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

Alexander Yamoldin

J4134c

Accepted by

Dr Petr Chunaev

St. Petersburg

2021

Goal

The use of the network analysis software Gephi

Formulation of the problem

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.*

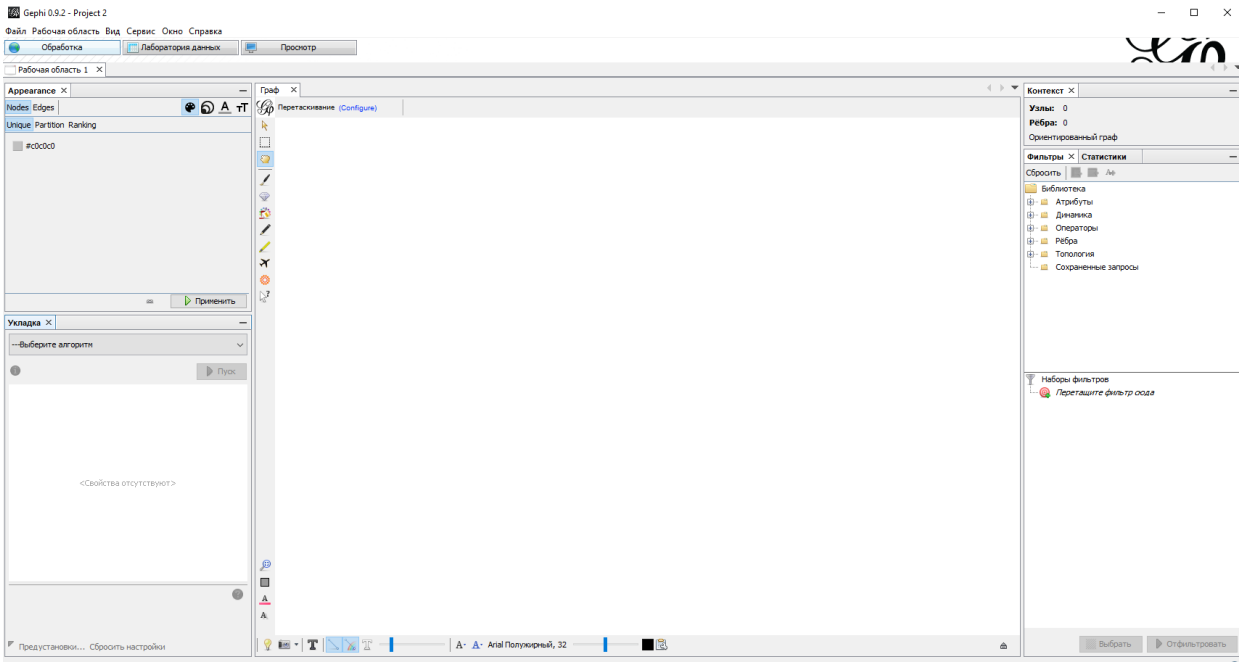
5. *Obtain a graph layout of at least two different types.*

6. *Calculate available network measures in Statistics provided by Gephi.*

7. *Analyze the results for the network chosen.*

Results and theoretical part within examples

Gephi is an interactive visualization and exploration platform for all kinds of networks and complex systems, dynamic and hierarchical graphs.
Gephi is a tool for people that have to explore and understand graphs. Like Photoshop but for graphs, the user interacts with the representation, manipulate the structures, shapes and colors to reveal hidden properties. The goal is to help data analysts to make hypothesis, intuitively discover patterns, isolate structure singularities or faults during data sourcing. It is a complementary tool to traditional statistics, as visual thinking with interactive interfaces is now recognized to facilitate reasoning. This is a software for Exploratory Data Analysis, a paradigm appeared in the Visual Analytics field of research.



First of all we need to find dataset with most of 10 000 number of nodes. We decide to use Dataset which describes a collaborations between authors papers submitted to a General Relativity and Quantum Cosmology categoty.

STANFORD UNIVERSITY

General Relativity and Quantum Cosmology collaboration network

Dataset information

Arxiv GR-QC (General Relativity and Quantum Cosmology) collaboration network is from the e-print arXiv and covers scientific collaborations between authors papers submitted to General Relativity and Quantum Cosmology category. If an author *i* co-authored a paper with author *j*, the graph contains a undirected edge from *i* to *j*. If the paper is co-authored by *k* authors this generates a completely connected (sub)graph on *k* nodes.

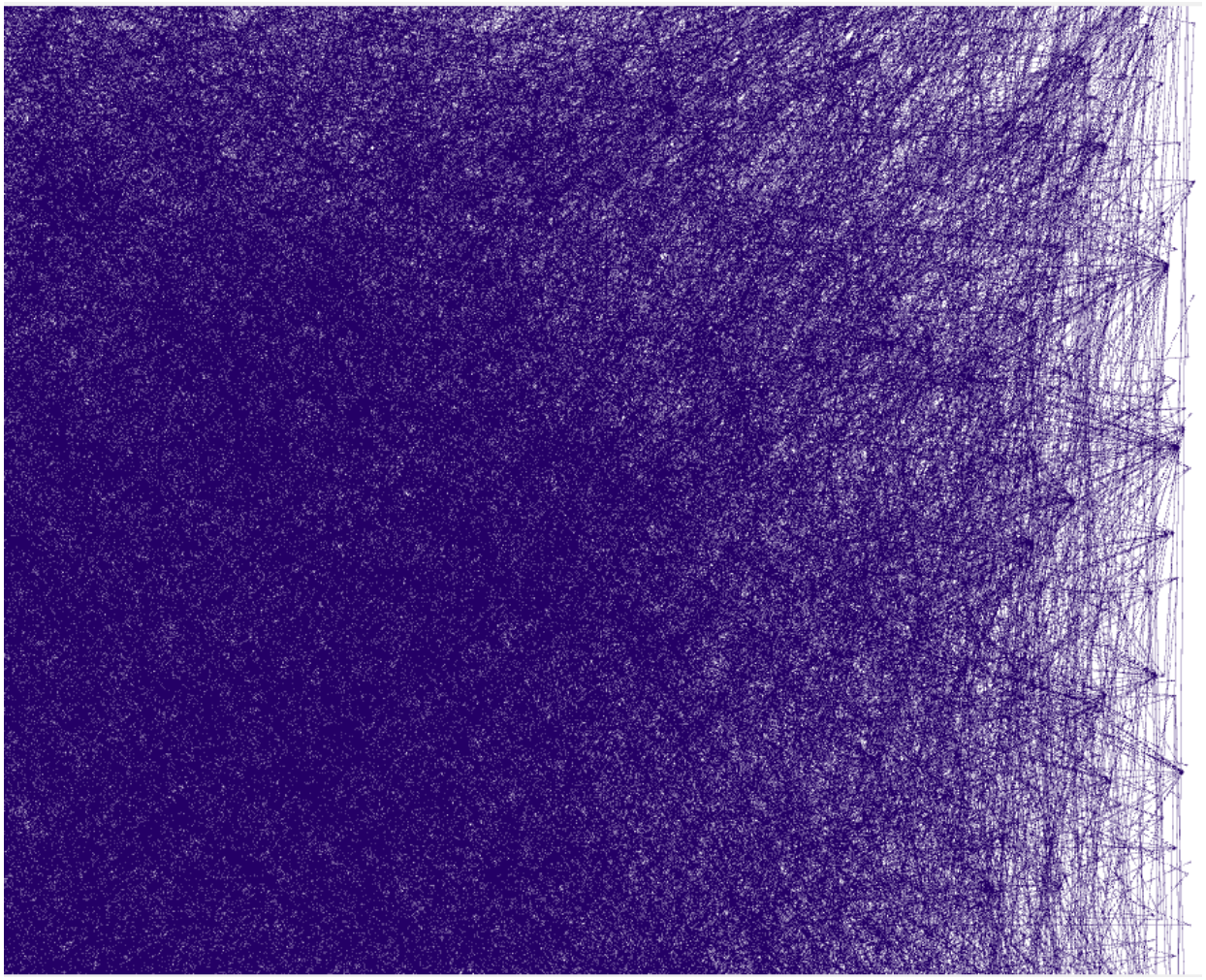
The data covers papers in the period from January 1993 to April 2003 (124 months). It begins within a few months of the inception of the arXiv, and thus represents essentially the complete history of its GR-QC section.

Dataset statistics	
Nodes	5242
Edges	14496
Nodes in largest WCC	4158 (0.793)
Edges in largest WCC	13428 (0.926)
Nodes in largest SCC	4158 (0.793)
Edges in largest SCC	13428 (0.926)
Average clustering coefficient	0.5296
Number of triangles	48260
Fraction of closed triangles	0.3619
Diameter (longest shortest path)	17
90-percentile effective diameter	7.6

Source (citation)

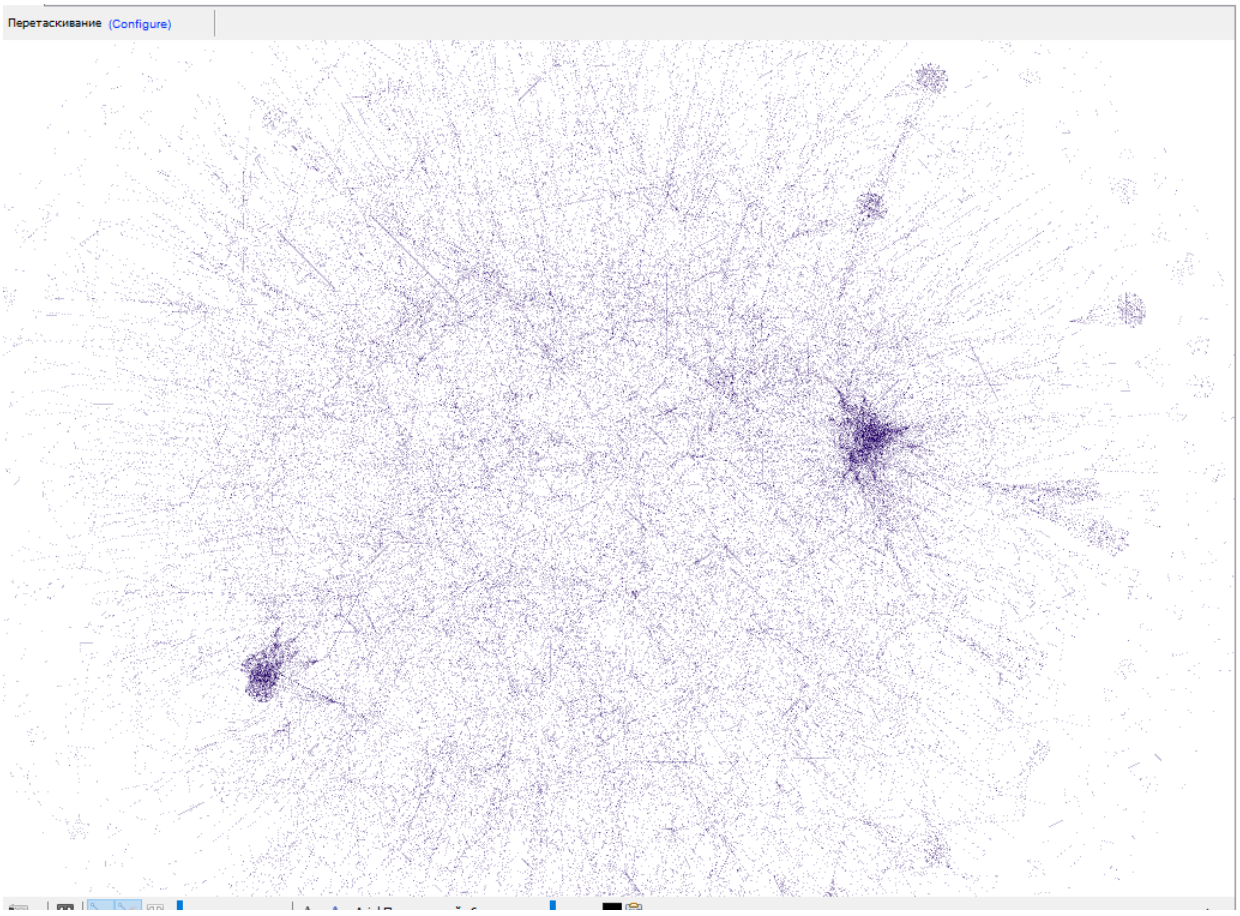
- J. Leskovec, J. Kleinberg and C. Faloutsos. Graph Evolution: Densification and Shrinking Diameters. ACM Transactions on Knowledge Discovery from Data (ACM TKDD), 1(1), 2007.

Upload and visualize the dataset



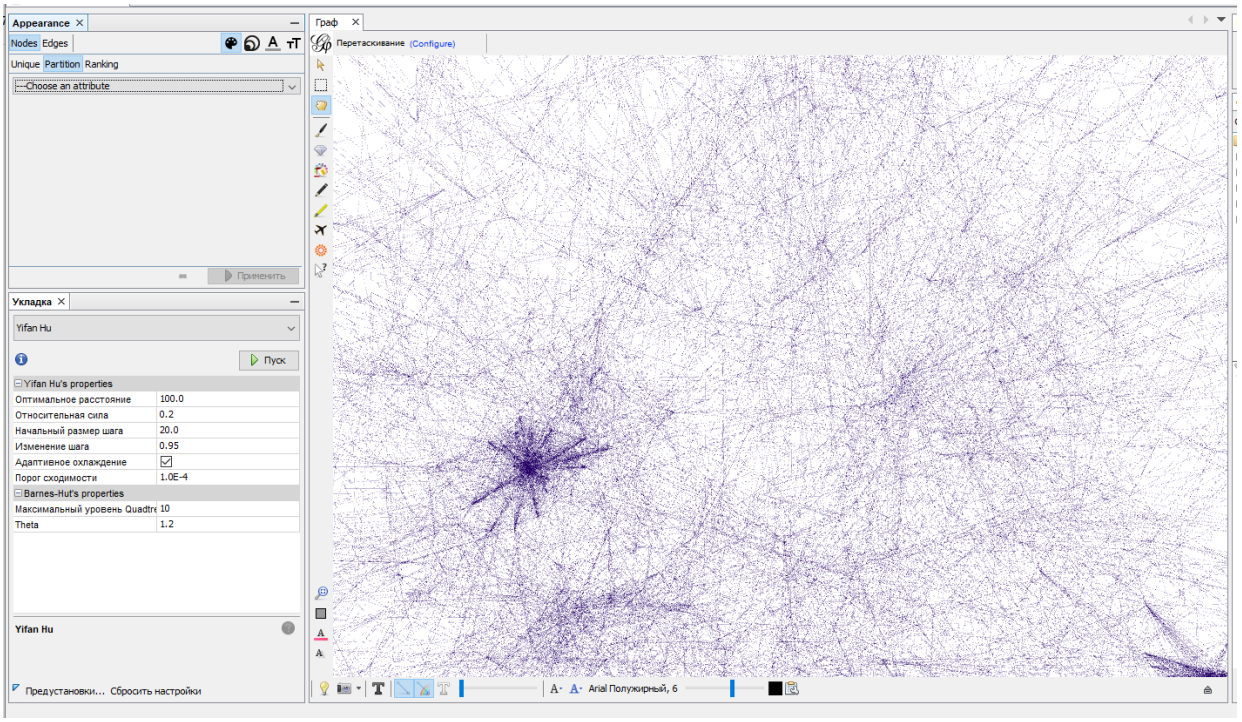
Obtain a graph layout by Frusherman-Reingold's algorithm.

Algorithm Frusherman-Reingold was develop in 1991. Complexity of algorithm is $O(N^2)$. Algorithm is from Forced-Directed family. Algorithm uses spring physical model in which nodes defines as components of system and edges as springs. Forces can doing only on nodes and weight of springs is ignore. The result of algorithm we can see below:



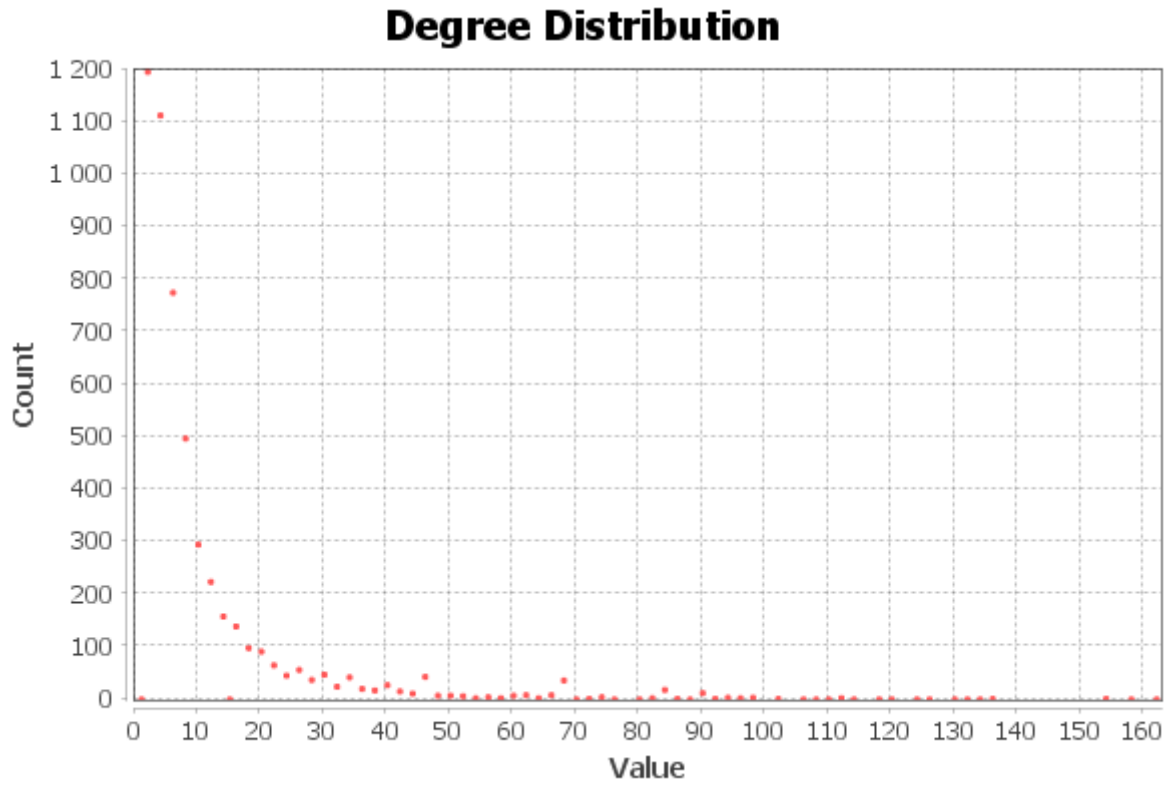
The next Algorithm is Yifan Hu's algorithm.

Algorithm Yifan Hu was develop in 2005. Complexity of algorithm is O(N * log(N)) so the algorithm works significantly faster than Frusherman-Reingold's algorithm. Repulsive forces of nodes evaluates with using forces of repulsive clusters from nodes. With the increasing of efficiently algorithm less the precision in visualizations. Another feature of the algorithm is it ability to return by reaching amplitude of fluctuations lower than threshold. The result of algorithm is on picture below:

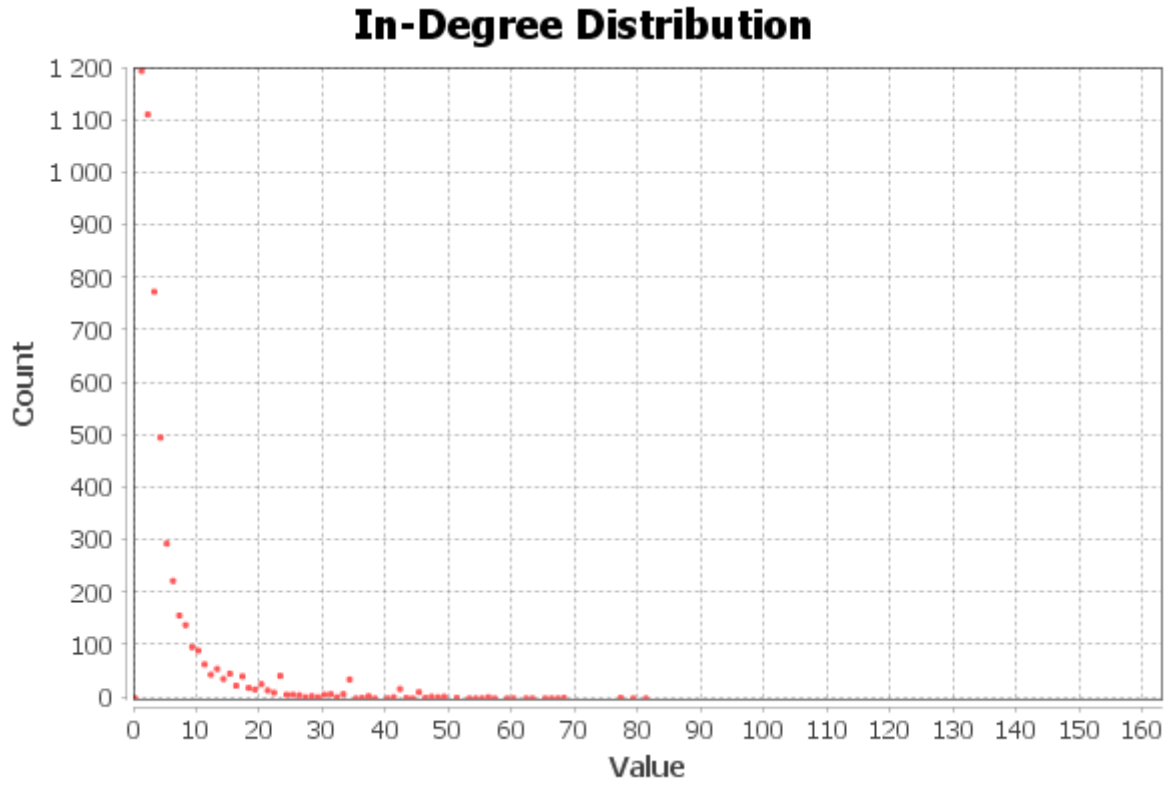


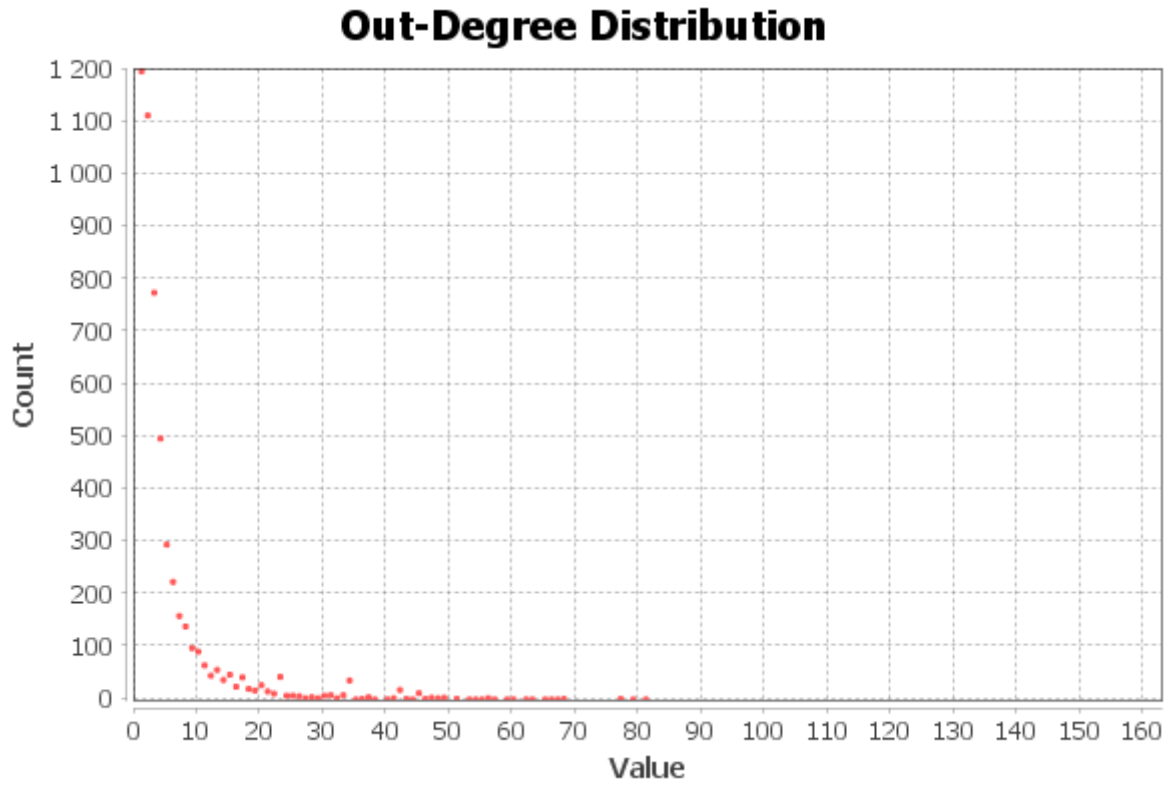
Analysis statistics

Degree – quantity links of node. We see a power law degree distribution of a network. We can see that some authors are hubs: with degrees in hundreds. Such degree distribution is said to have a long tail.
Average Degree is 5.527. It is means that each author has in average 5 collaborations with other authors.



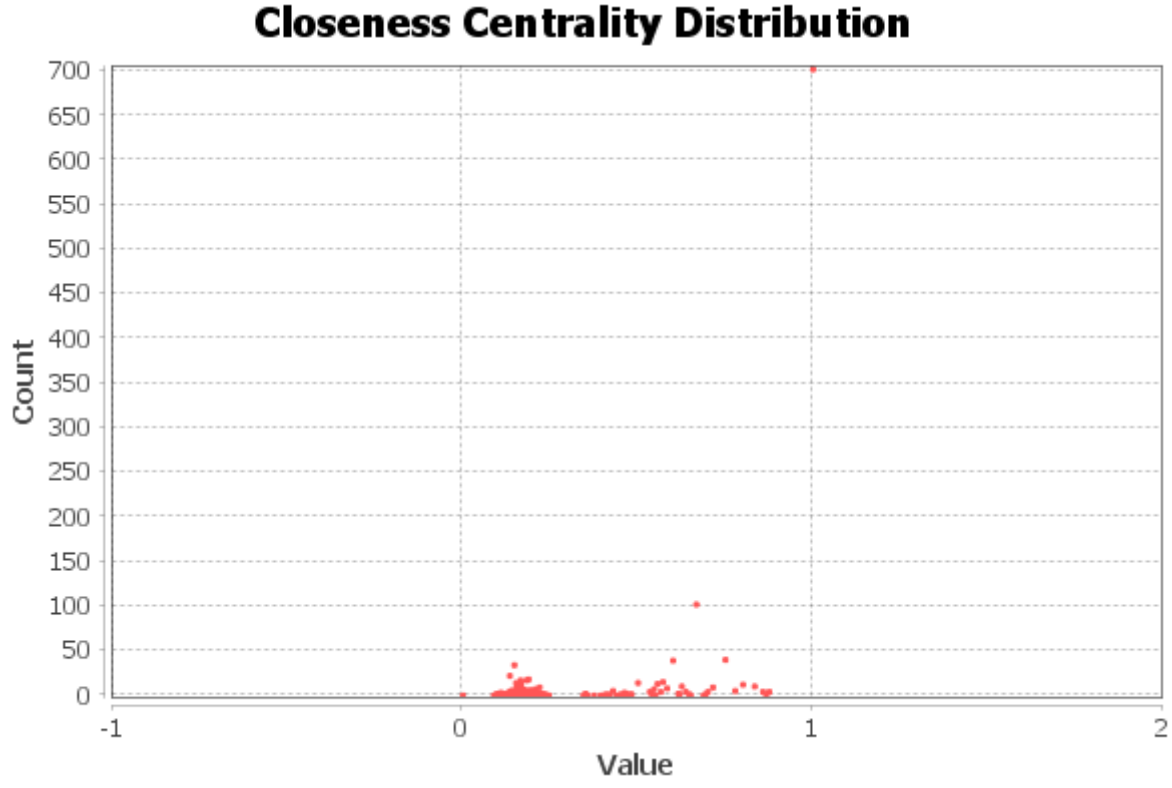
Comparing Indegree and Outdegree distribution shows the same results. It is conformed with theoretical concepts because our graph is undirected



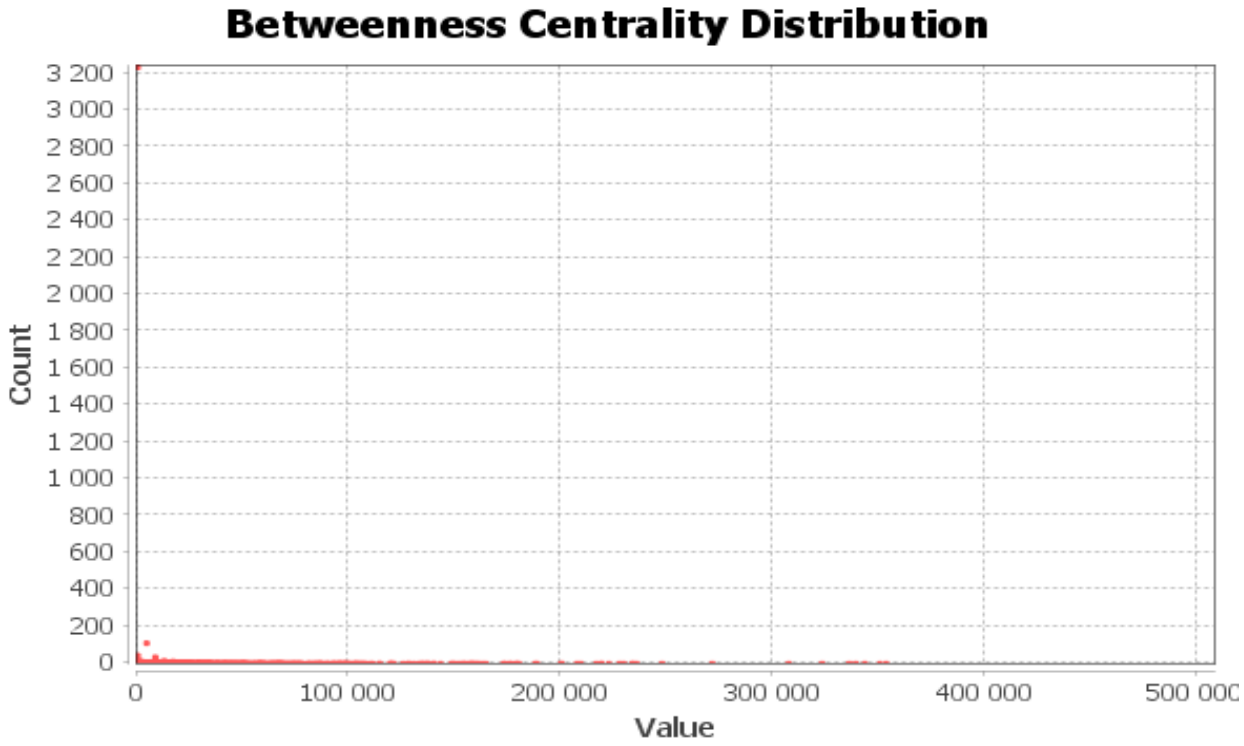


There are several methods to define ***node importance****:*

Closeness centrality – of a node is a measure of centrality in a network. The main idea is than node is closer to center than node is more important

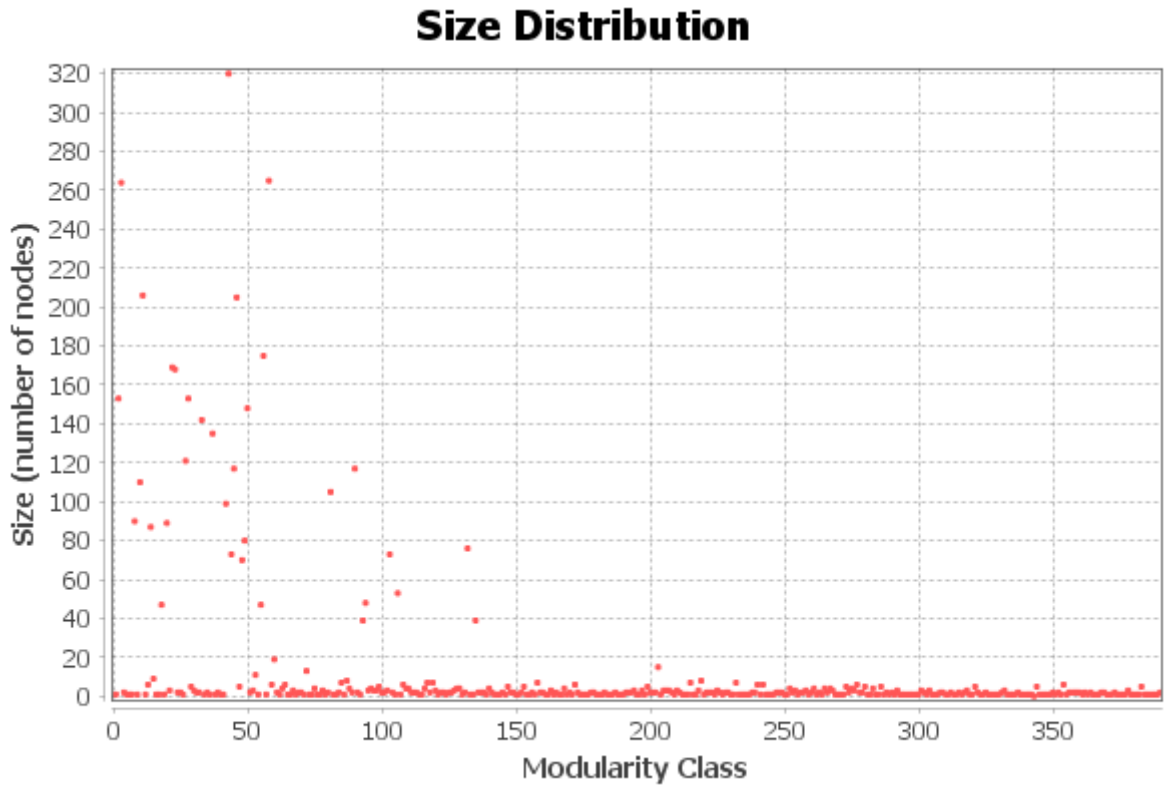


Betweenness centrality - for each vertex is the number of these shortest paths that pass through the vertex.
Diameter: 17
Average Path length: 6.048665544579825



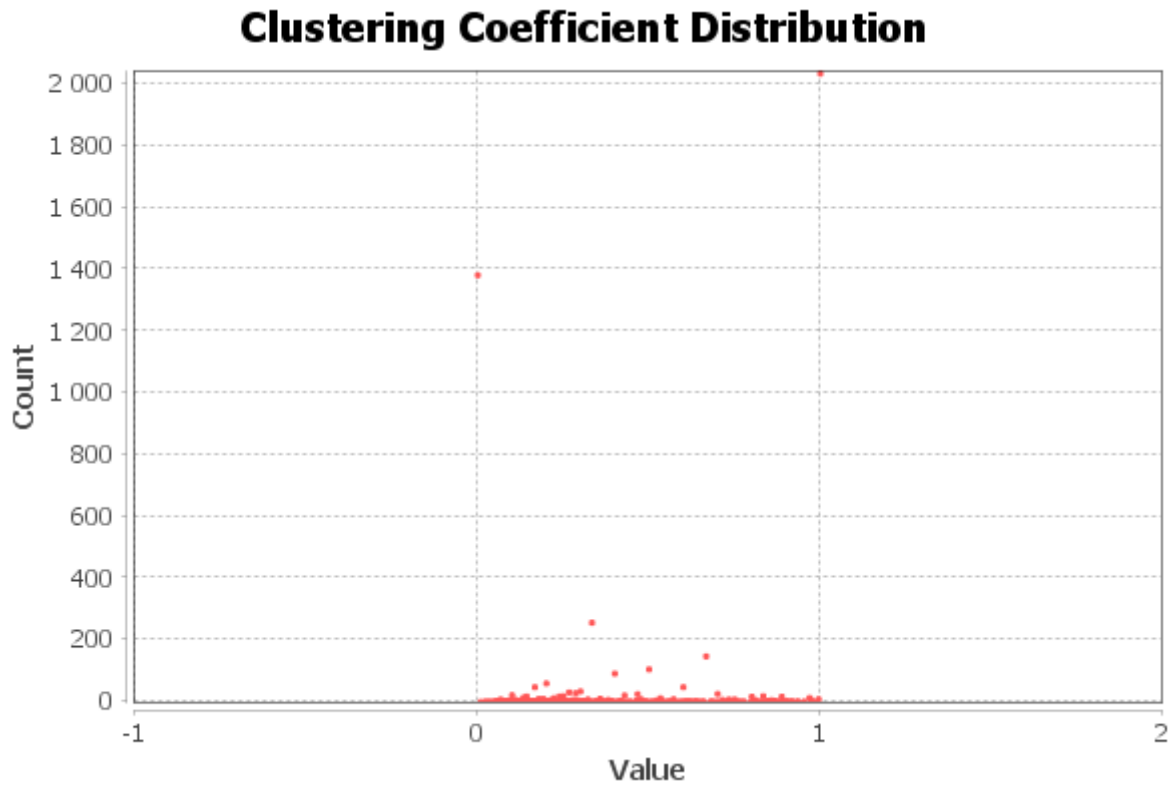
Modularity – shows how much on the given distribution density of the links between group bigger than density between groups. With that metric the graph divides on clusters

Modularity: 0,856
Number of Communities: 390

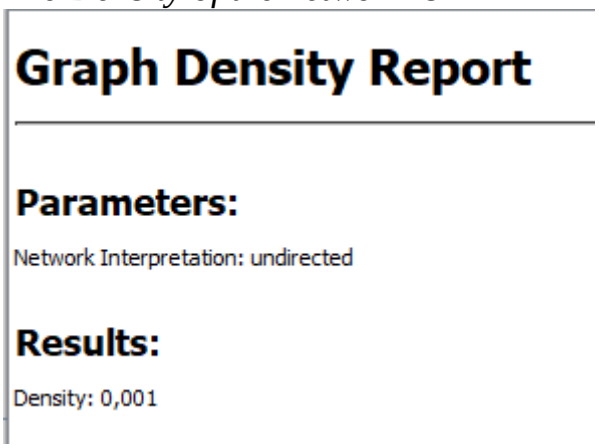


Clustering coefficient – amount of interactions between near neighbors of node. This is probability of links near neighbors of node between each other.

Average Clustering Coefficient: 0,687



The Density of the network is



Conclusions

In this lab we analysis collaboration network dataset. The results are: the graph was divided into 390 clusters. Average Degree is 5.527. It is means that each author has in average 5 collaborations with other authors. The averagepath length is 6, which means that authors collaborate with each other quite effectively. The diameter of graph is 17 it represents maximum distance between authors. This result means that General Relativity and Quantum Cosmology community is so tight. The density of the graph is 0.001, which indicates that it is sparse

Appendix

https://github.com/AAYamoldin/TrainingPrograms/blob/master/Python/ITMO_algorithms_lab/Task_7/project_1.gephi