Exploration of Amazon's Co-Purchase Network

Dhairya
Roll No: 2022157
dhairya22157@iiitd.ac.in

Harsh Vishwakarma
Roll No: 2022205
harsh22205@iiitd.ac.in

Divyansh
Roll No: 2022178
divyansh22178@iiitd.ac.in

1. Introduction

This project focuses on analyzing a real-world co-purchase network derived from product purchases. Each node represents a product, and a directed edge represents a co-purchase relationship. The study includes topological characterization, community detection, scale-free and small-world analysis, and predictive modeling.

2. Dataset Description

We have taken the datasets from Amazon Copurchase dataset, which contains two files products.csv and copurchase.csv. This study utilizes two interrelated datasets to analyze product characteristics and their co-purchase behavior:

2.1. Product Metadata (`products.csv`)

This dataset contains detailed metadata for a collection of products, primarily books and music. Each row corresponds to a unique product and includes the following attributes:

- **id**: Unique identifier for each product.
- **title**: Title of the product.
- **group**: Category of the product (e.g., Book, Music).
- **salesrank**: Sales ranking of the product (lower values indicate higher popularity).
- **review_cnt**: Number of customer reviews for the product.
- **downloads**: Number of times the product has been downloaded.
- **rating**: Average customer rating (on a scale of 0 to 5).

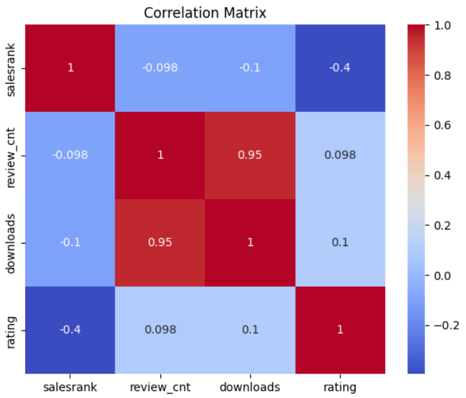


Figure 1: Product dataset Correlation Matrix

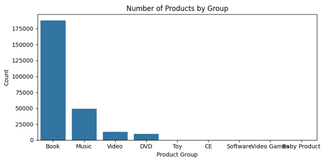


Figure 2: Number of Products by Group

This dataset enables analysis of product popularity, user engagement, and content ratings across different product categories.

2.2. Product Co-purchase Network (`copurchase.csv`)

This file represents a directed graph structure of co-purchase relationships between products. Each row contains:

- **Source**: The product ID of the primary item being purchased.
- **Target**: The product ID of another item frequently co-purchased with the source item.

This dataset models a *product recommendation network*, where a directed edge from product A to product B indicates that product B is often bought together with A.

Statistic	Value
Number of nodes	310,771
Number of edges	915,666
Average degree	5.89
Maximum degree	384
Minimum degree	1

Table 1: Key statistics of the co-purchase network.

3. Structural and Topological Analysis

3.1. 1. Degree Distribution (Log-Log)

We compute in-degree and out-degree distributions to assess how product popularity is distributed. The log-log plot reveals a long-tail behavior consistent with real-world networks, suggesting few products are highly popular.

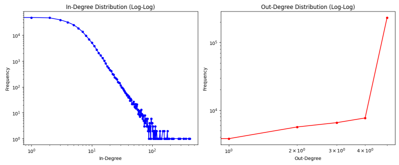


Figure 3: Degree Distribution (Log-Log)

3.2. 2. Average Clustering Coefficient

We compute the average clustering coefficient to measure the tendency of products to be co-purchased together. A relatively high coefficient (e.g., 0.42) indicates dense local neighborhoods.
Average Clustering Coefficient: 0.4198

3.3. 3. Assortativity

The assortativity coefficient evaluates whether nodes with similar degrees tend to connect. The value obtained indicates whether the network exhibits assortative or disassortative mixing.
Assortativity Coefficient: -0.0025

3.4. 4. Connected Components

We identify strongly and weakly connected components. The largest weakly connected component includes the majority of the nodes, showing the network is well-connected. Strongly connected components are fewer, due to directionality.

Metric	Value
No. of Strongly Connected Components	6,595
No. of Weakly Connected Components	1
Size of Largest SCC	241,761
Size of Largest WCC	262,110

Table 2: Connected Component Statistics of the Co-Purchase Network

3.5. 5. Comparison with Erdős–Rényi Random Graph

We generate a random graph of the same size (same number of nodes and edges) and compare key statistics (average degree, clustering, and path length). The real network displays a higher clustering coefficient and longer path length, indicating small-world structure.

Table 3: Comparison of Structural Properties: Real Network vs. Erdős–Rényi Random Network

Metric	Original	Random
Nodes	262,110	262,110
Directed Edges	1,234,870	1,234,870
Undirected Edges	899,787	1,234,861
Weakly Connected Components	1	20
Strongly Connected Components	6,595	4,819
Largest Strong Component Size	241,761	257,292
Average Clustering Coefficient	0.4198	0.000032
Clustering Ratio(Original/Random)	13,200.06×	
Estimated Avg Shortest Path Length	8.8626	5.8097
Path Length Ratio(Random/Original)	0.66×	
Average Degree	6.87	9.42
Maximum Degree	420	24

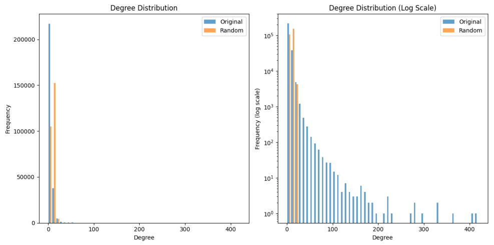


Figure 4: Degree Distribution Comparison

4. Community Detection and Modularity Analysis

4.1. 1. Analyze Community Structure

Community detection was performed using the Louvain method on the largest connected component of the undirected co-purchase network (262,110 nodes). A total of 164 communities were detected, achieving a high modularity score of 0.9016, indicating strong community structure.
Number of communities detected: 164
Modularity score: 0.9016
Size of largest community: 16517

4.2. 2. Community Visualization

Top communities are visualized with central products labeled, revealing thematic or category-based groupings.

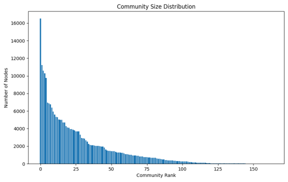


Figure 5: Community Size Distribution

Top 5 communities by size:

- **Community 10:** 16,517 nodes
  Top central products and their PageRank scores:
  Product ID: 1241 (0.0011), 3661 (0.0011), 9131 (0.0010), 3673 (0.0010), 1964 (0.0009)

- **Community 33:** 11,217 nodes
  Top central products and their PageRank scores:
  Product ID: 14949 (0.0041), 30171 (0.0013), 7303 (0.0012), 22073 (0.0010), 9955 (0.0010)

- **Community 42:** 10,596 nodes
  Top central products and their PageRank scores:
  Product ID: 12608 (0.0011), 6067 (0.0010), 14573 (0.0008), 13932 (0.0008), 36122 (0.0008)

- **Community 62:** 10,289 nodes
  Top central products and their PageRank scores:
  Product ID: 31037 (0.0019), 15934 (0.0017), 26010 (0.0015), 20898 (0.0011), 20899 (0.0009)

- **Community 36:** 9,770 nodes
  Top central products and their PageRank scores:
  Product ID: 61341 (0.0018), 15925 (0.0014), 19527 (0.0012), 70448 (0.0012), 42721 (0.0011)

5. Scale-Free and Small-World Properties

5.1. 1. Scale-Free Network Analysis

To evaluate whether the network exhibits scale-free behavior, we analyzed the degree distribution of the network and attempted to fit a power-law model. The following statistics summarize the findings:

- **Power-law exponent (α):** 3.6179
- **Minimum degree for power-law fit ($x_{\mathbf{min}}$):** 19
- **Log-likelihood ratio (power-law vs. log-normal):** -0.0698
- **p-value:** 0.8385

Although the degree distribution shows heavy-tailed characteristics, the log-likelihood ratio and high p-value suggest that it is not statistically significant enough to confirm a strict power-law behavior. Therefore, while the network may exhibit some scale-free tendencies, it does not definitively follow a power-law distribution.

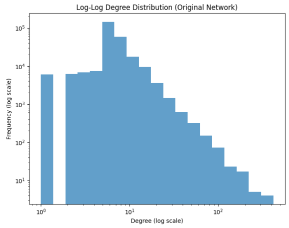


Figure 6: Power Law Check

5.2. 2. Small-World Characteristics

To examine whether the network demonstrates small-world properties, we analyzed the largest connected component (LCC), which contains 234,135 nodes and 711,194 edges. The small-world nature is evaluated by comparing the network's clustering coefficient and average path length to those of a random graph with the same number of nodes and edges.

- **Average shortest path length:** 9.8852
- **Average clustering coefficient:** 0.3647
- **Random graph clustering coefficient:** 0.0000
- **Clustering ratio (original/random):** 11,397.82
- **Random graph path length:** 5.8097
- **Path length ratio (random/original):** 0.5877

The network has a relatively high clustering coefficient compared to a random graph, and the average shortest path length remains small. These properties are consistent with the small-world phenomenon.

5.3. 3. Cumulative Degree Distribution

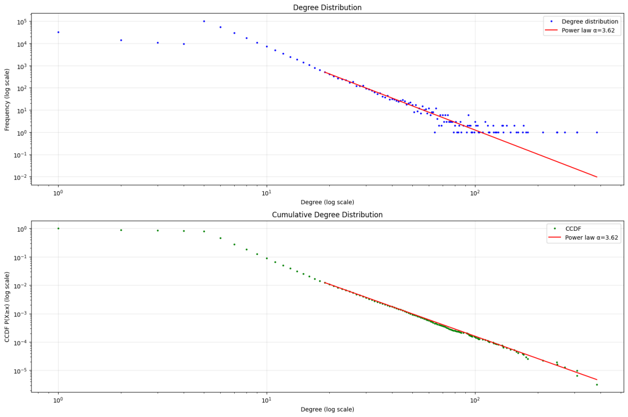


Figure 7: Cumulative and Degree Distribution

6. Predictive Modeling for Co-Purchases

In this section, we develop machine learning models to predict whether a given pair of products is likely to be co-purchased. The aim is to build a recommender system that identifies probable future co-purchases based on existing product network data and product metadata.

6.1. 1. Pre-processing

The dataset includes product metadata (`products.csv`) and co-purchase relationships (`copurchase.csv`). We filtered the dataset to retain only products in the "Book" category and with a valid sales rank. From the filtered product set, we extracted relevant co-purchase pairs to build a directed graph representing purchase relationships.

Each product pair (u, v) was represented using the following features:

- **Salesrank Difference**: Absolute difference in sales rank between the two products.
- **Preferential Attachment**: Product of node degrees.
- **Rating Product**: Product of product ratings.
- **Common Neighbors**: Number of shared neighbors in the co-purchase graph.
- **Jaccard Coefficient**: Ratio of common to total neighbors.
- **Resource Allocation Index**: Weighted sum over common neighbors.
- **Adamic-Adar Index**: Emphasizes rare shared neighbors.

6.2. 2. Classification Model

The following classifiers were trained using randomized hyperparameter search:

- **Random Forest Classifier(RF)**
  - `n_estimators`: [50, 300]
  - `max_depth`: [None, 5–30]
  - `min_samples_split`: [2, 20]
  - `min_samples_leaf`: [1, 10]
  - `max_features`: ['sqrt', 'log2', None]
- **Gradient Boosting Classifier(GB)**

  - `n_estimators`: [50, 300]
  - `learning_rate`: [0.01, 0.3]
  - `max_depth`: [2, 10]
  - `min_samples_split`: [2, 20]
  - `min_samples_leaf`: [1, 10]
  - `subsample`: [0.6, 1.0]

- **Logistic Regression(LR)**
  - `C`: [0.1, 10]
  - `penalty`: ['l1', 'l2', 'elasticnet', None]
  - `solver`: ['newton-cg', 'lbfgs', 'liblinear', 'saga']
  - `max_iter`: [100, 1000]

- **Multi-layer Perceptron (Neural Network-NN)**
  - `hidden_layer_sizes`: [(50,), (100,), (50,50), (100,50)]
  - `activation`: ['tanh', 'relu']
  - `alpha`: [0.0001, 0.01]
  - `learning_rate`: ['constant', 'adaptive']
  - `max_iter`: [200, 500]

- **K-Nearest Neighbors(K-NN)**
  - `n_neighbors`: [3, 15]
  - `weights`: ['uniform', 'distance']
  - `p`: [1, 2] (Manhattan or Euclidean distance)

Training Results:

Model	Accuracy	Precision	Recall
GB	0.6827	0.6652	0.7664
LR	0.6693	0.7959	0.4769
RF	0.6763	0.6809	0.6929
NN	0.5410	0.8150	0.1347
K-NN	0.4993	0.5121	0.4834

Table 4: Model Comparison on Accuracy, Precision, and Recall

6.3. 3. Evaluation Metrics

The model is evaluated using ROC-AUC and Precision-Recall metrics. AUC scores and plots indicate strong classification performance.

Model	F1 Score	AUC
Gradient Boosting	0.7122	0.7935
Logistic Regression	0.5964	0.7920
Random Forest	0.6869	0.7805
Neural Network	0.2312	0.5657
K-Nearest Neighbors	0.4973	0.5087

Table 5: Model Comparison on F1 Score and AUC

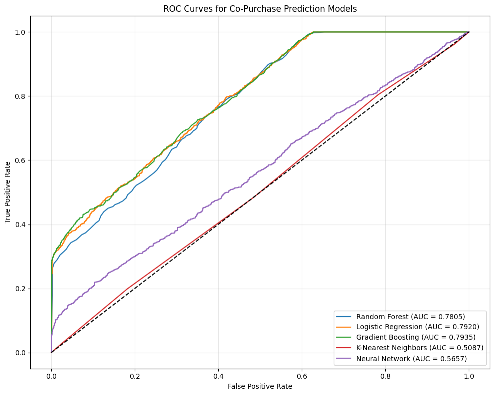


Figure 8: ROC curves of all the models

BEST MODEL:
The best model for the co-purchase prediction task is **Gradient Boosting**, with an AUC score of **0.7935**. Hyperparameter tuning was performed using cross-validation to find the optimal configuration for the model. After testing multiple candidate sets, the best hyperparameters were found to be:

- **Learning rate**: 0.0322
- **Max depth**: 8
- **Min samples leaf**: 9
- **Min samples split**: 8
- **Number of estimators**: 250
- **Subsample**: 0.8493

The best cross-validation score achieved with these parameters was **0.8062**, confirming that the tuned Gradient Boosting model significantly improved performance.

6.4. 4. Correlation Matrices for the models

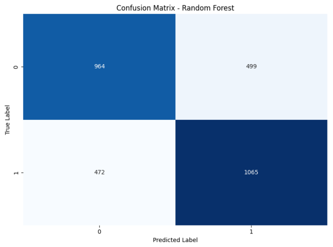


Figure 9: RANDOM FOREST

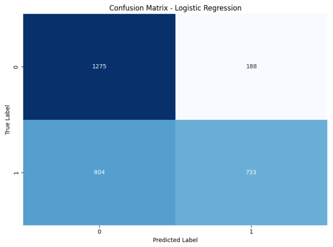


Figure 10: LOGISTIC REGRESSION

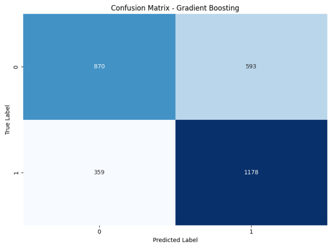


Figure 11: GRADIENT BOOSTING

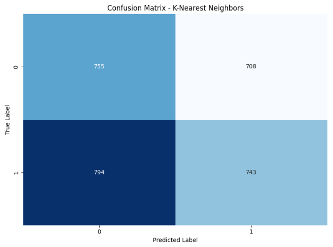


Figure 12: K-NEAREST NEIGHBOURS

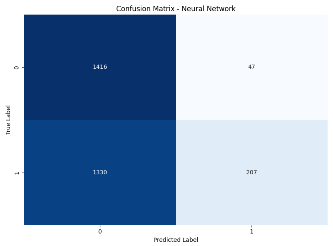


Figure 13: NEURAL NETWORK

6.5. 5. Co-Purchase Prediction

The `predict_co_purchase` function evaluates the likelihood that two books will be co-purchased in the future. The key steps involved in this prediction are:

1. **Feature Extraction:** Graph-based features are computed between the two specified book IDs.
2. **Model Inference:** The feature vector is passed to a pre-trained machine learning model, which has been trained on known co-purchase pairs.
3. **Probability Score:** The model returns a probability value between 0 and 1 indicating the likelihood of a future co-purchase between the two books.

This function allows users to compare any two books and assess their potential association based on historical data and network structure.

Feature	**RF**	**LR**	**GB**
Preferential Attachment	0.7001	0.610	0.150
Adamic-Adar Index	0.0876	0.120	0.150
Jaccard Coefficient	0.0665	0.080	0.150
Common Neighbors	0.0644	0.070	0.150
Salesrank Difference	0.0410	0.070	0.150
Resource Allocation Index	0.0338	0.040	0.150
Rating Product	0.0066	0.010	0.100

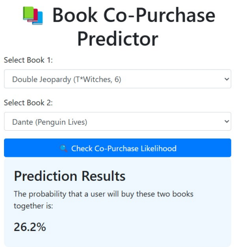
Table 6: Feature Importance Comparison Across Models

6.6. 5. Co-Purchase Recommendation

The `recommend_future_books` function is designed to generate personalized book recommendations using network-based link prediction. The process can be summarized as follows:

1. **Input:** A selected book for which recommendations are desired.
2. **Exclusion of Existing Neighbors:** All books that are already co-purchased with the selected book (i.e., direct neighbors in the co-purchase graph) are excluded from consideration.
3. **Candidate Sampling:** A set of candidate books is randomly sampled from the remaining nodes in the network that are not yet linked to the selected book.
4. **Feature Computation:** For each candidate book, a feature vector is computed representing the relationship between the selected book and the candidate.
5. **Probability Prediction:** A trained machine learning model predicts the probability that a co-purchase link will form between the selected book and each candidate.
6. **Top-K Recommendation:** The candidates are sorted by their predicted probabilities, and the top 5 books are returned as the most likely future co-purchases.

6.7. 5. GUI working of Prediction and Recommendation Tasks



Book Co-Purchase Predictor

Select Book 1: Double Jeopardy (T*Witches, 6)

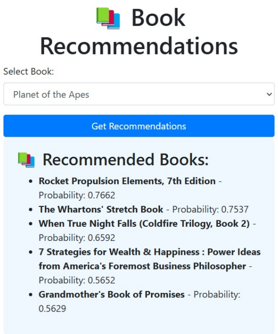
Select Book 2: Dante (Penguin Lives)

Prediction Results

The probability that a user will buy these two books together is:

26.2%

Figure 14: Copurchase Prediction



Book Recommendations

Select Book: Planet of the Apes

Recommended Books:

- **Rocket Propulsion Elements, 7th Edition** - Probability: 0.7662
- **The Whartons' Stretch Book** - Probability: 0.7537
- **When True Night Falls (Coldfire Trilogy, Book 2)** - Probability: 0.6592
- **7 Strategies for Wealth & Happiness : Power Ideas from America's Foremost Business Philosopher** - Probability: 0.5652
- **Grandmother's Book of Promises** - Probability: 0.5629

Figure 15: Book Recommendation

7. Conclusion

The Amazon co-purchase network shows key characteristics of real-world networks, including a heavy-tailed degree distribution, strong community structures, and small-world behavior. These features indicate that the network is efficient, with products being closely connected and relatively short paths between them. While the network has some scale-free properties, it doesn't fully follow a power-law distribution, possibly due to the diversity of product types.

The predictive models, especially Gradient Boosting, effectively forecast co-purchases, with important features like common neighbors and preferential attachment. These insights could improve recommendation systems for more accurate product suggestions. Although the analysis focused on books, the same approach can be applied to any product category, demonstrating the flexibility and significance of this work in broader contexts.

This analysis highlights the value of network science in understanding product relationships and provides a foundation for future work, such as incorporating more data and refining the models to improve predictions over time.

References

- Barabási, A. L. (2016). *Network Science.* Cambridge University Press.
- Newman, M. E. J. (2010). *Networks: An Introduction.* Oxford University Press.
- Erdős, P., & Rényi, A. (1959). On random graphs. *Publicationes Mathematicae.*