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
Zakhar Pinaev
J4132c
Accepted by
Dr Petr Chunaev

St. Petersburg
2021

Goal

The use of the network analysis software Gephi.

Problems and methods

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

Graph theory studies the relationships between objects in a group. Graph can be visualized the as a series of interconnected points, each of which represents a member of the group, for example, people on a social network. Lines drawn between the dots represent connections between participants, such as friendships on a social network. Analyzing graphs helps to identify things like the influence of a particular member on others, or who has more friends from two members of a group.

Gephi software has a set of graph tools. First of all, these are tools for visualization, for example, various types of layouts for network graphs. In addition, the software provides a number of statistics that can be calculated for the network under study.

The degree of a vertex is the number of edges in the graph to which this vertex belongs. Gephi allows to calculate the average degree of the vertices of a graph. To find the diameter of a graph, first find the shortest paths between all pairs of vertices. The longest shortest path is the diameter of the graph. The density of the graph reflects the degree of closeness of the graph to the complete one. Modularity was developed to measure the strength of a network partitioning into modules. High modularity networks have tight connections between nodes within modules, but weak connections between nodes in different modules. A connected component is a set of graph vertices, between any pair of which there is a path. The clustering factor is the clustering values for all nodes in the graph. When the clustering coefficient is high, it means that the graph is extremely densely clustered around several nodes; when it is low, it means that the connections in the graph are relatively evenly distributed among all nodes. The average path length is the average number of steps along the shortest paths for all possible pairs of network nodes.

Results

First of all, a network was selected for further visualization and analysis. It became an arxiv GR-QC (General Relativity and Quantum Cosmology) collaboration network, which 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 an 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). The network itself consists of 5242 nodes and 14496 edges. First of all, the data about this network was loaded into Gephi. Figure 1 shows the network graph before processing.

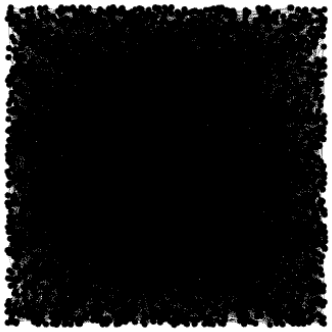

Figure 1 – graph obtained by network data before processing

Next, work was done with layouts for the network graph. First of all, the Fruchterman Reingold layout was obtained, the result of which is shown in Figure 2.

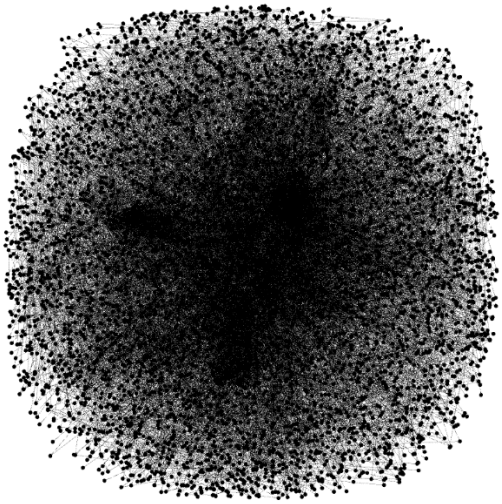

Figure 2 – Fruchterman Reingold layout for the graph of the network under study

Figure 2 shows that this layout made it possible to spread the graph vertices to the sides, which makes it easier to view in comparison with Figure 1 but does not provide additional information about the network structure. Therefore, further it was decided to obtain the Force Atlas layout for the graph of the network in question. The visualization of the resulting layout is shown in Figure 3.

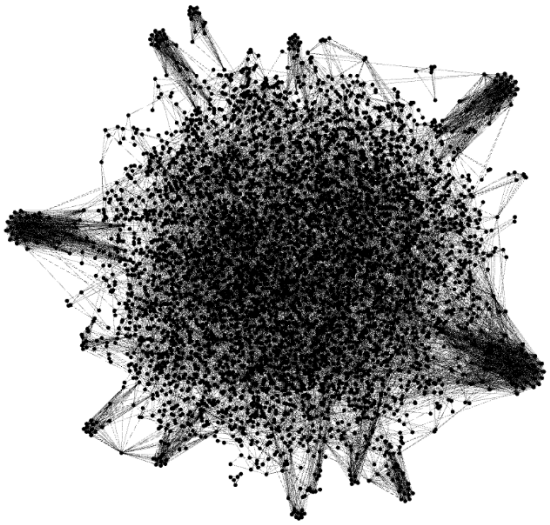

Figure 3 – Force Atlas layout for the graph of the network under study

Looking at Figure 3, more can be said about the structure of the network, for example, it can be seen that there are groups of scientists who closely interacted with each other, which is visually reflected on the graph in the form of clusters of points along the edges of the graph. However, in this graph, the vertices are too close to each other, so it was decided to combine the results of the two considered layouts. Figure 4 shows a graph obtained using the Fruchterman Reingold and Force Atlas layouts.

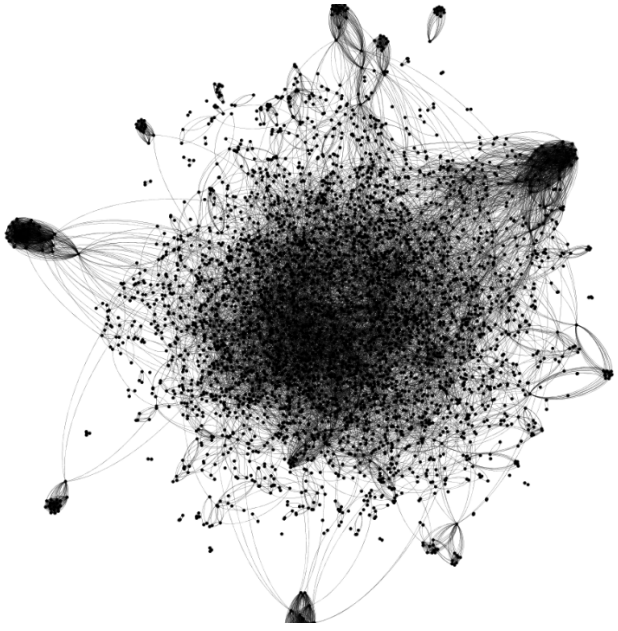

Figure 4 – Fruchterman Reingold and Force Atlas layouts for the graph of the network under study (the layout of the network graph is presented with curved edges as an experiment – then straight edges were used)

Then it was decided to change the size of the vertices depending on the number of edges to which the vertex belongs. Thus, the vertices of the authors with the most interactions will have a larger size and vice versa. Figure 5 shows the final graph with different sizes of vertices.

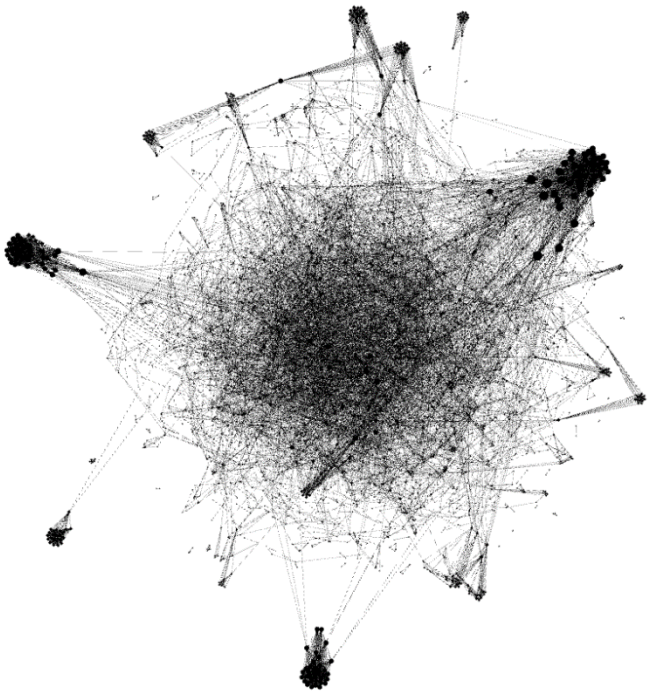

Figure 5 – the final view of the graph of the network under study, in which the size of the vertices reflects the degree of interaction of the author with other scientists

Thus, looking at the graph in Figure 5, it can be said that in the scientific community on this topic there are scientists who are in close cooperation (research teams), which visually looks like a cluster of large dark dots along the edges of the graph. In these teams, obviously, the degree of interaction with other scientists is higher than among scientists working independently (their vertices have the smallest size in the graph).

Further, for the studied network, the available statistical indicators were calculated. The calculated indicators were then compared with the given network parameters in the source. The comparison results are shown in Table 1.

Table 1 – Comparison of calculated and given in the source network indicators

	Average degree	Diameter	Density	Modularity	Connected components	Clustering factor	Average path length
Calculated indicators of the network	5.528	17	0.001	0.858	355	0.53	6.049
Indicators of the network from source	-	17	-	-	-	0.53	-

It follows from Table 1 that although the source does not contain many network parameters, all the available ones exactly coincide with the calculated ones, which indicates the correct operation with the network data.

As for the parameters themselves, then, for example, it was expected that the density of the graph would be rather low, since density expresses the closeness of the graph to complete. First, one can logically conclude that all subject scientists cannot write articles in conjunction with each of the others. Secondly, looking at the graph, it can be seen that there are much more small points than large ones, which suggests that there are many more scientists with a small number of interactions than closely cooperating scientists, therefore the graph is far from complete.

Further, for example, to understand why the network has a high degree of modularity, the graph can also be used. On the graph it can be seen that along the border of the graph there are groups of vertices – scientific teams that have close connections with each other (as can be seen from the size of these vertices). And outside these groups, the connections become less dense, and the vertices are much smaller.

Finally, the average value of the clustering factor is due to the fact that the network contains both grouped vertices with close connections and a large number of evenly distributed connections.

In addition to the calculated values, graphs of the distribution of degrees and the distribution of eccentricities were plotted (Figures 6,7 respectively).

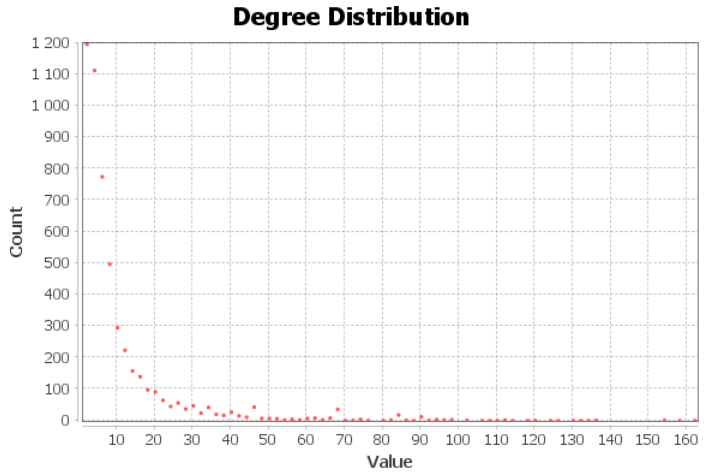

Figure 6 – degree distribution of the network under study

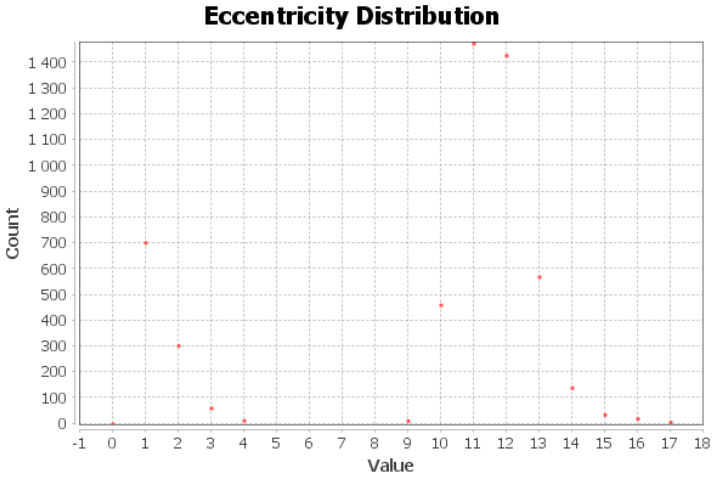

Figure 7 – eccentricity distribution of the network under study

As can be seen in Figure 6, and as mentioned earlier, most of the vertices (scientists) had no more than 20-30 connections (interactions with other scientists). According to Table 1, the average number of interactions is 5.528. The data from Figure 7 again shows that the mesh diameter is 17.

Conclusion

As a result of the work, an acquaintance with the Gephi software tools for network analysis was carried out. As an assignment, an analysis of the arxiv GR-QC collaboration network was carried out, which covers scientific collaboration between the authors of articles presented in the categories of general relativity and quantum cosmology. The processing, visualization and analysis of network data made it possible to find out that there are scientists in the scientific community on this topic who are in close cooperation (research groups). In these teams, it is evident that the degree of interaction with other scientists is higher than among scientists working independently.

In addition, the calculated statistical parameters completely matched the network parameters presented in the source, which indicates correct work with the data.