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

*Mariia Koroleva*

*J42322c*

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/*](https://gephi.org/)*.*
2. *Choose a network dataset from* [*https://snap.stanford.edu/data/*](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 two different types.*
6. *Calculate available network measures in Statistics provided by Gephi.*
7. *Analyze the results for the network chosen.*

*While performing the work, screenshot the main steps you are doing and insert in the report.*

**Brief theoretical part**

*Gephi* is the leading visualization and exploration software for all kinds of graphs and networks.

*Graph layout* is a way to assign coordinates to each vertex of a graph. To make a “good” layout, the so-called “aesthetic” metrics are usually optimized, which are formulated more objectively:

* Minimum intersection of edges
* Adjacent vertices are close to each other, non-adjacent ones are far
* Communities are grouped into clusters
* Minimum overlap of vertices and edges
* Distribute vertices and/or edges evenly

Here are the examples of several *Graph layout types,* supported by Gephi:

* Force Atlas – a layout focused on quality to allow a rigorous interpretation of the graph with the fewest biases possible, and a good readability. Because of this precision priority this algorithm is extremely slow
* Force Atlas 2 – an improved version of the Force Atlas to handle large networks while keeping a very good quality
* Frunchterman Reingold - simulates the graph as a system of mass particles. The nodes are the mass particles and the edges are springs between the particles. The algorithms try to minimize the energy of this physical system. It has become a standard but remains very slow.
* OpenOrd - expects undirected weighted graphs and aims to better distinguish clusters. It can be run in parallel to speed up computing, and stops automatically.

*Metrics and statistics* considered in this task.

Each graph has such basic measures as *a number of vertices* and *a number of edges*, which were you used in previous lab works.

Also, every vertex has a *degree* measure: d(v), which means the number of edges for vertex v. For all the vertices the *average degree* is counted. Also, we can consider the distribution of the degree in the sample of vertices – *degree distribution*, and see how the degree depends on the vertices overall number.

The *eccentricity* is the greatest distance between a vertex and any other vertex, i.e., how far a node is from the node most distant from it.

Изображение выглядит как текст

Автоматически созданное описание

The *diameter* D is the greatest distance between any pair of vertices.

Изображение выглядит как текст

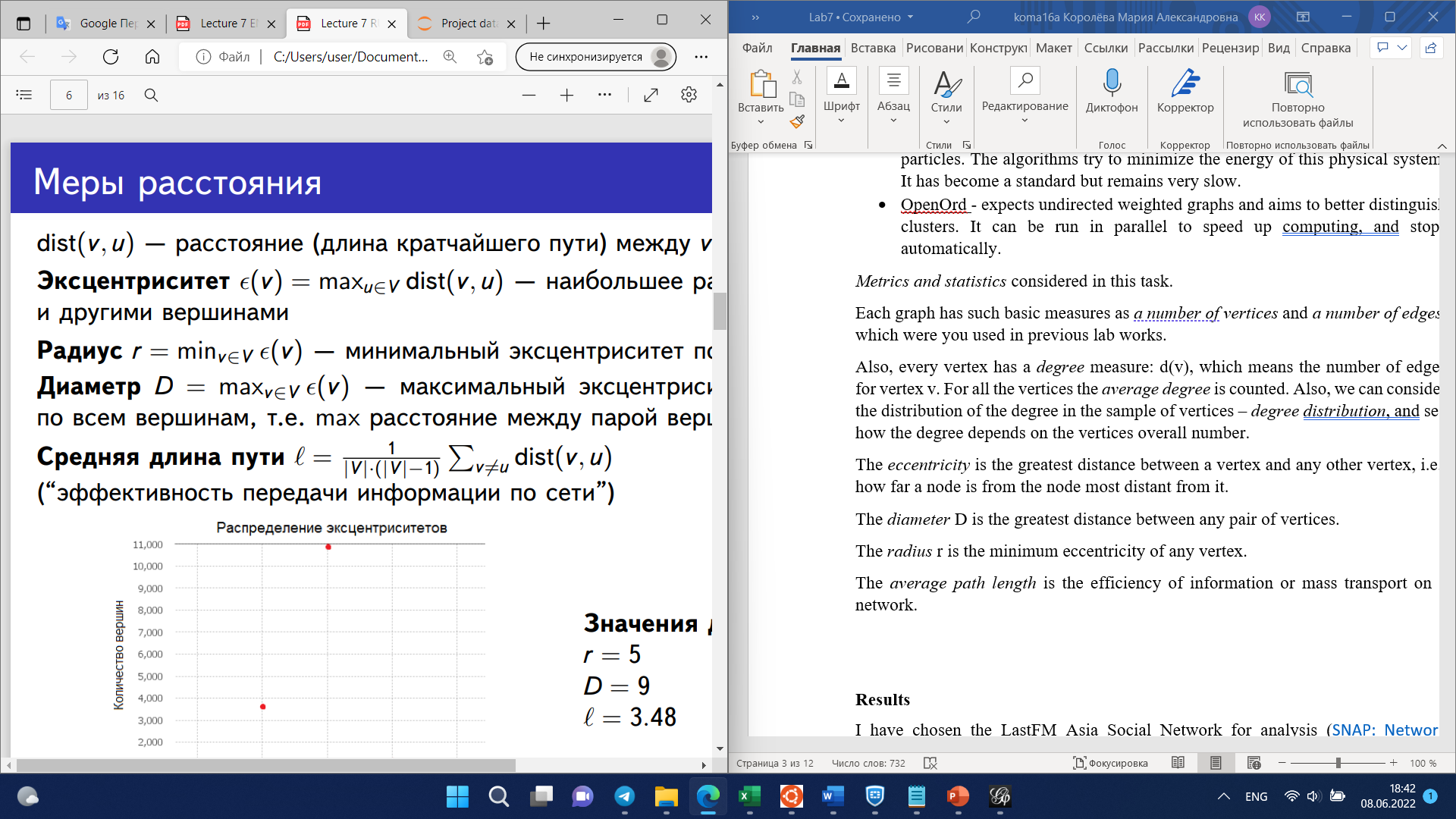
Автоматически созданное описание

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

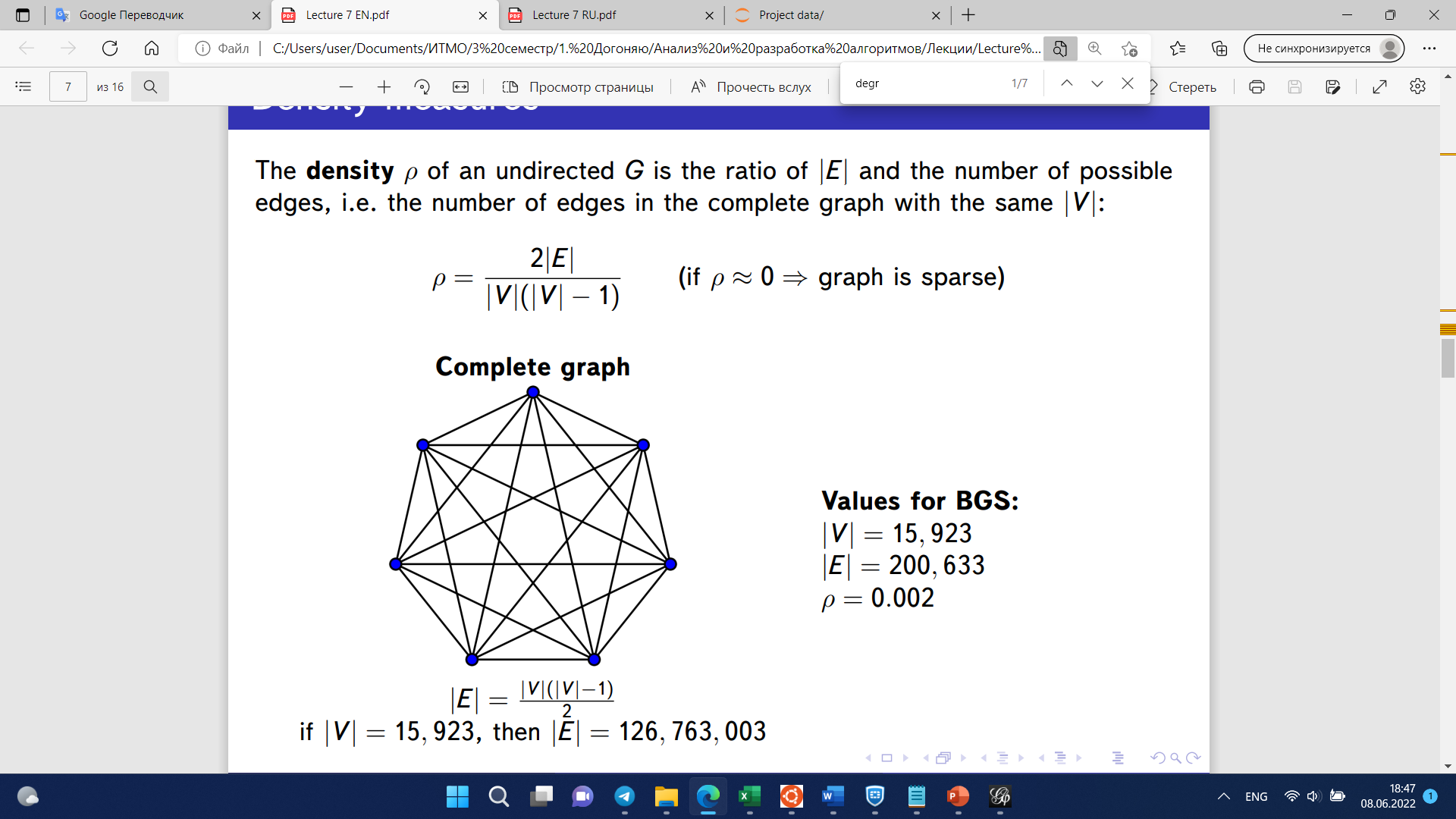
Изображение выглядит как текст

Автоматически созданное описание

The *average path length* is the efficiency of information or mass transport on a network.



The *density* of an undirected graph is the ratio of number of edges and the number of possible edges.



*Modularity* 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.

**Results**

I have chosen the LastFM Asia Social Network for analysis ([SNAP: Network datasets: LastFM Asia Social Network (stanford.edu)](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. Number of nodes: 7624.

Изображение выглядит как текст, снимок экрана, компьютер

Автоматически созданное описание

Then I imported the adjacency list (lastfm\_asia\_edges.csv) into the Gephi.

Изображение выглядит как текст

Автоматически созданное описание

Изображение выглядит как текст, снимок экрана, внутренний

Автоматически созданное описание

Here is our visualized graph:

Изображение выглядит как текст, снимок экрана, внутренний, компьютер

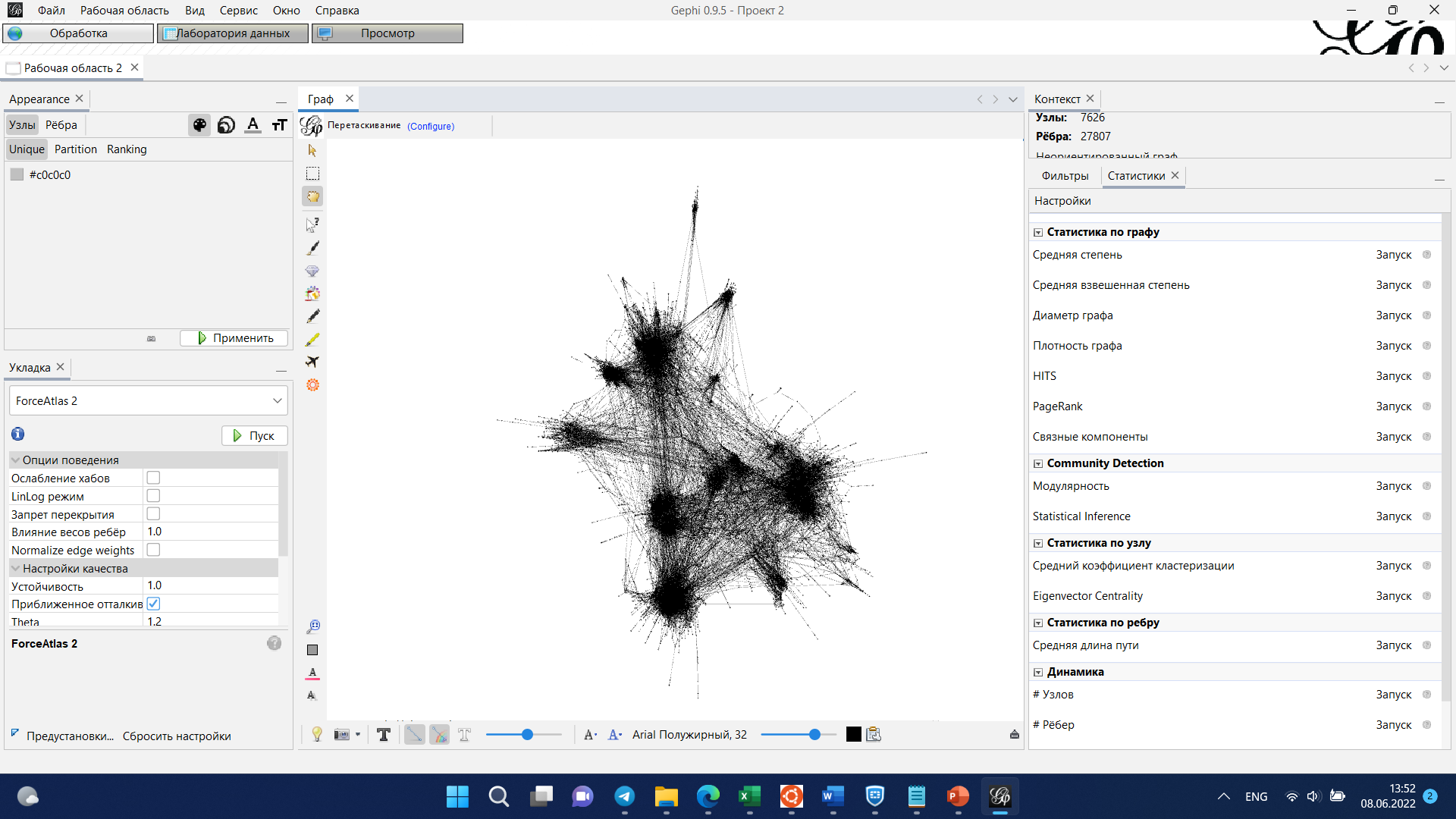
Автоматически созданное описание

In this lab I will demonstrate two layouts of the graph: ForceAtlas 2 and OpenOrd algorithms. They were chosen mainly because of their time effectiveness.

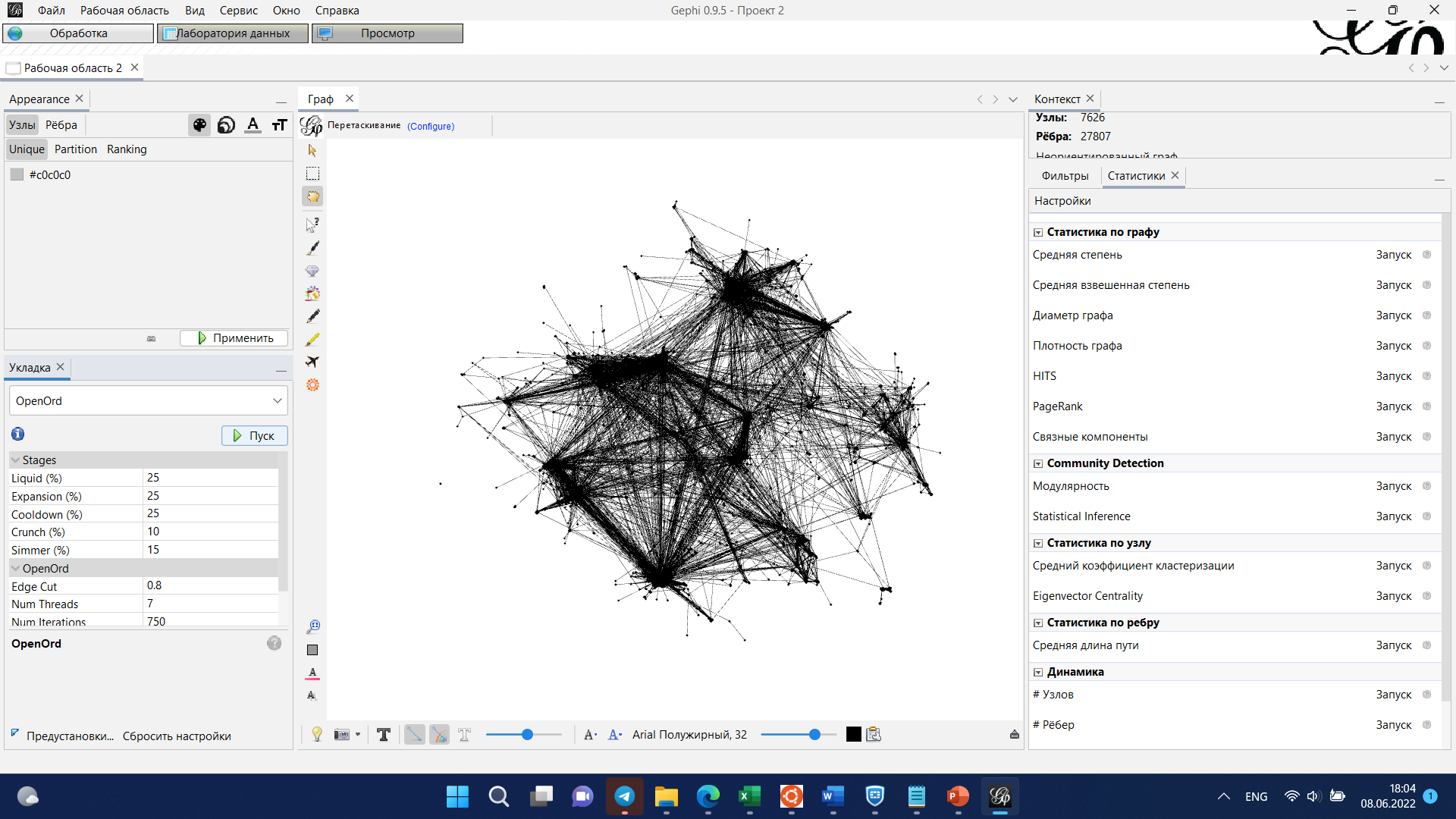
Here is the graph at the very beginning of its note visual relocation. Изображение выглядит как текст

Автоматически созданное описание

Here is the result of ForceAtlas 2 algorithm:



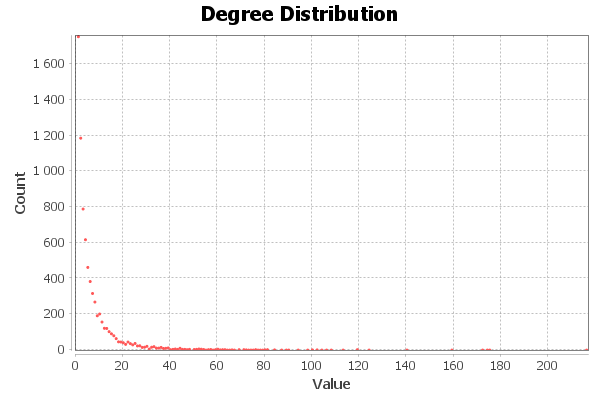
And here is the result of OpenOrd algoritm.

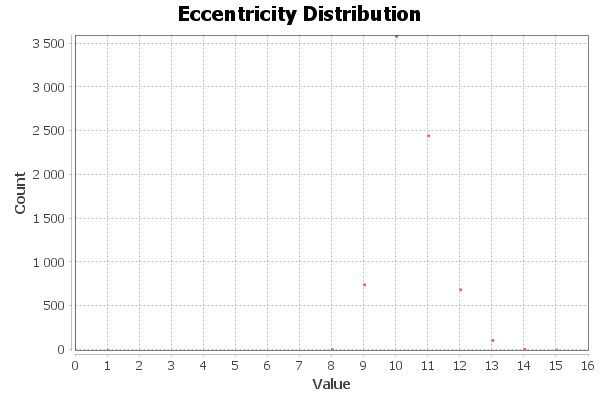


Now let's move on to counting statistics.

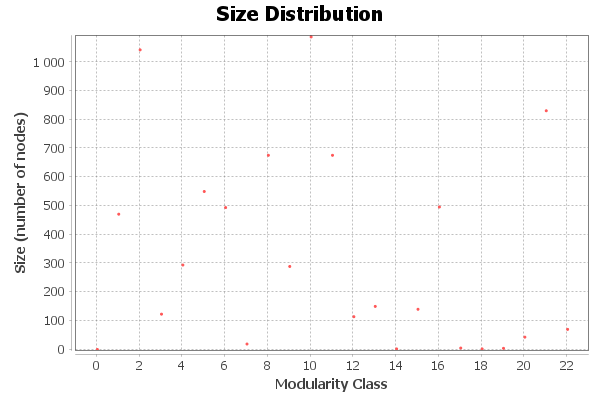
The number of vertices: |V| = 7626

The number of edges: |E| = 27807

Average Degree: 7,293  
  


Diameter: 15  
Radius: 1  
Average Path length: 5.232237123272176  
  


Density: 0,001

Modularity: 0,813  
Number of Communities: 23  
  


**Conclusions**

* The considered graph has vertices which in average have about 7 edges each.
* Visually, all the distribution plots are similar to those considered in a lecture which can mean that the considered network has typical features
* Graph density shows that it is far from being a complete graph
* Graph can be divided into 23 dense clusters of a widely varied size