William Clavier

5 May 2024

GraphicallyNetworked Final Report

This project got its inspiration from the idea of the interconnectedness of social media. I found it interesting to contemplate how social media represents real life social connections. That meant that if we analyze the connections one has on social media that we would be able to then likely have a decent understanding of their actual relationships. That is why I chose this dataset of Twitter connections because I felt that I could hopefully get an understanding of the average distance between users on Twitter. This would likely correlate pretty well with real life distance between person-to-person relationships.

The method for finding the average distance between nodes in the dataset was Breadth-First Search. I created a method that calculates the distances of all of the other nodes in the network from a particular node. This method iterates through all of the connections until the entire network is traversed. Each time a new step is taken, one is added to the associated distance for that node. That enables the algorithm to keep track of the distance each node is from the start node.

The next method that completes the average distance algorithm uses multi-threading for increasing the execution speed. The algorithm is slow because it iterates through a sample of the nodes in the dataset network and for each of them perform the distance method previously mentioned. With the sets of distances for each node, the code will average the values in order to return a overall average distance for the dataset. This means that there are a lot of calculations that have to be done in order to get the average distance for a large sample size.

The output of my program is as follows:

```
For samples of size: 10000
i=0     Average node distance: 5.547644334912133
i=1     Average node distance: 5.567798345126696
i=2     Average node distance: 5.560872420680197
i=3     Average node distance: 5.565487214677027
i=4     Average node distance: 5.564177950239673
```

This brings me to the conclusion that Twitter is a decently cohesive network. Most nodes are on average ~5.6 nodes apart from one another. That was found by performing 5 iterations of 10,000 random samples from my dataset of nodes. This is pretty low considering the diversity of the world and the size of most people's social network followings. This dataset has 81,306 nodes which is quite large to have such a dense network. The low average distance also will imply that users have a large reach because their tweets will be able to reach a large audience because a lot of the audience is not far from the user.

This project has 2 tests that test the critical functionality of the code. Specifically the distance from a given vertex method and the average distance method. For both of these methods, I used simplified graphs to test them. That way I could verify that the output makes sense with what I am expecting.

The output of the tests is:

```
running 0 tests

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

     Running unittests src\main.rs (target\debug\deps\project-bb1cbc173c48de30.exe)

running 0 tests

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

     Running tests\integration_test.rs
(target\debug\deps\integration_test-79ffa117a8fc7ab1.exe)

running 2 tests
```

```
test test_bfs ... ok
test test_avg_dist ... ok

test result: ok. 2 passed; 0 failed; 0 ignored; 0 measured; 0 filtered out; finished
in 0.04s

   Doc-tests project

running 0 tests

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

If you represented a company that is seeking advertising opportunities, then you could run this type of test on multiple social media platforms in order to find the best person to sponsor. This person's posts will likely reach the largest audience and therefore should be a better return on investment for the company.