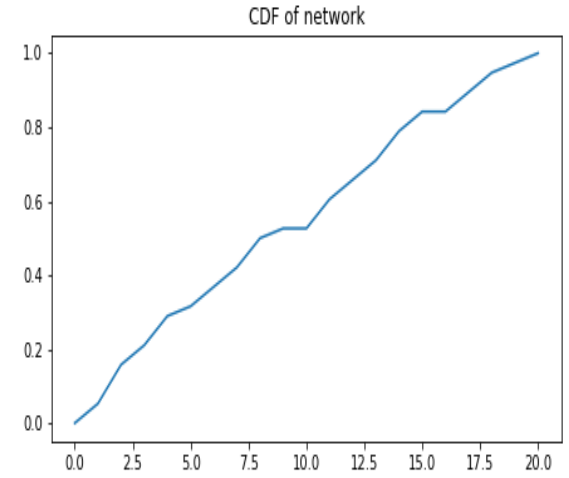
- Average shortest path length is 2.1038
- Radius is 3
- Diameter is 5

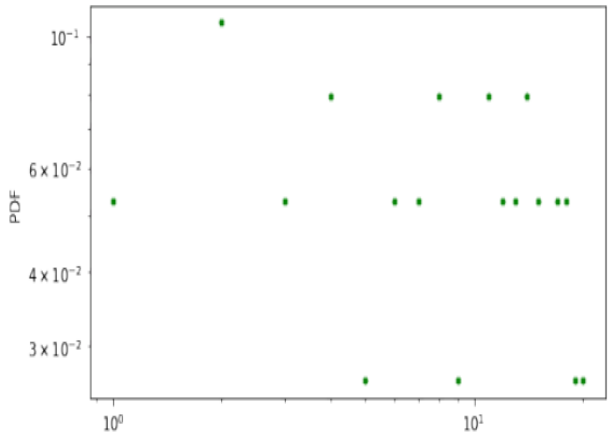


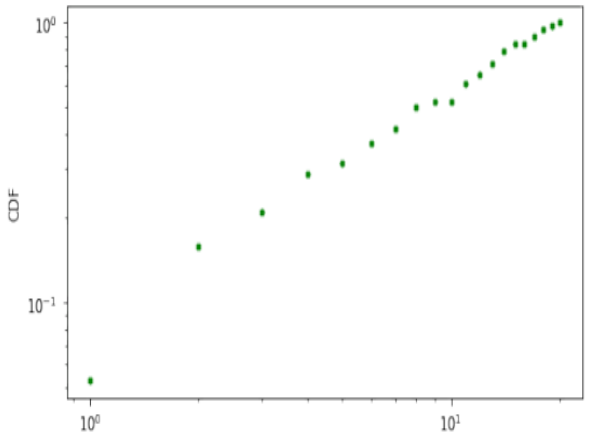
Probability degree distribution

The distribution of degree is like binomial.









Fitting models and coefficient from MLE (power law distribution)

The MLE consists of:

- Fix x_{min} as a minimal node degree (drop node degrees that less than x_{min})
- Calculate α via maximum likelihood estimation using fixed x_{min}
- $\alpha = 1 + n\left[\sum_i \frac{\log x_i}{x_{min}}\right]^{-1}$
-
- Calculate Kolmogorov-Smirnov test
- Fix x_{min} as the next node degree
- Repeat 2-4 by scanning all possible $x$min and find the best α and x_{min} with respect to Kolmogorov-Smirnov tes

In the result, best KSscore is 0.2024, best alpha is 4.7099, best x_{min} is 11.

Compare different model with our network

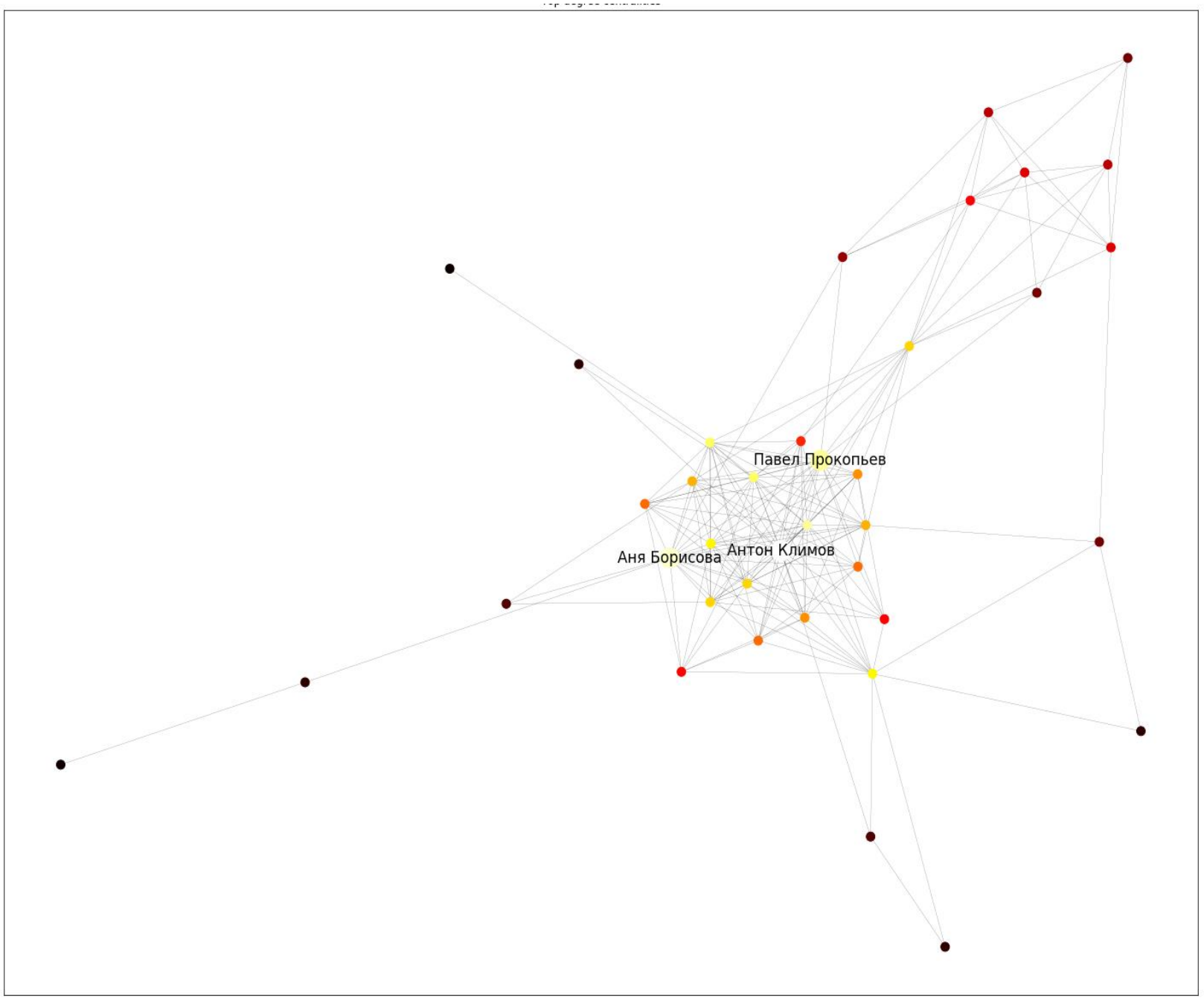
To do it, I generate 10 networks with estimated parameters to this model and aggregate some statistic (Average shortest path, average clustering coefficient, radius, diameter, kstest).

Clonest network model is Erdos-Renyi by Kstest. Barabasi-Albert network is close to my network by Average shortest path and Average clustering coefficient. Esimated probability for Erdos-Renyi model is 0.2532.

	Average shortest path	Average clustering coefficient	radius	diameter	ks_test
Barabasi-Albert	1.970555	0.326051	2.0	3.1	0.315789
Watts-Strogatz	1.925605	0.234315	2.4	3.1	0.392105
Erdos-Renyi	1.829018	0.249240	2.0	3.1	0.268421
Network	2.103841	0.604807	3.0	5.0	1.000000

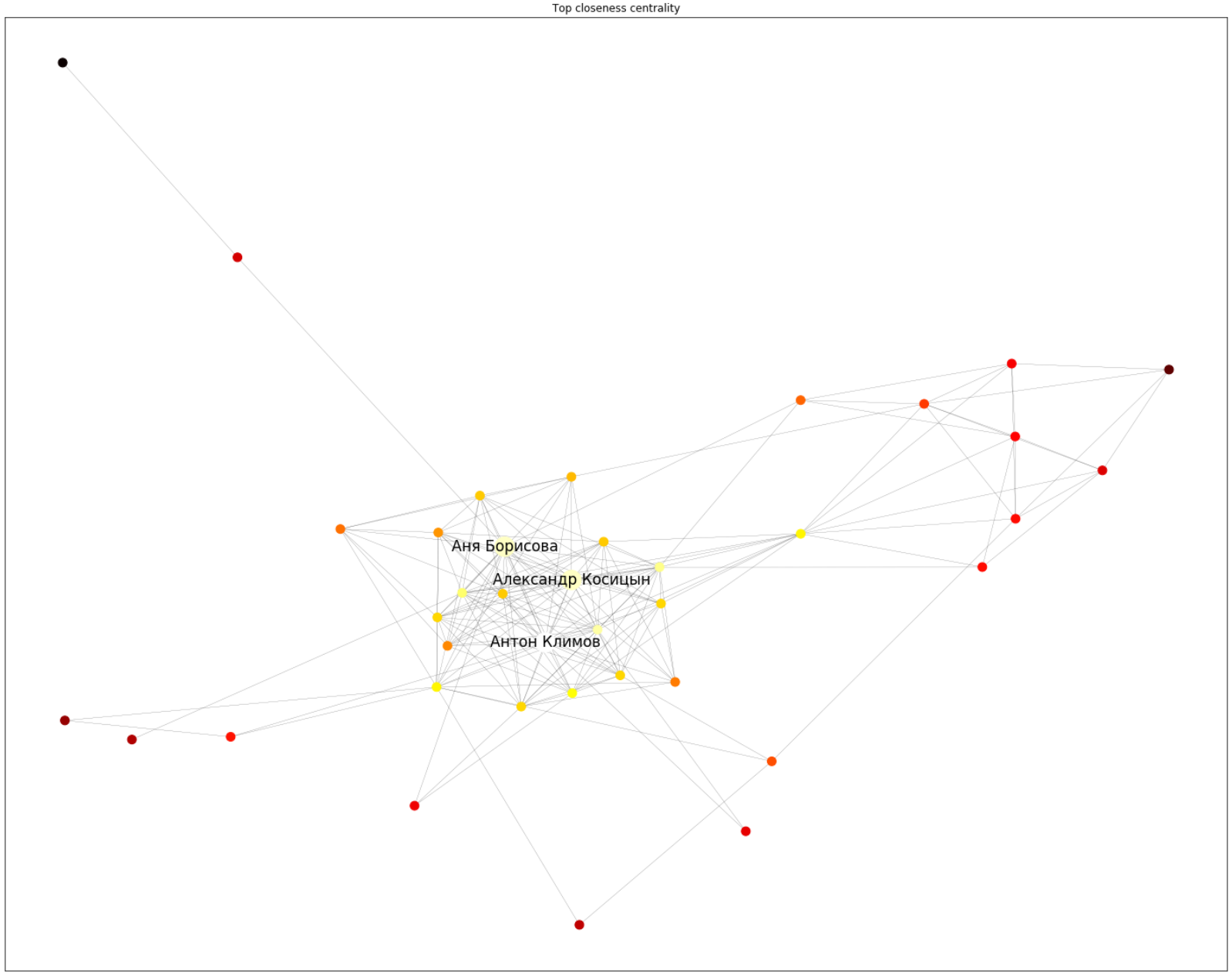
Degree centralities on the network

Аня Борисова, Павел Прокопьев, Антон Климов are my classmates from high schools. They went to middle school together. Moreover, they was popular in the schools and they were familiar with many people. Hince, they have high degree centrality.



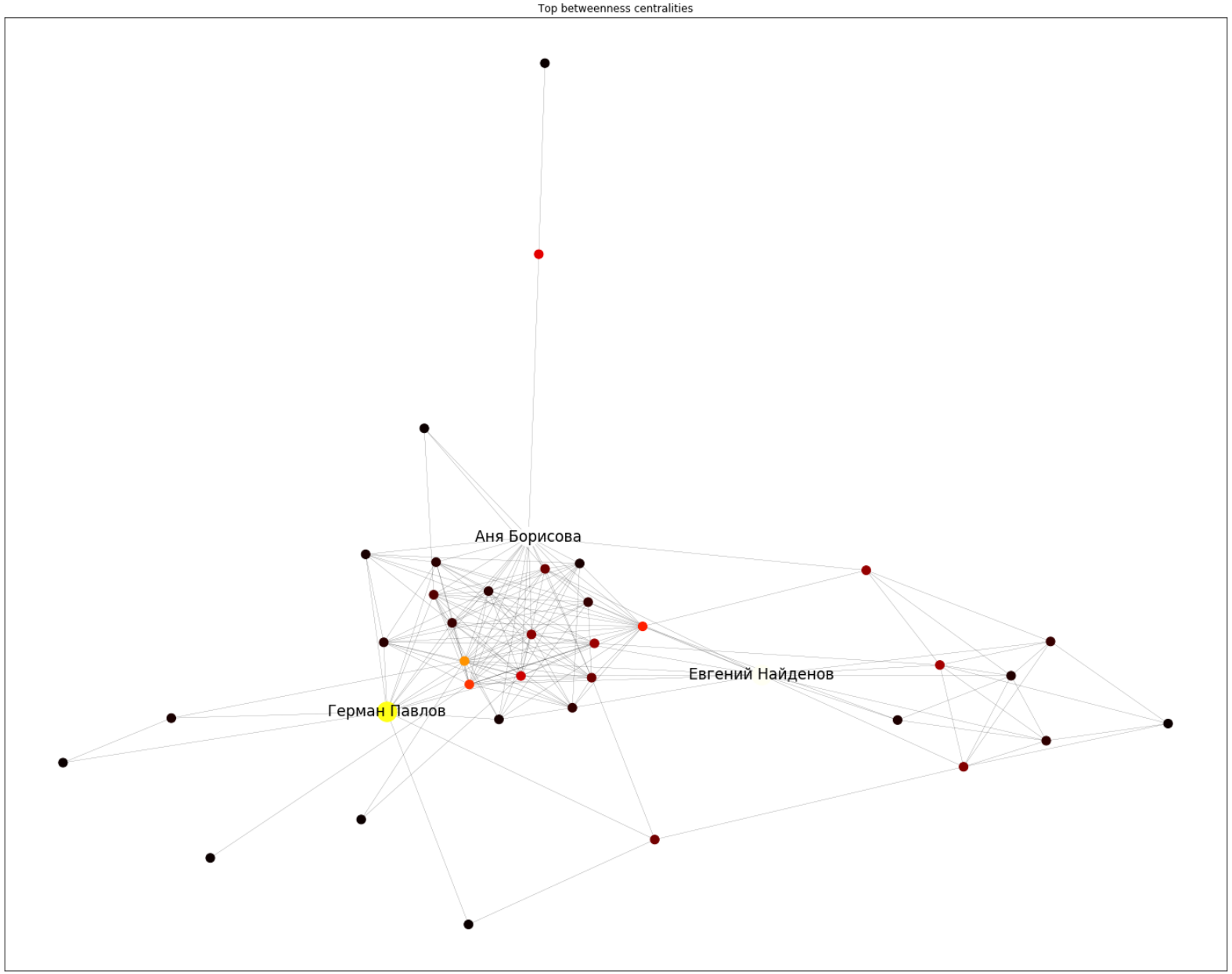
Closeness centrality on the network

Аня Борисова, Антон Климов, Александр Косицын went to middle school together. This friends was very popular in the schools. So, this friends are very close to my different friends from my schools. Morover, Антон Климов, Аня Борисова introduced me to people from another school. Hence, they have high closeness centrality.



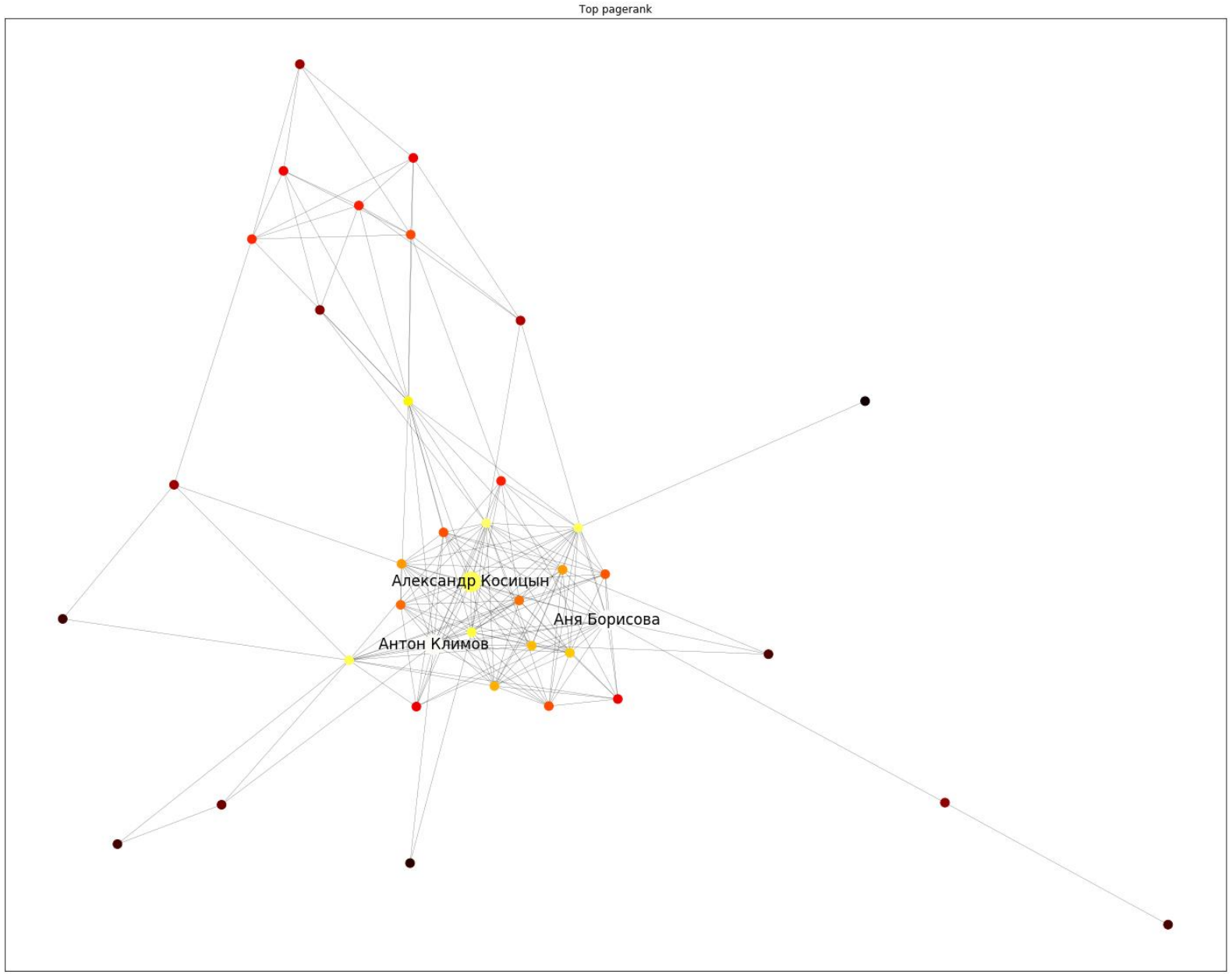
betweennes centrality on the network

Аня Борисова is very popular person in my schools. Moreover, she know my friends from another school. So, she know people from various communities. Евгений Найденов is my best friend from middle school and they know all my classmates from middle schools. Герман Павлов is my cousin. After ending middle schools, he went to another city (Stupino). After that, I met some of his new friends from Stupino and add they to vk friends. So, they have high betweenness centralities score.



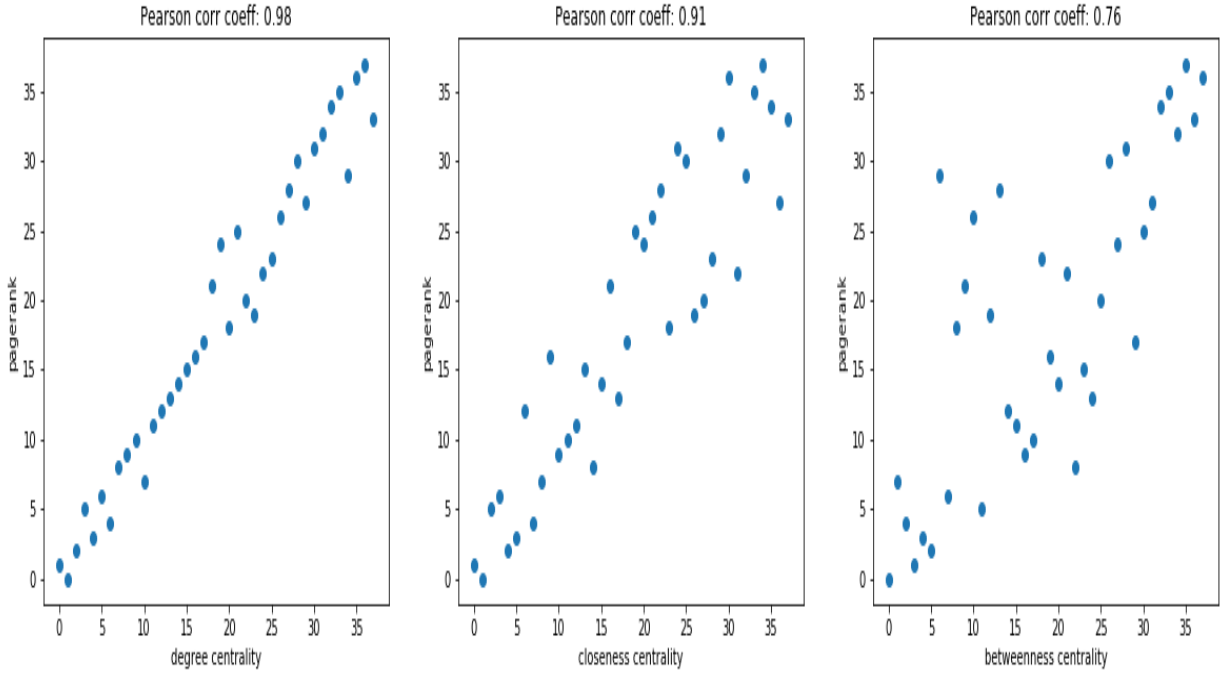
Pagerank on the network

Аня Борисова, Антон Климов, Павел Прокопьев was very popular in the schools. So, probability of stopping on this page after random walking on ego network is high.



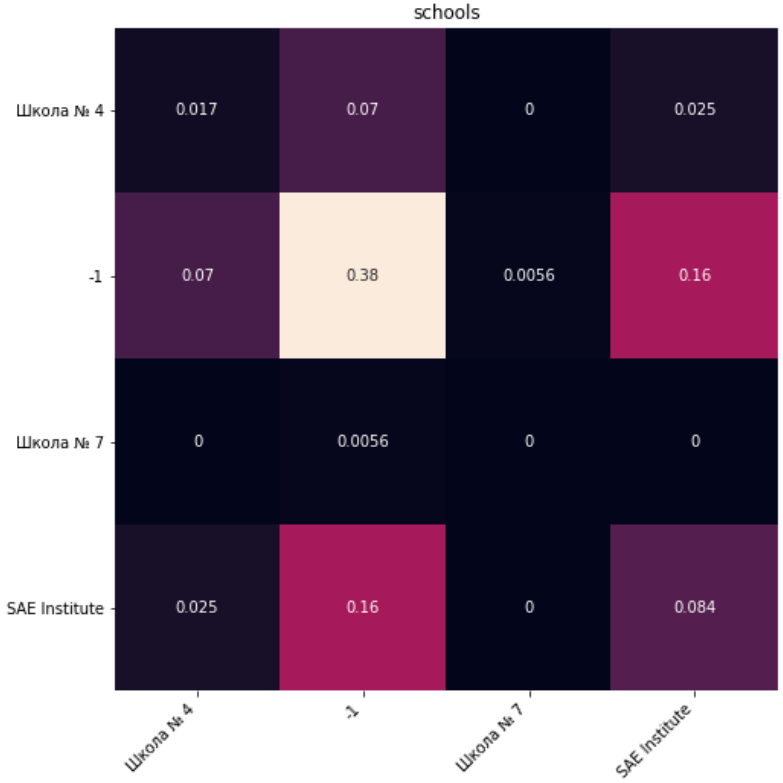
Compare ranking

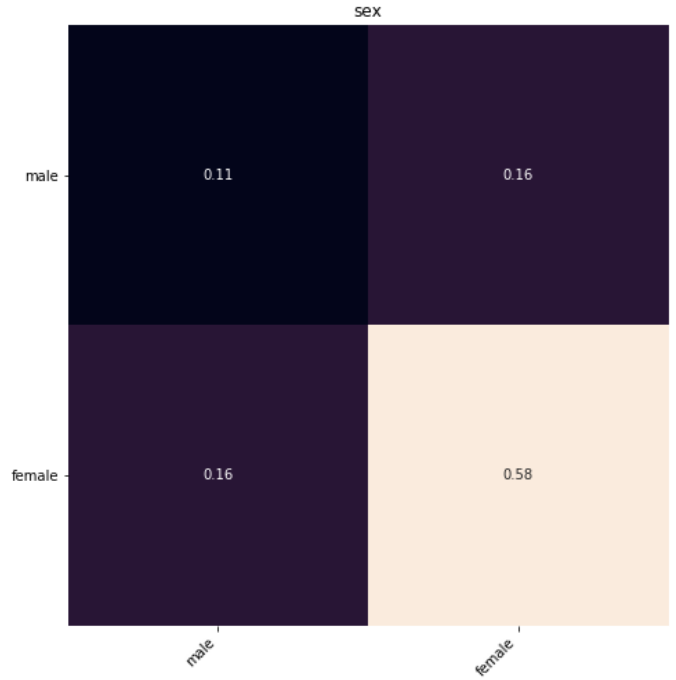
Ranking with pagerank correlates with ranking with different centralities (Pearson correlation coefficient is very high).

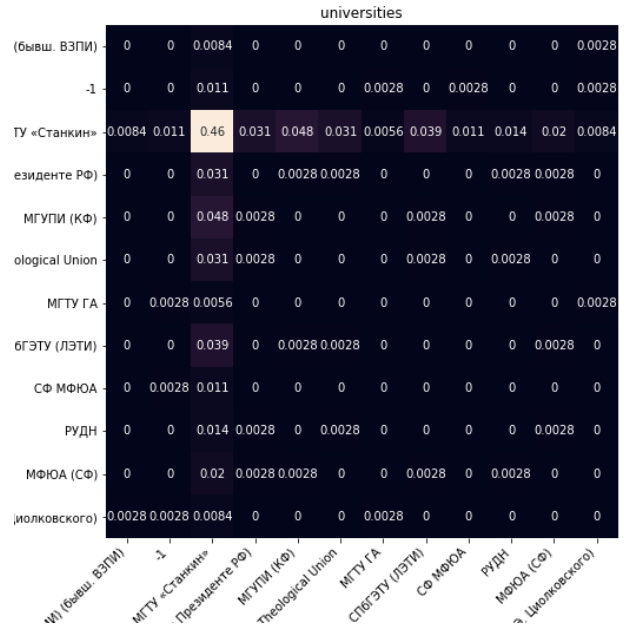


Assortative Mixing

Female friends are more friends with each other than male, according assortative mixing by sex attribute. A lot of friends from school study at МГТУ 'Станкин'. We can also see, that there are high assortative score between SAE institute. It is because one friend from schools №7 filled incorrect information about schools. Also, there are assortative with SAE institute and None filled information. I think, it is because people who don't want to fill information about schools like to put a random popular school in this field.

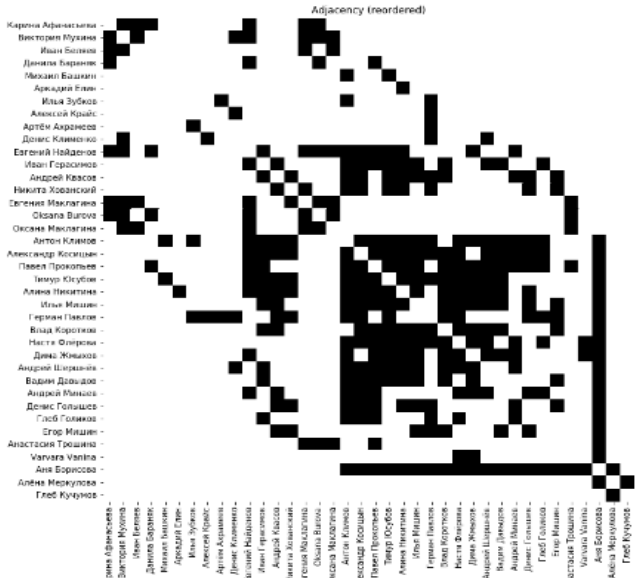


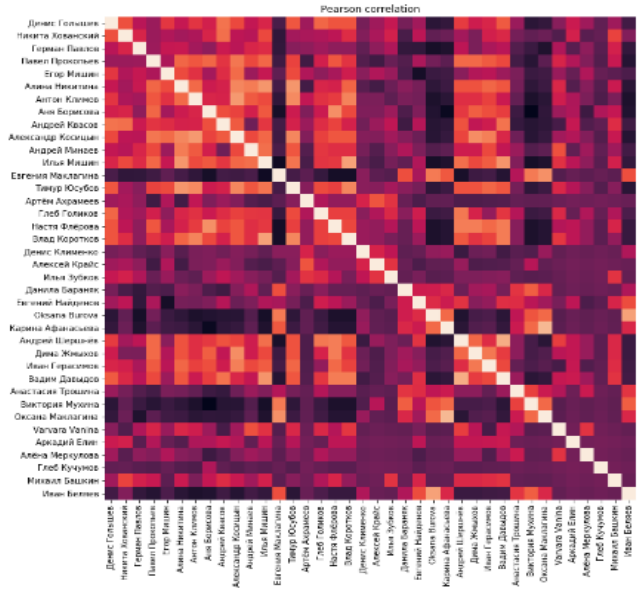


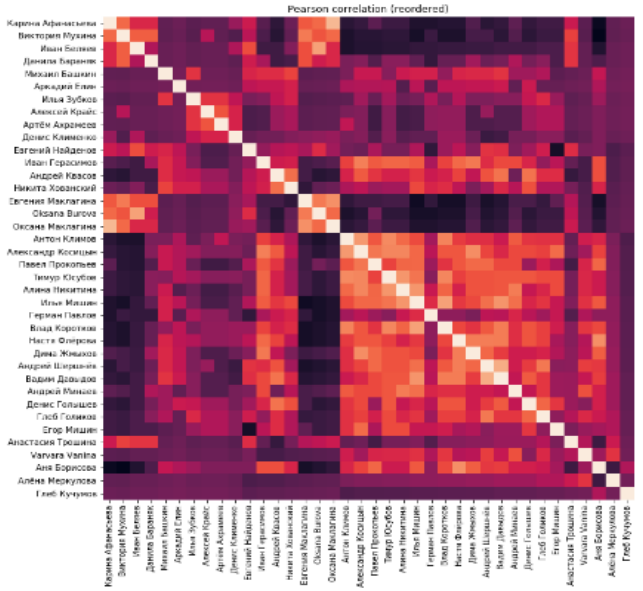


Assortative Mixing

Using information about similarity, we can detect communities with different size. There are big communities, this communities of friends from schools. This friends connect which other. Moreover, this communities can be dividing by some parts. Also, i can see some little communities. It is community from middle school. This friends have a small connection with big communities, because they prefer to be friends in their own small community and have a small number of connection with friends from big communities. Also, there are communities of friends of my brother from another city.

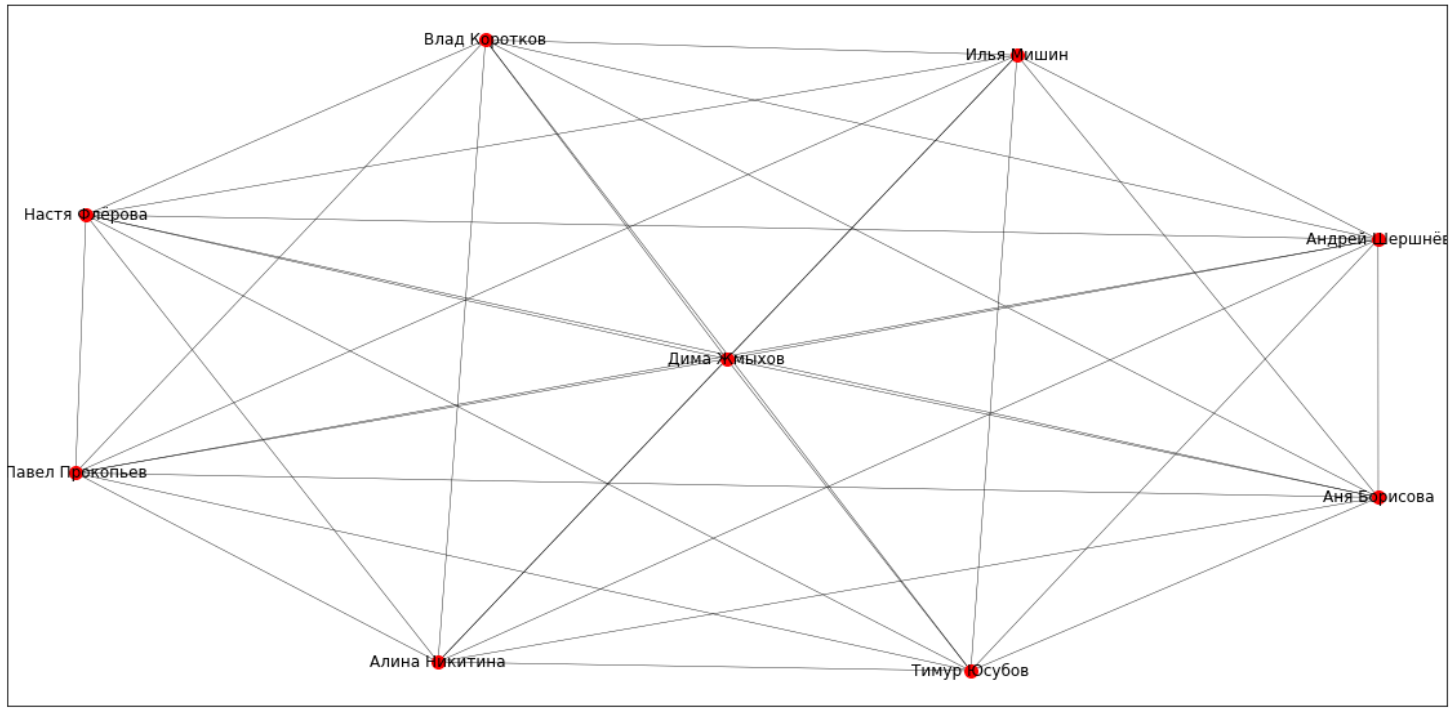






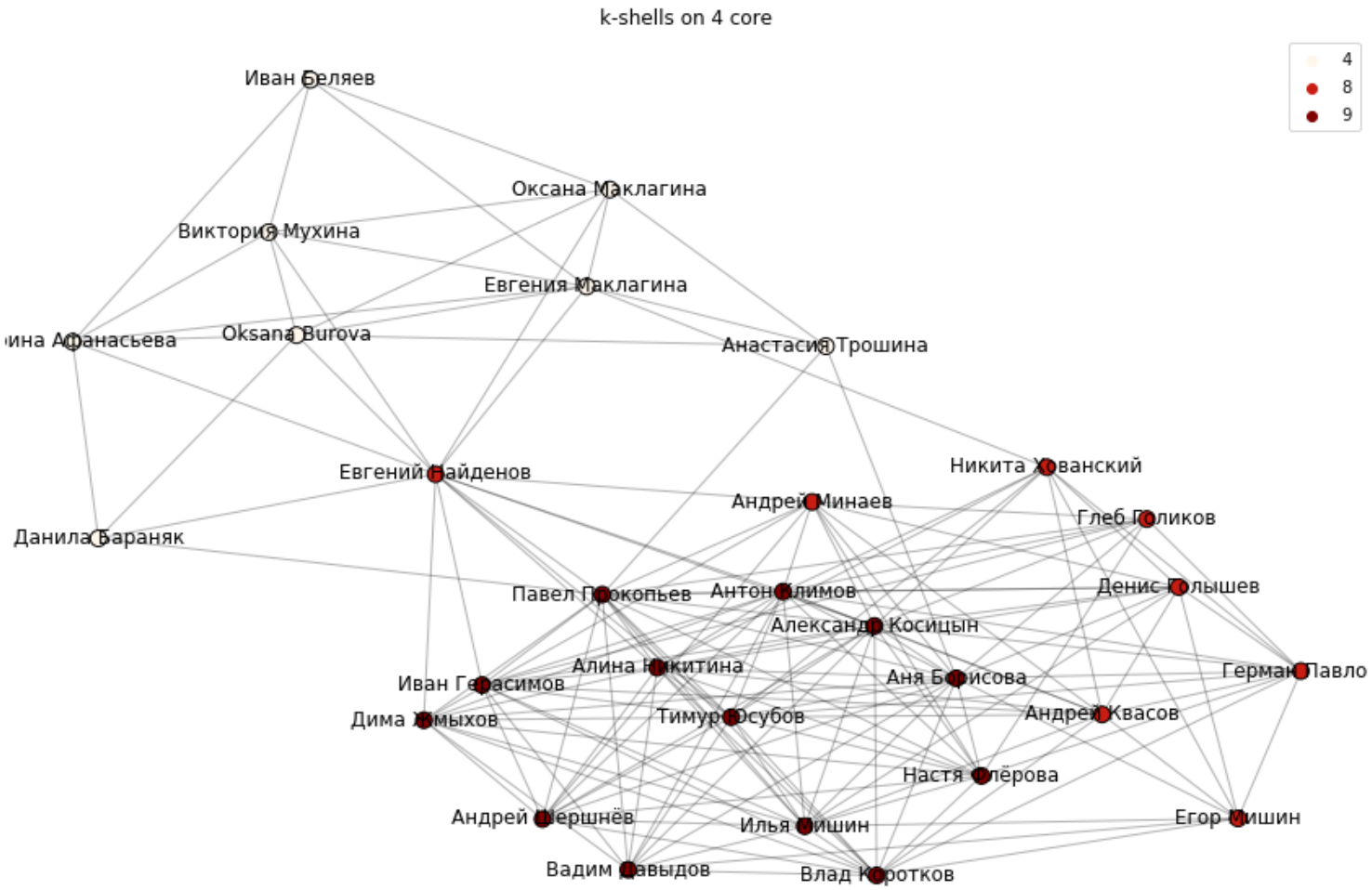
Maximal clique.

This clique is friends from my schools. Moreover, this clique contain friends from high schools and friends from parallel class. We was very friendship.



4 core subgraph

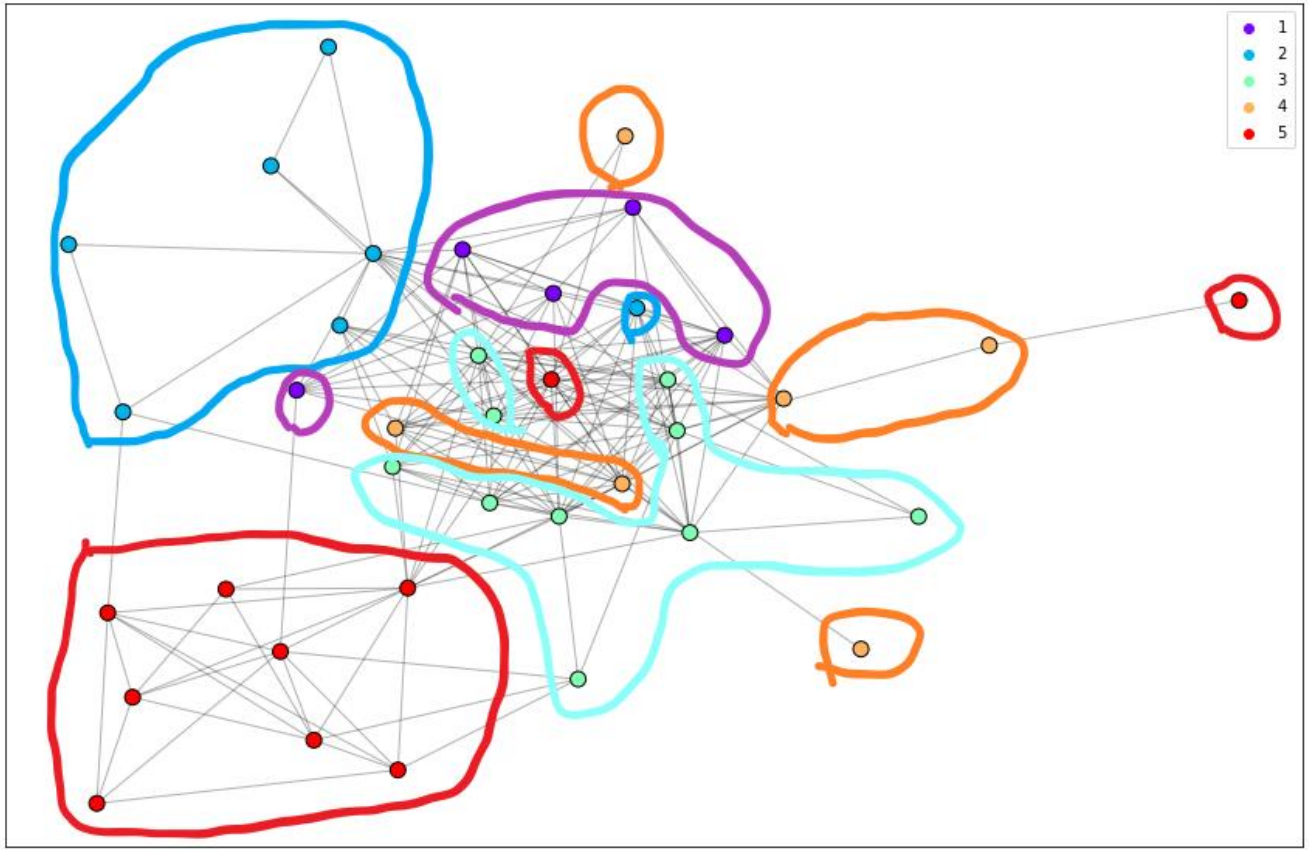
According k-shells on 4 core we can see 2 communities. It is friends from middle school (4 shell) and friends from high schools with another friends from schools (8,9 schools). My best friend from middle school (Евгений Найденов) have a lot some connections with friends from high schools, so he have 8 shell.



Assumption about communities

To detect communities on the ego network i will use Laplacian Eigenmaps, Agglomerative clustering, async update labels. Since, this components contain a few number of nodes (38 friends), i can detect communities from my assumptions and use it to calculate ground truth score.

From my assumptions i detect next communities:
1.Label 1. Friends from my school.
2.Label 2. Friends of my cousin.
3.Label 3. Classmates from high school.
4.Label 4. Friends from another schools.
5.Label 5. Classmates from middle school.



Best result with different clustering approach

First plot is communities obtained from the Laplacian eigenmaps.
Second plot is communities obtained from the algorithm clustering.
Third plot is ccommunities obtained from the async update labels.
Parameters for this model was selected by the best ground truth score.

Best result with different clustering approach

The best model by ground truth and modularity is Laplacian eigenmaps.
It is because, this model can divide community from my school's friend, but another model can not do it.
The best model by Silhouette score is async update labels. async update labels model detect communities which close to communities from 4 core subgraph.

Model	**Modularity**	**Silhouette score**	**Ground truth**
Laplacian eigenmaps	0,2286	0,0626	0,4336
algorithm clustering	0,2214	0,3329	0,2290
async update labels	0,1506	0,3745	0,0414