

FEDERAL STATE AUTONOMOUS EDUCATIONAL
INSTITUTION OF HIGHER EDUCATION ITMO
UNIVERSITY

Report
on the practical task No. 7
“Algorithms on graphs. Tools for network
analysis”

Performed by
Kozlov Alexey
Dmitriy Koryakov
J4133c
Accepted by
Dr Petr Chunaev

St.Petersburg
2022

TABLE OF CONTENTS

Goal	3
Problems	3
Brief theoretical part	3
Solution	4
Choose network dataset	4
Reformat dataset for gephi format	4
Upload the dataset in Gephi	4
Obtain a graph layout of at least two different types	6
Yifan Hu	6
Fruchterman Reingold	7
Network measures in Statistics	7
Degree	7
In-Degree	8
Out-Degree	9
Diameter	9
Graph Density	10
Modularity	10
Average Clustering Coefficient	10
Conclusion	11

Goal

The use of the network analysis software Gephi.

Problems

1. Download and install Gephi from <https://gephi.org/>.
2. Choose a network dataset from <https://snap.stanford.edu/data/> with number of nodes at most 10,000. You are free to choose the network nature and type (un/weighted, un/directed).
3. Change the format of the dataset for that accepted by Gephi (.csv, .xls, .edges, etc.), if necessary.
4. Upload and process the dataset in Gephi. Check if the parameters of import and data are correct.
5. Obtain a graph layout of at least two different types.
6. Calculate available network measures in Statistics provided by Gephi.
7. Analyze the results for the network chosen.

Brief theoretical part

Yifan Hu - algorithm for placing nodes of graphs. It is based on force-directed model with a graph coarsening technique to reduce the complexity. The repulsive forces on one node from a cluster of distant nodes are approximated by a Barnes-Hut calculation, which treats them as one super-node.

Fruchterman Reingold - algorithm for placing nodes of graphs. It is of a force-directed algorithm, which uses an analogy of physical springs as edges that attract connected vertices toward each other and a competing repulsive force that pushes all vertices away from one another, whether they are connected or not.

Directed weighted graph is $G = (V, E, w)$, where $V = \{v_1, v_2, \dots, v_n\}$ is the set of vertices, $E = \{e_1, e_2, \dots, e_m\}$ is the ordered set of edges and each vertex $v_i \in V$ is associated with a weight $w(v_i)$

Degree is the number of edges that are incident to the vertex.

Diameter is the length of the shortest path between the most distanced nodes.

Density is the ratio between the edges present in a graph and the maximum number of edges that the graph can contain.

Modularity is a measure of the structure of networks or graphs which measures the strength of division of a network into modules

Clustering is a local characteristic of a network. It characterizes the degree of interaction between the nearest neighbors of a given node.

Clustering coefficient is the probability that the two nearest neighbors of this node are themselves nearest neighbors.

Betweenness centrality is a measure of centrality in a graph based on shortest paths.

Closeness centrality of a node is a measure of centrality in a network, calculated as the reciprocal of the sum of the length of the shortest paths between the node and all other nodes in the graph.

Solution

Choose network dataset

Dataset - Bitcoin OTC weighted directed graph.

Dataset description - This is who-trusts-whom network of people who trade using Bitcoin on a platform called Bitcoin OTC. Since Bitcoin users are anonymous, there is a need to maintain a record of users' reputation to prevent transactions with fraudulent and risky users. Members of Bitcoin OTC rate other members on a scale of -10 (total distrust) to +10 (total trust) in steps of 1. This is the first explicit weighted signed directed network available for research.

Nodes count - 5881

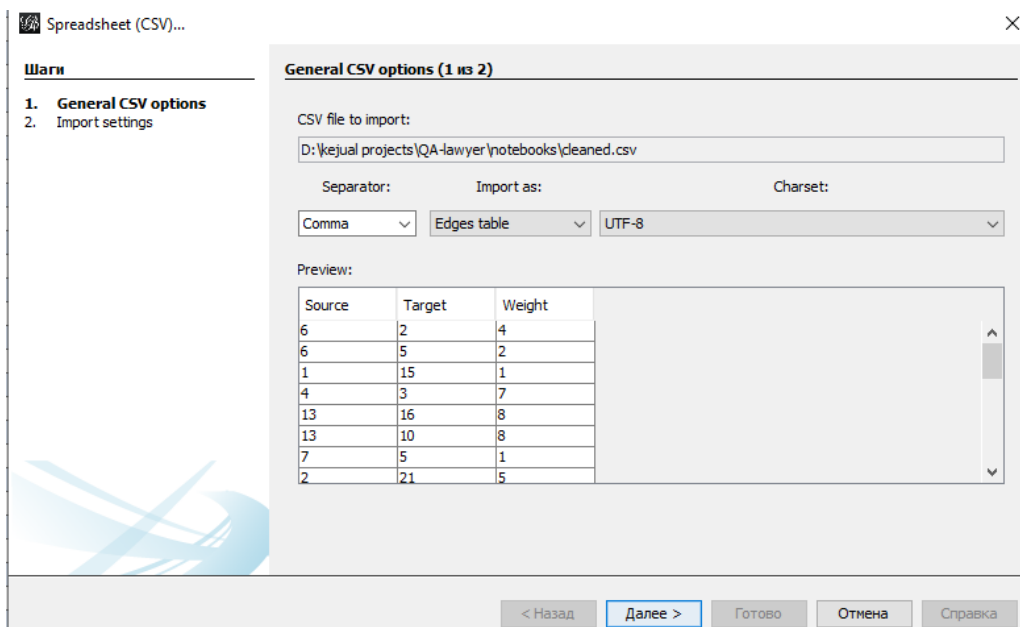
Edges count - 35592

Reformat dataset for gephi format

Source format - a csv file with SOURCE, TARGET, RATING, TIME columns and “,” as a sep character. Column names are not specified in the file.

Target format - csv file with Source,Target,Weight columns and “,” as a sep character. Column names are specified in the file.

Upload the dataset in Gephi



Spreadsheet (CSV)...

Warn

1. General CSV options
2. Import settings

General CSV options (1 из 2)

CSV file to import:

D:\kejual projects\QA-lawyer\notebooks\cleaned.csv

Separator: Comma Import as: Edges table Charset: UTF-8

Preview:

Source	Target	Weight
6	2	4
6	5	2
1	15	1
4	3	7
13	16	8
13	10	8
7	5	1
2	21	5

< Назад Далее > Готово Отмена Справка

Шаги

1. General CSV options
2. **Import settings**

Import settings (2 из 2)

Time representation

Intervals ▾

Imported columns:

☒ Source

☒ Target

☒ Weight

Double ▾

< Назад

Далее >

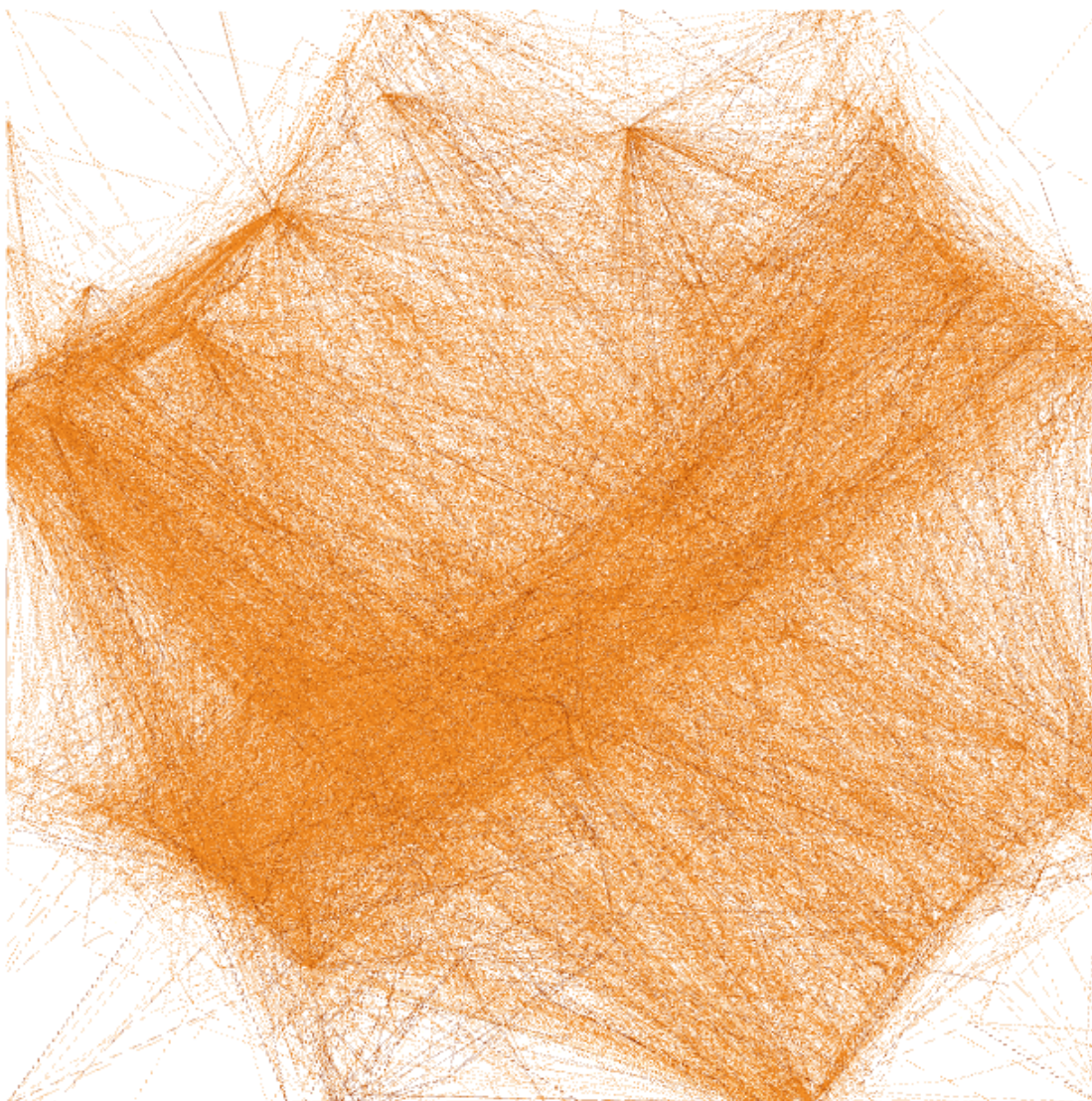
Готово

Отмена

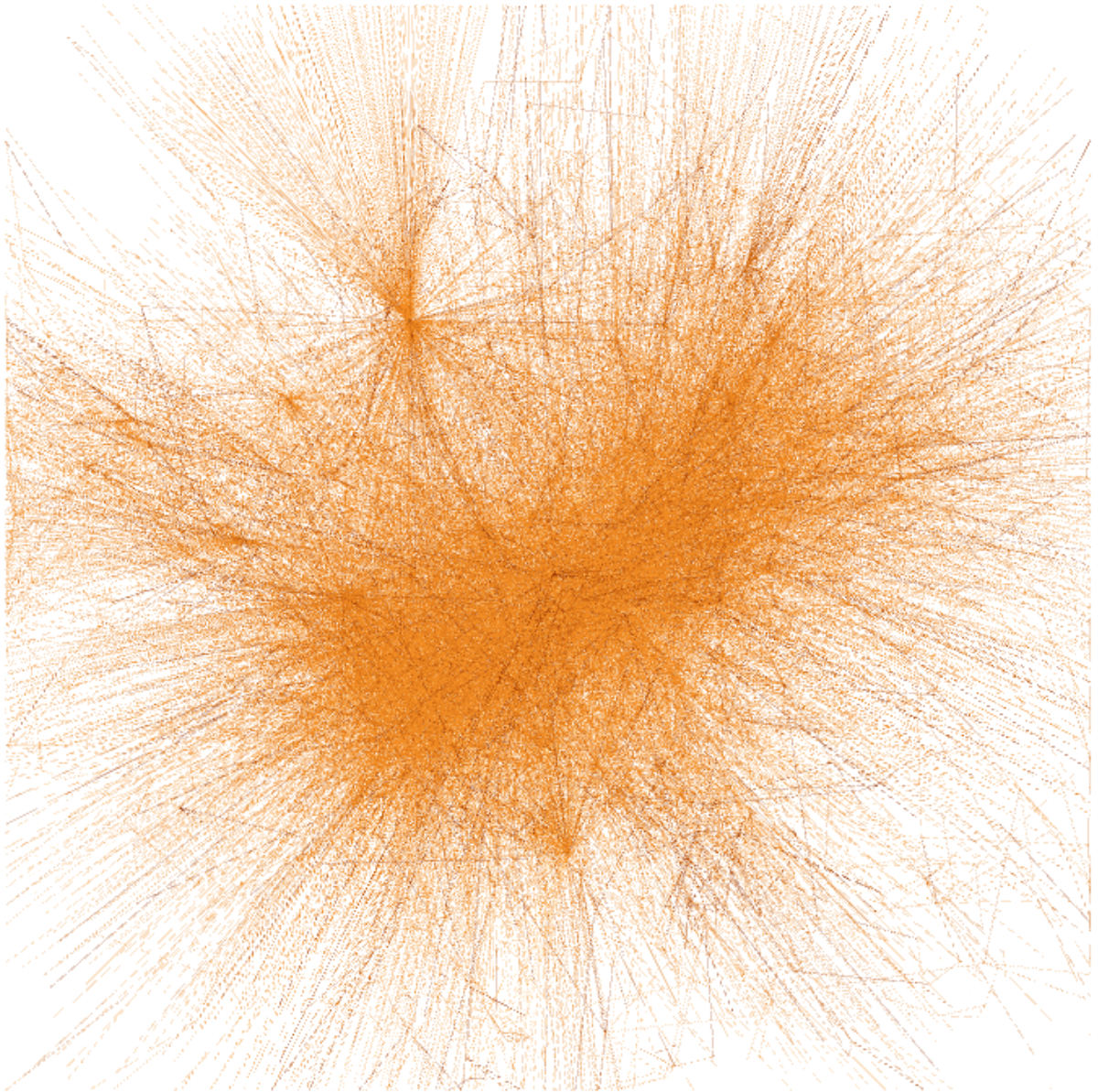
Справка

Obtain a graph layout of at least two different types

Yifan Hu



Fruchterman Reingold



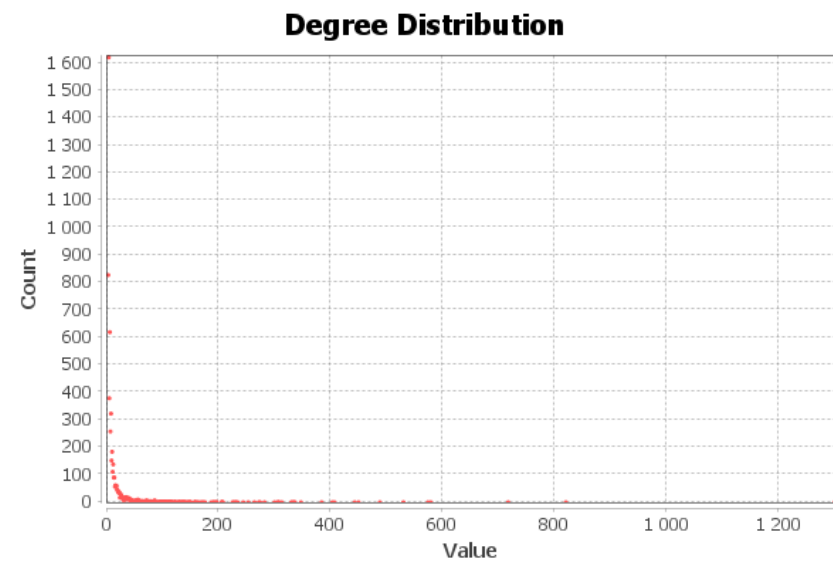
Network measures in Statistics

Degree

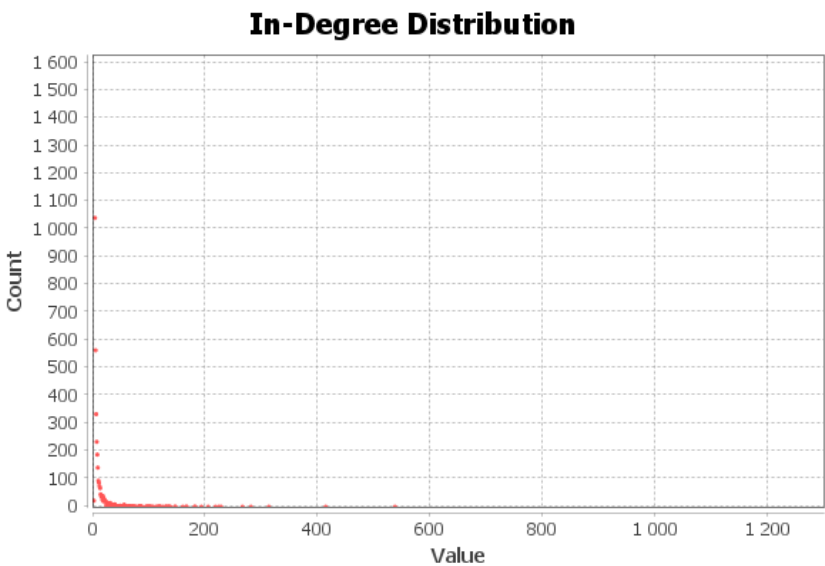
Degree - 6.052

Results:

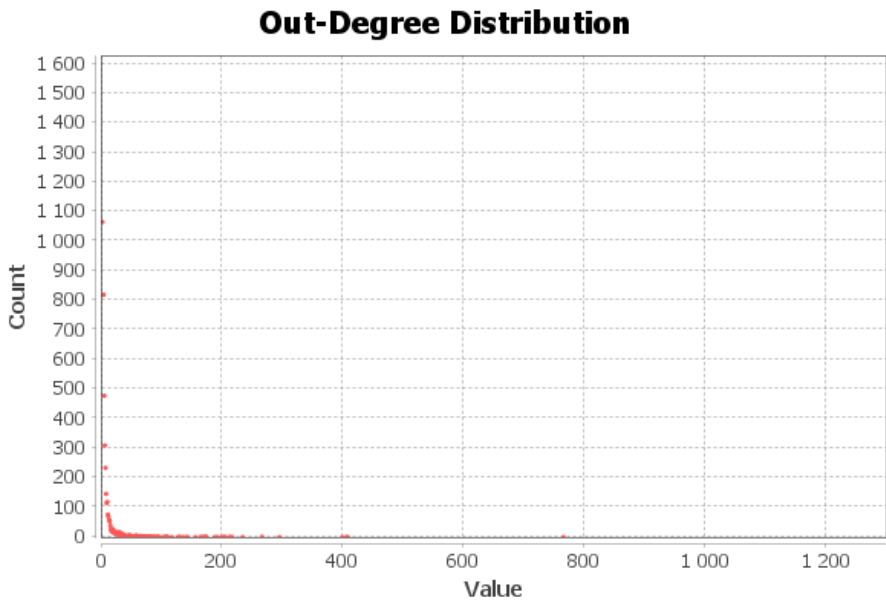
Average Degree: 6,052



In-Degree

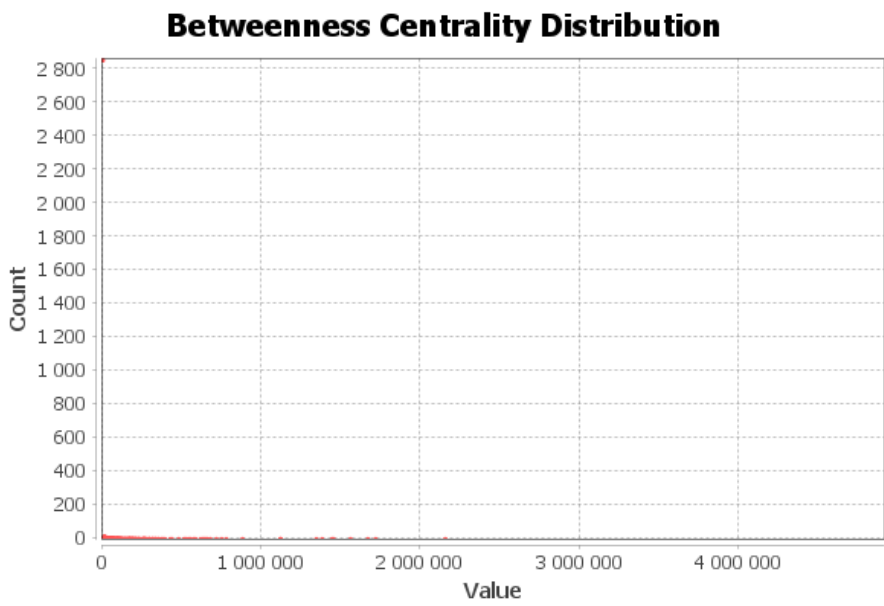


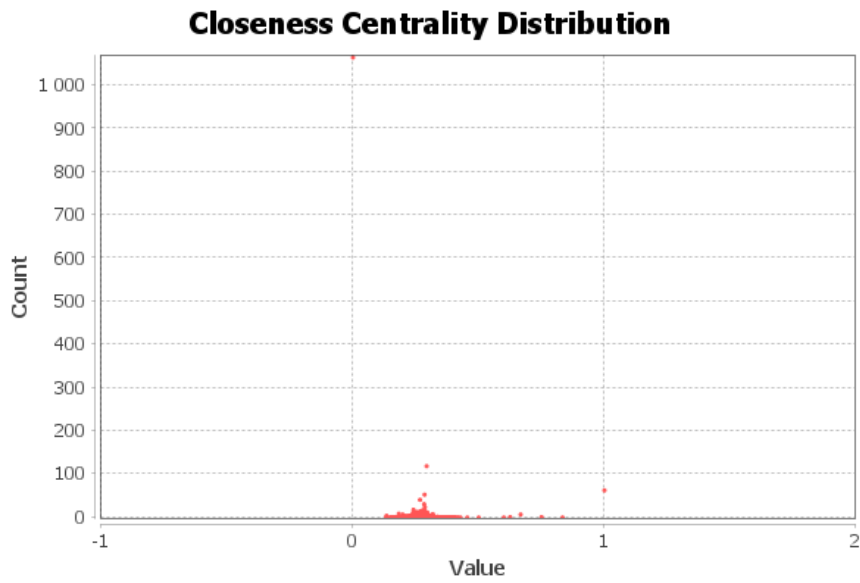
Out-Degree



Diameter

Diameter - 11
Average Path length: 3.7189





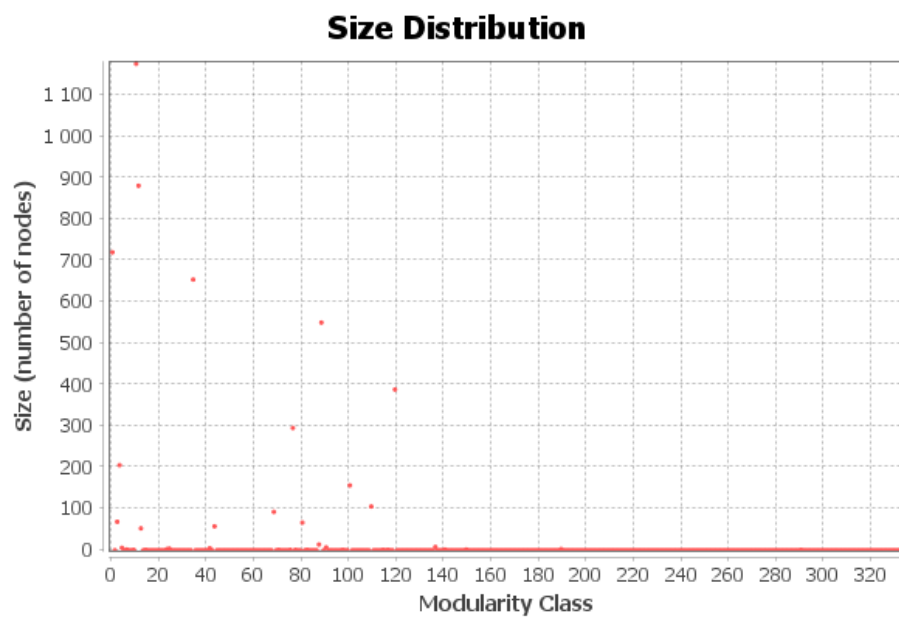
Graph Density

Density - 0.001

Modularity

Modularity - 1,402

Number of Communities - 334



Average Clustering Coefficient

Average Clustering Coefficient - 0,149

Conclusion

During this laboratory network of members ratings was analyzed. As a result we can conclude that the number of clusters is 334. Average number of marks is ~6 and the average mark is 6.125. Diameter of graph is 11, it means that maximum distance between users is 11. Average path length is 3.719. therefore members often make transactions with different members compared to past transactions. This graph is sparse due to its density, which is equal to 0.001. Based on average clustering coefficient, we can conclude that transaction scores are evenly distributed among users.