Analyzing Amazon Product Reviews Through Network Analysis

Aum Patel: 202101118 Sankalp Bohidar: 202101122
Abhi Thumar: 202101123 Aditya Prajapati: 202101133

***Abstract*—This research explores user interactions and movie product dynamics within the Amazon movie review dataset, which contains approximately 8 million reviews spanning over 10 years. Collected from the Stanford Network Analysis Project (SNAP), the dataset provides key attributes such as user IDs, product IDs, review scores, and timestamps, offering a foundation for various analytical methodologies, including network analysis and temporal dynamics. This study aims to understand how users engage with products over time and how their expertise evolves through interactions with differ- ent products. These networks will be used to examine interactions at both the individual and community levels. Key aspects of the analysis include identifying central users and popular products using degree centrality and other network metrics. Additionally, we will analyze temporal dynamics to model how user preferences change over time and the role of user expertise in shaping product ratings. The study will also focus on applying community detection methods to uncover groups of users with similar tastes or review behaviors and to understand how these communities evolve.**

I. Introduction

In the digital age, personalized recommendations and influencer analysis are critical components of user engagement on e-commerce and social media platforms. This project addresses two objectives:

1) Product Recommendation: Using Jaccard similarity, we recommend similar products to users based on common interaction patterns.
2) Influencer Identification: By leveraging centrality measures in a user-product network, we identify top influencers who can impact trends and drive engagement within the community.

The field of network analysis offers powerful methods to model and analyze interactions among entities, whether users, products, or other networked elements. In this project, a bipartite network structure is formed between users and products, with users represented as one set of nodes and products as another. The edges represent interactions such as reviews or ratings.

II. Dataset Description

The dataset comprises a detailed user-product interaction record, capturing essential data for our analysis. Key attributes include:

- Product ID (productId): A unique identifier for each product in the network.
- User ID (userId): A unique identifier for each user who has interacted with a product.
- Profile Name (profileName): Usernames as displayed on the platform.
- Helpfulness: A measure of how many users found a review helpful, indicating review impact.
- Score: The rating given to a product, providing insight into user satisfaction.
- Time: Timestamp of each user-product interaction.
- Review Text: Detailed content of the review, which can be analyzed further for sentiment analysis and keyword extraction.

This dataset allows us to analyze both user engagement (through ratings and review frequency) and product popularity (based on interaction volume and average rating).

III. Network Properties and Structure

The network is structured as a bipartite graph, where users and products form two distinct node sets. Edges represent interactions, which in this project are primarily user reviews and ratings of products.

Basic Network Statistics

- Nodes (Users): Total count of unique users in the dataset.
- Nodes (Products): Total count of unique products in the dataset.
- Edges (Interactions): Each edge represents a user interaction with a product, such as a rating or a review.

This bipartite structure enables us to use techniques like Jaccard similarity for product recommendation, as well as degree and eigenvector centrality to analyze user influence.

Density and Degree Distribution

The density of this network is low, typical of large social networks, indicating that most users interact with only a limited set of products. Degree distribution analysis reveals that a small subset of users interacts with many products, while most users interact with only a few products.

IV. Methodology

A. Product Recommendation Using Jaccard Similarity

The Jaccard Similarity Index measures the similarity between two sets and is defined as the size of the intersection divided by the size of the union of two sets. In this project,

Jaccard similarity is used to identify products that are frequently reviewed or rated by the same users, indicating similar appeal or utility.

$$\text{Jaccard Similarity } = \frac{|A \cap B|}{|A \cup B|} \tag{1}$$

where A and B represent sets of users who have interacted with two different products.
Steps:

1) User-Product Matrix Construction: A sparse matrix is created where rows represent users and columns represent products. Each entry records the interaction (e.g., a rating or review).
2) Similarity Calculation: For each pair of products, the Jaccard similarity is calculated based on user overlap.
3) Recommendation Output: For each product, the most similar products (based on a similarity threshold) are recommended.

B. Influencer Identification Using Centrality Measures

Two primary centrality measures are used:

1) Degree Centrality: Represents the number of direct connections a user has, indicating their immediate engagement within the network.
2) Eigenvector Centrality: Reflects the influence of a user based on their connections to other highly connected or influential users.

Steps:

1) Network Graph Construction: A directed graph is created where edges represent user interactions with products. The direction indicates the flow from user to product or vice versa.
2) Degree Centrality Calculation: Identifies users with high engagement through direct connections.
3) Eigenvector Centrality Calculation: Highlights users connected to other influential users, thus representing potential influencers.

To rank influencers, we combined both measures, using a weighted average to balance direct engagement with connectedness to influential nodes.

V. Data Preprocessing and Cleaning

The initial dataset required cleaning and transformation to facilitate analysis:

1) Missing Values: Rows with missing essential data (e.g., userId, productId, score) were removed.
2) Data Filtering: Only relevant columns were retained, and redundant data was filtered out.
3) Normalization: Scores were normalized to standardize ratings across different scales.
4) Downsampling: To manage computational efficiency, a subset of data was used for initial tests.

VI. Analysis and Results

A. Product Popularity Analysis

Two key metrics were derived:

1) Review Count: Counts the number of interactions for each product, reflecting its visibility and popularity.
2) Average Score: Calculates the mean rating for each product, providing a measure of quality based on user satisfaction.

A Combined Popularity Score was computed by normalizing and averaging these two metrics, allowing for balanced ranking that considers both frequency of interaction and user satisfaction.

Top 5 Most Popular Products:

Product ID	Review Count	Average Score	Popularity Score
B008Y5QTRG	413	4.312349	0.914044
B000K2EQBW	321	4.464174	0.821371
B00004YA6S	302	4.470199	0.799066
B00004TXKE	242	4.607438	0.743405
6305603847	221	4.488688	0.703076

B. Influencer Identification

Using centrality measures, the Top 5 Influencers were identified as users with the highest combined degree and eigenvector centrality.

User ID	Degree Centrality	Eigenvector Centrality	Average Rank
AJKWF4W7QD4NS	0.001351	0.041209	5.25
AWG2O9C42XW5G	0.001182	0.041378	7.25
A1JH5J1KQAUBMP	0.001013	0.042335	11.00
A1NI0VOVXZBM7Y	0.001351	0.037494	12.75
A1GHUN5HXMHZ89	0.001013	0.039528	13.50

C. Community Detection

We use community detection in this network to identify clusters of users with similar interests and related products(movies) based on shared reviews. By identifying these communities, we create a basis for personalized recommendations within each group, making suggestions more relevant. The significance of community detection lies in its ability to reveal natural groupings, allowing recommendations to be driven by users' and products' connections, which enhances recommendation accuracy by focusing on shared interests rather than general popularity.

D. Trusted Reviewers

1) Calculating the average score for each product across all reviews, establishing a benchmark for what an "average" score looks like.
2) Then we calculate the difference between the user's rating and the product's average score.
3) Users with an average deviation below a set threshold (0.5 in our case are considered trusted.
4) Finally, we examine the network influence of these trusted reviewers within the user-product network by calculating degree centrality scores.

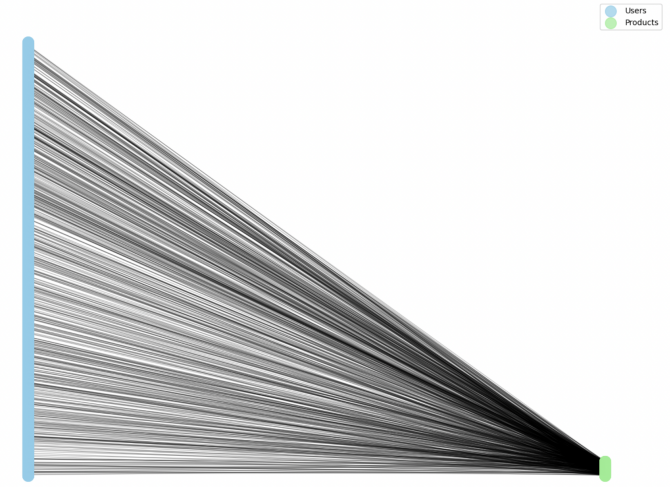

Fig. 1. User-Product Bipartite Network

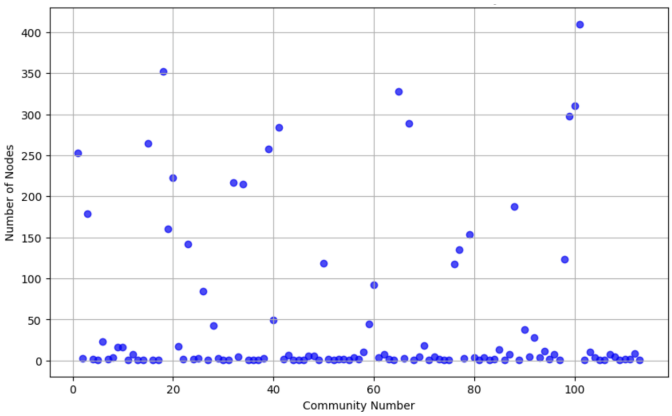

Fig. 2. Number of users in each detected community

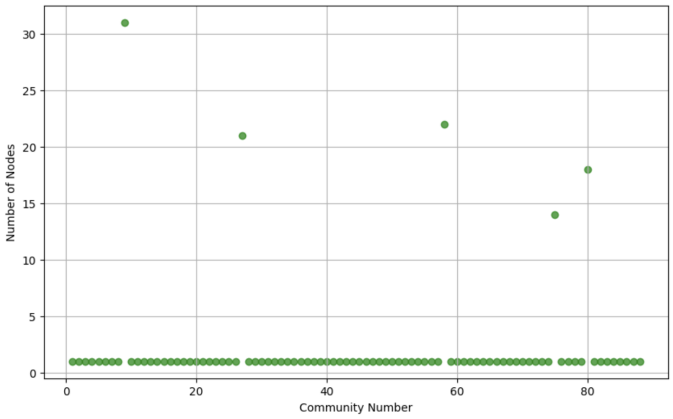

Fig. 3. Number of products in each detected community

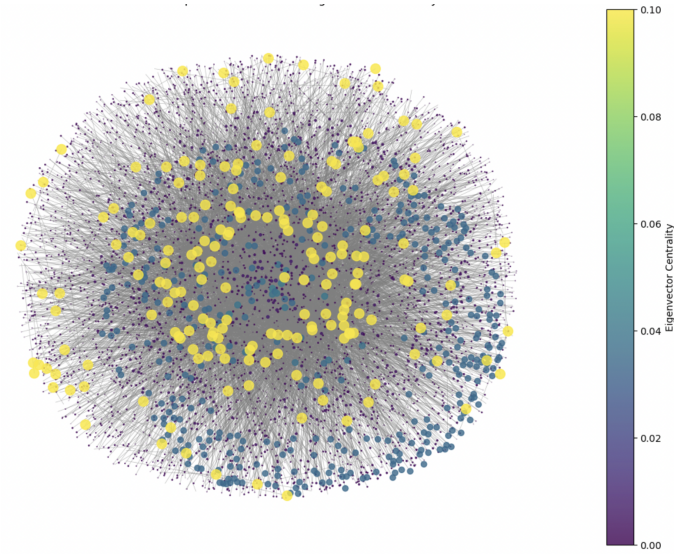

Fig. 4. Bipartite Network with Eigenvector Centrality

VII. Visualization and Insights

- Network Visualization: The network graph was visualized using Network and Matplotlib libraries, displaying major influencers and popular products. We can clearly see the Bipartite Network formed for our dataset (by using sampled edges for clarity) in the figure1.
- This graph helps visualize the direct connections between users and the products they interact with.
- The user projection reveals shared interests among users based on common product reviews, while the product projection reveals similarities among products based on common reviewers (in figure2 and figure3).
- Products associated with high-centrality users are likely to become more visible and influential within the network. Though product nodes have a fixed size and color, the presence of high-centrality users around certain product clusters suggests those items are not just popular but may also influence user connections within the network.
- Visualizing the community structure in figure4 helps understand the modularity of the graph and see how tightly or loosely communities are formed. It's also possible to identify isolated communities or highly connected communities based on the clustering.

VIII. Future Scope

1) User Segmentation and Targeted Recommendations: By analyzing user communities, segmented recommendations can be tailored to different audience preferences, enhancing user experience and engagement.
2) Movie Clustering for Genre Insights: Movie communities can reveal hidden genre or theme similarities, helping classify movies based on user-driven patterns and supporting genre-based recommendations.
3) Identifying Popular vs. Niche Content: Community size analysis can distinguish mainstream from niche movies, aiding in content prioritization and identifying long-tail effects for specialized audiences.
4) Enhanced Recommendation Systems: Cross-community analysis can improve recommendations, especially for new or less-reviewed movies, by associating them with similar user preferences to overcome the cold-start problem.

IX. Conclusion

This project demonstrates the application of network analysis techniques for practical e-commerce challenges:

1) Effective Product Recommendations: Jaccard similarity provides meaningful recommendations, enhancing user experience.
2) Influencer Identification: Centrality metrics allow for targeted marketing by identifying users with significant influence.
3) Community Insights: Community detection highlights clusters of similar interests, which can guide platform segmentation strategies.

X. References

[1] Amazon Movie Review Dataset, Stanford Network Analysis Project (SNAP). Available: https://snap.stanford.edu/data/movies.txt.gz

[2] J. McAuley and J. Leskovec, "From Amateurs to Connoisseurs: Modeling the Evolution of User Expertise through Online Reviews", in WWW, 2013. Available: http://i.stanford.edu/ julian/pdfs/www13.pdf

[3] Colab File: https://colab.research.google.com/drive/1-G6KJ-gnwaTSp4KCY-wnW5y8rvmq3pSk?usp=sharing