Network Analysis on Amazon metadata

Ehsan Ramezani, Master's Degree in Artificial Intelligence, 0001109969
Mohammad Pourtaheri, Master's Degree in Artificial Intelligence, 0001121324
Mehregan Nazarmohsenifakori, Master's Degree in Artificial Intelligence, 0001120680

A.A. 2024/2025

1 Introduction

In this report, we explore Amazon's data to understand how people behave when shopping online, how products are related, and how customers influence others. We use network analysis to create a graph that models interactions between customers and products, helping us find important patterns. Specifically, we look at how customer reviews, product similarities, and influential customers shape the network structure. Our analysis includes building detailed graphs, identifying key products and top customers, and examining how communities form. The findings of this study aim to improve our understanding of online buying behaviors, guide focused marketing techniques, and improve the user experience on online websites that sell goods or services to people.

2 Problem and Motivation

E-commerce platforms generate vast data on products, customers, and their interactions (e.g., reviews, co-purchases, similarities). We found that SNA can effectively harness product-customer relationships by modeling them as nodes and edges, revealing hidden structures. Degree and betweenness centralities highlight influential nodes, while community detection, k-core decomposition, clique analysis, and small-worldness expose clusters and overall network cohesion. Correlation and homophily analyses then link these features to performance metrics. Although businesses can refine recommendations and better predict sales with these insights, large-scale challenges remain:

- **Identifying Influential Products and Customers:** Traditional metrics (e.g., raw sales) do not explain why certain products or customers become focal points. By measuring degree centrality, we can pinpoint the most sought-after items or the most active shoppers. Moreover, betweenness centrality surfaces products or customers that bridge disconnected segments, highlighting "broker" opportunities for cross-promotion.

- **Detecting Communities and Network Cohesion:** Products and customers do not exist in isolation; they form clusters where behaviors, preferences, and co-purchasing habits converge. Community detection (e.g., Louvain) reveals these hidden subgroups, while modularity quantifies how strongly these subgroups are separated from the rest. This is essential for targeted campaigns since a tightly knit community can be approached with focused marketing strategies that would not work in a scattered or weakly connected group.

- **Uncovering Dense Purchase Patterns:** Certain products are consistently purchased or reviewed together, forming small, tightly knit cliques. Clique analysis identifies these maximal, fully connected subgraphs—useful for creating multi-product bundles, cross-promotions, or curated recommendations. Similarly, k-core decomposition reveals the "core" subset of the network where the most interactions happen, helping stakeholders invest in the most influential nodes.

- **Handling Fragmentation and Isolated Segments:** Connected components analysis points out pockets of the network that are disconnected from the main market. Understanding or re-engaging these niches (or isolated nodes) can unlock untapped potential, as they may represent highly specialized products or overlooked customer segments.

- **Tracking Evolution over Time:** Customer tastes and product performance vary across seasons and years. Temporal analysis helps detect surges, declines, or shifts in purchasing patterns, providing input for dynamic stock management and time-sensitive marketing decisions.

- **Linking Network Structure to Business Outcomes:** Finally, robust correlation analyses (Spearman/Pearson), permutation testing, and homophily measurements tie all these network-derived metrics (e.g., helpful reviews, centrality) to key performance indicators such as sales rank. Small-worldness underscores how efficiently information and influence spread across the network. Taken together, these techniques let us move beyond broad conjectures and toward quantitative, evidence-based insights on improving sales and customer satisfaction.

3 Datasets

The dataset was compiled by Jure Leskovec, Lada A. Adamic, and Bernardo A. Huberman for their study on viral marketing dynamics. They collected the data by crawling Amazon's website during summer 2006, focusing on product metadata and customer reviews across categories like Books, Music CDs, DVDs, and VHS tapes.

This dataset is publicly accessible through the Stanford Network Analysis Project (SNAP) at Stanford University. It includes information on 548,552 products, encompassing titles, sales ranks, similar product lists, categorizations, and customer reviews with specifics like review time, customer ID, rating, number of votes, and helpfulness votes.

We used dataset node and edge features and attributes without digitizing them. However, we removed categories due to their complex format, so we map them to Product IDs whenever we need them to calculate rate, vote, and helpfulness of reviews for the customer-product Network as follows:

$$w = \left(\frac{\text{votes}}{\max(\text{votes})} \cdot \frac{\text{helpful}}{\text{votes}+1} \cdot \text{rating} \right)$$

This formula balances three viewpoints: the review's popularity (through votes), the perceived utility of its content (through helpfulness ratio), and the customers' assessment (through rating).

Other formulations for this part have been implemented in various parts of our project to find the insights we intended.

Tools Utilized:
Several important libraries were used in this project. Pandas was used to work with data and do tabular analysis. NetworkX was used to make and study network graphs in social network analysis. Scikit-learn was used for structural equivalence analysis clustering algorithms like Agglomerative Clustering. Matplotlib and Seaborn were used to show data patterns and network structures. And the Community Package was used to use the Louvain Method to find communities. We employed all these libraries through Python, leveraging its versatile applications and efficient handling of large datasets.

4 Validity and Reliability

In order for our findings to be credible and reproducible, we carefully considered several factors, including accurately modeling real-world interactions between Amazon products and customers, employing consistent data preprocessing techniques such as handling missing values, normalizing data, and using standardized libraries like NetworkX and Scikit-learn for calculations. While subjective decisions, like cluster selection, and stochastic processes in algorithms can introduce variability, these were mitigated by using fixed random seeds to ensure consistent outcomes.

Methods and Justification for subgraph:
A strategic approach was employed to construct a representative subgraph comprising approximately 1,000 nodes from the extensive Amazon product-customer network. The process started with 40 seed nodes, evenly split between products and customers to ensure balanced representation. To ensure coverage of essential product types, one seed was allocated from each of the five groups: 'Music', 'DVD', 'Video', 'Toy', and 'Software'. This reservation ensures significant product segments are represented in the subgraph, even with a limited number of products. For the remaining product seeds, stratified sampling based on node degree was used to achieve even representation across degree ranges. This method partitioned the ordered list of product nodes into equal strata by their degrees, randomly selecting seeds from each stratum. The methodology covers a diverse array of products, including both highly connected and less connected items. Customer seeds were selected using stratified sampling to ensure customers with diverse interaction levels were represented. This involved partitioning customer nodes by degree and conducting random sampling within each partition to preserve diversity.

Following this, Breadth-First Search (BFS) expansion was conducted from each seed to a depth of three, incorporating a per-seed node limit to ensure diversity and prevent any single seed from dominating the subgraph. This technique captures the local structures and community dynamics of the network, yielding a balanced subgraph that accurately reflects the complexity of the original network.Figure 1 shows the subgraph.

5 Measures and Results

For our research, we set up two distinct networks: the product-customer network and the product-product network.

Product-customer: Without taking into account their categories, products are nodes with attributes that correspond to each product feature (ASIN ID, title, group, and sales rank) in the dataset. Other types of nodes we have are customers with the total number of reviews as an

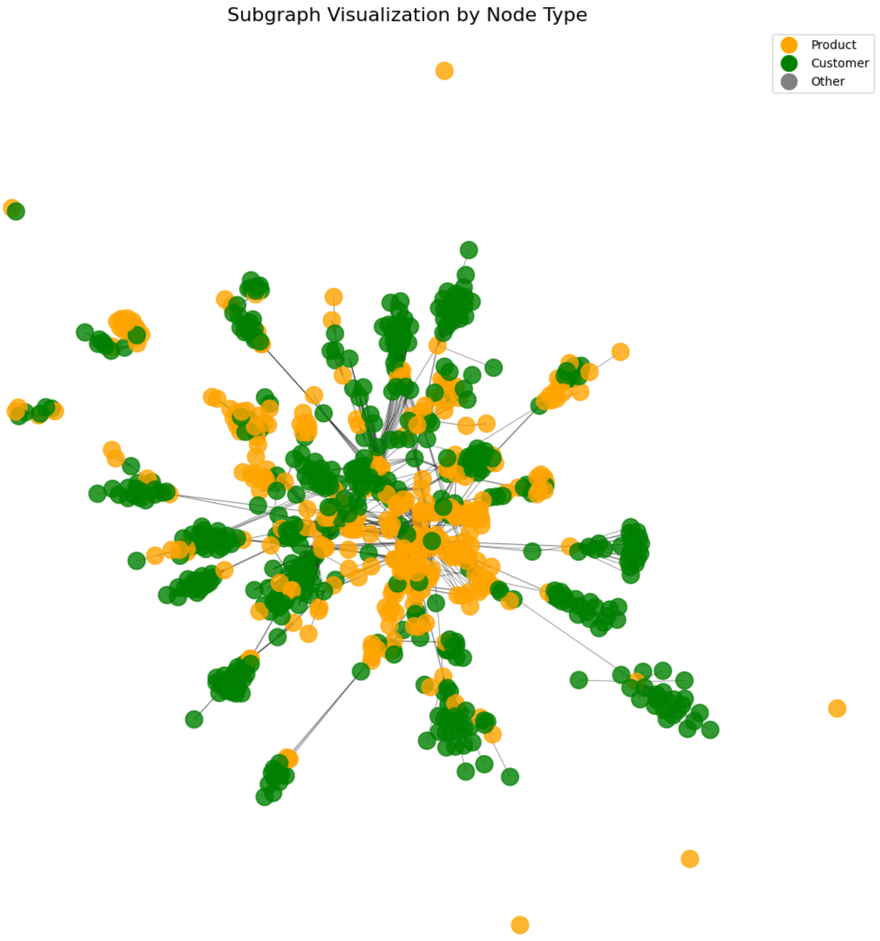

Figure 1: The subgraph used for the analysis had 817 nodes and 1332 edges.

attribute. The particular network has two edges: customers connect to products if they have reviewed them, and products connect to other products if they are related based on similar IDs in the dataset. The attributes of the customer-product edges include the weight of their reviews on the particular product with which they are associated and the time stamps indicating the product's purchase or review date.

product-product networks: in which a product is represented by each node. A network of relationships based on product similarity is created when nodes are connected by edges if the products are identified in the dataset as similar. and the attributes of the products are similar to those of the previous network, with the exception that we have the average rating of the product reviewers as an attribute.

There could be many metrics that apply to our dataset, however we have used the following ones for our analysis:

5.1 Degree Centrality Definition:

Degree centrality quantifies the number of direct connections a node possesses within a network. It measures the number of direct connections each node has.

5.1.1 Within the product-customer subgraph network:

Products with high degree centrality show strong popularity, attracting many customer interactions.These products are likely to be best-sellers and appeal to various customer segments.
Customers with high centrality show significant activity, engaging with many products.They can be regular consumers, fashion innovators, or have diverse buying preferences.

Results:

We calculated centrality for the entire network. Below is 1, showing the top 10 products and customers with the highest centrality degree. Average product ratings were calculated, and products were categorized as "too high," "medium," or "too low" based on set thresholds. Filtering ensured only products with enough reviews were analyzed. Examined the relation between high degree centrality (top 25%) and extreme ratings, showing a possible positive correlation with higher centrality and high average ratings. Visualization of extreme rating proportions for high-centrality products (2) supported the hypothesis that products with higher degree centrality are more likely to have higher average ratings.

Table 1: Top 10 Products and Customers by Centrality

#	Top 10 Products	#	Top 10 Customers
1	0807220299 (0.001562): Harry Potter and the Order of the Phoenix (Book 5 Audio CD)	1	ATVPDKIKX0DER (0.044150): 945065 reviews
2	043935806X (0.001534): Harry Potter and the Order of the Phoenix (Book 5)	2	A3UN6WX5RRO2AG (0.011367): 201770 reviews
3	0807220280 (0.001484): Harry Potter and the Order of the Phoenix (Book 5, Audio)	3	A14OJS0VWMOSWO (0.003352): 9795 reviews
4	0439567629 (0.001482): Harry Potter and the Order of the Phoenix (Book 5, Deluxe Edition)	4	A2NJ06YE954DBH (0.002173): 6324 reviews
5	0807282588 (0.001457): Harry Potter and the Goblet of Fire (Book 4, Audio)	5	AFVQZQ8PW0L (0.001898): 5441 reviews
6	0439139600 (0.001457): Harry Potter and the Goblet of Fire (Book 4)	6	A9Q28YTL7YE07 (0.001512): 4296 reviews
7	0439139597 (0.001454): Harry Potter and the Goblet of Fire (Book 4)	7	A1K1JW1C5CUSUZ (0.001204): 3576 reviews
8	0786229276 (0.001445): Harry Potter and the Goblet of Fire (Book 4)	8	A3LZGLA88K0LA0 (0.000804): 2276 reviews
9	0939173379 (0.001430): Harry Potter and the Goblet of Fire (Book 4 Audio CD)	9	AU8552YC005QX (0.000758): 2864 reviews
10	0807282596 (0.001430): St. Anger (with Bonus DVD)	10	A20EEWSFMZ1PN (0.000753): 2181 reviews

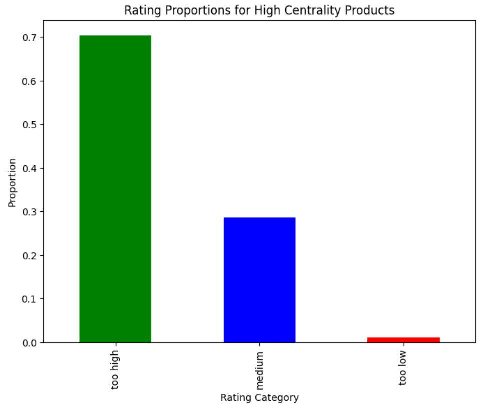

Figure 2: Rating Proportions for High Centrality Products

5.1.2 Within the product-product network:

In Predictive Analysis, degree centrality was used to quantify the connectivity of products based on their similarity to other products. Higher centrality indicates products that are more interconnected with other items, suggesting greater visibility and potential influence within the network.

5.2 Betweenness Centrality

Betweenness centrality quantifies the number of shortest paths between all pairs of nodes in the network that pass through a particular node. It identifies nodes that serve as critical bridges or brokers within the network.

5.2.1 Within the product-customer subgraph network:

Products with high betweenness centrality can be leveraged in cross-promotional strategies to connect different product lines. Customers with high betweenness centrality can be targeted for referral programs, reviews, or as brand ambassadors to enhance network cohesion and information dissemination.
Results:
the Figure 3 demonstrates 17 nodes with high betweenness centrality within our subgraph (**top 95th percentile > 0.0294**)

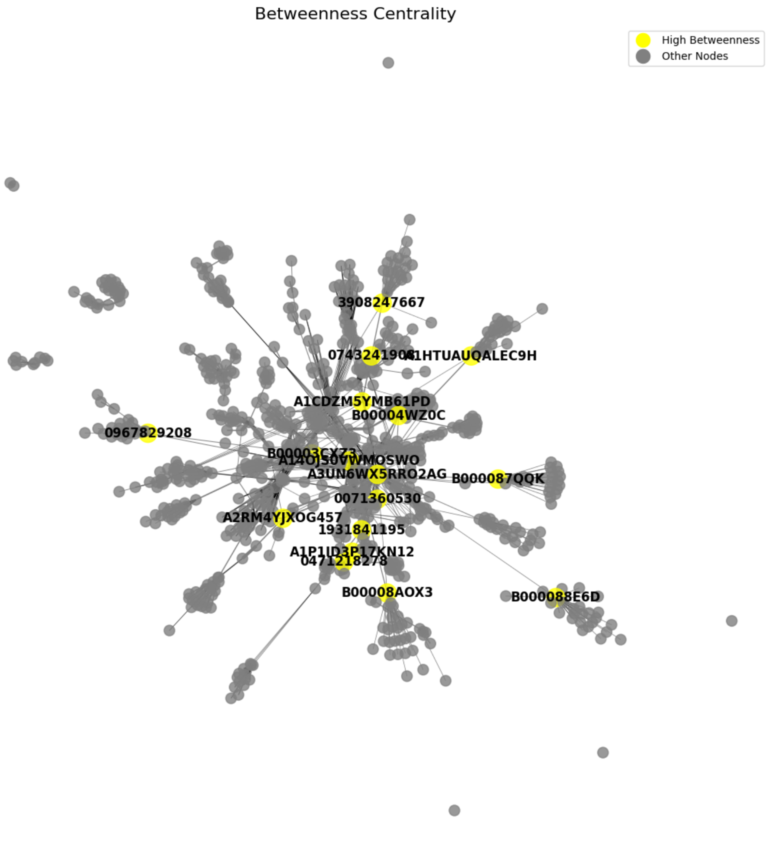

Figure 3: Betweenness Centrality

5.3 Communities

Community detection involves partitioning a network into clusters or groups where nodes within the same group are more densely connected to each other than to nodes in other groups. These communities often represent modules with shared characteristics or interactions.
Louvain method: The Louvain and Infomap algorithms were applied to identify product communities within the similarity network. These communities represent groups of products that

are often compared or co-purchased. Understanding these clusters helps analyze the impact of proximity to popular product groups on sales performance.

5.3.1 Within the product-customer subgraph network:

Leveraging community structures can enhance recommendation algorithms by suggesting products within the same community to users. Tailored marketing campaigns can be designed for distinct customer communities, improving engagement and conversion rates.

Results:

We identified 41 communities within our subgraph, represented by distinct colors for improved visualization. The communities exhibit various patterns between products and customers. For instance, the investing category in community number 4 is a common category across all products, indicating a similar product structure within this community. Another example is community 0, which contains a single product, music in the hip-hop genre while the remaining nodes are customers exhibiting their behaviors and preferences, allowing for future recommendations of additional hip-hop music (Figure 4).

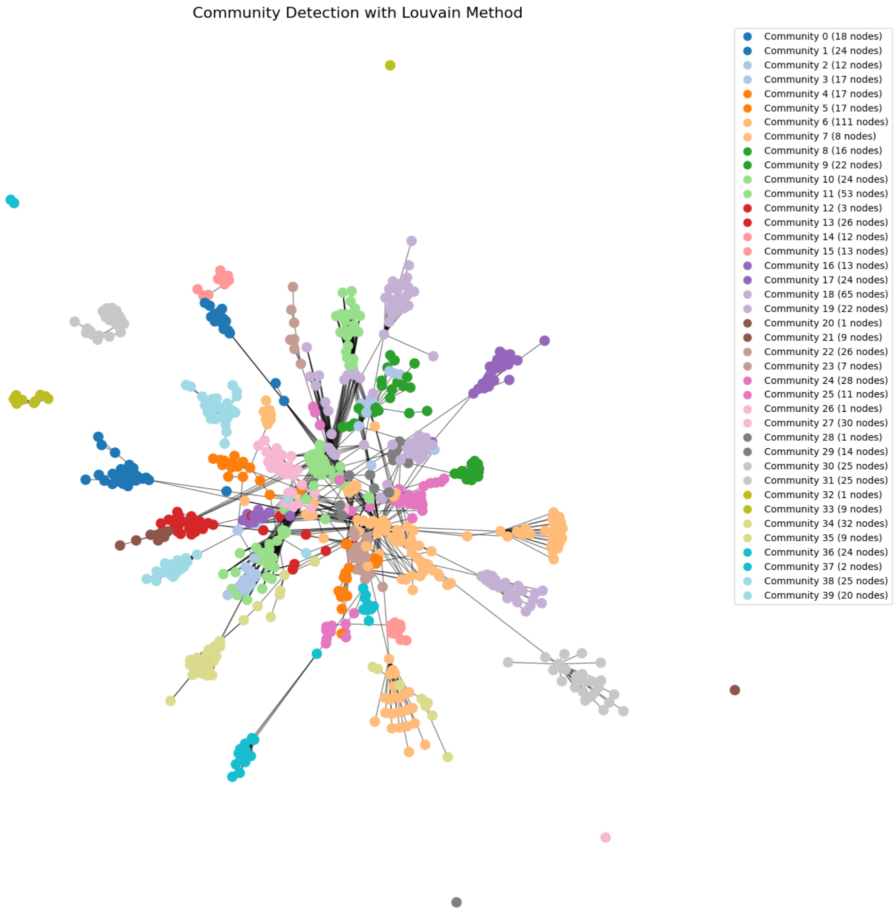

Figure 4: Communities detected within the subgraph based on the Louvain method

5.4 Modularity

Modularity is a metric that quantifies the strength of division of a network into communities. High modularity indicates dense connections within communities and sparse connections between them, signifying well-defined community structures.

5.4.1 Within the product-customer subgraph network:

Calculating modularity helps evaluate how effectively the network has been partitioned into distinct product and customer communities. High modularity can validate the effectiveness of targeted marketing strategies within well-defined communities. Low modularity may necessitate revisiting community detection parameters or exploring hybrid community structures to better capture network intricacies.
Results:
We calculated the modularity of the partitions and obtained a value of 0.9295, indicating that targeted marketing strategies will be effective within the communities identified in the the previous step.

5.5 Clique Analysis

Clique analysis identifies fully connected subgraphs within a network, where every node is directly connected to every other node in the clique. These cliques represent tightly-knit groups with maximal interconnectivity.

5.5.1 Within the product-customer subgraph network:

Product Cliques: Sets of products that are frequently co-purchased or co-reviewed by the same customers, indicating strong product associations. Customer Cliques: Groups of customers who interact with exactly the same set of products, reflecting shared purchasing patterns. Identifying product cliques facilitates the development of multi-product promotions and enhances the user experience through curated recommendations.
Results:
Our clique analysis reveals important structural characteristics of the dataset's network. We identified 1316 maximal cliques, indicating the presence of numerous small, densely connected subgroups where all nodes are directly connected, but these groups cannot be expanded without losing their complete connectivity. The size of the largest cliques, which is 3, shows that the largest fully connected subgroups in our network consist of only three nodes, highlighting a relatively sparse structure where larger cohesive groups are rare. Furthermore, we found 14 distinct largest cliques, suggesting that there are 14 separate instances of these fully connected triads. This analysis suggests that our network is characterized by localized dense connections, such as those formed by product similarity or customer-product interactions, rather than large, globally cohesive clusters.

5.5.2 Within the customer-customer subgraph network:

The customer-customer projection network was constructed by identifying connections between customers who reviewed common products. Two customers are linked by an edge if they have reviewed at least one product in common, and the weight of the edge represents the number of shared products reviewed. For Using this approach, a subgraph was selected based on degree centrality, focusing on products (in the product-customer network) randomly chosen from high, medium, and low centrality categories.

Results:

In the initial analysis focused on finding customers with the highest edge weight sum for identifying key customers(Figure 5). Further analysis focused on identifying the largest clique within the customer-customer projection network and calculating the weighted degree for each customer in the clique. To understand the relationship between customer behavior and network connections, category diversity was measured by counting the unique categories of products reviewed by each customer. Pearson and Spearman correlations(Table 2) between category diversity and weighted degree revealed a clear trend: customers who review products from more diverse categories tend to have stronger connections, reflected in higher weighted degrees within the network(Figure 6).

Table 2: Results of correlations

#	Name of correlations	Results of correlations	Results of p-value
1	Pearson	0.91	0.0000
2	Spearman	0.95	0.0000

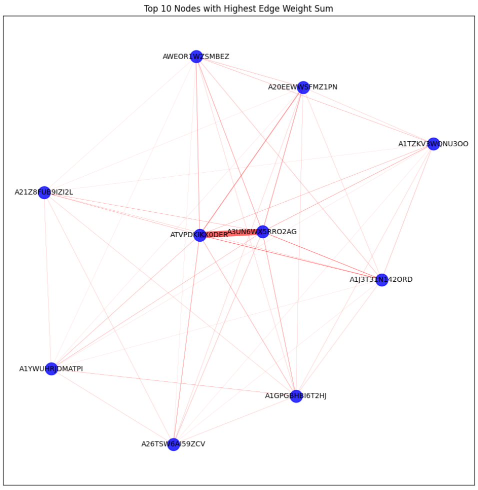

Figure 5: Top 10 Nodes with Highest Edge Weight Sum (Thicker edge indicates more weight.)

5.6 K-Core Decomposition

K-core decomposition involves peeling away layers of the network by iteratively removing nodes with a degree less than k. The k-core is the maximal subgraph where every node has at least k connections within the subgraph. This analysis reveals the network's core-periphery structure.

5.6.1 Within the product-customer subgraph netxwork:

Product Core: Products within higher k-cores are highly interconnected with numerous other products and customers, indicating centrality in the selling ecosystem. Customer Core: Cus-

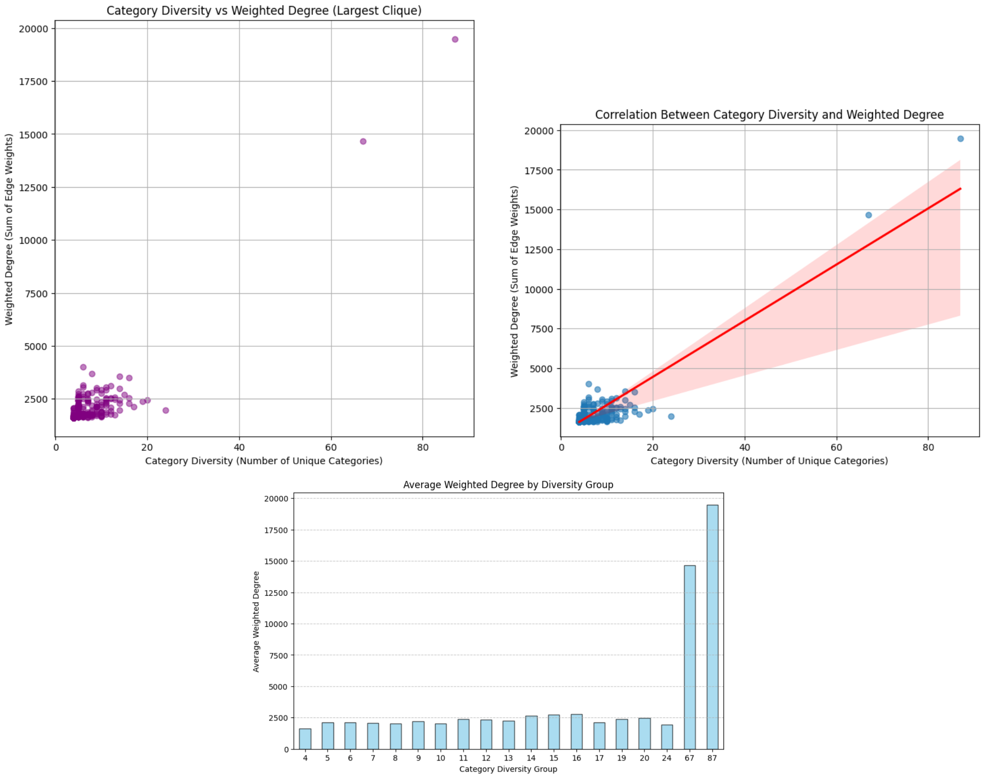

Figure 6: Relation between category diversity and weighted degree.

tomers in higher k-cores interact with many products and are central within the customer network.

Results:

The k-core decomposition(7 reveals important structural insights into our customer-product network. The presence of high-core nodes (e.g., in the 4-core and 5-core) highlights influential customers and popular products that form tightly interconnected subgroups, making them key targets for engagement and promotional strategies. On the other hand, the 1-core's large number of nodes shows a long tail of customers and products that are only loosely connected. This means that personalized suggestions or incentives could be used to get these customers and products more involved. This study shows how important it is to use high-core nodes to gain power and strategically involve low-core nodes to increase network activity and connectivity.

5.7 Components Analysis

Components analysis involves identifying connected components within a network, where each component is a subset of nodes that are interconnected, and no node in a component is connected to nodes in other components. This analysis highlights isolated or semi-isolated subgroups within the network.

5.7.1 Within the product-customer subgraph network:

Connected Components: Determine distinct clusters where products and customers are interconnected, but there's no interaction between different components.

Isolated Nodes: Identify products or customers that do not interact with any other nodes, indi-

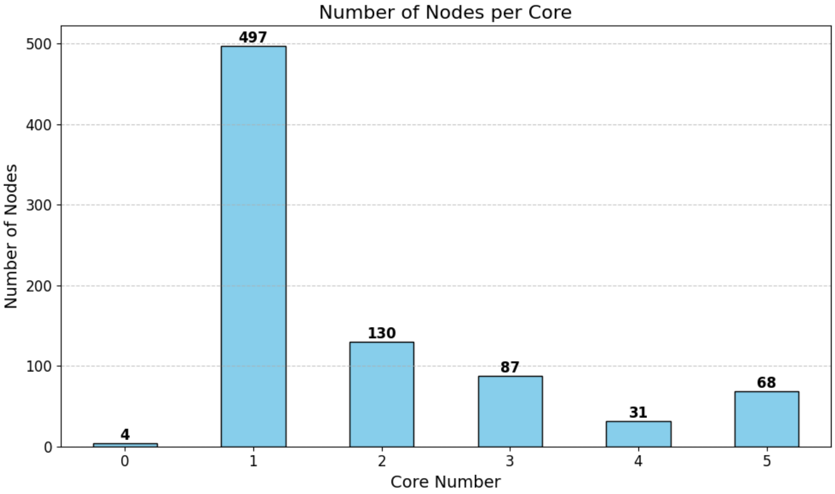

Figure 7: K-core decomposition

cating potential issues or niche cases. Addressing isolated nodes can enhance overall network engagement and ensure that all products and customers contribute to the network's value.
Results:

The component analysis highlights that the network's largest component, with 709 nodes, forms a cohesive core where most customer-product interactions occur, representing the primary focus for influence and engagement strategies. The presence of several smaller components, ranging from 25 nodes to isolated nodes, indicates fragmented groups likely tied to niche products or inactive customers. These disconnected or minimally connected nodes present opportunities to enhance integration through cross-promotion, personalized recommendations, or strategies to re-engage less active users. Strengthening connections between smaller components and the core network could significantly improve overall network cohesion and activity.

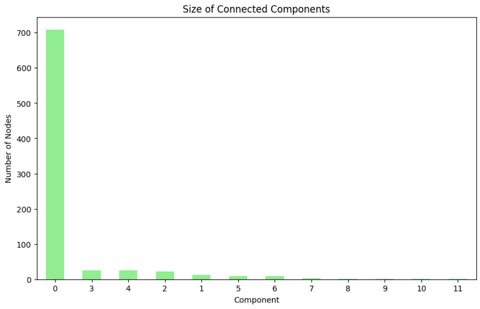

Figure 8: size of connected components

5.8 Temporal Analysis

Temporal analysis in Network Analysis examines how the structure and relationships within a network evolve over time, capturing dynamic patterns and trends. It lets us study things that change over time, like how influences spread, how communities form, and how connections between nodes form or break down.

5.8.1 Within the product-customer subgraph network:

Temporal Trends: Keeping track of review activity over time helps find trends in products, sales patterns that change with the seasons, and changes in how customers feel about a product. These details can help with planning promotional efforts and keeping track of stock.
Results:
As shown in Figure 9, the varying patterns and structures observed across different years highlight shifting trends and behaviors. This represents only a subset of the primary network, where periods of reduced engagement or popularity in some parts are contrasted by increased activity and prominence in others.

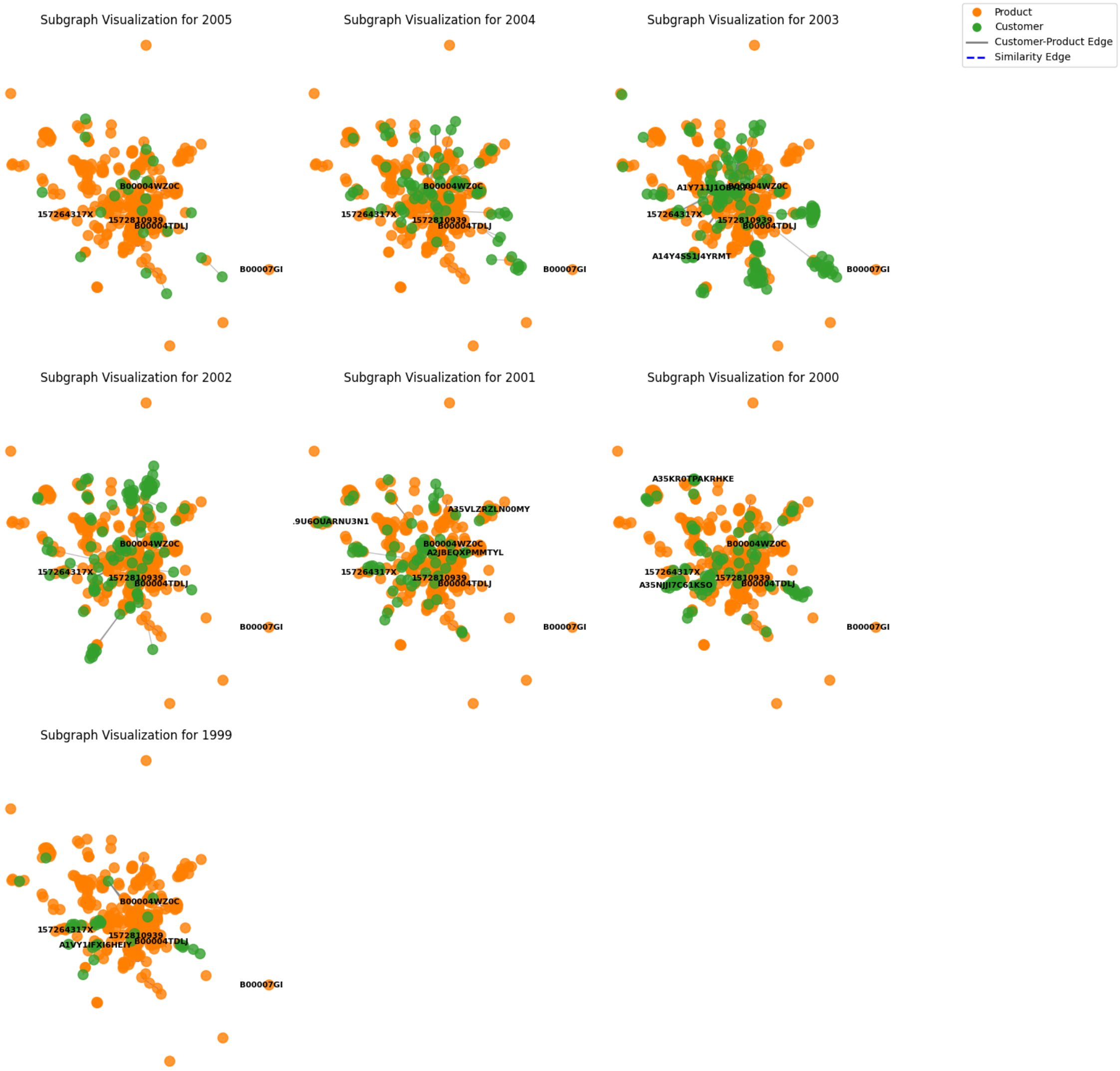

Figure 9: subgraph during different years from 1999-2005

5.9 Correlation Analysis

Spearman's rank correlation assesses the monotonic relationship between two variables, providing insight into the direction and strength of their relationship. Spearman correlation was used to examine the relationships between sales rank and network characteristics such as helpful reviews, degree centrality, and proximity to clusters. This method helped quantify how

these factors are associated with sales performance.
Pearson correlation measures the strength and direction of a linear relationship between two variables.

5.10 Permutation Testing

Permutation testing involves creating a null distribution by repeatedly shuffling the data and comparing the observed results to this distribution to derive statistical significance. Permutation tests were used to validate the significance of observed correlations. By generating a null distribution through random shuffling, we ensured that the observed relationships between network features and sales ranks were not due to chance.

5.11 Proximity to Clusters

The proximity of a product to densely connected clusters within the similarity network was measured using the shortest path distance from non-clustered products to clustered nodes. Clusters were defined by products with high degree centrality (top 20% based on centrality values). The distance to the nearest cluster was calculated using a breadth-first search (BFS) approach. This metric reflects how isolated or integrated a product is within the overall product network, with products closer to clusters expected to have better sales performance.

5.12 Homophily Analysis

Homophily measures the tendency of nodes in a network to connect with similar nodes based on a specified attribute. Positive homophily indicates a preference for similar nodes, while negative homophily suggests connections between dissimilar nodes.
Within the product-customer subgraph network:
In the Amazon product-customer network, nodes consist of two primary types: "customer" and "product." The network's structure is analyzed to understand how these nodes interact based on their types, reflecting the dynamics between customers and products.
Results:
The assortativity coefficient for the network based on node type is -0.9152, indicating a strong negative homophily. This demonstrates that customers primarily connect with products and vice versa, confirming the bipartite nature of the network with minimal intra-group connections. These findings support the structure of the network and can help improve recommendation systems by focusing on how customers and products interact with each other and creating targeted marketing campaigns that use the different roles of customers and products.

5.13 Small-Worldness Analysis

Small-worldness measures whether a network combines high clustering with short average path lengths. Networks with small-world properties are efficient at spreading information while maintaining tightly knit clusters. The **Sigma Score (σ)** is a metric used to quantify the "small-worldness" of a network by comparing the clustering coefficient and average path length of the real network to those of a random network.

Formula:

$$\sigma = \frac{C_{real}/C_{rand}}{L_{real}/L_{rand}}$$

The **Omega Score** (**ω**) is a metric that quantifies how close a network is to being a **random**, **regular (lattice-like)**, or **small-world** network by comparing its clustering coefficient and average shortest path length with those of a regular and a random network.

Formula:

$$\omega = \frac{L_{rand}}{L_{real}} - \frac{C_{real}}{C_{latt}}$$

Within the product-customer subgraph network:
This analysis evaluates the clustering and average path length in the Amazon product-customer network. Clustering measures how nodes form tightly knit groups, while path length determines the average number of steps required to connect any two nodes. These metrics assess how efficiently information or influence spreads within the network.
Results:
The study found that the average clustering coefficient was 0.0132, which means that there was very little clustering. This is in line with the fact that the network is a bipartite network, with customers mostly connecting to products. We found the average shortest path length to be 5.5786, indicating that the largest connected component can reach most nodes in approximately 5-6 steps. The network exhibits small-world properties, characterized by efficient information flow and short path lengths compared to a random graph. This is evidenced by a **sigma score of 3.1661**, significantly higher than 1, indicating enhanced clustering relative to a random graph. Additionally, the **omega score of 0.9605**, close to 1, reinforces these characteristics by showing that the network's path length is nearly as efficient as a random graph while its clustering surpasses that of a regular lattice. Together, these metrics confirm a robust small-world structure, balancing local clustering with global connectivity.

5.14 Predictive Analysis

Objective: This study aims to analyze product networks to identify factors that optimize product sales. By evaluating review characteristics, network connectivity, and proximity to communities, we aim to provide actionable insights for businesses like Amazon.

Initial Hypotheses: We initially hypothesized that the number of similar products and the quantity of helpful comments could influence sales rank. This was based on the assumption that positive reviews and higher connectivity in the product similarity network were associated with improved sales performance.

Analysis of Initial Hypotheses: Through correlation analysis (Pearson correlation and permutation tests), we observed a relationship between sales rank and both the number of reviews and their perceived helpfulness. Visualizations revealed that products with more reviews tend to have lower sales ranks, indicating better sales performance. The permutation test showed a negative correlation between sales rank and the number of reviews (observed correlation: -0.0792, p-value: < 0.001). Pearson correlation analysis further confirmed a weak inverse relationship between high-quality reviews and sales rank, suggesting that while an increase in high-quality reviews was associated with slightly improved sales, the effect was minimal. However, these findings raised an important concern: the potential fallacy of correlation implying causation. The observed correlation might be confounded by other factors, such as product quality, which could influence both sales performance and review characteristics. Products of inherently higher quality may naturally receive better reviews, creating a spurious relationship.

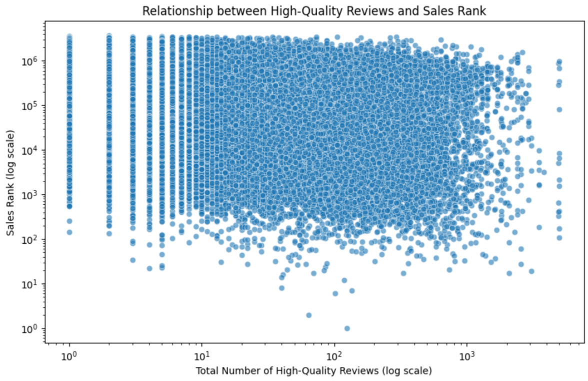

Figure 10: This plot raises concerns about the potential fallacy of assuming that correlation implies causation, as the relationship is not clearly observable.

Refinement of Hypotheses: To address this issue, we categorized products by average review scores into three quality groups: low-quality (0–3), medium-quality (3–4), and high-quality (4–5). By isolating the effect of product quality, we could better examine the influence of product similarity and review helpfulness on sales rank. This refinement led to the updated hypothesis: Products with medium or low quality are expected to have better sales ranks when they are closer to highly clustered regions in the product similarity network, compared to products located farther from these clusters. Additionally, medium- and low-quality products with fewer helpful reviews and weaker network connections tend to have worse sales ranks compared to products with higher connectivity or a greater number of helpful reviews.

5.14.1 Methodology for Hypothesis:

Data Preparation: Products were categorized into three quality groups based on average review ratings. High-quality products were excluded to focus on medium- and low-quality groups, reducing bias. Two metrics were calculated: (1) degree centrality, to measure product connectivity, and (2) the total number of helpful reviews, as a proxy for social influence.

Community Detection: We applied the Louvain and Infomap algorithms to identify clusters in the product similarity network. Each product was labeled with a community ID to facilitate cluster-based analysis.

Proximity Analysis: Shortest distances to highly clustered regions were precomputed as a measure of proximity. Products closer to these clusters were hypothesized to exhibit better sales performance. The logic behind this variable is based on how businesses like Amazon offer a number of similar products, which consumers often explore until they find what they want. If the proximity to a cluster of similar products is too large, we expect consumers to be less likely to explore those products.

Testing: The significance of observed correlations was validated using permutation tests. For each hypothesis, a null distribution was generated by shuffling sales rank values 1,000 times,

and observed correlations were compared to this distribution to derive p-values.
Validation and Visualization: Scatter plots were generated to illustrate the relationships between cluster proximity, degree centrality, and sales rank. (Figures 11, 12, 13)

5.14.2 Results:

- **Helpful Reviews:** The Spearman correlation of -0.2499 ($p < 0.001$) suggests that products with a higher number of helpful reviews tend to have better sales ranks. (Figure 11)

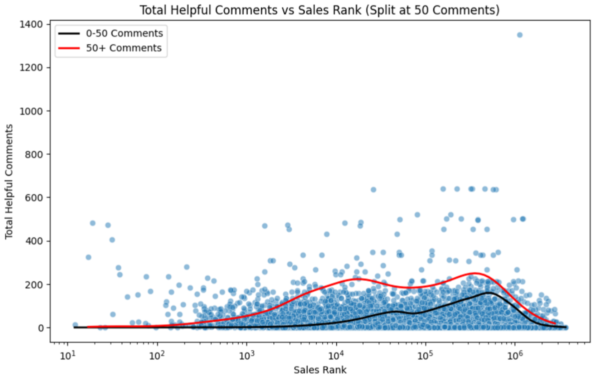

Figure 11: This plot illustrates that, particularly after refining the hypothesis, an increase in the number of comments is associated with better sales performance.

- **Degree Centrality:** The correlation of -0.3403 ($p < 0.001$) indicates that products with higher network connectivity tend to perform better in terms of sales. (Figure 12)

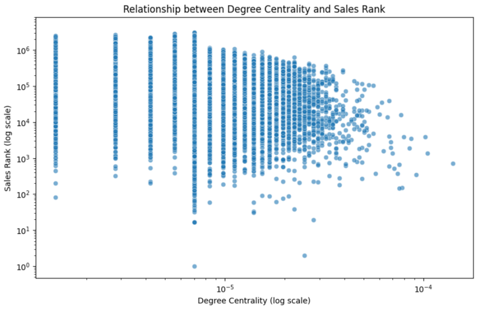

Figure 12: This plot demonstrates that as the Degree Centrality increases, the likelihood of achieving a better sales rank decreases.

- **Shortest Distance to Cluster:** The positive correlation of 0.3150 ($p < 0.001$) confirms that products closer to densely clustered regions perform better in terms of sales. (Figure 13)

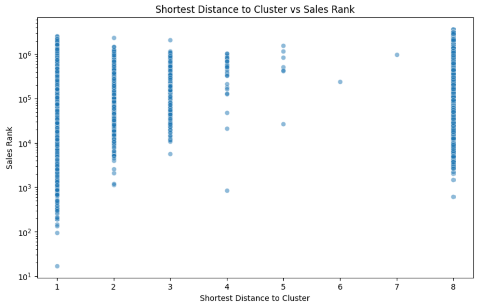

Figure 13: This plot demonstrates that as the distance from the cluster increases, the likelihood of achieving a higher sales rank decreases. Specifically, nodes with a distance greater than 7 are grouped under distance 8.

The statistically significant p-values indicate that these relationships are not due to chance. Products near highly clustered regions performed better, confirming our prediction. Additionally, higher degree centrality and more helpful reviews were linked to improved sales ranks, supported by the permutation test results.

6 Conclusion

This study analyzed the Amazon metadata through network analysis, revealing key insights into customer-product interactions and their impact on sales. Our findings are structured around several key themes:

- **Identifying Influential Products and Customers:** Degree centrality analysis highlighted products like the Harry Potter series as highly central, attracting a broad range of customer interactions and suggesting their potential as best-sellers. High-centrality customers were identified as exceptionally active reviewers, significantly influencing product visibility. Betweenness centrality further identified 17 “broker” nodes, comprising both products and customers, that bridge otherwise disconnected segments of the network. These nodes are crucial for cross-promotional strategies and enhancing network cohesion.

- **Detecting Community Detection and Network Cohesion:** Application of the Louvain algorithm revealed 41 distinct communities within the network, characterized by a high modularity score of 0.9295. This indicates well-defined clusters, with Community 4 focusing on investment-related products and Community 0 centered around hip-hop music and its consumers as examples of communities. The high modularity suggests that targeted marketing and specialized recommendation campaigns within these communities could be highly effective.

- **Uncovering Dense Purchase Patterns:** Clique analysis uncovered 1316 maximal cliques, though the largest clique size was limited to 3. This suggests a network structure dominated by small, tightly-knit groups rather than large, cohesive clusters. K-core decomposition further illustrated this structure, revealing a small core of highly interconnected customers and products (4-core and 5-core) surrounded by a larger, less connected periphery (1-core). These findings suggest opportunities for multi-product promotions and strategies to foster deeper connectivity within the network.

- **Handling Fragmentation and Isolated Segments:** Component analysis identified a large connected component of 709 nodes, alongside several smaller components and isolated nodes. These smaller fragments often represent niche products or less-engaged customers. Integrating these segments into the main network through cross-promotion or personalized recommendations could enhance overall network engagement and activity.

- **Tracking Evolution over Time:** Temporal analysis spanning from 1999 to 2005 revealed shifting engagement patterns, with some products experiencing surges in reviews and prominence while others declined. These fluctuations underscore the dynamic nature of customer preferences and the importance of real-time network monitoring for effective inventory management and marketing strategies.

- **Linking Network Structure to Business Outcomes:** Correlation analyses, including Spearman and Pearson, alongside permutation testing, demonstrated that higher degree centrality, more helpful reviews, and closer proximity to highly clustered product regions correlate with improved sales ranks. Homophily analysis confirmed the bipartite nature of the network, while small-worldness measurements indicated efficient information flow despite relatively low clustering. These findings highlight the importance of maximizing network engagement and leveraging social proof to enhance sales performance.

In summary, this study demonstrates that products and customers in key network positions (high centrality, tight communities, strong cliques) or closer to influential clusters generally exhibit better sales performance. The presence of fragmented segments highlights opportunities for targeted interventions to enhance network synergy. By integrating structural insights from network analysis with predictive analyses, businesses can develop more effective recommendation systems, marketing strategies, and long-term growth plans. The combination of these approaches provides a robust framework for understanding and optimizing customer-product interactions within the Amazon marketplace.

7 Critique

The analysis is limited by available metadata, excluding external factors like social media trends or seasonal effects that could enhance understanding of sales performance. The focus on centrality and community detection may miss key factors such as product quality, pricing strategies, and customer sentiment, which could impact the findings' robustness.
Conduct separate tests for low, medium, and high-quality products to assess how comments, similarity, and proximity influence sales rank at various quality levels.
Subgraph analysis is useful but can result in biased or incomplete network views. Future work may use advanced sampling techniques or high-performance computing to explore larger network segments. Incorporating machine learning models can uncover non-linear relationships and boost predictive power.
In conclusion, the findings on network dynamics are promising but could be improved with recent sales data or experiments to test marketing strategies. This would strengthen evidence for applying findings and connect theory with actionable business recommendations.

8 Reference:

1-Leskovec, Jure, Lada A. Adamic, and Bernardo A. Huberman. "The dynamics of viral marketing." ACM Transactions on the Web (TWEB) 1, no. 1 (2007): 5-es.
2-https://snap.stanford.edu/data/amazon-meta.html

3-Newman, M. E. J. (2018). Networks. Oxford University Press.