Yilan Hu Project Report

What does this project do

The dataset that is used is a social network in nodes and edges of Facebook, downloaded from SNAP. This project aims to write code to investigate whether most of the data in the data set fits into the idea of six degrees of separation and to dig what is the usual distance between paired data points. The project wants to find the distribution of distance for every paired node in the data set and conclude what is the most usual distance occurred. It will also control the number of neighbors to see what influence it has on the outcome.

Code

The design of the code is roughly cut into 5 areas: 1) file reading; 2) graph creation; 3) computation of distance (through Breadth-first Search); 4) output the longest distance; 5) output data distribution; 6) calculate statistical significance. The logic is first to read the text file. Then a graph has to be created, with the pre-request of finding the vertices of the data set and creating the adjacency list. Then we need to output the longest distance available in the data set in order to create a distribution collection. The distance distribution is a collection of tuples, where the first position of each tuple represents the distance, and the second position represents the occurrence frequency of the distance. Finally, based on the distribution vectors, we can output the median, average, and most occurred distance of the dataset.

The code contains 6 modules. The first module (“file.rs”) is created for importing and reading the text file. Then the creation module (“creation.rs”) contains functions that could output the vertices and adjacency list of the data set. In the main program, a struct called Graph is also created. The vertices needed to be output before the implementation of the struct Graph as the adjacency and creation function is inside the implementation. Then there is a module

(“distance.rs”) completing the Breadth-first search, which is the essential part of the code as it outputs the degree of distance for each node to its every possible pair. Even though this function is not in the main file, it is repeatedly implemented in other modules. The maximum module (“maximum.rs”) contains two functions: the node_maximum function iterates each node's output distance that is created through BFS and outputs the longest distance that appears for each node. Then the total_maximum function will output the longest distance of the dataset. The distribution module (“distribution.rs”) will return the vector of tuples of distribution. The distribution function also divides each output by two before returning them, because the function will calculate the paired twice (if node a and b’s distance is x, node b and a’s distance will be calculated again and x will return 2 times; though it has no influence on distribution). The last module (“result.rs”) is a toolbox of returning average, median, and most occurred frequency in the distribution collections of tuples. In addition, the module “genfile.rs” is also created to produce a new data set with a controlled number of neighbors. The generation file would not produce any node without a neighbor.

The design of the main file is by creating the graph first with struct, with the output of file reading and vertices finding prior to its establishment. Then I put every possible interesting feature that I want to explore in the implementation, rather than just putting in the function for creating the graph. As the project aims to test different characteristics of the degree of distance, I consider it legitimate to include different aspects of the graph inside as well.

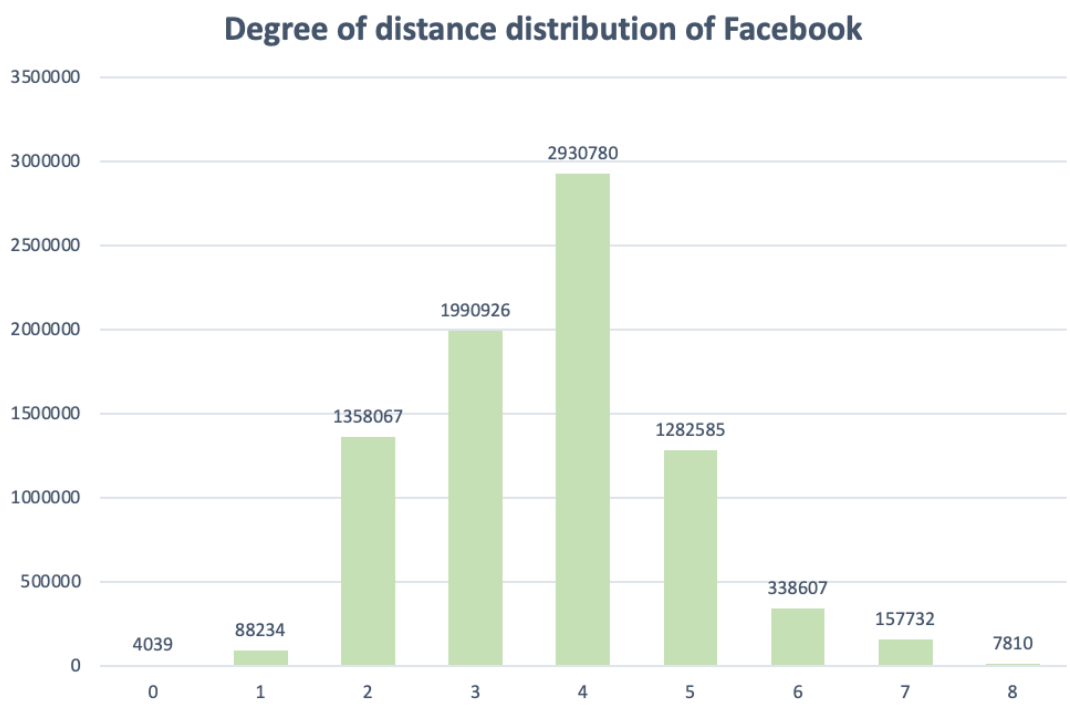
Then in the main function, I generate some extra data sets and repeat the same process to see whether the change of variable has any effect on the outcome, as well as does the average distance fits into six degrees of separation while differentiating from the Facebook data set. The first variable tested is the number of neighbors a node is allowed to have. In the original

facebook dataset, a single node is allowed to connect to many neighbors, while the first control group (selfgen_data1) eliminates it to 10. Then I shrunk the data size with the same elimination of the number of neighbors as dataset 2, and this would work as the second control group (selfgen_data2). The file generation function is put into comments after use.

Output

```
the longest degree of distance is 8
the degree of distance 0 has the frequency of 4039 times
the degree of distance 1 has the frequency of 88234 times
the degree of distance 2 has the frequency of 1358067 times
the degree of distance 3 has the frequency of 1990926 times
the degree of distance 4 has the frequency of 2930780 times
the degree of distance 5 has the frequency of 1282585 times
the degree of distance 6 has the frequency of 338607 times
the degree of distance 7 has the frequency of 157732 times
the degree of distance 8 has the frequency of 7810 times
the distribution of degree is [(0, 4039), (1, 88234), (2, 1358067), (3, 1990926), (4, 2930780), (5, 1282585), (6, 338607), (7, 157732), (8, 7810)]
the average degree of distance for facebook dataset is 3.6906788
the median degree of distance for facebook dataset is 4.0
the most occurred degree of distance for facebook dataset is 4
--------------------------------------------------
the longest degree of distance for the first self generated dataset is 7
the degree of distance 0 has the frequency of 4039 times
the degree of distance 1 has the frequency of 19999 times
the degree of distance 2 has the frequency of 198238 times
the degree of distance 3 has the frequency of 1691696 times
the degree of distance 4 has the frequency of 5193176 times
the degree of distance 5 has the frequency of 1044027 times
the degree of distance 6 has the frequency of 7576 times
the degree of distance 7 has the frequency of 29 times
the distribution of degree for the first self generated dataset is [(0, 4039), (1, 19999), (2, 198238), (3, 1691696), (4, 5193176), (5, 1044027), (6, 7576), (7, 29)]
the average degree of distance for first self generated dataset is 3.8645558
the median degree of distance for first self generated dataset is 4.0
the most occurred degree of distance for first self generated dataset is 4
--------------------------------------------------
the longest degree of distance for the second self generated dataset is 5
the degree of distance 0 has the frequency of 500 times
the degree of distance 1 has the frequency of 2446 times
the degree of distance 2 has the frequency of 21523 times
the degree of distance 3 has the frequency of 78586 times
the degree of distance 4 has the frequency of 22107 times
the degree of distance 5 has the frequency of 88 times
the distribution of degree for the second self generated dataset is [(0, 500), (1, 2446), (2, 21523), (3, 78586), (4, 22107), (5, 88)]
the average degree of distance for second self created dataset is 2.955034
the median degree of distance for second self generated dataset is 3.0
the most occurred degree of distance for second self generated dataset is 3
```

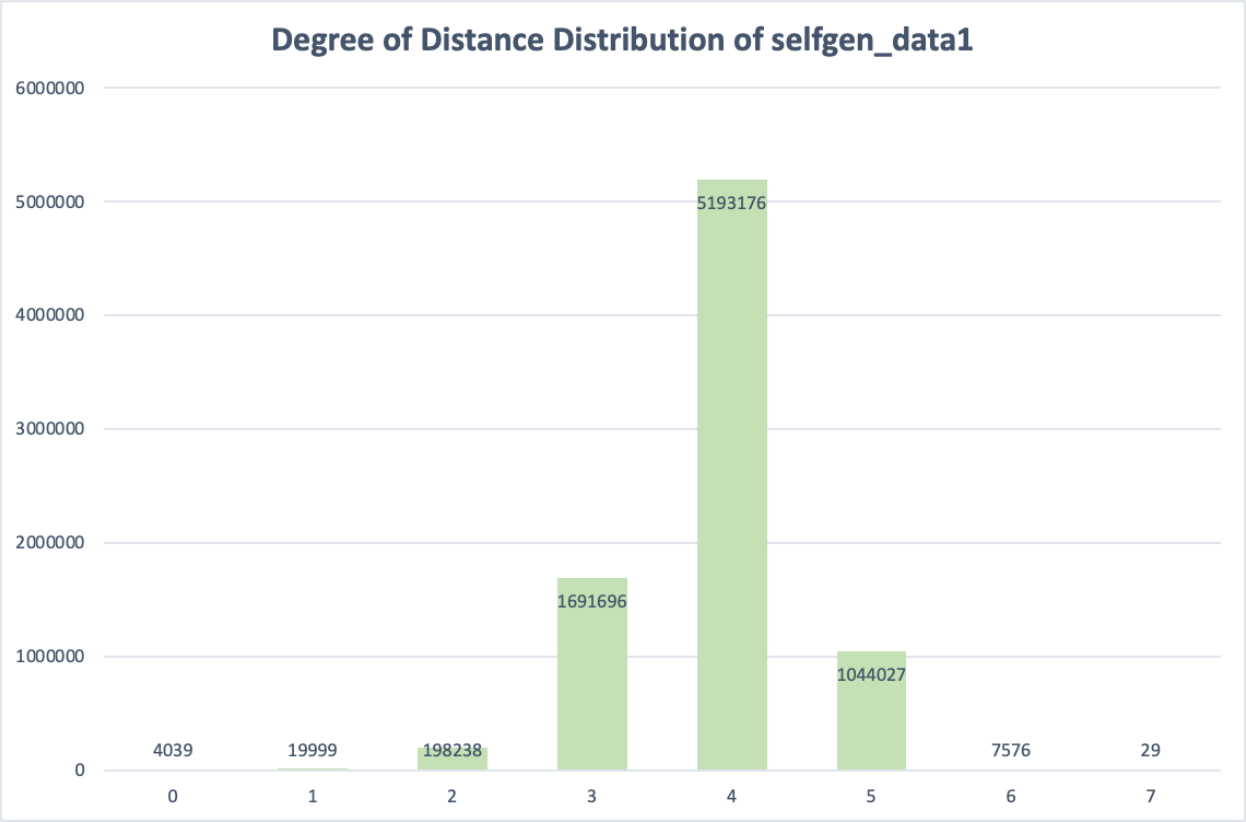
The output of the data shows that the longest degree of distance for the Facebook dataset is 8. I used Excel to plot the data distribution, which is the following:



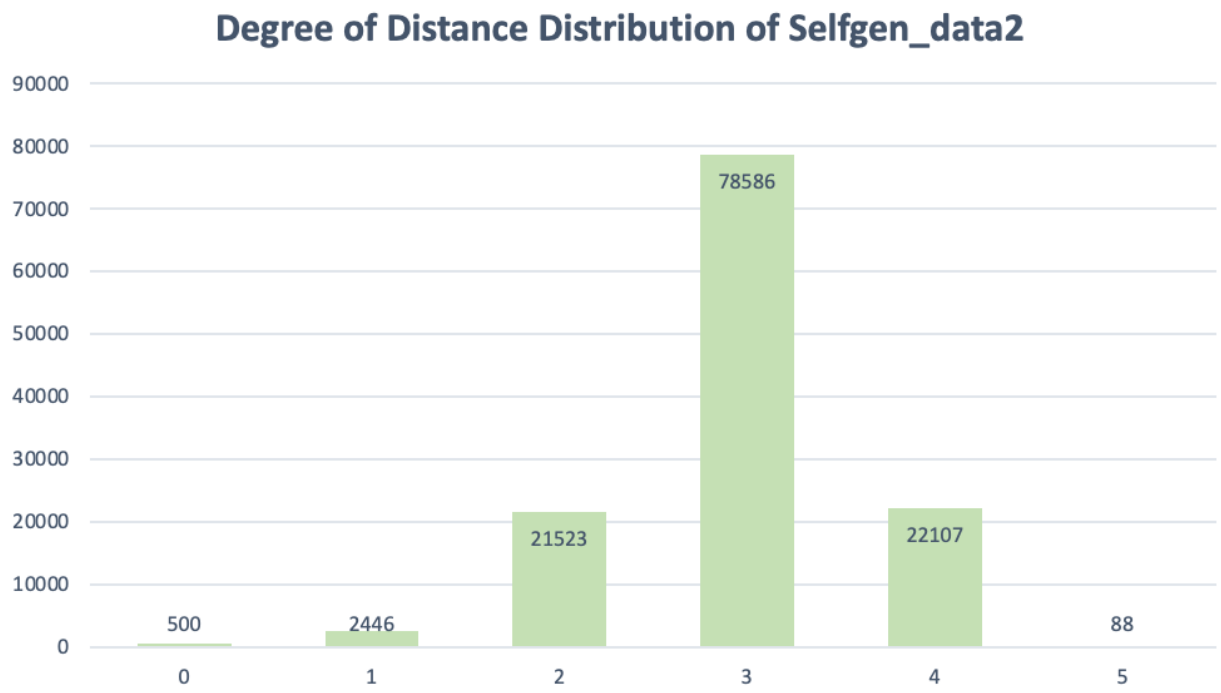
This is a roughly normal distribution, though slightly skewed to the right. Since it is roughly in the shape of normal distribution, the average degree returned is quite meaningful. It represents that the average distance one person can reach the other on Facebook is 3.6906788. However, both the median degree of distance and the most occurred degree of distance is 4. Such a phenomenon is due to the outliers of degree 7 and 8, which is outside of the range of six degrees of separation. It might have something to do with the size of the data, or the number of nodes connected is not large enough for nodes that have a distance of 8 returns. It requires further study. In general, for Facebook, most people reach the others within 3 to 4 other nodes.

The first control dataset (selfgen_data1) shows a similar pattern as the original Facebook dataset. The greatest distance between nodes is 7. The average, median, and most occurred degree of distance is also within 6. The average is 3.8645558, and the median and most occurred is 4. However, it is more concentrated on 4. The neighbor number variable's influence probably leads to the high concentration as there are fewer people connected to each node, it takes more

effort to build a connection for any paired nodes, resulting in the decrease of distribution for degree 2 and 3.



The second self-generated dataset controls both the number of neighbors and the size of the dataset. With both the size and number of neighbors restricted, it appears that the range of distribution shrunk to 5 and the degree concentrated on 3. The average is 2.955034 while the median and most occurred is 3. The size of the dataset and the number of neighbors certainly have an influence on the number of steps needed to take to reach the paired node. As the size of the dataset shrunk, the probability that people are connected to each other increased, and so it is more likely to output a shorter usual degree of distance for each paired matrix.



Conclusion

The average degree of distance in the Facebook dataset is 3.6906788. Most of the time, a node can reach its paired node within 6 steps. While generally fitting into the model, different datasets produce different results. The size of the dataset and the number of neighbors all have an influence on the distribution of distance, thus influencing the most occurred / median / average degree for each data set.

Cargo test result:

```
    Finished test [unoptimized + debuginfo] target(s) in 0.04s
     Running unittests src/main.rs (target/debug/deps/procode-3512b857568b4c9b)

running 3 tests
test result::test_average ... ok
test result::test_median ... ok
test creation::test_vertrices ... ok

test result: ok. 3 passed; 0 failed; 0 ignored; 0 measured; 0 filtered out; finished in 0.00s
```

Dataset citation:

J. McAuley and J. Leskovec. Learning to Discover Social Circles in Ego Networks. NIPS, 2012.