Haoyang(Polly) Peng

Prof K

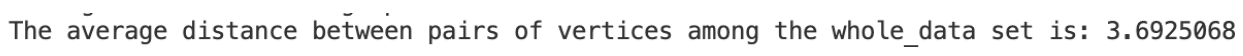
DS210 Final Project Report

7 May 2023

The data set I chose is called: “Social Circles: Facebook” which contains circles or friends lists from Facebook with 4039 nodes and 88234 edges. In its description, it includes “node features(profiles), circles, and ego networks.” Showing the connection between both users that mutually follow each other. I chose this data set because I am really interested in making friends and also learned that networking is a very important part of one’s career through job finding. And interestingly enough, I usually found that any new friends that I make now will have at least one common friend or person that we both know. So by analyzing this data set, I wonder what’s the average distance between a random person to another for connections, what would be the shortest path to knowing another person, how connected are people in this data set and what are some uniqueness of the most popular component?

Digging into the analysis through coding, I used an adjacency list after importing and reading the data, making each node connects to all its connected profiles. Since the graph is undirected, we also need to reverse the edges making sure they are mutual relationships. First of all, I used the Breadth First Search algorithm which it starts from one node and visits all the reachable nodes, returning a vector containing the distance of each vertices. Then the ‘compute_average_distance_bfs” function will collect all the distances together and divide them by the total number of pairs which then we get the average distance between pairs of vertices among the whole data set. The number, in the end, is 3.692508, indicating that within this

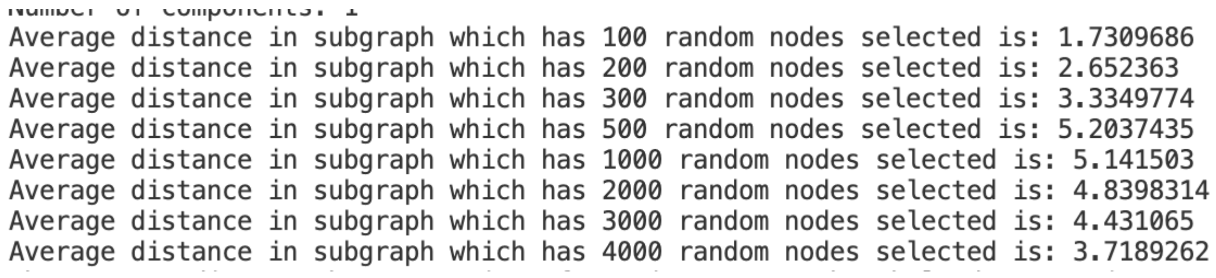
Facebook dataset which has 4000+ nodes, the average distance of getting to know another person is about 4 people away.



```
The average distance between pairs of vertices among the whole_data set is: 3.6925068
```

Figure 1: average distance: 3.6925069

I expected the number to be near 6 since the theory of ‘six degrees of separation’, therefore I do want further experiments for testing out how would the number change with different amounts of node numbers randomly chosen. So I created another function ‘subgraph’ that is based on the number of nodes that I chose, randomly picking nodes and then constructing a subgraph for calculating the average distance between pairs of vertices it has. I then created a number of node lists that contain these amounts: 100,200,300,500,1000,2000,3000,4000. Then for each amount, we run 15 times, each time creating a random subgraph for calculating the average distance, therefore in the end, we have 8 outputs with the average - average distance for each amount.

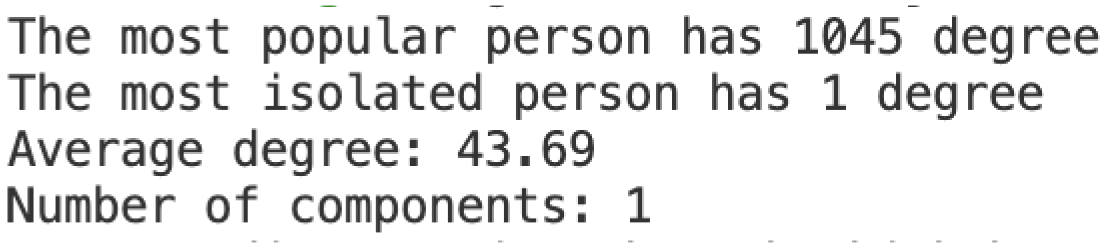


```
Number of components: 1
Average distance in subgraph which has 100 random nodes selected is: 1.7309686
Average distance in subgraph which has 200 random nodes selected is: 2.652363
Average distance in subgraph which has 300 random nodes selected is: 3.3349774
Average distance in subgraph which has 500 random nodes selected is: 5.2037435
Average distance in subgraph which has 1000 random nodes selected is: 5.141503
Average distance in subgraph which has 2000 random nodes selected is: 4.8398314
Average distance in subgraph which has 3000 random nodes selected is: 4.431065
Average distance in subgraph which has 4000 random nodes selected is: 3.7189262
```

Figure 2: parabolic pattern

From the 8 outputs, we can see a parabolic pattern for the average distance in which the number increases first from 1.7 with 100 random nodes to 5.2 with 500 random nodes, then it decreases at a slower path to 3.7 when the subgraph has 4000 nodes.

On the other hand, via Breadth First Search, I calculated the number of components that the whole graph has, the function ‘mark_component_bfs’ starts from one node and then marks all the vertices that are in the same component. Surprisingly, the output is 1, meaning that all nodes in the graph can reach each other in a different number of steps. It suggests a high level of social cohesion and interconnectivity among the individuals within the dataset. I have also found the most popular person(highest degree) has 1045 degrees and the most isolated person only has 1 degree. Despite the huge difference between them, there weren’t any total isolated nodes(with a degree of 0), the highly popular person within this social network can create a large amount of information flow for introducing all his/her connections to each other, forming new relationships. So if in this case, we are new to this community, the most effective way is to connect with the most popular person first which immediately made us just 1 step away from his/her other 1045 connections. The large difference between these two degrees however owning an average distance of around 4 people as well as on average each node would connect to around 44 other nodes again emphasizing the tight interconnectivity within the dataset.



```
The most popular person has 1045 degree
The most isolated person has 1 degree
Average degree: 43.69
Number of components: 1
```

After analyzing all these data, it’s better to understand the parabolic shape of the average distance for different amounts of nodes. To put us in a real-world situation with a community that has high connectivity, while the community size remains small, it is naturally easier to reach another person since if take the average amount of people that each node is connected to 44, it’s on average just 1 or 2 steps away. However, while the size of the community gets larger without a matching speed of edges growing, it expands the distance between people therefore it takes more steps for us to reach another person. And then in the end, as the community grows even larger with more edges and possibly with more ‘super-connected’ people like those who have 1045 degrees, to gather more people together adding a great amount of information flow, it decreases at a slower path as more people comes in, adding more stability. However, it’s unlikely that the number would keep decreasing but will most likely change slower and slower either increasing or decreasing for complicated reasons which we can discover more about in bigger and more complex data sets continued collected from the real world.