Myat Thadar
DS210
Professor Leonidas Kontothanassis

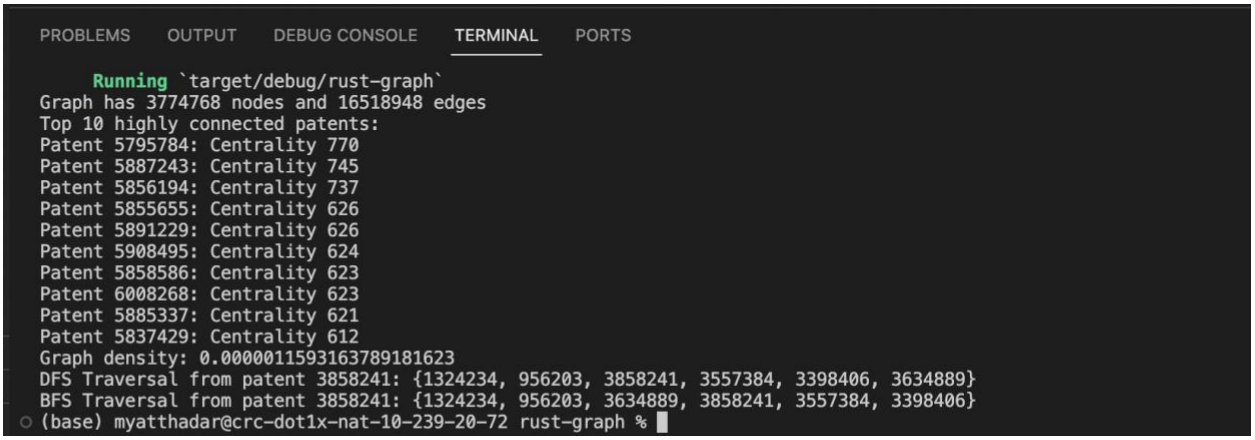
Analysis of US Patent Data Set

Introduction:

In this project, I analyzed a data set of almost 4 million utility patents that span across 37 years to discern and understand patterns in the rate of innovation. To understand the data, I have implemented multiple varying algorithms to

Data Set: https://snap.stanford.edu/data/cit-Patents.html

Output:



```
     Running `target/debug/rust-graph`
Graph has 3774768 nodes and 16518948 edges
Top 10 highly connected patents:
Patent 5795784: Centrality 770
Patent 5887243: Centrality 745
Patent 5856194: Centrality 737
Patent 5855655: Centrality 626
Patent 5891229: Centrality 626
Patent 5908495: Centrality 624
Patent 5858586: Centrality 623
Patent 6008268: Centrality 623
Patent 5885337: Centrality 621
Patent 5837429: Centrality 612
Graph density: 0.0000011593163789181623
DFS Traversal from patent 3858241: {1324234, 956203, 3858241, 3557384, 3398406, 3634889}
BFS Traversal from patent 3858241: {1324234, 956203, 3634889, 3858241, 3557384, 3398406}
○ (base) myatthadar@crc-dot1x-nat-10-239-20-72 rust-graph % █
```

Algorithms Implemented:

Degree Centrality

This function calculates the degree centrality of each node to measure the number of connections a node has. In this data set, the function calculates the number of other patents each patent cites or is cited by. My analysis shows that the top 10 patents with the highest centrality scores range from 612 to 770, which means even the most influential patents only cite, or are cited by, 612 to 770 patents. However, the fact that the highest centrality is 770 despite the network being over 3.7 million nodes and 16.5 million edges indicates an extremely sparse network overall with little connection amongst nodes.

Graph Density

This function calculates the density of the data set to measure how many connections between edges exist in the graph compared to the maximum number of possible edges. The value of 0.0000011593163789181623 reaffirms my previous analysis of the sparse network and could potentially be attributable to focused research. The patents in the data set could be research on individual niche topics that build upon a small number of prior work and are less connected amongst different research topics. The high sparsity and low density also indicates a diverse field of patents with limited fields.

<u>Depth-First Search (DFS) Traversal</u>

This graph traversal algorithm explores as far along each individual branch as possible before backtracking and is implemented to explore the full reach of each node's connections to the network. In my data set, it begins at a specific patent and iteratively follows each citation to the next patent to trace the lineage of innovations where the patents build upon one another. As the algorithm iterates through the patents, it marks each patent to make sure it does not revisit the same patent until it reaches a dead-end and backtracks to the most recent patent with unexplored connections. With this data set in particular, this algorithm serves the practical purpose of tracing the lineage of patents over time and allows us to see how innovations are formed on the basis of the previous one.

Limitation: large and complex network can become intensive, especially when you start from a highly-cited patent.

<u>Breath-First Search (BFS) Traversal</u>

This graph traversal algorithm explores neighbor nodes first before moving to next level neighbors to find the shortest path between nodes. It begins from a starting patent and visits all patents that are directly cited by, or cites, the starting patent before progression to the next layer. This allows us to identify immediate influence and connectivity and understand how connected a patent is to the immediate network. The algorithm also provides insight into the layers of connection within the dataset and shows us how the ideas are spread across other patents through direct and secondary citations. BFS Traversal is also useful to identify the shortest path in a network and highlight patents that are directly influencing one another.

Limitation: large and complex network can become intensive, especially when you start from a highly cited patent.