W-209 - Assignment 5 – Gephi

In this assignment you will use Gephi (https://gephi.org/) to generate a network visualization for data from one of the datasets in the link below. Import the data and use Gephi to examine relationships among the nodes, edges, and communities in the data. Utilize several encoding techniques (e.g., color, size, directionality, etc.) to highlight patterns in the network, looking for several different patterns. During this exploration process:

- Experiment with different layout algorithms and examine how these affect the network visualization
- Explore centrality measures, community structures, network density, and paths
- Use filtering and labeling techniques as appropriate.

Select **three** distinct views that you've created that showcase compelling patterns in the data. At least one of these views should utilize an additional plugin (Tools > Plugins) in Gephi. Turn in a PDF document to the ISVC submission page that shows a screen capture of each of these three views (with a labeled heading for each).

Please select a dataset from: https://github.com/gephi/gephi/wiki/Datasets
Do **not** use the "Jazz musicians network" dataset as this is difficult to work with for the assignment.

Overview

For my visualization I chose the diseasome dataset from http://gephi.org/datasets/diseasome.gexf.zip. The diseasome is a network of disorders and disease genes linked by known disorder-gene associations, indicating the common disease origin of many diseases.

The code and data sources for assignment 4 is available in GitHub at https://github.com/dorairajsanjay/w209assignments/tree/master/assignment5

Gephi Pre-requisites

The below additional plugins are loaded and used in this assignment

1. Yifan Hu
2. Force Atlas 2

Data Sources

Shown below are examples of the original data sources for the node and edges table

Example: Original Nodes table

Id	Label	type	disclass
55	Deafness	disease	Ear,Nose,Throat
888	Enlarged vestibular aqueduct	disease	Ear,Nose,Throat
889	Pendred syndrome	disease	Ear,Nose,Throat
1329	MTP	gene	gene
1338	CNGA3	gene	gene

Example: Original Edges table

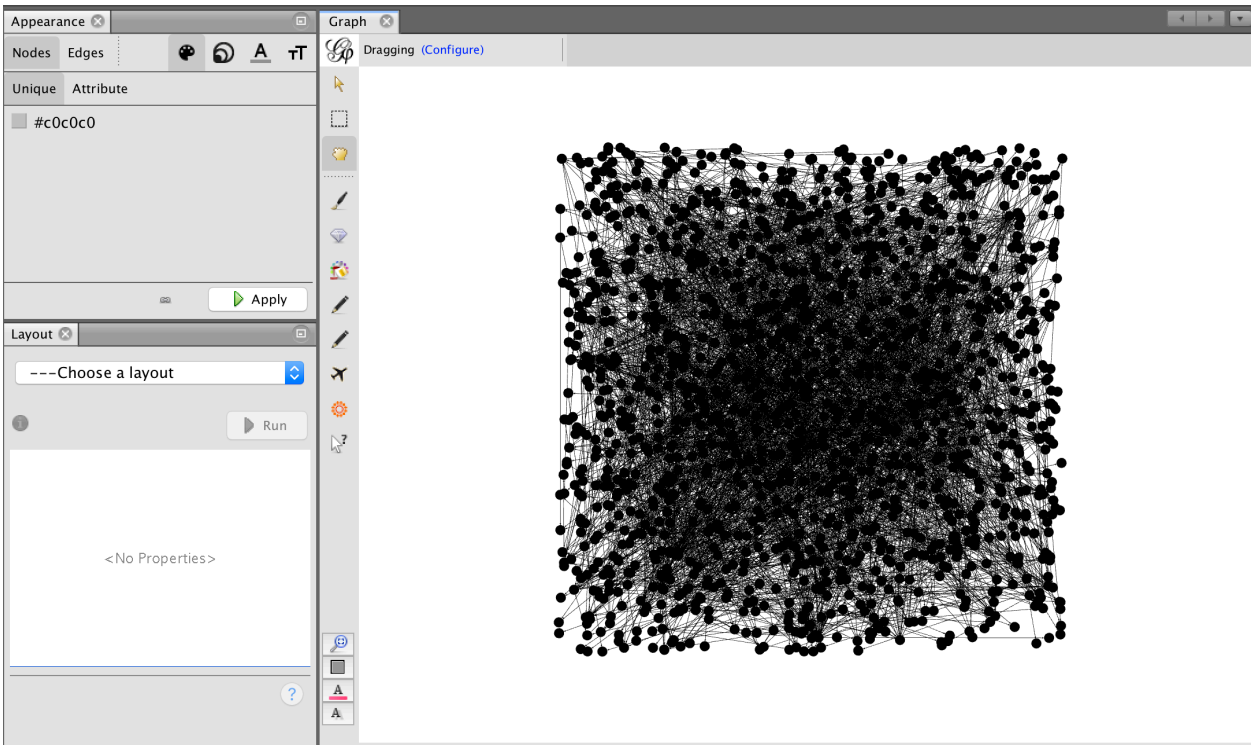
Source	Target	Type	Id
30	1420	Directed	36
30	1421	Directed	37
30	3553	Directed	38
30	3539	Directed	39

Visualization clusters of related diseases and genes with focus on visualization the Disorder Network

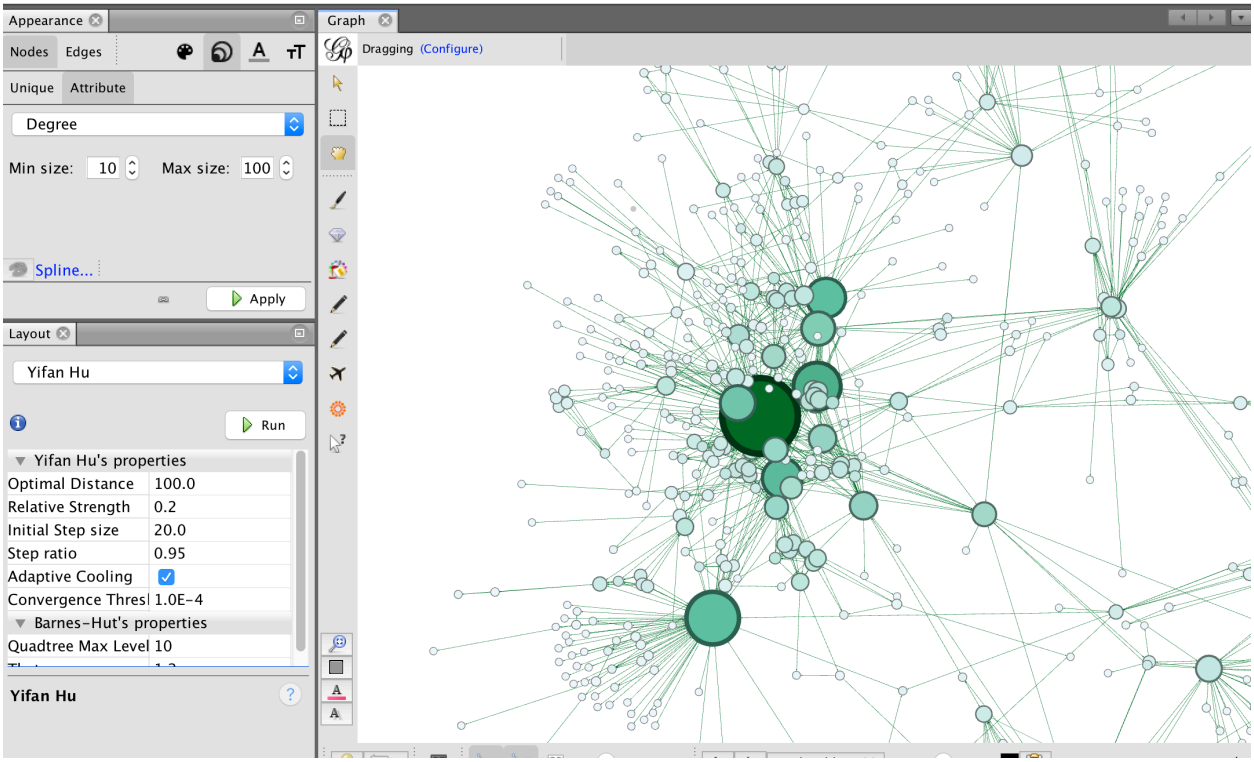
For the first visualization, I use the original dataset and extract clusters of related disorders and genes.

Steps

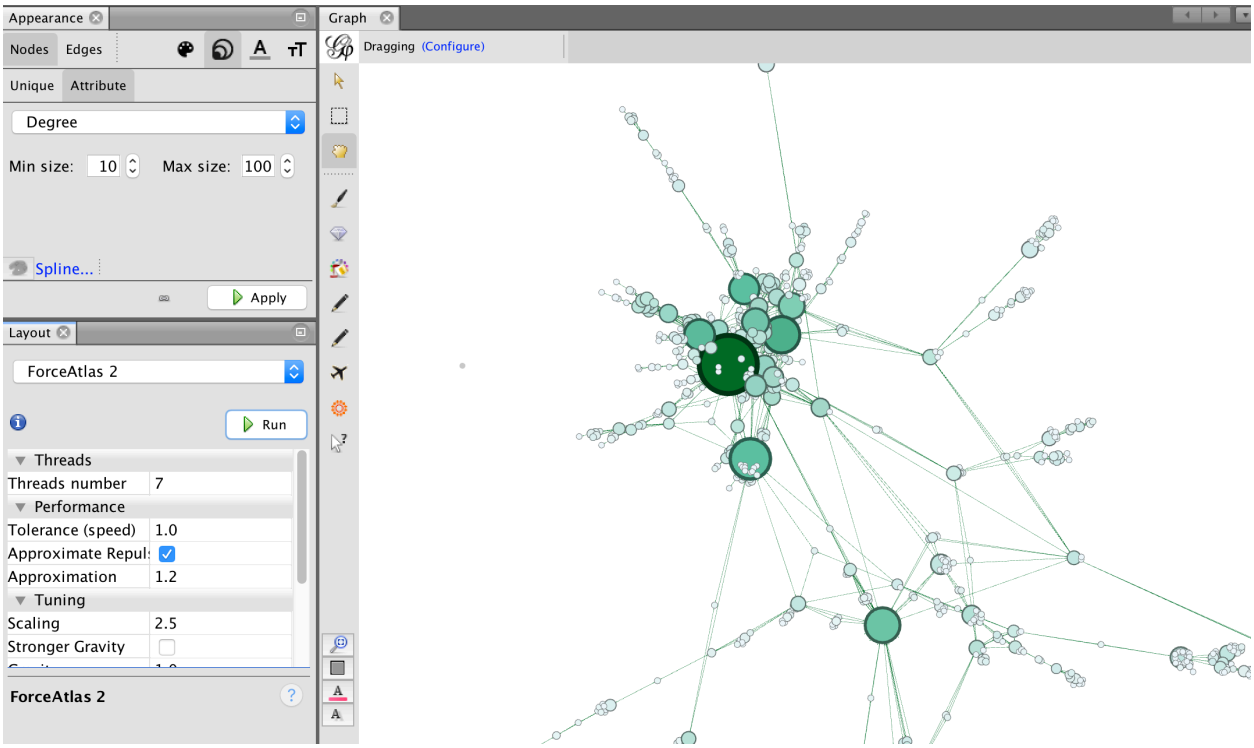
1. Create a new workspace
2. Import the nodes and edges datasets CSV file
3. View the Graph in the Overview tab



4. Set node size based on in-degree count
5. Set node color based on disease class.
6. Execute the Yifan Hu algorithm in order to generate clusters based on the number of connected edges between nodes and edge.



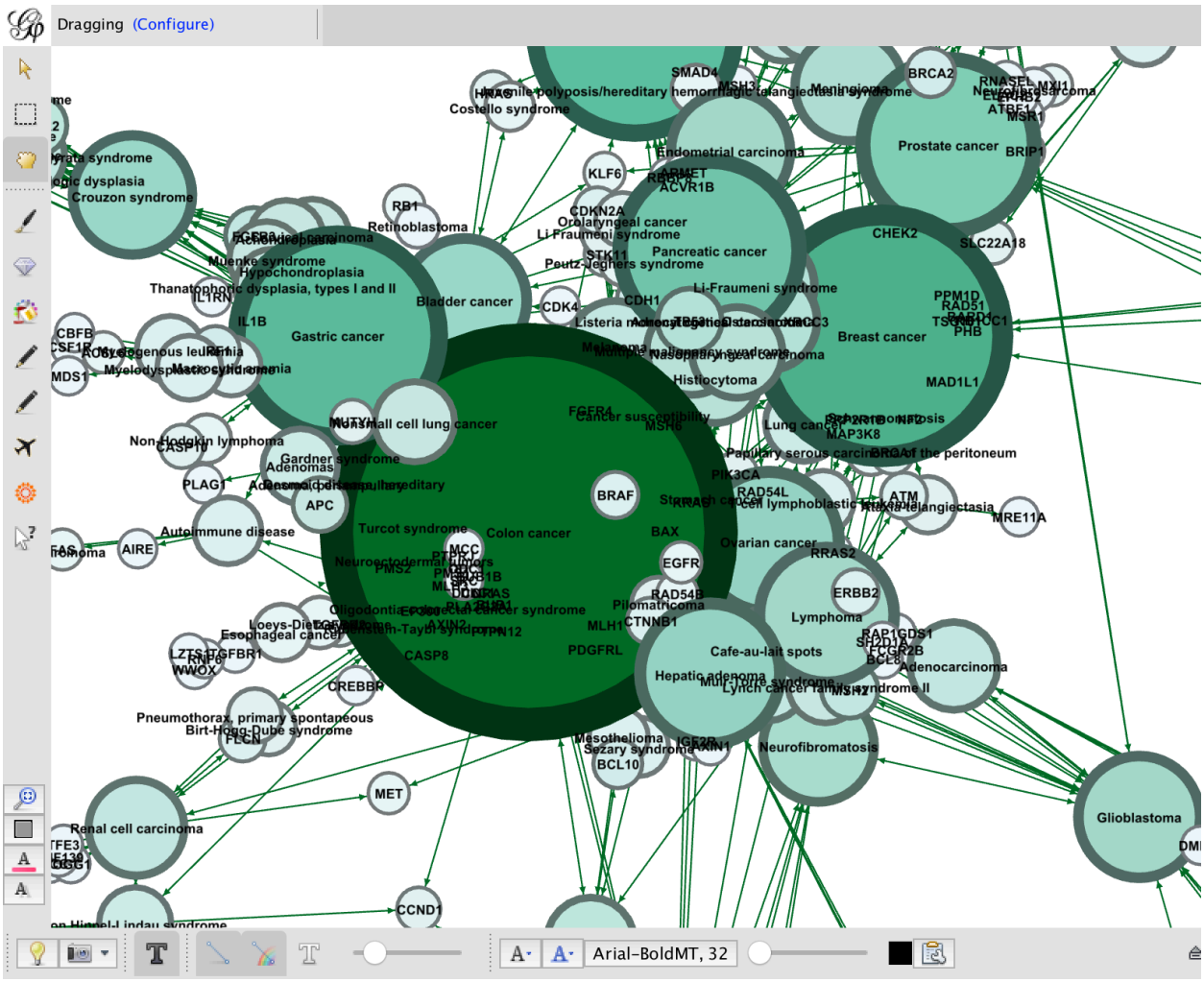
7. Execute the Force Atlas 2 algorithm using a scaling size of 2.5 to maximize separate between clusters



Insight

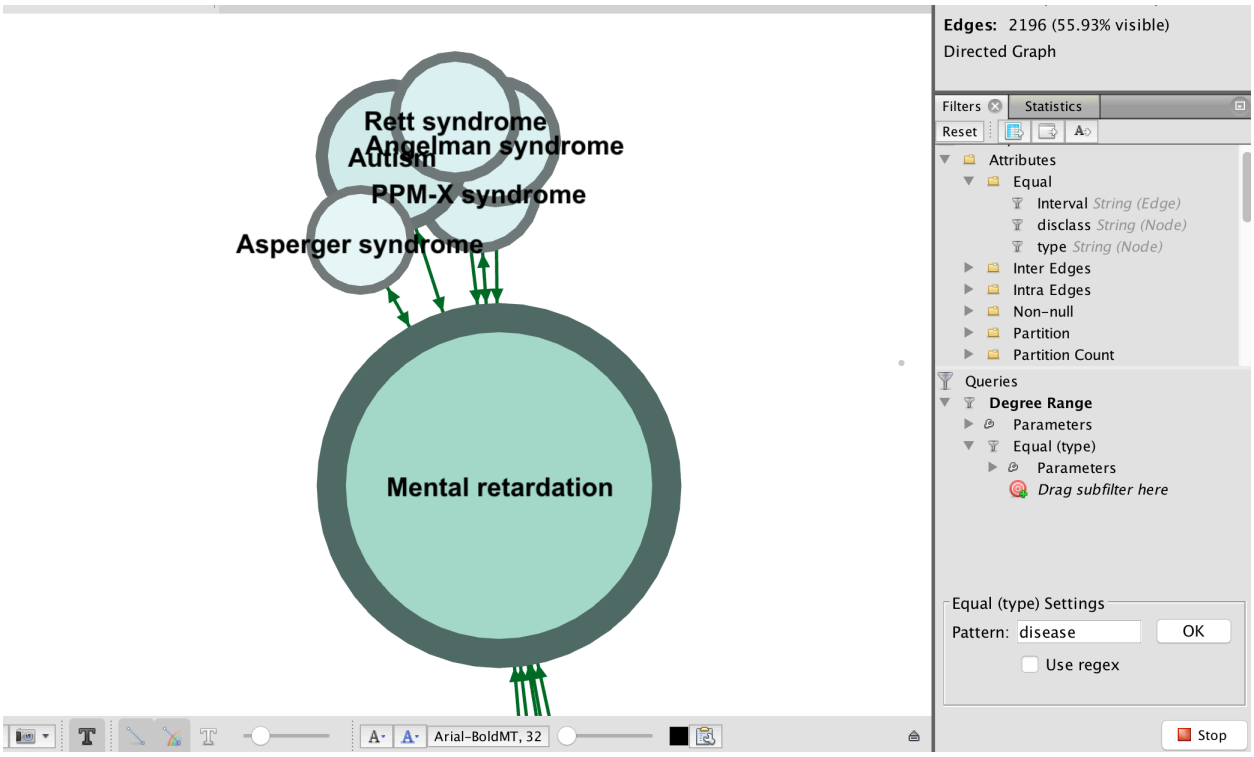
This visualization reveals very interesting groupings between disorders and their related genetic mutations by identifying well defined clusters of logically related disorders, such as cancers, hearing disorders, neuro disorders and so on.

Cancer Group



Mental disorders group

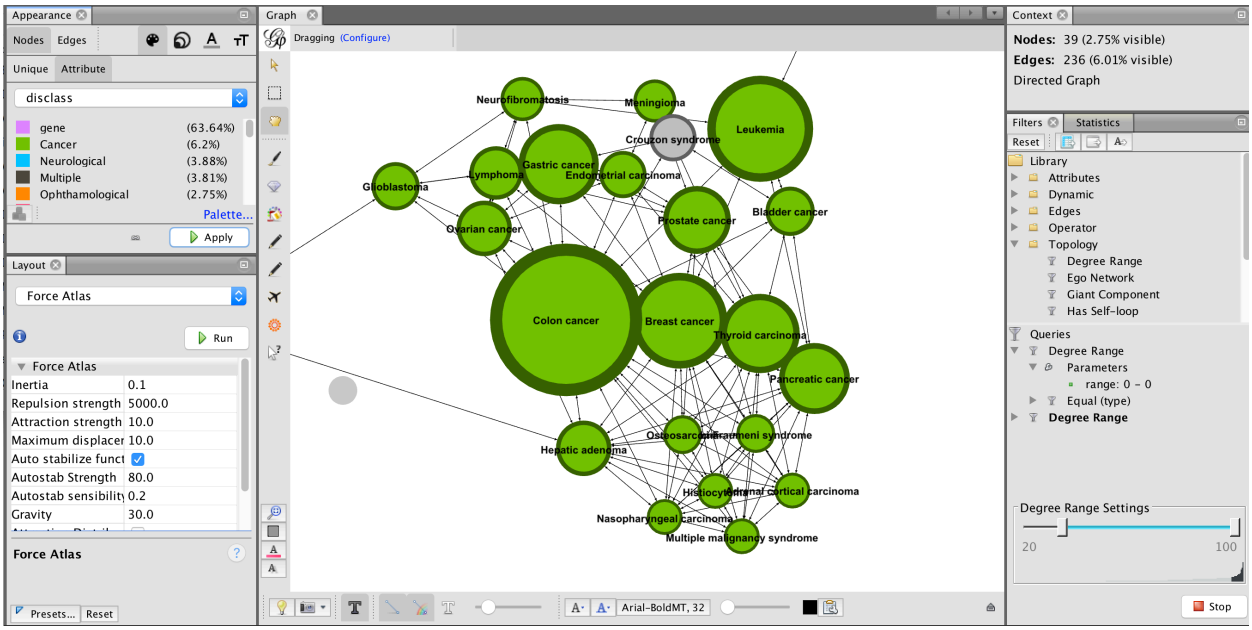
For this group, we look at cluster of mental disorders. In order to get a better visualization we set a multi-level filter to filter based on (1) degrees and (2) disease type. This limits nodes connected with 3 or more edges and removes nodes that represents genes, showing only disorders



Visualization of relationship based on different types of centrality

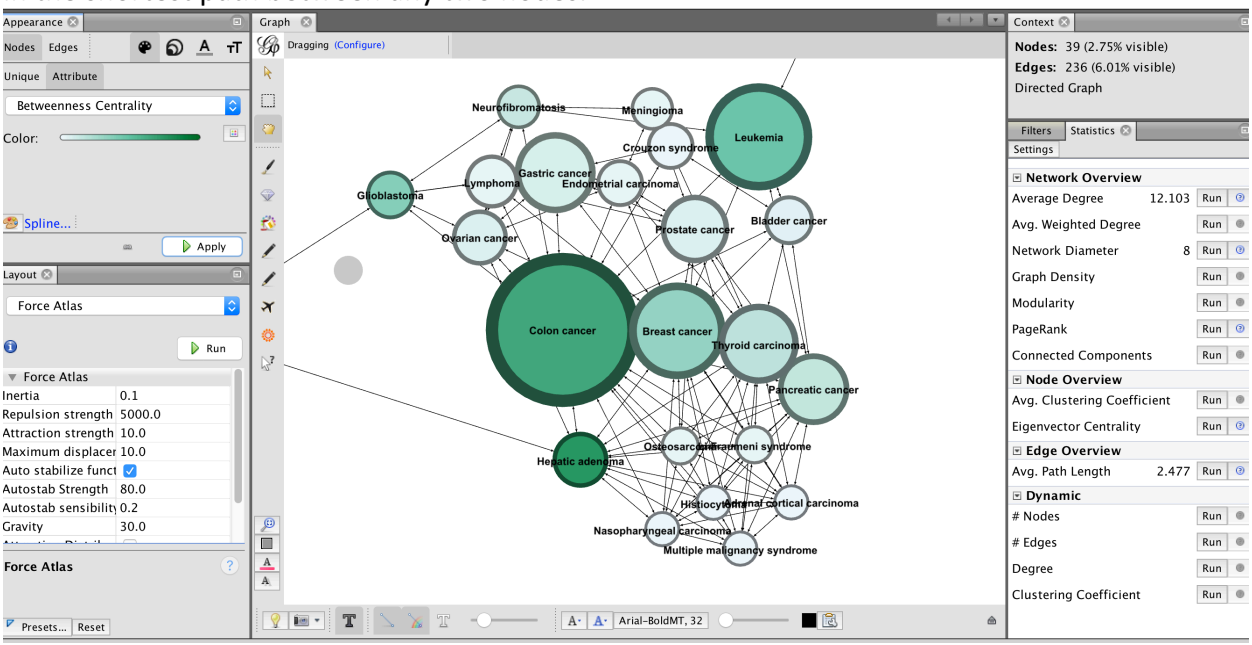
Here, we look at the relationship between disorders based on different types of centrality to see if there are any obvious differences based on centrality. We focus on disorder named Colon Cancer for this experiment. We use Force Atlas with a repulsion strength of 5000 to visualize cluster separation.

Degree based centrality



Betweenness Centrality

Betweenness Centrality is centrality based on the number of times a particular node shows up in the shortest path between any two nodes.

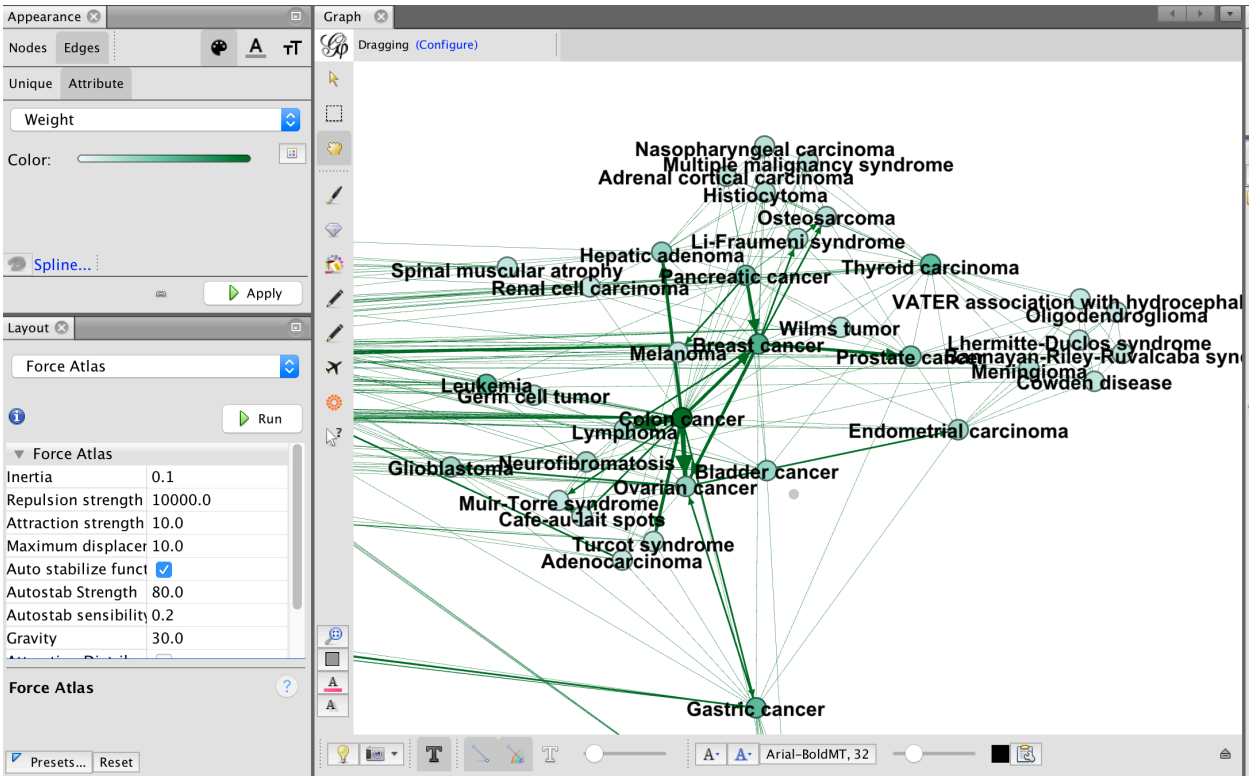


Note that the type of centrality does not significantly impact the visualization of nodes on the graph. Other algorithms like Page Rank do not yield any significant difference.

Visualization 3 – Gene Network

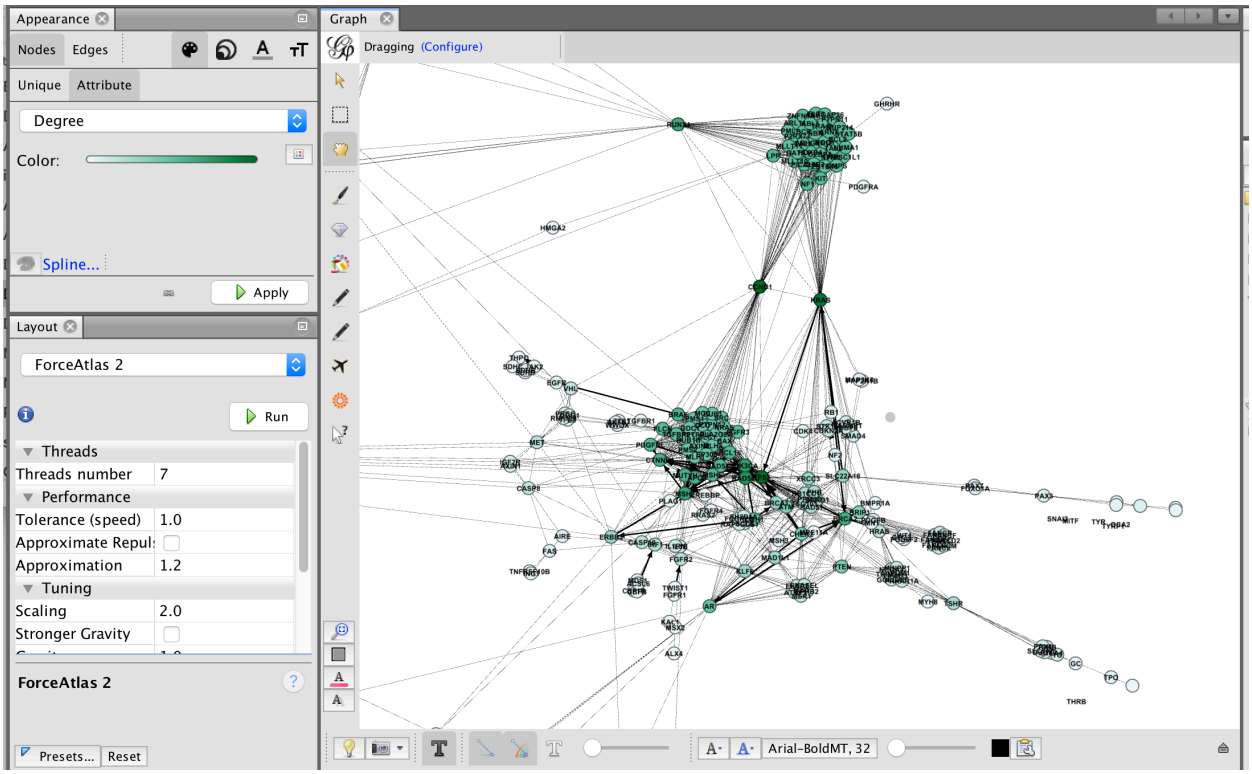
In this last experiment, we show the relationship between genes based on disorders. For this, we create a new set of derived edges files that show the relationship between genes based on the number of shared disorders and disorders based on the number of shared genes. Two new data files – nodes_disorders.csv and nodes_genes.csv were generated for this purpose. (See Python code in GitHub here https://github.com/dorairajsanjay/w209assignments/blob/master/assignment5/diseasome.ipynb)

Visualizing the Disorder Network



Visualizing the Gene Network

This visualization shows several interesting clusters of genes based on the number of shared disorders.


Appearance
Nodes
Edges
Unique
Attribute
Degree
Color:
Spline...
Apply
Layout
ForceAtlas 2
Run
Threads
Threads number
7
Performance
Tolerance (speed)
1.0
Approximate Repul
Approximation
1.2
Tuning
Scaling
2.0
Stronger Gravity
ForceAtlas 2
Presets...
Reset
Graph
Dragging
(Configure)
Arial-BoldMT, 32