Zihao Guo

DS210

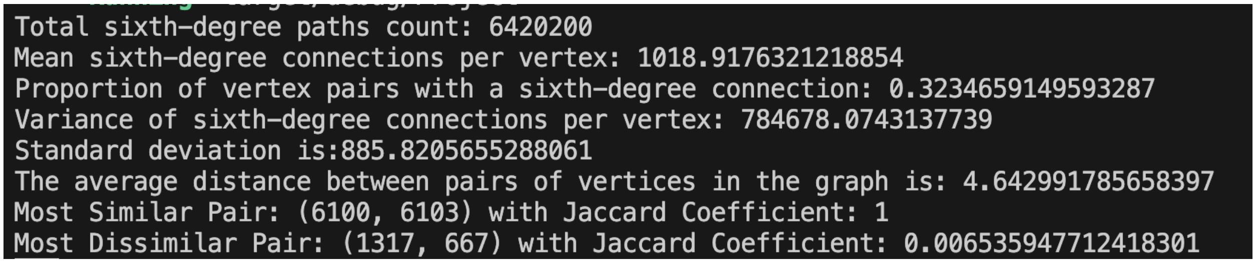
12/12/2023

Project Report

Data Link: https://snap.stanford.edu/data/p2p-Gnutella08.html

This data is connection data with 2 columns. The left column is the subject ID and the right column is the object ID. It outputs that the left ID knows the right ID in a **DIRECTED** way.

Output data:



```
Total sixth-degree paths count: 6420200
Mean sixth-degree connections per vertex: 1018.9176321218854
Proportion of vertex pairs with a sixth-degree connection: 0.3234659149593287
Variance of sixth-degree connections per vertex: 784678.0743137739
Standard deviation is:885.8205655288061
The average distance between pairs of vertices in the graph is: 4.642991785658397
Most Similar Pair: (6100, 6103) with Jaccard Coefficient: 1
Most Dissimilar Pair: (1317, 667) with Jaccard Coefficient: 0.006535947712418301
```

Data explanation:

Total sixth degree: In all the combinations that are listed by node, there are only 6420200 paths which is exactly 6 steps from one ID to another.

Mean sixth degree: The total number of six steps reached divided by the number of nodes, which means that each node will have about 1019 paths on average.

Proportion: Total number of six steps divided by the total number of paths. 32.35 percent of the total path is a six-step path.

Variance: The difference between each node’s sixth-degree path compared to the mean of the sixth-degree path.

Standard deviation: The measure of the average distance from the mean, in the same units as the data itself.

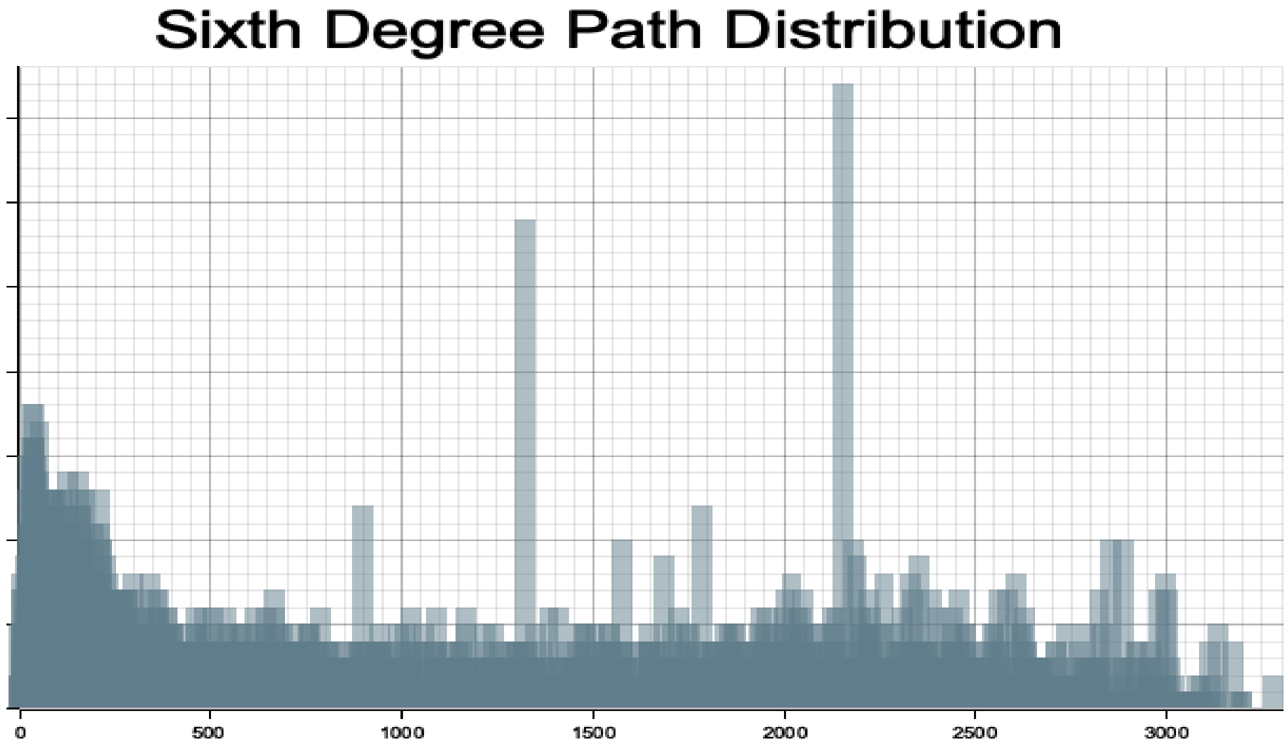
Avg steps: The average steps that one node needs to reach another node.

Similar and dissimilar pair: The similar and dissimilar pair will use the Jaccard coefficient from the domain of 0 to 1 to find the most related and unrelated pairs. 0 would be considered the most not related and 1 would be the most relative.

Data analysis:

From the mean and proportion of vertex pairs with a sixth-degree connection, we can conclude the connection map of the data is a dense network, which means that if a person wants to reach out to another person, it may cost a lot of steps to connect. The high variance and standard deviation shows that there are different type of people in the data. Some people are very socialized and they can connect with people directly or within a few steps (at least less than six steps). Also, some people may not be good at social networking, and it’s hard for them to connect with people, so they may need to connect with more than six steps. When we connect both groups of people, we can see that there are more socialized people than unsocialized people, from the average distance between pairs of vertices in the graph, we can see that the average step cost is 4.64, which is less than six steps. We also made a graph that analyzes the number of sixth steps that each person needs and the frequency of people.

(The image is on the next page):



The peaks that appear on the graph show that there may be some underlying structure or community clusters where nodes within the same cluster have a similar number of sixth-degree connections. The spread of width distribution also indicates that people may be socialized or unsocialized, which means that some people need more sixth-degree paths and some don't. The skewness shows that it's a slightly right skew. We can conclude that more nodes with fewer sixth-degree connection, and we can also confirm the average number of steps that a node need to another node. There are some outliers on both the left and right sides. We can conclude that there are some people knows a lot of connection that are in the data. Some people need a lot of sixth-degree paths and only know a few people in the dataset.