**Virushi Patel**

**DS 210**

**11th December 2023**

**Amazon Project Report**

**Project Introduction:**

The dataset, which comes from the "Customers Who Bought This Item Also Bought" feature on the platform, was gathered by crawling the Amazon website. The dataset is a directed graph, with directed edges connecting nodes representing frequent co-purchasing behaviours and nodes representing products sold on Amazon. The dataset is a snapshot of the relationships between co-purchasing that were seen on the Amazon website on March 2, 2003. It offers insightful information on the relationships between various products and reflects trends in consumer preferences and purchase behaviour. Analysing such a network can yield information about product recommendations, market trends, and the overall structure of the Amazon product ecosystem as it existed at that particular point in time.

**Code Description:**

The Graph struct uses a HashMap to store edges, where each node is associated with a vector of its neighbours. The program supports loading graph data from a file, with the file format representing edges between nodes. The main function demonstrates loading a graph from a file ("Amazon0302.txt") and then performing various analyses on it. Analysis functions include computing graph density, average degree, degree distribution, clustering coefficient, and various degree centrality measures. Additionally, there are test cases defined in the tests module to ensure the correctness of the implemented graph operations. The program uses the std::fs and std::io modules for file handling, and it depends on external crates such as rand and petgraph. The rand crate is likely used for testing random graph scenarios, and petgraph might be considered for more advanced graph operations in future developments.

**Metrics Analysis and Output: Screenshot of the output is showcased below.**

Density of the Graph: The concept of graph density is a fundamental metric in graph theory that quantifies the proportion of edges present in a graph relative to the total number of possible edges. It provides insights into how interconnected and complete a graph is. The output of the code reveals that the density of the graph loaded from the file "Amazon0302.txt" is approximately 0.000017974419114806206.

Density = Number of Edges / Total Possible Edges

For a directed graph with N nodes, the total possible edges is N×(N−1), as each node can be connected to every other node except itself. In the output, the density is quite low, indicating that the graph is sparse, with only a small fraction of possible edges realized. The low density in this particular instance suggests that there are few relationships between the entities that the nodes represent. This data can help direct future research into the characteristics of the data and clarify the particular context of the graph's application. Comprehending graph density is essential for describing the architecture of networks and for guiding further research and modelling choices. In summary, the Rust code's output of the graph's density is an important statistic that sheds light on the network's nodes' interconnectedness. The low density in this instance points to a sparse graph, and the information's importance varies according to the network's particular domain and context. The low density in the particular instance of the "Amazon0302.txt" graph suggests that there are few relationships between the entities that the nodes represent. This data can help direct future research into the characteristics of the data and clarify the particular context of the graph's application. Comprehending graph density is essential for describing the architecture of networks and for guiding further research and modelling choices. The low density in this instance points to a sparse graph, and the information's importance varies according to the network's particular domain and context.

Average Degree: The average degree is calculated by dividing the total number of edges in the graph by the number of nodes. The average degree value of 4.71 suggests that, on average, each node in the graph is connected to approximately 4.71 other nodes. This information is valuable in understanding the overall connectivity and complexity of the network. A higher average degree often indicates a denser and more interconnected graph, as each node has more edges, representing stronger connections between nodes. Conversely, a lower average degree might imply a sparser network with fewer connections per node. In the specific output, the value of 4.71 falls in between, suggesting a moderately dense graph. In summary, the output of the average degree at approximately 4.71 provides valuable information about the connectivity of the graph. It suggests a moderately dense network with nodes, on average, having around 4.71 connections. However, a comprehensive analysis involves considering this metric in conjunction with other measures such as degree distribution, graph density, and centrality to gain a holistic understanding of the network's structure and connectivity.

Degree Distribution: The degree distribution, as output by the provided Rust code, is a key indicator of the structural properties of the graph. In the output, the degree distribution is represented as a HashMap, where each key corresponds to a degree value, and the associated value is the frequency of nodes with that degree in the graph. Analysing this degree distribution provides valuable insights into the connectivity patterns within the graph. The degree distribution reveals the diversity of node degrees present in the graph. In this specific case, the degrees range from 0 to 5, indicating that nodes in the graph exhibit a variety of connectivity patterns. The most prevalent degree in the graph is 5, with an astonishing 233,871 nodes possessing this degree. This suggests a highly connected core within the graph, possibly representing central nodes in a network or hubs that play crucial roles in information flow. Conversely, lower-degree nodes (0, 1, 2, and 3) are also present, but their frequencies are noticeably lower, indicating a tiered structure where a majority of nodes are densely connected. It's also notable that there are nodes with a degree of 0. These nodes show that there are detached elements in the graph. Investigating these solitary nodes might reveal information about possible disconnected clusters or subgraphs with unique traits or purposes inside the larger network. The network's robustness and resilience may be inferred from the degree distribution. The elimination of low-degree nodes might not have a major effect on overall connectivity in a scale-free network. However, because of their key function, attacking high-degree nodes—in this case, those with a degree of 5—may possibly cause a more serious disruption to the network. Consequently, determining the degree distribution helps determine how vulnerable the network is to specific assaults. In conclusion, the degree distribution output of {2: 5654, 4: 7685, 1: 3803, 5: 233871, 0: 4541, 3: 6557} paints a vivid picture of the underlying graph's structure. It signifies a diverse network with varying degrees of connectivity, showcasing characteristics typical of a scale-free network. It is possible to identify key components, comprehend the scale-free nature of the network, and detect potential vulnerabilities by analysing the degree distribution. Using insights from the connectivity patterns within the enormous network of items and customer interactions, Amazon may use this information to influence strategic decisions about product suggestions, marketing plans, and logistical optimization. Through the lens of the degree distribution, Amazon is able to improve its comprehension of the dynamics of the network and optimize several facets of its e-commerce ecosystem.

Graph Clustering: The clustering coefficient, as calculated in the provided Rust program for a given graph, provides valuable insights into the structural organization of the network. In this specific analysis, the graph in question comprises 262,111 nodes, and the resulting clustering coefficient is approximately 0.905. This coefficient is a numerical representation of the extent to which nodes in the graph tend to cluster together. A clustering coefficient close to 1 suggests a high level of interconnectedness among the neighbours of individual nodes. In practical terms, a clustering coefficient of 0.905 indicates a network where nodes are strongly inclined to form local clusters or communities. This phenomenon aligns with the characteristics commonly observed in social networks, biological systems, and certain technological networks. Nodes with high clustering coefficients imply that their neighbouring nodes are likely to be interconnected, forming cohesive groups within the larger network. Furthermore, the variance of the node degrees, another aspect of the analysis, offers additional depth to our understanding of the network's structure. The variance in degrees measures the spread of node degrees around the average degree. In this context, a higher variance suggests a greater diversity in the number of connections that nodes possess. The substantial variance, coupled with the high clustering coefficient, indicates a network where some nodes act as hubs with numerous connections, while others have fewer connections, resulting in a heterogeneous distribution of degrees. In the context of a social network or a recommendation system like Amazon, this could mean that products (nodes) tend to be connected in groups. In the case of Amazon, these groups might represent clusters of products frequently bought together or related in some way.

Degree Centrality: One of the key metrics examined is the average degree centrality, a measure that assesses the importance of nodes within the graph based on their connectivity. The output for the average degree centrality is 1.797441911483109e-5, signifying a very small centrality value. This suggests that, on average, nodes in the graph are not highly central, meaning that there is no single node that dominates in terms of connections. In this specific instance, the node with the maximum degree centrality are identified with multiple nodes 98870, 69541,188396 with a centrality of 1.9075960474609896e-5. These nodes possess the highest level of connectivity within the graph. In a social network or information flow context, these nodes might represent pivotal individuals or crucial hubs that significantly impact the flow of information or influence throughout the network. Conversely, the node with the minimum degree centrality is labelled as multiple nodes 262062, 248403, 65460… with a centrality of 0.0. These extremes shed light on the structure of the graph and can provide insights into the distribution of influence or connectivity within the network. These nodes have minimal or no connections to other nodes in the graph. Their lack of connectivity might indicate peripheral or isolated elements within the network, possibly suggesting nodes that are disconnected from the main structure or play insignificant roles in the overall network dynamics. The analysis's conclusions draw attention to the graph's structural variety. The low average degree centrality indicates a more evenly distributed connectivity pattern in a network where nodes do not show significantly different amounts of influence. The existence of nodes with varying degrees of significance or impact, on the other hand, is demonstrated by the presence of nodes with noticeably higher and lower degree centralities (such as the maximum and minimum centrality nodes), suggesting the presence of important hubs in addition to more peripheral or isolated elements within the network. The node centrality metrics distribution shows a diverse structure, with some nodes having significant effect and others being marginal. Understanding such centrality metrics aids in comprehending network dynamics, identifying critical nodes for information flow or control, and characterizing the overall resilience or vulnerability of a network to targeted node removal or disruptions.

Overall Analysis: Beginning with the fundamental concept of graph density, which stands at approximately 0.000017974419114806206, it becomes evident that the graph exhibits a sparse nature with limited relationships among its represented entities. This low density hints at the potential presence of disconnected clusters or subgraphs within the dataset, prompting further investigation into these isolated elements. Complementing the density metric, the average degree of 4.71 signifies a moderately dense network. On average, nodes are connected to approximately 4.71 other nodes, painting a picture of interconnectedness within the graph. This insight is enriched by the degree distribution, showcasing a diverse range of connectivity patterns. Notably, nodes with a degree of 5 emerge as a highly connected core, potentially representing pivotal hubs influencing information flow. Conversely, the presence of nodes with a degree of 0 indicates detached elements, ripe for exploration to understand their role within the larger network. The high clustering value of about 0.905 reveals a propensity for nodes to form local communities or clusters, which is similar to what is observed in social and technological networks and emphasizes the cohesiveness of the network even more. Significant variation in node degrees is entwined with this clustering tendency, suggesting a diverse distribution of connections where some nodes function as hubs and others have less connections. Node relevance inside the network is more complicated when degree centrality measurements are evaluated. The low average degree centrality (1.797441911483109e-5) indicates that nodes are connected in an even manner with little variation in influence. On the other hand, nodes that possess the highest degree of centrality become significant nodes that could be important players or crucial components in the dynamics of the network. Conversely, nodes with minimum degree centrality signal peripheral or isolated elements with minimal connections, shedding light on less significant areas within the network. Essentially, the combination of these metrics reveals the intricate structure of the "Amazon0302.txt" dataset. The network's low density and rather high average degree reveal a variety of connectivity patterns, including both isolated nodes and heavily connected hubs. A scale-free network structure is shown by the degree distribution, while cohesive clustering tendencies among nodes are highlighted by the high clustering coefficient.

**Connection to Amazon Dataset:** Knowing these dynamics gives Amazon strategic information that it can use to improve its e-commerce ecosystem, provide better product recommendations, hone its marketing approaches, and adjust to changing interactions between customers and products. These graph theory metrics provide a thorough analysis that gives Amazon the means to explore and leverage the complex relationships found in its enormous network of products and customer interactions, which in turn affects many aspects of its operational strategy. There can be several promotional offers such as buy one get one free, or bundled discounts to increase sales of certain products. Isolated nodes which have a degree distribution of 0 can be further analysed to find why they are solitary. Company website can design their page such that related products immediately pop up while buying a particular product. Amazon can improve its comprehension of the interactions between customers and products by coordinating marketing initiatives, promotional methods, and logistics optimization with the recognized network features. Increased revenues, higher consumer satisfaction, and a more responsive and dynamic e-commerce ecosystem are all possible outcomes of this calculated approach.

**Testing Output:**

**A screenshot of a computer screen

Description automatically generated**

**Output:**

A screenshot of a computer

Description automatically generated

There is a list of nodes with maximum degree centrality that ends in the next screenshot

A computer screen shot of a number

Description automatically generated

There is a list of nodes with Minimum degree centrality that ends in the next screenshot

A screenshot of a computer

Description automatically generated