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

*Vdovkina Sophia*

*Syrchenko Arina*

*Academic group: J4133c*

Accepted by

Dr Petr Chunaev

St. Petersburg

2022

**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.

**Brief theoretical part**

Gephi is an open-source network analysis and visualization software package written in Java on the NetBeans platform. The user interacts with the representation, manipulate the structures, shapes and colors to reveal hidden patterns. 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.

Social network analysis (SNA) is the process of investigating social structures through the use of networks and graph theory. It characterizes networked structures in terms of nodes (individual actors, people, or things within the network) and the ties, edges, or links (relationships or interactions) that connect them.

The *degree of a vertex* in a graph is the number of edges that are incident to the vertex. The *in-degree of a vertex* is the number of in-edges for vertex v, the *out-degree of vertex*, is the number of out-edges for vertex.

The *eccentricity* of is the greatest distance between and any other vertex: .

The *diameter* is the maximum eccentricity of any vertex, i.e. the greatest distance between any pair of vertices: . So, this characteristic shows how far is the node from the most distant node.

The *radius* is the minimum eccentricity of any vertex .

The *average path length*

*Modularity* Q measures the strength of division of a graph into clusters (subgraphs, modules). Graphs with high Q > 0 have dense connections between the vertices within clusters but sparse between those in different clusters. Q compares the number of edges within clusters in a graph with the expected number of edges in a random graph regardless of clusters.

The *density* ρ of an undirected G is the ratio of |E| and the number of possible edges, i.e. the number of edges in the complete graph with the same |V|.

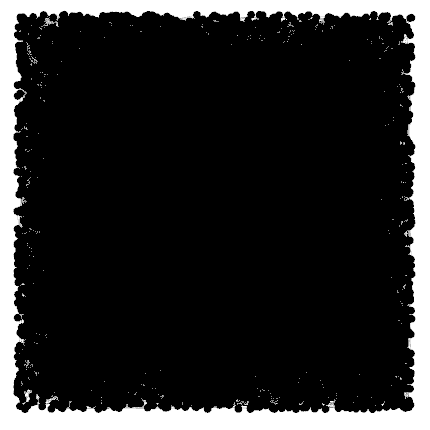
**Results**

The program gephi 0.9.7 was used to analyze the LastFM Asia Social Network, that can be found on <https://snap.stanford.edu/data/feather-lastfm-social.html>.

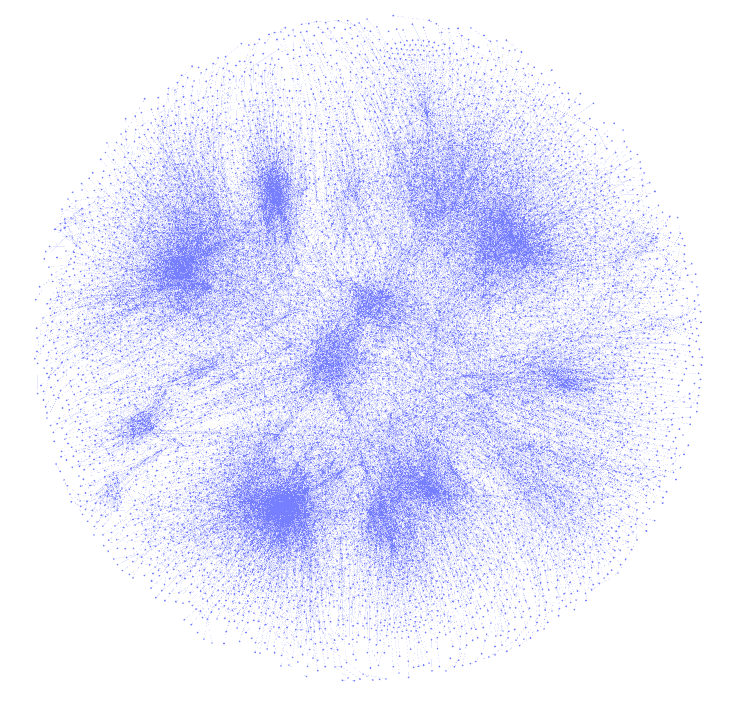
A social network of LastFM users which was collected from the public API in March 2020. Nodes are LastFM users from Asian countries and edges are mutual follower relationships between them.

This undirected graph contains 7624 nodes and 27806 edges.

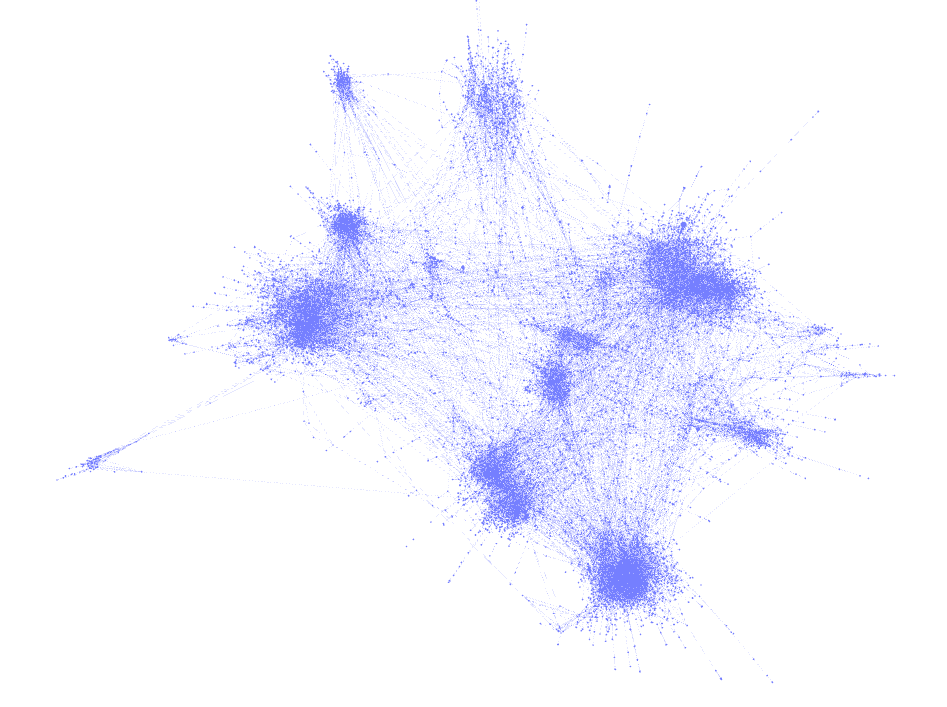
The graph (original graph is shown on picture 1) was loaded into Gephi and visualized using Fruchterman Reingold (shown on picture 2) layout and Force Atlas 2 layout (picture 3).



Picture 1 – original graph



Picture 2 – Fruchterman Reingold layout

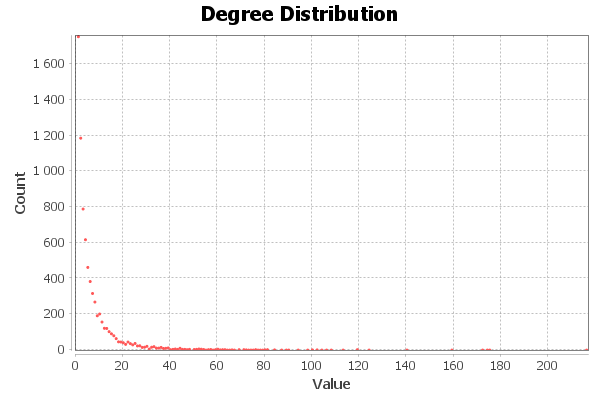


Picture 3 – Force Atlas 2 layout

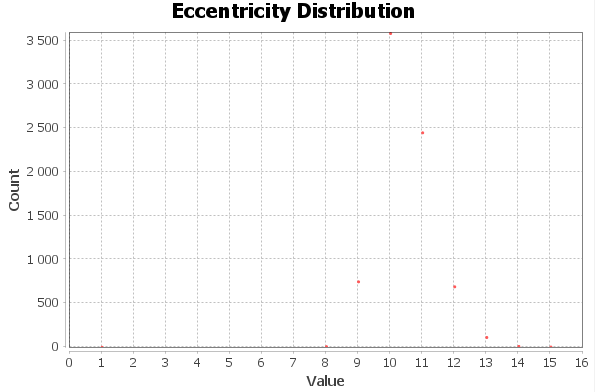
Layouts on pictures 2 and 3 are more visual and easier to examine. More can be seen about the structure of the network: for example, there are closely connected points, which is visually reflected on the graph in the form points clusters.

Further, for the studied network, the available statistical indicators were calculated.

The average degree value is 7.293. Most of the vertices have the degree below 200, but there are some outliers, for example, a very little number of people has the degree over 1600, it means they have over 1600 friends,.



The diameter of the graph is 15. The average path length is approximately 5.23. The radius equals 1. The diameter of the graph can be explained as two users were connected via at most 13 other users, while average path length is three times lower.



Density of the graph is equal to 0.001. The density expresses the closeness of the graph to complete. It expresses the fact that there is hardly anybody who can be mutual followers with everyone.

The modularity of the graph equals 0.809 and the number of communities is 21. It is understandable why the graph has a high degree of modularity. As it was concluded, while examining the graph, there some closely connected points, forming clusters, so people may form some groups and outside these groups, the connections become less dense.

**Conclusion**

As a result of the work, the LastFM Asia Social Network was analyzed, using network analysis and visualization program Gephi. This graph has several connected components, low density, but the value of modularity is quite high and positive, some of the components are connected more strongly than the others.

**Appendix**

Project file is available on <https://github.com/sophia-vdovkina/Analysis-and-development-of-algorithms/blob/main/Task%206/Task6.ipynb>