Network Analysis

Project 1

Beketova Anna

Group: мНОД18_ИССА

About the data

Data Extraction

- This is my friends network. My profile is https://vk.com/anyamb . I'm an active person and have lots of friends from different events, school, conferences, sport trips.
- Data is extracted using VK API from VK.com – one of thebiggest Russian Social Network
- Receiving the info about my friends: their “first_name”, “last_name”, “gender”, “city”, “education”, ”relationship status” and some personal features.
- Field 'personal' in VK includes various parameters: “political views”, “languages”, “religion”, “inspired_by”, “people_main” (improtant in others), “life_main” (personal priority), “smoking” (views on smoking), “alcohol” (views on alcohol). I decided to use only those coded as positive numbers, not strings, as this information is structured.

Attributes collected

alcohol	city	friends	gender	id	label	life_main	name	people_main	political	relation	smoking	university
0	Moscow	646	2	0	13572109	0	Dmitry Menshenin	0	0	0	0	250
2	Moscow	1495	2	1	54622222	5	Roman Faynshmidt	2	8	2	1	128
0	Moscow	253	1	2	205223952	6	Darya Rykova	2	0	0	0	128
0	Moscow	454	2	3	13801489	0	Alexey Chernov	2	4	1	0	2
0	Moscow	173	1	4	17930258	0	Katya Anderson	0	0	0	0	128

As for main fields:

- Name
- Gender (is coded as 1 — female; 2 — male; 0 — not specified);
- City (user's current city name);
- University id;
- Relationship status (such as: 1 – single; 2 – in a relationship; 3 – engaged; 4 – married: 5 – it's complicated; 6 – actively searching; 7 – in love).

0 in most cases means 'not specified'

Attributes collected

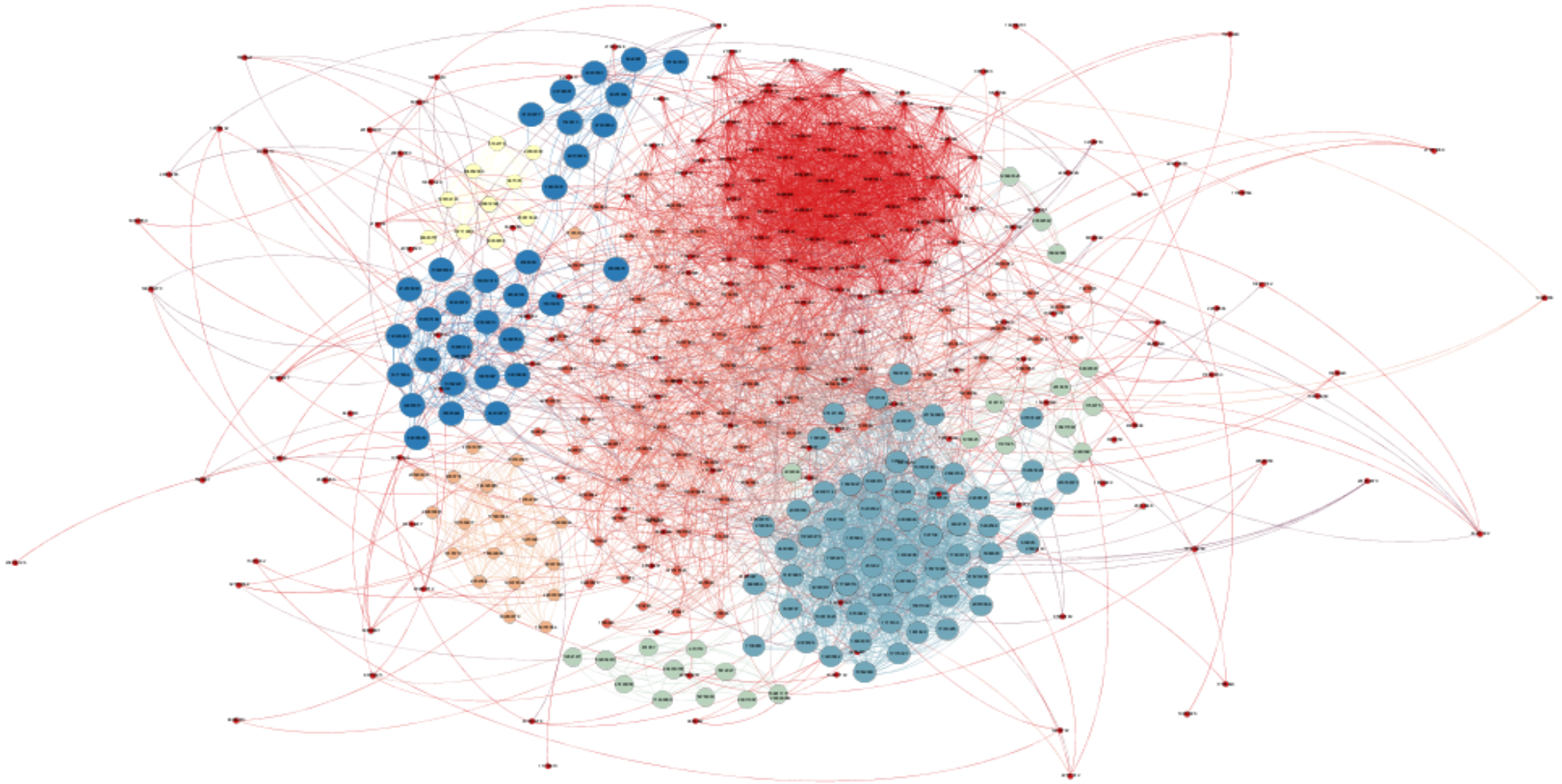
alcohol	city	friends	gender	id	label	life_main	name	people_main	political	relation	smoking	university
0	Moscow	646	2	0	13572109	0	Dmitry Menshenin	0	0	0	0	250
2	Moscow	1495	2	1	54622222	5	Roman Faynshmidt	2	8	2	1	128
0	Moscow	253	1	2	205223952	6	Darya Rykova	2	0	0	0	128
0	Moscow	454	2	3	13801489	0	Alexey Chernov	2	4	1	0	2
0	Moscow	173	1	4	17930258	0	Katya Anderson	0	0	0	0	128

As for additional fields:

- Political. 1 : "Communist", 2 : "Socialist", 3 : "Moderate", 4 : "Liberal", 5 : "Conservative", 6 : "Monarchist", 7 : "Ultraconservative", 8 : "Apathetic", 9 : "Libertian", 0 : "not specified";
- People_main. 1 : "intellect and creativity", 2 : "kindness and honesty", 3 : "health and beauty", 4 : "wealth and power", 5 : "courage and persistance", 6 : "humor and love for life", 0 : "not specified";
- Life_main. 1 : "family and children", 2 : "career and money", 3 : "entertainment and leisure", 4 : "science and research", 5 : "improving the world", 6 : "personal development", 7 : "beauty and art", 8 : "fame and influence", 0 : "not specified";
- Smoking.1 : "very negative", 2 : "negative", 3 : "neutral", 4 : "compromisable", 5 : "positive", 0 : "not specified";
- Alchohol. 1 : "very negative", 2 : "negative", 3 : "neutral", 4 : "compromisable", 5 : "positive", 0 : "not specified".

My Network Summary

Network



- My network consists of 399 friends with 3790 connections between each other. Number of connected components = 14

Network description

14 - number of connected components :

- 1 - number of nodes: 374
- 2 - number of nodes: 9
- 3 - number of nodes: 4
- 4 - number of nodes: 2
- 5-14 - number of nodes: 1

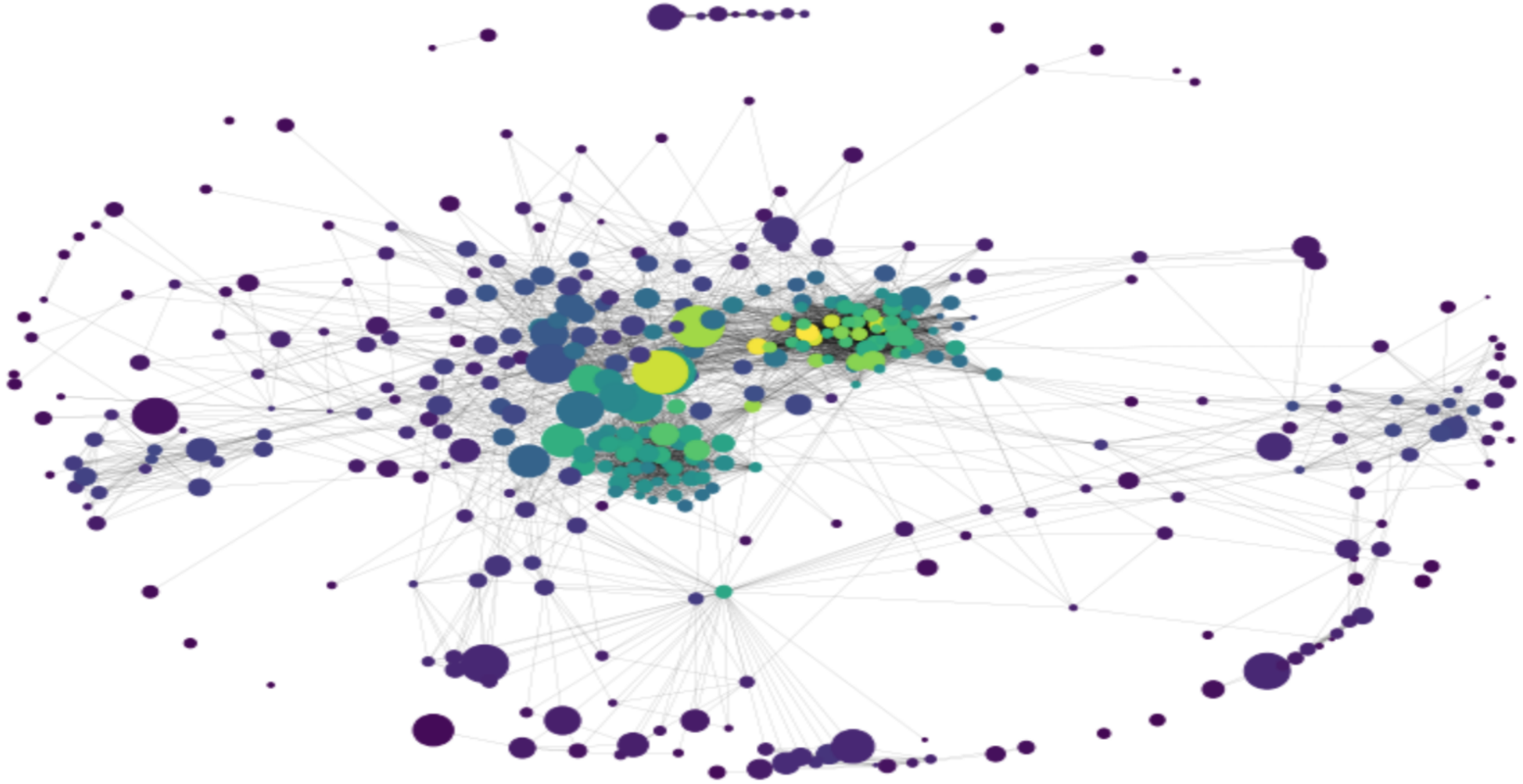
- Min value of node degree: 1
- Median value of node degree: 13.0
- Mean value of node degree: 20.08
- Max value of node degree: 77

Network description

- The network Diameter: 8
- Graph's diameter is the largest number of vertices which must be traversed in order to travel from one vertex to another

- The network average clustering coefficient: 0.51
- Average clustering coefficient is average measure of the degree to which nodes in a graph tend to cluster together. In our case, Average clustering coefficient is of mean value - not so high, not so low

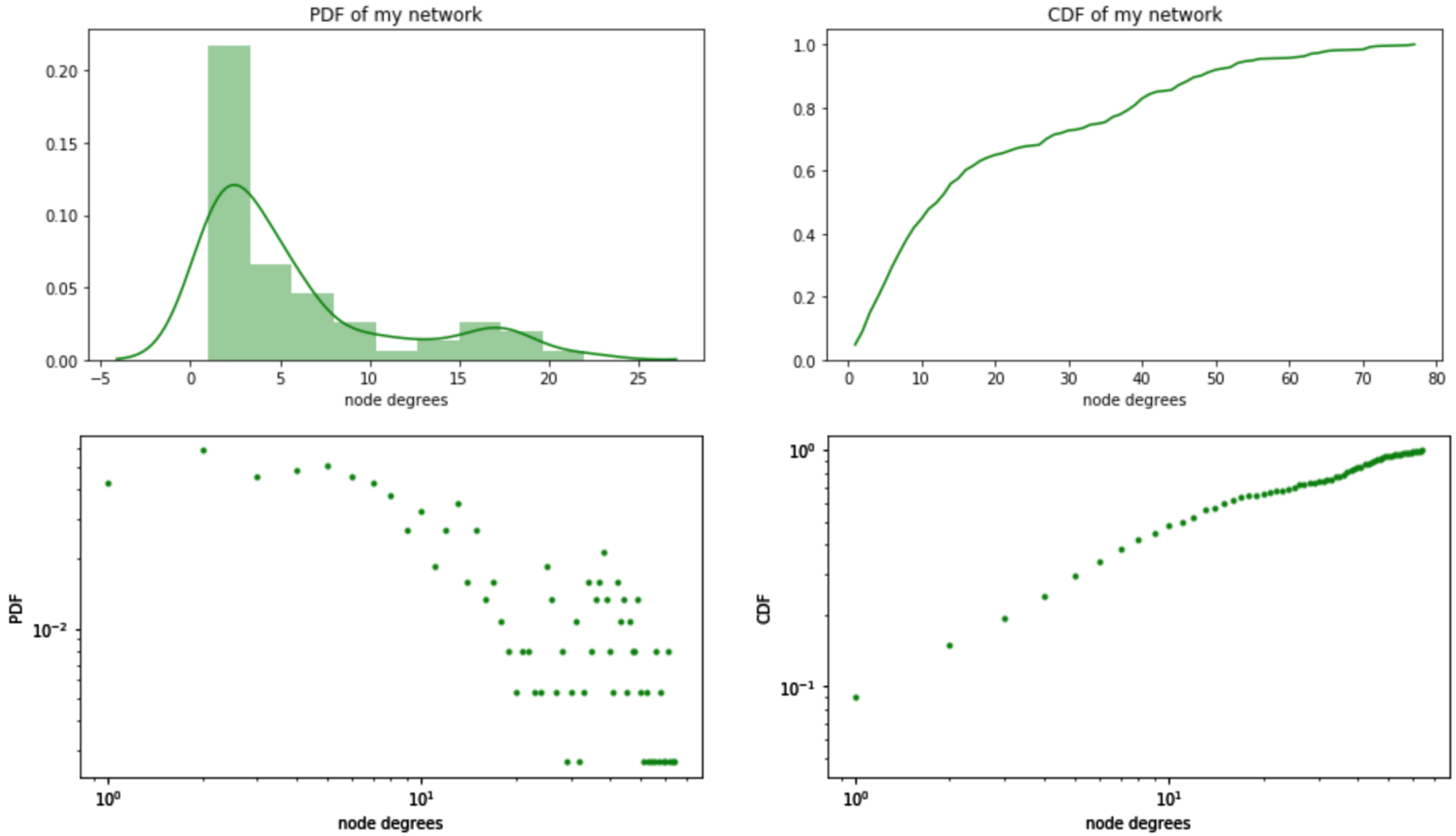
- Average shortest path length: 3.10
- So in average my friend needs 2 people more to find another person (approximately 3 with that friend)



- The darker a node the less connection it has in the graph – degree value
- The bigger a node the more friends overall it has – “popularity”

- So that small and green/yellow nodes are ”more like me” nodes – they have most of their friends in my network and they are tightly connected with my friends. Big and dark nodes on opposite is more popular but not so popular between my network

Degree distribution



- It is clear from PDF, that there are 2 main kinds of my friends - those who have 15-20 mutual friends on average, and those with 0-5 mutual friends.

Structural Analysis

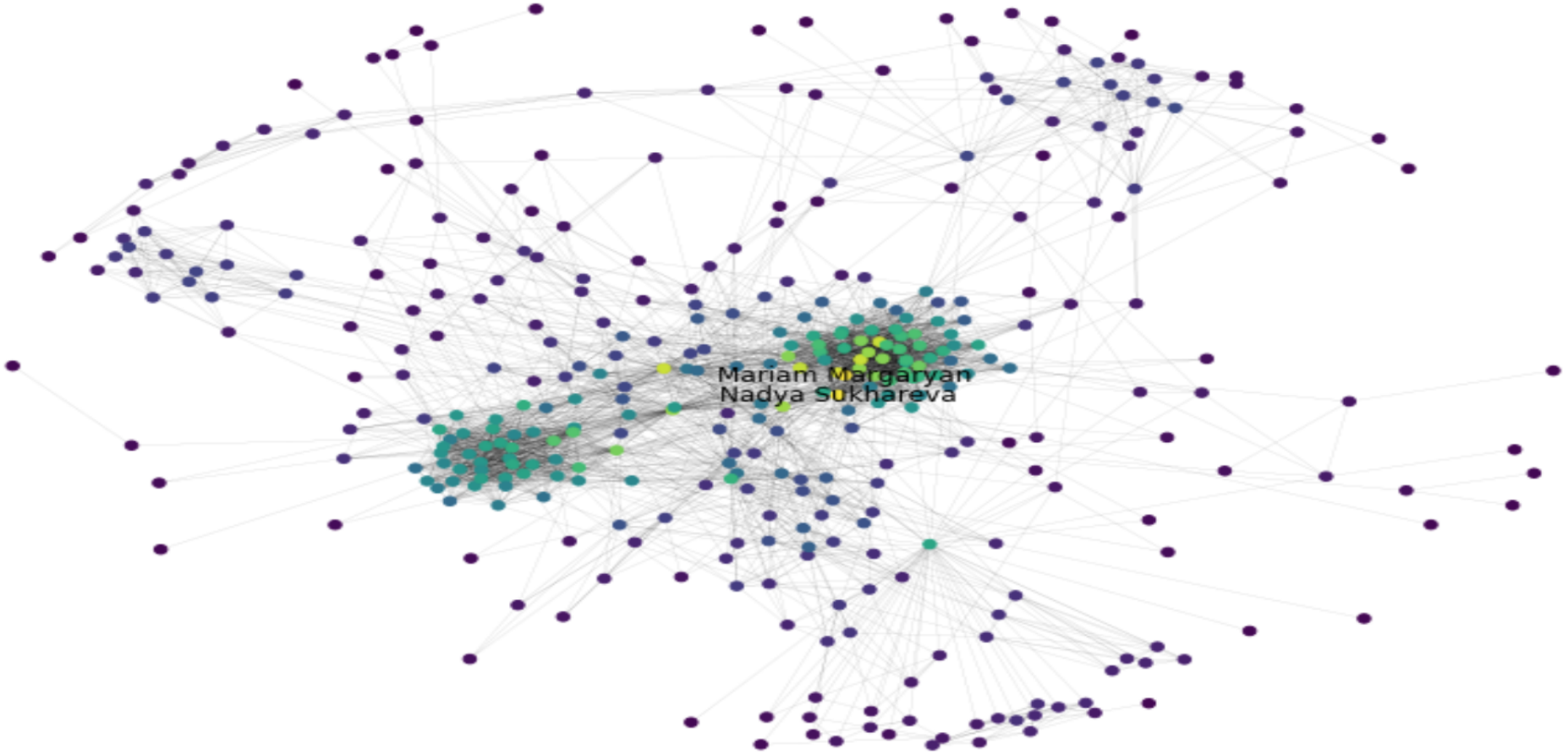
Degree centralities

DegreeCentrality	DegreeCentralityRank	name
0.206434	1	Mariam Margaryan
0.203753	2	Nadya Sukhareva
0.193029	3	Daria Kuznetsova
0.190349	4	Oleg Pavlyuk
0.190349	5	Darya Sidoruk
0.190349	6	Viktor Kozlov
0.187668	7	Nina Panova
0.176944	8	Nikolay Tesla
0.174263	9	Anna Rezyapova
0.174263	10	Olya Gritsenko

- Degree centralities represent how many connections in a given network a node has
- Top nodes are:

Mariam Margarian and Nadya Sukhareva, Daria Kuznetsova, Oleg Pavlyuk, Daria Sidoruk – my University friends and groupmates from Bachelor degree, really sociable people

Degree centralities



- Color indicates the number of degrees of the node (the darker the lower degree)
- It is seen that only a few people have plenty of connections, the majority of the network is connected with less nodes

Closeness centralities

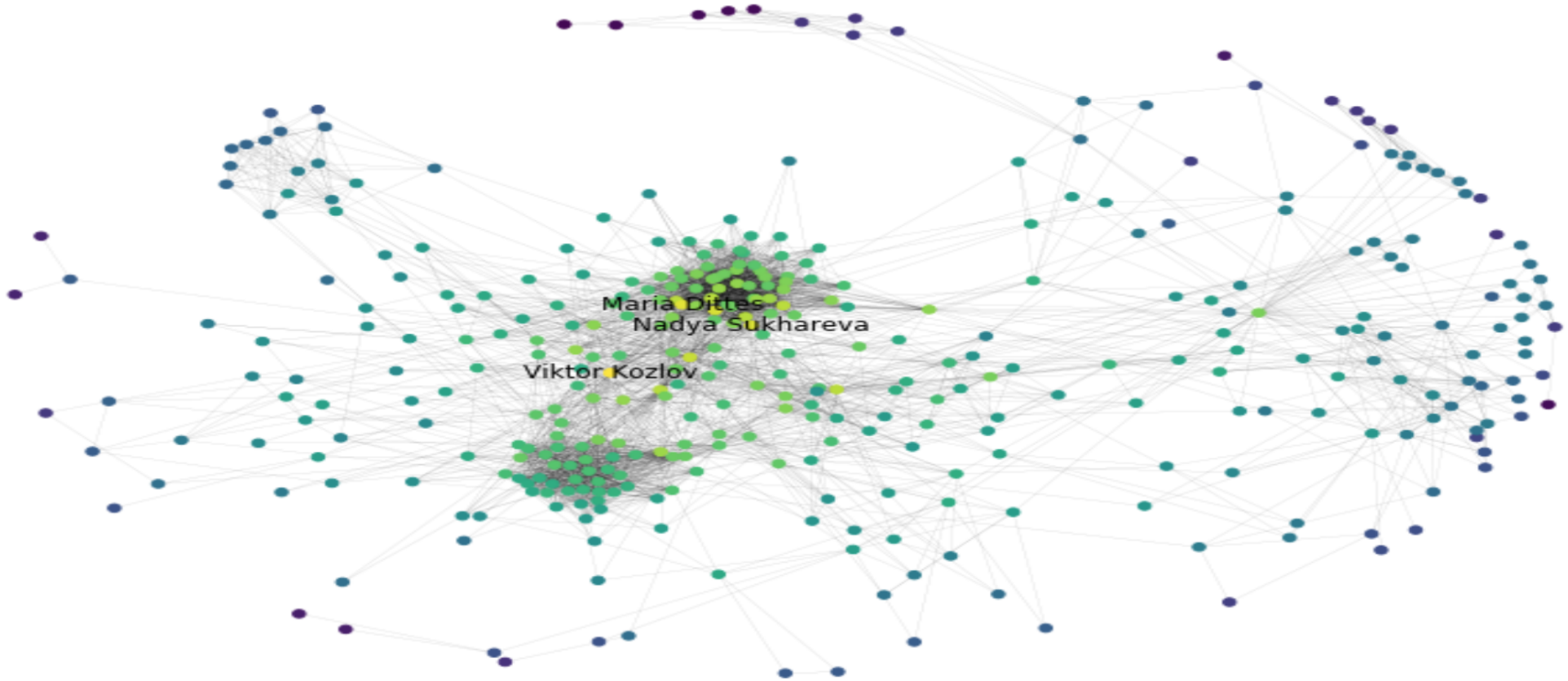
ClosenessCentrality	ClosenessCentralityRank	name
0.459926	1	Viktor Kozlov
0.447779	2	Maria Dittes
0.44352	3	Nadya Sukhareva
0.44352	4	Mariam Margaryan
0.436257	5	Nikolay Tesla
0.430716	6	Oleg Pavlyuk
0.430716	7	Nina Panova
0.429724	8	Darya Sidoruk
0.429724	9	Garri Rutberg
0.429229	10	Anna Rezyapova

- Closeness Centrality is calculated as the average distance from a given starting node to all other nodes in the network, identify ”central nodes”
- Top nodes are:

Viktor Kozlov (financial expert who shares lots of courses and career opportunities, therefore has lots of people in friends)

Nadya Sukhareva and Maria Dittes – my groupmates from Bachelor, really sociable people

Closeness centralities



- Color indicates the average distance from a given starting node to all other nodes (the darker the higher distance)
- According to closeness centrality, there are a few nodes that are “hard starting points” (darker nodes) to find new connection, most of my friends connected within 1-3 friends

Betweenness centralities

BetweennessCentrality	BetweennessCentralityRank	name
0.175587	1	Anatoly Burtsev
0.106836	2	Viktor Kozlov
0.0612637	3	Maria Dittes
0.0559108	4	Yulia Ershova
0.0538501	5	Alexandra Voronkova
0.0497915	6	Garri Rutberg
0.0449215	7	Nikolay Tesla
0.0399048	8	Anna Rezyapova
0.0332819	9	Veronika Taxir
0.0326058	10	Taisia Kulikova

- Betweenness Centrality measures how often a node appears on shortest paths between nodes in the network
- Top nodes are like connections:

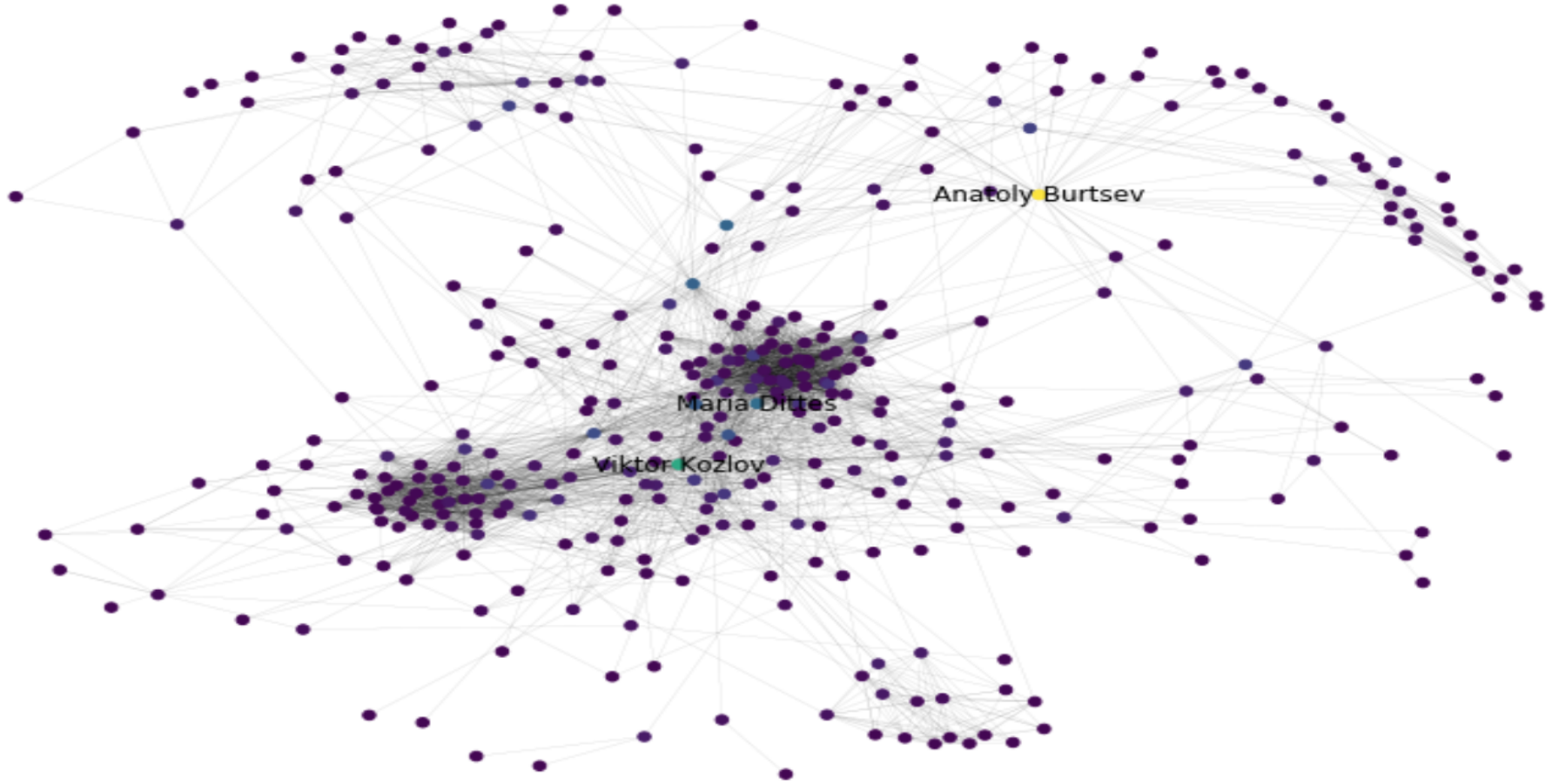
Anatoly Burtsev is my boyfriend, ”the connection to Yaroslavl and MSU friends”

Viktor Kozlov - financial expert, ”the connection to finance network”

Maria Dittes – is my university friend, “the connection to my Bachelor degree friends”

Yulia Ershova – “the connection to my home town school friends”

Betweenness centralities



- Color indicates how often a node appears on shortest paths between nodes (the darker the rarer)
- There are a few nodes (light colors) that connect main communities. More often they are people how appear in several different communities and are “connections” between these communities

PageRank

PageRank	PageRankRank	name
0.0123477	1	Anatoly Burtsev
0.00812842	2	Viktor Kozlov
0.00746788	3	Nikolay Tesla
0.00691662	4	Nadya Sukhareva
0.00682567	5	Garri Rutberg
0.00668513	6	Mariam Margaryan
0.00637849	7	Nina Panova
0.00628429	8	Oleg Pavlyuk
0.00627101	9	Daria Petrova
0.00626042	10	Daria Kuznetsova

- PageRank shows those nodes where you finally will be if randomly choose another connected nodes of current nodes

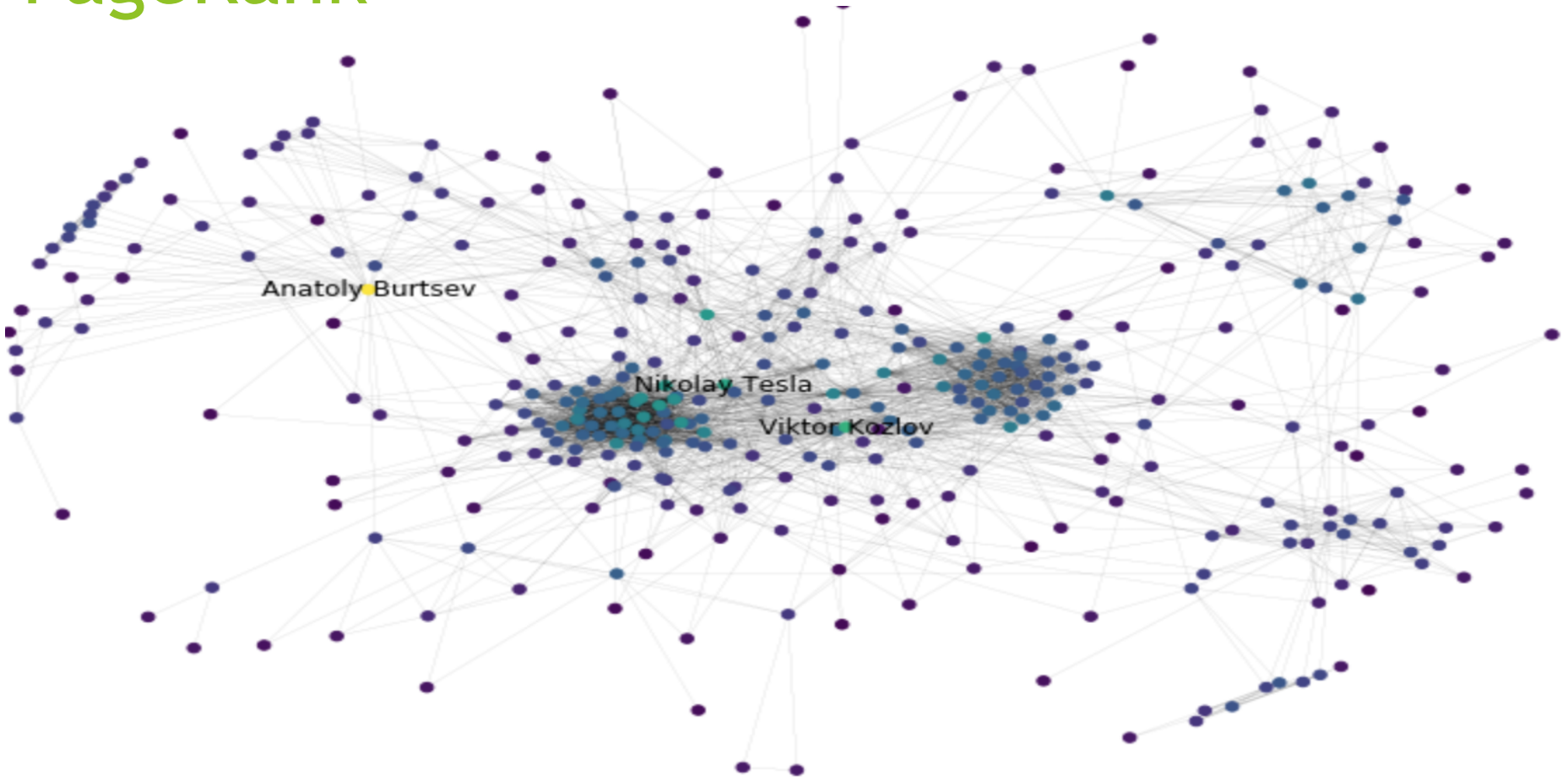
- In my graph all links lead to:

Anatoly Burtsev as the main person in my network

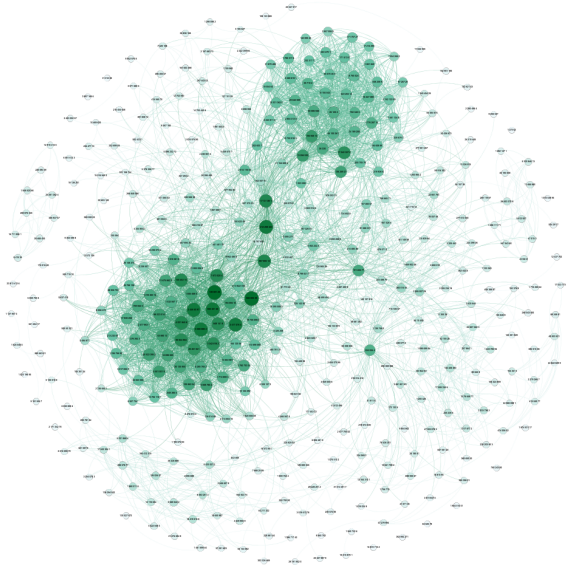
Viktor Kozlov as the person who share a significant amount of useful information

Nikolay Tesla as the main connection of university friends and sport activities

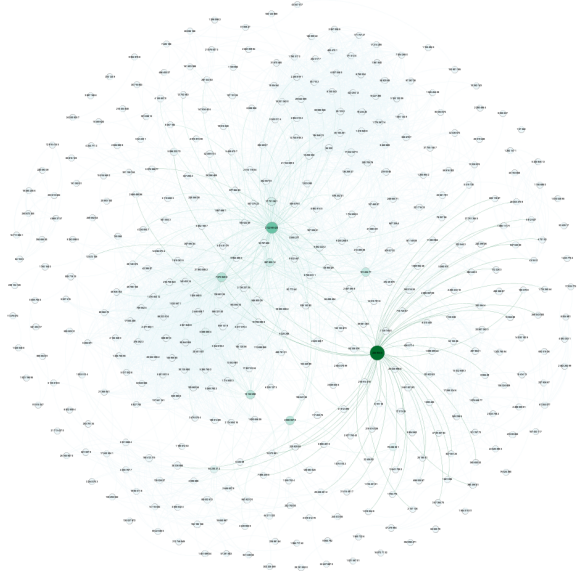
PageRank



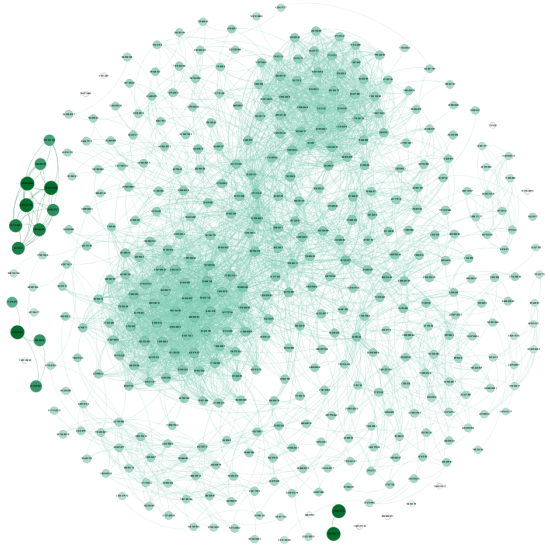
- PageRank shows that there are a few main “final points” – people with bright light nodes and a lot of not significant “final points” – people with dark nodes



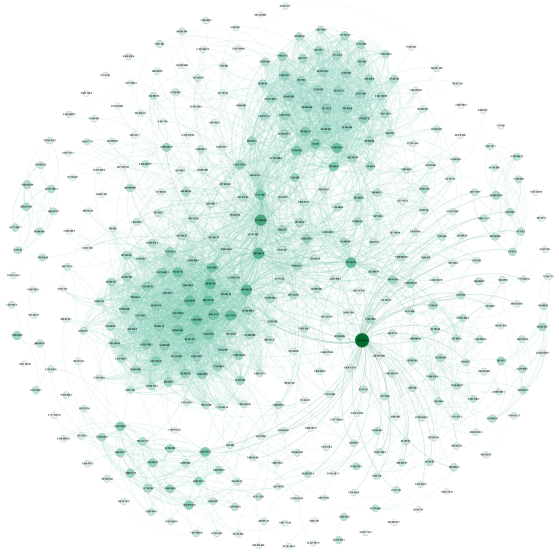
- Degree Centrality



- Betweenness Centrality

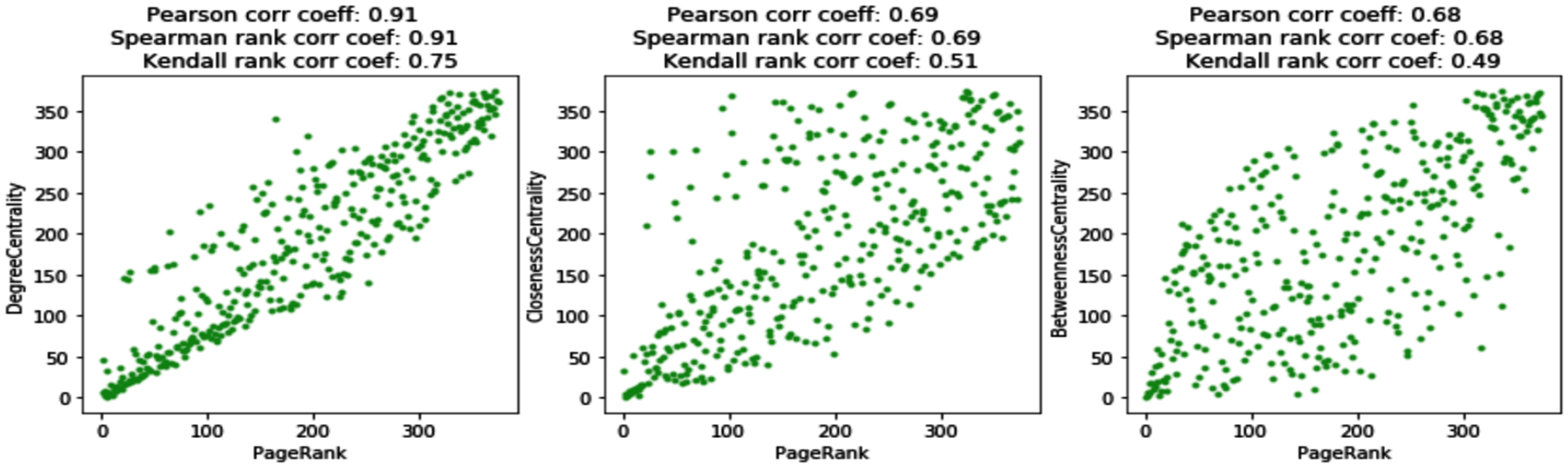


- Closeness Centrality



- PageRank

PageRank comparison with centralities



- Spearman rank correlation coefficient assesses how well the relationship between two variables can be described using a monotonic function.
- Kendall rank correlation coefficient uses metric that counts the number of pairwise disagreements between two ranking lists.
- There is strong direct monotonic association between ranking combinations. Degree Centrality results correlate with PageRank highly. The least connection (still high) is between PageRank and Betweenness Centrality.
- According to general layout (not taking into account ordering as above) of networks colored by these measures, the most similar picture for PageRank is Betweenness Centrality.

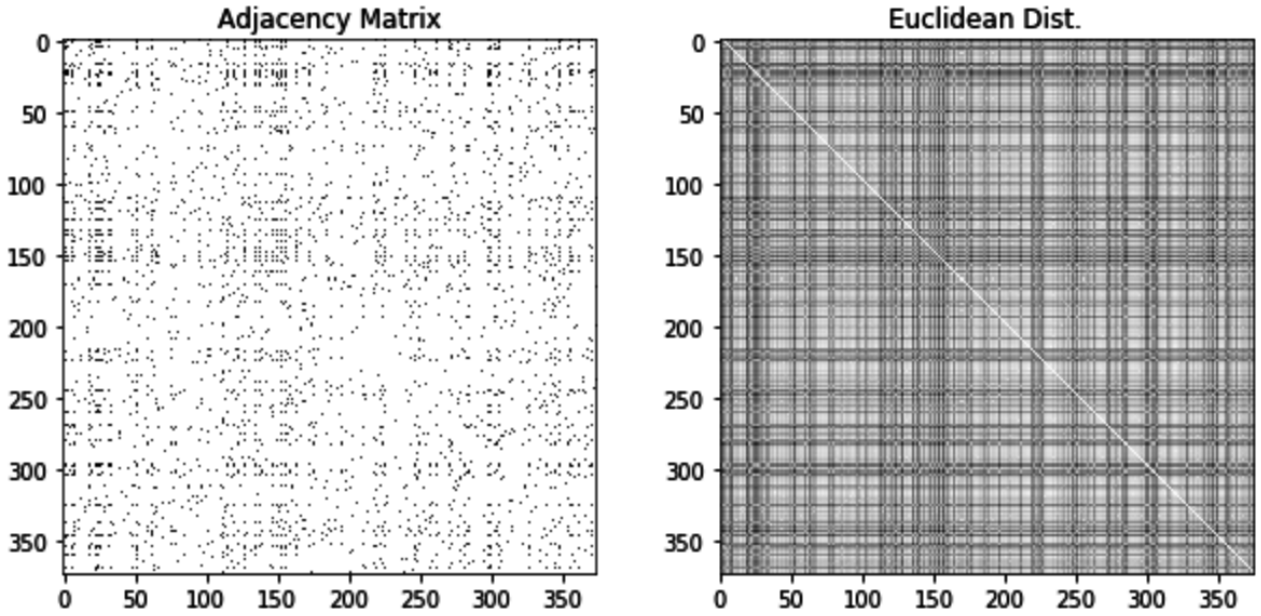
Assortative Mixing according to node attributes

	alcohol	city	friends	gender	life_main	people_main	political	relation	smoking	university
Modularity	0.006065	0.051775	-0.004263	0.062999	-0.014584	-0.013953	-0.015491	-0.011684	0.006915	0.092875
Number of NaN	315.000000	0.000000	0.000000	0.000000	289.000000	282.000000	309.000000	291.000000	305.000000	103.000000

- r = 1 - Assortative network ("like links with like"): interconnected high degree nodes - core, low degree nodes - periphery
- r = -1 - Disassortative network (""like links with dislike"): high degree nodes connected to low degree nodes, star-like structure
- r = 0 the network is non-assortative

- Here we have lots of missing data in personal characteristics
- All parameters could be considered as non-assortative – with no exact tendency. It could make sense: I'm really active and my network consists of people from summer schools, sport trips, university, courses, some organizations. I add people from different events not looking at their city, number of friends gender and so on. We just need to keep in touch.

Node structural equivalence/similarity



- Two nodes of a network are structurally equivalent if they share many of the same neighbors.
- For node structural equivalence calculation let's use one of algorithm - Euclidean Distance. Euclidean distance is larger for vertices which differ more.

Node structural equivalence/similarity

- Two nodes of a network are structurally equivalent if they have the same neighbors.

- Result is:

```
Members 40 and 339 have structural equivalence with neigbors: [43]
Members 47 and 107 have structural equivalence with neigbors: [32, 240, 98, 342, 247]
Members 171 and 174 have structural equivalence with neigbors: [375]
Members 230 and 245 have structural equivalence with neigbors: [176]
```

- Not so many structurally equivalent nodes. All of them are different, so no central element could be found.
- This means that the network is diverse and there exists only a few nodes with the same neighbors. Almost every node is unique and has his/her own contacts.

The closest random graph model similar to my social network

	# edges	Clustering coeff	Path length	Diameter	K-S stat	K-S p_val
My network	3755	0.510949	3.101905	8	-	-
Erdos-Renyi model	3736	0.054692	2.273000	4	0.467914	1.17415e-36
Barabasi-Albert model	3640	0.124854	2.253975	3	0.419786	1.37946e-29
Watts-Strogatz model	3740	0.127392	2.335250	3	0.542781	3.53332e-49

- Kolmogorov-Smirnov is a two-sided test for the null hypothesis that 2 independent samples are drawn from the same continuous distribution.
- If the K-S statistic is small or the p-value is high, then we cannot reject the hypothesis that the distributions of the two samples are the same.

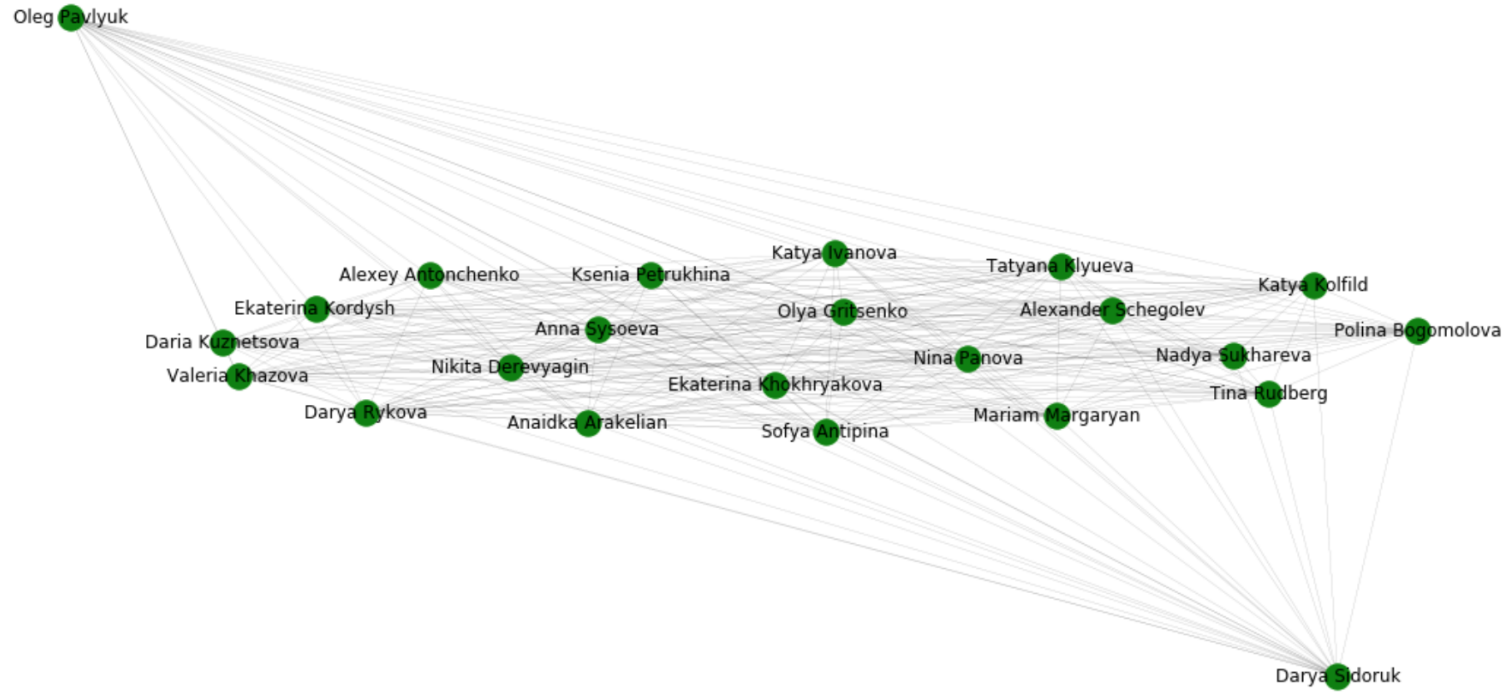
- According to KS test the closest from these models random graph is Barabasi-Albert model. Still there are huge differences in other parameters

Community Detection

Clique search

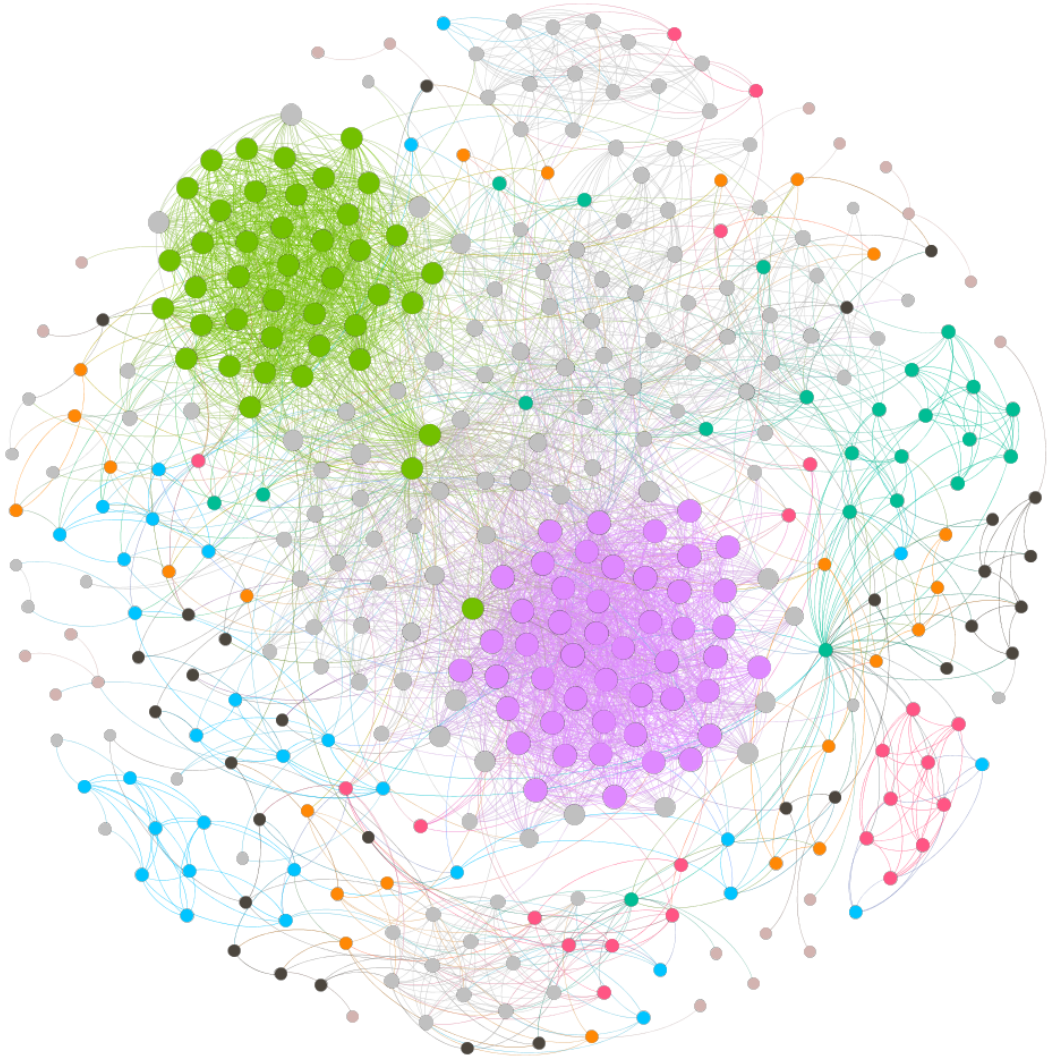
- A clique in a graph G is a complete subgraph of G. That is, it is a subset K of the vertices such that every two vertices in K are the two endpoints of an edge in G. A maximal clique is a clique to which no more vertices can be added.
- Number of maximal cliques of my friends network is: 3567
- Each of them contains up to 23 people

Clique search

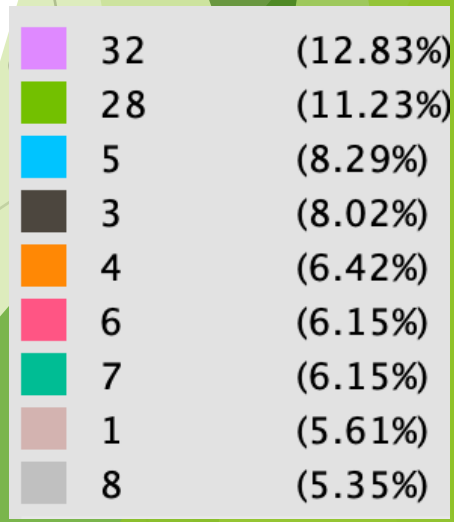


- Here is one of the maximal cliques with 23 members
- All these people are my Bachelor groupmates or people from that course

K-core decomposition



- The k-core decomposition identify subgraphs of increasing centrality (the property of being more densely connected)
- Here we can see clusters – communities based on their interconnections
- It is clear, that my network has lots of several clusters: the pink and green ones are the most tightly connected while others have significantly less number connections within.
- Main clusters are my friends from University and winter school



Community detection by modularity

Modularity score

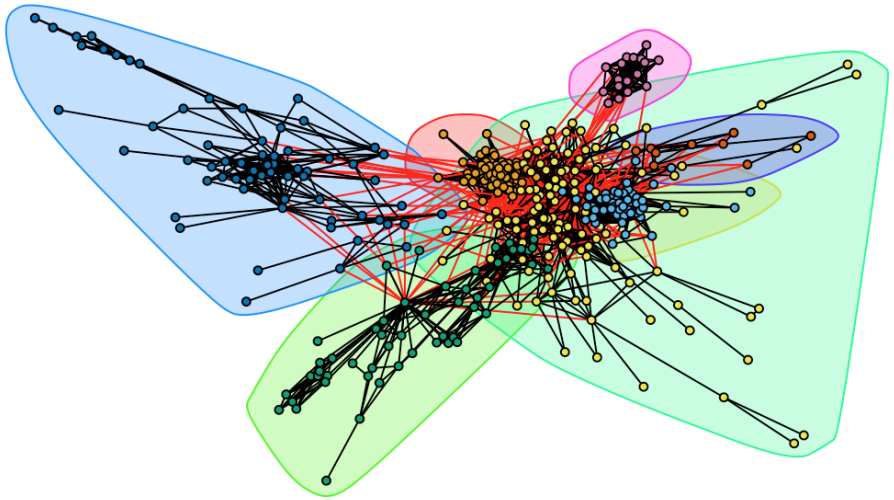
$$Q = \frac{1}{2m} \sum_{ij} \left(A_{ij} - \frac{k_i k_j}{2m} \right) \delta(c_i, c_j), = \sum_{u} \left(\frac{m_u}{m} - \left(\frac{k_u}{2m} \right)^2 \right)$$

m_u - number of internal edges in a community u,
k_u - sum of node degrees within a community

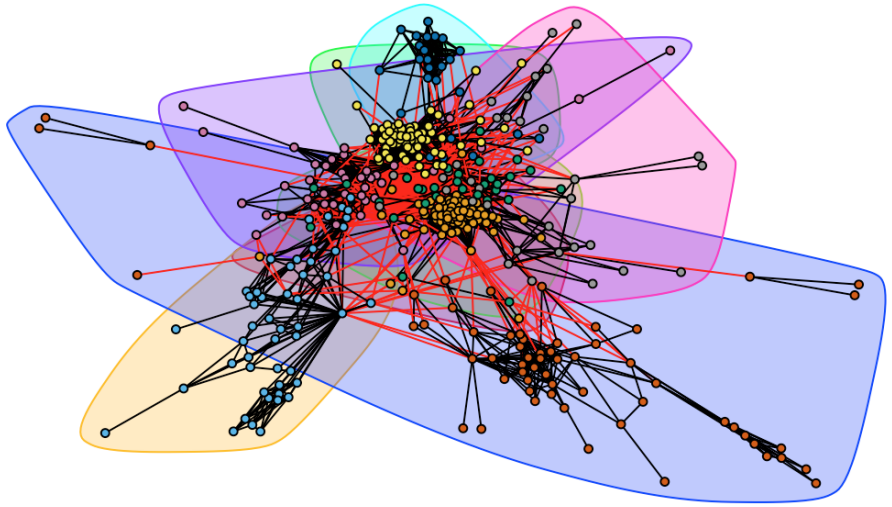
Modularity score range $Q \in [-1/2, 1)$, single community $Q = 0$

The higher the modularity score - the better are communities

Community detection in R
Top 2 by modularity



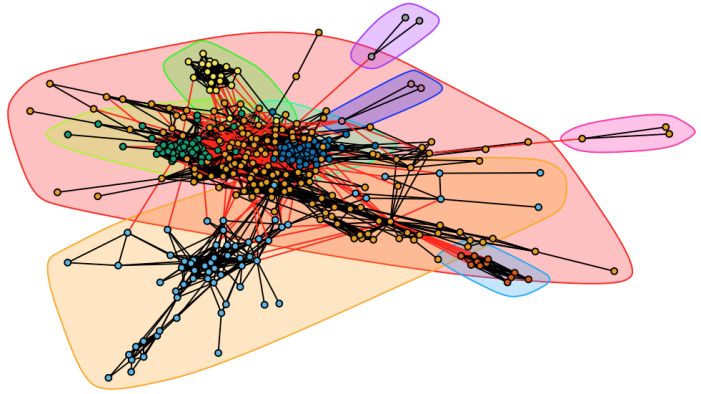
cluster_louvain, modularity = 0.572



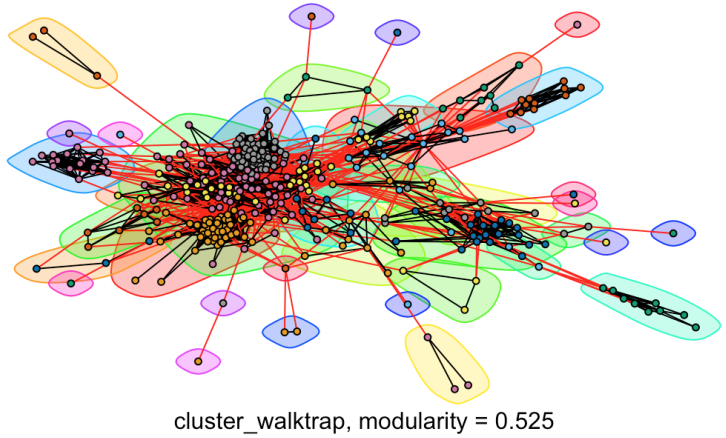
cluster_leading_eigen, modularity = 0.541

- cluster_louvain implements the multi-level modularity optimization algorithm for finding community structure. It is based on the modularity measure and a hierarchial approach.
- cluster_leading_eigen tries to find densely connected subgraphs in a graph by calculating the leading non-negative eigenvector of the modularity matrix of the graph.
- It is seen that partitions with fewer but larger clusters had higher modularity score for my network. So in my network larger clusters dominate

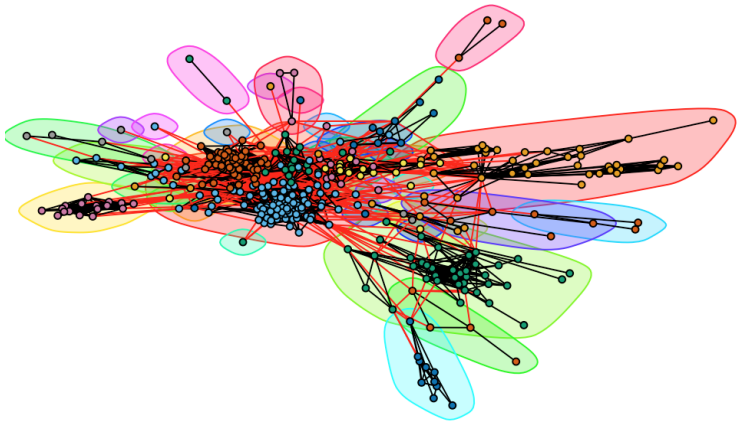
Community detection in R
Other partitions



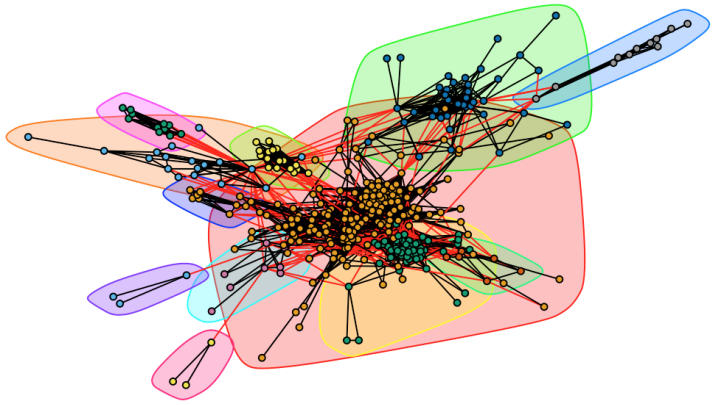
cluster_fast_greedy, modularity = 0.539



cluster_walktrap, modularity = 0.525



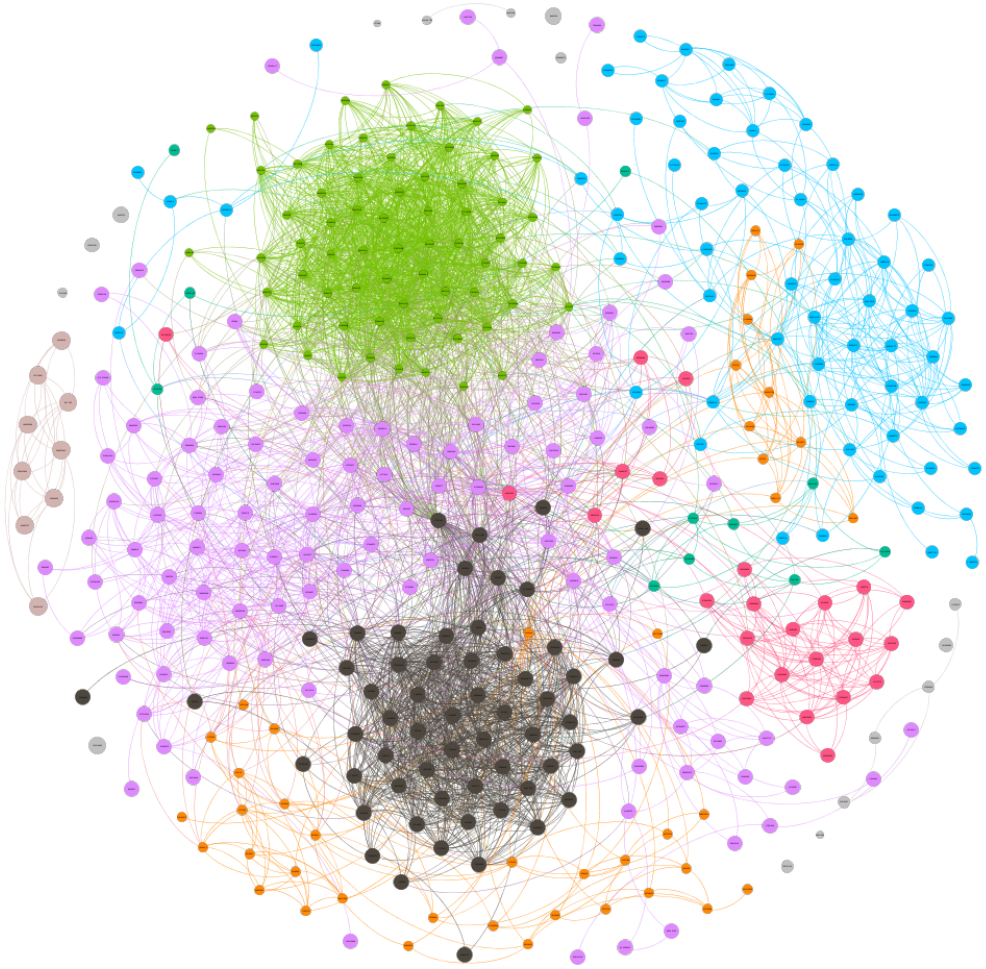
cluster_edge_betweenness, modularity = 0.520



cluster_label_prop, modularity = 0.459

- It is seen that partitions with fewer but larger clusters had higher modularity score for my network. So in my network larger clusters dominate

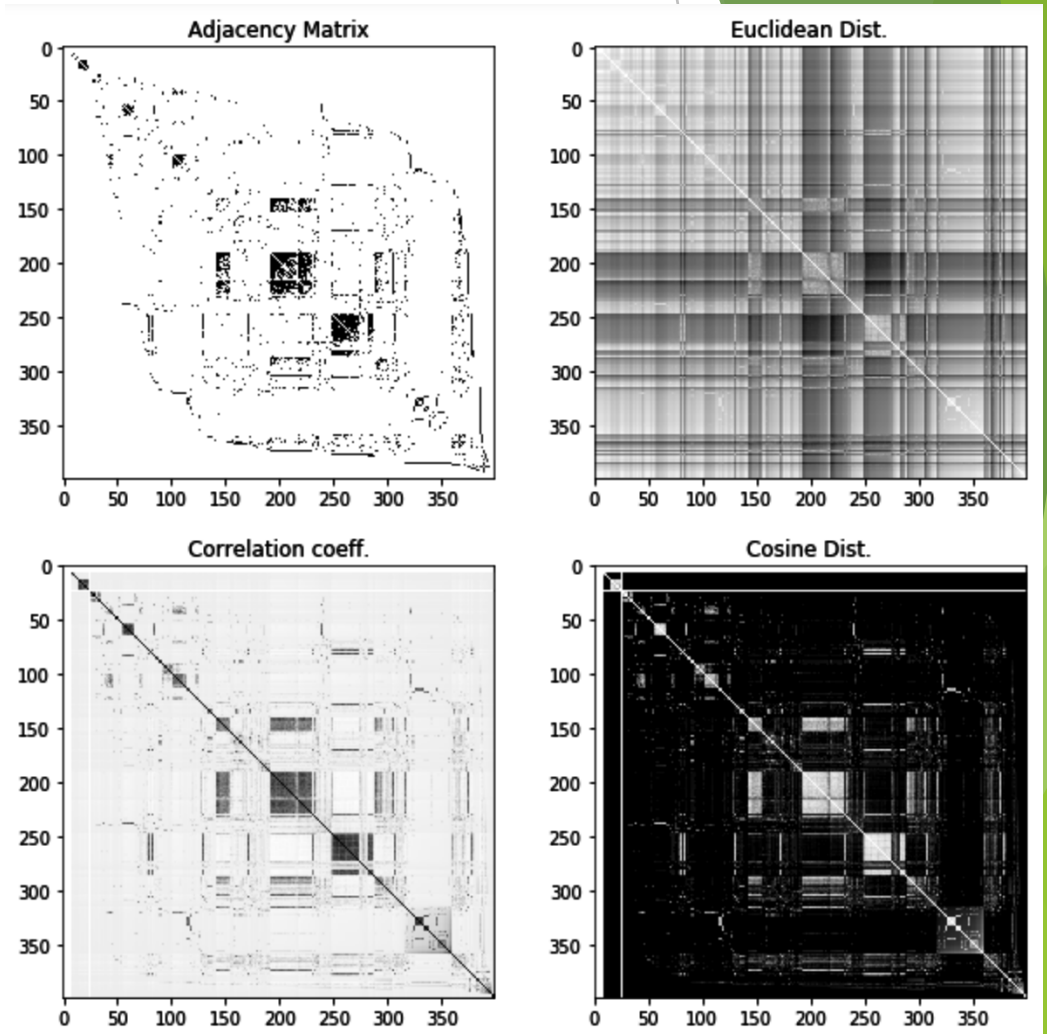
Community detection in Gephi



- Those in green are my groupmates from Bachelor and coursemates
- Those in pink – are my university related friends (also Master degree friends)
- Those in black – are my friends from Changellenge winter school
- Those in light blue on the right are my school/home town friends
- Those in light brown – are people from speaking club
- Those in orange on the bottom – are friends from snowboarding camp
- Those in orange on the right – are friends from Javabootcamp
- Those in red on the right bottom – are friends from conference

Community detection in Python

- After reordering the nodes the following matrices were received.
- It is clear, that there are 2-3 large communities – big squares on the middle of diagonal. Also, there are many small communities – small squares on the diagonal.
- As it was mentioned: these main communities are my friends from University, from Changellenge winter school, from my school and hometown.
- Tiny communities may be my school friends, sport trips friends, some courses friends, speaking club and mafia club friends.



Thank you for your attention!