DS210 Final Report

In this Final project, I used the data set “Facebook Large Page-Page Network”, which is an undirected graph network that consists of Facebook pages and their connections. In particular, The dataset classified each page to one of the 4 categories: “tvshow”, “government”, “company”, and “politician”.

I chose this dataset because I would like to explore more on community detection and wanted to see if the communities that I classified are made based on the previous 4 types. My hypothesis is that each cluster of community have one type that is significantly higher than the other types of pages. For example, one community should have tv show pages significantly more than the other three pages, as they are communities based on tv shows.

Below is my approach:

I used the Girvan-Newman Algorithm to detect communities. The algorithm first calculates the betweenness centralities for each edge in the graph and remove the edge with the highest betweenness centrality measure. These edges are more likely to be connections between communities and removing them would expose existing communities. The algorithms continue to remove edges for hundreds of iterations to make each community clearer.

Secondly, I randomly chose 20 nodes to be the starting point and used BFS to record down the nodes 2-unit distance within the starting point, since these nodes are more likely to form a community. Thus, by randomly choosing 20 nodes I randomly chose 20 communities. I recorded down the type of page these nodes are and see if there’s a type that’s more dominant.

For testing, I wrote a few unit tests for functions to ensure that the functions are working intendedly.

Result:

After deletion of 200 edges and running 20 random point selections, I found that the result aligned with my initial hypothesis. For most of the points randomly picked, their nearby units all have one type significantly higher than the other three types, below is an example:

```
In this iteration 11 ,The community has 77 unit
Their types are: [("tvshow", 0), ("government", 2), ("company", 0), ("politician", 75)]
In this iteration 12 ,The community has 432 unit
Their types are: [("tvshow", 25), ("government", 365), ("company", 16), ("politician", 26)]
In this iteration 13 ,The community has 1050 unit
Their types are: [("tvshow", 26), ("government", 851), ("company", 116), ("politician", 57)]
In this iteration 14 ,The community has 18 unit
Their types are: [("tvshow", 2), ("government", 2), ("company", 1), ("politician", 13)]
In this iteration 15 ,The community has 261 unit
Their types are: [("tvshow", 27), ("government", 68), ("company", 15), ("politician", 151)]
In this iteration 16 ,The community has 10 unit
Their types are: [("tvshow", 8), ("government", 0), ("company", 2), ("politician", 0)]
In this iteration 17 ,The community has 308 unit
Their types are: [("tvshow", 0), ("government", 226), ("company", 8), ("politician", 74)]
In this iteration 18 ,The community has 104 unit
Their types are: [("tvshow", 4), ("government", 99), ("company", 0), ("politician", 1)]
In this iteration 19 ,The community has 56 unit
Their types are: [("tvshow", 0), ("government", 4), ("company", 52), ("politician", 0)]
```

Although the size of the community depends on the randomly picked starting point, the units in the community generally have one dominant type with numbers more than the other three types combined. Overall, I can conclude that my hypothesis is correct that the communities are made based on their types.