Naeco Le
DS210
4 May 2023

For this project, I used the hollywood-2011.graph dataset from https://law.di.unimi.it/webdata/hollywood-2011/. This dataset contains information about actors and movies from the Hollywood industry in 2011. The graph consists of 2,180,759 nodes, and it is undirected. I chose this dataset because it is large and complex, and it provides a real-world scenario that can be used to test various graph analysis algorithms. I also chose this dataset because of my fondness for film.
From my analysis of this dataset, I discovered that most actors are very connected in terms of the films they do. This is evident in nearly every part of my analysis, from Degree Distribution to—most importantly—Distance Between Vertices.

My code is set up without any modules, something I regret doing. I also lack test cases for some of my analysis. This is a product of time constraints and my own inexperience coding. I am still struggling with writing modules and test cases, though I was able to implement a—albeit simple—testing strategy for some of my analysis. Each of my pieces of analysis is in the src folder. The names of each are in the cargo toml folder. The first step is to run **iconv -f ISO-8859-1 -t UTF-8 hollywood-2011.graph > hollywood-2011-utf8.graph** in terminal to get the hollywood-2011-utf8.graph file from the hollywood-2011.graph file downloaded from the link above. From there, you need to run the prep.rs file to get the nodes_and_edges.txt file. This is the file I used to initially visualize the data.

In my analysis of the data, I first looked at degree distribution. Here is the degree distribution:
-9 1
0 36
1 32
2 43
3 23
4 34
5 39
6 36
7 24
8 44
9 25
14 1
16 1
25 1
41 1

56 1
74 1
86 1
89 2
44 nodes with 8 connections
36 nodes with 0 connections (isolated nodes)
There are few nodes with greater than 9 connections.
Gamma: -20621.013755360796
Power-law distribution confirmed.
Since the degree distribution of the movie actor graph follows a power-law function, this indicates that there are a relatively large number of actors with many connections, and that these high-degree actors play an important role in the connectivity and structure of the graph.

Next is the graph density. Using density.rs I calculated the density.
Graph density: 1.0116959064327486
Densest subgraph: {16, -9}
Densest subgraph density: 2

The graph density for the movie actor graph is 1.0116959064327486, which is greater than 1, indicating that there are more edges in the graph than the maximum possible number of edges. This is likely due to the fact that the graph is not a simple graph, since it contains self-loops (i.e., edges connecting a node to itself) and multiple edges between some node pairs.

In the case of the movie actor graph, the densest subgraph is {16, -9}, which consists of two nodes. The densest subgraph density is 2, which means that there are two edges between the two nodes in the densest subgraph.

It is worth noting that the node with ID -9 appears in the degree distribution with a single connection, which suggests that it is a relatively isolated node in the graph. The fact that this node is included in the densest subgraph may indicate that it has a strong connection to the other node in the subgraph (node 16), or that the densest subgraph is not representative of the overall structure of the graph. It is also possible that the presence of self-loops and multiple edges in the graph is affecting the calculation of the densest subgraph and its density.

Using centrality.rs I calculated degree centrality
Degree centrality:
Node 89: 2
Node 86: 1
Node 74: 1
Node 56: 1

Node 41: 1
Node 25: 1
Node 16: 1
Node 14: 1
Node 9: 25
Node 8: 44
Node 7: 24
Node 6: 36
Node 5: 39
Node 4: 34
Node 3: 23
Node 2: 43
Node 1: 32
Node 0: 36
Node -9: 1

In the context of the movie actor graph, the degree centrality of each node is a measure of how many movies the corresponding actor has appeared in with other actors in the network. Nodes with higher degree centrality are actors who have worked with a larger number of other actors in the network, and are therefore potentially more influential or well-connected within the industry.

In the provided degree centrality list, Node 8 has the highest degree centrality, with a value of 44. This means that the actor represented by Node 8 has appeared in movies with 44 other actors in the network. Node 2 has the second highest degree centrality, with a value of 43, indicating that the actor represented by this node has appeared in movies with 43 other actors in the network.

Finally, I calculated the average distance between pairs of vertices.

Average distance between pairs of vertices: 1.7445255474452555

The average distance between pairs of vertices in the graph of movie actors means that on average, any two actors in the graph can be connected by a path of 1.74 edges, or one or two steps in the graph. In other words, the average shortest path between any two actors in the graph is 1.74. In the context of the movie actor graph, this suggests that actors are often connected through shared movie appearances, indicating that the movie industry is a tightly knit network where actors tend to work together frequently.