Exploring Consumer Behavior and Market Trends in Amazon Products With Network Analyses

Guillermo Sanchez Lamas (membriux), Xinqi Guo(xinqiguo),
Yifei Sun (yifeisun), Alexander He (alexhe)

Abstract

This study aims to explore data from Amazon's e-commerce platform, analyzing complex consumer behavior and market trends. The project uses network analysis techniques to focus on Amazon's extensive product review data collected from 29 categories. Key research questions include "What do co-purchase and viewing patterns reveal about consumer behavior?" and "How can network analyses of such patterns inform our understanding of product relationships?" Various Python libraries were used for data processing and machine learning, emphasizing data cleaning, and employing the k-core algorithm for data reduction. While network characteristics and consumer behavior varied across different categories, a significant influence of certain interconnected products was observed. However, the study findings are exploratory and limited due to the imbalanced nature of the data and its focus on specific categories. The project suggests a need for more extensive, balanced datasets for future research in e-commerce.

Introduction

In our project, we set out to explore data from Amazon's e-commerce platform, seeking to learn about complex consumer behavior and diverse market trends. As students exploring the vast landscape of data analysis, we focus on employing network analysis techniques to delve into Amazon’s extensive product review data. Our aim is to glean insights into consumer purchasing patterns and understand how products are interconnected within this online retail giant’s ecosystem.

This project is an academic endeavor to understand the practical application of network analysis in e-commerce. We are motivated by the desire to see how consumer behavior can be deciphered from the relationships between products, particularly what customers 'also bought' or 'also viewed'. This exploration is grounded in the Amazon Review Data (2018), an extensive dataset that offers a rich field for applying data science techniques.

Our approach involves a careful examination of product metadata combined from 29 categories. We seek to understand how products relate to each other in terms of consumer interest and preference. The project is guided by questions such as "What do co-purchase and viewing patterns reveal about consumer behavior?" and "How can network analyses of such patterns inform our understanding of product relationships?"

While we do not aim to revolutionize the field of network analysis or e-commerce studies, our project brings forth a fresh perspective by merging data from various Amazon product categories to conduct a comprehensive analysis. This student-led endeavor focuses on applying network analysis techniques and a machine learning model to decipher complex patterns in consumer behavior and product relationships.

Our report outlines the process of data cleaning, the application of the k-core algorithm for data reduction, data analyses, and a machine learning model. We hope that our findings, while exploratory in nature, contribute to a foundational understanding of e-commerce dynamics and practical applications of network analysis in the digital marketplace.

Dataset Description

Our project utilized the Amazon Review Data (2018) dataset, curated by Jianmo Ni from UCSD.[1] This comprehensive dataset, an update from the 2014 release, encompasses a broad spectrum of Amazon product reviews, metadata, and transactional information. Spanning from May 1996 to October 2018, it includes a staggering 233.1 million reviews across various product categories.

The dataset is particularly rich in its inclusion of diverse product metadata, offering insights into product features, images, brand information, and co-purchasing patterns. Additionally, it expands its scope by encompassing five new product categories, thus providing a more extensive and varied analytical playground.

For our analysis, we focused on the metadata component of the dataset, which contains detailed information on 15.5 million products. This metadata includes but is not limited to, product descriptions, pricing information, image URLs, brand details, and categorical classifications. Notably, each product entry features unique identifiers (ASIN), titles, descriptive features, and lists of products that users “also bought” and “also viewed,” thus forming a network structure.

Data Processing

Initial Data Loading and Merging

The initial step in our data processing journey involved loading and merging data from different sources. We utilized Python and its powerful libraries, such as Pandas, for efficient data handling. Each category's data, stored in CSV format, was read into a DataFrame, focusing only on the 'asin' and 'reviewerID' columns to minimize memory usage. These partial DataFrames were then concatenated into a single DataFrame using `pd.concat`, providing us with a unified view of our dataset and its underlying network.

K-Core Filtering

The processing of this vast dataset presented challenges. Our objective was to refine and reduce the data to a manageable size for cross-category analysis while maintaining its integrity and richness. To achieve this, we combined review data from all categories into a single DataFrame and applied a k-core filter with a size of 32. In other words, we ensured that each product and user in our final dataset had at least 32 reviews. The unfiltered dataset is ~9 GB, while our 32-core dataset is ~300 MB, which is a much more manageable size, especially when using the NetworkX library. However, please note that some of our analyses did not use the k-core; instead, they used the full data of individual categories.

The k-core algorithm is a crucial aspect of our data processing strategy, particularly for its application in refining and ensuring the quality of our large-scale dataset. Fundamentally, a k-core of a graph is a maximal subgraph in which each vertex (node) has at least k connections (edges). In the context of our project, this

[1] https://jmcauley.ucsd.edu/data/amazon_v2/index.html

translates to retaining only those products and users in the dataset that have at least k reviews. The necessity for an iterative approach stems from the interdependent nature of products and users. Initially, when we filter out products with fewer than k reviews, this may affect the review count of certain users, potentially dropping them below the k threshold. Similarly, removing these users might affect the review counts of other products. Thus, the dataset needs to be repeatedly filtered until it stabilizes – a state where all remaining products and users have k or more reviews. This iterative process ensures the robustness and reliability of our dataset, as it systematically weeds out sparsely connected nodes, resulting in a denser and more interconnected network of reviews. The convergence of this process is critical; it signifies that we have achieved a consistent and dependable subset of our data, making it particularly suitable for subsequent analyses that rely on the interconnectedness of users and products.

Data Cleaning and Transformation

Along with k-core filtering, data cleaning was an essential part of our process, given the varied nature of the data. We implemented specific checks and transformations to maintain data quality:

- **Title Format Check:** We removed entries with unformatted titles, identified by the presence of strings like "getTime", ensuring that our dataset only contained well-structured data.
- **Price Processing:** The 'price' field underwent thorough processing. We handled different formats, including ranges (e.g., "$13.79 - $36.67") and empty strings, translating them into a standardized numeric format. For price ranges, we calculated and stored the midpoint price along with the minimum and maximum prices for a more granular analysis. Products with no listed price were appropriately handled by assigning a value of $0.00.
- **Field Exclusion:** Unnecessary fields such as 'imageURL' and 'description' were excluded since we're focusing on the most impactful data for our analysis.

Our refined dataset aligns with the scope of our course project and sets a strong foundation for comprehensive analytical and modeling tasks.

Category-Wise Processing

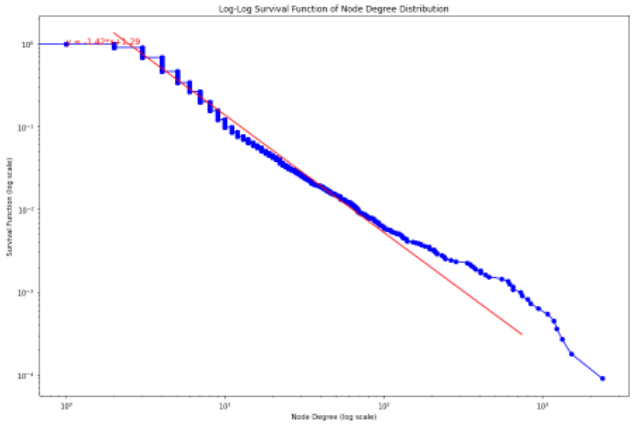
After constructing the 32-core using only the product and user pairs, we still needed to extract the full data. Our processing script was designed to handle each category independently while maintaining the correct 32-core. We downloaded the full data file for each category, filtered out the products that don't exist in the 32-core, and then conducted final data cleaning and transformation.

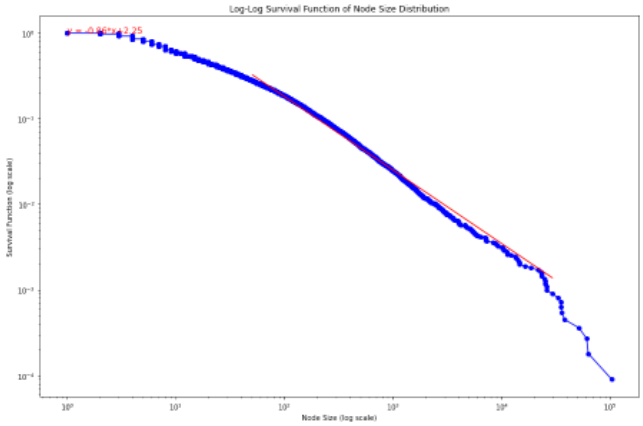
Category Network Analysis

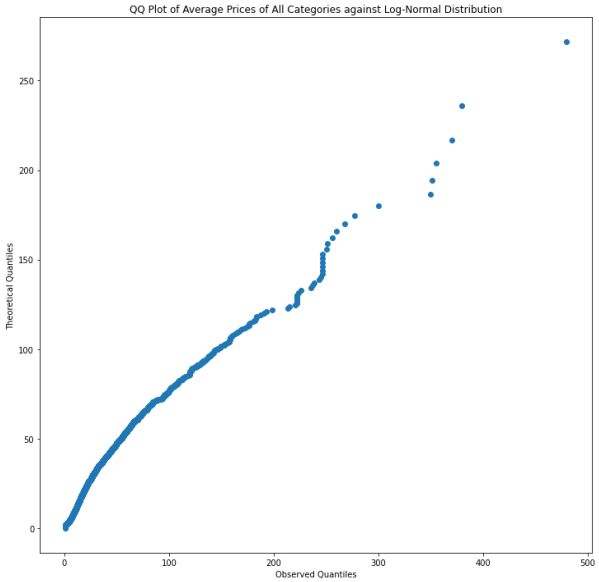
We constructed a network of category tags based on the co-appearance of two category tags in a product. The node size represents the number of products with this category tag. The weight of an edge represents the number of products that have both end nodes of this edge as category tags. We also calculated the average price of all products in a node and the average price in an edge. We finally constructed a network of 11082 nodes and 48307 edges by filtering out errors.

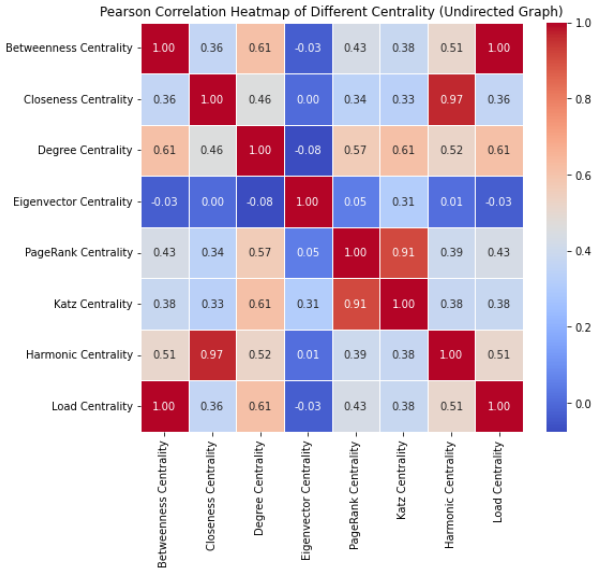
A scale-free network is a network whose degree distribution follows a power law[2]. The degree distribution of this network clearly shows the power law tail. The alpha value of this power law tail is around 2.42, which is within the typical range for power law distributions, indicating that this is indeed a scale-free network. These node sizes also show a power law distribution in the tail, with an alpha of around 1.86. The power law indicates there are more likely to be extreme values in both the node sizes and node degrees.

However, the distribution of average prices does not follow the power law[3]. It is an asymmetrical unimodal distribution, similar to a log-normal distribution.









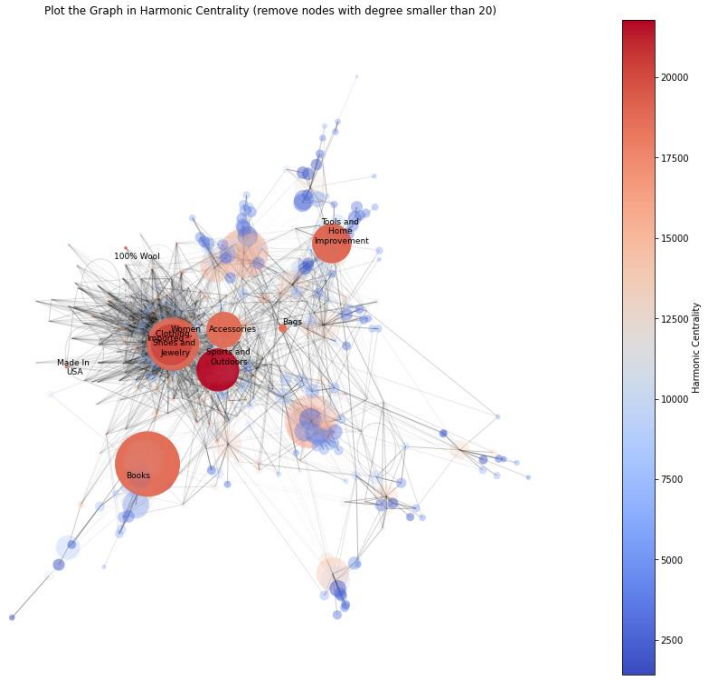
This network has a degree assortativity coefficient of -0.20124, which means dissimilar degrees are more likely to be connected, and there are a lot of high-degree and low-degree pairs in the network. Meanwhile, the assortativity coefficient for average price is near 0, indicating there is no correlation between the average price of pairs of two categories.

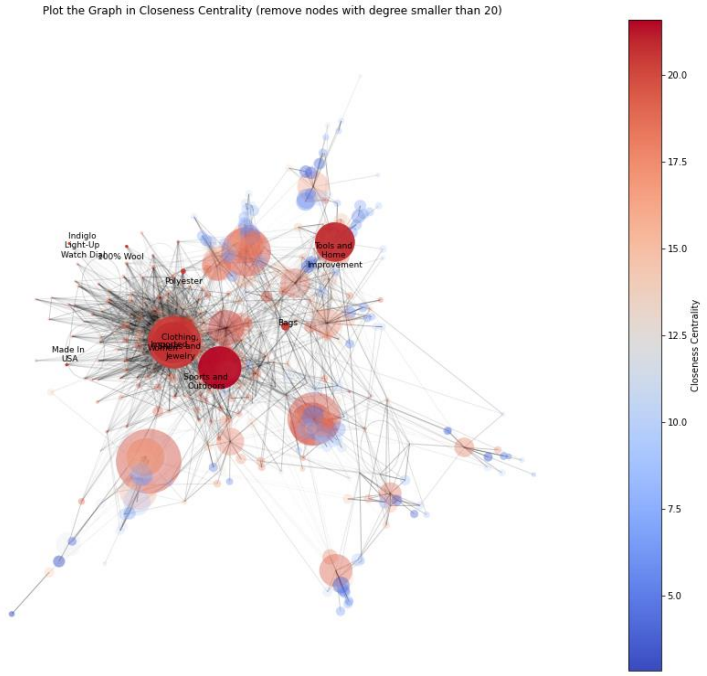
Then, we want to identify some of the most important categories in the network. Due to the limitation of computation, we first deleted points with smaller degrees. The weight of an edge in this undirected graph is calculated as the Jaccard similarity of the two end nodes. The distance between two nodes is defined as the square root of the inverse of this weight. After computing nine centrality measurements, we found they give

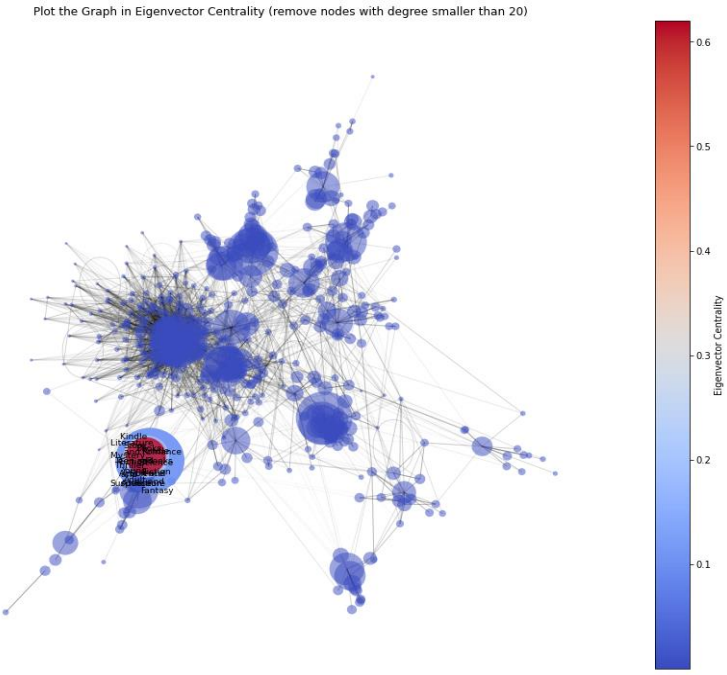
[2] https://en.wikipedia.org/wiki/Scale-free_network

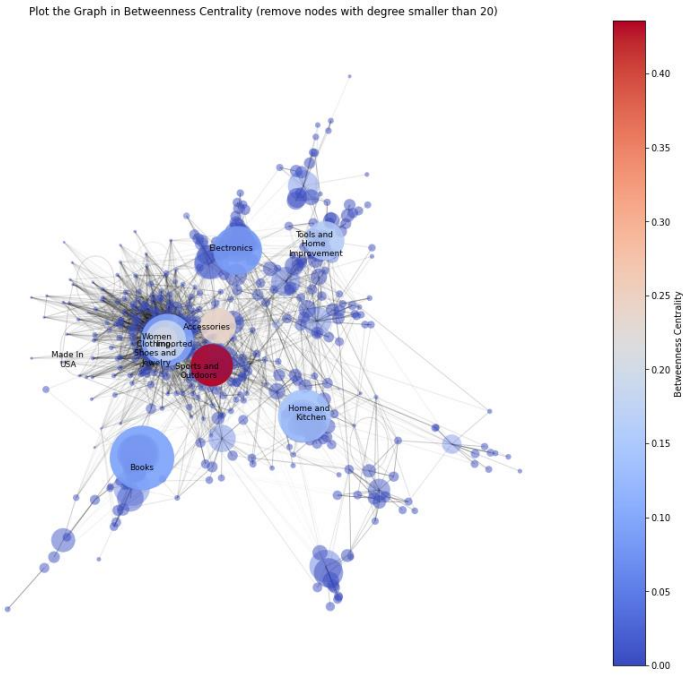
[3] https://www.google.com.hk/url?sa=t&rct=j&q=&esrc=s&source=web&cd=&ved=2ahUKEwix5vy_4ZCDAxUfIlkEHdEYASMQFnoECBIQAw&url=https%3A%2F%2Fanalystprep.com%2Fcfa-level-1-exam%2Fquantitative-methods%2Flognormal-distribution%2F%23%3A~%3Atext%3DWhen%2520the%2520returns%2520on%2520a%2Cappropriate%2520model%2520for%2520stock%2520prices.&usg=AOvVaw2F6jbybxCMk-iPseG5fBX9&opi=89978449

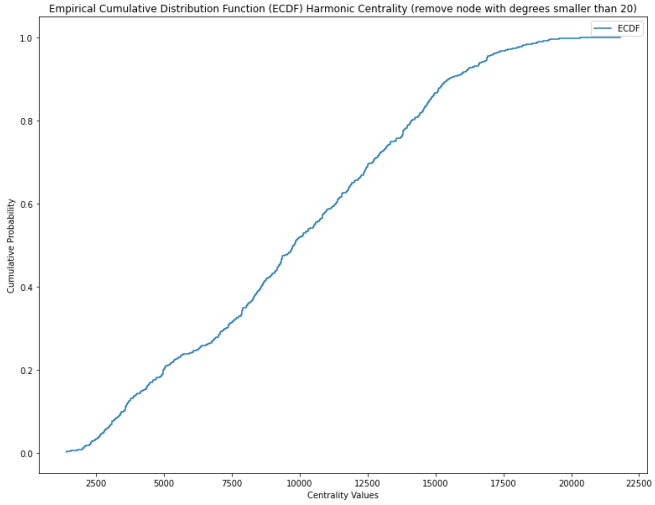
very different results. The results for eigenvector centrality especially do not seem related to all other centralities. Load centrality, eigenvector centrality, and betweenness centrality provide very skewed centrality distributions, with most of the values being near 0 and a few large values. In contrast, closeness and harmonic centrality give more even distributions. By calculating the average of the scaled centrality scores, we found the 10 most important categories. These are shown in the plot below.

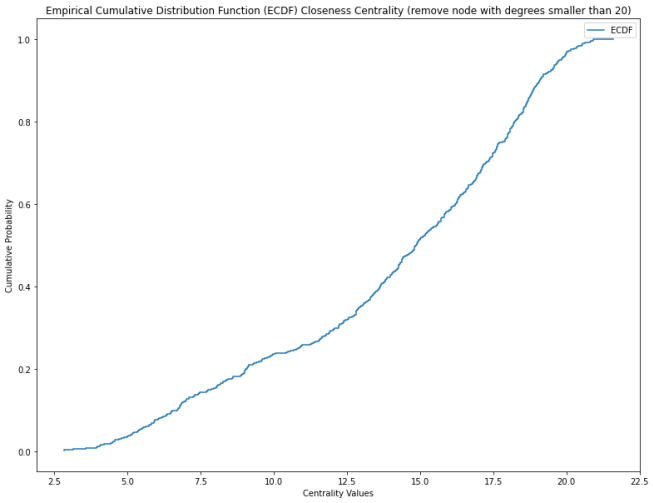


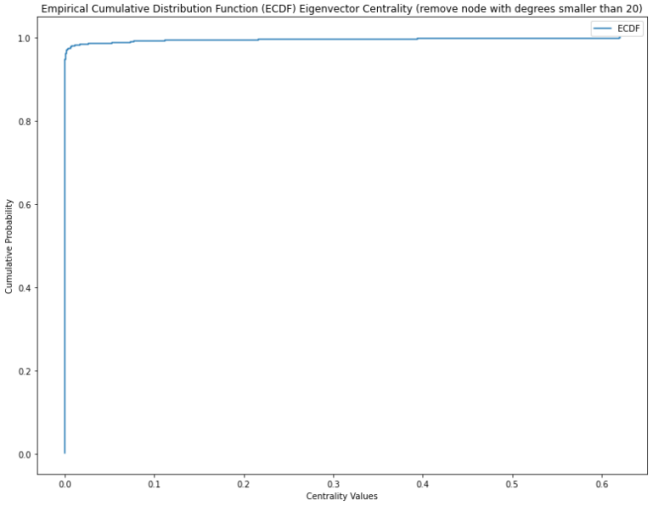


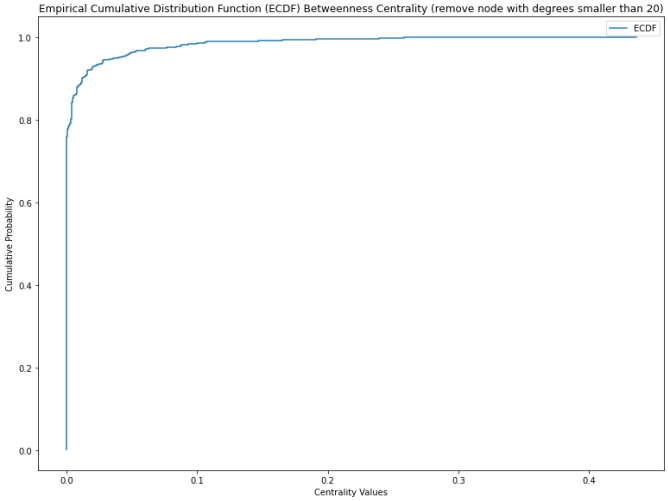


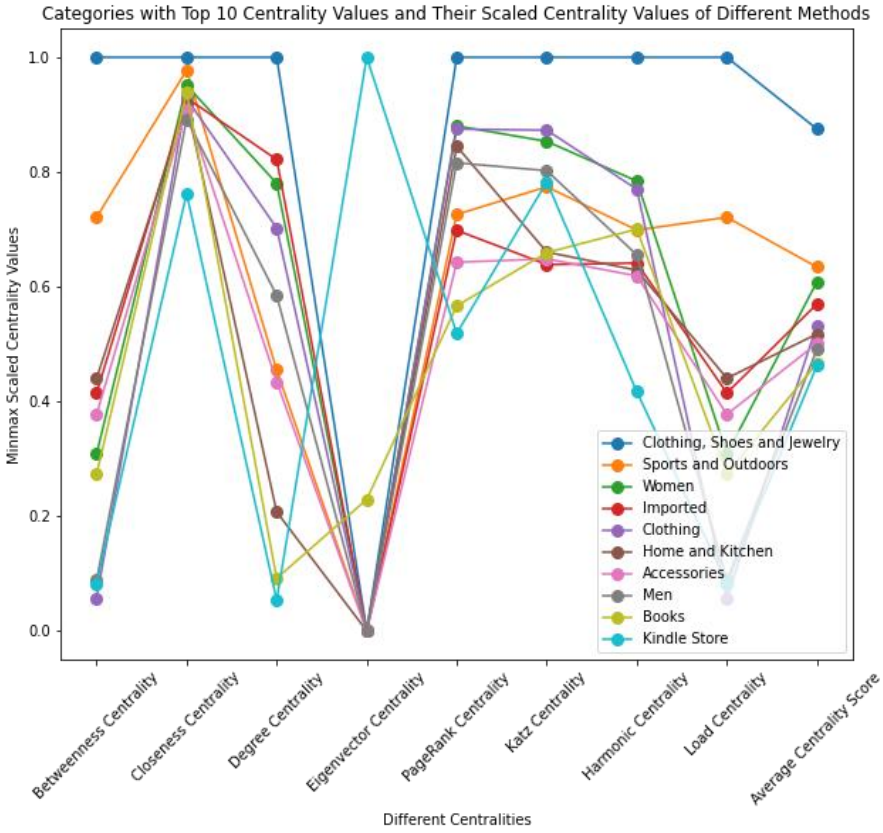



Empirical Cumulative Distribution Function (ECDF) Harmonic Centrality (remove node with degrees smaller than 20)
ECDF
Cumulative Probability
Centrality Values
0.0
0.2
0.4
0.6
0.8
1.0
2500
5000
7500
10000
12500
15000
17500
20000
22500

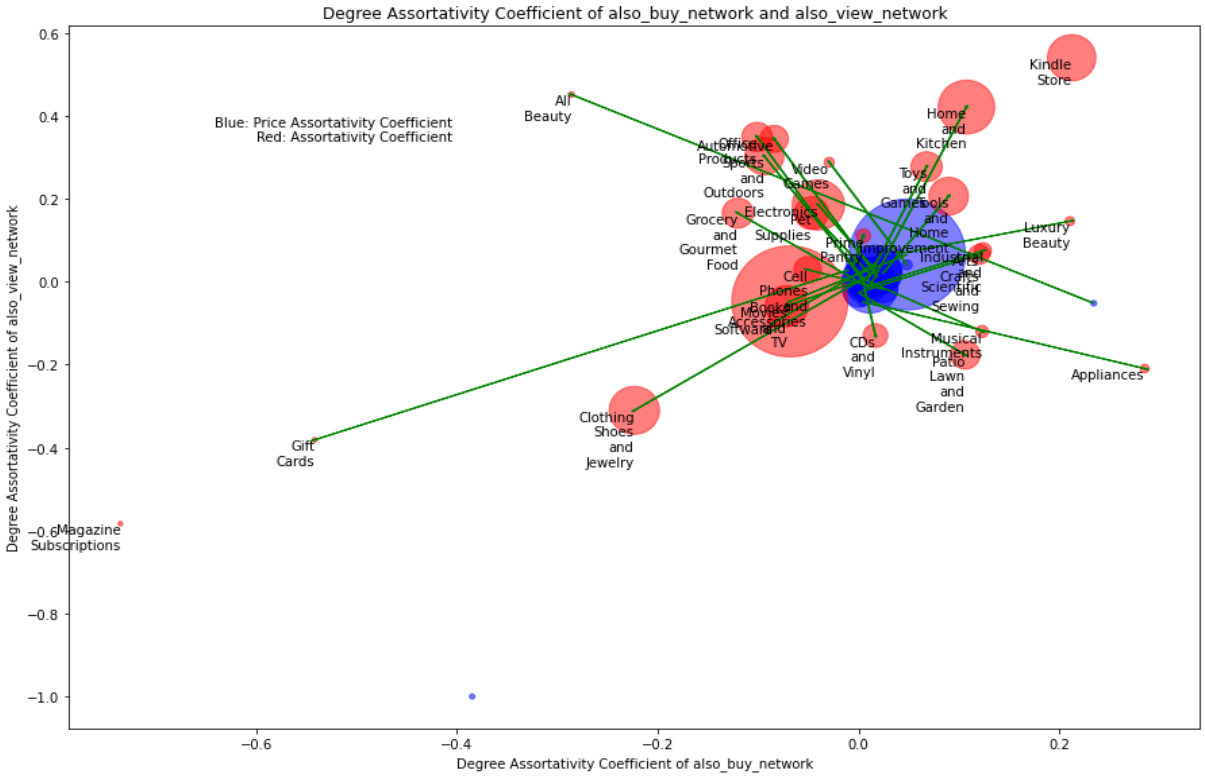

Empirical Cumulative Distribution Function (ECDF) Closeness Centrality (remove node with degrees smaller than 20)
ECDF
Cumulative Probability
Centrality Values
0.0
0.2
0.4
0.6
0.8
1.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
20.0
22.5


Empirical Cumulative Distribution Function (ECDF) Eigenvector Centrality (remove node with degrees smaller than 20)
ECDF
Cumulative Probability
Centrality Values
0.0
0.2
0.4
0.6
0.8
1.0
0.0
0.1
0.2
0.3
0.4
0.5
0.6

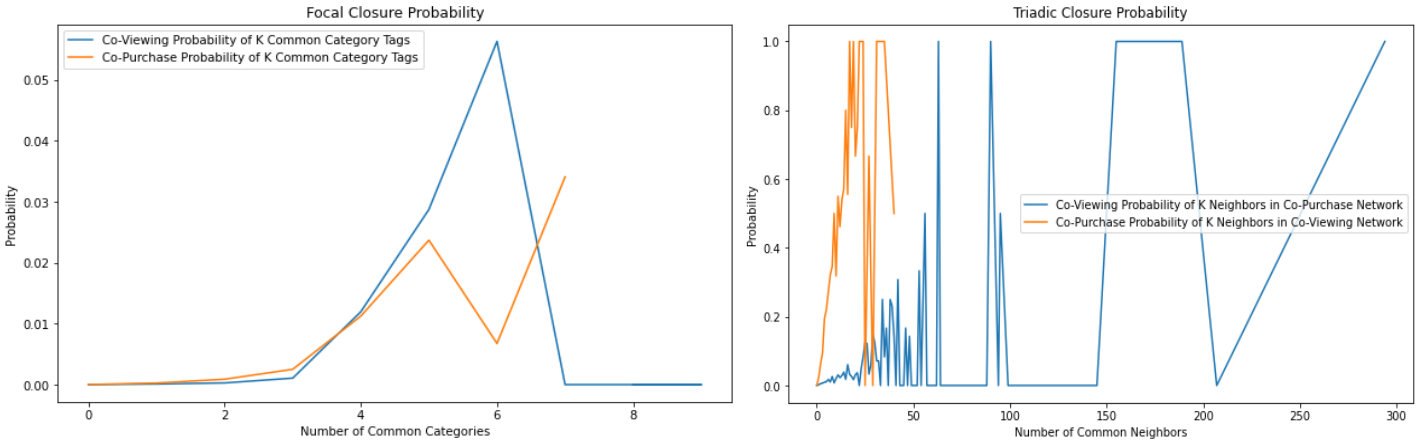

Empirical Cumulative Distribution Function (ECDF) Betweenness Centrality (remove node with degrees smaller than 20)
ECDF
Cumulative Probability
Centrality Values
0.0
0.2
0.4
0.6
0.8
1.0
0.0
0.1
0.2
0.3
0.4



For every category, we constructed two networks, also_buy, and also_view, using co-purchase and co-viewing information of products. The calculated degree assortativity coefficients and price assortativity coefficients show that though there is a large variation in the degree similarity of node pairs, the prices of co-purchased and co-viewed products are similar.



By constructing a bipartite graph of both products and categories, we can calculate the empirical probability of two products being co-purchased or co-viewed when these two products have k category tags in common. We can see that when the number of common category tags increases for pairs of two products, the co-purchase probability and co-viewing probability also increase. By constructing two networks with edges representing co-purchase and co-viewing, we could calculate the empirical probability of co-viewing probability of k neighbors in co-purchase network and the empirical probability of co-purchase probability of k neighbors in co-viewing network. The trend for both line is generally increasing. Because of the scale of the network (~200,000 product nodes and ~ 10,000 category nodes), sampling (5000 nodes out of 200,000 nodes) is used in the calculation of these values.



Inter-Category Co-Purchase Network

To characterize consumer purchasing patterns, we developed an Inter-Category Co-Purchase Network using product category tags. This network connected categories based on shared product associations, with edges representing co-purchase and co-viewing relationships between categories. We calculated transition probabilities as the likelihood of purchasing another product from the same or a different category, given that the product was viewed. Since probabilities are calculated at the category level, this offers insights into the dynamics of consumer choice and the interplay between different product categories on Amazon.

To identify co-purchase patterns and group categories into meaningful clusters, we tested Spectral Clustering[4], Louvain Community Detection[5], and Greedy Modularity Maximization[6]. We used the Silhouette Score[7] and Davies-Bouldin Score[8] metrics to assess clustering quality. The Silhouette Score, which ranges from -1 to 1, evaluates how well an object fits within its cluster compared to other clusters. A higher score suggests better-defined clustering. The Davies-Bouldin Score measures the average similarity between clusters, considering both the distance between clusters and their individual sizes. Here, a lower score indicates more distinct, well-separated clusters. These metrics provided a comprehensive view of clustering performance, with Silhouette Score focusing on intra-cluster cohesion and Davies-Bouldin Score on inter-cluster separation.

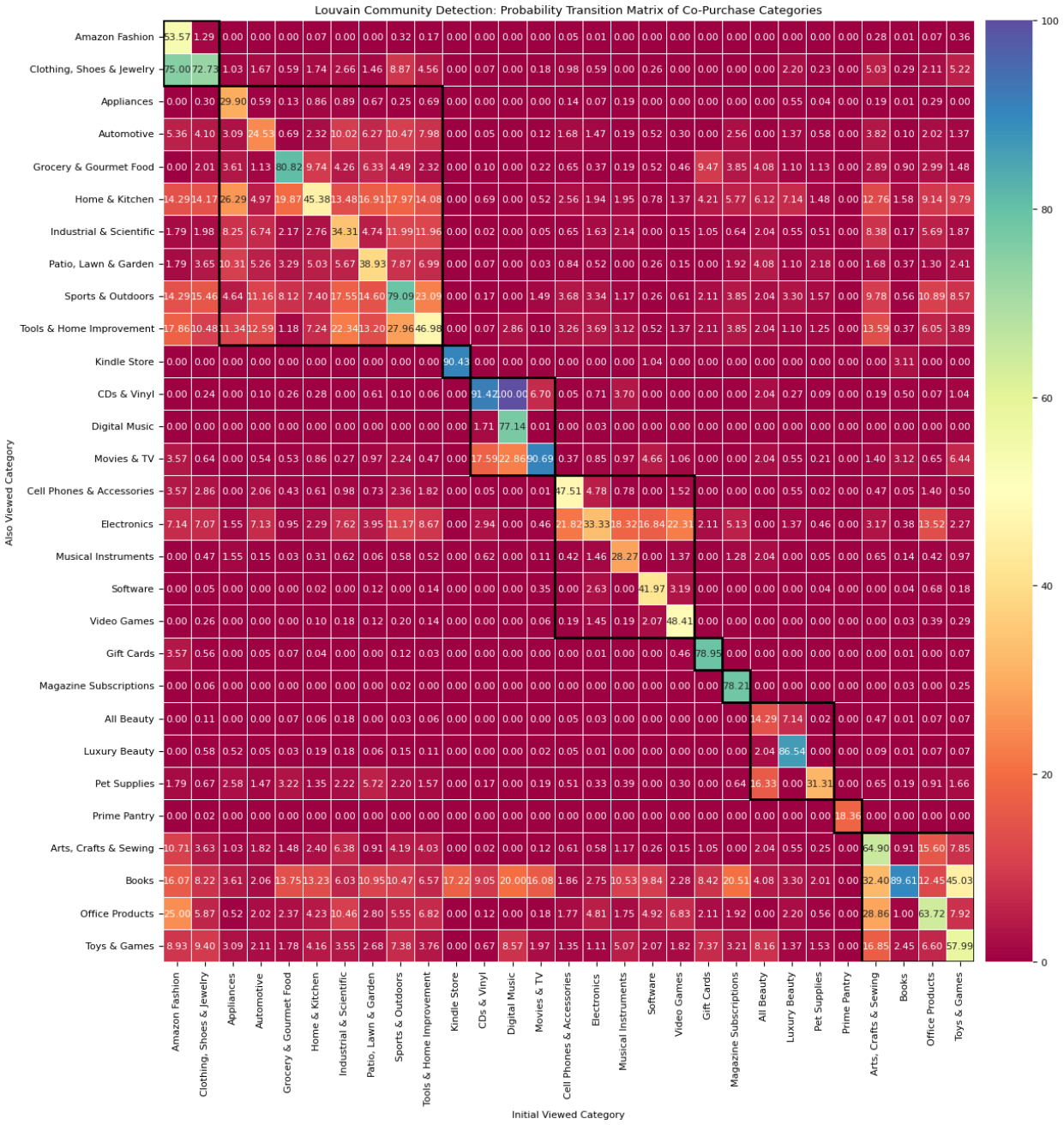
Even with tuning across multiple specified numbers of clusters and/or random seeds, Spectral Clustering did not perform well, likely because the algorithm assumes a symmetric matrix. Greedy Modularity Maximization

[4] sklearn.cluster.SpectralClustering — scikit-learn 1.3.2 documentation
[5] louvain_communities — NetworkX 3.2.1 documentation
[6] greedy_modularity_communities — NetworkX 3.2.1 documentation
[7] sklearn.metrics.silhouette_score — scikit-learn 1.3.2 documentation
[8] sklearn.metrics.davies_bouldin_score — scikit-learn 1.3.2 documentation

gave the lowest scores. Louvain Community Detection performed the best, achieving an ideal balance between the two metrics. Unlike the previous two algorithms, this one finds the best partition, so no tuning was necessary.

	Best number of clusters	Silhouette Score	Davies-Bouldin Score
Louvain Community Detection	10	-0.107457	1.233066
Spectral Clustering	11	-0.118110	1.300155
Greedy Modularity Maximization	5	-0.073076	2.914430

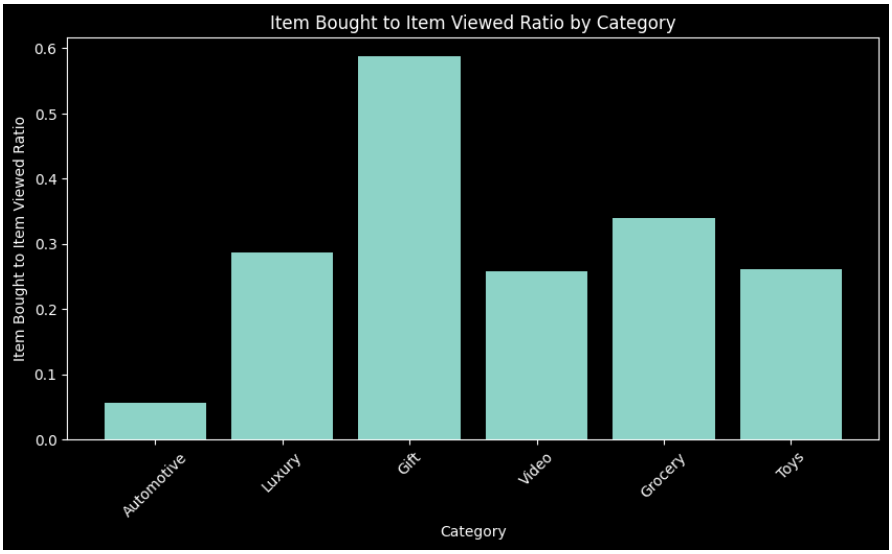
To visualize clustering results, we generated the directed transition matrix of the network with clusters grouped together. The columns represent the category of the initially viewed product, the rows indicate the category of products that users also viewed, and the values indicate the probability that these products were also bought. Thus, this illustrates the likelihood of co-purchasing items in the same or different categories.



Price Market Trends Analysis

Item Bought to Item Viewed Ratio by Category

To answer the main question “Do people buy what they see?”, we first analyze the relationship between the items that people buy (‘also_buy’) and see (‘also_view’). We started by analyzing the Item Bought to Item Viewed Ratio for each of the categories

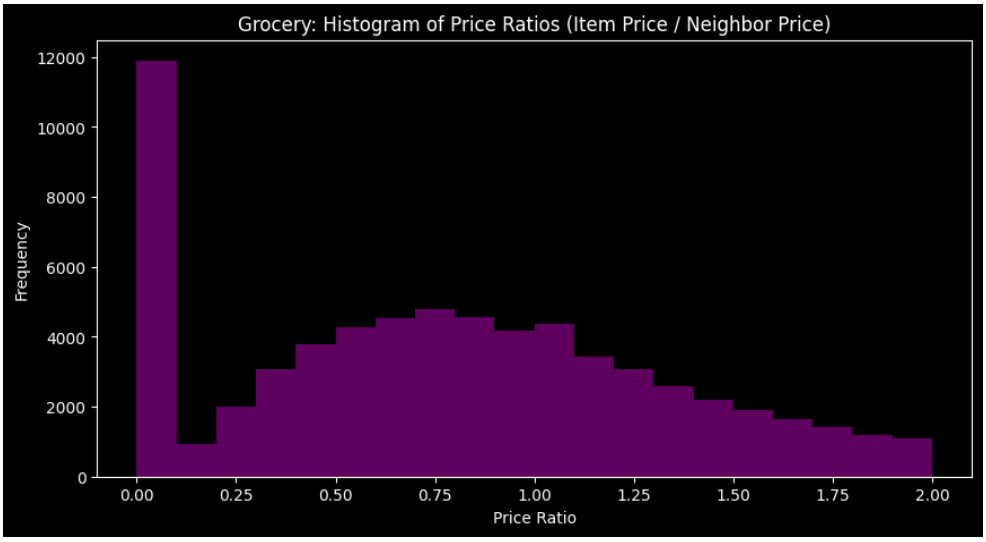


Here is what we found in this chart:

1. **Gifts** have the highest ratio of items bought to items viewed, nearly reaching 0.6. This suggests that customers are more likely to purchase a gift item after viewing it compared to other categories. This could be due to various reasons such as the urgency of the purchase, lower price points, or the emotional value associated with gifts.
2. The **Automotive** category has the lowest ratio, indicating that customers are less likely to buy automotive items after viewing them. This could be due to higher price points, the need for more research before making a purchase, or the infrequency of such purchases.
3. Other categories like **Luxury**, **Video Games**, **Grocery**, and **Toys** have varying ratios, suggesting different levels of customer purchase intent after viewing items in these categories.

Histogram of Price Ratios (Item Price / Neighbor Price)

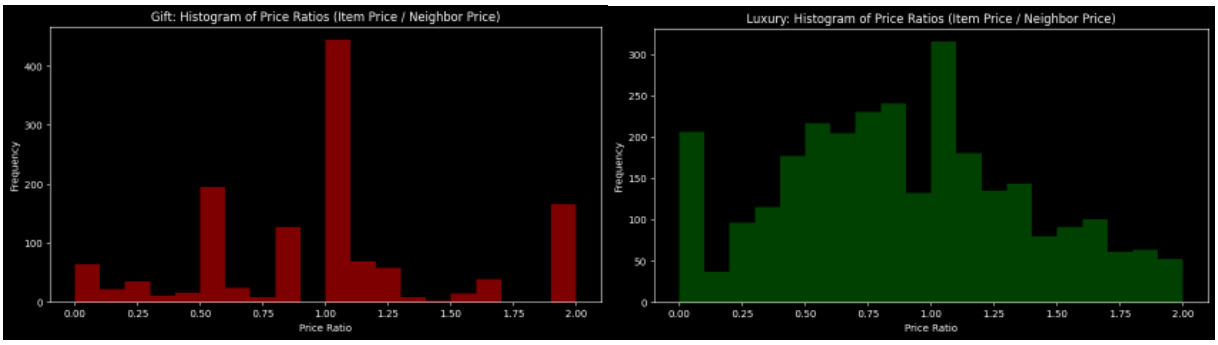
Next, we analyzed the histograms of price ratios between the item price and the neighbor price. Below is a sample histogram from the Groceries category, which we will use as a representative chart of these categories (since they are very similar): video games, automotive, and toys.



This is a histogram showing the distribution of price ratios (item price / neighbor price) for grocery items (but are very similar to video games, automotive, and toys categories as well, so we will use this histogram representative of the previously mentioned categories). The x-axis represents the “Price Ratio” ranging from 0 to 2, and the y-axis represents the “Frequency”.

From the histogram, it’s evident that a significant number of items have a price ratio less than 0.25. This indicates that many grocery items are priced much lower than their neighboring items. As the price ratio increases, there are fewer items in each category, creating a descending pattern in the histogram. This suggests that items with a higher price compared to their neighboring items are less common.

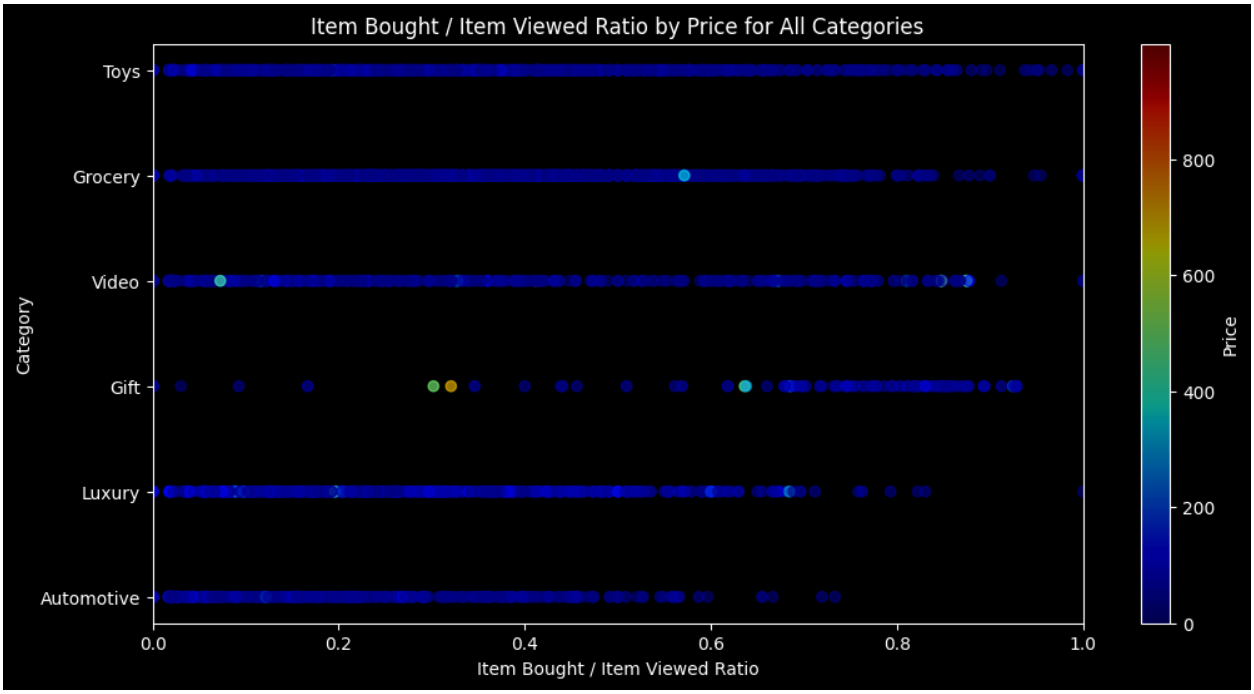
This analysis can be useful for understanding pricing strategies and consumer behavior in the grocery sector. For instance, items with a low price ratio might be more likely to be purchased due to their relative affordability. However, this would need to be confirmed with additional data, such as sales figures.



From the Gift and Luxury category histogram, it’s evident that there is a prominent peak at a 1.25 price ratio with a frequency reaching almost up to 400. This indicates that a significant number of gift items are priced 25% higher than their neighboring items.

This analysis can be useful for understanding pricing strategies and consumer behavior in the gift sector. For instance, items with a higher price ratio might be perceived as more valuable or desirable gifts. This can be confirmed with the Bought to Item Viewed Ratio chart mentioned previously.

Items Bought / Item Viewed Ratio by Price



Here's a short description of the histogram

- The y-axis lists six different categories: Toys, Grocery, Video, Gift, Luxury, and Automotive. These categories likely represent the types of items being analyzed.
- The x-axis represents the item bought/item viewed ratio, which ranges from 0 to 1. This ratio provides insight into consumer behavior, indicating how often viewed items are actually purchased.
- Each dot on the graph corresponds to a specific ratio and price in its category. The position of the dot indicates the specific ratio for that category.
- The color of each dot represents the price of the item, according to the color scale provided on the right. The scale ranges from blue (low price) through green and yellow to red (high price), indicating prices from $0 to over $800.

Based on the scatter plot, there are some outlier data points that deviate significantly from the other observations in their category. Let's analyze the outliers in the Grocery, Video, and Gift categories:

1. **Grocery**: The outliers in this category seem to be the dots that have a high item bought/item viewed ratio (close to 1) but are colored blue or green, indicating a low to medium price. This suggests that there are certain low to medium-priced grocery items that are bought almost every time they are viewed.
2. **Video Games**: The outliers in the Video Games category are the dots with a high item bought/item viewed ratio (close to 1) and are colored red, indicating a high price. This suggests that there are certain high-priced video game items that are frequently bought when viewed.
3. **Gift**: In the Gift category, the outliers appear to be the dots with a low item bought/item viewed ratio (close to 0) but are colored red, indicating a high price. This suggests that high-priced gift items are often viewed but not bought as frequently.

Price Difference Prediction Model

Data Preprocessing

In this section, we built a model to predict if the items in also_view will be bought based on price differences and node features, using the full data of a couple individual categories (not the 32-core). We first mapped each item in also_buy and also_view to their corresponding items, which is the key variable of the data. Then, we mapped each also_buy and also_view to labels 0 and 1 to indicate if the item was brought or not. Next, we removed all the intersections between buy and view from view data and turned the view data as didn't-buy. Therefore, we built a directed graph where the items are nodes and the edges are the pairs of the connected items. We also added the price difference sign as the sign for the graph.
Finally, we added some features to the data. The features include price_difference, price_sign, triads_count, total_degree, in_degree_positive, in_degree_negative, out_degree_positive, and out_degree_negative. Below is the description of the data:

'Asin': the id of the item
'Associated_asin': the id of also_buy or also_view
'Price_difference': the price difference between the items
'Label': buy:1, didn't buy: 0
'Price_sign': + or -
'Triads_count': the count of triads
'Total_degree': total degree of the node
'In_degree_positive': in degree with sign +
'In_degree_negative': in degree with sign -
'Out_degree_positive': out-degree with sign +
'Out_degree_negative': out-degree with sign -

The table below shows the sample training data.

	price_difference	price_sign_encoded	triads_count	total_degree	in_degree_positive	in_degree_negative	out_degree_positive	out_degree_negative
1582	-0.36	0	1	8	2	4	0	2
775	-18.44	0	10	4	0	0	0	4
1072	26.33	1	84	9	0	1	7	1
661	-18.00	0	4	4	1	0	0	3
373	0.00	0	56	8	1	0	2	5

Model Training

In the training, we decided to use logistic regression to predict the potential purchased items from the also_view. We set the train iterations as 10 to ensure the results converged. First, we fitted the video_game data to the model, and below is the F1 score by iterations:

The video game dataframes have 1800 records and the F1 score finally converges around 0.76. We are glad to see the convergence here even though the F1 score is not quite good. We can conclude that the F1 score could be high if we have more data for the model.

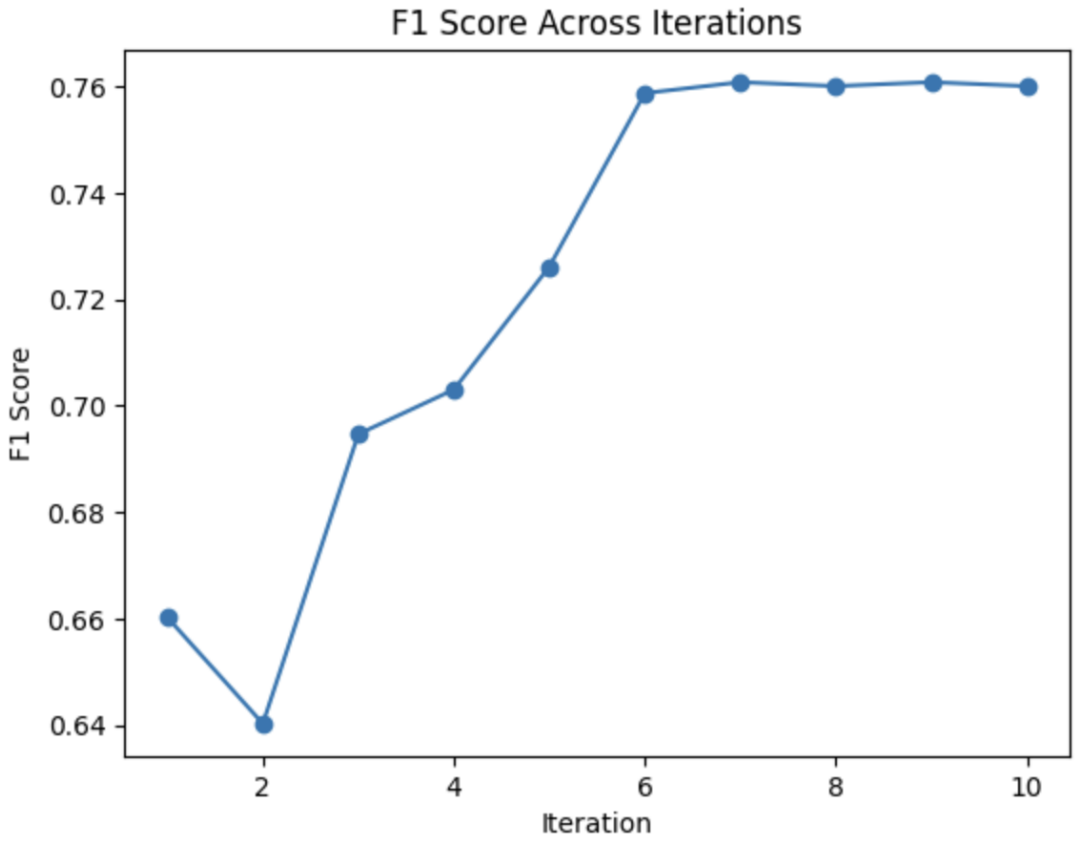


Fig: Video_Game F1_score

The result of the video_game is promising to explore more on the model. We then tested other data from other categories. However, we found many imbalanced data across different categories after we tested the model. Here is the label distribution of some sample data (see Fig Label Counts):

As the plot shows, we can see that label 0 is far more than label 1. The fact indicates that across all the viewed and bought items, customers browsed (viewed) items more than purchased, especially in the Toys and Games and Grocery categories. We can infer that there are many options in these two categories, and prices are not very high; thus, people tend to browse and compare more when shopping for grocery and toy items. In contrast, the luxury_beauty and gift_card have much fewer purchased items. The reason why there are bigger gaps is that these items are not everyday items. Customers may take a look at that but seldom purchase one. We can also draw reasonable conclusions from video games and automotive categories. Items under video games and automotive are the items people need, but the opinions are not as rich as categories like grocery. People typically view these items when they have purchased intents. Therefore, a fair amount of the items are purchased, proportional to the number of viewed items.

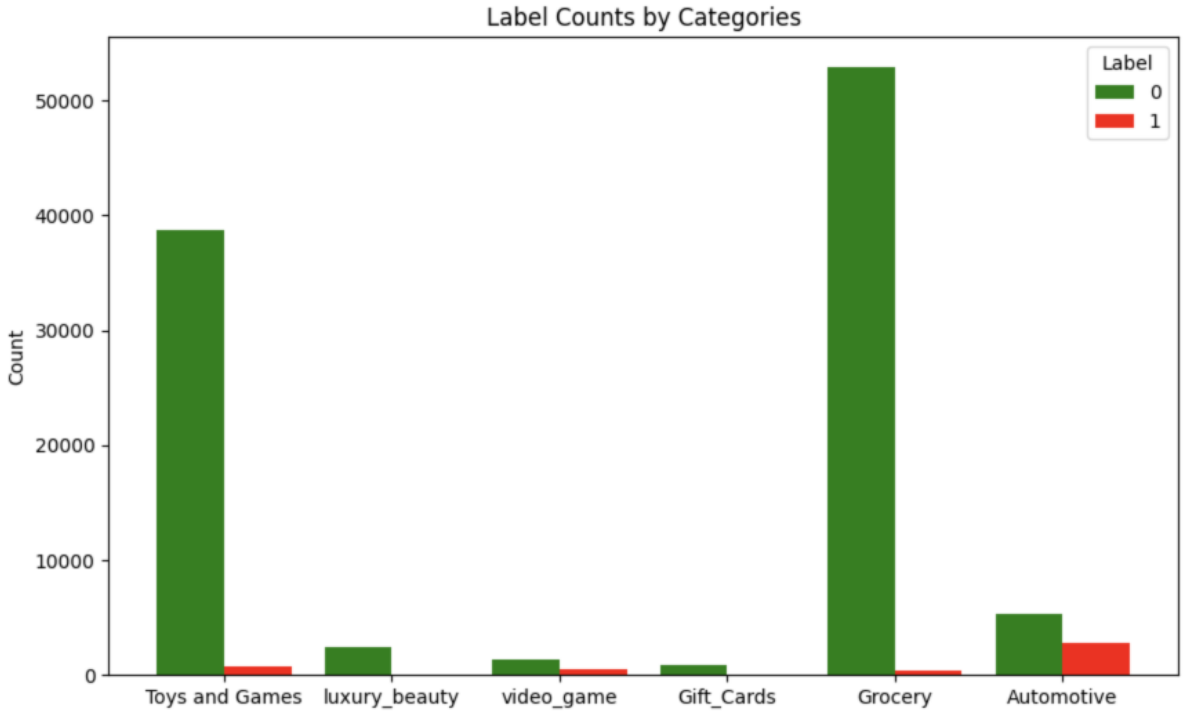


Fig: Label Counts

Considering the model prediction of video games and the data balance, we then trained the model on the automotive category to test our confidence in the model. The model performance of automotive shows results similar to those of video games (see Fig: F1_score of Automotive Data). The F1 score starts converging around the 15th iteration, and the F1 score finally falls between 0.78 and 0.8, which is similar to the video game performance. Since the automotive data has 8140 records, about 4~5 times the video game data, we concluded that the data size is still not enough to fit a good model. In future model-building work, we need more data to keep the labels balanced and then use the model to predict potential purchases for promotion purposes.

Limitations and Future Directions on Models

- Imbalanced Data: The primary limitation identified is the imbalance in the labeled data across different categories, where the count of label 0 significantly outweighs label 1. This imbalance can lead to biased model predictions and may impact the model's generalization to different categories.

- Limited Data Size: The available dataset, particularly for categories like video games, automotive, and luxury beauty, may need to be bigger to train a robust model. The conclusion that more data is needed suggests a potential limitation in the current dataset's size and diversity.

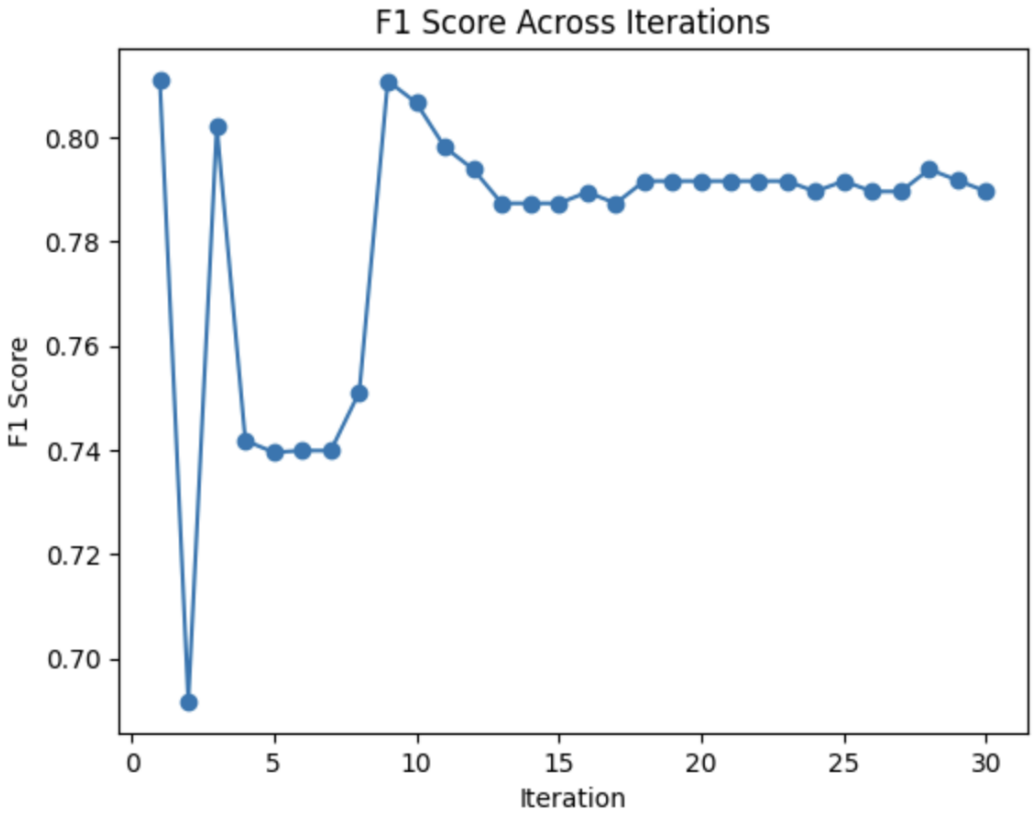


Fig: F1_score of Automotive Data

- Assumption of Convergence: The assumption that the F1 score converges after a certain number of iterations might be premature. Convergence depends on multiple factors, and the choice of the number of iterations could influence the results. A more thorough convergence analysis, considering multiple runs and evaluation metrics, could provide a clearer picture.

- Category-Specific Behavior: The observed behavior and purchasing patterns might be specific to certain categories, and generalizing the findings across diverse product categories may not be accurate. The model's performance on automotive data may not be directly applicable to other categories with different customer behaviors.

Discussion

Interpreting Network Analysis Insights

Our network analysis of Amazon's product categories has offered valuable insights, particularly in understanding consumer behavior and product relationships. The application of k-core filtering and centrality measures has helped us map out the intricate connections within Amazon's product ecosystem. The scale-free nature of the network, indicated by the power law distribution in degree, suggests a pattern of a few highly interconnected products or 'hubs.' This aspect of the network provides a fundamental understanding of how certain products might influence consumer choices more than others.

The assortativity coefficients highlight interesting aspects of the network: the negative degree assortativity suggests a diverse range of product associations, reflecting varied consumer interests. The negligible price

assortativity coefficient indicates that pricing might not be a significant factor in how products are associated within the network.

Consumer Behavior and Market Trends

The Item Bought to Item Viewed Ratio and the analysis of price ratios across different categories have given us a peek into varying consumer purchase behaviors. Categories like Gifts exhibit a high ratio, possibly due to impulse buying or the need-based nature of such purchases, as opposed to more deliberate decision-making in categories like Automotive. These insights, while limited by the scope of our data, provide an initial understanding of how pricing and other factors might influence purchasing decisions in different product categories.

Conclusion

Key Findings and Implications

This student-led project has successfully employed network analysis to gain insights into the complex dynamics of Amazon's e-commerce platform. We found that the product network on Amazon shows a scale-free characteristic, with some products serving as influential nodes. The study suggests that consumer purchasing behaviors vary across different categories, influenced by factors like the type of product and its price.

However, there are limitations to our findings, primarily due to the imbalanced nature of the data and its focus on specific categories, which might limit the breadth of our conclusions. Our observations provide an initial understanding rather than a comprehensive overview of the e-commerce landscape.

Future Directions

The project highlights the potential of network analysis in e-commerce, but it also points to the need for more balanced and extensive datasets for future research. Continuous updating and data analysis are crucial to keep pace with evolving market trends and consumer behaviors. This study serves as a stepping stone for further research in this area, offering a foundational perspective for students and academics interested in exploring the applications of network analysis in digital marketplaces.