Lize Chen

Professor Leonidas Kontothanassis

DS 210

7th May 2023

Six Degree of Separation on Facebook Dataset

In the final project, I use BFS (Breadth-First Search) to explore the degrees of separation of Facebook users' social circle according to the dataset from Stanford Network Analysis Project (SNAP) https://snap.stanford.edu/data/ego-Facebook.html. The six degree of separation is an interesting topic for me since primary school while one classmate said that he could know the president of United State through connected to 6 friends' friends. Before, I did not have any way to prove whether this is a true statement or not. After learning relative approaches such as BFS and DFS(Depth-First Search), it becomes accessible for me to analyze whether it is possible for one person to get to know almost every other person in the world within connections to 6 people, and I would like to research about it this project.

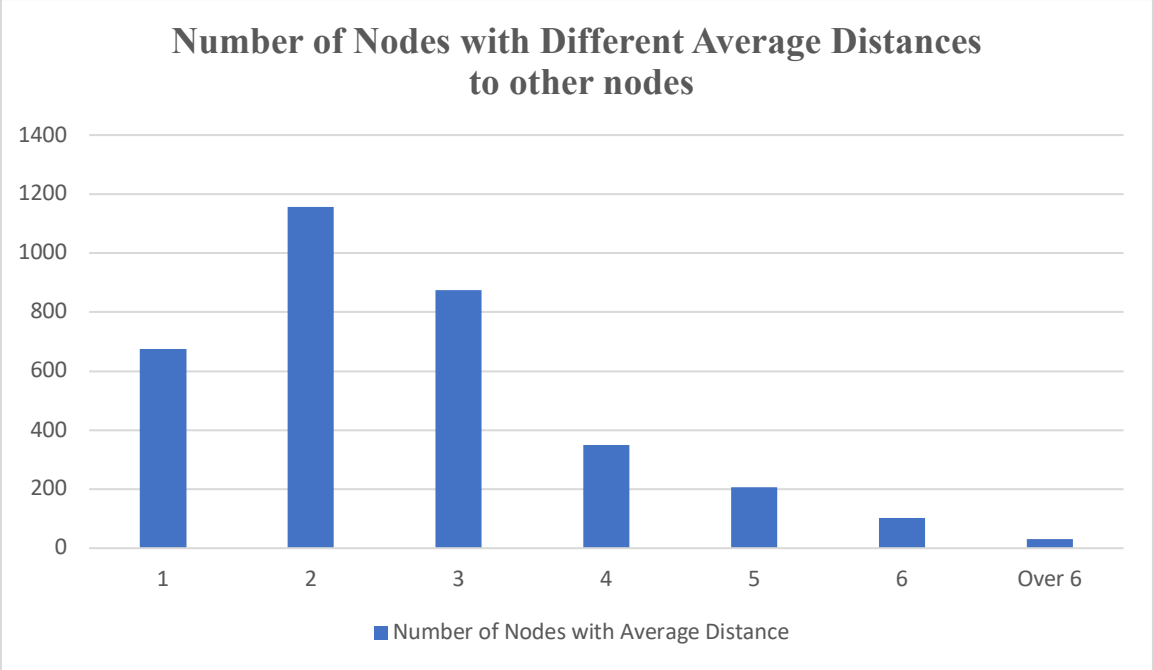
The dataset called "facebook_combined.txt" of Facebook I used contains 4039 nodes connected to 88234 edges in total, representing 4039 anonymous accounts on Facebook and their friends they follow and also follow them. The main methodology of my project is to turn the text file into a graph formatting first, and then perform BFS to calculate distances from each node to reach out to other nodes. Since it would be difficult for me to draw a conclusion with every distance from one person to another, I decided to use the average distance from every node to all other nodes they connected with to show whether six degree of separation exists in general. Then, I could determine the existence of the concept depending on the majority of average distances from nodes to nodes by seeing whether they are below six.

To complete this project effectively, I separated modules into 4 parts: the main function including tests, the read_file module which is able to access files for multiple times for different projects, the graph module implemented with methods to construct graph structure, and a BFS module to calculate the average distance from nodes to nodes. Finally, the result of my code

shows that there are 675 nodes with an average distance of 1 to others, 1157 nodes with 2, 857 nodes with 3, 350 nodes with 4, 207 nodes with 5, and 103 nodes with average distance of 6. There are also 31 nodes with a larger average distance over 6, as shown in the screenshot out output in terminal attached below. What's more, I also get the maximum and minimum connections throughout the dataset, showing that there is a difference larger than 1000 showing that the smallest one is 0.

```
From statistics aspect: [[0, 3380], [1, 675], [2, 1157], [3, 857], [4, 350], [5,
 207], [6, 103], [7, 31]]
```

To be more direct, a histogram based on the data is also attached:



It is clear that most people could access to most others through an average distance of two or there, and in general, I proved the correctness of six degree of separation by finding out that most people could connect to others in the rest of world by no more than 6 connections among people, at least on Facebook.