Task 7

December 7, 2019

Task 7. Algorithms on graphs. Tools for network analysis

Sarafanov Mikhail, Big Data and Machine Learning, C4134

1. Import data and display

The data (Wikipedia Article Networks) for the graph was loaded from the page - https://snap.stanford.edu/data/wikipedia-article-networks.html (07.12.2019). These datasets represent page-page networks on specific topics (chameleons, crocodiles and squirrels). Nodes represent articles and edges are links between them.

For this laboratory work, we used a graph for articles about chameleons.

Type - unweighted, directed graph.

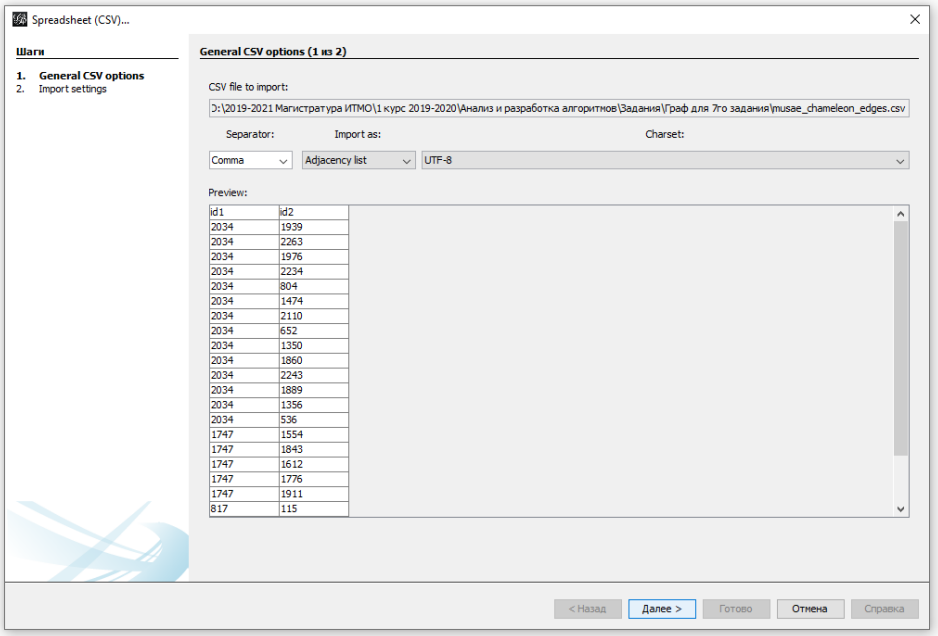

Figure 1.Loading a graph in Gephi

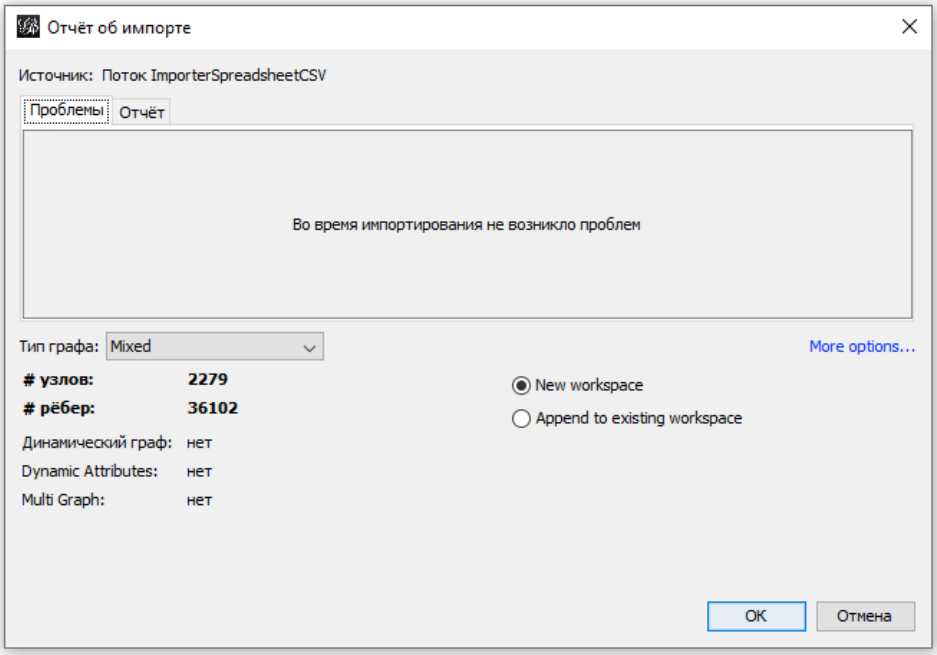

Figure 2. Checking the correctness of data import

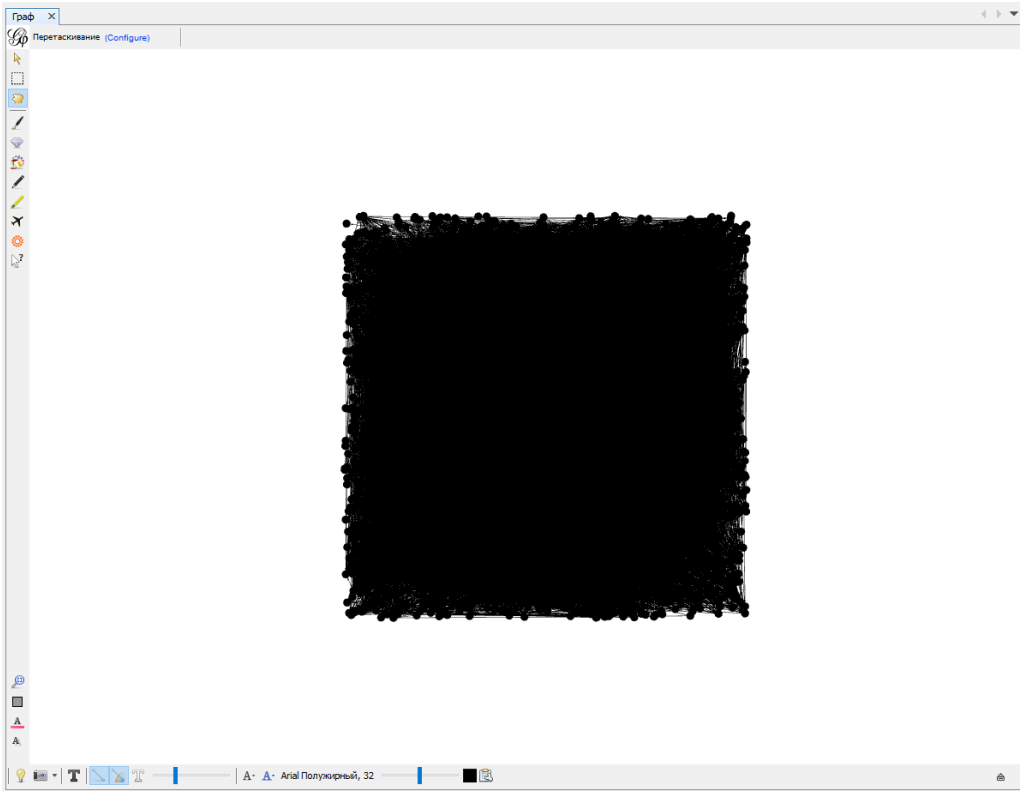

Figure 3. The visualization of the graph

Figure 3 shows that this representation of the graph is not representative. Therefore, the graph layout was performed (fig. 4, 5).

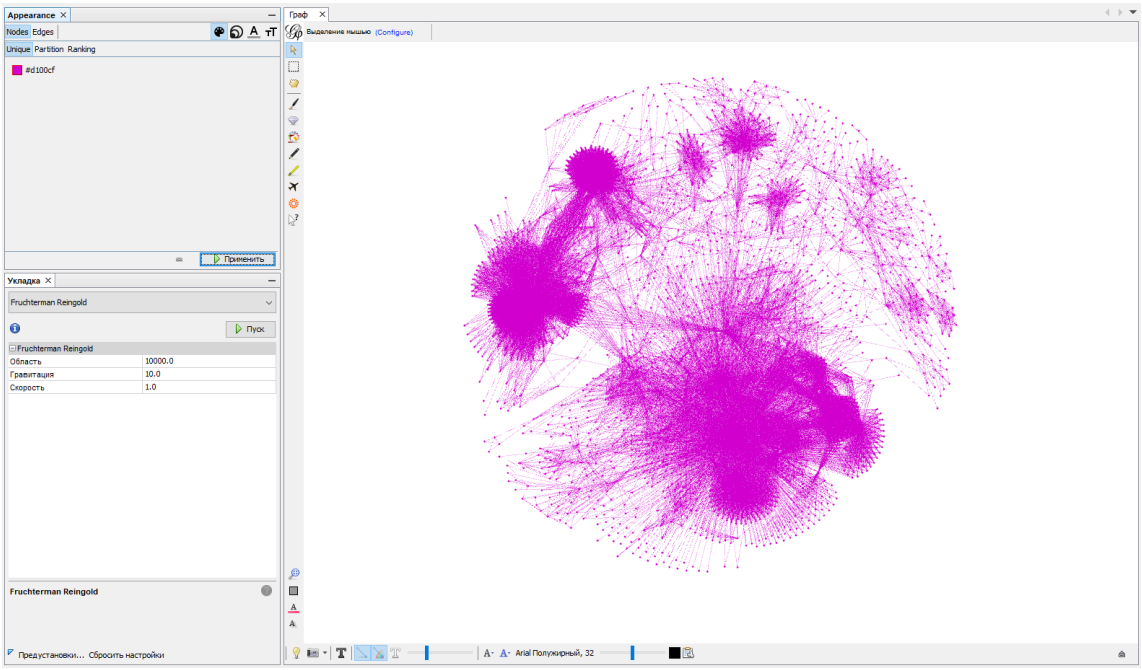

Figure 4. Fruchterman Reingold graph layout

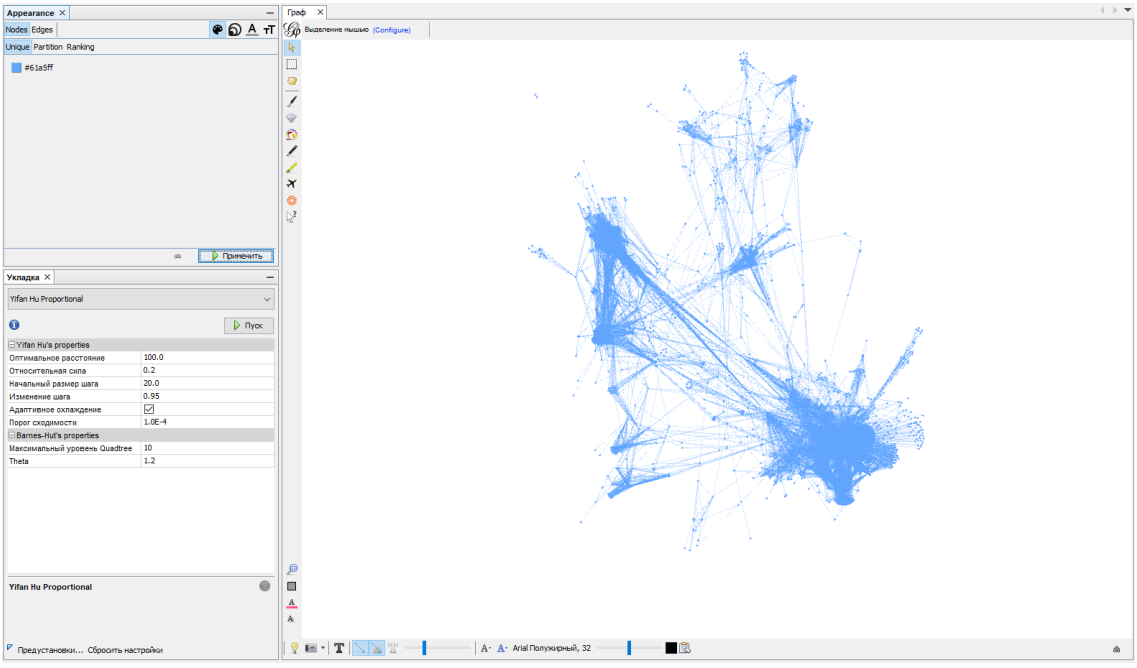

Figure 5. Yifan Hu Proportional graph layout

The graph has two connectivity components. Moreover, in the drawings it is seen that there are several groups (sub-graphs) in our graph. It is possible that these centers are articles devoted only to chameleons. Other articles refer to them when covering broader topics.

2. Statistics

|V| = 2279

|E| = 36102

So, the graph has 2279 vertices and 36102 edges.
The average degree of graph is 15.8

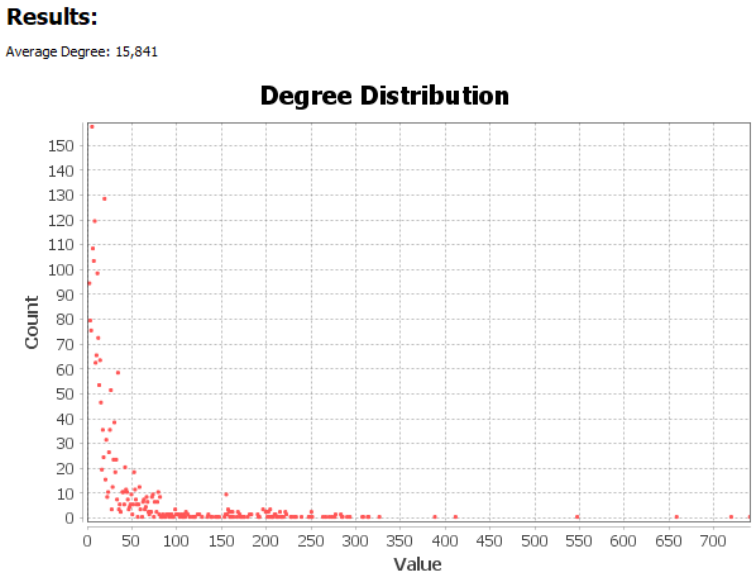

Figure 6. Degree distribution

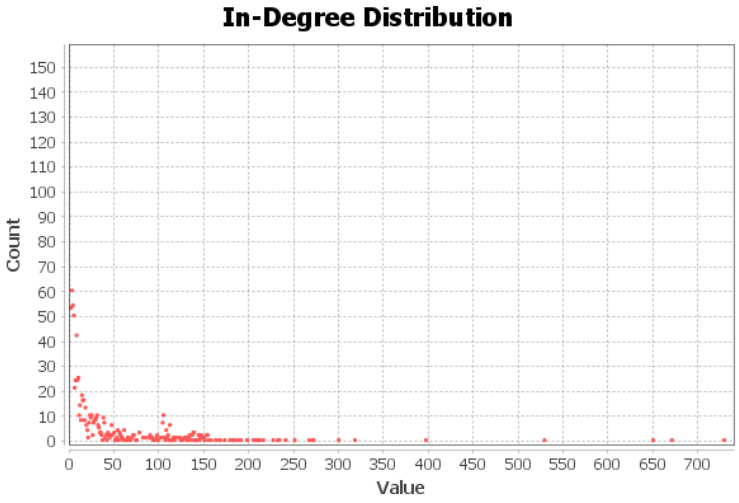

Figure 7. In-Degree distribution

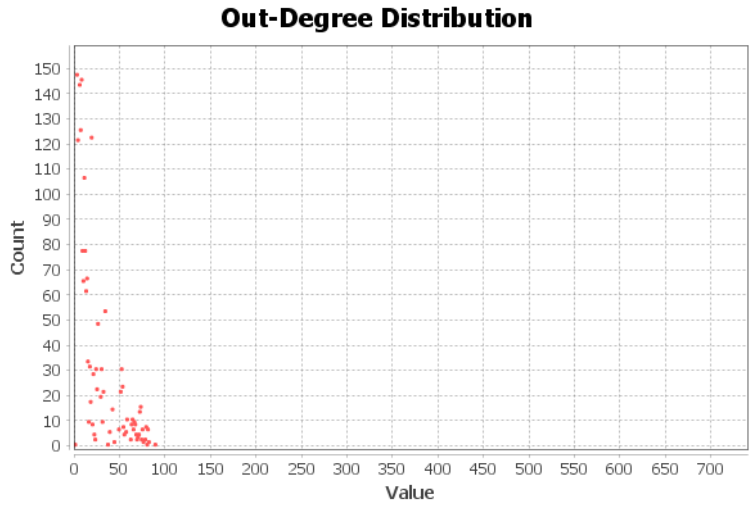

Figure 8. Out-Degree distribution

As can be seen from the graphs, the In-Degree distribution has the highest density in the range of 1 to 25 edges, after which it slowly decreases to the value of 1 vertex (the value of 700 edges on the x-axis). While Out-Degree distribution has a range of 0 to 100 edges.

The diameter D of the graph (maximum eccentricity of any vertex) is 23.

The radius r (minimum eccentricity of any vertex) is 0.

The average path length of the graph is 5.8.

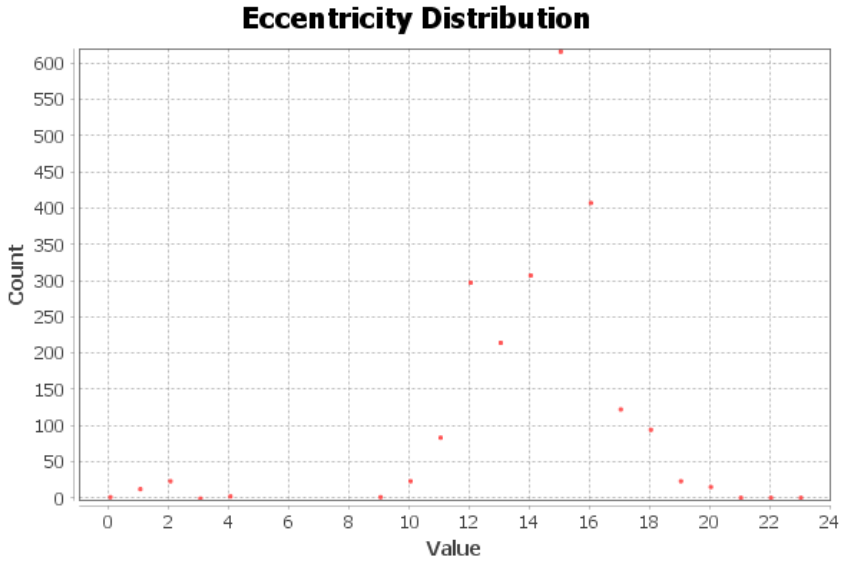

Figure 9. Eccentricity distribution

The density of the graph is 0.007. So, the graph is sparse.

Modularity of the graph is 0.696. This indicates that our graph has dense connections between the vertices within clusters.

Number of Communities is equals to 13.

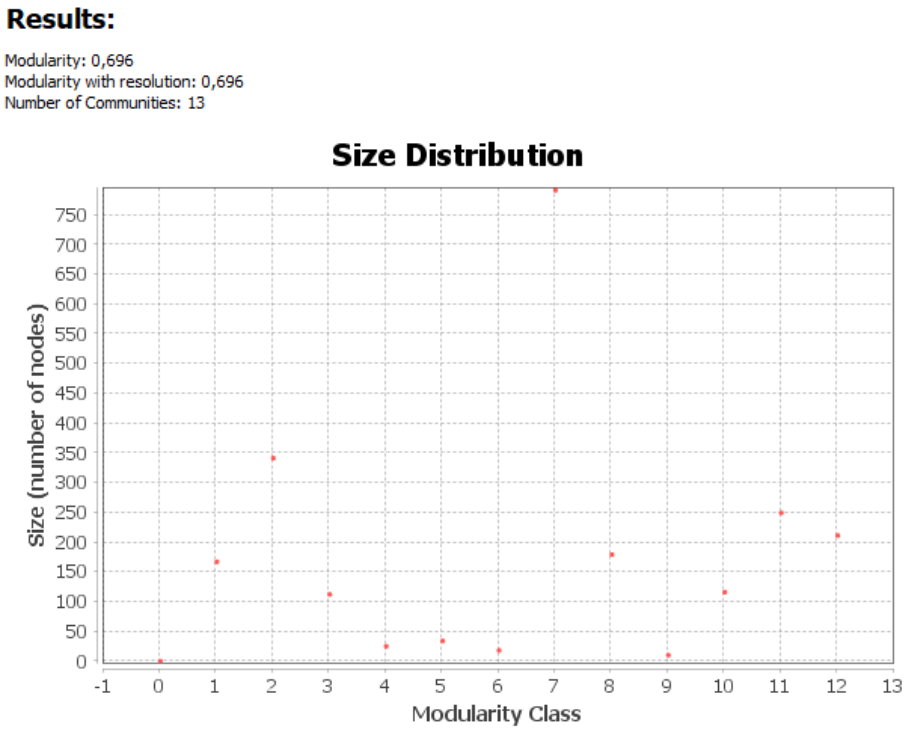

Figure 10. Size distribution

Conclusion

Thus, the graph has 13 clusters. Each article has an average of about 16 links to other Wikipedia articles. The vast majority of articles have no more than 100 references. The average path is 5.8. So, each article is linked to the other through about 6 articles.