DS210 Final Project Report By: Marcos Sasson

For this final assignment, I examined a dataset taken from "fb-pages-company_edges.txt," which shows the relationships between various company-related Facebook pages as an undirected graph. In this graph, each node represents a company's Facebook page, and an undirected edge between two nodes denotes a mutual link or "like" relationship between those pages. Understanding how these pages are linked reveals the fundamental structure of their network and can provide insights into how information or influence may travel across them.

To learn about the connectedness and "distance" between these pages, I utilized the Breadth-First Search (BFS) technique. BFS not only identifies all vertices reachable from a given starting node, but it also automatically calculates the shortest path distances from that node to all other reachable nodes in an unweighted graph. Using BFS, I was able to effectively compute the shortest path between pairs of randomly selected nodes, allowing me to assess how tightly the network is connected.

While the dataset can be rather extensive, doing shortest path calculations on every potential pair of nodes would be extremely computationally costly. Instead, I used a sample approach to make the study more manageable. Specifically, after creating the graph from the edge list, I followed the procedures below:

1. **Constructing the Graph:**
   I read each line of "fb-pages-company_edges.txt," which comprised pairs of integers representing edges connecting two vertices. From these edges, I created an adjacency list to ensure that the graph structure captured the mutual "likes" between pages.
2. **Reachability via BFS:**
   I used BFS to identify which nodes were reachable, beginning with the dataset's initial vertex. As a result, I was given a subset of vertices that make up at least one connected graph component. I had a well-defined reachable subgraph from which to take samples since BFS created a distinct "frontier" of distances from the start vertex for these reachable nodes.
3. **Random Sampling of Pairs:**
   I chose up to 1000 different pairings of reachable vertices at random in order to calculate the total degree of connection. I could obtain a representative sample of path lengths without processing the complete dataset by looking at a comparatively small, randomized subset of node pairs.
4. **Computing Shortest Paths and Average Distances:**
   I used BFS to calculate the shortest path distance between each sampled pair of nodes. This step required calculating the fewest number of edges that connect one page to another. I then calculated the "average distance" metric by averaging the shortest path distances across all observed pairs. The theory behind this sampling strategy is that, while not every pair of nodes is investigated, a large enough random sample can approximate the network's true average distance. Over several runs, the average

distances will most likely settle to a stable value, reflecting the network's fundamental structure.

5. **Statistical Interpretation:**
   After finishing the analysis, I executed the code several times. Each run calculated the average shortest path distance for the sampled pairs. Although the actual values produced may differ significantly due to unpredictability in pair selection, the results frequently cluster around a single average which was 5 roughly around 5 edges. This means that any given company's page may be accessed from another company's page by traversing approximately five mutual "like" relationships.

Conclusion:

By using BFS for reachability and shortest path computations, and selectively sampling just a portion of node pairings, I was able to estimate the average shortest path distance between various company-related Facebook sites. Even though the dataset was vast and computationally demanding to evaluate in its whole, sampling approaches and BFS algorithms enabled me to efficiently draw relevant findings.

If we discover that the average distance between two random Facebook pages is consistently around a tiny integer (such as 5), this signals a highly interwoven network. Despite their apparent diversity, corporate pages are not isolated; instead, they can be linked together via a relatively short chain of intermediary relationships.

Overall, this investigation reveals the effectiveness of BFS and sampling methods in comprehending complicated networks. The method allowed for the estimate of the sampling distribution of the mean distance between pages, which provided insights into the underlying connectivity. With these methodologies, one may clearly say that, in a massive network of company pages, the social graph frequently remains "small," reflecting a "small-world" phenomena in which every given node is only a few steps away from any other.