DS210 Final Project Writeup

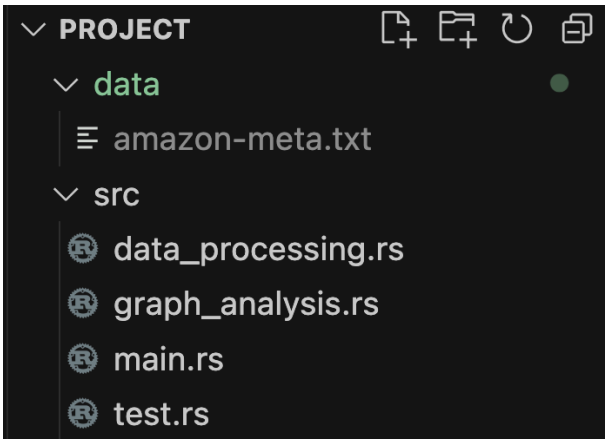
Project description:

For this project I analysed the Amazon product co-purchasing network metadata and produced a graph. The graph is an undirected graph with the nodes representing an Amazon product represented by their ASIN number (the products unique Amazon code number) and each node it is connected to represents products that it is often co-purchased with i.e if node A is connected to node B, people who buy product A they are likely to buy product B and vice versa. Because this Amazon data is from the early 2000s when Amazon, the main products with a large number of products are DVDs, music, videos and books, thus, most of the analysis will be centred around these 4 categories despite there being other categories in the dataset due to the low item counts of products in other categories.

To explore the co-purchasing relationship between different categories of products. I implemented 3 main tests.

1. The average degree centrality for each category. High average degree centrality in a category would mean that the products that are frequently bought together with many other products, suggesting high popularity or versatility.
2. The likelihood of someone purchasing a product of the same category based on the category of the initial product. A high likelihood of someone purchasing another product from the same category would suggest the product is self-contained, meaning people who bought that category of product are most likely to buy the same category of product instead of other products.
3. What are the average number of items in other categories people bought depending on the category of their initial purchase. If another category of item is bought frequently with a specific category, it could suggest that these two categories are complementary goods. Meaning if someone bought this category they are likely to buy products of a specific category because they work well together or are relevant to one another.

How to run project:



I split the data processing, graph analysis and tests into 3 separate modules all run in the main file. To run the project, the amazon-meta.txt file should be located in a separate folder called “data” in the main file along with src as seen on the right. Due to the large size of the data file, I recommend using cargo run --release to speed up the run time. The run command should be executed while being in the project repository containing src and data and should run all of the data processing, graph analysis and tests in one command. The data file was too large to be uploaded to github and can be downloaded from this link:

https://drive.google.com/file/d/1E9_B9O7dOF6mXUVmHuwwQUU7BBZDDn_u/view?usp=drive_link

Output:

Categories
CE - Item Count: 4
Video - Item Count: 26406
Book - Item Count: 397901
Sports - Item Count: 1
Video Games - Item Count: 1
Music - Item Count: 104210
Toy - Item Count: 8
Baby Product - Item Count: 1
Software - Item Count: 5
DVD - Item Count: 20014
 - Item Count: 1

For number of categories, the output should look like the above. Because the number of items in other categories excluding DVD, music, video and Books are so low, I focused my analysis only on DVD, music, video and Books and their relation with one another.

Avg Degree Centrality for each category
Category: Book - Average Degree Centrality: 7.769407465676135
Category: Video - Average Degree Centrality: 6.896803756721957
Category: DVD - Average Degree Centrality: 10.815778954731687
Category: Music - Average Degree Centrality: 6.72018040495154

Avg Degree Centrality for each category based only on nodes with neighbours
Category: DVD - Average Degree Centrality: 12.064147578442848
Category: Book - Average Degree Centrality: 11.051923166297847
Category: Music - Average Degree Centrality: 10.852135374697824
Category: Video - Average Degree Centrality: 9.878335864612714

For degree centrality, the output should look like the above. I ran two separate degree centrality tests. One including nodes with neighbours and one including nodes without neighbours because certain products did not have any purchases possibly due to factors such as being new to Amazon thus I did not want these factors to affect the analysis. From this, we can see that on average DVDs had the highest degree centrality, followed by Books then Music then Videos. Meaning that people who bought DVDs and books were more likely to buy a wider range of other DVDs and books compared to music and video. Suggesting people were more interested in purchasing multiple DVDs and Books compared to other items. This could also reflect that there was a wider range of selection for DVD and books on amazon leading to their higher degree centrality as people would be more likely to buy other books and DVDs if there was a wider selection to choose from. Additionally, because of how amazon connects products through recommendations, it could suggest that people were more likely to recommend books and DVDs leading to a wider network of co-purchasing for books and DVDs.

Likelihood of each category purchasing another product of its own category:
Category: Music - Likelihood: 0.7178806528537363
Category: DVD - Likelihood: 0.5854887811999058
Category: Video - Likelihood: 0.3077801633016138

Category: Book - Likelihood: 0.7377814006673233

For the likelihood of each category purchasing another product of its own category, we can see that people who bought Books from Amazon were more likely to only buy books whereas people who bought videos were more likely to buy products from other categories not just from videos. This could mean that people who bought books were less interested in other product categories whereas people who bought videos were more interested in other categories such as DVDs or music. The lower likelihood of same category purchases of videos and DVDs could suggest that the two products were interconnected as people who bought DVDs or videos were more likely to purchase from other categories which could mean people buying videos were more likely to also buy movies and vice versa or buy books that were related to movies. Suggesting that videos and DVDs are good products to encourage consumers to purchase other categories.

Average product category co-purchases per Category:
Category: Video
  Co-purchases Video: 2.12
  Co-purchases Book: 0.16
  Co-purchases Music: 0.07
  Co-purchases Sports: 0.00
  Co-purchases DVD: 2.37
Category: Book
  Co-purchases Music: 0.01
  Co-purchases Toy: 0.00
  Co-purchases Video Games: 0.00
  Co-purchases Book: 5.73
  Co-purchases Software: 0.00
  Co-purchases DVD: 0.02
  Co-purchases Video: 0.01
Category: Music
  Co-purchases DVD: 0.05
  Co-purchases Video: 0.01
  Co-purchases Music: 4.82
  Co-purchases Book: 0.02
  Co-purchases Software: 0.00
  Co-purchases Toy: 0.00
Category: DVD
  Co-purchases Video: 0.75
  Co-purchases DVD: 6.33
  Co-purchases Book: 0.17

Co-purchases Music: 0.16

From the above, I analysed the average number of items people bought from each category. We can see for most categories, people would only purchase items within the same categories. However, for videos, the average user would buy approximately 2 videos and 2 movies. Suggesting that movies and videos are complementary goods. Meaning people who were interested in buying videos were also very keen on buying DVDs. However, people who bought DVDs were not very interested in buying videos. Additionally, we can see that on average, books and DVDs have the highest number of average co-purchases with books at 5.73 books and DVD at 6.33 DVDs. Meaning people who bought these items often bought similar or related books and DVDs at a much larger number compared to other categories which suggest their popularity.

Average Degree Centrality:
Category 2: 3.00
Category 1: 2.50
Category 3: 2.00

Category purchasing likelihood:
Category 1: 0.40
Category 2: 0.00
Category 3: 0.50

For tests, I implemented two tests to test the average degree centrality function and the Category purchasing likelihood function. The reasoning behind their solution to prove they are working correctly are included in comments within the test.rs file.