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
Tiulkov Nikita
J4133c

Accepted by
Dr Petr Chunaev

St.Petersburg
2021

1. Goal

The use of the network analysis software Gephi

2. 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 two different types.
6. Calculate available network measures in Statistics provided by Gephi.
7. Analyze the results for the network chosen.

3. Brief theoretical part

Basic measures (statistics):
|V| - the number of vertices
|E| - the number of edges

Degree measures:
$d(v)$ - degree of v, i.e. the number of edges for vertex v
$d_{in}(v)$ - in-degree of v, i.e. the number of in-edges for vertex v
$d_{out}(v)$, out-degree of v, i.e. the number of out-edges for vertex
$d_{avg}(v) = \frac{1}{|V|}\sum d(v), v \in V$, average degree over all vertices

Given a connected G, dist(v, u) is the distance (shortest path length) between v and u.

The eccentricity $\epsilon(v)$ of v is the greatest distance between v and any other vertex: $\epsilon(v) = max_{u \in V} dist(v, u)$ ("how far a node is from the node most distant from it").

The diameter D is the maximum eccentricity of any vertex, i.e. the greatest distance between any pair of vertices: $D = max_{v \in V}\epsilon(v)$.

The average path length $l = \frac{1}{|V|\cdot(|V|-1)}\sum_{v \neq u} dist(v, u)$ ("the efficiency of information or mass transport on a network").

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|: $\rho = \frac{2|E|}{|V|(|V|-1)}$

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.

4. Results

In next sections I will describe all steps from problem formulation.

1. The Gephi was downloaded and installed on local machine.

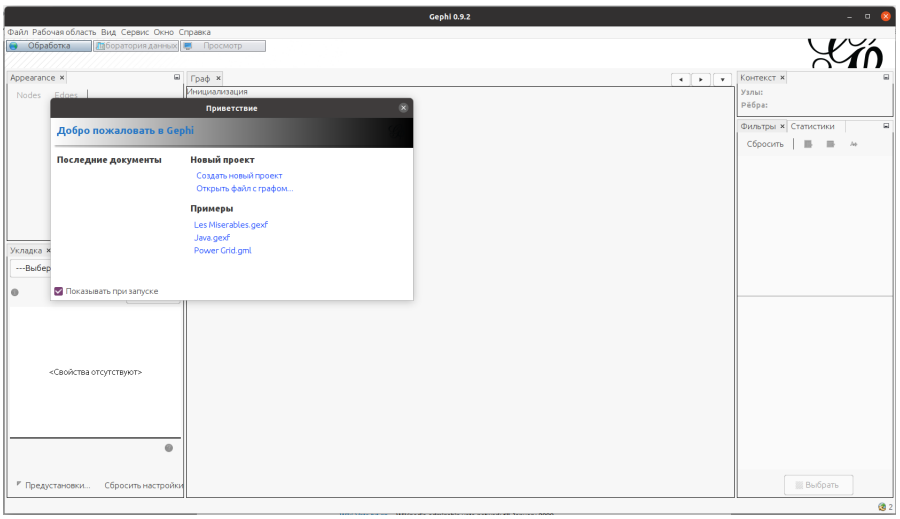

Figure 1: Gephi

2. For this task the «Bitcoin OTC trust weighted signed network» dataset was chosen. This is who-trusts-whom network of people who trade using Bitcoin on a platform called Bitcoin OTC. Since Bitcoin users are anonymous, there is a need to maintain a record of users' reputation to prevent transactions with fraudulent and risky users. This directed network has 5,881 nodes and 35,592 edges.

3. Dataset was in csv format, but has timestamp column and column with weights, these columns was dropped by using Pandas. So we got directed unweighted network.

4. After processing data was correctly uploaded to Gephi.

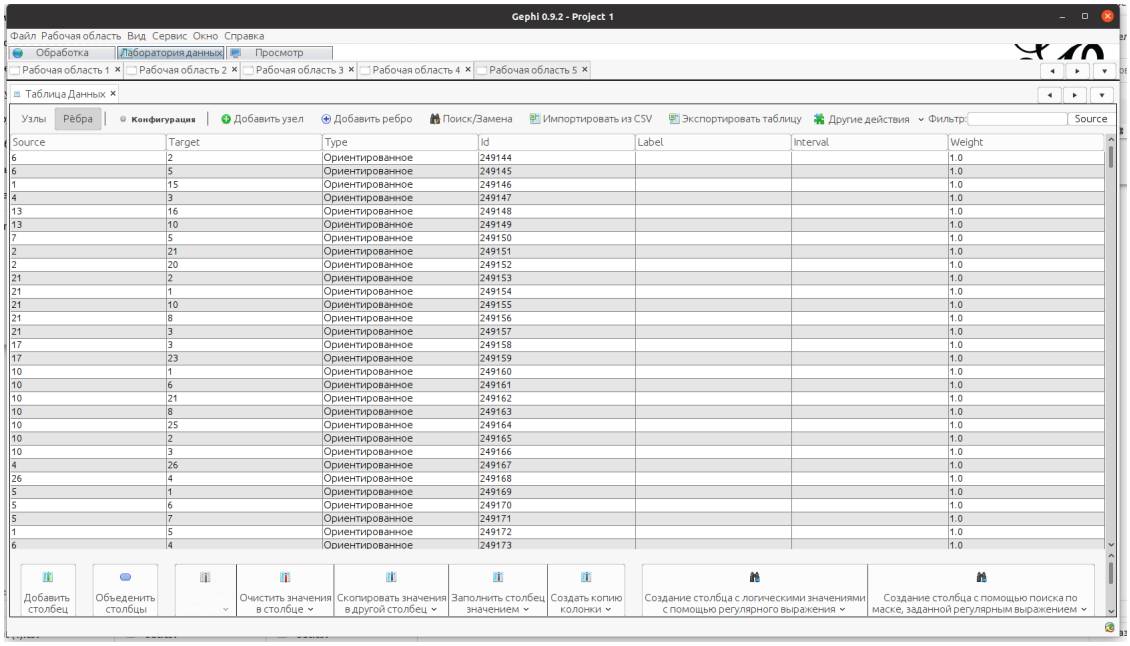

Figure 2: Data

5. Before applying graph layout network is shown on Figure 3.

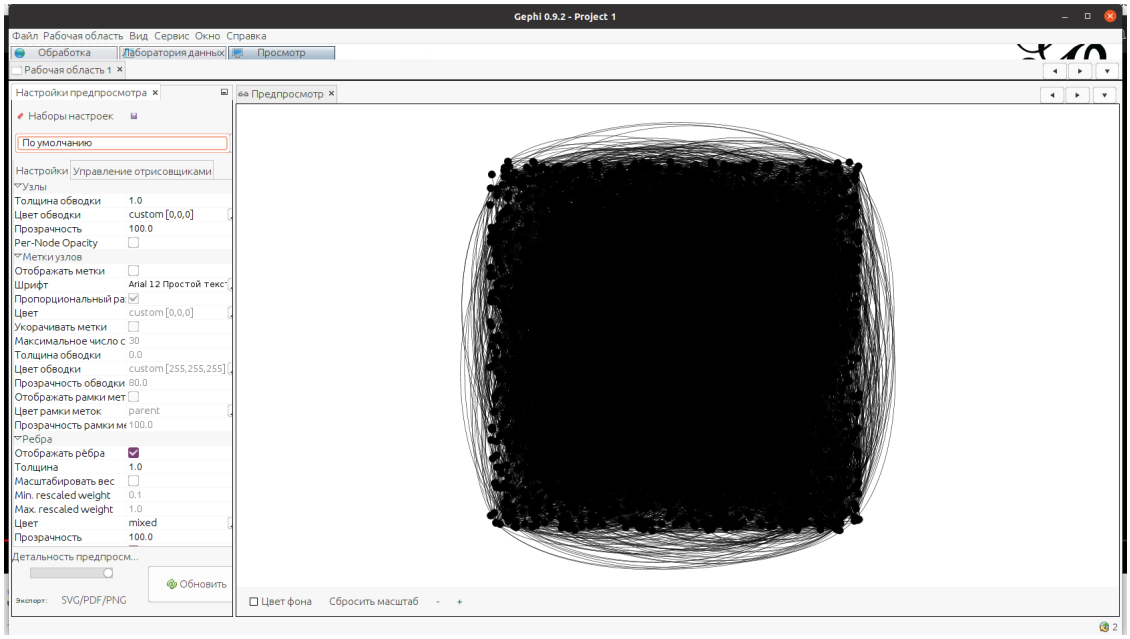

Figure 3: Network before graph layout

Next I have applied ForceAtlas 2 algorithm (Figure 4) and Yifan Hu algorithm (Figure 5).

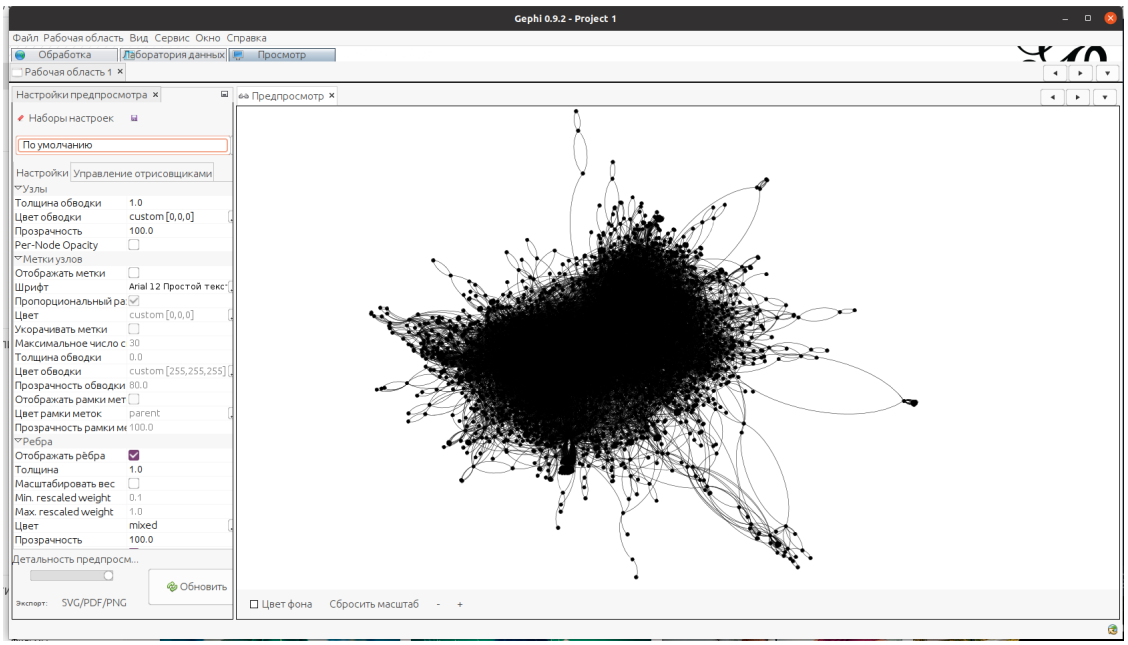

Figure 4: ForceAtlas 2 algorithm

Yifan Hu algorithm was faster than ForceAtlas 2 and shown better visual result. Perhaps it depends on dataset size.

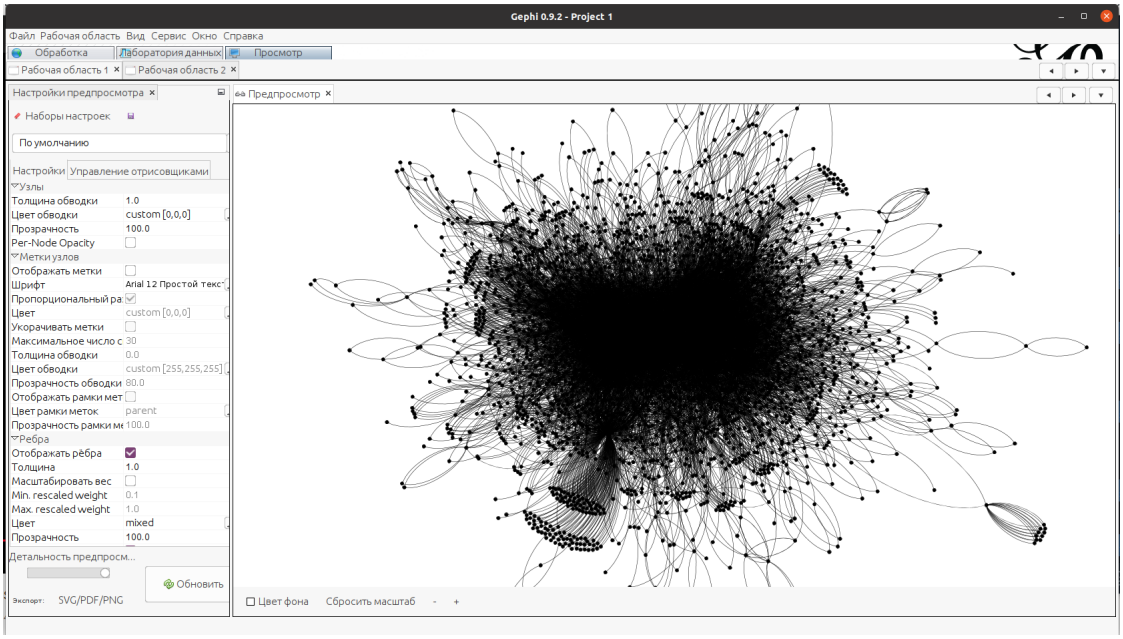

Figure 5: Yifan Hu algorithm

6. Results of calculation of network measures by Statistics.

- $|V| = 5881$
- $|E| = 35592$
- Average degree: 5.417
- Diameter: 11
- Density of the graph: 0.002
- Average path length: 3.814
- Modularity: 0.463
- Modularity with resolution: 0.463
- Number of Communities: 13

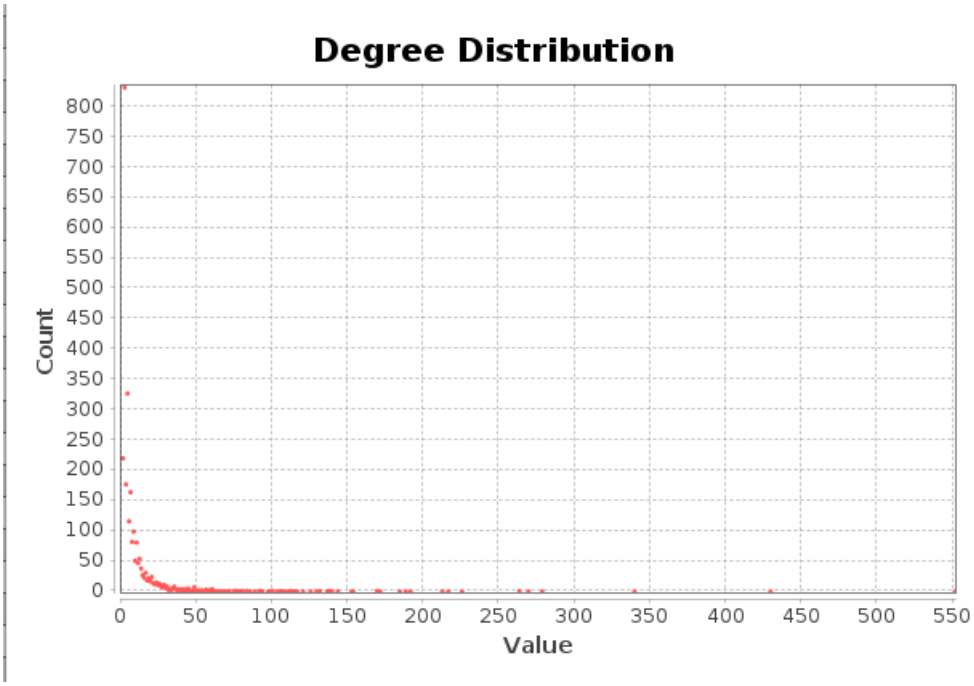

Figure 6: Average degree distribution

7. Let's analyze the results of calculation of network measures.

Since average degree is 5.4 we can make the assumption that common user trust to 5 another users. But some users trust around 300-400 another users. The average path length is 3.8, that means that two common users are connected through 4 another users. Maximum users between two is our diameter minus 2 = 9. There is really small density (it is equal 0.002), so for two users probability of their connection is equal to 0.002, perhaps via dataset size (it is large). Since modularity more than 0, we can make the assumption that some of components connected stronger than others.

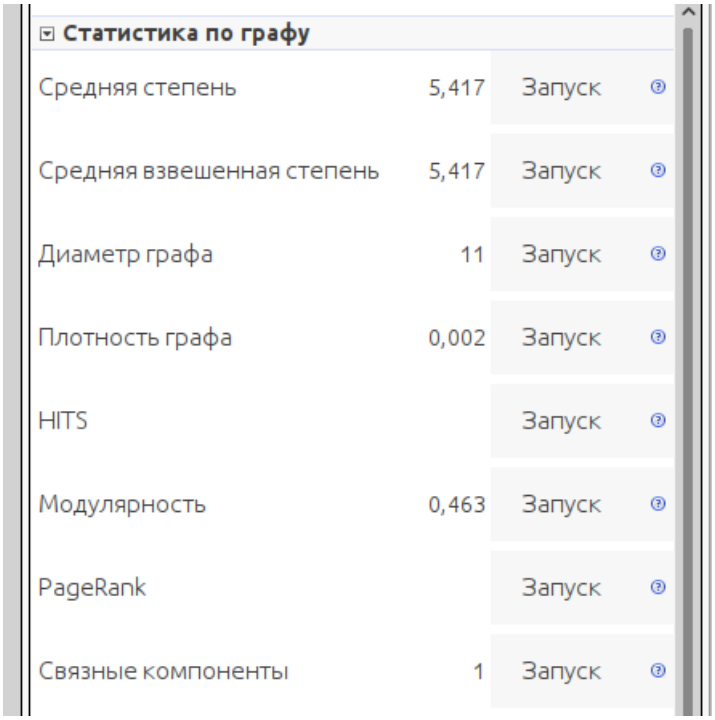

Figure 7: Statistics

5. Conclusions

As a result of this work, network analysis by using Gephi was performed.

Appendix

The code of algorithms you can find on GitHub: https://github.com/FranticLOL/ITMO_Algorithms