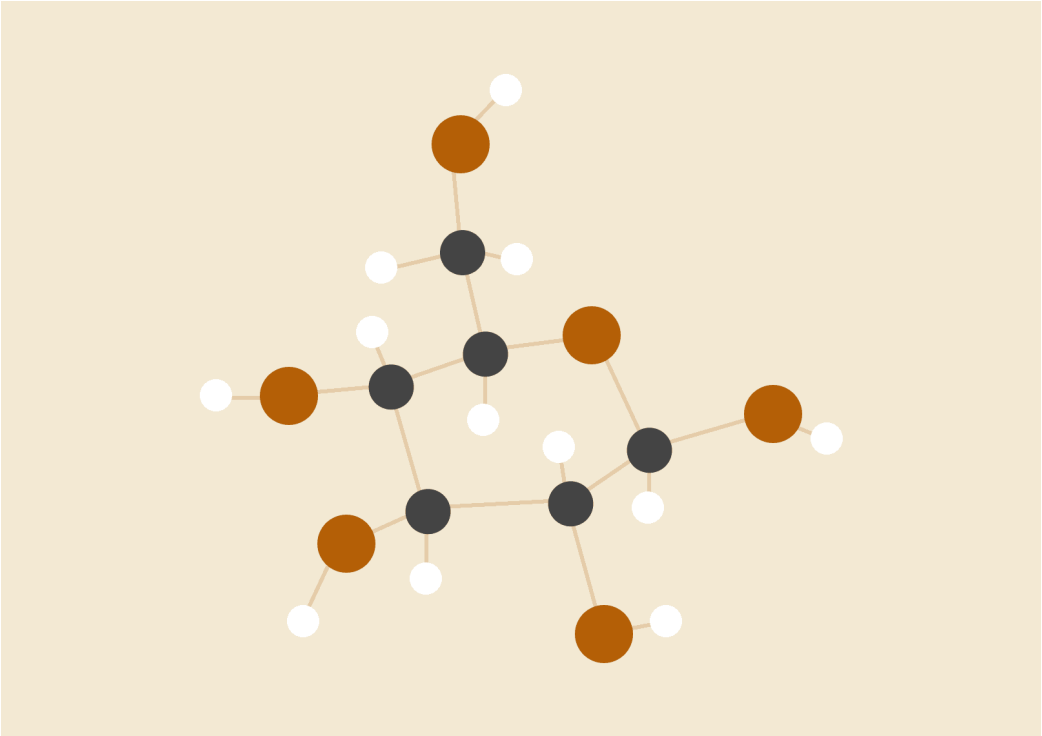
A NETWORK ANALYSIS OF HUMAN DISEASES

GROUP-3 | FODBMS | TERM-3 | BDA-02-G

FORE SCHOOL OF MANAGEMENT, NEW DELHI



Ayushi Arora - 025010

Dhruv Khosla - 025012

Akhil Sharma - 025046

15th April 2022

Fundamentals of Database Management Systems

Submitted to
Prof. Ashok Harnal

INTRODUCTION

For the network analysis, we chose to focus on a network of human diseases and gene types. We wanted to not only explore the visualization but also to discover specific gene types and their relationship to diseases. We also wanted to understand the links between neurological diseases because many people were diagnosed with these types of diseases specifically Alzheimer's and Dementia.

MATERIALS

When we first started this lab, we knew that we wanted to find a dataset that is related to specifically human diseases including neurological diseases. Over the past few years, we have heard several news stories about people who were diagnosed with Alzheimer's and Dementia. We not only wanted to understand the disease but to also see how it links to various other diseases. We started the project by doing a Google search using the keywords "network human diseases". We found a network visualization of human diseases from the Exploring Data website which focuses on interactive data visualizations from open source tools.

Next, we went to the Gephi Wiki which houses a variety of datasets on GitHub, a development platform. We then chose the Diseasome dataset and downloaded it as a zip file. We then downloaded Gephi, an open-source visualization and exploration software for networks and graphs.

PROCEDURE

We started the lab by opening the Diseasome file in Gephi. We did an inspection of the data to make sure there weren't any issues. We then wanted to take a look at the nodes in the data set so we clicked on the Data Laboratory button and then clicked on Nodes. We saw that there was a label column and type column that specified either the disease or gene type.

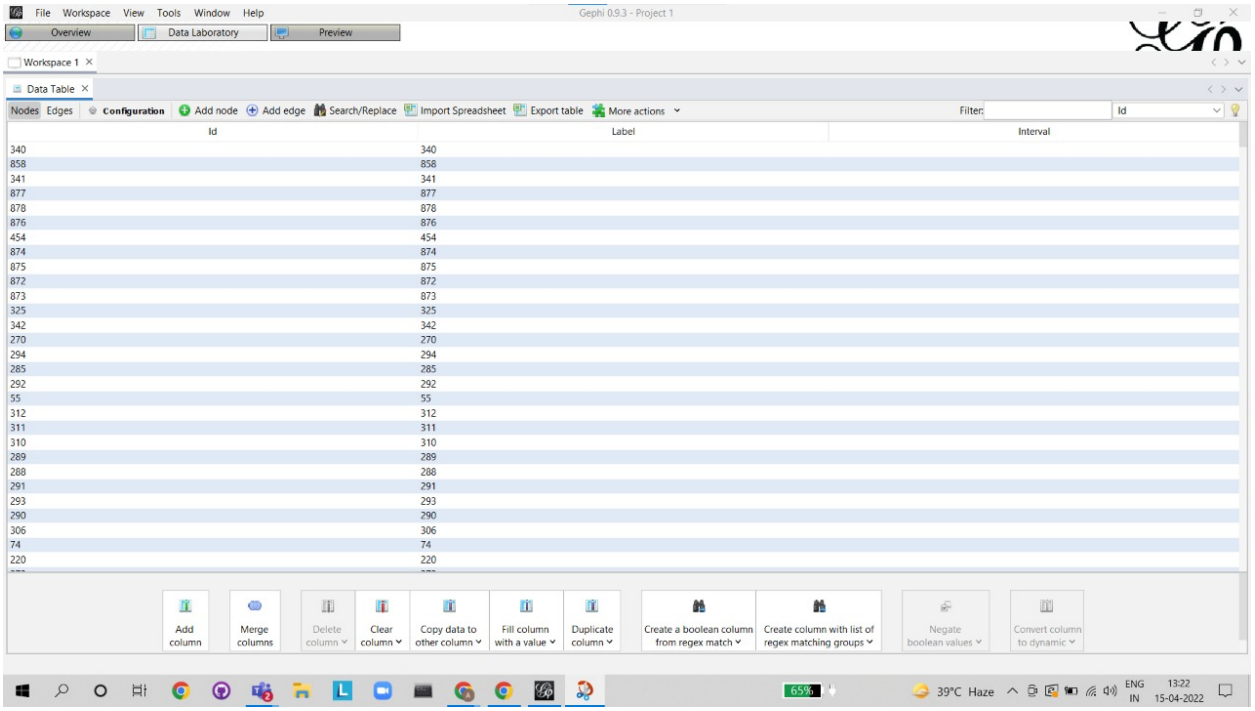


Fig. 1

Next, we began to experiment with the overall visualization of the data set. In the Appearance section, we clicked on Nodes then Size. We wanted to adjust the size of the nodes so that they could stand out more. We clicked on ranking and changed the attribute to “Betweenness Centrality” then adjusted the minimum size to 25 and the maximum size to 400. I then changed the layout of the visualization by clicking on “Force Atlas” under the Layout section. We adjusted the Repulsion strength to “10,000”.

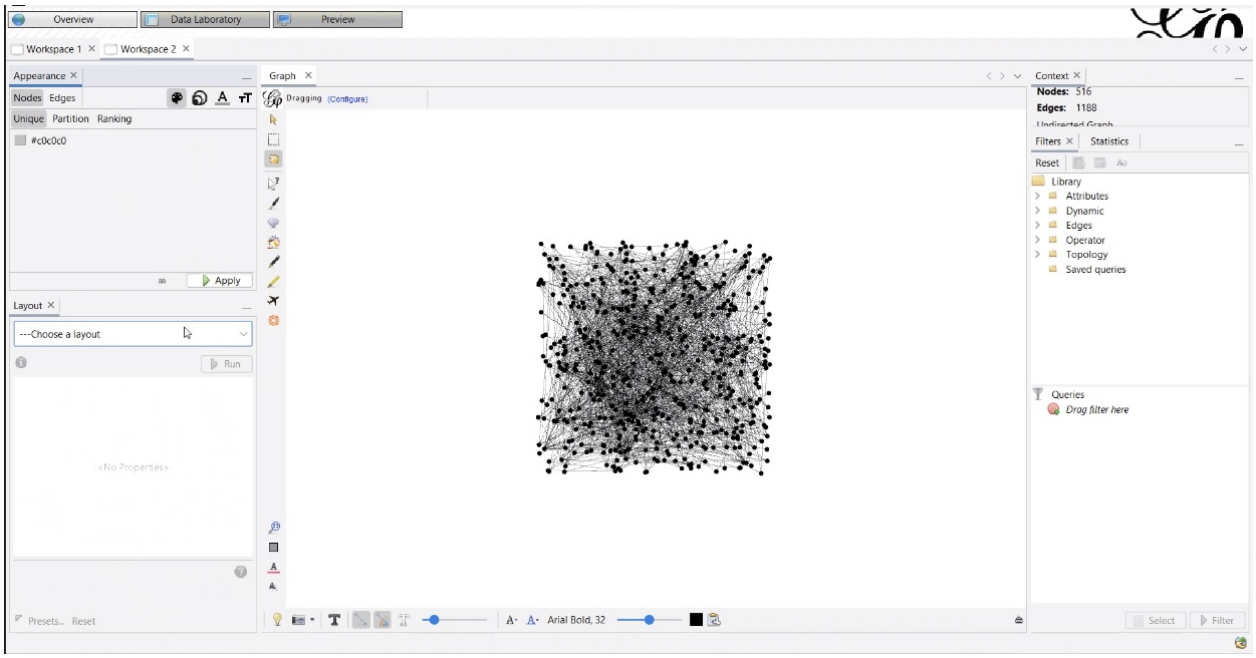

Overview
Data Laboratory
Preview
Workspace 1
Workspace 2
Appearance
Nodes Edges
Unique Partition Ranking
#c0c0c0
Apply
Layout
---Choose a layout
Run
<No Properties>
Presets... Reset
Graph
Dragging (Configure)
Arial Bold, 32
Context
Nodes: 516
Edges: 1188
Filters
Statistics
Reset
Library
Attributes
Dynamic
Edges
Operator
Topology
Saved queries
Queries
Drag filter here
Select
Filter

Fig. RAW

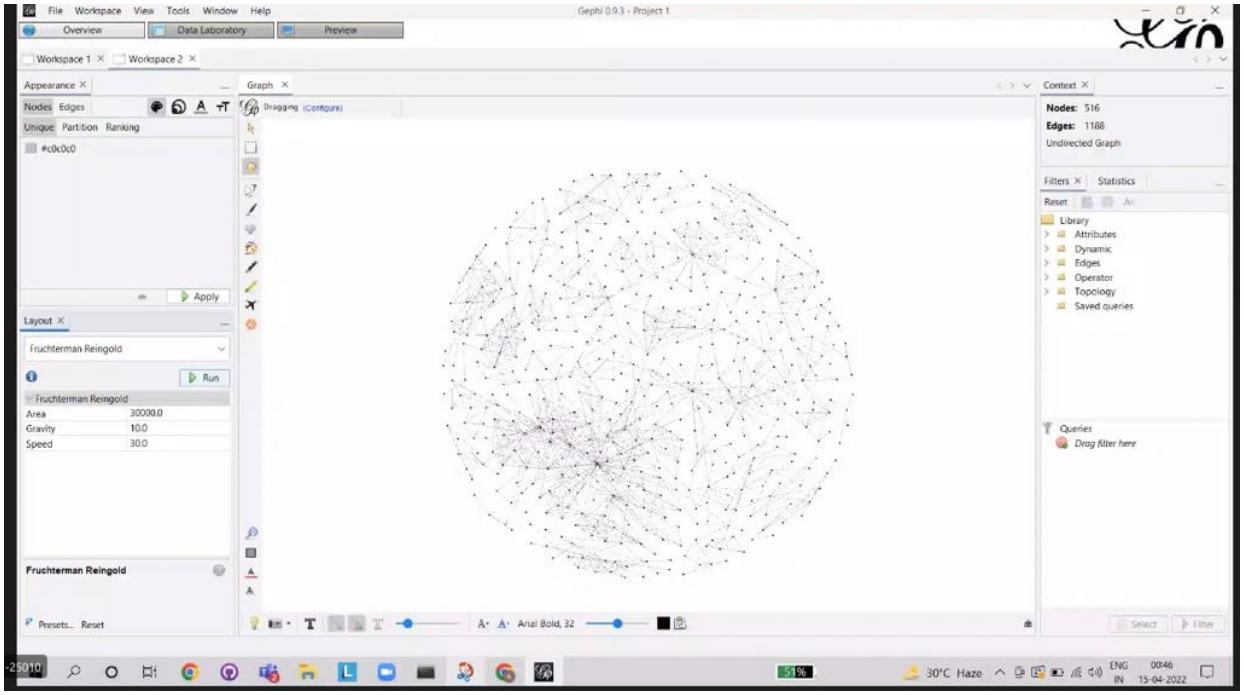

File Workspace View Tools Window Help
Gephi 0.9.3 - Project 1
Overview
Data Laboratory
Preview
Workspace 1
Workspace 2
Appearance
Nodes Edges
Unique Partition Ranking
#c0c0c0
Apply
Layout
Fruchterman Reingold
Run
Fruchterman Reingold
Area 30000.0
Gravity 10.0
Speed 30.0
Fruchterman Reingold
Presets... Reset
Graph
Dragging (Configure)
Arial Bold, 32
Context
Nodes: 516
Edges: 1188
Undirected Graph
Filters
Statistics
Reset
Library
Attributes
Dynamic
Edges
Operator
Topology
Saved queries
Queries
Drag filter here
Select
Filter
30°C Haze
ENG
IN
00:46
15-04-2022

Fig. 2

We then added the text labels to the visualizations and adjusted the text size for the nodes by clicking on the drop down field at the far right of the bar and changing the Size field to “Node size”. We used the algorithm of ForceAtlas 2 to make the nodes separate and form a prominent looking network, which is easier to study.

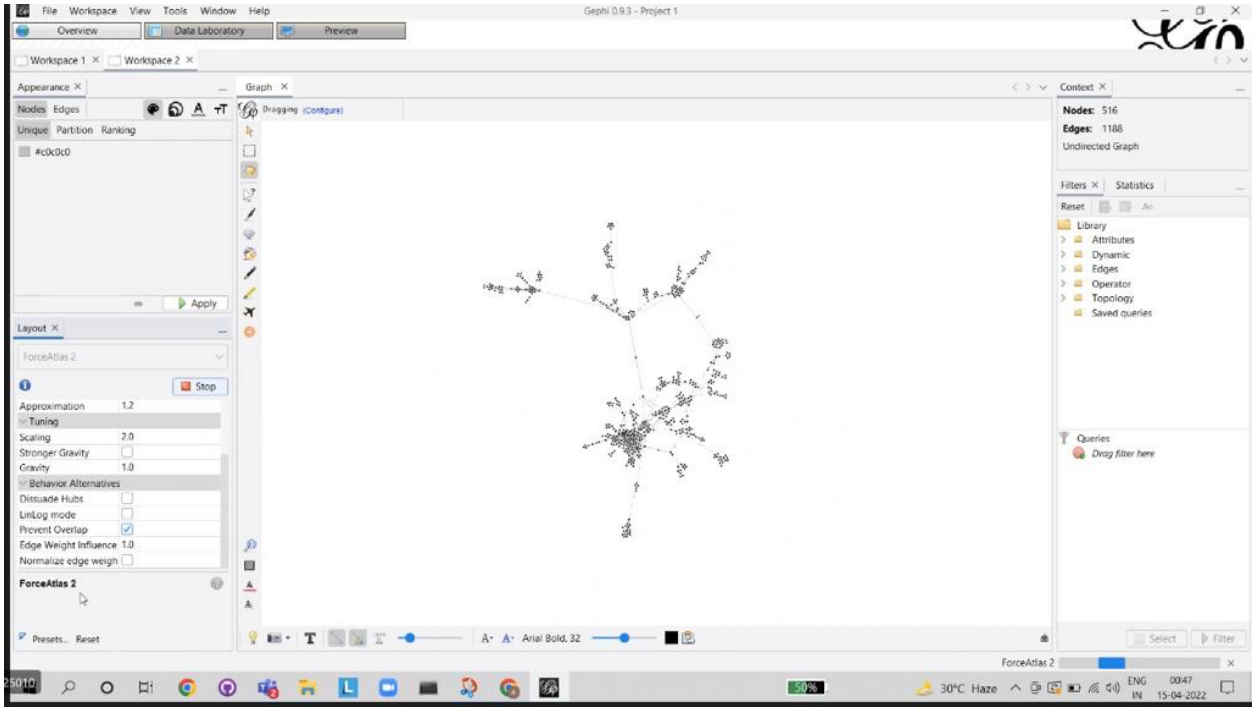


Fig. 3

We found that the value of Modularity was 0.819 and the number of communities were found to be 24 in number (refer Fig. 4).

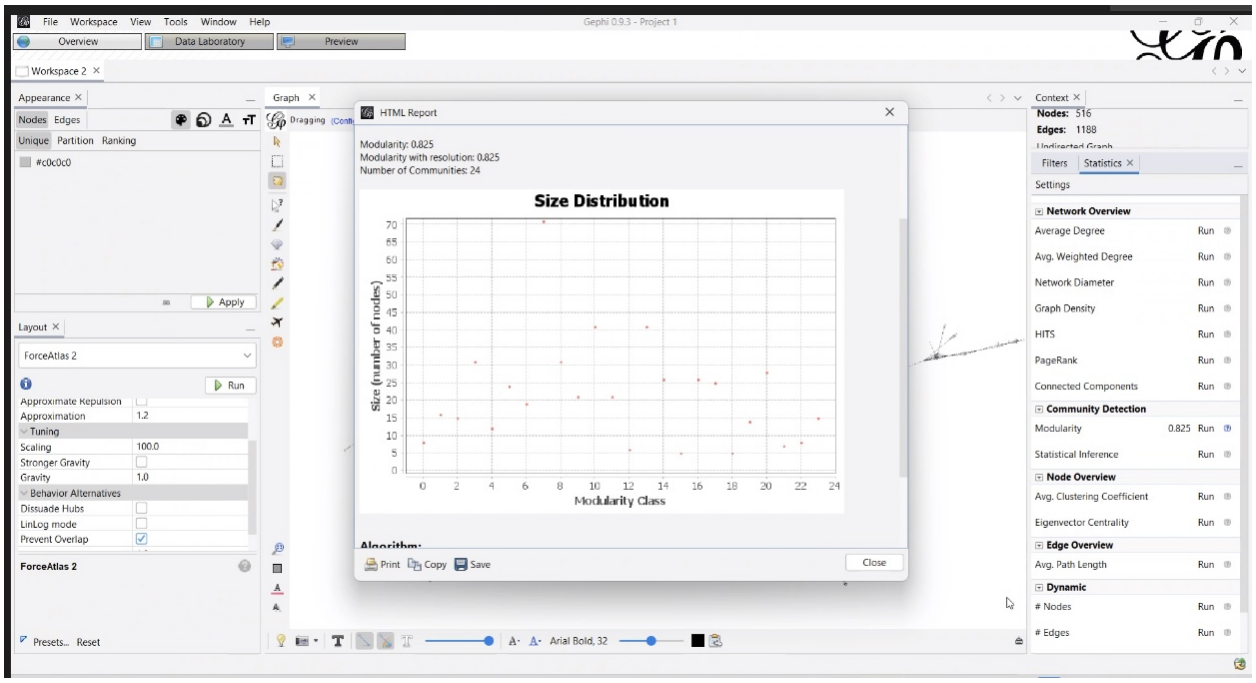


Fig. 4

Next, we took a look at the overall visualization and made adjustments by moving specific nodes around in order to have a prominent standout.

The following image (Fig. 5) shows the final result of the visualization.

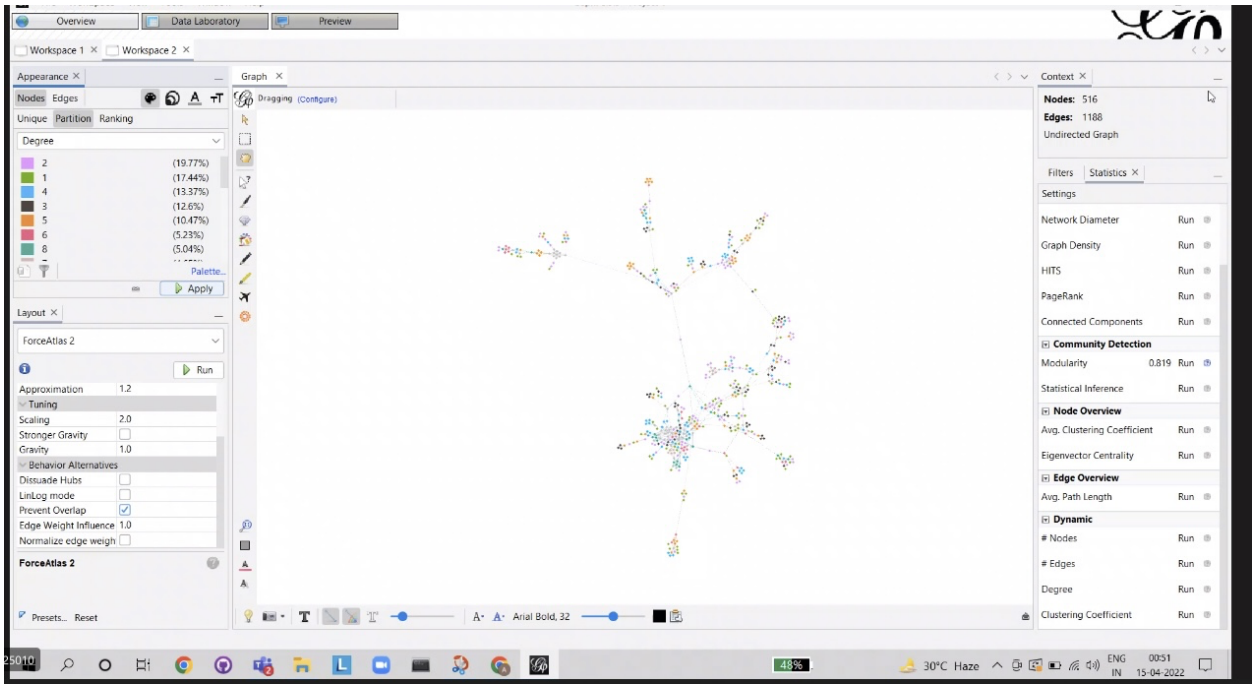


Fig. 5

We then wanted to see the linked diseases to Alzheimer's, so we clicked on the respective node. We saw that dementia and amyloidosis were both linked to Alzheimer's diseases. However, we were surprised to discover that schizophrenia was also linked to Alzheimer's.

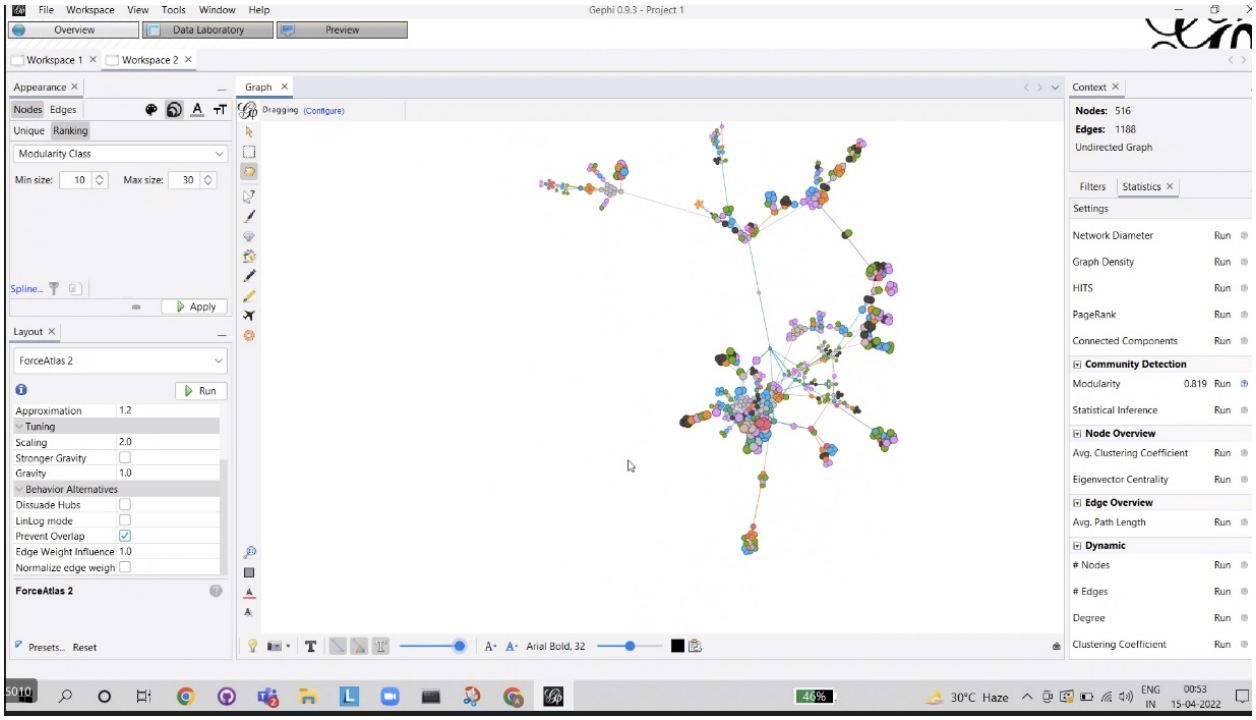

File Workspace View Tools Window Help
Gephi 0.9.3 - Project 1
Overview
Data Laboratory
Preview
Workspace 1
Workspace 2
Appearance
Nodes Edges
Unique Ranking
Modularity Class
Min size: 10
Max size: 30
Spline...
Apply
Layout
ForceAtlas 2
Run
Approximation 1.2
Tuning
Scaling 2.0
Stronger Gravity
Gravity 1.0
Behavior Alternatives
Dissuade Hubs
LinLog mode
Prevent Overlap
Edge Weight Influence 1.0
Normalize edge weigh
ForceAtlas 2
Presets... Reset
Graph
Dragging (Configure)
Arial Bold, 32
Context
Nodes: 516
Edges: 1188
Undirected Graph
Filters
Statistics
Settings
Network Diameter Run
Graph Density Run
HITS Run
PageRank Run
Connected Components Run
Community Detection
Modularity 0.819 Run
Statistical Inference Run
Node Overview
Avg. Clustering Coefficient Run
Eigenvector Centrality Run
Edge Overview
Avg. Path Length Run
Dynamic
Nodes Run
Edges Run
Degree Run
Clustering Coefficient Run
46%
30°C Haze
ENG
IN
00:53
15-04-2022

Fig. 6

File Workspace View Tools Window Help
Gephi 0.9.3 - Project 1
Overview
Data Laboratory
Preview
Workspace 2
Appearance
Nodes Edges
Unique Ranking
Degree
Min size: 50
Max size: 150
Spline...
Apply
Layout
ForceAtlas 2
Run
Approximate Repulsion
Approximation 1.2
Tuning
Scaling 100.0
Stronger Gravity
Gravity 1.0
Behavior Alternatives
Dissuade Hubs
LinLog mode
Prevent Overlap
ForceAtlas 2
Presets... Reset
Graph
Dragging (Configure)
Arial Bold, 32
Context
Nodes: 112 (21.71% visible)
Edges: 323 (27.19% visible)
Filters
Statistics
Reset
Attributes
Equal
Inter Edges
Intra Edges
Non-null
Partition
Modularity Class (Node)
Partition Count
Range
Dynamic
Edges
Operator
Topology
Degree Range
Ego Network
Queries
Partition (Modularity Class)
Partition (Modularity Class) Settings
7 (13.76%)
10 (7.95%)
13 (7.95%)
3 (6.01%)
5 (6.01%)
Select
Stop

Fig. 7

We used the filters of Topology, Degree Range and Attributes, Partition, Modularity to check the strength of communities and their network.

Finally, we wanted to see the overall visualization similar to the Explore Data visualization.

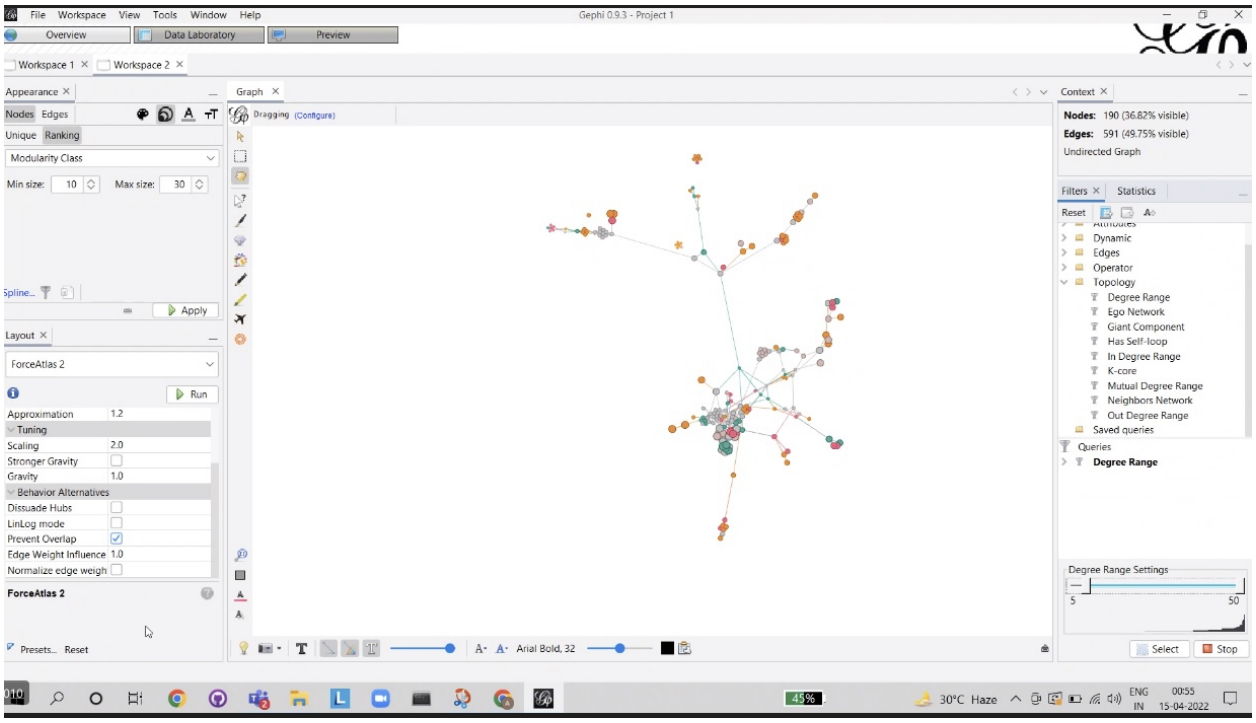


Fig. 8

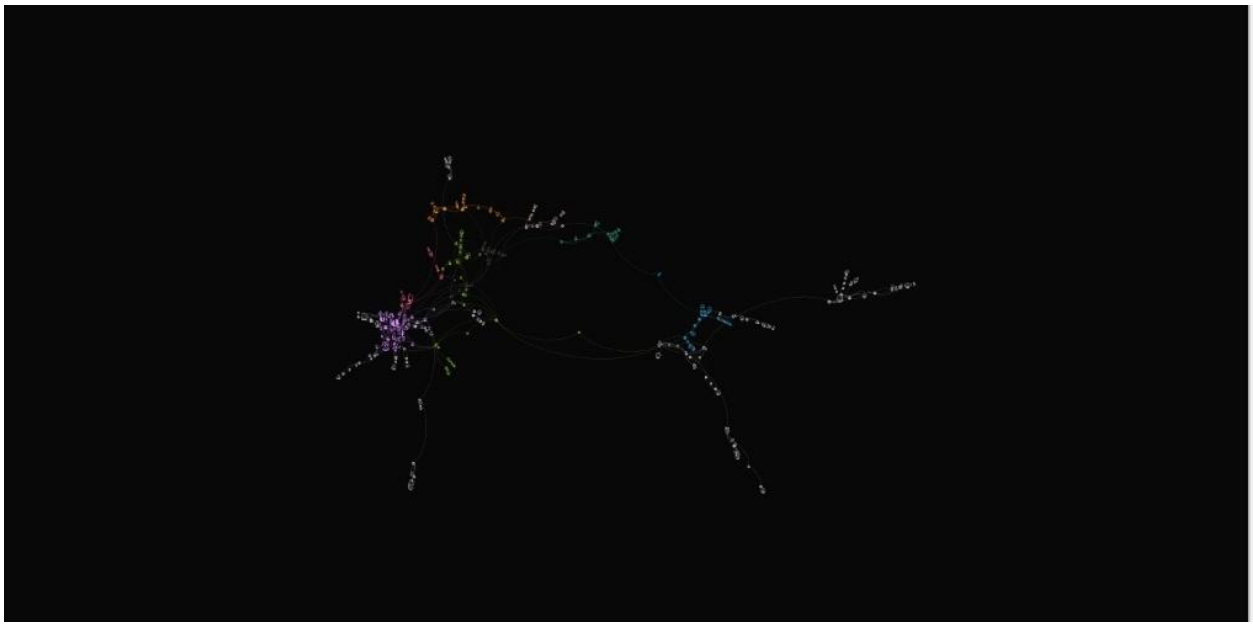


Fig. 9

REFLECTIONS

Overall, we found this lab to be quite challenging as well as informative. We were not so precisely familiar with Gephi before this lab assignment and found the software to be not particularly user-friendly. It took us several attempts to truly create the visualization that we wanted. We did like the ability to adjust the layout and appearance of the visualization with a variety of properties. We also liked being able to drill down on a particular disease to see what was linked to it and being surprised by these findings.

In the future, we would like to see what other layouts we could use for this dataset. We think it would be interesting to see how I could display the networks of diseases in a variety of ways. We would also like to use Gephi to explore other datasets to see what insights we could derive. We also want to explore other network visualization software particularly those that are more user-friendly and easier to navigate.

CONCLUSION

1. Exploring Data Website
2. Gephi Wiki
3. Gephi

RELATED POSTS

1. My Experience with Gephi – Visualizing Game of Thrones: A Song of Ice and Fire
2. Les Misérables: Network of Characters
3. unveiling GLOBAL trade patterns using network analysis