The dataset I decided to use for this project is a sequence of snapshots of the Gnutella peer-to-peer file sharing network from August 2002. Each node represents a host or user in the network, while each edge represents a connection or link between hosts. With this dataset, we are able to analyze the connectivity, efficiency and structure of the Gnutella peer-to-peer network. Through analysis of centrality, shortest paths, and other measures, we can gain insights on how well-connected nodes are as well as other implications of the network.

I decided to use centrality measures on this data set as they provide insight on the influence or importance of notes in a network. I separate centrality into in-degree centrality and out degree centrality. In-degree centrality is the number of incoming edges to that node, while out-degree centrality is the number of outgoing edges to that node. This is a useful algorithm, as it gives us insight, as it tells us how influential a certain node is, and how important of an information-receiving hub a node is. Essentially, centrality shows us how heavily connected nodes are and how important each node is in the distribution of files within the network.

Along with centrality (in-degree and out-degree), I also performed further statistical analysis on the dataset through finding shortest path between nodes, clustering coefficient, mean degree, degree variance, and global clustering coefficient.

```
In-Degree Centrality (Top 5):
6450: 0.0010131712259371835
3481: 0.0008105369807497467
10729: 0.00020263424518743666
961: 0.00425531914893617
2550: 0.0008105369807497467
Out-Degree Centrality (Top 5):
7318: 0.002026342451874367
8743: 0.002026342451874367
890: 0.002026342451874367
5699: 0.00020263424518743666
2119: 0.002026342451874367
Shortest Path between nodes 4780 and 5049: [4780, 2455, 4813, 1534, 2684, 5049]
Mean Degree: 8.10
Variance of Degree: 17.45
Global Clustering Coefficient: 0.0027
```

My results show the top 5 nodes and their respective centrality for in-degree and out-degree. We see that 6450 is a very influential node as it has the highest number of incoming connections. On the other hand, node 7318 is a very important node as it has the highest number of outgoing connections. This shows that certain nodes on the network have a higher level of importance than others based on the measures of their centrality.

Shortest path between two nodes shows the minimum number of connections that need to be moved through to get to the final node. In my code, I chose to analyze the shortest path between nodes 4780 and 5049. I chose these nodes as 4780 as it was among the top out-degree centrality nodes, and 5049 was among the top in-degree centrality nodes in my

previous tests. This makes these nodes a good selection to see what the shortest path would be from a very influential source of information to a very important hub that gathers information. According to the result, having to traverse only four nodes to get to the destination node, suggests that there is a very direct route between these two nodes, confirming the out-degree and in-degree centrality. This gives us insight into the efficiency and directness among the connections of the top centrality nodes in the network.

I also perform statistical analysis on the mean degree of the dataset. The mean degree is a measure that shows how many connections each node has. According to the result, each node in the network has (on average) roughly 8 other nodes that are connected to it. A high mean degree would show that the network is very dense, while a low mean degree would show that the network is very sparse. With 8.10 being the mean degree, we can see that the dataset is moderately dense, where the network is interconnected but not to an extreme extent.

Moreover, I calculate the degree variance. Degree variance shows the spread in the connections that nodes have in the network. According to my result, the degree variance of the dataset is 17.45, meaning that the degree of nodes vary quite a bit. This high variability may be a sign that nodes have a higher number of connections and also might mean that there are nodes that are highly connected, and there are nodes with very few connections.

The last part of statistical analysis I performed was finding the global clustering coefficient of the dataset. The global clustering coefficient shows the level of interconnectedness in a network. It shows how likely two neighbors of a node are to be connected. With a global clustering coefficient of 0.0027, the network has low clustering and connectivity. This means that the nodes sin the network are not very connected and do not fall in tight groups. This could also be an indication that nodes in the network have a random pattern rather than having local clusters.

```
running 0 tests

test result: ok. 0 passed; 0 failed; 0 ignored; 0 measured; 0 filtered out; finished in 0.00s
```