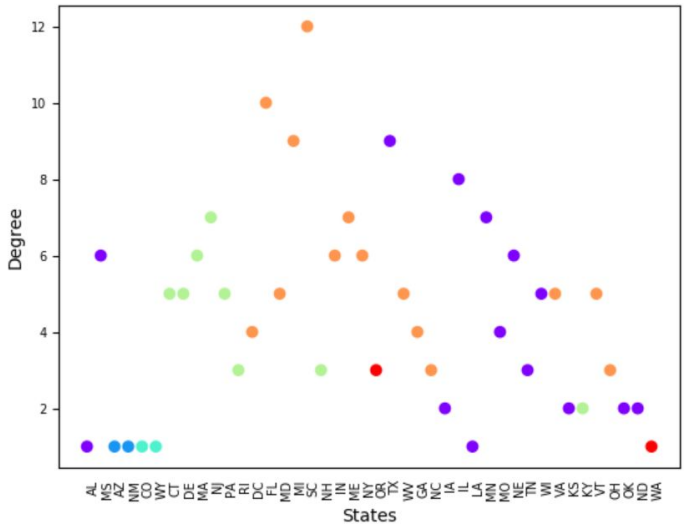
Network Analysis

To create a network from our merged dataset, we generated a similarity network between states (50 states and DC). For this similarity network, we connected each state to every other state in an undirected fashion and weighted the edges by the following characteristics (0 - 1 scale):

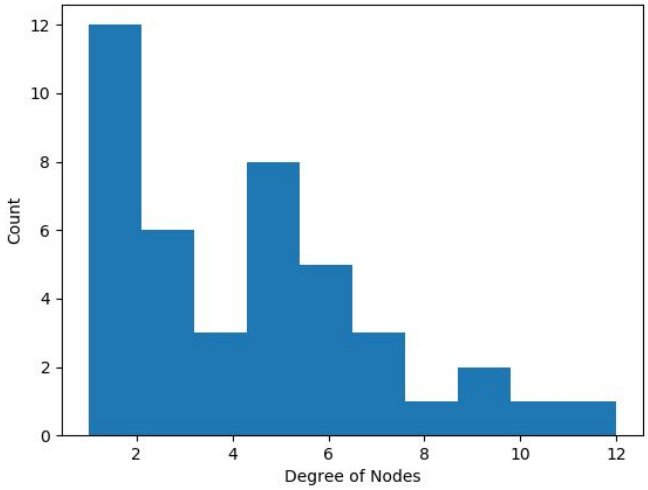
10% for the same region
30% for the same z-score cancer level
30% for the same z-score total chemicals level
15% for the correlation between the states' cancer rates over time
15% for the correlation between the states' total chemicals level over time.

However, due to the high density of this artificial graph, which includes low similarity connections, we only included the edges with a weight greater than 0.8 in the final network. Thus, the final network has some states and edges excluded to show the strongest relationships that existed between the states in our merged dataset.

Node Degree:

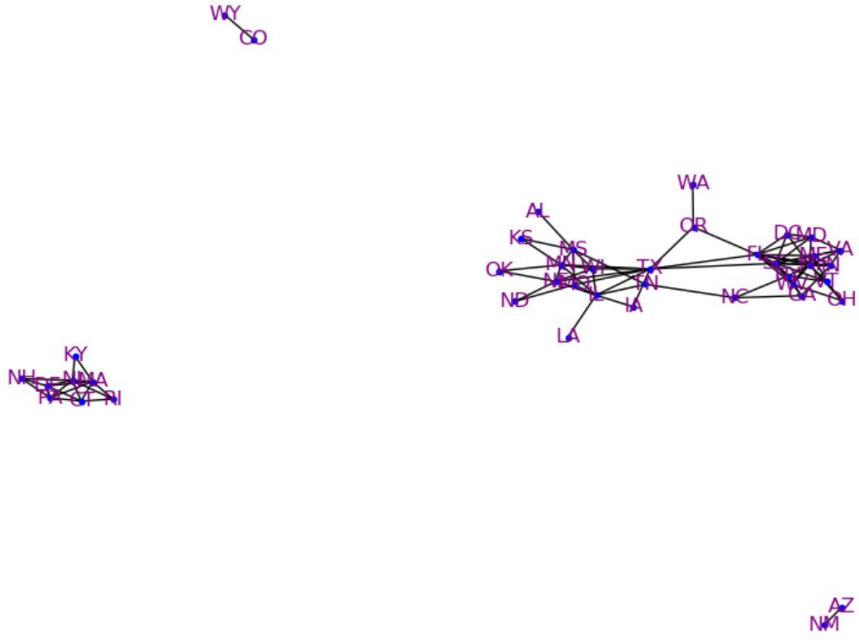


This graph displays the degree of each state in the network. Closer analysis reveals a pyramidal shape to the scatter plot, indicating that there are groups of states that share a similar degree. This might mean that these states have comparable levels of similarity to other states. Overlaying the partition labels reveals that states with similar degrees are not particularly included in the same best partition of the network, indicating that the degree alone was not a very large influence on the best partitioning of the network.



This histogram shows the distribution of node degrees in the network, displaying how most states had a fairly low degree given the edge weight cutoff of 0.8.

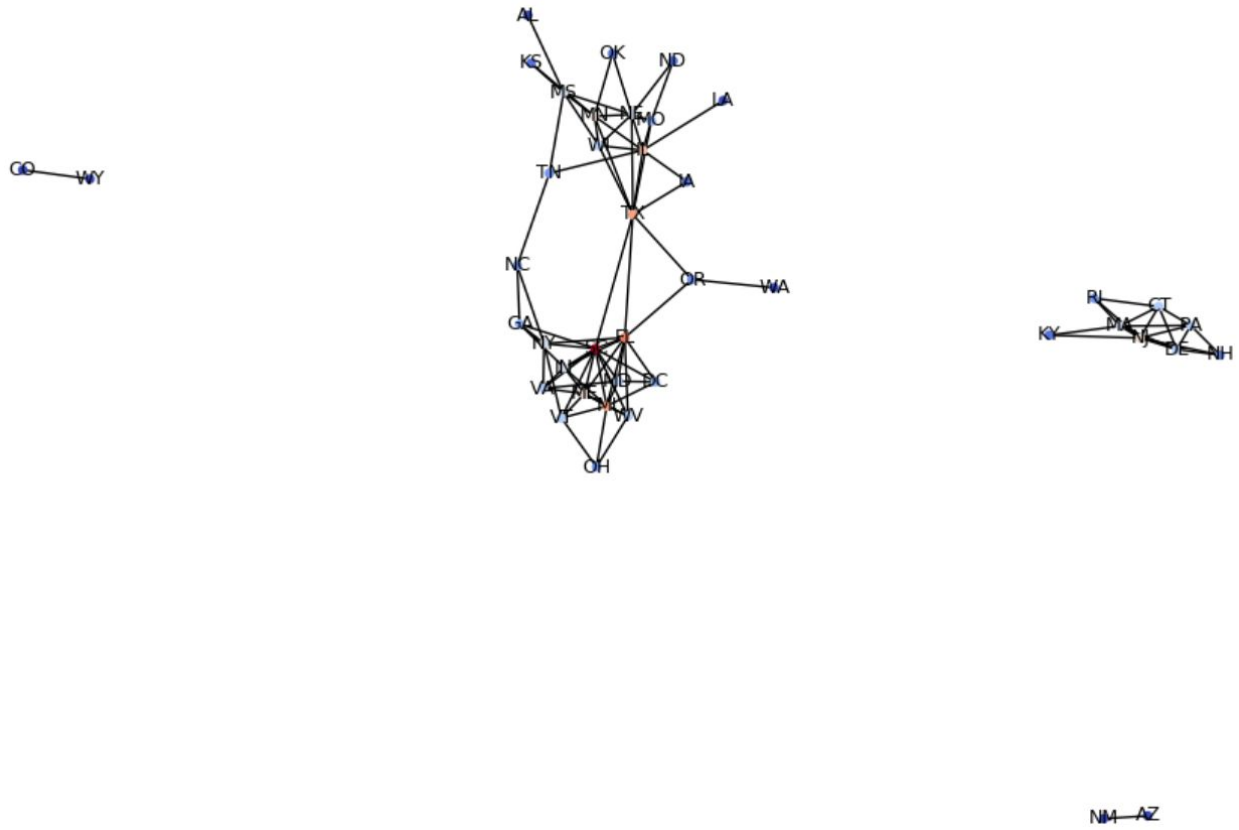
Whole Network:



The network graph displays the 4 connected components of the graph and the states in those components. Since low edge weight connections are taken out, this graph essentially shows the self-partitioning of the network based upon the sharing of characteristics for cancer and chemical release levels over time. This also displays how the largest connected component contains 30 states that are distributed all over the country, rather than in a specific region. This highlights how regionality is not an important factor in determining the cancer rate or total

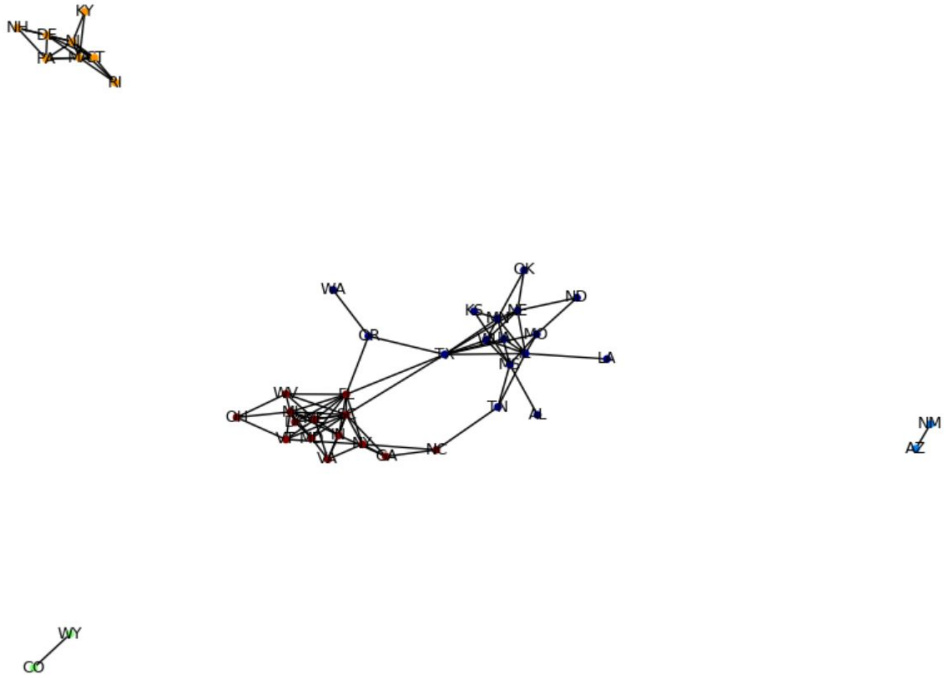
chemical release levels per capita. More likely, the relationships may be due to the types of industries in these states, but further examination of this is necessary.

Degree centrality:



This network graph displays the degree centralities of the nodes, which points out states with high connectivity with a more red color. This also highlights states with more similarities to other states, with South Caroline (SC) having the highest degree centrality.

Partitioning:



Here, the partitioning of the network into separate communities is displayed with the colors representing different “clusters.” As is clearly visible, the 4 connected components have different colors, but more interestingly, the largest connected component in the middle is split into two communities by the partitioning algorithm. This indicates that there is less connectivity between groups than at first-glance. Perhaps the 5 “clusters” represent states with similar chemical industries or some unknown factor that is not easily determined. In terms of modularity, this measure was found to be 0.609, which is fairly high. This indicates that there are more edges within the clusters than expected due to random chance.

Network Characteristics:

Density: 0.108

Number of nodes: 42

Number of edges: 93

Number of connected components: 4

Size of largest connected component: 30 nodes

Number of triangles: 74

Degree centrality average: 0.108

Betweenness centrality average: 0.0208

Total number of clusters: 6

Average clustering coefficient: 0.477

Modularity: 0.609

Overall Network Characteristics:

Overall, the network is fairly dense for a real-world based network as seen in the statistics above. This indicates that there are more connections than expected for such a network. Further, looking at the average degree centrality, the mean indicates that the distribution of degree centrality is centered around 0.108 which represents 10.8% of the highest degree values for a node. This is approximately 1.3. Since this is not really an information related network, the betweenness centrality does not apply much to our dataset, since a node with high betweenness does not really represent a significant characteristic of the similarity network.