Final Project Report

DS 210 Cixin Gao

In this final project, I want to prove six degrees of separation by using email connections in a company. Six degrees of separation is a magic theory that people worldwide can contact others within six people (or less than five other members) in their networking chain. I am interested in whether it could use the emails in a company that could prove this theory. Therefore, I decide to use <https://snap.stanford.edu/data/email-Eu-core.html> email-Eu-core network dataset found in Stanford Network Analysis Project (SNAP) to prove the concept of six degrees of separation. This dataset is about the email networking of a large European research institution. This dataset contains two files. The first one is [email-Eu-core.txt.gz](https://snap.stanford.edu/data/email-Eu-core.txt.gz). It let every member in the institution as the nodes and let the others that the member emailed to as the edges. The second one is [email-Eu-core-department-labels.txt.gz](https://snap.stanford.edu/data/email-Eu-core-department-labels.txt.gz). It labels the department that each member belongs to.

In my project, I first used the function of read profile and read two datasets separately. Using the Graph impl, I built two graphs based on two data txts. The first one (named graph) is to let the member be the nodes and let the member that the node used to send the email be the edges, The second graph (named graph\_l) is to let the department be the graph and let the member that belongs to the department be the edges. For the first graph, I use BFS to find out the path length that each member can contact the other members. If member A does not have any contact networking based on email in data with member B, the bfs showed the length would be None. By building up a 2d vector, I counted the path number classified by the different path lengths and the number that two members did not have any networking with each other based on the email sending.

The output is as follows.



Based on the vector, I build up a histogram that could show the normal distribution of the path length. The histogram does not show the None path number.

Based on the distribution above we could find out that the vast majority of the existing paths ‘length are followed the six degrees of separation. However, there’s still a mass of networking paths that do not exist between the two institution members as the vector shows in the [7,216591], which represents how many members did not have any networking email or path with another member.

Therefore, based on the graph\_l, I build up a vector to find out which department that have the least networking with other members. Based on the 216591 Nones I classified them based on the second graph (graph\_l) into the vectors and build up a histogram as follows to show the data. From what we could see, department 5 has the least networking with other members.

The output is as follows. I sorted this 2d vector which can help us find the least social networking department quickly.



In the end, for the test part, I test whether the function read\_files, impl graph, and bfs are working. And in the main function, I checked or tested the sum of the data to ensure that the coding did not miss any data during the algorithms.

In conclusion, the data does not successfully prove the six degrees of separation, since there are a lot of paths that do not exist, which means by the emails there are a lot of numbers that cannot find others. However, by the existing path, it could prove six degrees of separation, because the most of path’s length is within 6. By counting the None number, it could find that department 5 is the least social department that has the least social networking with other members.