Yutong Wu

U98858372

Collaborators: none

Final Project write-up

**Dataset:**

Because the dataset is too big to upload to Git Hub, it can be downloaded from this link:

https://snap.stanford.edu/data/amazon-meta.html

**Project Description:**

This project explores co-purchasing behavior across product categories using Amazon product metadata. The analysis leverages directed graphs to model these relationships, with nodes representing unique products identified by their ASIN (Amazon Standard Identification Number). Edges between nodes, derived from the dataset's "similar" field, represent co-purchase patterns, where a link from product X to product Y indicates that customers who purchase X frequently also purchase Y. The objective of this analysis is to uncover patterns in consumer purchasing behavior and product associations.

Given the dataset's large size and the significant time required to process it fully, I selected a random sample of 100,000 products for analysis. From this sample, I created two graphs to examine different aspects of co-purchasing behavior.

The first graph, generated using the create_graphs_for_top_categories function, focuses on the top three categories with the highest number of products. To ensure category specificity, I restricted both nodes and edges to products within the same category. I then calculated the Average Degree Centrality for each category using

the calculate_average_degree_centrality function. This involved counting each node's connections, summing the degrees of all nodes within the graph, and dividing the total degree by the number of nodes. This analysis provided insights into the connectivity and co-purchase patterns within individual categories.

The second graph, created using the create_global_graph function, includes all products in the sample as nodes and edges, without filtering by category. Using this graph, I calculated the in-category and cross-category ratios for each category with the calculate_co_purchase_ratios function. For this calculation, I traversed all edges in the graph and examined the categories of the source and target nodes connected by each edge. Edges were classified as in-category if both nodes belonged to the same category, and cross-category if the nodes belonged to different categories. From these classifications, I derived two key metrics:

1. In-Category Ratio: the number of in-category edges divided by the total number of edges.
2. Cross-Category Ratio: the number of cross-category edges divided by the total number of edges.

These metrics offered valuable insights into how co-purchase behavior is distributed—whether it remains concentrated within specific categories or extends across multiple categories—highlighting patterns of consumer preferences and cross-category associations.

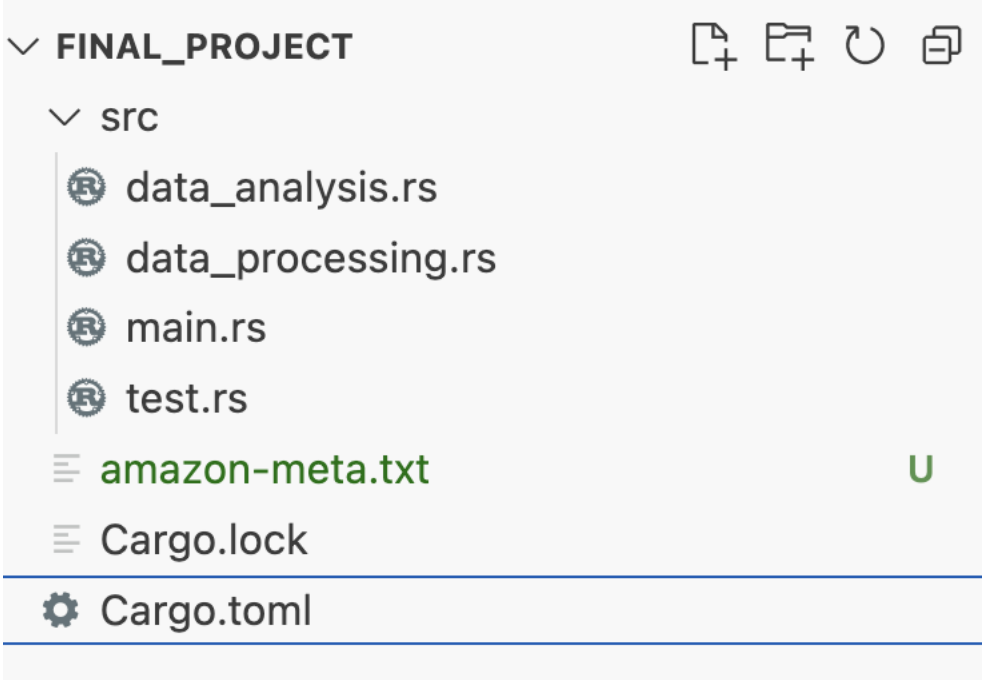
**File and Structure:**

1. data_processing.rs: Contains logic for loading and cleaning the dataset, extracting product data, and building both category-specific and global graphs.
2. data_analysis.rs: Implements functions for analyzing graphs, including calculating in-category and cross-category co-purchase ratios.

3. main.rs: The main entry point for the application.

4. test.rs: contains three test functions to test their related functions: test_adjacency_list, test_calculate_average_degree_centrality and test_calculate_co_purchase_ratios.

**How to run the project:**




I split the project into three separate modules: data_processing, data_analysis, and test, all of which are executed from the main file. To run the project, the amazon-meta.txt file must be placed in a folder named final_project, as shown above. Due to the large size of the dataset, I recommend using cargo run --release to improve runtime performance. The run command should be executed from the project repository containing the src and data directories, and it will handle both data processing and graph analysis. To test the functions, use the cargo test command.

**Output:**

Top Categories in Random Sample:

Category: Book

Number of Products: 71800

Average Sales Rank: 631909.61

Average Review Rating: 4.16

Category: Music

Number of Products: 18781

Average Sales Rank: 155563.89

Average Review Rating: 4.34

Category: Video

Number of Products: 4679

Average Sales Rank: 31418.92

Average Review Rating: 4.12

In the dataset, some categories have only a small number of products. To ensure sufficient data and co-purchasing connections for analysis, I selected the top three categories with the highest number of products.

Average Degree Centrality for Category Book: 0.43

Average Degree Centrality for Category Music: 0.36

Average Degree Centrality for Category Video: 0.15

The average degree centrality values highlight the density of connections (edges) within each category-specific graph, reflecting the extent of co-purchasing behavior within categories. The highest average degree centrality is observed in the "Book" category (0.43), indicating a densely connected network where customers frequently co-purchase books, creating a robust internal

structure. The "Music" category shows a moderately high average degree centrality (0.36), suggesting common co-purchases within the category, though with slightly less dense connections compared to "Books," potentially reflecting more specialized or dispersed buying behavior. In contrast, the "Video" category has the lowest average degree centrality (0.15), pointing to sparse connections and suggesting infrequent co-purchases or more niche-specific buying patterns within the category.

Global Graph created with 100000 nodes and 41427 edges.

Co-Purchase Ratios:

Category: DVD - In-Category Ratio: 0.94, Cross-Category Ratio: 0.06

Category: Video - In-Category Ratio: 0.38, Cross-Category Ratio: 0.62

Category: Music - In-Category Ratio: 0.98, Cross-Category Ratio: 0.02

Category: Book - In-Category Ratio: 0.99, Cross-Category Ratio: 0.01

The results of the co-purchase analysis reveal distinct patterns of in-category and cross-category behavior across different product categories within the global graph, which consists of 100,000 nodes and 41,427 edges:

1. **DVD**: With an in-category ratio of 0.94 and a cross-category ratio of 0.06, co-purchases within the DVD category are highly concentrated within the category itself. This suggests that consumers purchasing DVDs tend to stick to other DVDs rather than products from different categories.

2. **Video**: The in-category ratio of 0.38 and cross-category ratio of 0.62 indicate that co-purchases in the video category are more likely to span across different categories than remain within the category. This highlights a diverse co-purchasing behavior for video-related products, possibly due to overlaps with other entertainment or media categories.
3. **Music**: The in-category ratio of 0.98 and cross-category ratio of 0.02 show an overwhelming preference for co-purchases within the music category. This suggests that customers buying music products are strongly inclined to purchase other music-related items, showcasing a focused and specialized buying pattern.
4. **Book**: With an in-category ratio of 0.99 and a cross-category ratio of 0.01, books demonstrate the highest level of in-category co-purchasing behavior. This reflects a clear trend where consumers purchasing books are almost exclusively interested in other books, indicating minimal overlap with other categories.