# Visualisation and Visual Analytics

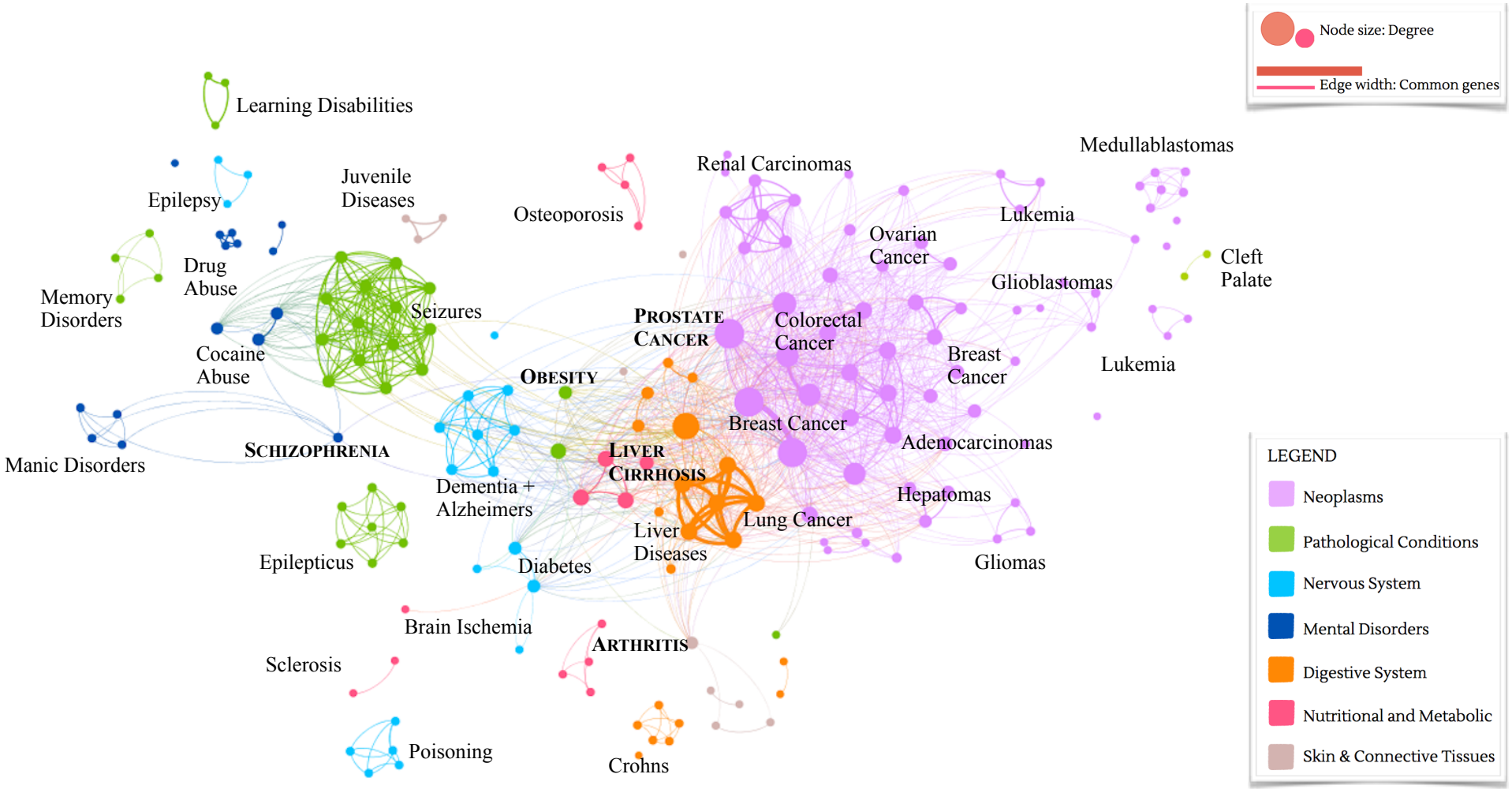
## Network Analysis

*A study of gene-linked disease-disease association*

Mishty Negi

---

## A. Human Disease Network for Gene-linked Disease-Disease Associations on Comparative Toxicogenomic Dataset



## B. Visualisation Method

### B.1 Network Dataset Preparation

A database of gene-disease association was downloaded from DisGenet. Top 300 diseases with the most number of gene influences from CTD (Comparative Toxicogenomics) subset were chosen. CSV files for edges and nodes were prepared using SQL queries (files and SQL queries included in the submission folder).

### B.2 Tool and Algorithm

Gephi with Force Atlas 2 as the layout algorithm was used to build the network. As the top 300 nodes (weighted by number of genes) were chosen with edge weight proportional to the number of connections, the attraction between nodes was quite substantial resulting in close proximities. In order to disperse the nodes and make clustering more prominent, **LinLog Mode** was used with adjustment of scaling and gravity parameters. Additionally, **Dissuade mode** was activated to push hubs to the periphery while keeping authorities in the centre.

### B.3 Filters

Three filters were applied on the result in order to observe more meaningful patterns and relationships. The number of partitions was reduced to include disease type (disease class) with representation >= 4% only. The nodes were filtered to exclude nodes with weight >=40 to dissuade isolated hubs from being formed by diseases with less diversity of connections. The edge weight was filtered to include edges with weight >=15 to remove less influential connections (noise), reducing number of edges from ~ 34,000 to ~1,200.

## C. Image Caption

**Humane Disease Network for Gene-linked Disease-Disease Associations on Comparative Toxicogenomic Dataset**. Each node represents a particular disease with size proportional to its degree. The degree of a node means the number of connections to other nodes. The width of an edge is proportional to its weight, which is the number of connections based on common genes between two nodes. The colour of the node is the class of disease it is categorised under (Please refer to the legend). The labels correspond to disease sub-classes of the clusters with the capitalised labels representing an individual disease.

## D. Discussion

The most important objects are the nodes with higher degree because they represent higher connections (gene links). The importance can also be determined by analysing the diversity of connections by observing the edges.

**Intra Category Trends:** Diseases in the same category tend to be more connected than others, but this can vary based on the category e.g. Neoplasms are quite well connected in general while Nervous System Diseases do form clusters but are spread all over the network.

**Cluster Trends:** The connections between clusters reveal an interesting set of relationships. Some clusters here seem to show a strong association based on shared genetic components (see the complex of Seizures, Juvenile Disorders, Learning Disabilities and Epilepsy or Dementia and Brain Ischemia ), while others seem to potentially be connected by risk factors as well (Manic Disorders, Substance Abuse and Memory Disorders, Diabetes and Liver Disease). Still others tend to be isolated with relatively few connections with other diseases (Osteoporosis, Cleft Palate, Crohns) implying perhaps a very specific set of genetic expressions responsible for the same.

**Disease Hubs:** It can also be seen that certain individual diseases seem to be hubs that are connected to a number of different clusters, e.g. Schizophrenia, Obesity, Liver Cirrhosis) again perhaps indicating that they share more than just common genetic expressions (risk factors, causal relationships).