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
Mikhail Grigoryev (370852)
Semenova Valeria (370061)
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.
2. Choose a network dataset from https://snap.stanford.edu/data/ with number of nodes at most 10000. 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. It widely used both in academia and journalism as well as digital humanities such as history, political sciences and literature.

It was originally developed in the University of Technology of Compiègne in France, then was noticed by Google. In 2010 a non-profit organization called Gephi Consortium was formed for further development of the package.

Results

Steps 1-2. Gephi was installed, then, from the stanford library a social network of LastFM users collected from the public API in March 2020 was selected. Nodes are LastFM users from Asian countries and edges are mutual follower relationships between them. The vertex features are extracted based on the artists liked by the users. The task related to the graph is multinomial node classification – one has to predict the location of users. This target feature was derived from the country field for each user. Network statistics (graph is unweighted and undirected):

Parameter	Value
Nodes	7 624
Edges	27 806
Density	0.0009
Transitivity	0.1787

Steps 3-4. The initial dataset was in `.csv` format supported by Gephi. Import parameters were checked on import:

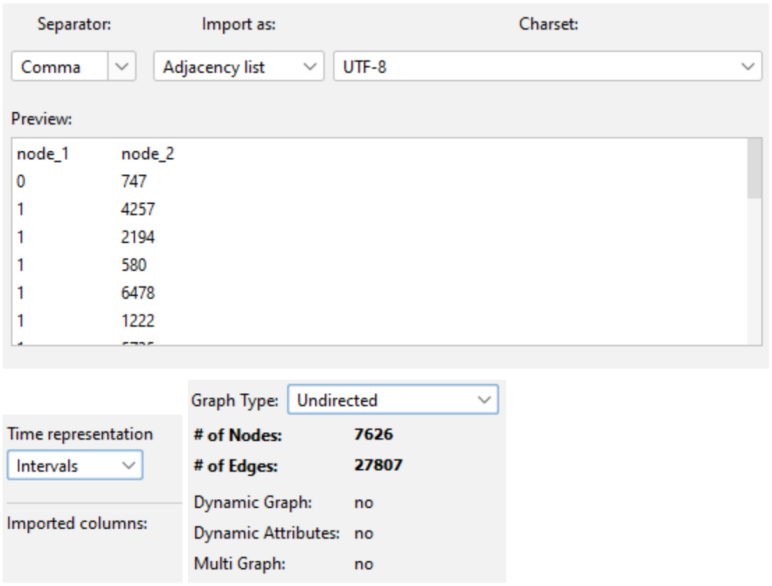


Figure 1: Import parameters.

Step 5. Two layout types were used – OpenOrd and YifanHu.

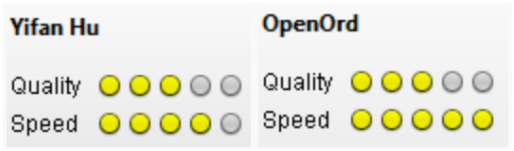
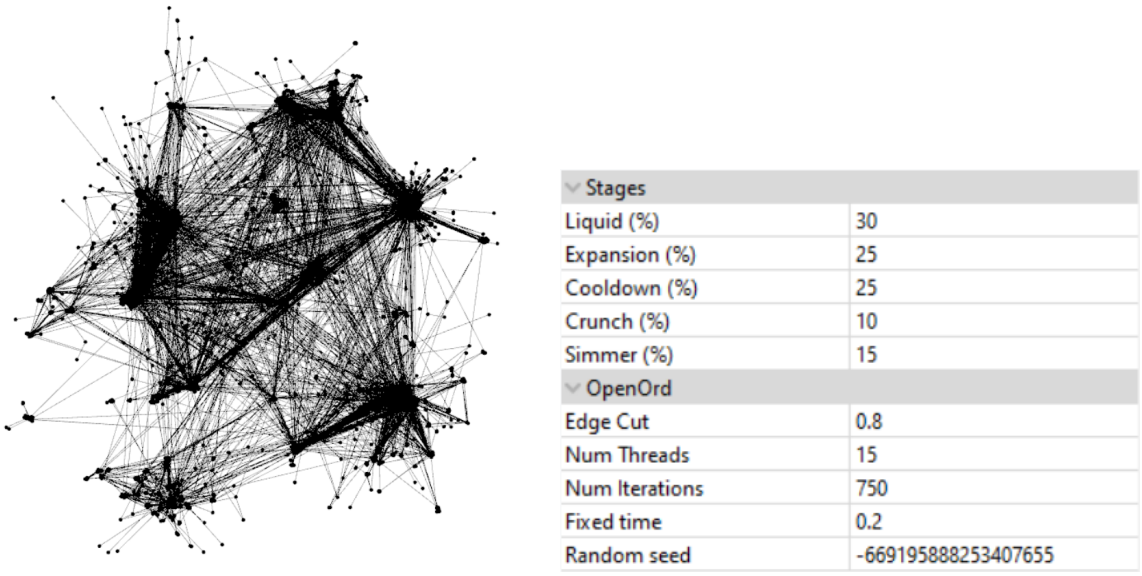


Figure 2: Visualization type evaluations from Gephi.

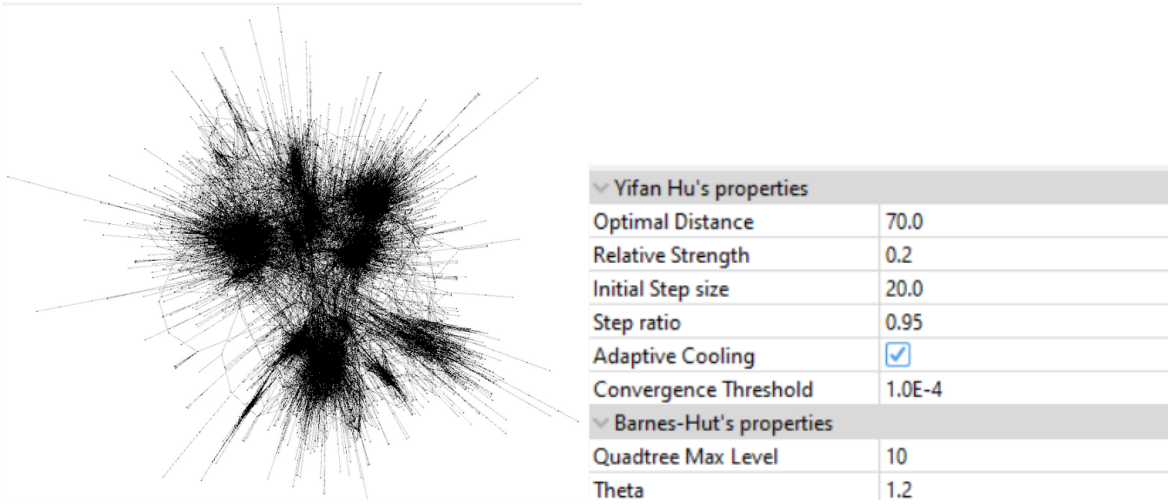
OpenOrd is Force-Directed layout algorithm for real-world large-scale undirected graphs. It can scale to over 1 million nodes, making it ideal for large graphs. However, small graphs (hundreds or less) do not always end up looking good. This algorithm expects undirected weighted graphs and aims to better distinguish clusters.



Stages	
Liquid (%)	30
Expansion (%)	25
Cooldown (%)	25
Crunch (%)	10
Simmer (%)	15
OpenOrd	
Edge Cut	0.8
Num Threads	15
Num Iterations	750
Fixed time	0.2
Random seed	-669195888253407655

Figure 3: OpenOrd layout visualization and preset.

Yifan Hu is the original Yifan Hu's attraction-repulsion model. It reduces the computational cost by restricting force calculation to the neighborhood. The algorithm stops itself, as it has an adaptive cooling scheme.



Yifan Hu's properties	
Optimal Distance	70.0
Relative Strength	0.2
Initial Step size	20.0
Step ratio	0.95
Adaptive Cooling	☑
Convergence Threshold	1.0E-4
Barnes-Hut's properties	
Quadtree Max Level	10
Theta	1.2

Figure 4: YifanHu layout visualization and preset.

Step 6. Available network measures were taken from Gephi's Statistics. Reports are presented in the figures below.

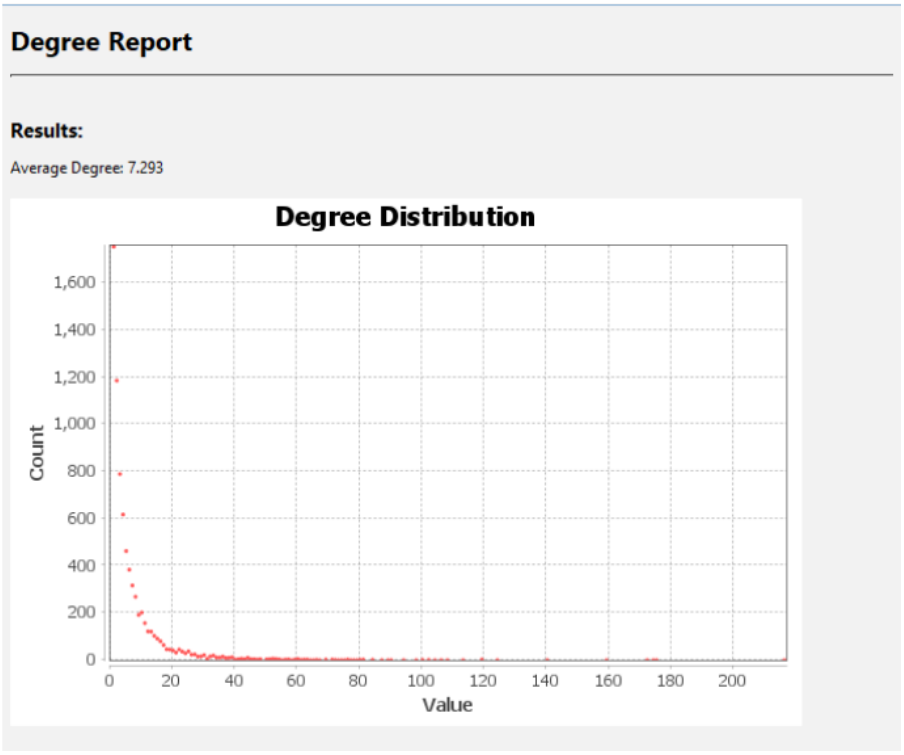


Figure 5: Statistics – average degree.

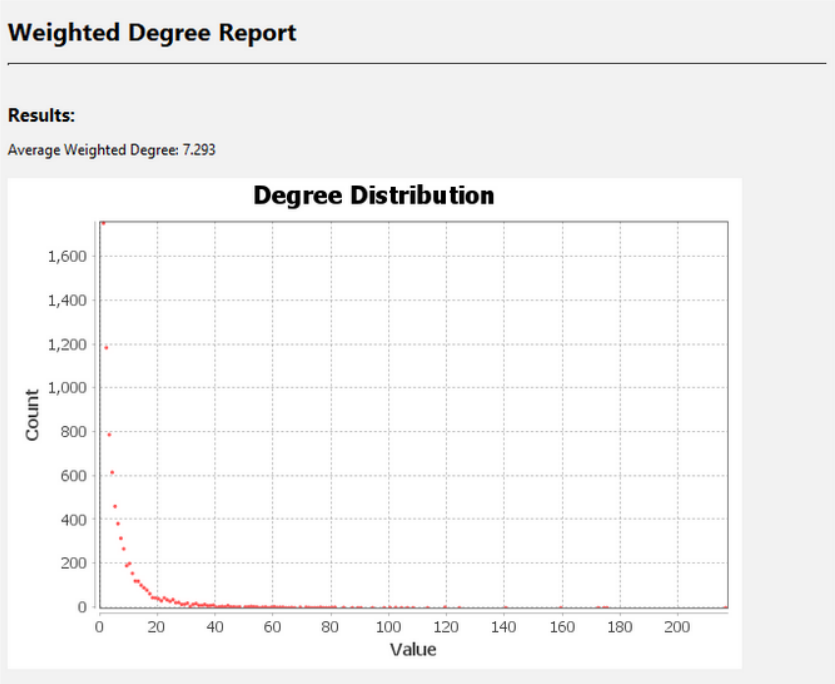


Figure 6: Statistics – average weighted degree.

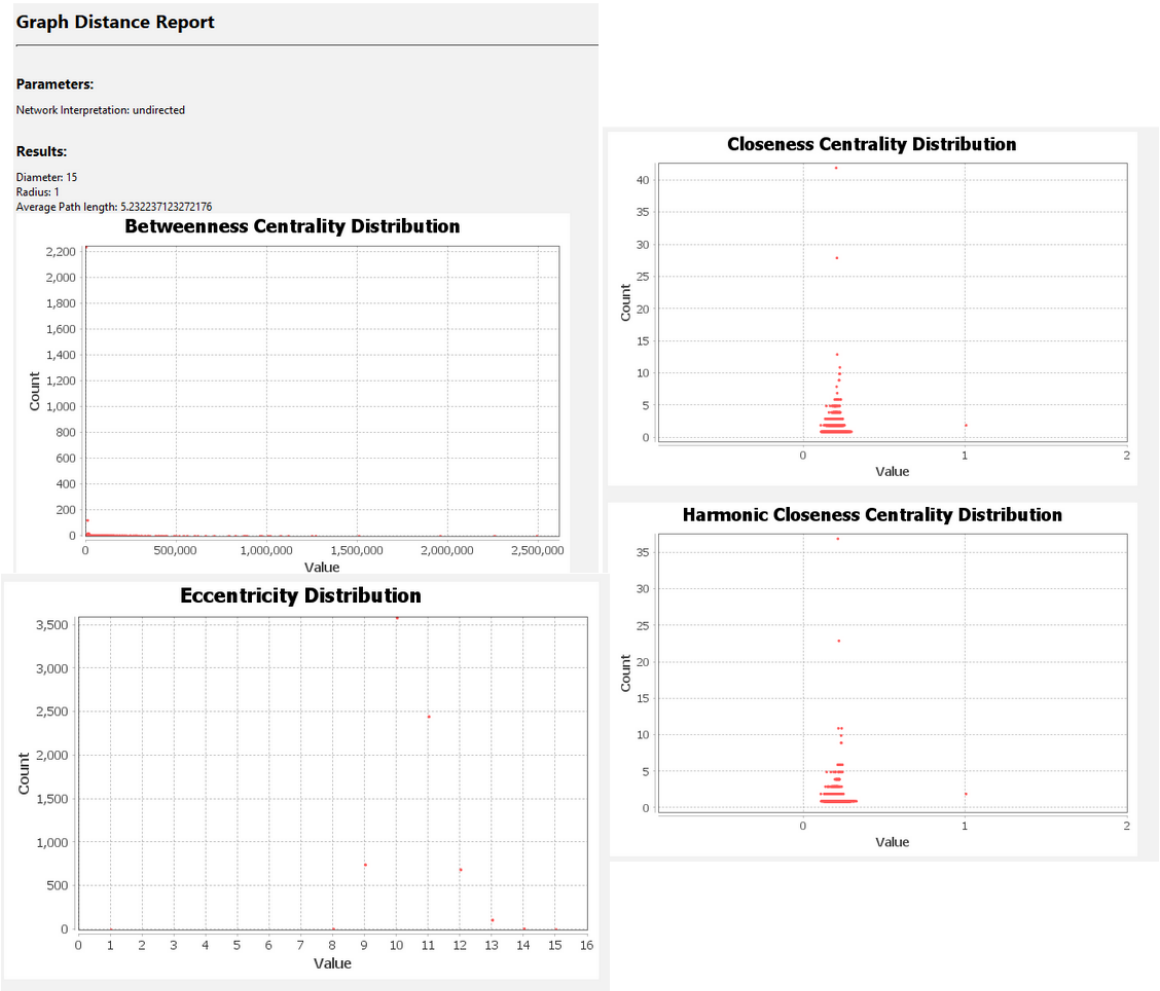


Figure 7: Statistics – network diameter.

Graph density measures how close the network is to complete. A complete graph has all possible edges and density equal to 1.

Figure 8: Statistics – graph density.

HITS computes two separate values for each node. The first value (called Authority) measures how valuable information stored at that node is. The second value (called Hub) measures the quality of the nodes links.

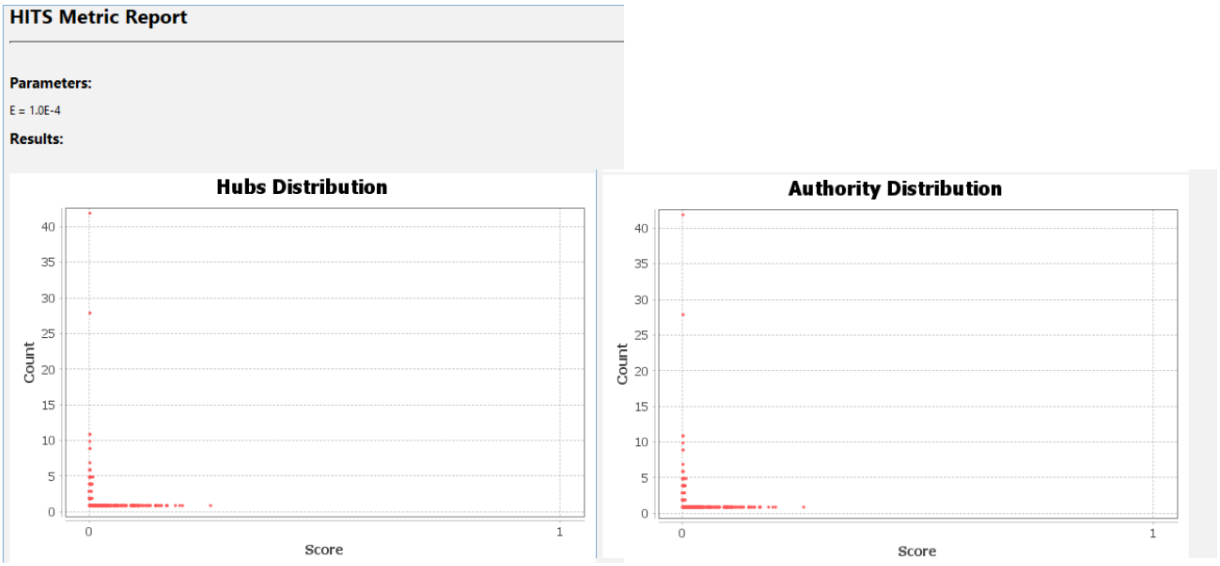


Figure 9: Statistics – HITS.

PageRank ranks nodes ”pages” according to how often a user following links will non-randomly reach the node ”page”.

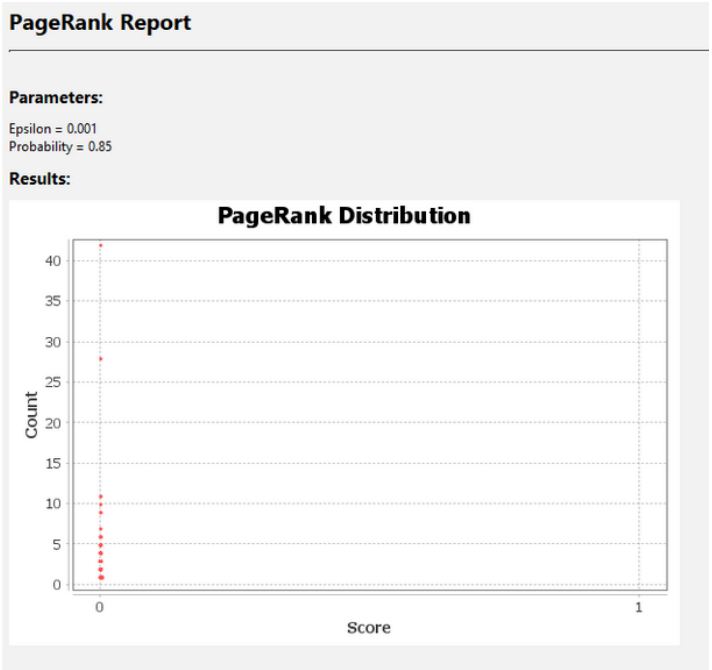


Figure 10: Statistics – PageRank.

Connected Components determines the number of connected components in the network. For undirected graph detects only weakly connected components.

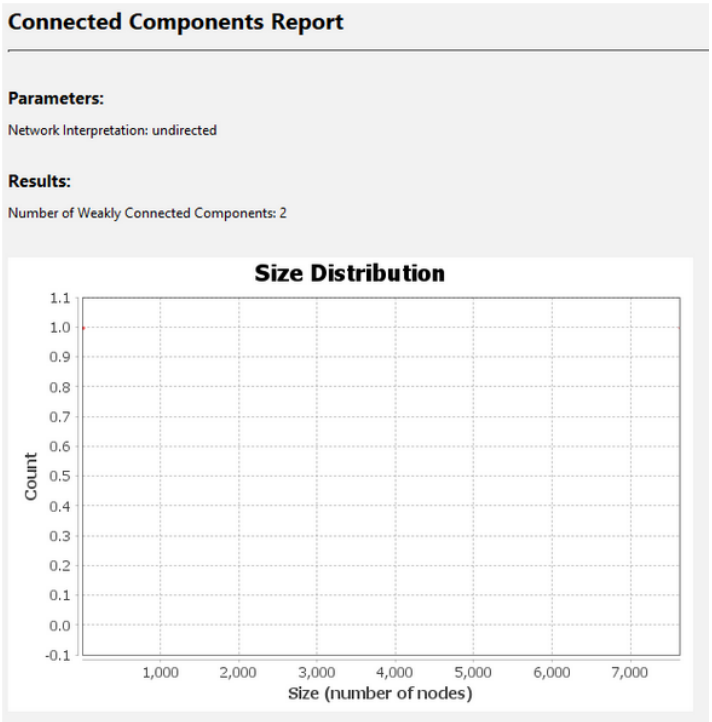


Figure 11: Statistics – connected components.

Modularity is a community detection algorithm.

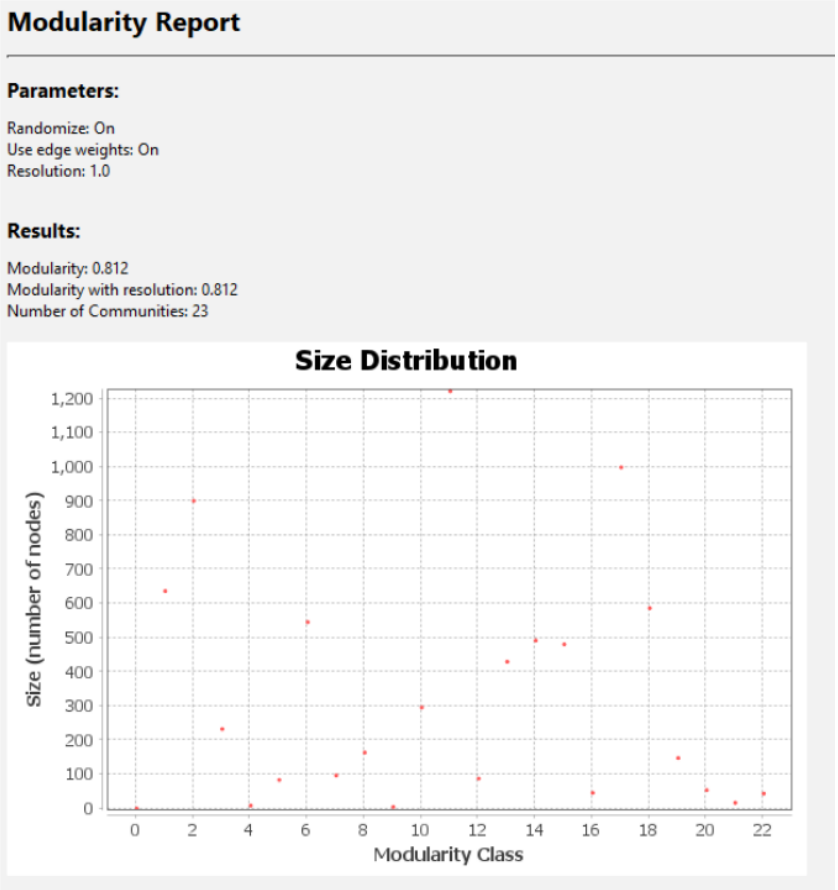


Figure 12: Statistics – modularity.

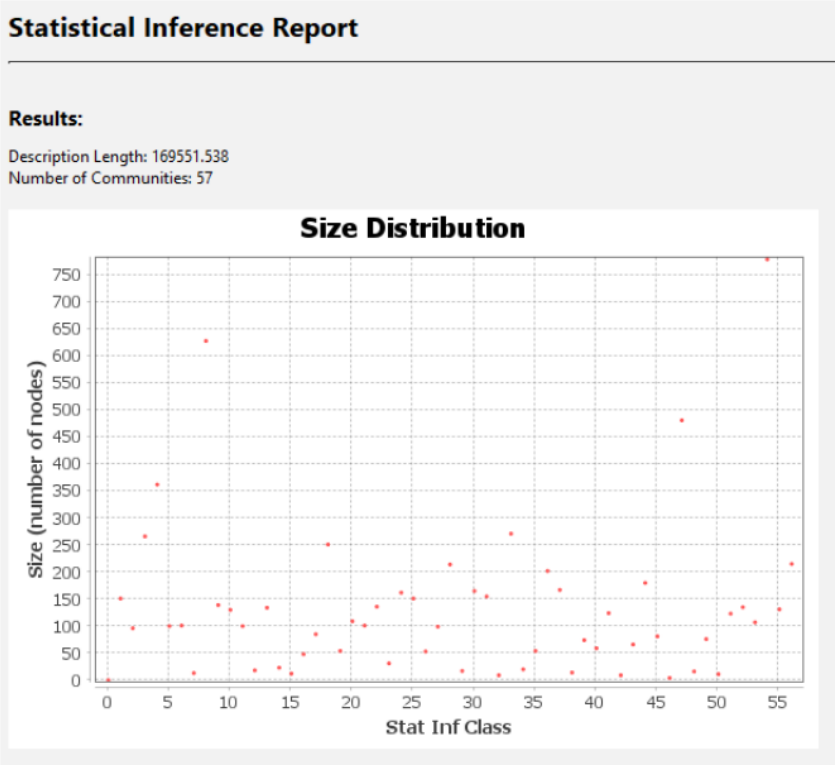


Figure 13: Statistics – statistical inference.

The *Average Clustering Coefficient*, along with the mean shortest path, can indicate a ”small-world” effect. It indicates how nodes are embedded in their neighborhood. The average gives an overall indication of the clustering in the network. The Average Clustering Coefficient is the mean value of individual coefficients.

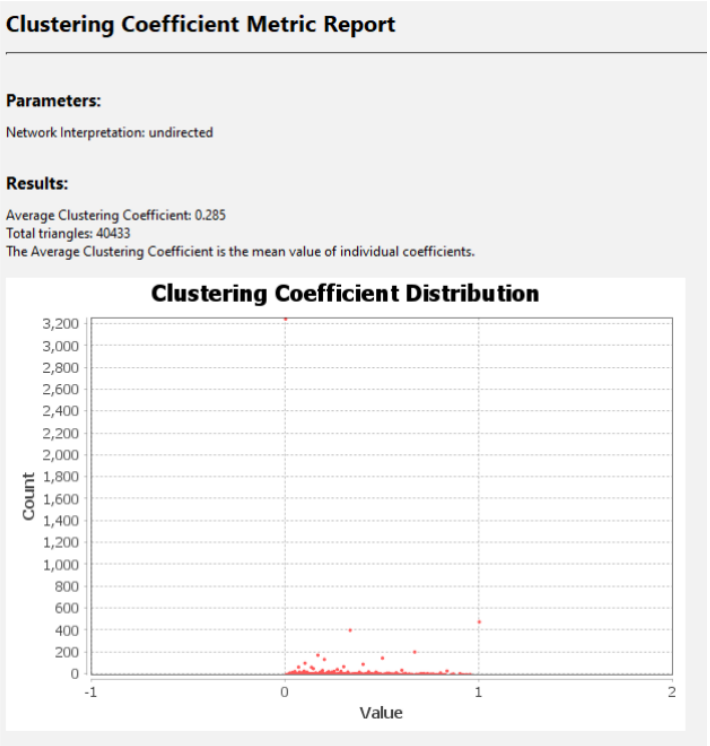


Figure 14: Statistics – Average Clustering Coefficient.

Eigenvector Centrality is a measure of node importance in a network based on node’s connections.

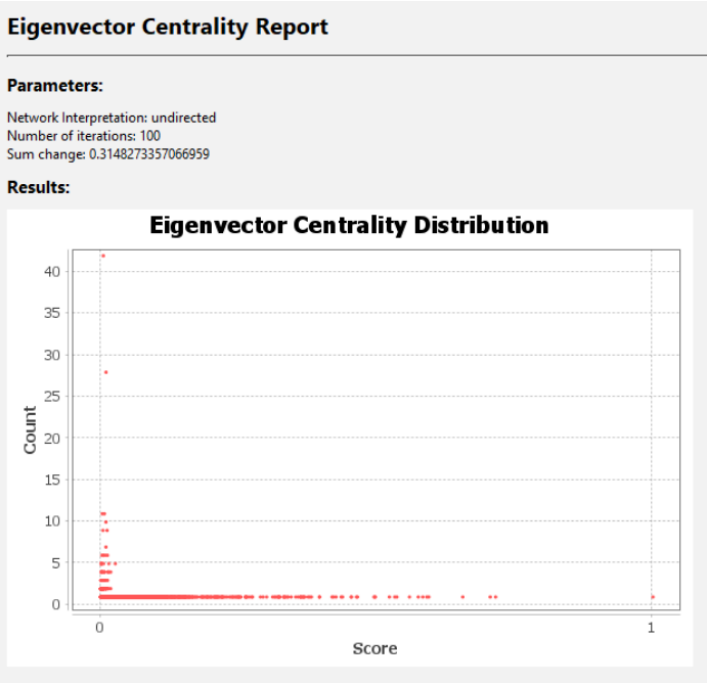


Figure 15: Statistics – Eigenvector Centrality.

Network Diameter. Distance is the average graph distance between all pairs of nodes. Connected nodes have graph distance 1. The diameter is the longest graph distance between any two nodes in the network. Related metrics:

- *Betweenness Centrality*: Measures how often a node appears on shortest paths between nodes in the network.
- *Closeness Centrality*: The average distance from a given starting node to all other nodes in the network.
- *Eccentricity*: The distance from a given starting node to the farthest node from it in the network.

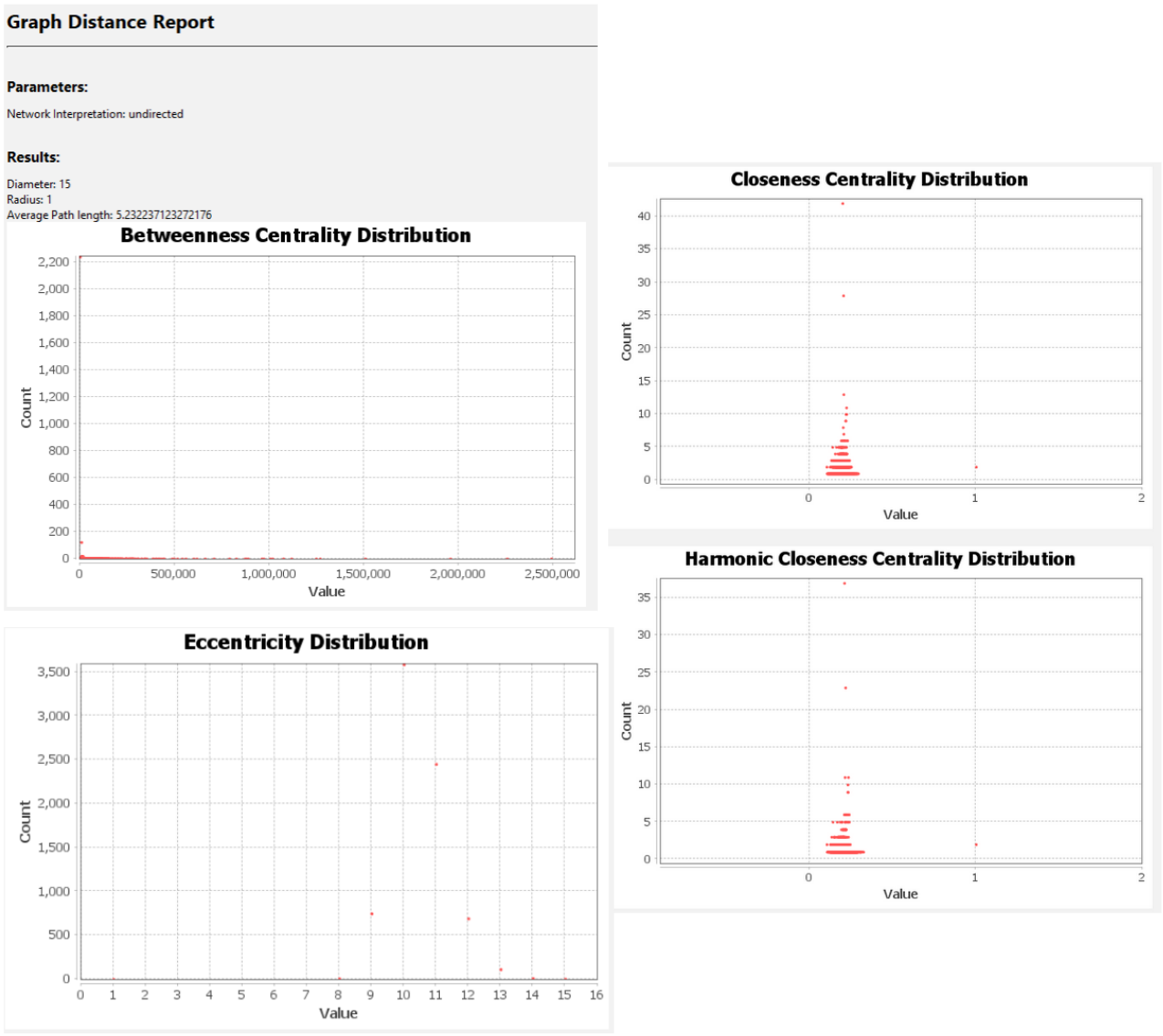


Figure 16: Statistics – Network Diameter.

Step 7. All network measures are summarized below.

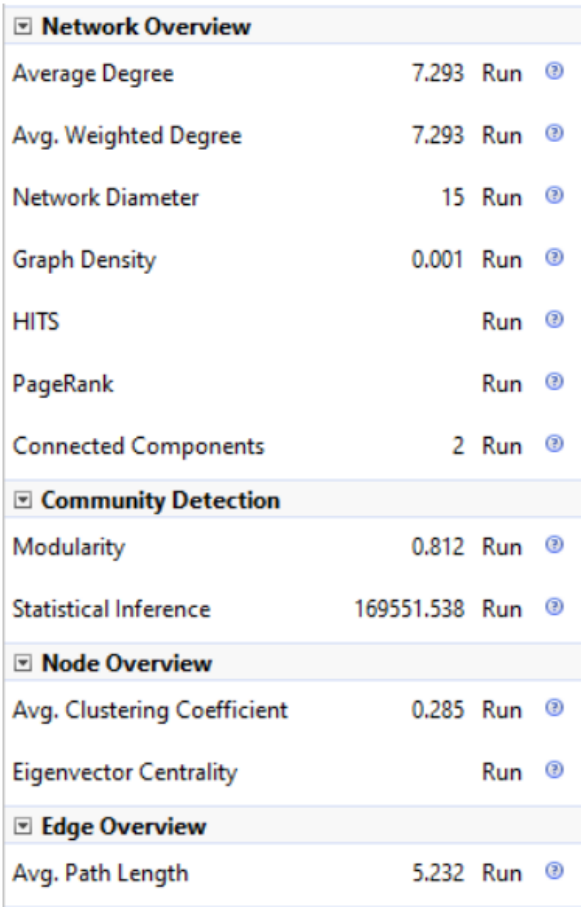


Figure 17: All network measures.

Judging by the Density, the graph is incomplete. Average Degree and Average Weighted Degree are equal because the graph is unweighted.

Conclusions

Gephi was used for analysis of a LastFM user network. It showed itself as a fast and optimized software package that lets user visualize large graphs and gather their metrics, both basic and complex.