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

Victor Voronin

J4132c

Accepted by

Dr Petr Chunaev

St. Petersburg

2020

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

# Brief theoretical part

**Main definition**

* degree − the number of incoming and outgoing edges of the vertex
* degree − the number of incoming edges of the vertex
* degree − the number of incoming edges of the vertex
* −average degree of vertexes
* − shortest path length between vertices v and u in connected graph G
* eccentricity − the greatest distance between vertex v and other vertices
* diameter − max eccentricity for all vertexes, in other words, max distance between any pair of vertexes.
* Graph density ‒ the number of in completed graph with |V| vertexes.
* Modularity is one measure of the structure of networks or graphs which was designed to measure the strength of division of a network into modules (also called groups, clusters, or communities)
* PageRank ‒ is one measure of the structure of networks shows to how often a user following links will non-randomly reach the node “page”.
* eigenvector centrality is a measure of the influence of a node in a network. Relative scores are assigned to all nodes in the network based on the concept that connections to high-scoring nodes contribute more to the score of the node in question than equal connections to low-scoring nodes. A high eigenvector score means that a node is connected to many nodes who themselves have high scores.
* The Average Clustering Coefficient is the mean value of individual coefficients.

**Yifan Hu layout**

It combines a force-directed model with a graph coarsening technique (multilevel algorithm) to reduce the complexity. The repulsive forces on one node from a cluster of distant nodes are approximated by a Barnes-Hut calculation, which treats them as one super-node. It stops automatically

Time complexity is

**OpenOrd layout**

It expects undirected weighted graphs and aims to better distinguish clusters. It can be run in parallel to speed up computing and stops automatically. The algorithm is originally based on Frutcherman-Reingold and works with a fixed number of iterations controlled via a simulated annealing type schedule (liquid, expansion, cool-down, crunch, and simmer). Long edges are cut to allow clusters to separate.

Time complexity is

# Results

1. Follow data set was used for performing of task

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Name | [Type](https://snap.stanford.edu/data/#netTypes) | Nodes | Edges | Description |
| [amazon0302](https://snap.stanford.edu/data/amazon0302.html) | Directed | 262,111 | 1,234,877 | Amazon product co-purchasing network from March 2 2003 |

1. Yifan Hu layout was applied (fig. 1)

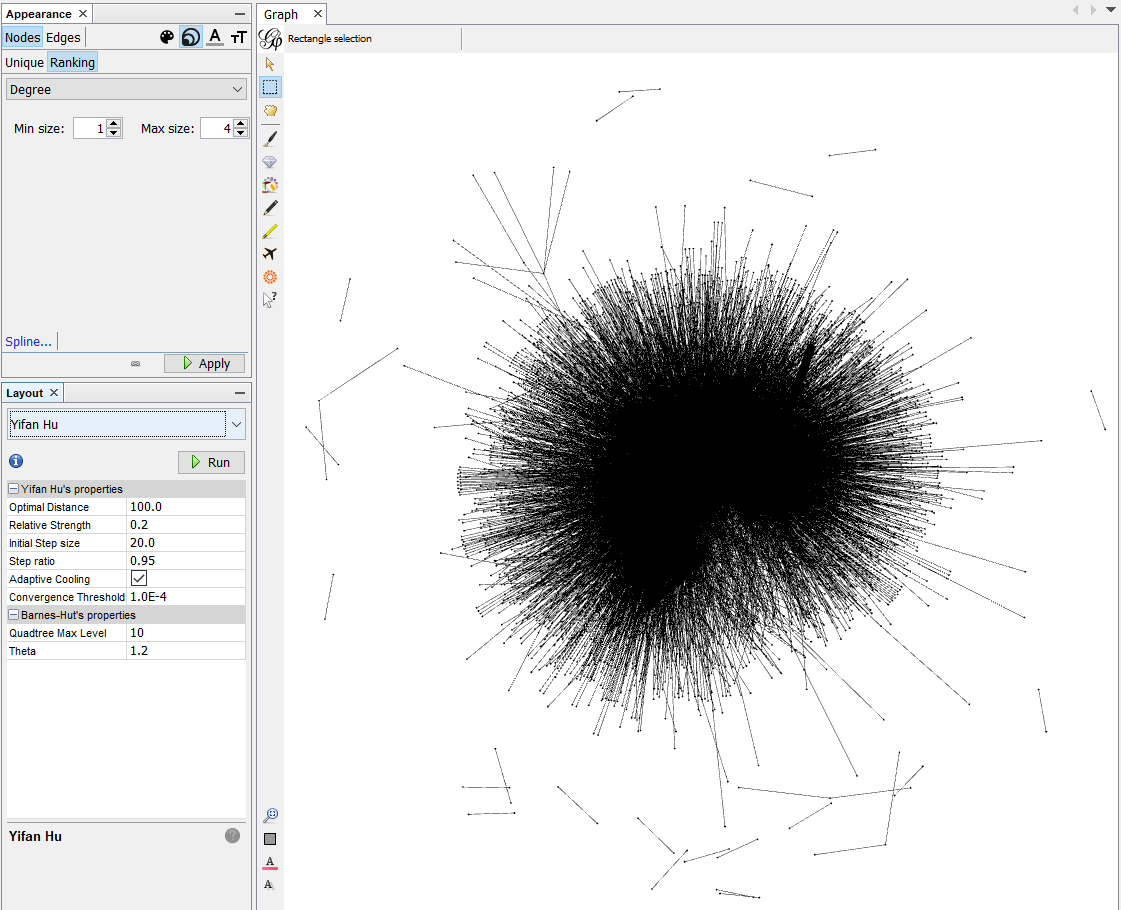


Figure 1 Yifan Hu layout

1. OpenOrd layout was applied (fig.2)

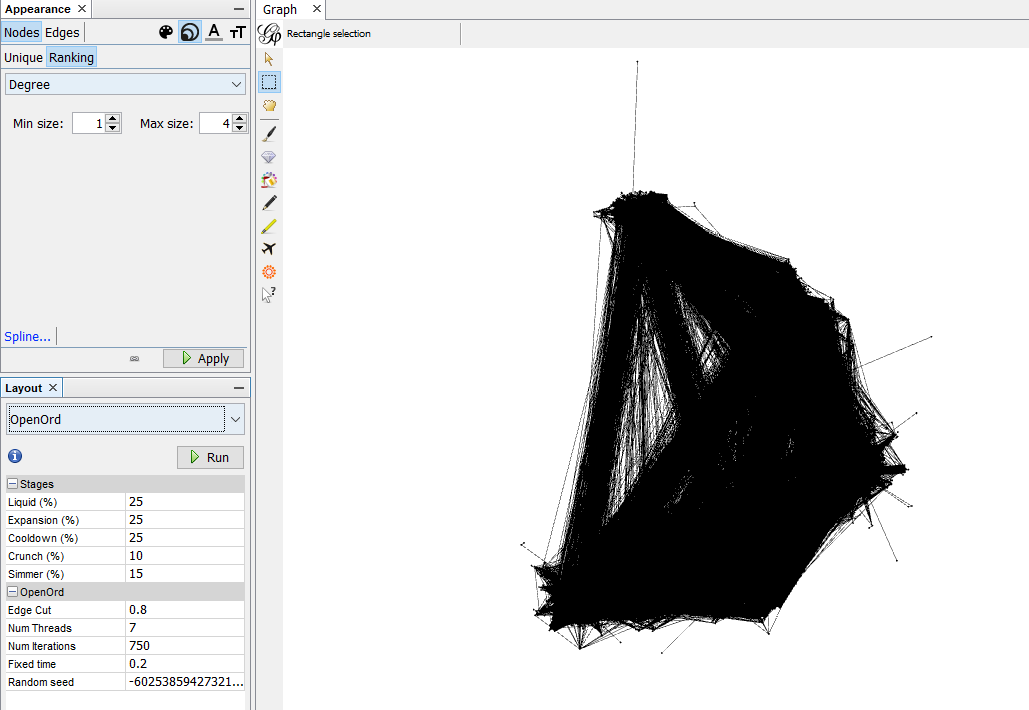


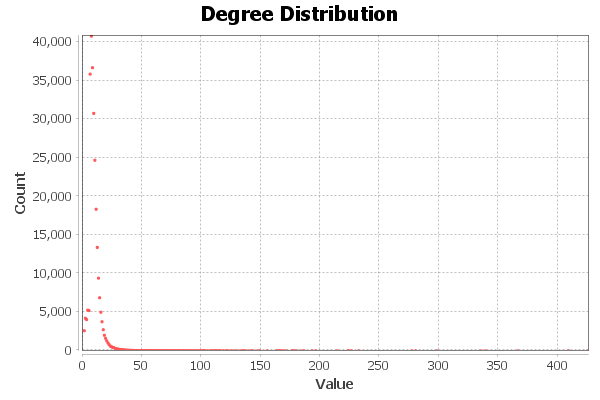
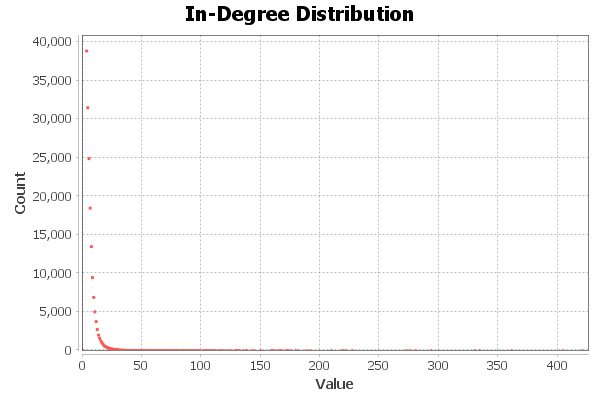
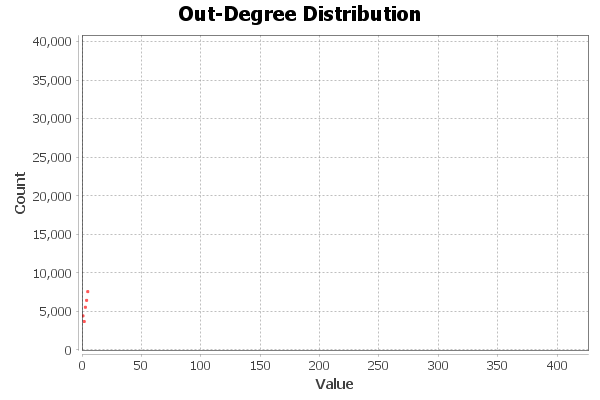
Figure 2 OpenOrd layout

1. The statistic

# Degree Report

## Results:

Average Degree: 4.711

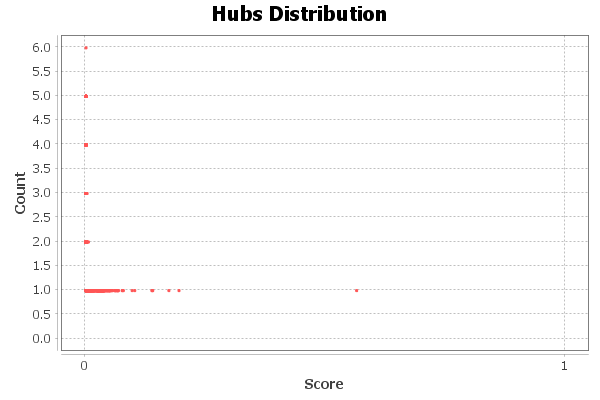
  
  
  


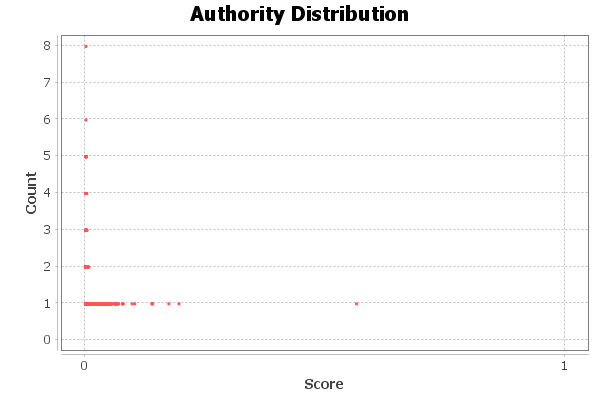
# HITS Metric Report

## Parameters:

Ε = 1.0E-4

## Results:





# Modularity Report

Modularity: 0.895; Modularity with resolution: 0.895; Number of Communities: 93

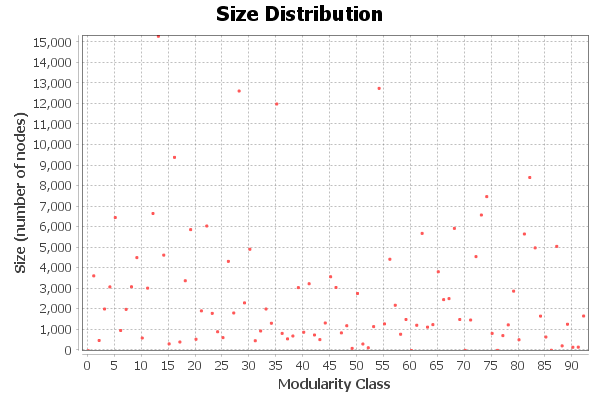
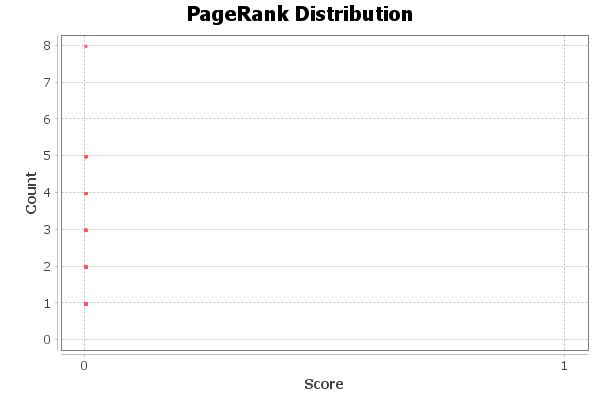


Figure 3 Modularity

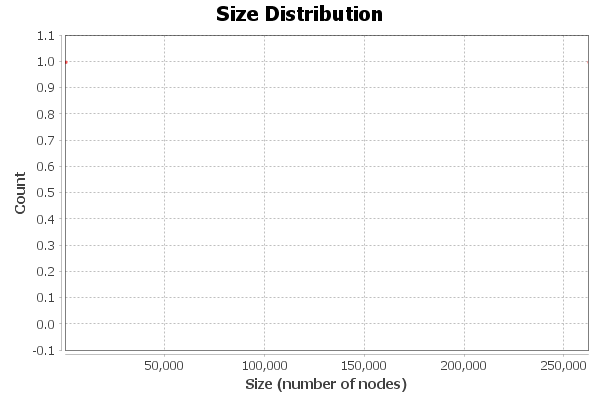
# PageRank Report

Epsilon = 0.001; Probability = 0.85



# Connected Components Report

Number of Weakly Connected Components: 2



# Graph Density Report

Network Interpretation: undirected

Density: 0.2

# Eigenvector Centrality Report

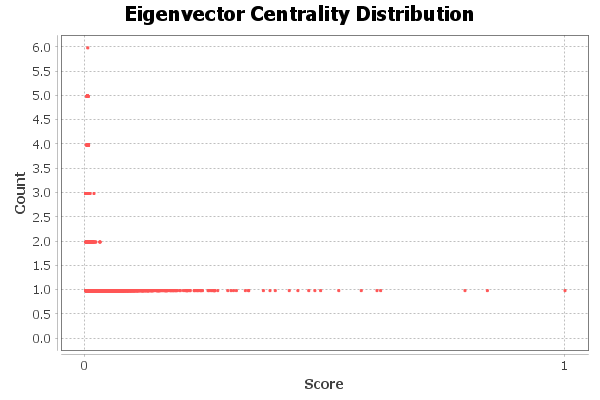
## Parameters:

Network Interpretation: undirected

Number of iterations: 100

Sum change: 8.215262841307267

## Results:



# Clustering Coefficient Metric Report

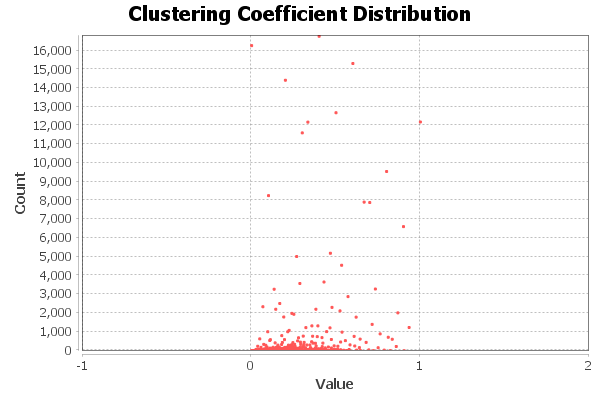
## Parameters:

Network Interpretation: undirected

## Results:

Average Clustering Coefficient: 0.430

Total triangles: 717719



Modularity ~0.9 allows to talk about a relatively simple selection of classes in the graph.

There are few nodes with high value of eigenvector centrality. This suggests that the graph has a few ‘popular” nodes, which linking to others, but others to link to them. The same nodes were distinguished by HITS metric and PageRank.

Clustering coefficient mostly ~1. It is enough low and allow to talk that the nodes in the graph “know” some common node but don’t “know” each other.

The graph has low density and is sparse.

# Conclusions

During the execution of the task, Amazon product co-purchasing network was visualized with Gephi. Different layouts were applied and some measurements of graph were calculated. The results obtained were analysed.