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
Arina Shinkorenok (itmo id: 412704, group: j4132c) &
Nikita Matveev (itmo id: 411831, group: j4133c) &
Fyodor Prazdnikov (itmo id: 412136, group: j4132c)
Academic groups j4132c, j4133c
Accepted by
Dr Petr Chunaev

St. Petersburg
2023

Goal

The use of the network analysis software Gephi.

Formulation of problem

We need to use software Gephi in order to analysis network and calculate network measures in Statistics.

Brief theoretical part

A graph is fundamentally defined as a pair G = (V, E), where V represents a set of vertices, and E signifies a set of paired vertices forming edges. Graphs are quantified by measures such as |V|, denoting the number of vertices, and |E|, indicating the number of edges. The degree of a vertex, denoting the number of edges connected to it, is crucial; vertices with odd degrees are termed odd, while those with even degrees are termed even.

Additional degree characteristics for unweighted graphs include:

- d(v): degree of vertex v, i.e., the count of edges for vertex v,
- $d_{in}(v)$: in-degree of vertex v, representing the number of in-edges,
- $d_{out}(v)$: out-degree of vertex v, representing the number of out-edges,
- $\overline{d} = \frac{1}{|V|}\sum_{v \in V} d(v)$, average degree across all vertices.

For weighted graphs, analogous characteristics factor in edge weights, distinguishing such graphs by assigned numerical values on edges.

Graph analysis extends to characteristics like distance metrics (e.g., shortest path length dist(v,u)) and graph efficiency:

- Eccentricity ε(v) signifies the greatest distance between vertex v and any other vertex.
- Radius r is the minimum eccentricity of any vertex: $r = \min_{v \in V} \varepsilon(v)$.
- Diameter D is the maximum eccentricity among vertices: $D = \max_{v \in V} \varepsilon(v)$.
- Average path length $l = \frac{1}{|V|(|V|-1)} \sum_{v \neq u} dist(v,u)$ assesses information transport efficiency on the network.

Graph density ρ (for undirected graph G) calculates the ratio of existing edges |E| to potential edges in a complete graph with the same |V|: $\rho = \frac{2|E|}{|V|(|V|-1)}$

A graph is sparse when $\rho \approx 0$, while a complete graph has a density of one. These characteristics are invaluable when analyzing intricate real-world networks. Gephi software serves as a powerful tool for conducting such analyses and visualizing networks.

Gephi is a software for graph visualization and analysis. It allows users to analyze and explore various types of data such as social networks, web page link graphs, network graphs, biological networks, and more. Gephi provides a wide range of features including data import from various formats, computation of different graph metrics, and powerful visualization tools for exploring and presenting graphs.

Graph construction schemes and understanding of key statistics:

Graph layouts are essential in visualizing complex networks by positioning nodes and edges in a way that highlights patterns and structures. Two commonly used algorithms are Force Atlas and Yifan Hu. Force Atlas utilizes a physics-inspired approach, treating nodes as charged particles and edges as springs, balancing repulsive forces between nodes and attractive forces along edges to create an aesthetically pleasing layout. It's useful for large networks and helps reveal global structures. Yifan Hu layout, on the other hand, is a force-directed algorithm that optimizes the graph's energy to achieve a clear layout. It is particularly effective for untangling complex networks and is useful when the focus is on local structures.

Main Parameters for Layout Algorithms: for Force Atlas, key parameters include «gravity» (affects the attraction between nodes and the center of the layout), «scaling ratio» (determines the strength of attractive and repulsive forces), and «dissuade hubs» (controls the repulsion between high-degree nodes). Yifan Hu layout generally requires parameters such as

«ideal edge length» (specifying the desired length of edges) and «initial temperature» (controlling the node movement in the initial phase).

Graph Statistics:

- Average Degree: Indicates the average number of connections per node. High average degree suggests a well-connected network.
- Average Path Length: Represents the average shortest distance between all pairs of nodes. Short path length signifies efficient communication.
- Density: Measures the proportion of actual connections to all possible connections. Low density indicates a sparse network.
- Average Clustering Coefficient: Measures the degree to which nodes in a graph cluster together. High clustering coefficient indicates the presence of tightly-knit communities.
- Modularity: Measures the strength of division of a network into modules or communities. High modularity suggests distinct, well-defined communities.
- Number of Communities: Indicates the count of distinct groups in the network. Higher numbers signify a more fragmented network.

Interpreting Graph Statistics:

- A high average degree implies a well-connected network, indicating robust communication channels.
- Short average path length signifies efficient information flow, essential for quick dissemination.
- Low density suggests that only a small fraction of potential connections are realized, revealing network sparsity.
- High clustering coefficient indicates strong local connections, implying the presence of cohesive groups.
- High modularity highlights community structures, demonstrating specialized interactions within groups.
- A large number of communities showcases network fragmentation, with distinct subgroups having limited interaction outside their community.

Employing appropriate graph layouts and understanding key statistics enable researchers to visualize, analyze, and interpret complex networks effectively, providing valuable insights into communication patterns, community structures, and overall network dynamics.

Results

Gephi has been downloaded and installed.

We have taken «email-Eu-core network» dataset from here. The network was generated using email data from a large European research institution. We have anonymized information about all incoming and outgoing email between members of the research institution. There is an edge (u, v) in the network if person u sent person v at least one email. The e-mails only represent communication between institution members (the core), and the dataset does not contain incoming messages from or outgoing messages to the rest of the world.

Table.1 – Dataset information

Name	Value
Nodes	1005
Edges	25571
Nodes in largest WCC	986 (0.981)
Edges in largest WCC	25552 (0.999)
Nodes in largest SCC	803 (0.799)
Edges in largest SCC	24729 (0.967)
Average clustering coefficient	0.3994
Number of triangles	105461
Fraction of closed triangles	0.1085
Diameter (longest shortest path)	7
90-percentile effective diameter	2.9

After loading this dataset into Gephi we have got following picture (Fig.1).

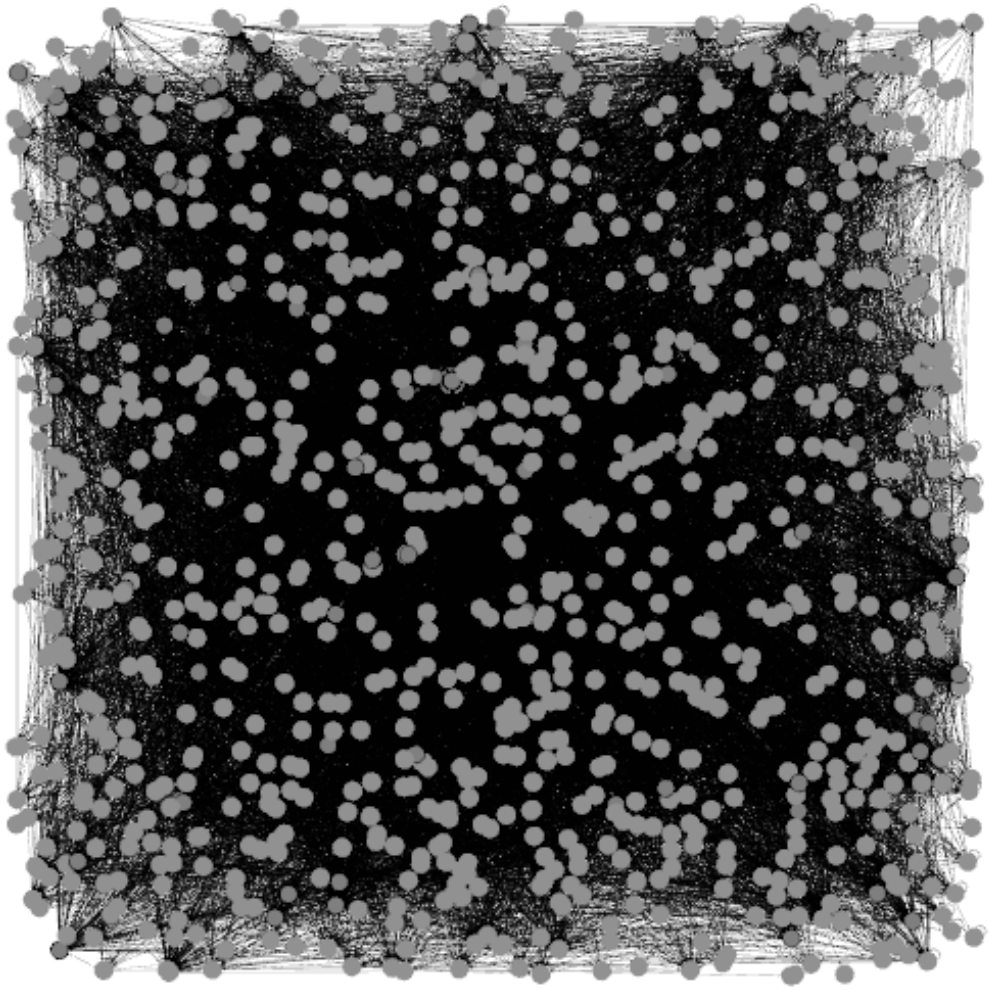

Fig.1 – email-Eu-core network

Try to use Force Atlas layout and Yifan Hu layout in order to get clearer and more comprehensible structure of the graph in the visualization (Fig.2-4).

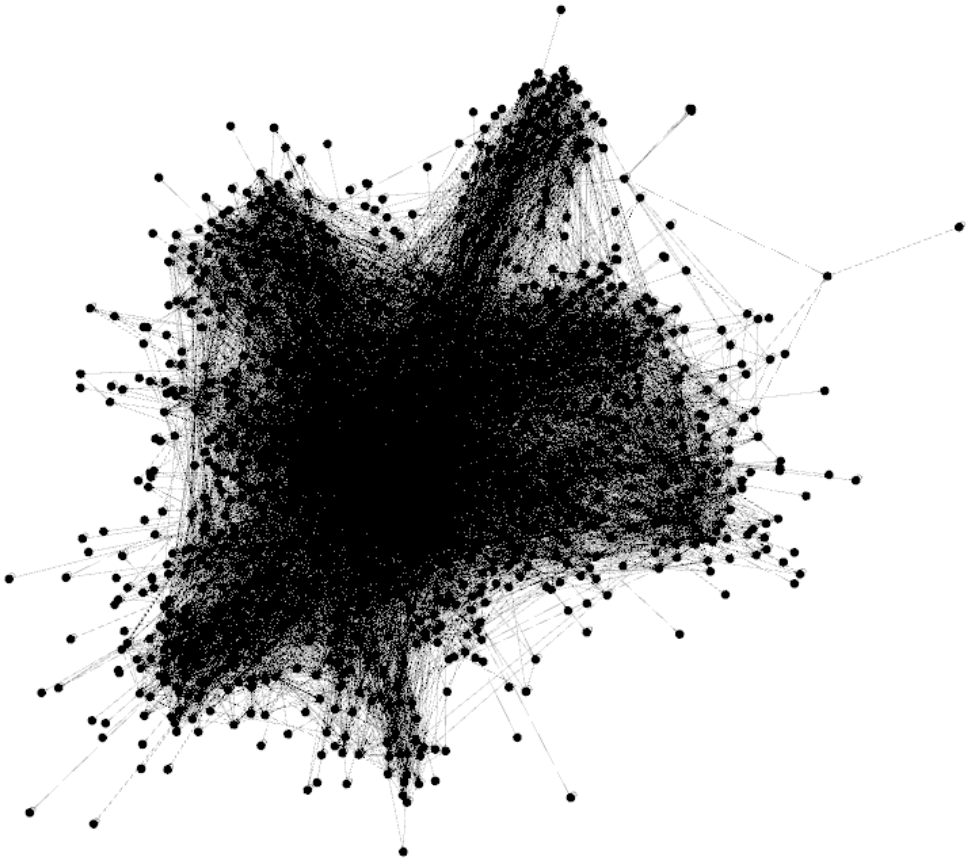

Fig.2 – network after Force Atlas layout (zoomed)

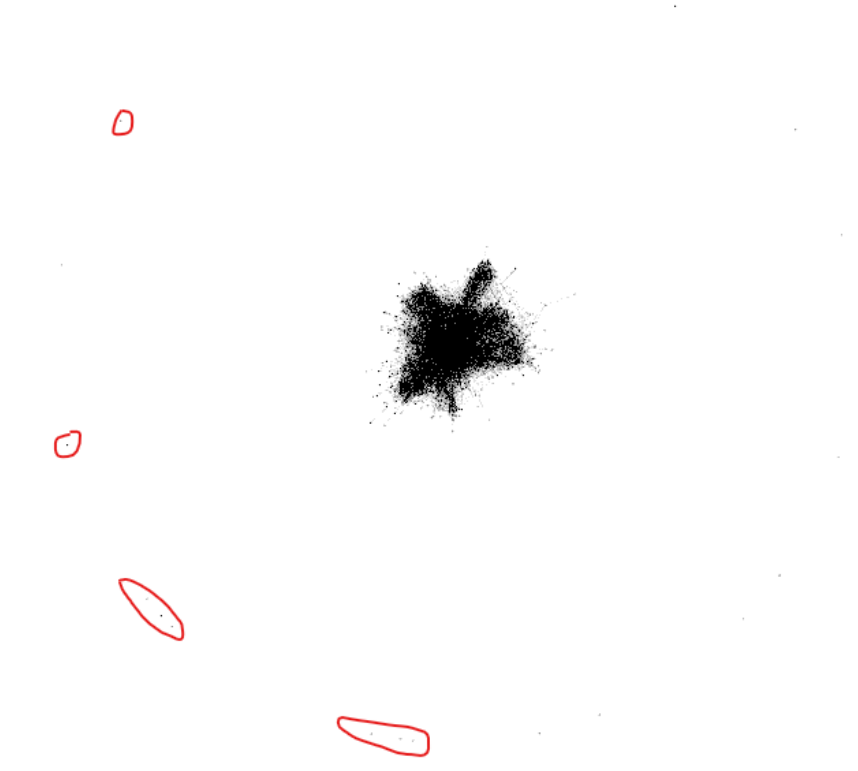

Fig.3 – network after Force Atlas layout (full screen)

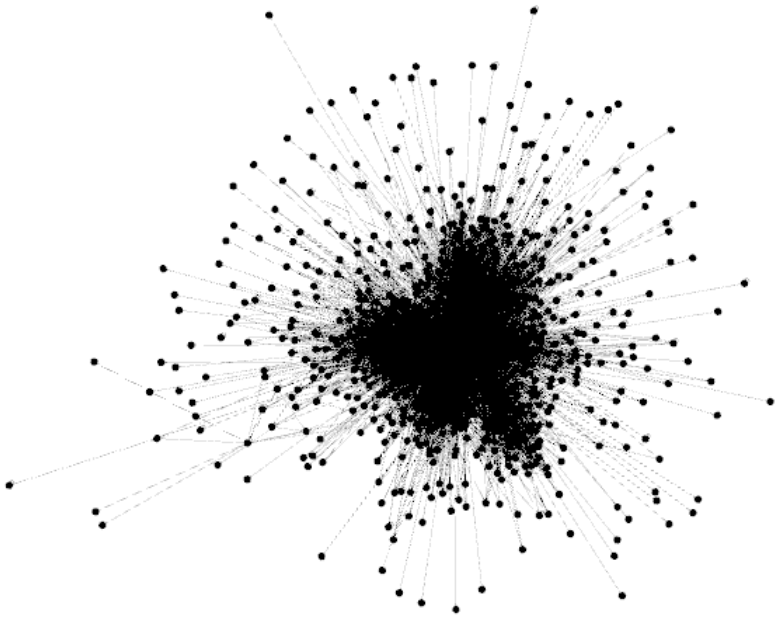

Fig.4 – network after Yifan Hu layout (zoomed)

Using these views, we can notice that some people did not send emails at all.

Next step we try to detect clusters using community detections (Fig.5).

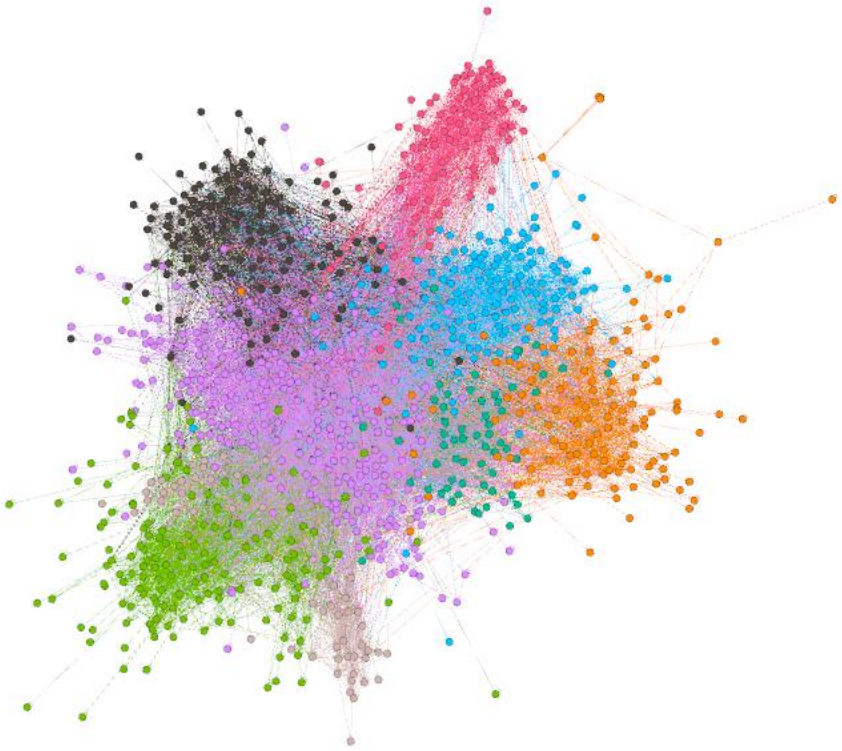

Fig.5 – Clusters

Here we can see that clusters are well defined. It seems like people from same departments communicate more often.

Now we can calculate statistics available in Gephi and analyze results (Fig.6-8 in appendix):

1) Average Degree: 25. This means that on average each node in the graph has about 25 connections to other nodes;

2) Average Path length: 2.66. It is the average of the number of steps required to reach one node from another in the graph.

3) Density: 0,025. This means that the connections between nodes in the graph are relatively sparse, since only about 2.5% of the possible connections between nodes actually exist. The graph is not densely connected, and most pairs of nodes do not have direct connections.

4) Average Clustering Coefficient: 0,372. This means that, on average, neighboring nodes in the graph have a high degree of connectivity.

5) Modularity: 0,416. This means that the graph has a high degree of separation into communities or modules. The modularity value is close to 1, indicating that nodes in the graph are highly clustered within their communities and have few connections to nodes from other communities.

6) Number of Communities: 26. This indicates that the graph is divided into 26 densely connected groups of nodes that form distinct communities.

Conclusions

In our analysis of the «email-Eu-core network» dataset, we utilized Force Atlas and Yifan Hu layouts to visually enhance the graph's structure, revealing that certain individuals did not engage in any email communication within the institution. Through community detection, we identified well-defined clusters, indicating frequent communication among people within the same departments.

Statistical analysis further illuminated the network's characteristics: an average degree of 25 highlighted the interconnectedness of nodes; an average path length of 2.66 signified the short distances between nodes; a density of 0.025 indicated sparse connections; an average clustering coefficient of 0.372 showcased high connectivity among neighboring nodes; and a modularity of 0.416 demonstrated strong community separation, with 26 distinct communities identified.

Our group's conclusion emphasizes the network's segmented nature, where specific departments fostered dense communication, while also highlighting the overall sparse connections within the institution. These findings not only shed light on the communication patterns within the organization but also underscore the significance of community-based interactions in shaping the network's structure.

Appendix

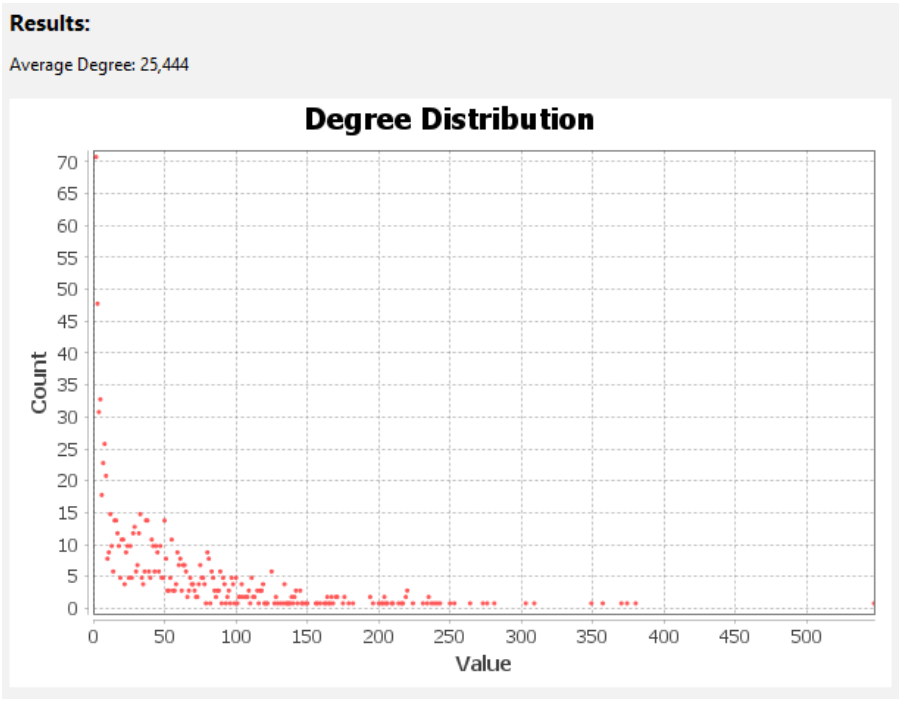

Fig.6 – Degree Report

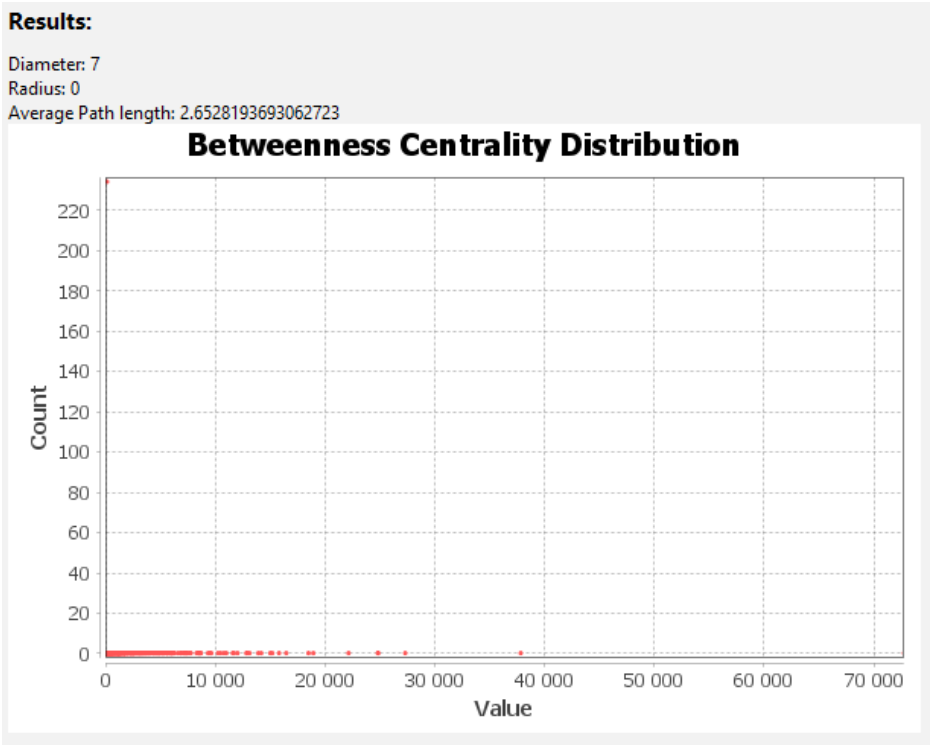

Fig.7 – Graph Distance Report

Results:

Modularity: 0,416
Modularity with resolution: 0,416
Number of Communities: 26

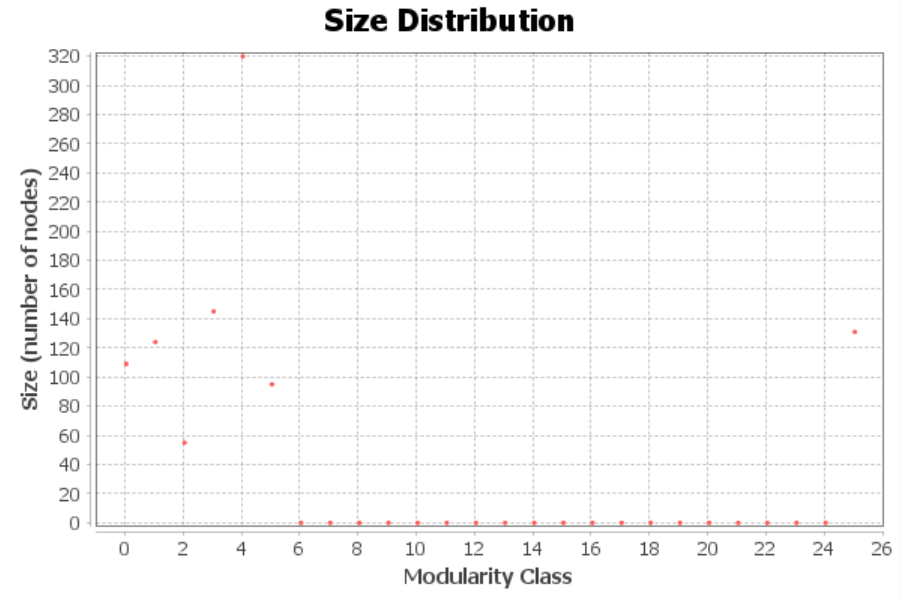

Fig.8 – Modularity Report