A Deep Dive into the Network of Book Co-Purchases on Amazon

Ammar Bagharib A0218111X
Chen Yang A0201905R
Ng Qing Yu Randy A0199577B

National University of Singapore AY2023/24 Semester 1

(*Github Repo Here*)

Contents

1 Introduction **2**

2 About the Dataset **2**
2.1 Network Sampling 2

3 Network Structure **3**
3.1 Degree Distribution 3
3.2 Density and Transitivity 4
3.3 PageRank 4

4 Network Metrics on Amazon's Recommendation Engine **6**
4.1 Objective of our Models 7
4.2 Methodology 8
4.3 Findings 9
4.4 Further Analysis: Recency of Data 10

5 Conclusion **11**

6 Limitations and Considerations **11**

7 References **12**

1 Introduction

Established in 1995 as a pioneering online bookstore, Amazon has been a cherished destination for book enthusiasts for several decades. While the company has evolved and expanded into a diverse range of industries, it has consistently maintained its status as a go-to hub for book shoppers.

In the scope of this project, our objective is to delve into the evolving co-purchase patterns of books over time, understand it's characteristics and underlying patterns. With this information, we're looking to identify key products which drive purchasing behaviour among consumers, devise marketing strategies, as well as to assess the efficacy of incorporating network analytics into the development of product recommendation engines.

In doing so, we anticipate that our efforts will yield a mutually beneficial outcome. By facilitating book discoveries and guiding readers to titles they might not have encountered otherwise, we aim to stimulate sales on the platform, benefiting both Amazon and its thriving community of book lovers.

2 About the Dataset

In our project, we used the amazon co-purchase dataset from Stanford University's SNAP library. The dataset was compiled through an extensive crawl of the Amazon website in the summer of 2006, capturing a vast array of product metadata and review details for 548,552 diverse items, spanning categories such as Books, music CDs, DVDs, and VHS video tapes. It encompasses comprehensive information for each product, including titles, sales ranks, and genre/category information. In our analysis, we restricted the product scope to books only.

Specifically, we delved into the structure of the network as of March 2003 and developed our model around it. In our network, each node represents a book and each undirected edge represent the co-purchase relationship between two books. To supplement our analysis, we extracted the product ID, name, genre, product rank and review count information for all books to be included as node attributes.

2.1 Network Sampling

2.1.1 Methodology:

To enable us to compute some of the key network metrics efficiently, we opted to sample 100,000 nodes from the network. This approach offers a practical means of analysis while retaining a substantial portion of the network's complexity. Steps taken in our sampling process are as follows:

Step 1:

Load graph, simplify and remove multiple loops, and set it as undirected. Filter for nodes with degree > 1, and set this as the graph we'll be working with.

Step 1:

Filter for books which have reviews > 1 per 1.25 months. The time period is defined as from when the co-purchases network was derived e.g., March 2003, to 7 August 2006, which is the date of collection of the metadata file.

Step 2:

Filter for books which have valid genres i.e., non-empty strings

Step 3:

Filter for links in which both 'from' and 'to' books follow the conditions outlined in the steps 1 and 2

Step 4:

Extract the indices of nodes from step 3, lets define this as nodes_1, and get the count of nodes. Let's define the counts of these nodes as `x`

Step 5:

Get count of remaining nodes needed to sample: y = (100,000-x)

Step 6:

Sample y indices of nodes from setdiff(1:length(V(g)), nodes). Let's label these nodes as nodes_2

Step 7:

Combine nodes_1 and nodes_2 indices and obtain subgraph from g using these indices.

Step 8:

Generate network metrics for subgraph, after which we filter only for the nodes present in nodes_1, and save it in a data frame.

As we can see in Figure 1, the subgraph formed by our sampled nodes in step 7 exhibits a degree and local transitivity distribution that closely mirrors those of the original, complete network. This congruence strongly suggests that the sampled network effectively captures the essential structural characteristics of the larger network.

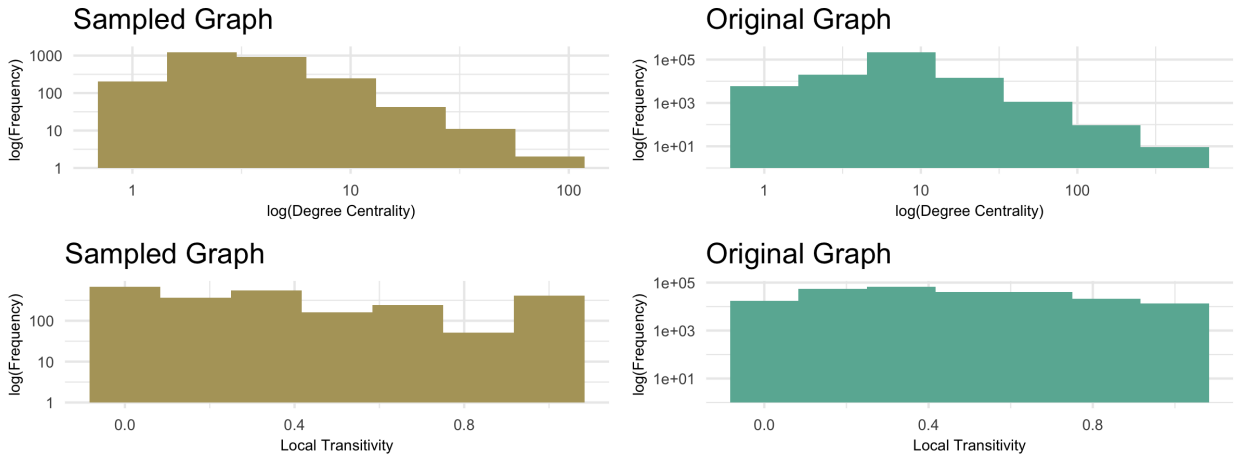

Figure 1: Degree and Transitivty Distributions in Sampled vs Overall Graph

By selecting a significant subset, we aim to strike a balance between computational efficiency and preserving the network's fundamental properties. This method allows us to derive meaningful insights and draw reliable conclusions about the overall network dynamics without the computational burden of analyzing the entire network.

3 Network Structure

We performed some basic analysis to draw some insights on the purchasing pattern of books on Amazon.

3.1 Degree Distribution

As we can see in Figure 2, the degree distribution of our network exhibits a power-law distribution, a characteristic often seen in naturally occurring networks.

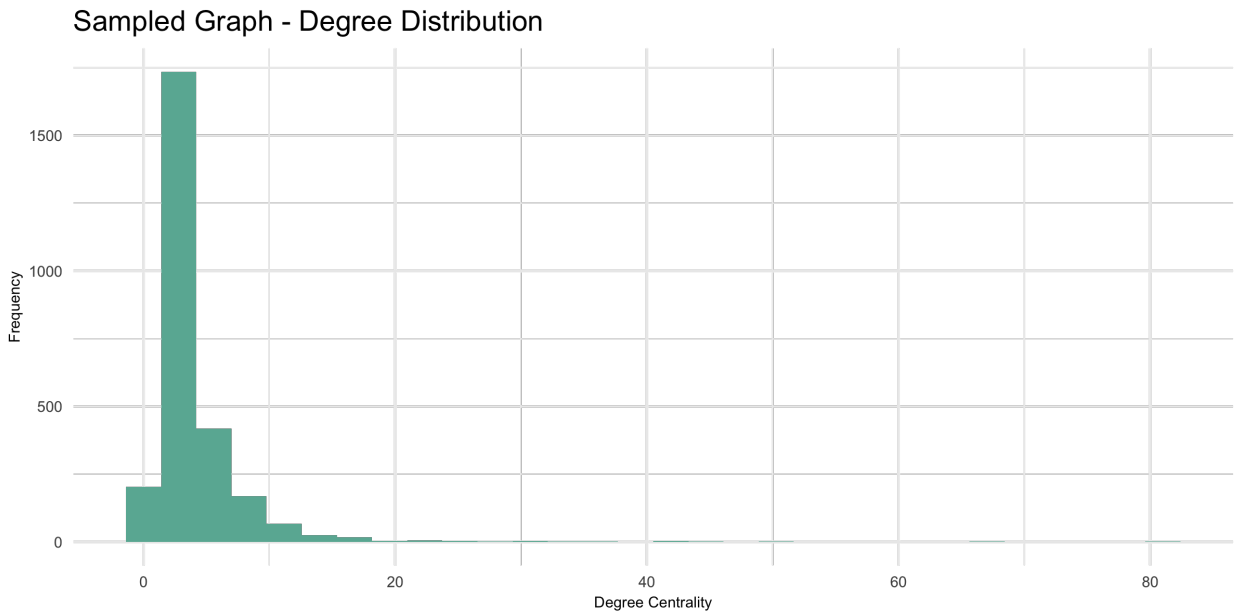

Figure 2: Degree Centrality Distribution

Majority of books in our network have low degree centrality, with approximately 3 or fewer connections. On the far right of the distribution, we observe a small group of books with approximately 70 first-degree connections, indicating a significant influence on the purchase of other books. This suggests that most books within the network neither strongly influence the purchase of other books nor are significantly influenced by other books.

3.2 Density and Transitivity

In a co-purchase network, the density and transitivity metrics offer valuable insights into the interconnectedness of products through customer co-purchases. With a global transitivity of 0.233, our co-purchase network demonstrates a moderate level of clustering tendency, indicating that some books do tend to form clusters in terms of co-purchase behaviour as seen in Figure 2.

Table 1: Density and Global Transitivity

Metric	Value
Global Transitivity	0.2327958
Density	0.0000277

Simultaneously, the network exhibits a low density of 0.0000277, signifying a sparse web of connections between products. This low density indicates that even though a small number of clusters exist, the overall network is very sparse.

Taken together, this means that books within the network are not frequently co-purchased together, and individual products seem to have limited influence on the purchasing behaviour of others. This disconnectedness further highlights the challenge in seeking out cross-selling opportunities, as there are few evident connections or influences between various types of books.

3.3 PageRank

We observed a noteworthy similarity between the distribution of PageRank and degree centrality in our network as seen in Figure 3. Both distributions exhibit a power-law pattern, wherein a small number of

books boast significantly higher PageRank scores than the rest.

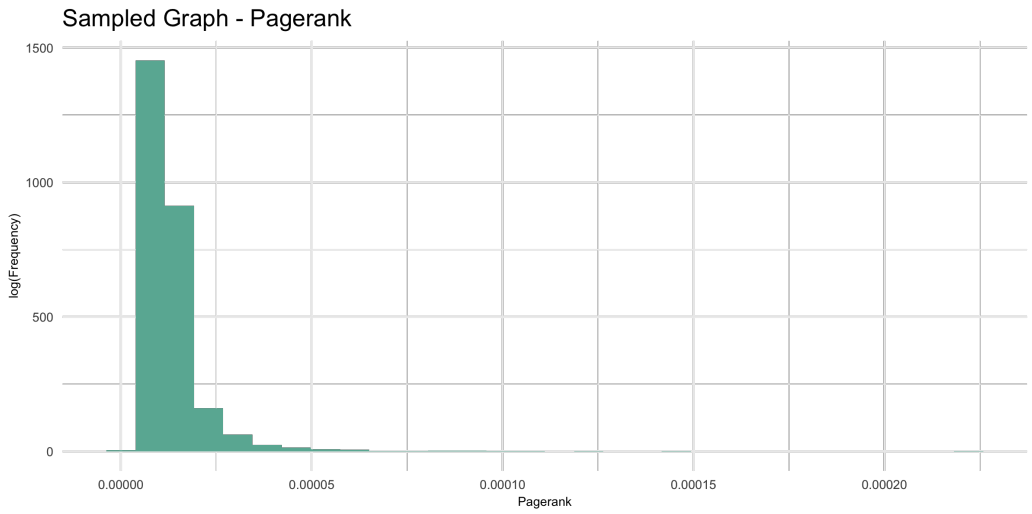

Figure 3: Pagerank Distribution

High PageRank indicates that a given book enjoys strong interconnectedness and popularity within our co-purchase network. While the highest PageRank score in our network is modest at 0.000231, the existence of a select few titles with distinctly higher PageRank suggests that a small subset of books still wield considerable authority and influence over copurchase behaviours.

Building on this insight, we propose a strategic initiative to prominently feature these high PageRank books in our homepage recommendations. For instance, Amazon can promote the top-10 books in descending PageRank as seen in Figure 3, in a carousel on the homepage. This allows Amazon to leverage the influence of these books and foster a more interconnected and dynamic co-purchase network on the platform, thereby improving sales.

Table 2: Top 10 Books by Pagerank

vertex_name	pagerank
5765	0.0002256
37780	0.0001477
56817	0.0001233
2555	0.0001090
19526	0.0000958
2563	0.0000933
2423	0.0000921
976	0.0000881
3247	0.0000867
36122	0.0000832

The popularity of books, as determined by PageRank, may be attributed to their association with specific genres that inherently possess broad mass appeal. To explore this hypothesis further, we tabulated the count of books within each genre among the top 100 PageRank books. Subsequently, we organized the genres in descending order based on book count as seen in Table 2. We excluded the first two genres which are very broad and offer no analytical value (“General”, “Authors, (A-Z)”, “Literature & Fiction”).

Table 3: Top 10 Genres by Count

Genre	Count
Science Fiction & Fantasy	14
Biographies & Memoirs	12
Contemporary	12
Nonfiction	12
Series	11
Fantasy	10
Mystery & Thrillers	10
Books on Tape	9
Children's Books	9
Genre Fiction	8

Sci-Fi & Fantasy emerged as the most popular genre. This is likely due to their unique ability to transport readers to imaginative worlds, offering an escape from reality and fostering a sense of wonder. These genres encourage creativity, exploration, and reflection on societal issues, with complex characters and timeless themes that resonate across cultures and generations. The genres' success could have been further fueled by strong fan communities, mainstream media adaptations, and the emotional engagement derived from characters facing extraordinary challenges. In essence, the allure of Sci-Fi and Fantasy lies in their capacity to captivate readers through boundless imagination, adventure, and a sense of belonging to a vibrant literary community.

In second place is Biographies & Memoirs, possibly driven by a profound interest in real-life stories and the human experience. Readers are drawn to the compelling narratives that chronicle the lives of individuals, offering insights into personal journeys, challenges, and triumphs (Knox, 2023). The genre's appeal lies in its ability to provide inspiration, lessons, and a genuine connection to the diverse and authentic experiences of others. Biographies & Memoirs serve as a powerful medium for readers seeking to understand, empathize, and find common ground with the remarkable lives detailed within the pages, making the genre a compelling and enduring choice for a wide audience.

We also noticed that "Books on Tape" is rather low on the list. Given that the dataset was from 2006, this does not come as a surprise since smartphones weren't as common back then. Audiobooks are convenient mediums which allow users to absorb literary content effortlessly while on the move or while performing mundane chores. Given the hectic lifestyles of many readers today, we expect it to be higher up on the list today, since it allows readers to "read" on the move.

The insights derived from the analysis of PageRank popularity among book genres not only shed light on the overarching trends in readership preferences but also provide actionable strategies for platforms like Amazon. Understanding that certain genres, such as 'Science Fiction & Fantasy' and 'Biographies & Memoirs,' dominate the PageRank metrics due to their broad appeal, Amazon can leverage this information to enhance its recommendation system. By tailoring book suggestions based on users' demonstrated interests, Amazon has the opportunity to elevate user experience and engagement. For example, recognizing a user's penchant for biographies, Amazon can deploy a refined recommendation algorithm akin to the concept of topic-sensitive PageRank (Haveliwala, 2003), ensuring that users receive personalized suggestions aligned with their literary preferences. This strategic approach not only acknowledges the popularity hierarchy of genres but also translates it into a more individualized and enjoyable browsing journey for users.

4 Network Metrics on Amazon's Recommendation Engine

Considering Amazon's position as a leading tech company renowned for employing diverse cutting-edge techniques and algorithms within its recommendation engine, one noteworthy approach utilized is the integration of network analytics in its item-to-item collaborative filtering system (Muralidharan, 2023). However, as outsiders, we are unsure of the impact which network metrics have on a predictive model's performance – and

the second goal of our project is to answer this question. To further narrow down our scope of analysis, we decided to only focus on books which have at least 1 review per month and valid genre information. We hypothesize that the incorporation of network analytics is a valuable method capable of enhancing predictive accuracy. To test this hypothesis, we have developed two models for the purpose of link prediction.

The first model type serves as a baseline, incorporating solely item-specific features including genre, number of reviews, and item ranking. The second model type, an alternative approach, uses only network metrics such as degree centrality, pagerank centrality, community etc., just to name a few. Our aim is to assess and compare the predictive performance of these models to determine the potential impact and effectiveness of incorporating network-based metrics within the recommendation system.

Our algorithm of choice is a logistic regression. We believe that it is the most appropriate model because our problem at hand is binary in nature, and logistic regression outputs predicted probabilities that can be rounded to 0 or 1 based on some criteria. Additionally, logistic regression is not a 'black-box' model where we cannot see the training process, but rather we can fully understand the estimation process as it typically uses maximum likelihood estimation (MLE). Lastly, the coefficients of the logistic regression are also easily interpretable, where we can infer and gauge the contribution of each independent variable to the dependent variable.

4.1 Objective of our Models

The primary objective for our model was to determine whether the network metrics of the books serve as better predictors of co-purchases, when compared against book-specific features.

Baseline Features:

These features represent the book-specific information obtained from August 2006.

1. *num_common_genre*: This refers to the number of common genres in the entire genre string between 2 books. We do not make any distinction between possible main vs sub-genres as we did not have access to such data.
2. *mean_salesrank*: This refers to the average salesrank of 2 books, collected from the metadata file
3. *mean_reviews*: This refers to the average number of reviews left by users on Amazon, collected from the metadata file
4. *mean_ratings*: This refers to the average rating collected from the metadata file. Ratings are on a scale of 1-5.

Network Metrics

All the following network metrics of the books, would be collected from Amazon's product co-purchasing network, at a specific point in time. For example, from the Amazon product co-purchasing network from March 2 2003 *amazon0302.txt*

1. *mean_degree*: The average degree centrality of 2 books
2. *mean_betweenness*: The average betweenness centrality of 2 books
3. *mean_eigen*: The average eigen centrality of 2 books
4. *mean_transitivity*: The average local transitivity of 2 books
5. *mean_closeness*: The average closeness centrality of 2 books
6. *mean_pagerank*: The average pagerank of 2 books
7. *mean_eigen*: The average Eigenvector centrality of 2 books
8. *same_community*: Whether or not the 2 books belonged to the same community.

For the feature *same_community*, we tried Louvain against the Greedy community detection algorithms but ultimately decided to go with the Louvain method as it resulted in clusters with higher modularity scores (i.e., better-defined cummunity structure).

4.2 Methodology

First, we exhaustively created all conceivable pairings among the books within our network. After this stage, we specifically isolated connected pairs to serve as our positive data points, while the unconnected pairs were designated as our negative data points. Each data point captured information such as the number of shared genres between the paired books, along with individual book attributes including sales rank and review counts.

4.2.1 Downsampling of Negative Data Points

However, the disparity in numbers between the negative and positive data points, mainly owing to the sparse nature of our network, posed a significant challenge. To address this imbalance, we adopted a strategic approach, opting to randomly down-sample the excessive negative data points, thereby achieving a balanced representation of positive and negative instances. We opted for downsampling instead of upsampling based on 2 considerations, with the first being downsampling requiring much less computational resources. Secondly, upsampling methods such as Synthetic Minority Oversampling Technique (SMOTE), which generates new samples of the minority class based on an unsupervised machine learning method, is not appropriate. This is because the counts of positive class are too little, which would render the algorithm to not be able to learn properly, and thus, also disrupt the overall network structure. Therefore, we believe that downsampling is the appropriate choice. Subsequently, the datasets have an equal proportion of positive and negative classes.

Given the downsampling of our negative data points, the conventional use of metrics such as F1 score, precision, and recall became inappropriate, as the dataset no longer accurately represented the true distribution of positive and negative instances. Consequently, we directed our focus solely towards evaluating the predictive accuracy concerning the positive cases in our model, recognizing the need to adapt our evaluation approach in response to the modified dataset composition.

4.2.2 Handling of NA values

Handling of NA values are essential in ensuring that model training and model prediction can occur. In the training and the test datasets, with regards to networks metrics, namely local transitivity and closeness centralities had NA values. It is not unreasonable for there to be NA values because it means that it is an isolated node. As such, we decided that the most appropriate value to represent NA values with the value: –1. This is because transitivity is a numerical variable with values that are usually positive. Thus, to not confound with the numeric meaning of typical transitivity and closeness values, we believe that the number –1 will still be able capture the essence of the datapoints with 'NA' values.

4.2.3 Scaling of data

Scaling features before fitting a logistic regression model is crucial because it ensures that each feature contributes equally to the determination of the outcome. Logistic regression uses gradient descent as an optimization technique, which converges much faster when features are on similar scales. Without scaling, features with large values can disproportionately influence the model and result in an unstable training process, where the algorithm might oscillate and take a long time to find the global minimum, or potentially not converge at all. Thus, we used to use standard scaling for our dataset, which subtract the mean of the variable from each datapoint then divide it by its standard deviation. This ensures that each variable has a mean of 0 and a standard deviation of 1. There are alternative forms of scaling such as min-max scaling, but we believe that standard-scaling is sufficient for our case.

4.2.4 Train-Test Split

To construct our testing set, we did not utilise the traditional train-test-split approach. Instead, we gathered node pairs that formed **only new connections** between March and June. In essence, these pairs represent books that established co-purchase relationships only during June. This was achieved by getting all node pairs which are connected in June and further filtering out pairs already linked in March. From the newly formed links in June, we then assign the nodes' network metrics based on that of March. Meaning, although our test set was newly formed links in June, we still utilised network metrics in March, as we believed it would be counter intuitive to use the network metrics of June. For the sake of our hypothesis, we believe regarding the test set: the the network structure of the co-purchases have not been formed yet. We're essentially looking at the perspective of the recommendation engine and attempting to see whether our the past network metrics could predict a co-purchase.

4.3 Findings

Just to re-emphasize, we want to test the hypothesis that network metrics serve as better predictors than item-specific information. Consequently, to test this hypothesis, the first model which is based on the March *0302* dataset, only contains book-specific information, and the second model contains only network metrics.

We then compare the predictive performance, evaluated using accuracy as the metric, of both models against that of the newly formed links in June *0602* dataset to determine the outcome of our hypothesis.

Table 4: Regression Results

Model	Accuracy
baseline 0302	0.4491811
network metrics 0302	0.4932563
baseline + network metrics 0302	0.4879576

As seen from Table 4 above, we first refer to the first two rows, which is in direct relation to our hypothesis. Although both model's accuracy is less than 50%, which suggests that it is worse than a coin-flip, our hypothesis is still correct in that model two's accuracy is higher than model one's accuracy, by approximately 5%. Plausible explanations for the low accuracy could be that the model was not tuned, where we merely used a rule of thumb in that predicted probabilities above the value 0.5 was rounded to 1, and 0 otherwise. Nonetheless, these results are indicative that network metrics serve as better predictors than item-specific information.

Thereafter, we combined the 2 sets of information for a third model that makes use of both item-specific information and network metrics, as seen in row 3 in table 5 above. Although the performance was better than model 1, it was worse than model 2. A plausible explanation is that overfitting occurred due to there being too many independent variables, thus producing lousier out-of-sample performance. One solution to this is to perform regularization in the form of L1-norm regularization, but due to the time constraints and direction of the project, it was put on hold and can be researched in the future as an extension.

4.4 Further Analysis: Recency of Data

We further hypothesize that incorporating more recent network metrics' information could yield improvements in the model's performance. To test this hypothesis, we opted to utilize network metrics from May *0505* dataset. Consequently, we re-trained our model, anticipating that this adjustment would provide the algorithm with a more current and nuanced understanding of the network's dynamics. The testing data also followed suit, utlising network metrics of nodes from May, instead of March as was done in the first hypothesis.

Table 5: Added Regression Results

Model	Accuracy
baseline 0302	0.4491811
network metrics 0302	0.4932563
baseline + network metrics 0302	0.4879576
network metrics 0505	0.6180154

As seen in Table 5, the accuracy of the model improved significantly to approximately 62%. This is a marked improvement because now the model predictive accuracy is better than a random coin flip and is usable. Thus, this re-affirms the notion that networks are constantly evolving, and thus it is essential to use the most recent information in model training to achieve the best results possible.

Table 6: Comparison of Regression Coefficients using March and May Network Metrics

V1	network_0302	network_0505	change
mean_closeness	-0.2837302	-0.1077686	Increase
mean_degree	0.0284803	1.9621518	Increase
mean_transitivity	-0.5349795	-0.5431613	Decrease
mean_pagerank	0.1882880	-1.6621105	Decrease
mean_eigen	0.2143290	-0.0466116	Decrease
mean_betweenness	0.2506547	0.4525081	Increase
same_community	3.6142644	4.1384405	Increase

As seen in Table 6 above, the variable *same_community* emerges with the most substantial coefficient, signifying its pivotal role in predicting link formation. The positive coefficient indicates that node pairs within the same community, as identified by the louvain community detection algorithm, are more inclined to establish connections in the future. Notably, the magnitude of this coefficient increased over time (between March and May of 2006), suggesting that, as the network grows, there is an increasing tendency for connections to form among members of the same community. This underscores the escalating significance of community cohesion in shaping linkages within the evolving network.

In contrast, the variable *mean_pagerank* exhibits a negative correlation with the log-odds of a co-purchase relationship between two books. This finding diverges from the anticipated outcome, as one might intuitively expect a higher mean pagerank to positively influence the likelihood of a linkage between the paired books. One possibly reason could be that books with higher pageranks may belong to different genres or cater to diverse interests. Readers with an interest in one highly ranked book may not necessarily have the same preferences for another book, reducing the likelihood of a co-purchase.

As for the drastic change in coefficient of *mean_pagerank* between March and may, one possible reason could be the introduction of new books or changes in literary trends between March and May, which could have altered how readers make purchasing decisions. New releases may attract different audiences with distinct preferences, affecting the co-purchase patterns. Alternatively, there could also be seasonal trends influencing book purchasing behavior. Preferences, reading habits, or marketing strategies might vary between March and May, leading to different patterns of co-purchase relationships. However, we'll require extensive data spanning a much longer time horizon to test these hypotheses.

5 Conclusion

In conclusion, our project has underscored the transformative potential of network analytics in reshaping marketing strategies. By leveraging tools like PageRank, marketers can effectively identify influential products affecting consumers' copurchasing behaviour. This insight allows for the formulation of effective sales strategies, such as prominently featuring high PageRank products in strategic locations frequented by consumers.

Our analysis of the baseline model, reliant solely on product-specific information, and the alternative model integrating network metrics also underscored the significant enhancements attainable through the incorporation of the latter. The infusion of network metrics imparts a wealth of additional insights into the intricate relationships among different books on Amazon—insights that remain elusive when relying solely on product-specific information. This augmentation goes beyond the surface, unravelling the nuanced dynamics and interconnections within the copurchase network, thereby enriching the model's understanding of the complex web of influences at play in consumer behaviour.

Lastly, an additional key observation is the significance of incorporating recent network metrics. Our findings emphasize the dynamic nature of Amazon's copurchase network, where constant changes occur. Recognizing this dynamism, we advocate for the importance of regularly updating the model with the most recent data. This practice ensures not only accurate link predictions but also the delivery of relevant personalized recommendations, aligning marketing strategies with the evolving landscape of consumer behaviour on Amazon's platform.

6 Limitations and Considerations

While our study yielded promising results, we acknowledge several potential weaknesses that could be addressed in future iterations. One primary limitation is the absence of edge weights in our dataset. As a result, products with numerous co-purchases are considered equal to those with only a few, disregarding the differential influence levels these products might have on consumers' purchasing patterns. Incorporating edge weight information would significantly enhance our ability to make informed decisions regarding product bundling and recommendations. Additionally, it would bolster the predictive power of our models, showcasing the tangible impact of including network metrics as features. However, obtaining such detailed data is restricted to Amazon's internal and confidential information. Hence, we are constrained to using moderately representative features, such as sales rank, to approximate product popularity.

Moreover, our dataset is from almost two decades ago, which raises concerns about its relevance. Given the rapid evolution of consumer preferences and the influx of new products since 2006, the insights derived from the top influential products might now be outdated. Nevertheless, our study lays the groundwork for a framework that illustrates how specific product selections and bundling strategies can effectively drive consumer purchasing behaviour, as well as the positive impacts of network analytics on predictive models

7 References

1. Knox C. Why are biographies so popular? Because humans are enthralled by the lives of others. The Guardian [Internet] 2023;Available from: https://www.theguardian.com/commentisfree/2023/may/09/biographies-popular-stories-true-books-boswell-book-festival
2. Haveliwala TH. Topic-sensitive PageRank: A context-sensitive ranking algorithm for web search. IEEE [Internet] 2003;Available from: https://www.researchgate.net/publication/3297186_Topic-sensitive_PageRank_A_context-sensitive_ranking_algorithm_for_Web_search
3. Muralidharan G. Decoding amazon's recommendation system. Argoid.ai [Internet] 2023;Available from: https://www.argoid.ai/blog/decoding-amazons-recommendation-system