Amazon Product Co-Purchasing Network Analysis

Mu Niu, Kai Wa Ho, Jingtang

2024-06-07

Introduction

In this project, we conduct a comprehensive social network analysis on a dataset derived from Amazon's product co-purchasing records. The raw edge data, which was collected on June 1, 2003, from the Amazon website, includes 403,394 nodes representing individual products and 3,387,388 unweighted directed edges indicating frequently co-purchased product pairs. The direction of an edge signifies the order of purchase, with an outward edge from product A to product B indicating that product B is frequently purchased after product A. This dataset is available at **Amazon Product Co-Purchasing Network.**

Additionally, the raw nodes data was collected by crawling Amazon's website and encompasses metadata and review information for 548,552 different products, including books, music, DVDs, and video tapes. This dataset provides detailed information such as the product title, sales rank, list of similar products, detailed product categorization, and reviews. This data was collected in the summer of 2006, and it can be accessed at **Amazon Product Metadata.**

After a thorough data cleaning process, we integrated the edge data and node information data, resulting in an igraph object containing 398,688 nodes. Each node has four attributes: ID (unique product ID), Title (product name), Group (class the product belongs to), and Category (subcategory of the class).

The primary objectives of this analysis are to understand the relationships between products, identify co-purchasing patterns and customer shopping behaviors through network data visualization, and explore various network, node, and edge metrics. We aim to interpret these metrics within the network context, apply community detection algorithms, and analyze the resulting communities. Additionally, we will examine the network's adjacency matrix, reorder vertices to highlight patterns, and observe changes in the matrix to identify clearer patterns based on community or connected components.

Methodology

In this project, we followed a structured approach to process the data, analyze the network, and visualize the results. The data processing procedures began with mapping the nodes information dataset, which was a text file with issues of missing data and inconsistent formats, into a dataframe with four columns: ID, Title, Group, and Category. Subsequently, we filtered out all edges in the edge dataset that connected nodes with missing IDs, product titles, or product groups. This ensured that only valid and complete data was included in the analysis. The mapped nodes information data and filtered edge data was then integrated into an igraph object containing 398,688 nodes with four vertex attributes: ID, Title, Group, and Category. The data cleaning process employed packages such as `dplyr` for data manipulation and `igraph` for constructing and managing the igraph object in R.

For sampling methodology, we generated 4 induced subgraphs by using a node-centric approach. For each sub-network, we randomly selected one node as the starting point and retrieved all nodes connected to it. We then iteratively expanded our sample by including nodes connected to the current set of nodes, continuing this process until the specified number of nodes was reached. This sampling method was designed to capture densely connected nodes, facilitating better visualization and understanding of the relationships between products.

This comprehensive methodology enabled us to effectively analyze and interpret the Amazon product co-purchasing network, uncovering valuable insights into customer shopping behaviors and product relationships.

Analysis

Description and Visualization of Subgraphs:

In the study of co-purchase networks for categories such as Music, Book, Video, and DVD, we began by visualizing subgraphs representing these networks. Using tools like the Plotly package in R, we generated both static and interactive plots to reveal the co-purchasing patterns among products within these categories. For instance, the music subgraph highlighted dense connections among various genres, suggesting demographic overlaps. Similarly, the book subgraph showed strong links between genres like non-fiction and history. These visualizations allowed us to identify distinct clusters and understand the relationships between different types of products, offering insights into customer preferences and potential cross-selling opportunities.

We further analyzed these subgraphs to uncover deeper insights into customer behavior. For example, in the DVD subgraph, we noticed that customers who bought comedy DVDs also tended to purchase a wide range of music genres. This observation highlights the diverse tastes of these customers and suggests opportunities for targeted marketing strategies. By examining these visualizations, we gained valuable knowledge about how products are interconnected through customer purchases, which can be leveraged to enhance product recommendations and marketing efforts.

Metrics and Community Detection Results:

To understand the structural properties and influential nodes within each network, we examined key network metrics like density, path length, degree distribution, and centrality measures. The Music network, with a density of 0.237, showed moderate connectivity with frequent co-purchases, and its influential nodes were identified through high betweenness and eigenvector centrality values. The Book network was more dispersed, with a density of 0.137, indicating fewer direct connections but significant roles for a few key nodes.

Propagation and Community Influence in Co-Purchase Networks: The propagation of recommendations in viral marketing, as detailed in the paper, aligns closely with the co-purchase behaviors observed in our analysis of the Amazon product networks. In the Music network, nodes like "Movimiento Music" act as hubs, driving further purchases and forming dense co-purchase chains. These clusters, detected through community detection algorithms, mirror the tightly-knit groups described in viral marketing, where frequent co-purchases create robust communities. Understanding these propagation dynamics helps in identifying how recommendations can spread effectively through co-purchase networks, enhancing marketing strategies and product recommendations.

Applying community detection algorithms such as Walktrap, Infomap, and FastGreedy provided a more nuanced view of the network's structure. These algorithms revealed how nodes grouped into communities, with Walktrap and Infomap detecting smaller, tightly-knit communities, while FastGreedy captured broader structures. The Music network's modularity scores indicated moderately strong community structures, with nodes forming clear clusters around influential hubs. In contrast, the Book and DVD networks showed a more complex interplay of smaller, detailed communities and larger, overarching ones.

Interpretation of Adjacency Matrix Reorderings:

Reordering the adjacency matrices based on hierarchical and community structures allowed us to visualize the organization of the networks more clearly. For the combined network, the reordered matrices highlighted dense intra-network connections within each category and minimal inter-network interactions. This finding aligns with the inherent properties of each network, showing that products within the same category are more likely to be co-purchased together.

Examining the Music network's reordered matrices, we observed five distinct clusters centered around influential nodes. These clusters represent subgroups where specific genres or types of music are frequently co-purchased. The reordering also made apparent how community detection algorithms like Walktrap and FastGreedy highlight different aspects of the network's structure. While Walktrap captured smaller, cohesive communities, FastGreedy's hierarchical approach revealed broader groupings. These insights underscore the importance of choosing the right algorithm based on the analysis goals, whether for detailed community detection or broader structural understanding.

Results:

Key Findings from the Analysis:

The analysis revealed that the networks for Music, Book, Video, and DVD categories each have unique structural characteristics and influential nodes. The Music network, with its high density and short average path length, emerged as the most interconnected, highlighting robust co-purchase relationships within this category. In contrast, the Book network, though more dispersed, showed the critical role of a few key nodes in maintaining its connectivity. The Video and DVD networks were moderately dense and compact, with influential nodes that played significant roles in their community structures.

Practical Applications for Marketing Strategies: Insights from the viral marketing study are particularly relevant for devising marketing strategies in the context of our co-purchase network analysis. The dense clusters and influential nodes identified in networks like Music and Book suggest potential for targeted marketing and product bundling. For example, leveraging the strong intra-community ties and rapid co-purchase propagation can inform effective promotional campaigns. The temporal dynamics observed, such as the short average path length in the Music network, highlight opportunities for timely and impactful product recommendations, reflecting the rapid influence spread seen in viral marketing.

Community detection across these networks uncovered distinct clusters that reflect the underlying co-purchasing behavior. Algorithms like Walktrap and Infomap identified tightly-knit communities within each network, while FastGreedy highlighted broader, less distinct groupings. Notably, the consistent appearance of certain clusters across different algorithms indicates the stability and robustness of these networks' structures. These findings provide valuable insights into how products are linked through customer purchases, offering potential avenues for targeted marketing and product recommendation strategies.

Visual and Textual Summaries:

Visual representations, including adjacency matrices and network graphs, effectively captured the structure and dynamics of the co-purchase networks. The matrices, both in their original and reordered forms, highlighted dense intra-category connections and revealed how community detection algorithms define clusters. For example, the reordered matrix of the Music network showed clear clusters centered around influential nodes like "Movimiento Music," reflecting cohesive groups of frequently co-purchased items. These visualizations were complemented by detailed network graphs that depicted the central roles of key nodes and the intricate patterns of connections within each network.

Textual summaries provided contextual explanations for these visual findings. They elaborated on how the identified clusters correspond to actual customer behavior, such as the diverse musical tastes linked to comedy DVD buyers or the demographic overlaps in book and music purchases. The textual analysis also discussed the practical implications of these insights, suggesting how understanding the network's structure can enhance marketing strategies and optimize product recommendations. Together, these visual and textual summaries offered a comprehensive understanding of the networks' dynamics, highlighting their potential for influencing customer behavior and driving sales.

Conclusion

We analyzed the Amazon product co-purchasing network to understand its structure and dynamics. Initially, we looked at the basic characteristics of the network, which showed moderate density with significant co-purchasing connections among music products. This setup allows us to see clear patterns and clusters, indicating that while many products are connected, they're not all directly linked.

Our study used community detection algorithms like Walktrap, Infomap, and FastGreedy to identify community structures within the network. Walktrap and Infomap were particularly good at finding smaller, closely-knit groups, showing specific buying behaviors among certain products. On the other hand, FastGreedy helped us see larger community structures, providing a broader view of how the network is organized.

We also used visualizations like reordered adjacency matrices to illustrate the network's dense connections and the formation of communities around frequently bought items. These visual tools confirmed that our community detection methods were effective and that the network structure was both robust and stable.

Our analysis of the Amazon product co-purchasing network provides valuable insights for enhancing marketing strategies. We identified clusters and key products that illustrate how recommendations and purchases propagate through the network. This information is crucial for targeting marketing efforts effectively and optimizing product recommendations. The study underscores the importance of understanding community dynamics within these networks, offering essential guidance for developing targeted marketing approaches and improving recommendation systems in e-commerce and marketing.

References

Dataset:

- **Amazon Product Co-Purchasing Network**: The edge dataset representing products and their co-purchasing patterns on Amazon.
- **Amazon Product Metadata**:Metadata for products listed in the Amazon Product Co-Purchasing Network.

Academic Reference:

- J. Leskovec, L. Adamic, and B. Adamic. The Dynamics of Viral Marketing. ACM Transactions on the Web (ACM TWEB), 1(1), 2007.

Further Elaboration:

Since we do not have enough space in this report, further analysis and interpretation can be accessed via this **link**

Acknowledgments:

We extend our heartfelt thanks to Isaiah Katz and Dr. Uma Ravat. Their assistance was invaluable, and we couldn't have completed this project without their support.