FODBMS REPORT

ON

LastFM Asia Social Network

Submitted in Partial Fulfillment for the Course

of

Fundamentals of Database Management System

in Term-3

For

Academic Session 2021-2022

SUBMITTED TO:

Prof. Ashok Harnal

SUBMITTED BY:

Anisha Siwas 025007

Suchit Katyal 025033

Kartik Mohan Sinha 025017

LastFM Asia Social Network

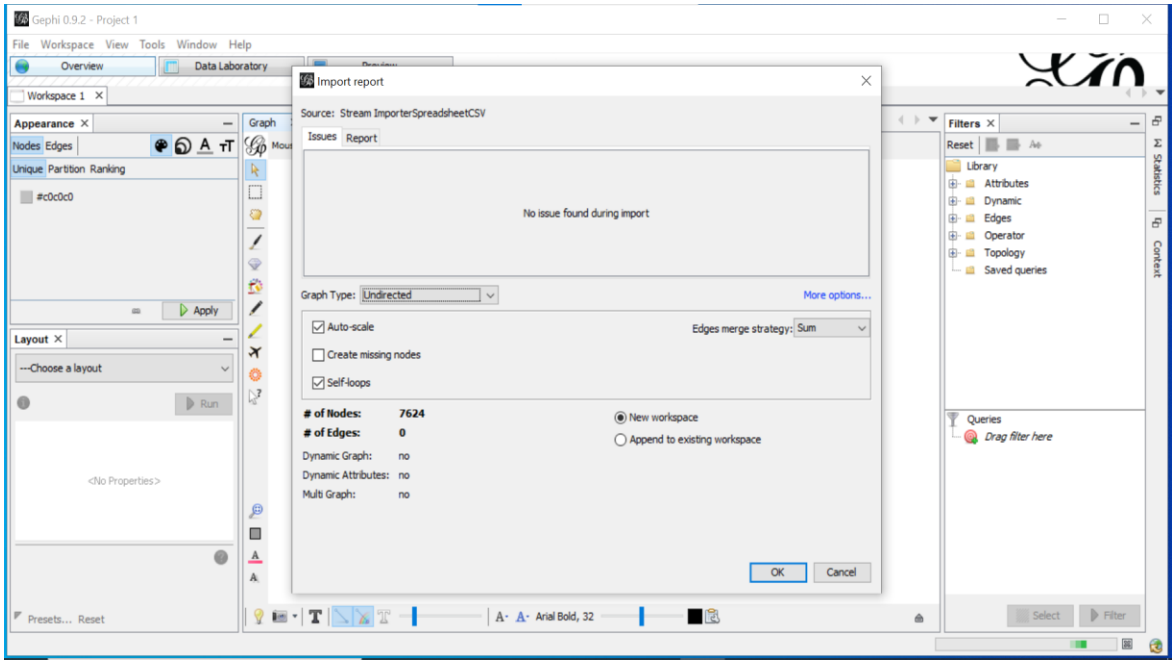
A social network of LastFM users which was collected from the public API in March 2020. Nodes are LastFM users from Asian countries and edges are mutual follower relationships between them. The vertex features are extracted based on the artists liked by the users. The task related to the graph is multinomial node classification - one has to predict the location of users. This target feature was derived from the country field for each user.

DATASET STATISTICS

Directed	No
Node Features	Yes
Edge Features	No
Node Labels	Yes. Multi Class.
Temporal	No
Nodes	7,624
Edges	27,806
Density	0.001
Transitivity	0.179

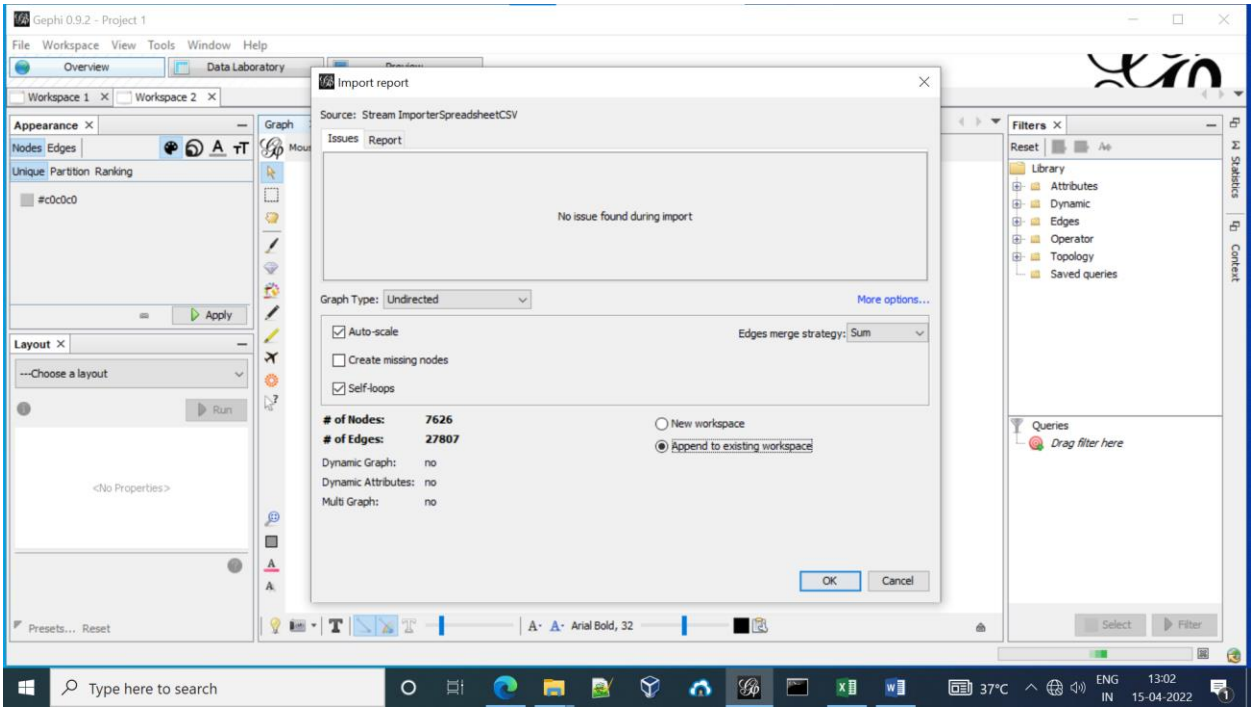
Importing target file to Gephi

While importing, making sure that all data-types of features being imported are correct and also unchecking ‘Create missing nodes’ option. Finally creating a new workspace.

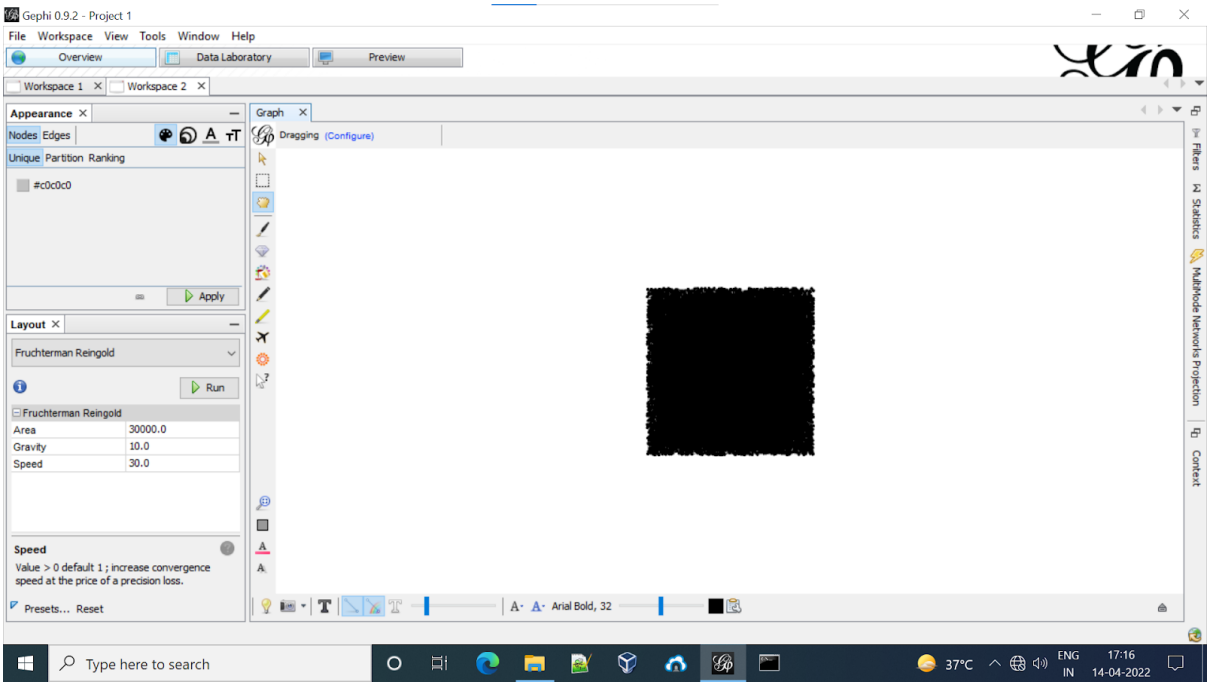


Importing edges file to Gephi

While importing, making sure that all data-types of features being imported are correct and also unchecking ‘Create missing nodes’ option. Finally appending to the existing workspace.



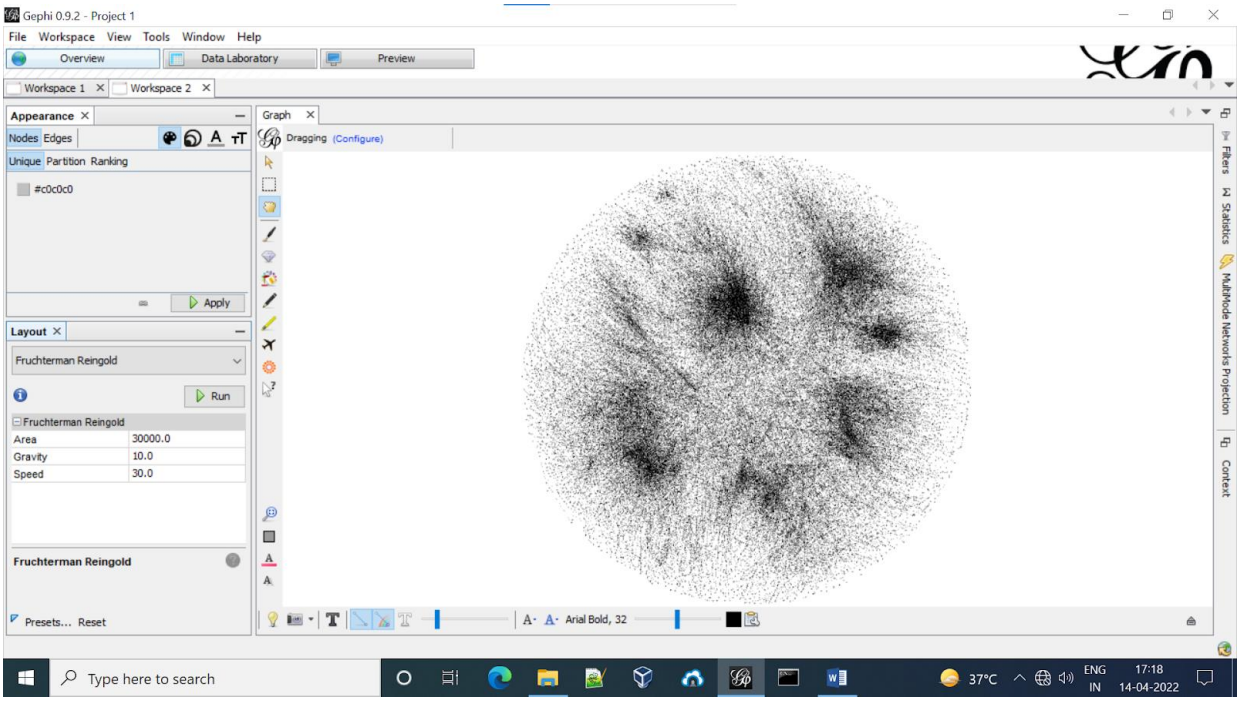
Finale output after importing both files is as below.



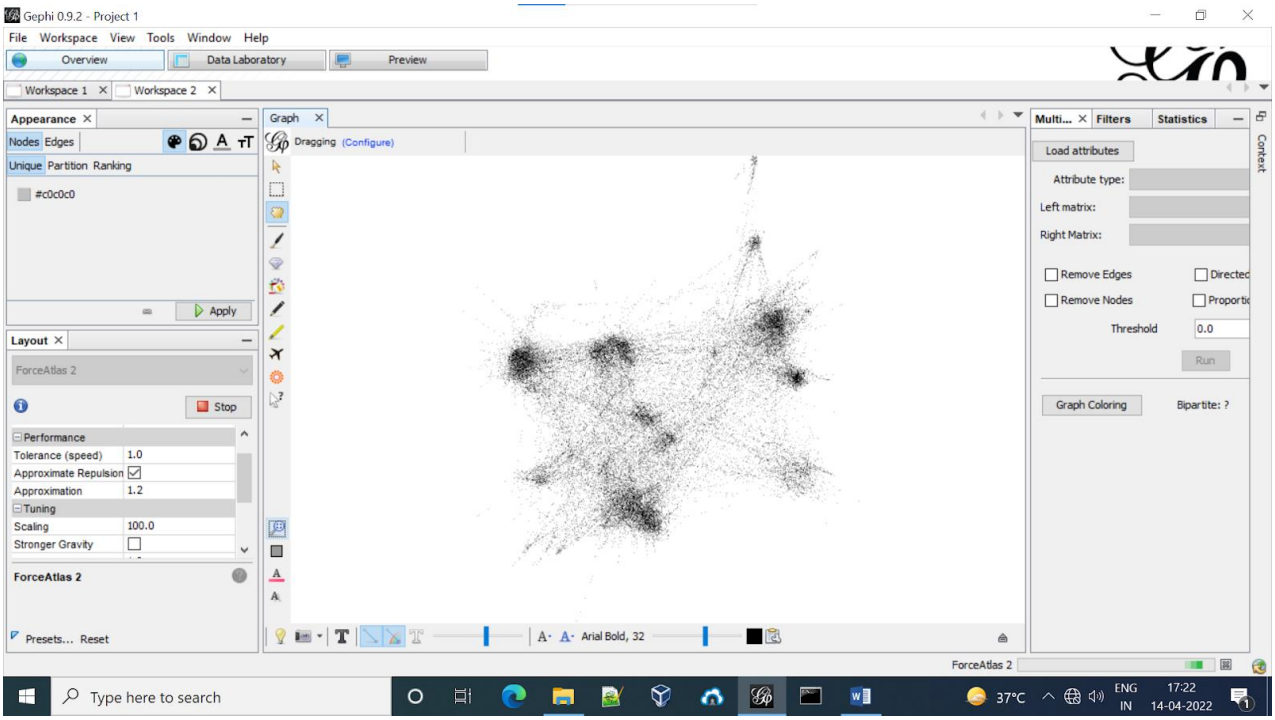
Graph processing

1. Select from Layout dropdown, option Fruchterman Reingold option with parameters as mentioned in the figure below. We set ‘Area’ to 30000 and ‘Speed’ to 30 as the number of nodes is very large. Large speed sacrifices precision. Click ‘Run’. Wait for at least for three minutes.

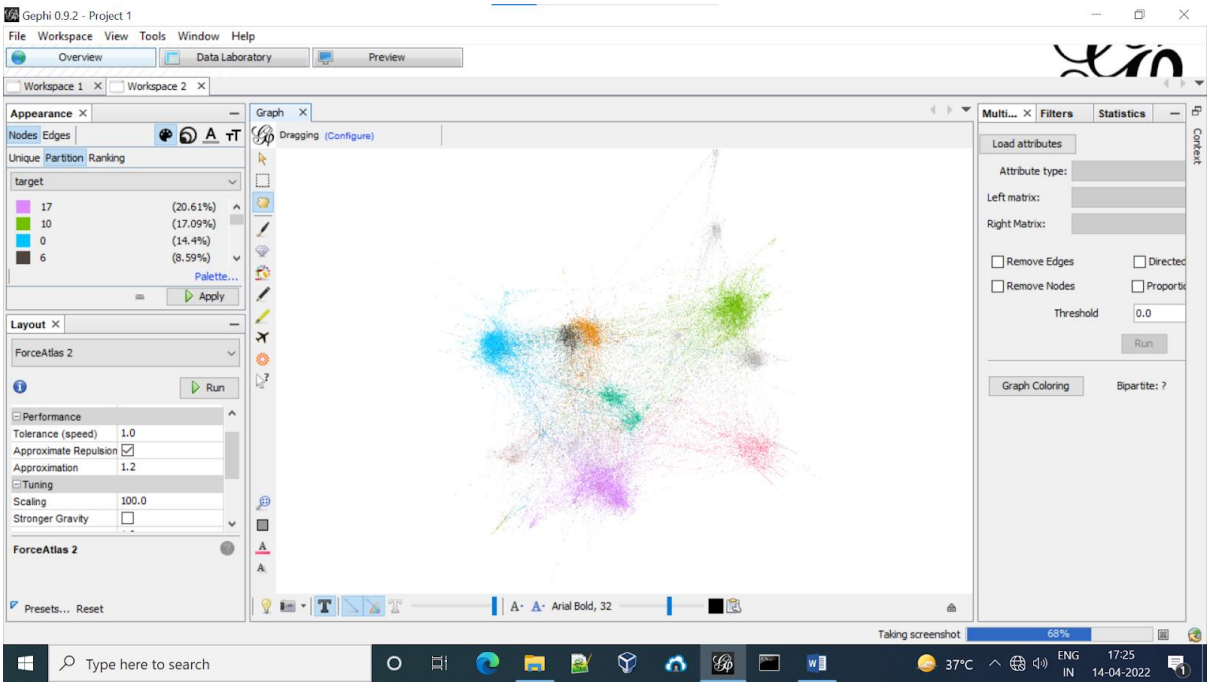
   Here is the graph after 3 minutes.



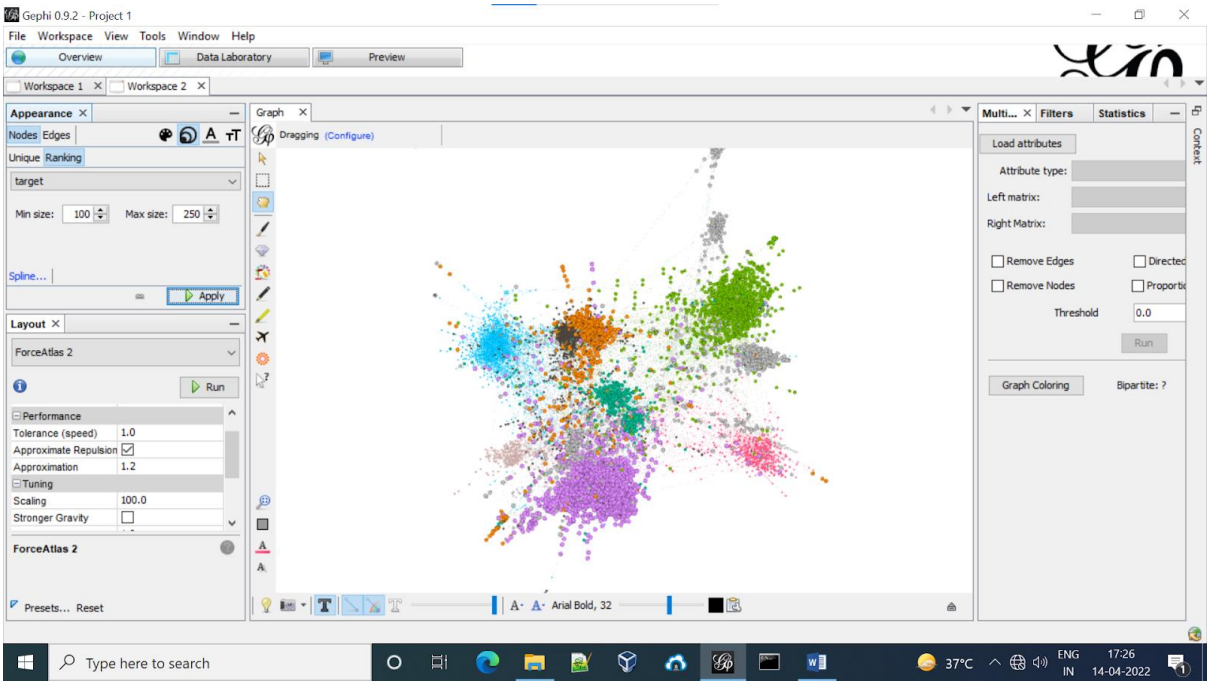
2. Next, select Layout ForceAtlas2. Set Scaling to 100. Scaling decides how much repulsion one wants. More scaling makes a more sparse graph. Also check Prevent Overlap.



3. Differentiating nodes on the basis of target and adding colors to them.



4. Increasing the size of nodes.



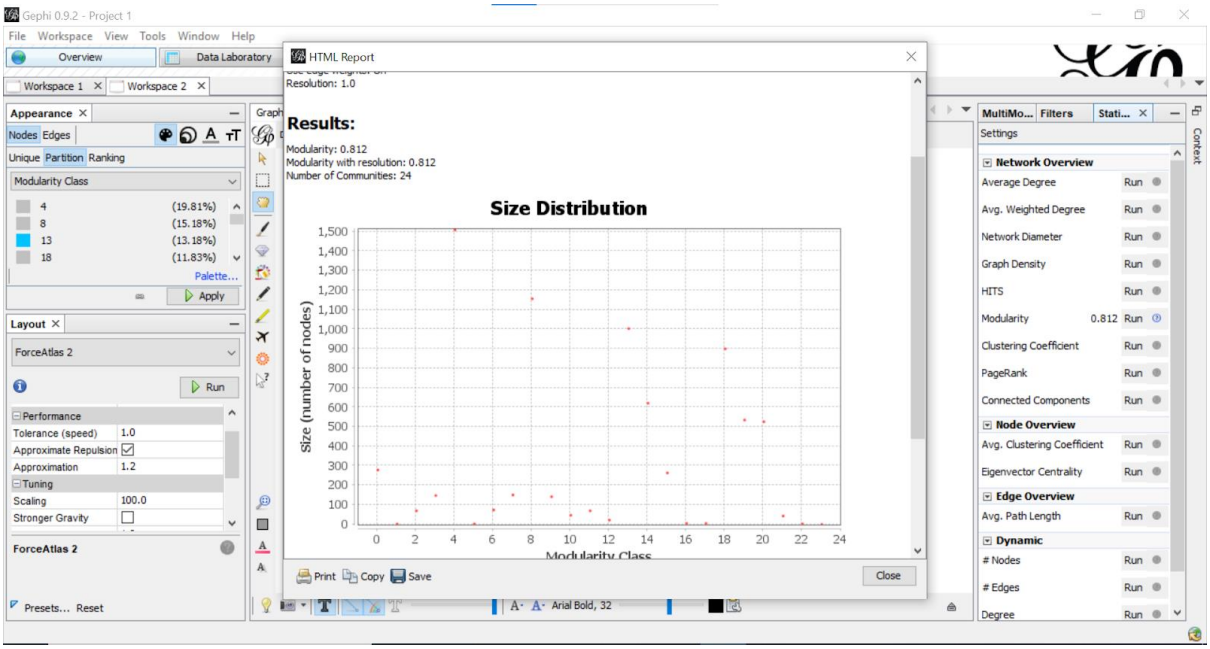
Filtration

Analyze networks with filters and communities.

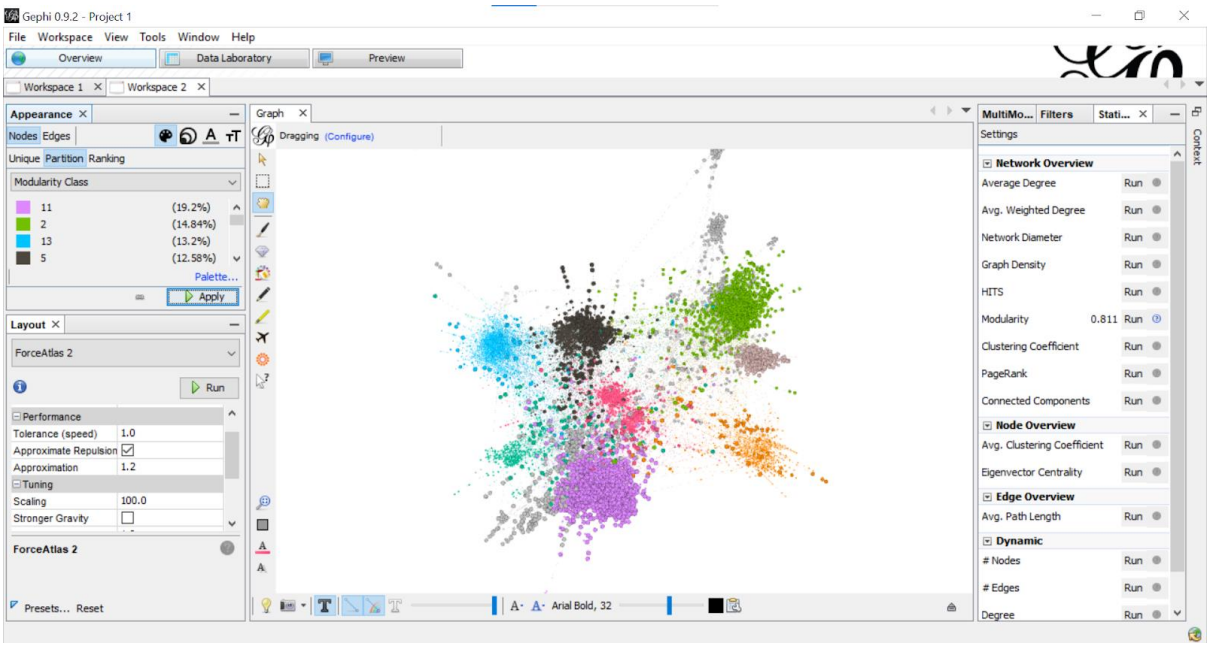
Modularity

Modularity is a measure of the structure of networks. It measures the strength of division of a network into modules or groups or communities. Networks with high modularity have dense connections between the nodes within modules but sparse connections between nodes in different modules. Modularity is often used in detecting community structure in networks.

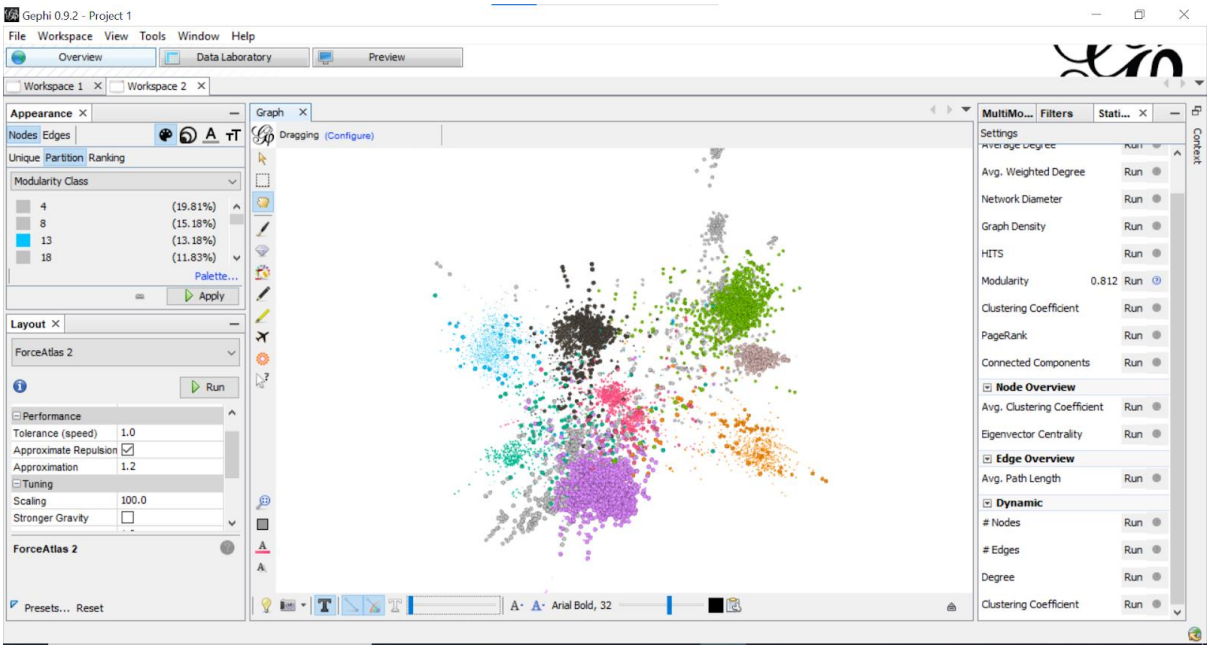
In the Statistics tab, click on Modularity. Repeated clicking on Modularity gave us a different number of communities. At 0.812. we got 24 communities. Keep communities to a minimum.



Coloring nodes as per communities. Mainly 5 communities out of total 24 communities dominate. These 5 communities are coloured as Magenta, Green, Black, Blue and Orange.



Removing blackness to have a better graph.

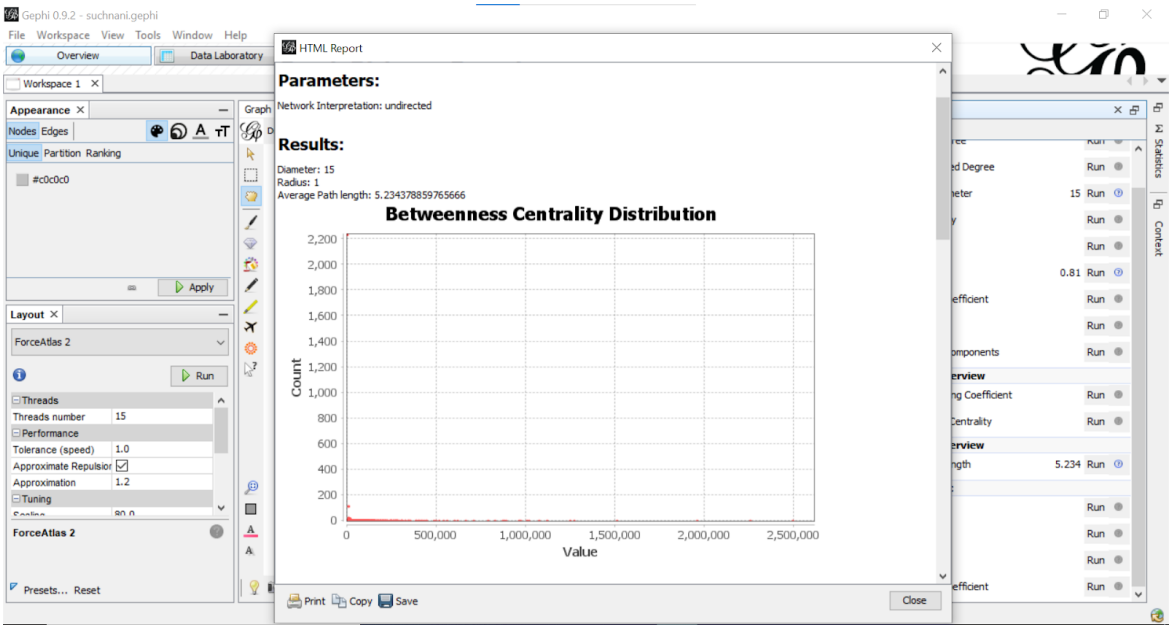


Checking betweenness

Betweenness centrality scores each node as the number of times it lies on the shortest path between other nodes. Betweenness centrality is a measure based on the number of shortest paths between any two nodes that pass through a particular node. Nodes around the edge of the network would typically have a low betweenness centrality.

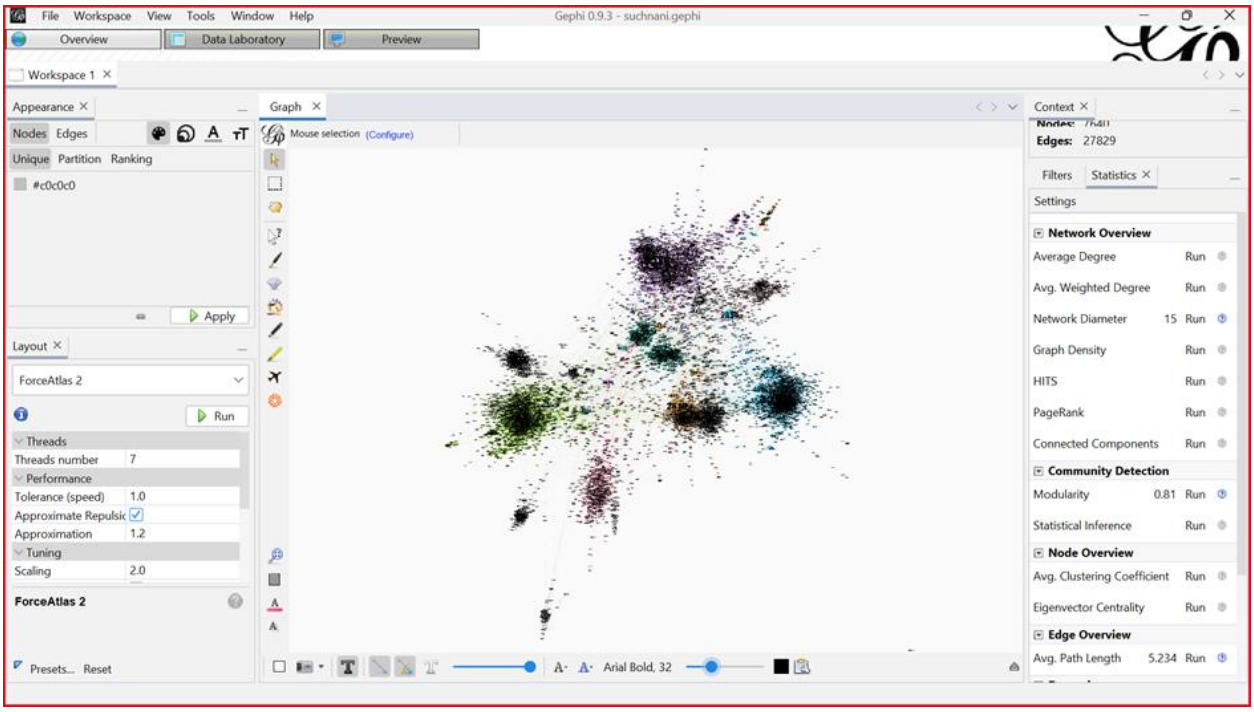
A high betweenness count could indicate someone holds authority over disparate clusters in a network, or just that they are on the periphery of both clusters.

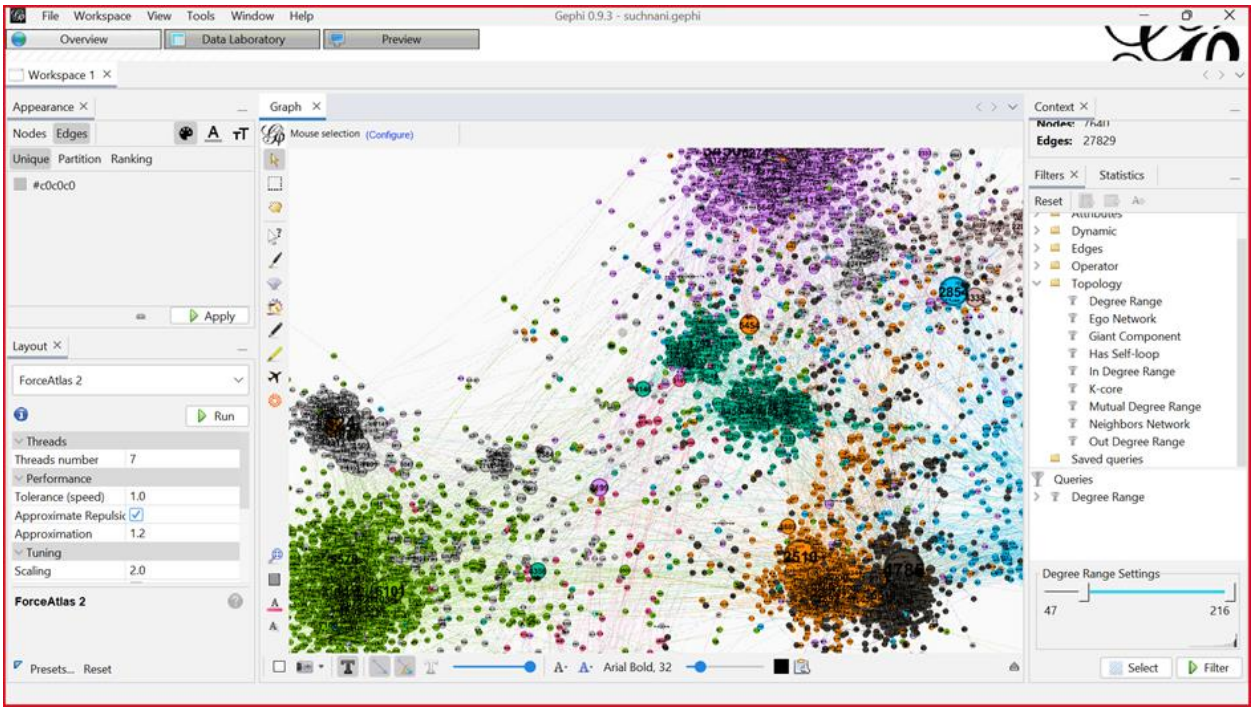
Click the Statistics tab in the top right module. Click Run next to Average Path Length. Here are results of betweenness centrality distribution. Average path length is 5.23.



Labels

Click the bold black T in the toolbar at the bottom of the window to turn labels on. Click the black letter A in the same toolbar to select the Size Mode for the labels, and choose the Node Size option.

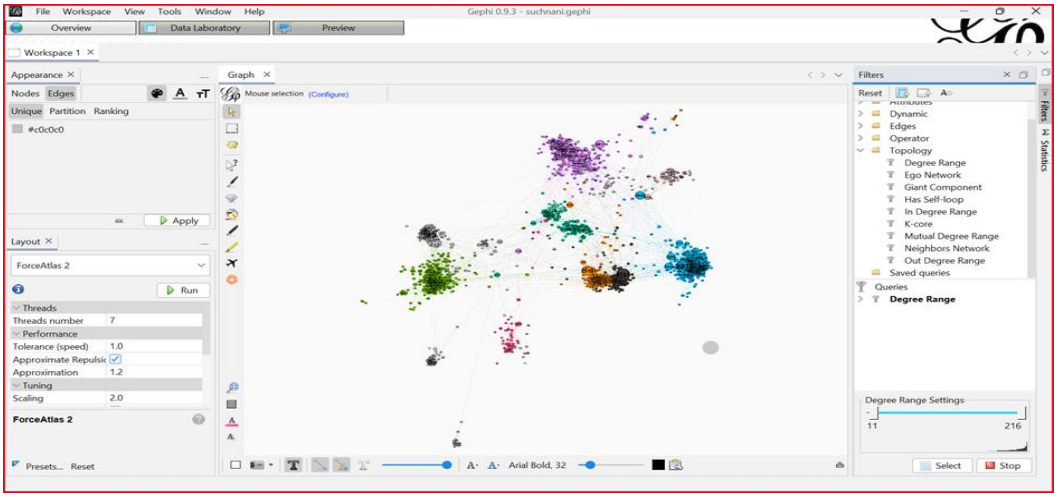




Using topology filter

Go to Filters in the top right module and open the Topology folder. Drag the Degree Range filter to the box below ("Drag filter here"). Click on Degree Range to open the Parameters, then edit the degree range settings by clicking on the "0" and changing it to 11. This option basically removes the "leaves" in the network that are not connected to many other nodes. We set the lower range to 11, meaning that it hides all nodes with less than 11 connections. Click Filter to apply.

Here are the results.

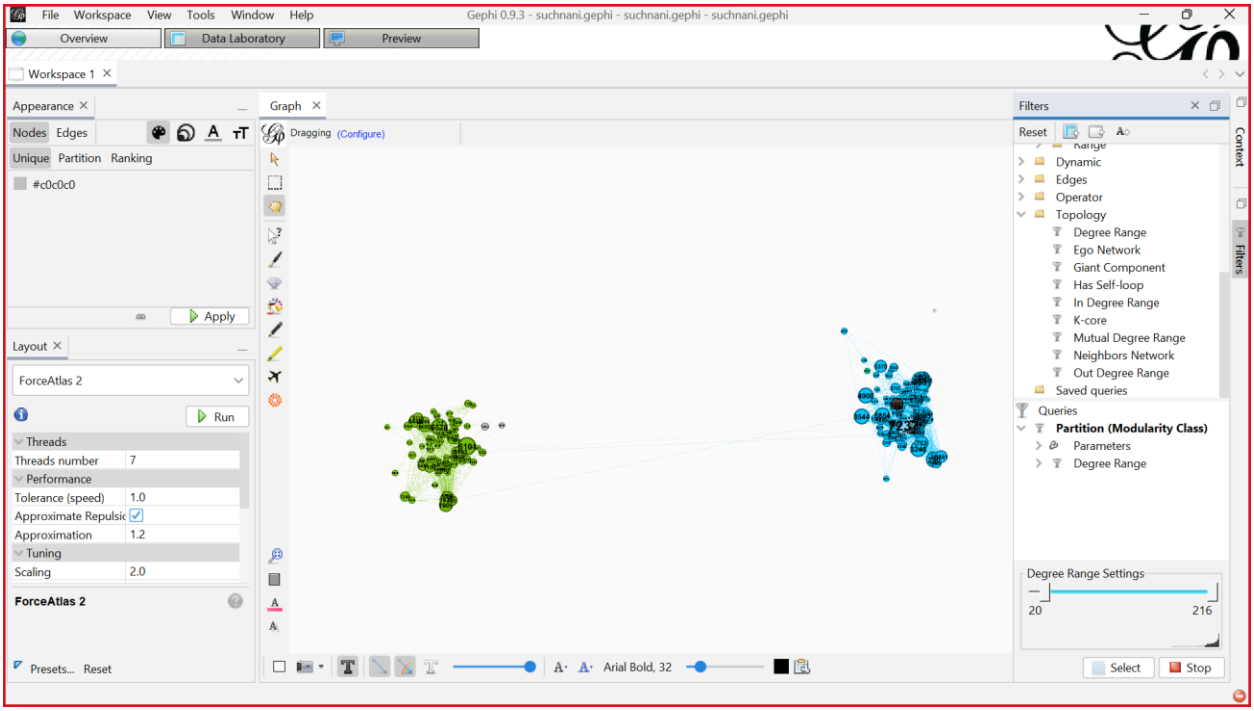


Using modularity class filter

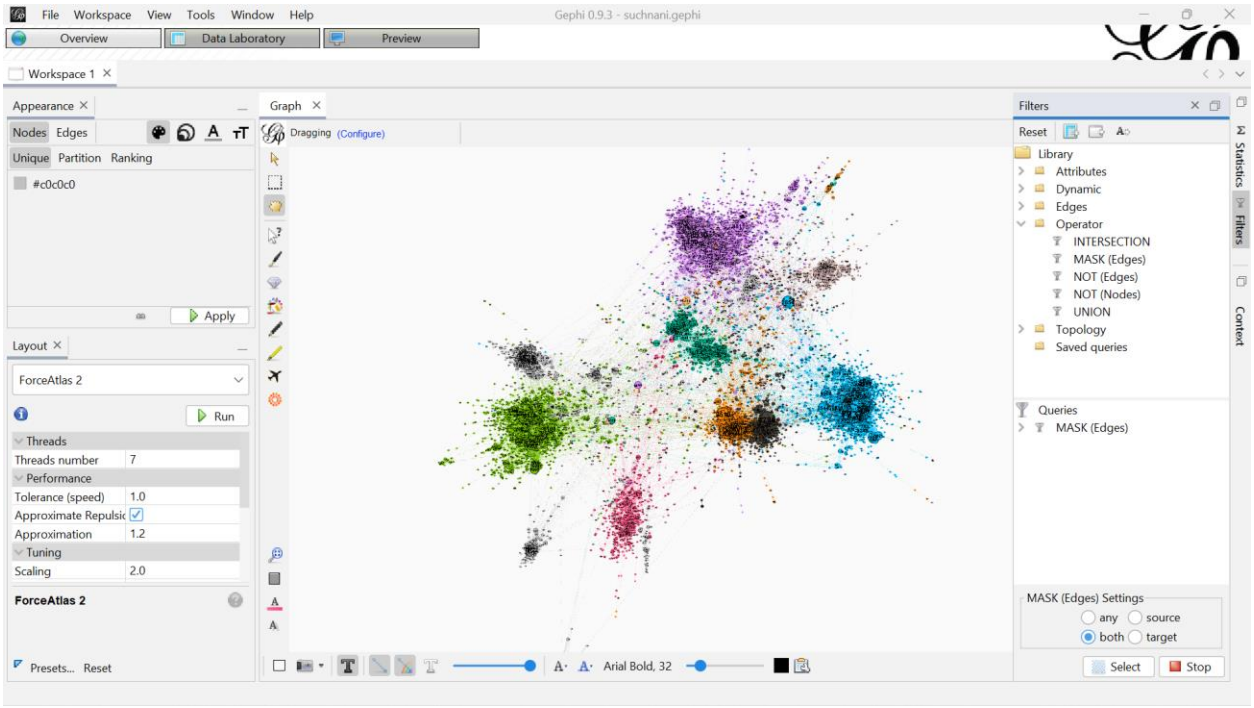
Combining 2 filters

Using 2 filters of modularity class and topology (20-216).

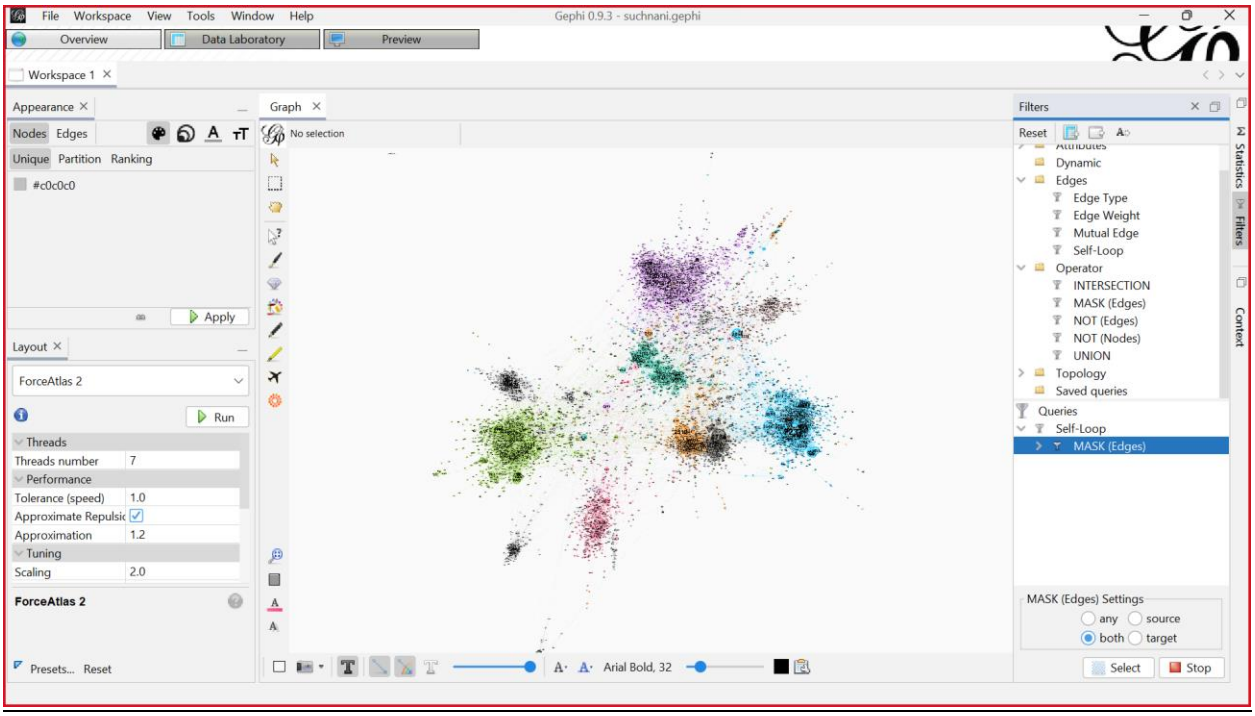
These 2 communities of green and blue color are most important and the nodes represented in below graph are the most important nodes from these communities having connections between minimum 20 to maximum 216.



Operator filter

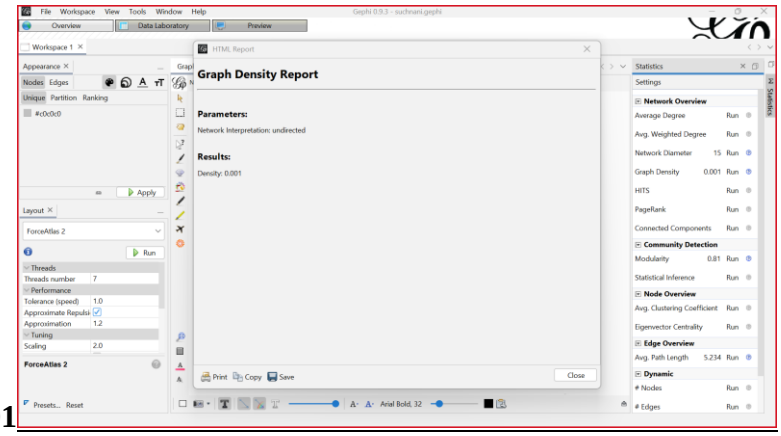


Edges self-loop filter



DENSITY

Density is an aggregate network metric used to describe the level of interconnectedness of the vertices. Density is a count of the number of relationships observed to be present in a network divided by the total number of possible relationships that could be present. It is a quantitative way to capture important sociological ideas like cohesion, solidarity, and membership.

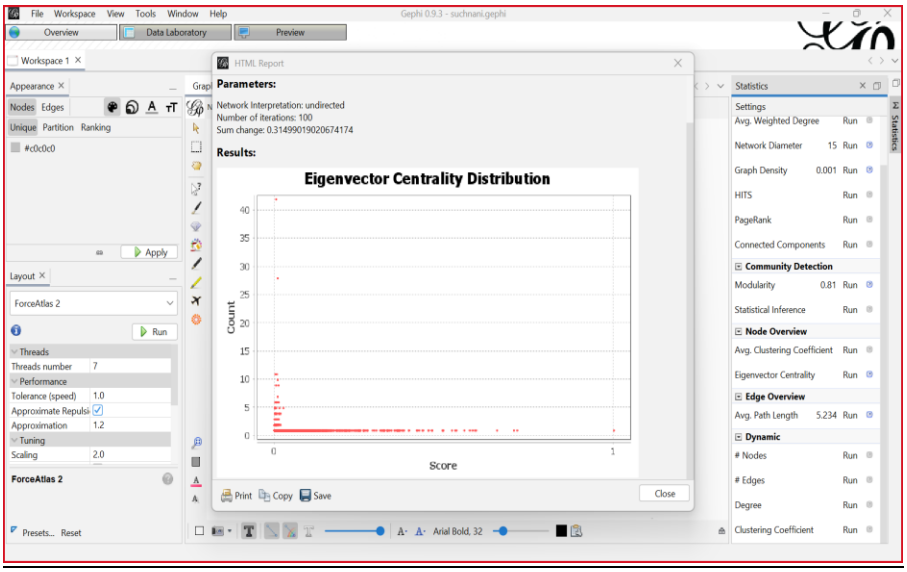


Density of our network is 0.001

EIGENVECTOR CENTRALITY DISTRIBUTION

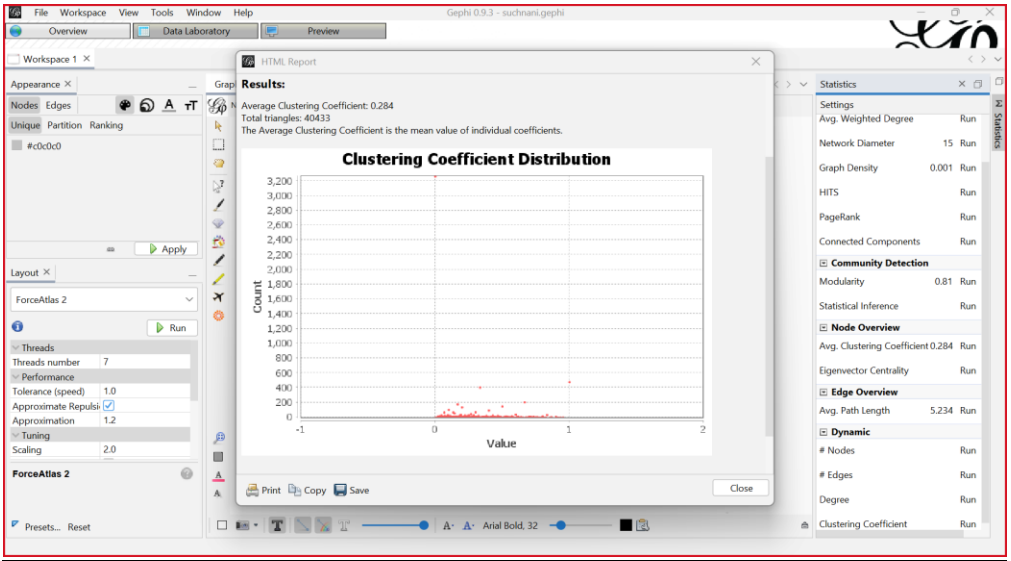
Eigenvector centrality (also called eigen centrality) is a measure of the influence of a node in a network. It assigns relative scores to all nodes in the network based on the concept that connections to high-scoring nodes contribute more to the score of the node in question than equal connections to low-scoring nodes.

Sum change of this network is 0.3149.

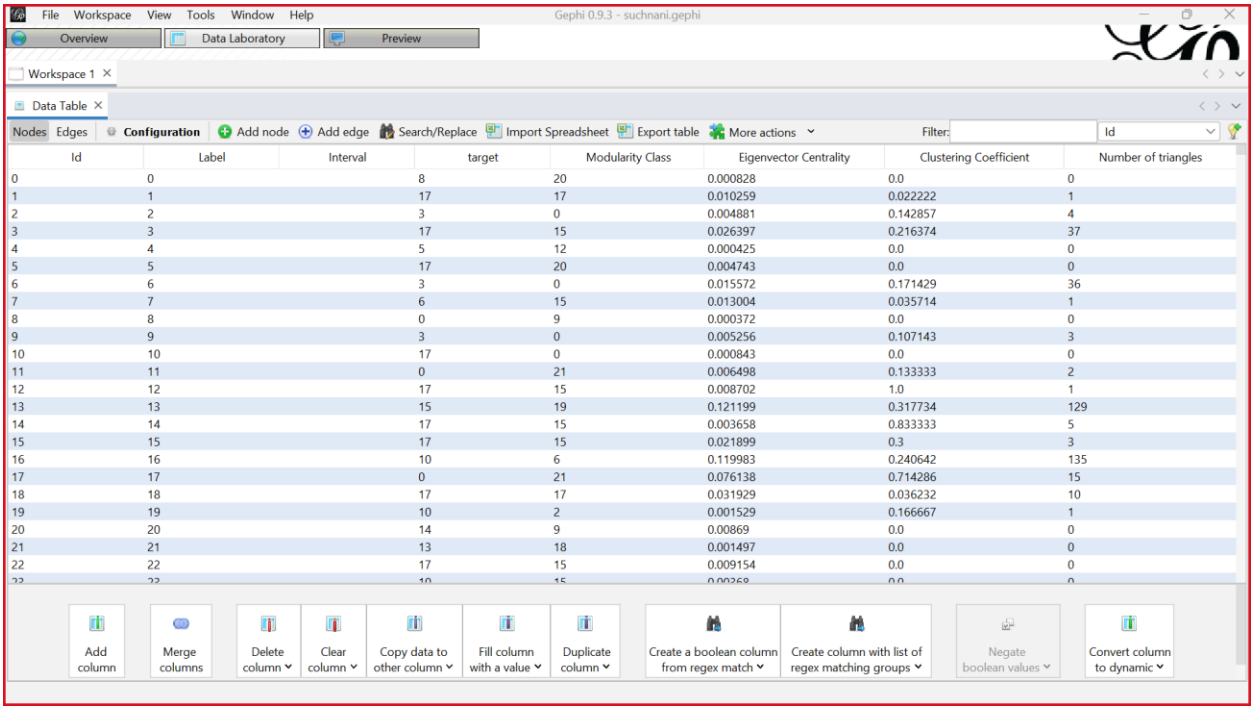


CLUSTERING COEFFICIENT

Clustering coefficient is a method used to detect community.

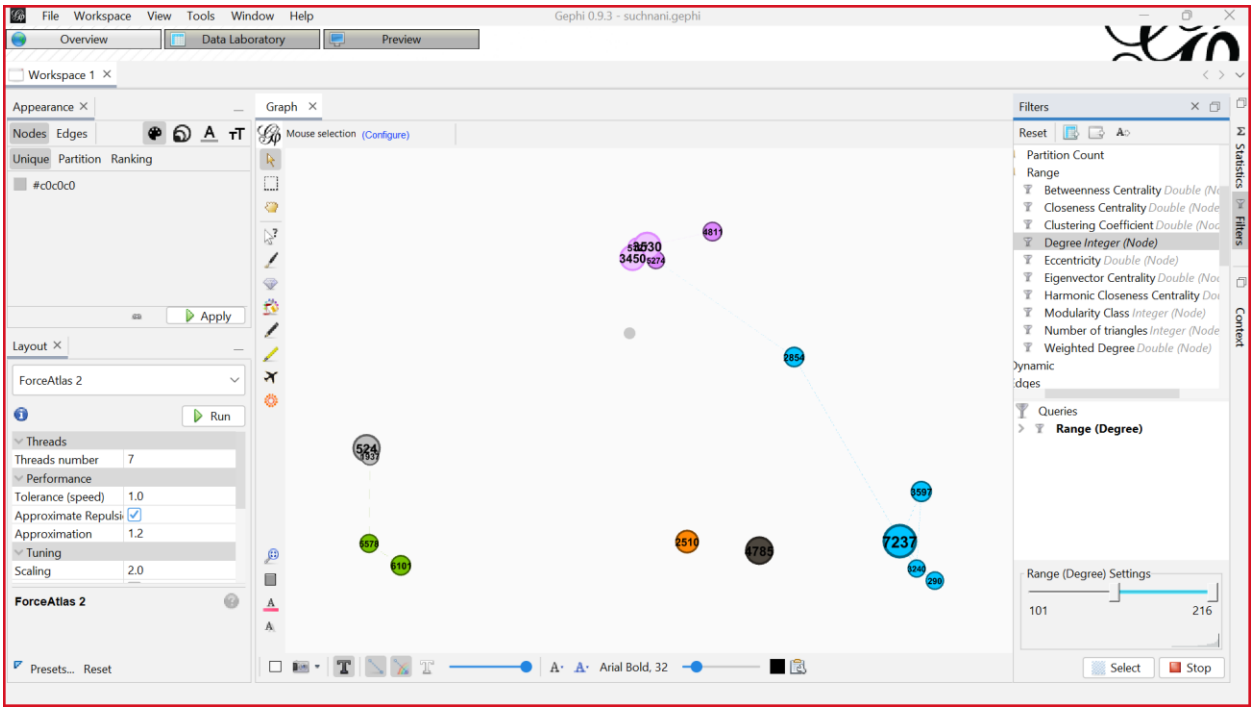


DATA LABORATORY GEPHI

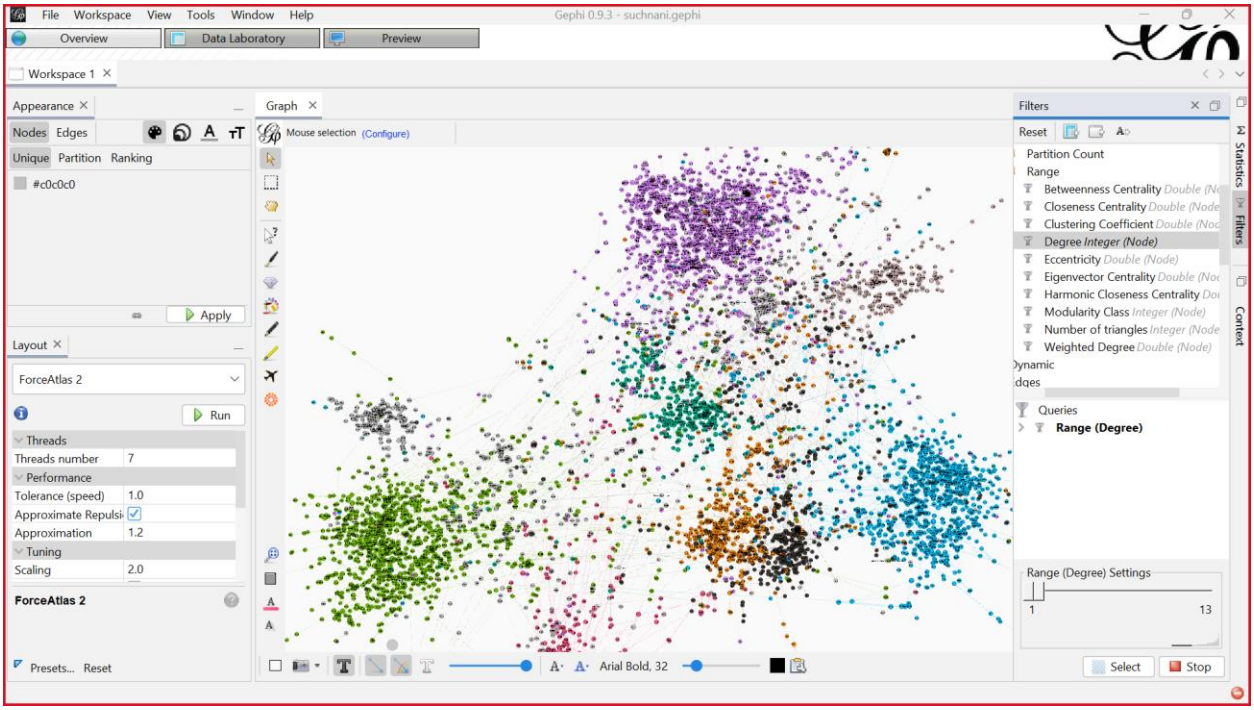


Id	Label	Interval	target	Modularity Class	Eigenvector Centrality	Clustering Coefficient	Number of triangles
0	0		8	20	0.000828	0.0	0
1	1		17	17	0.010259	0.022222	1
2	2		3	0	0.004881	0.142857	4
3	3		17	15	0.026397	0.216374	37
4	4		5	12	0.000425	0.0	0
5	5		17	20	0.004743	0.0	0
6	6		3	0	0.015572	0.171429	36
7	7		6	15	0.013004	0.035714	1
8	8		0	9	0.000372	0.0	0
9	9		3	0	0.005256	0.107143	3
10	10		17	0	0.000843	0.0	0
11	11		0	21	0.006498	0.133333	2
12	12		17	15	0.008702	1.0	1
13	13		15	19	0.121199	0.317734	129
14	14		17	15	0.003658	0.833333	5
15	15		17	15	0.021899	0.3	3
16	16		10	6	0.119983	0.240642	135
17	17		0	21	0.076138	0.714286	15
18	18		17	17	0.031929	0.036232	10
19	19		10	2	0.001529	0.166667	1
20	20		14	9	0.00869	0.0	0
21	21		13	18	0.001497	0.0	0
22	22		17	15	0.009154	0.0	0

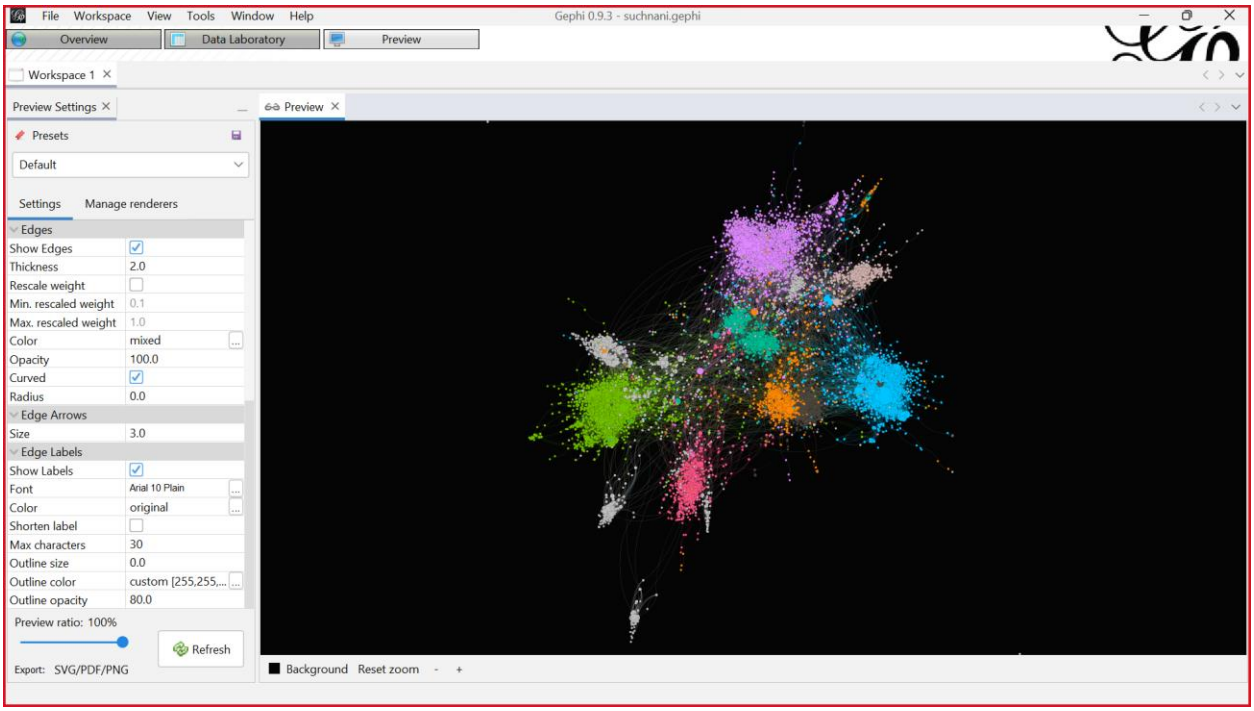
Important nodes of the network as these have high degree and have a number of connections more than 100.



These are less important nodes as these have low degree and have a number of connections less than 13.

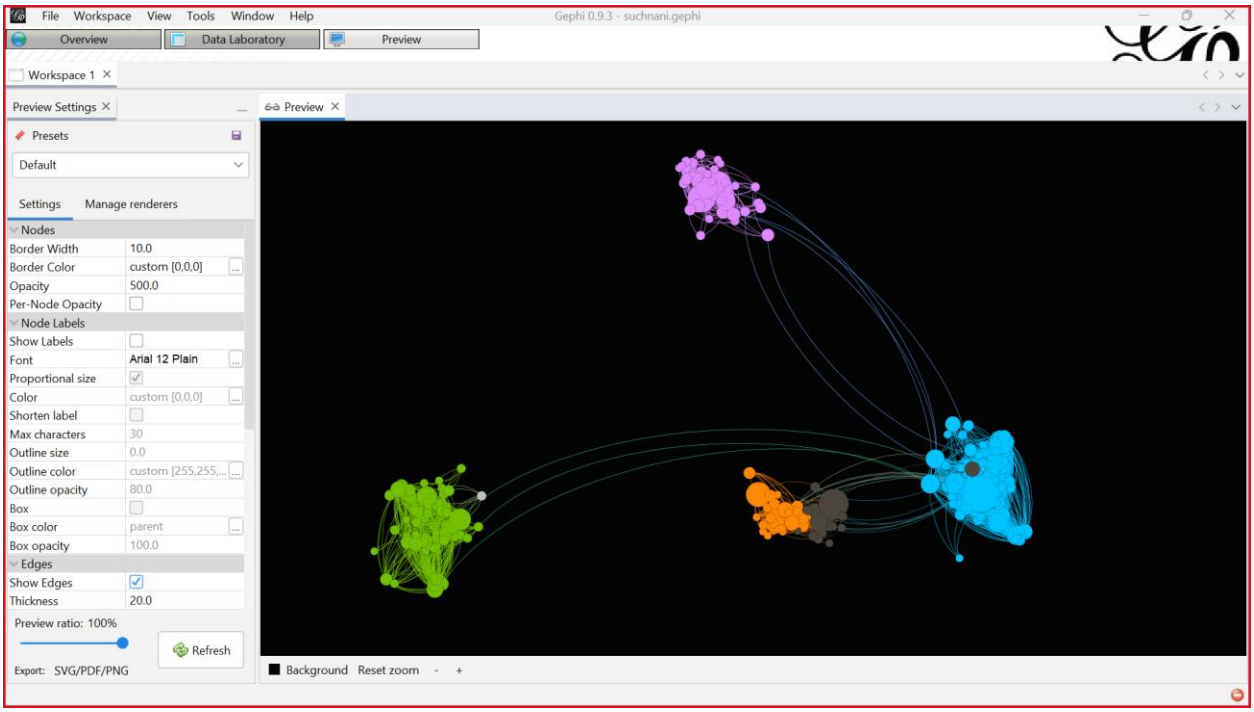


Preview



Filtered preview

Most important 4 communities with degree range more than 10.



Examining the final results

Betweenness - Diameter - 15

Radius - 1

Average Path Length - 5.23

Density - 0.001

Average Degree - 7.28

EigenVector Centrality - Sum Change - 0.3149

Average Clustering Coefficient - 0.284