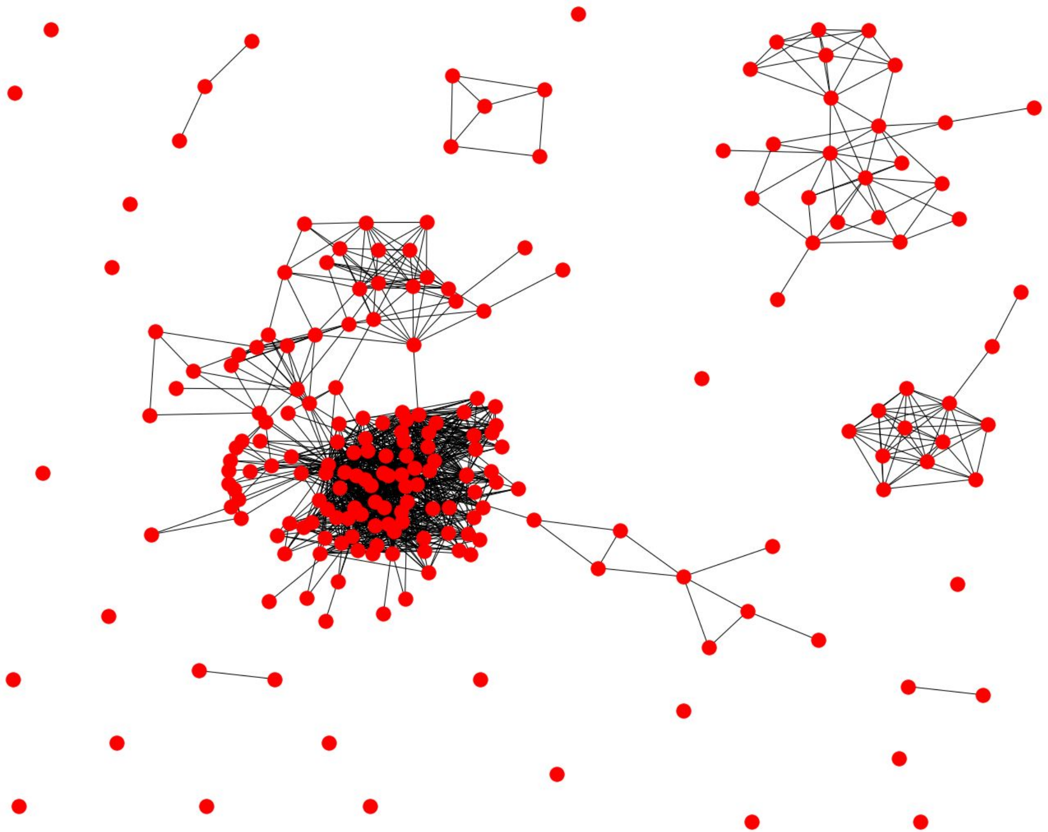
Project 1: Network analysis of VK friends

by Nikolaeva Anna, 10.2019

1. Network summary



- Network source page - https://vk.com/annnyway

  The graph was extracted using http://vk.com/app3861133

  Nodes are persons, edges are friendship connections.

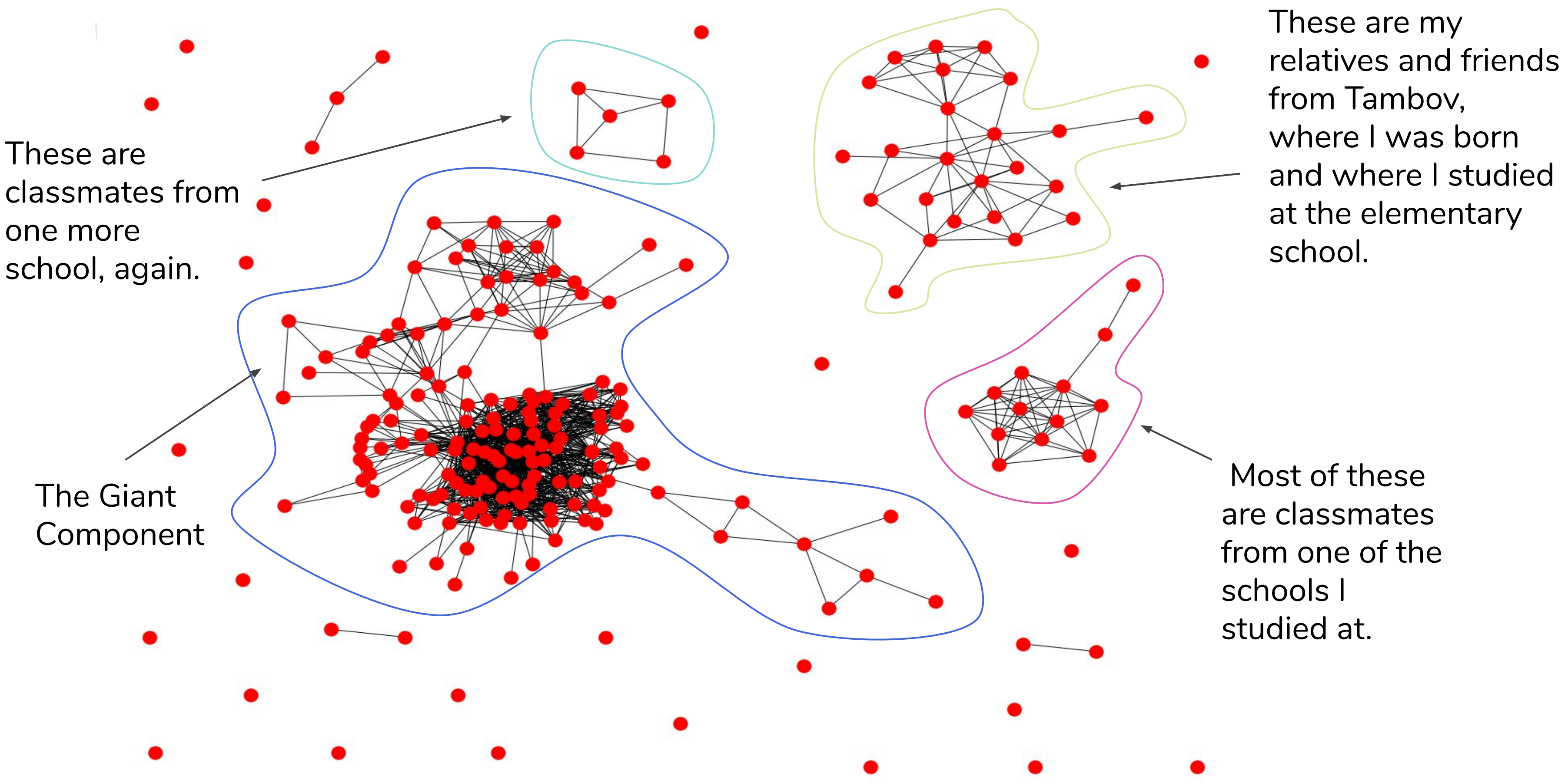
  Total number of nodes - 222

  Total number of edges - 1333

  This is an original graph with 7 subgraphs and standalone nodes (the friends which connected only to me).

  During preprocessing I delete these nodes and then focus on analyzing the biggest subgraph - the Giant component of the graph.

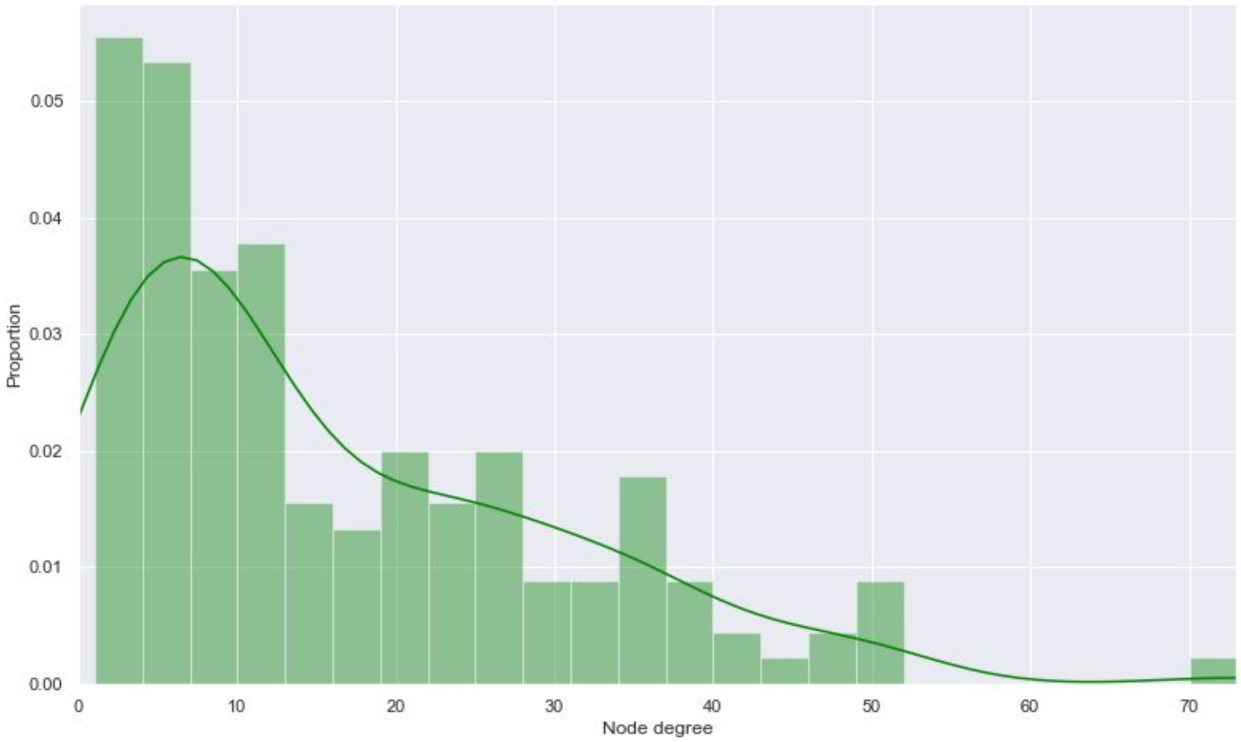
1. Network summary



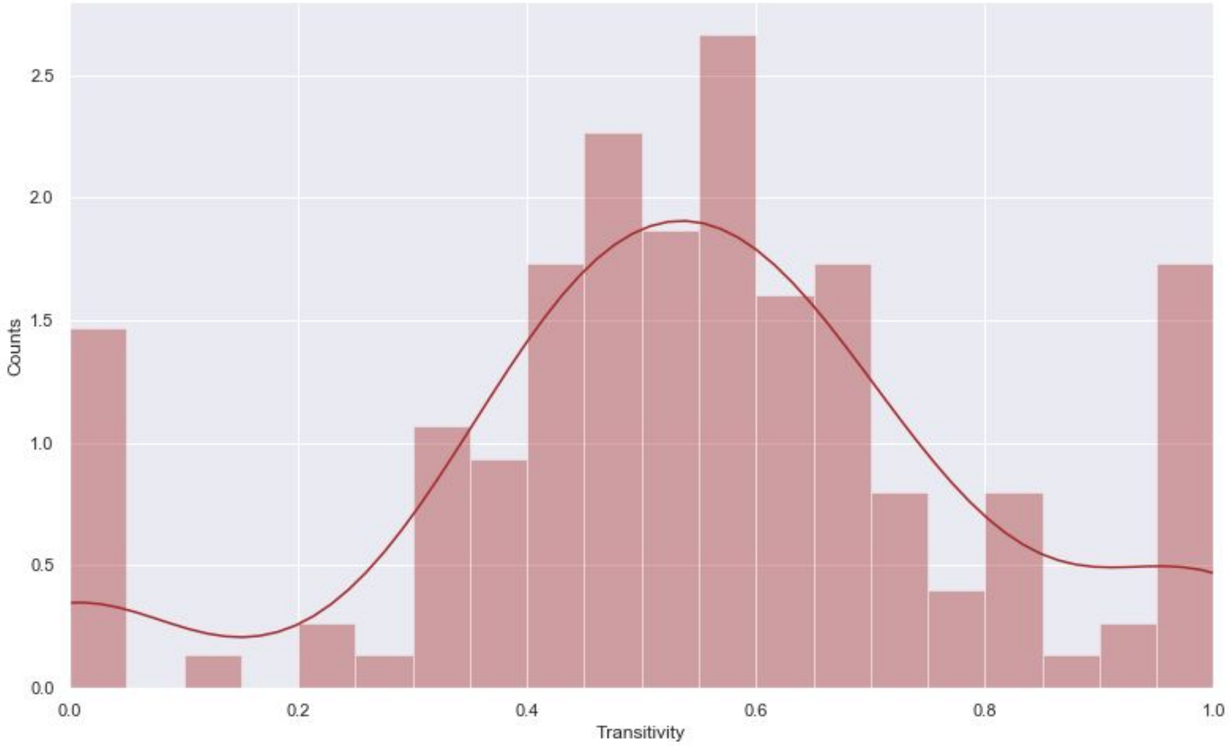
1. Network summary. The Giant Component

- number of nodes - 150
- number of edges - 1217
- average node degree - 16.23
- median node degree - 11
- diameter - 9
- radius - 5
- average shortest path length - 2.88
- average clustering coefficient - 0.55
- global clustering coefficient - 0.48

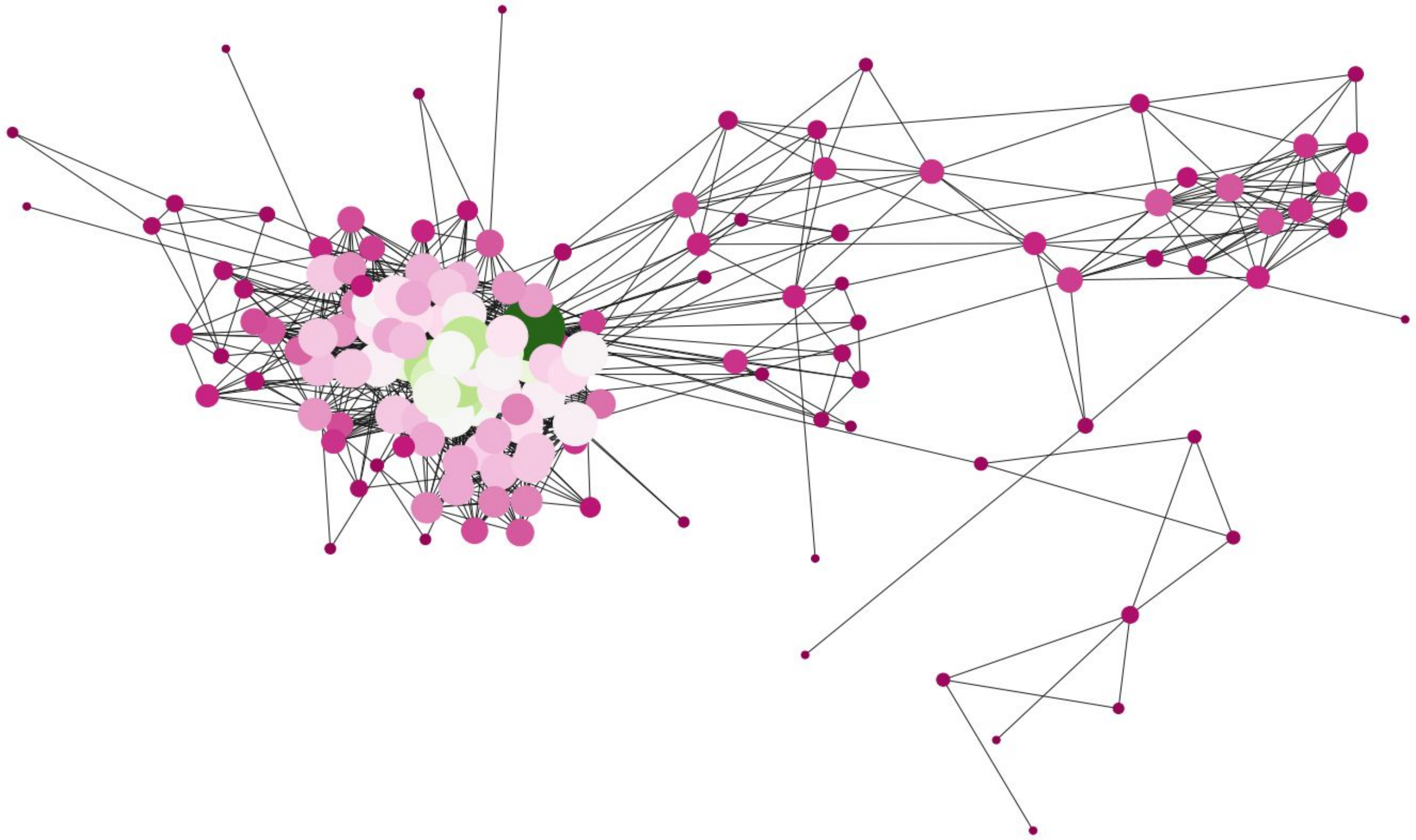
1. Network summary. Degree distribution



1. Network summary. Transitivity



The Giant Component



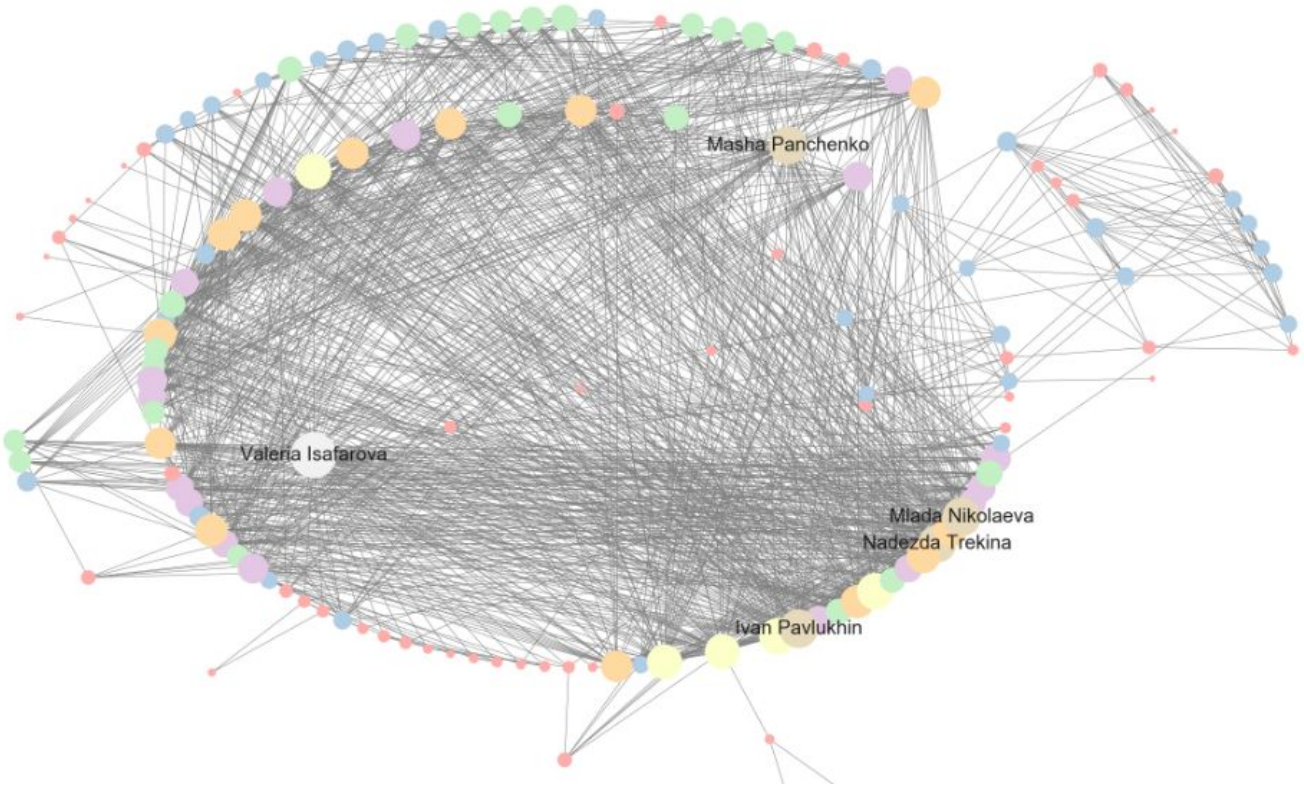
2. Structural analysis. Centrality (top-5)

	Degree centrality	Closeness centrality	Betweenness centrality	Eigenvector centrality
1	Valeria Isafarova	Valeria Isafarova	Valeria Isafarova	Valeria Isafarova
2	Ivan Pavlukhin*	Ivan Pavlukhin*	Aleksandra Voropaeva	Mlada Nikolaeva
3	Mlada Nikolaeva*	Mlada Nikolaeva*	Lil Cinderella	Masha Panchenko
4	Masha Panchenko	Aleksandra Voropaeva**	Maria Prokhorova	Angelina Kovalenko
5	Nadezda Trekina	Masha Panchenko**	Nadya Katricheva	Ivan Pavlukhin

* the stars near names mean that these people share the same centrality value within the column

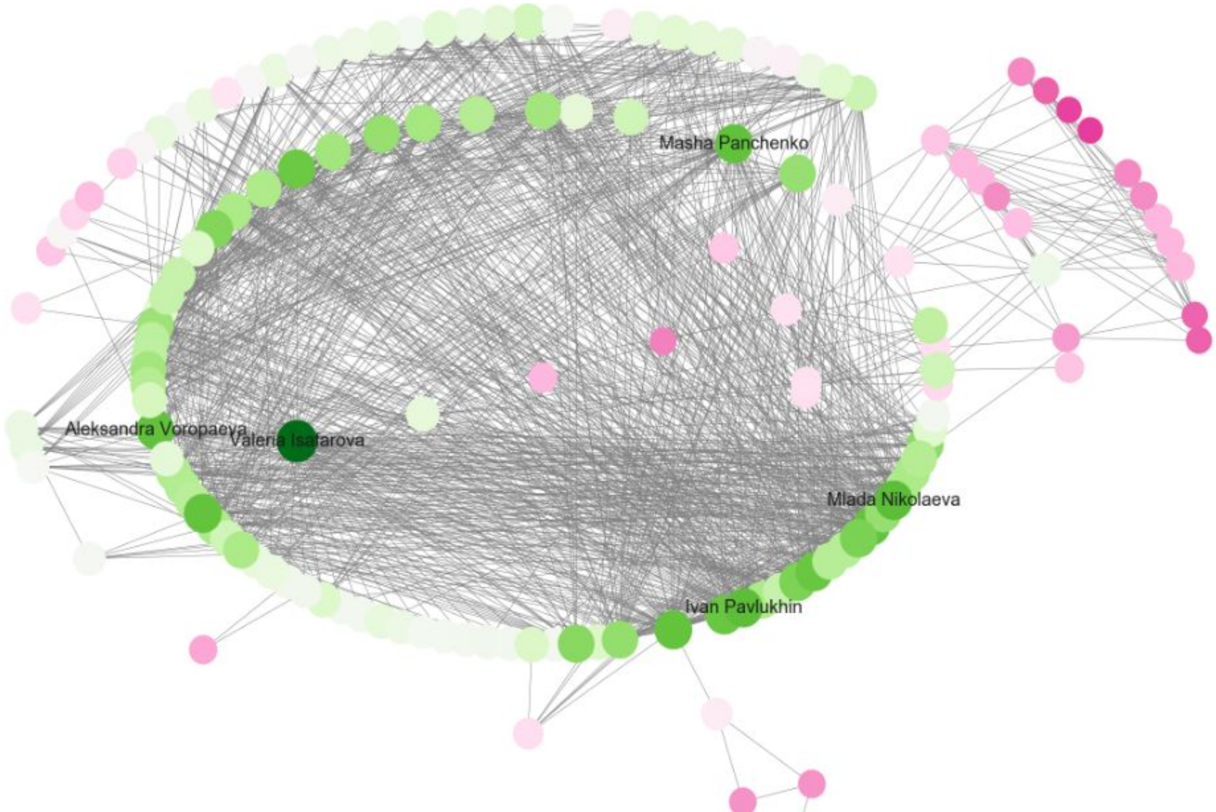
2. Structural analysis. Degree centrality

These people are my best friend (Valeria Isafarova) and university groupmates with lots of connections. All of them are from MGPU (the Moscow City Teacher’s Training University). I used Vkontakte very often when I studied there, so degree centrality reflects the pattern of my connections in that period of time.



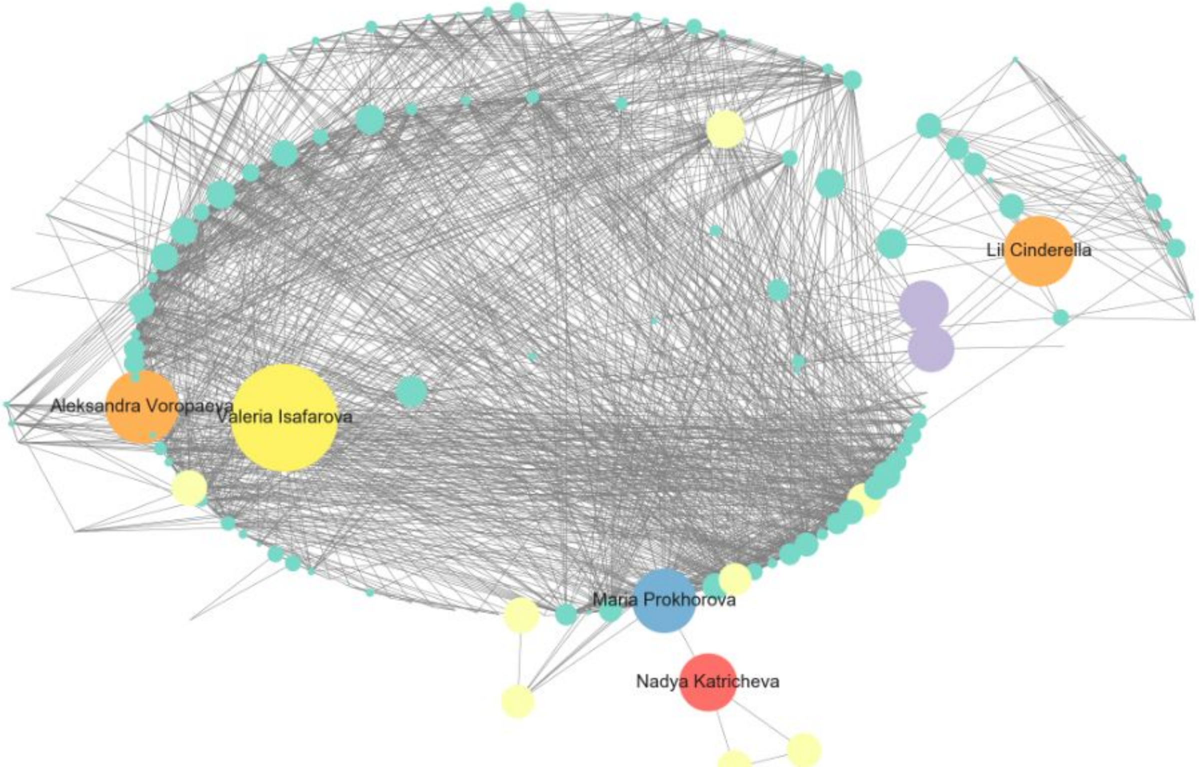
2. Structural analysis. Closeness centrality

These people are again from the same university, but in top-5 there is Alexandra Voropaeva, who is not my groupmate and who I do not know well. She is just a popular person, and she is rather close to other nodes.



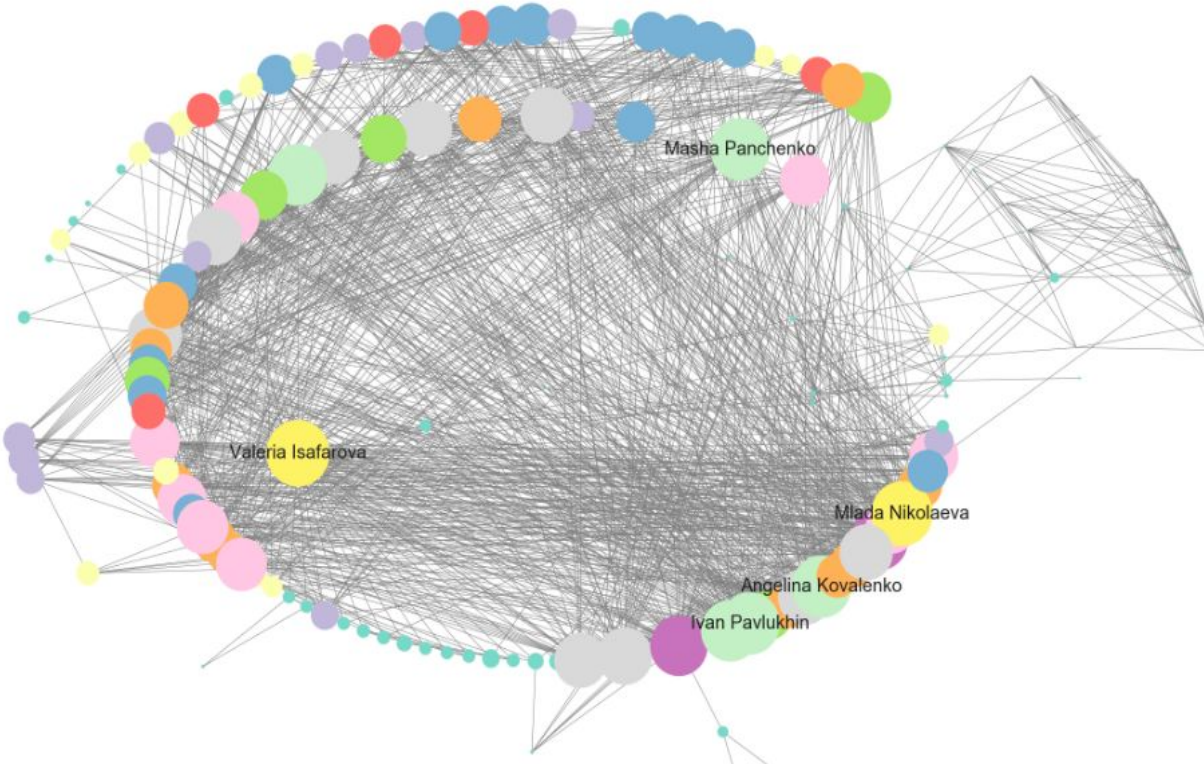
2. Structural analysis. Betweenness centrality

People with the highest betweenness centrality can be called “bridges” through which I can reach most of the friends in my network. For example, Nadya Katricheva is my friend from HSE, but she studied with Maria Prokhorova at the same school, and I know Maria from another university. So she is a “bridge” between different groups of friends.



2. Structural analysis. Eigenvector centrality

People with highest eigenvector centrality seem to be central in their clusters.



2. Structural analysis. PageRank

- Valeria Isafarova
- Masha Panchenko
- Ivan Pavlukhin
- Mlada Nikolaeva
- Nadezda Trekina

No wonder, that the top-5 from PageRank overlap with degree centrality's top-5, but have a slightly different order. The "importance" of a node in PageRank reflects how often one can "visit" the node (e.g. open this person's page) making a lot of steps on the graph. The higher the degree of the node, the more often one will visit it.

2. Structural analysis. Assortative mixing


```
{'first_name': 'Sergey',
 'last_name': 'Portnoy',
 'nickname': 'セル',
 'screen_name': 'homelleon',
 'sex': 2,
 'photo_50': 'https://sun1-30.userapi.com/c846322/v846322289/1ded44/0pZ7O8FJE2s.jpg?ava=1',
 'relation': 4,
 'country': '[object Object]',
 'city': '[object Object]',
 'bdate': '2.11.1989',
 'label': 'Sergey Portnoy'}
```


Assortative mixing deals with node attributes and measures the similarity of connections in the graph with respect to the given attribute. Among all the node attributes we can use only two of them to some extent: sex and date of birth.

The assortativity coefficient with respect to sex is rather small: **0.08**. So we cannot conclude that male tend to connect to male and female - to female.

2. Structural analysis. Assortative mixing

I have also tried to get the assortativity coefficient with respect to the year period people were born. I preprocessed the node attribute of birth date in order to associate every person with a decade: if one was born in 1993, his decade is 1990's, and for those who was born in 1996 the decade is 2000's.

The assortativity coefficient with respect to such decades turned out to be **-0.03**. The negative value indicate relationships between nodes of different degree, but it is not reliable as not so many people in my network mentioned their year of birth.

2. Structural analysis.
Comparison to random graphs

graph_name	n_edges	n_nodes	avg_node_degree	avg_clust_coef	avg_short_path	diameter	density
My Graph	1217	150	16.23	0.55	2.88	9	0.11
Erdős-Rényi (p=0.11)	1182	150	15.76	0.11	2.06	3	0.11
Barabasi Albert (m=9)	1182	150	15.76	0.11	2.06	3	0.11
Watts–Strogatz (k=17, p=0.05)	1200	150	16.00	0.61	2.66	5	0.11
Powerlaw Graph (m=9, p=0.95)	1266	150	16.88	0.47	2.10	4	0.11

The closest random graph is Watts–Strogatz graph

3. Community detection. Clique search

- **number of maximal cliques - 706**
- **length of the maximum clique - 16**

- Valeria Isafarova
- Masha Panchenko
- Yulia Ugrina
- Kirill Yakovlev
- Maria Prokhorova
- Kristina Morozova
- Maria Mironova
- Irina Smolyanaya
- Angelina Kovalenko
- Alexandra Syschikova
- Kirill Isafarov
- Nika Golubeva
- Mlada Nikolaeva
- Polina Sigachyova
- Nadezda Trekina
- Alexandra Kuznechenkova

These people are my group mates from the Moscow City Teacher's Training University, where I studied a bachelor program.

3. Community detection. Modularity scores

Girvan–Newman method:

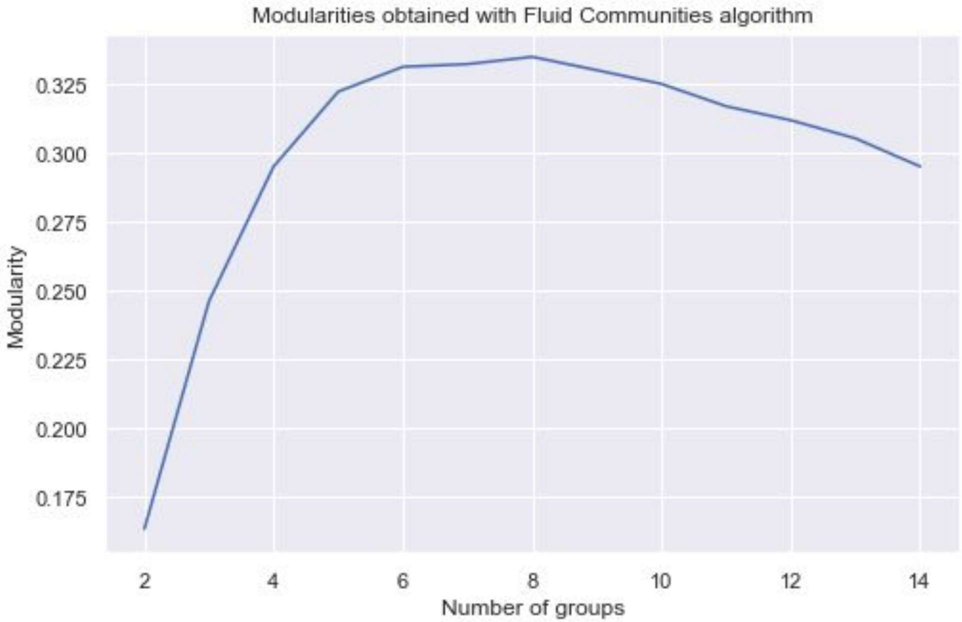
modularity - 0.18, 3 communities

Clauset-Newman-Moore greedy modularity maximization:

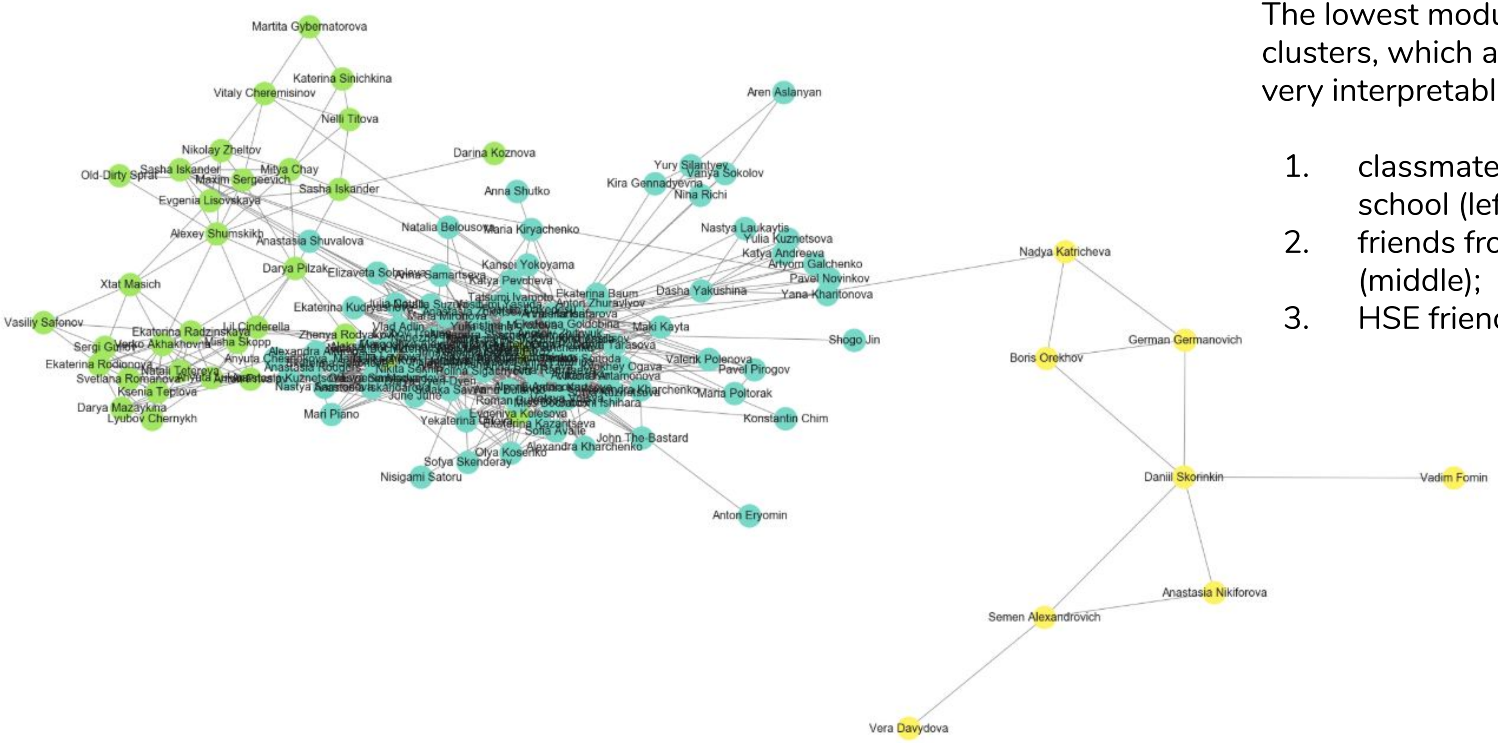
modularity - 0.31, 6 communities

Fluid Communities algorithm:

best modularity - 0.32-0.35 - with 8 communities



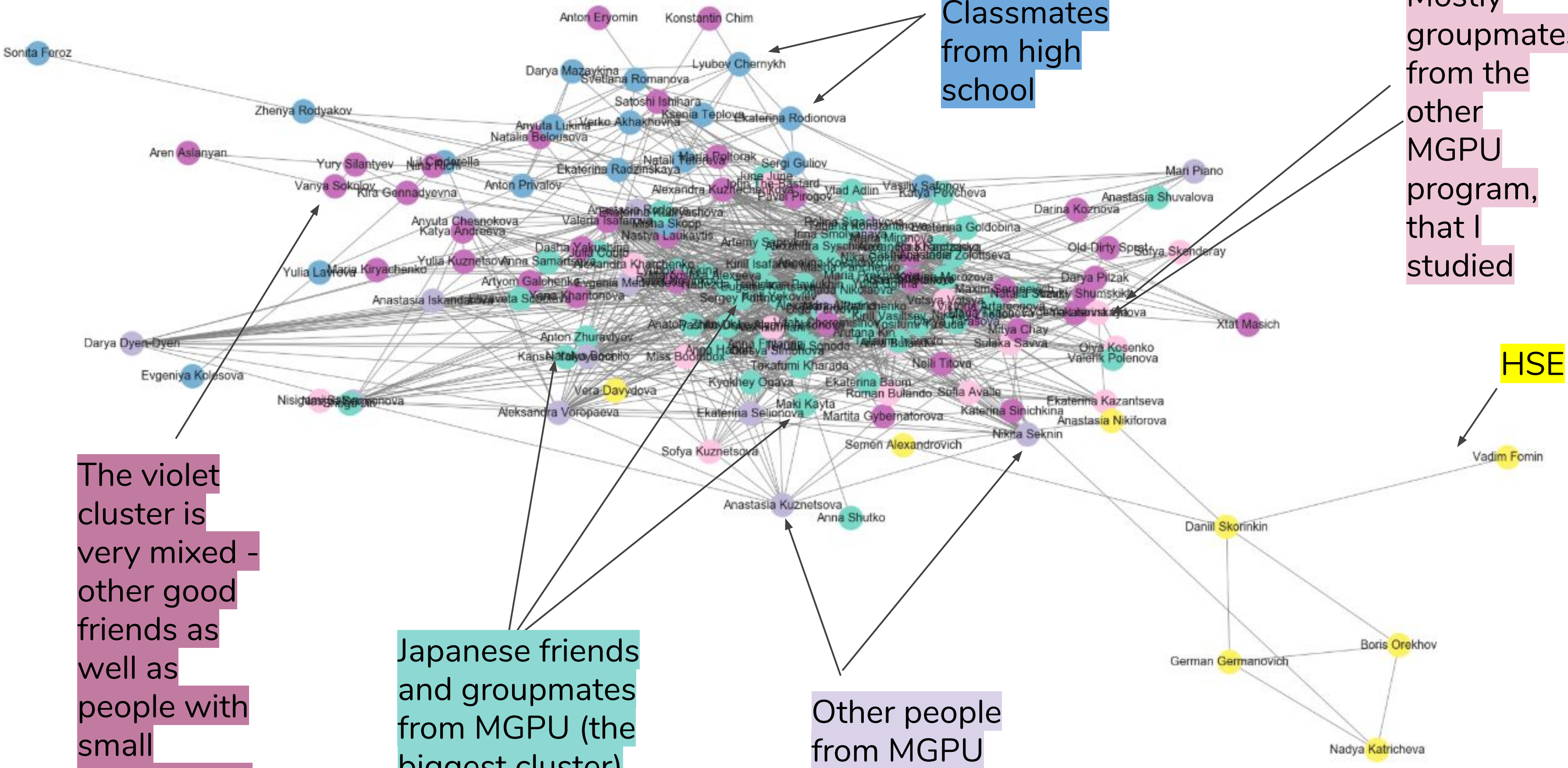
3. Community detection. Girvan-Newman method



The lowest modularity and 3 clusters, which are nevertheless very interpretable:

1. classmates from my last school (left);
2. friends from MGPU (middle);
3. HSE friends (right)

3. Community detection. Clauset-Newman-Moore maximization - 6 clusters



Classmates from high school

Mostly groupmates from the other MGPU program, that I studied

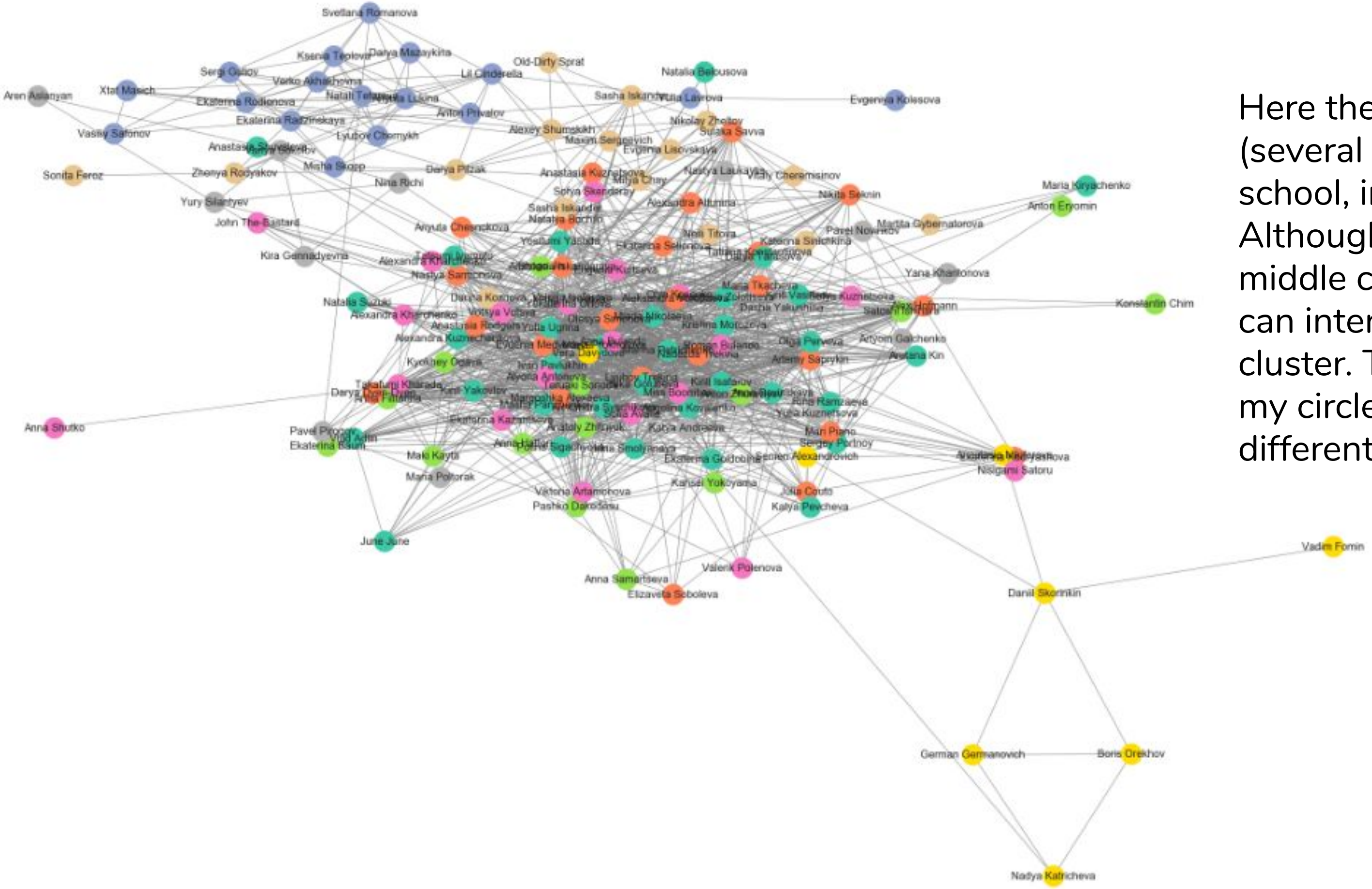
HSE

The violet cluster is very mixed - other good friends as well as people with small connections

Japanese friends and groupmates from MGPU (the biggest cluster)

Other people from MGPU

3. Fluid Communities algorithm - best modularity, 8 clusters



Here the partition is bigger (several communities in high school, in MGPU etc.). Although many people in the middle connected to many, I can interpret every small cluster. The groups reflect my circle of connections at different stages of life.