Recommendation Engine for Amazon Magazine Subscriptions

Sushil Khairnar[1], Deep Bodra[2]
Virginia Tech, Blacksburg, USA[1]
Harrisburg University of Science and Technology, USA[2]

***Abstract*—Recommender systems play a crucial role in enhancing user experience and engagement on e-commerce platforms by suggesting relevant products based on user behavior. In the context of Amazon's extensive catalog of over 8,000 magazines spanning more than twenty-five categories, providing personalized magazine subscription recommendations poses a significant challenge. This study addresses the problem of identifying potential future associations between magazine reviewers and products using a graph-based approach. Specifically, we aim to predict unseen but likely links between users and magazines to improve recommendation quality. To achieve this, we construct an undirected bipartite network connecting reviewers and magazine products based on review data. We perform network analysis using measures such as centrality, modularity, and clustering, and apply sentiment analysis and topic modeling to extract behavioral and thematic insights from user reviews. These insights inform a series of link prediction techniques including Common Neighbors, Adamic-Adar, Jaccard Coefficient, and Preferential Attachment evaluated using cross-validation and ROC curves. Our results show that the Preferential Attachment model outperforms other approaches, attributed to the skewed degree distribution inherent in the dataset's structure.**

***Keywords*—Sentiment analysis; topic modeling; recommender system; link prediction**

I. Introduction

A recommendation system is a data filtering system that generates data predictions based on user data linked with the item. Both merchants and consumers benefit from recommendation systems in internet marketing. On the one hand, recommendation algorithms are employed by sellers in order to maintain high levels of user engagement. Consumers, on the other hand, benefit from recommendation systems because they can quickly find what they're looking for. For example, Amazon suggests products to consumers based on their past purchases to encourage future purchases, while news websites suggest related content to viewers to optimize ad revenue. These businesses rely significantly on recommendation systems to improve user experience.

Both service providers and users benefit from recommender systems. Users are able to lower the resource consumption and expenses of searching for and selecting the right things to buy in an online environment. Also, service providers reduce the cost of marketing large amount of products to large user base. If a person enjoys reading scientific periodicals, he or she might be interested in reading magazines about scientific research. Recommendation systems can help them find related products and improve the quality of their decisions. Recommender systems increase revenue in an e-commerce scenario since they are efficient at selling more things. Recommender systems in scientific libraries assist users by allowing them to go beyond catalog searches. In addition to this, indirect connections between the users can be identified to recommend the products purchased by one user to another user. As a result, the importance of employing efficient and accurate recommendation algorithms inside a system that provides consumers with relevant and dependable recommendations cannot be overstated.

Many e-commerce companies offer a diverse selection of products to their customers. Providing users with the most suited items makes the purchasing process more efficient and boosting user happiness. Improved customer happiness keeps users coming back to the website, increasing sales and profitability for products. As a result, more companies are beginning to recommend products to customers, necessitating the efficient analysis of user product interests. Standard recommendation algorithms, such as content-based filtering and collaborative filtering, use a matrix to model user ratings and anticipate customer ratings for unrated goods based on user/item similarity.

In the context of Amazon's magazine subscription platform, which features more than 8,000 magazines across over twenty-five categories, recommending personalized content is particularly challenging due to the diversity and volume of data. This study focuses on addressing the following research question: How can a graph-based recommendation system leveraging user reviews, sentiment analysis, and topic modeling improve the prediction of potential future links between reviewers and magazine subscriptions?

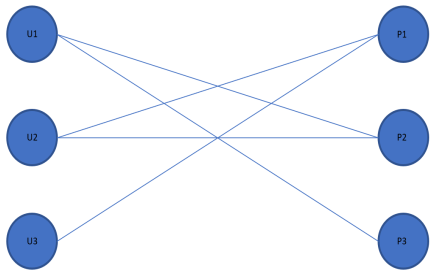


Fig. 1. Bipartite network.

To answer this question, we build a bipartite network (Fig. 1) between users and products using Amazon review data and explore the utility of various link prediction algorithms in identifying new recommendations. We incorporate techniques such as sentiment analysis to assess review polarity, and topic modeling to identify key themes in user preferences. These components help enrich the graph and guide the recommendation process.

The main contributions of this study are as follows:

- We construct and analyze a large-scale bipartite graph between Amazon reviewers and magazines based on review data.
- Integrate sentiment analysis and topic modeling to extract contextual and thematic features from textual reviews.
- Apply and compare multiple link prediction methods Common Neighbors, Adamic-Adar, Jaccard Coefficient, and Preferential Attachment—to identify potential reviewer-magazine associations.
- Evaluate the performance of these models using ROC-AUC and demonstrate that the Preferential Attachment model yields superior results due to the skewed degree distribution in the network.

The implications of this work lie in enhancing personalized recommendation engines for online subscriptions and providing a scalable framework for integrating behavioral and semantic data in graph-based recommender systems.
The rest of the study is organized as follows: Section II reviews the related work. Section III details the methodology, including network construction, sentiment analysis, and topic modeling. Section IV presents experimental results and analysis. Section V presents the discussion of the study. Section VI concludes the study and discusses possible directions for future work.

II. Background and Related Work

Recommender systems have significantly improved user engagement and satisfaction on e-commerce platforms. Particularly, graph-based methods have become prominent for modeling user-product interactions as networks. Li et al. explored the efficacy of bipartite graphs for recommendations, demonstrating their superiority over conventional collaborative filtering. Similarly, Zhang and Chen highlighted the potential of network centrality and clustering to deliver explainable recommendations, enhancing user trust and transparency.

Link prediction is a core component of graph-based recommendations, providing insights into potential future interactions within networks. The foundational works by Adamic and Adar [1], Liben-Nowell and Kleinberg [8], Newman [11], and Barabási et al. [2] introduced seminal link prediction approaches including Common Neighbors, Adamic-Adar, Jaccard Coefficient, and Preferential Attachment. Preferential Attachment, in particular, emphasizes that nodes with higher degrees are likelier to form new connections, a concept strongly relevant to real-world scale-free networks.

Incorporating textual data through sentiment analysis and topic modeling has further enriched recommender systems. Pak and Paroubek [20] leveraged sentiment analysis to refine recommendations based on user-generated textual feedback. Additionally, Blei et al. introduced Latent Dirichlet Allocation (LDA) for topic modeling, enabling a more nuanced understanding of user interests through thematic analysis of reviews. McAuley et al. extended these techniques, combining textual and visual features to enhance the contextual relevance of recommendations.

Parallel to recommender system advancements, significant research has been conducted in predictive analytics within other application domains, such as air traffic delay prediction. Flight delays have been a persistent issue, with only 79.63% of flights landing on time over the past decade, according to the Bureau of Transportation Statistics (BTS). Ye's study employed multiple machine learning models (Linear Regression, SVM, ExtraRT, and LightGBM) to forecast flight departure delays at Nanjing Lukou International Airport, identifying critical relationships between meteorological factors and delays [6]. Similarly, Atlioglu [1] analyzed flight operational data using diverse machine learning methods, emphasizing the importance of data selection for accurate delay prediction.

Esmaeilzadeh and Mokhtarimousavi utilized Support Vector Machine (SVM) analysis to examine air traffic delays at major airports in New York City, focusing on operational and flow management factors [3]. Xu et al. [5] applied Bayesian networks for modeling delay propagation across airports, providing insights into the systemic nature of delay spread through an empirically informed Bayesian approach. Bratu and Barnhart [2] introduced the Passenger Delay Calculator, shifting focus from aircraft-centric to passenger-centric delay analysis, thus providing a more nuanced understanding of the impact of delays. Kim and Choi [4] proposed advanced deep-learning models, including deep RNN architectures, for predicting delays with greater accuracy, emphasizing the efficacy of neural networks in capturing complex temporal patterns in delay data [9].

While the application domains of recommender systems and air traffic delay prediction differ substantially, methodologies such as machine learning, sentiment analysis, and data-driven modeling are common threads linking these fields. Our research leverages these interdisciplinary insights, employing advanced analytical techniques such as sentiment analysis, topic modeling, and graph-based link prediction methods to provide personalized magazine recommendations on Amazon. This integrated approach not only enhances recommendation quality but also demonstrates the adaptability and robustness of predictive methodologies across diverse contexts.

III. Approach

We analyze the dataset for amazon magazine subscriptions which contain the data about magazines subscribed by a certain user and reviews written for the particular magazine. We generate a bipartite graph using the reviewerID and ProductID as the two independent sets of vertices [13]. This is a bipartite graph because the edges are between the reviewerID and productID and no edges from and to the same set. Once the graph is generated, we perform network analysis which involves centrality analysis, modularity analysis, finding connected components, small world network analysis and heavy tail analysis. This analysis supplies crucial insights about the graph. After the network analysis, we perform sentiment analysis to capture the sentiment of the reviewers in relation to that productID. This step gives us insights about the relevance of the edges while conducting link prediction. We then perform topic modeling to capture the group of topics the reviewer is discussing about. We select the highest-ranked reviewer Id and apply topic modeling to the reviews written by that user. Following that, we apply link prediction to find a link between the reviewer and the products with which aim we recommend magazines to users.

A. Network Analysis

The use of networks and graph theory to investigate group structures is known as network analysis [14]. It describes networked structures in terms of nodes, which can be individual actors, people, or items inside the network, and the ties, edges, or connections that connect them, as well as the relationships and interactions that exist between them. It aids us in comprehending the structure of a relationship in social networks, as well as a structure or change process in natural phenomena. Identifying the critical node in a network is an important application of network analysis. This task is called Network Centrality measurement. It can relate to the task of determining the most influential person or the group's representative in social network analysis. We make use of the networkx [10] library in the python to capture the insights about the network. Fig. 2 presents the Gephi visualization.

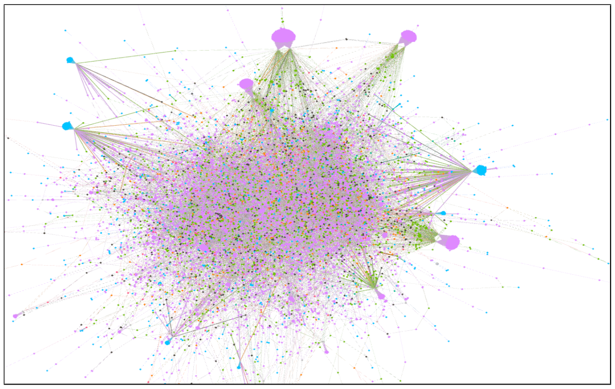


Fig. 2. Gephi visualization.

B. Sentiment Analysis

Sentiment Analysis, also known as opinion mining, is the computational study of opinions, sentiments, subjectivity, or evaluations. It is widely used across domains such as market research, public relations, reputation management, financial trading, and customer experience analysis. By examining the polarity (positive, negative, or neutral) of user-generated text, sentiment analysis helps organizations understand public perception and detect shifts in opinion.

In our study, sentiment analysis was used to determine the tone of user reviews and to assess whether sentiment polarity could improve link prediction performance. Specifically, each review was analyzed to extract a sentiment score, which we considered as a potential weight on the edge between a reviewer and a product in the bipartite graph.

To prepare the text for sentiment analysis, we applied a multi-step preprocessing pipeline: we used regular expressions to remove links and special characters, eliminated stop words using the NLTK corpus, and performed stemming using the Porter Stemmer [15]. These steps ensured that irrelevant tokens and variations of the same word did not distort sentiment scores. We used the VADER sentiment analyzer, which outputs a compound score between -1 (most negative) and +1 (most positive). These scores were mapped onto a 3-class label system (positive, neutral, negative) using VADER's default thresholds. Although we initially considered using these scores as edge weights in the graph for weighted link prediction, the review distribution was heavily skewed: out of 88,318 edges, only 2.8% were negative. This imbalance would have led to biased edge weighting and diminished performance in graph-based models. Therefore, we opted not to use sentiment scores as weights, instead using an unweighted graph for better generalizability and efficiency.

C. Topic Modeling

Topic modeling is a powerful text mining technique used to uncover hidden thematic structures in a corpus. It is especially useful in organizing and summarizing large collections of unstructured text data. We employed Latent Dirichlet Allocation (LDA), a widely adopted probabilistic model that represents each document as a mixture of topics and each topic as a distribution over words. For our study, we applied topic modeling to the set of reviews written by the most active reviewer in the dataset. This helped us gain insights into the types of content that frequent reviewers engage with, which in turn aids in better personalization for recommendations [17]. We selected four topics for the LDA model based on empirical evaluation using coherence scores, which measure the semantic consistency of words within a topic. We evaluated topic numbers ranging from 3 to 7 and found that k=4 yielded the highest coherence score (0.51), indicating that four topics offered a good balance between granularity and interpretability. The topics identified included themes such as food and travel, business and income, technology and education, and writing and products, reflecting the diverse interests of the reviewer. This thematic categorization allowed us to associate reviewers with topic clusters, which could be used for finer-grained recommendation strategies beyond graph-based link prediction.

IV. Experiment

A. Data Analysis

The analysis was conducted using the Amazon magazine subscription dataset. The provided dataset contains data about amazon magazine subscriptions, publications subscribed by a certain user (reviewerID), and reviews published for that magazine (reviewText). The overall score ranges from 0 (worst) to 5 (excellent) (best). Amazon assigns a unique identifying number (Asin) to each magazine. The data set has a total of 89,689 rows, with a total of 72098 distinct reviewers and 2428 unique Asin's.

Features:

- Overall - States the user rating of the product
- Vote - Total votes/ratings given
- Verified - Is the user verified
- reviewTime - Time at which the product was reviewed
- reviewerID - Unique ID of the reviewer

- Asin - Amazon Standard Identification Number (Product ID)
- reviewerName - Name of the review
- reviewText - product review written by the user
- Summary - summary of the user review
- unixReviewTime - timestamp of the review

The bipartite graph was subjected to network analysis, and key parameters such as density, connected components, and clustering coefficient were computed. Below describes a more detailed investigation of network.

B. Preprocessing

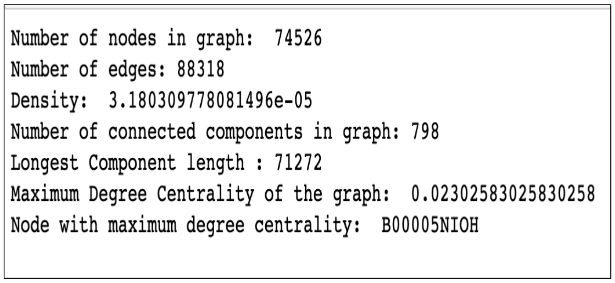
Prior to analysis, the dataset underwent several preprocessing steps to ensure data quality and suitability for graph construction and text-based modeling. The original dataset contained 89,689 records. After preprocessing, a total of 88,318 valid review entries remained, which were used for graph construction. The following preprocessing steps were performed:

- Text Cleaning: URLs, HTML tags, special characters, and excessive whitespace were removed using regular expressions.
- Stop Word Removal: Common English stop words were removed using the NLTK stopword corpus.
- Lowercasing: All review text was converted to lowercase to maintain consistency.
- Stemming: Words were reduced to their base/root form using the Porter Stemmer to normalize vocabulary.
- Filtering Empty Reviews: Reviews with fewer than 10 characters after cleaning were discarded.
- Verified Purchases: Only reviews from verified purchasers were retained to improve reliability.
- Reviewer Filtering: For topic modeling, we selected the reviewer with the highest number of reviews, filtering the dataset to just their entries, resulting in 146 reviews for that user.

The remaining dataset was then used to build a bipartite graph between reviewerID and Asin, where each edge represents a review. This cleaned and filtered dataset formed the basis for subsequent network, sentiment, and topic modeling analyses.

C. Graph Analysis

The strength and direction of links between nodes in a graph are determined via graph analysis. The graph analysis in this study began with determining the density and average clustering coefficient of the graph [18]. The centrality analysis was then performed, which determined how essential a node or edge is for the network's connectedness or information flow. The graph's maximum centrality was found to be 0.023. Fig. 3 shows the basic graph analysis.



```
Number of nodes in graph:  74526
Number of edges: 88318
Density:  3.180309778081496e-05
Number of connected components in graph: 798
Longest Component length : 71272
Maximum Degree Centrality of the graph:  0.02302583025830258
Node with maximum degree centrality:  B00005NIOH
```

Fig. 3. Basic graph analysis.

Following that, small world network identification and heavy tail analysis were performed. Fig. 4 shows the result of the small world network analysis. The modularity analysis was then used to determine how thick the network's connections are. The optimal partition's modularity was found to be 0.8767 and Fig. 5 reflects the outcome.

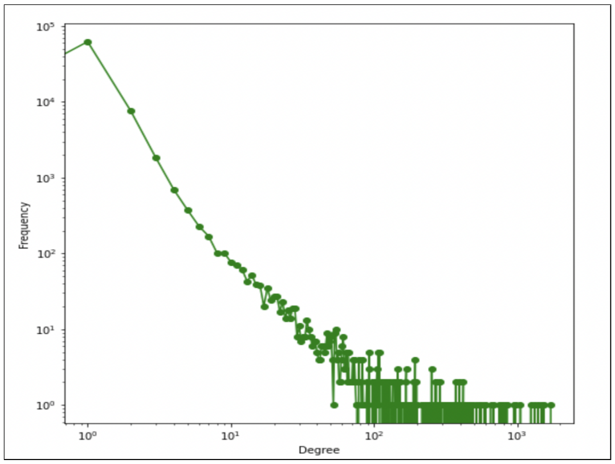


Fig. 4. Small world network analysis.

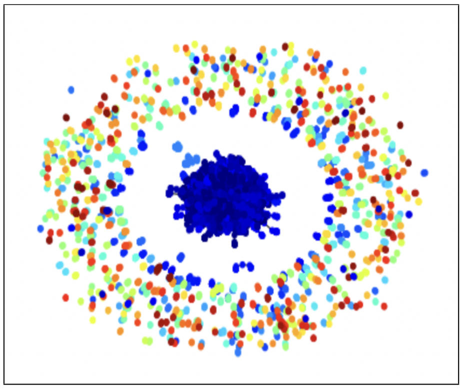
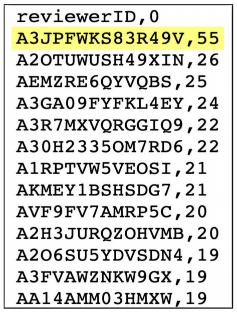


Fig. 5. Modularity plot for the best partition.

D. Topic Modeling

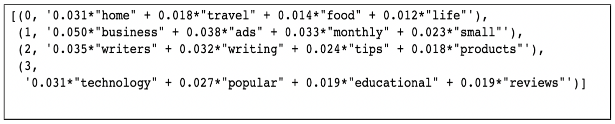
A user with the maximum number of Amazon magazine reviews was chosen for the purpose of topic modeling. The data of reviewers with the most reviews was sorted, and the reviewerID with the most reviews was chosen. The data list that was utilized to select the user is shown in Fig. 6.



```
reviewerID,0
A3JPFWKS83R49V,55
A2OTUWUSH49XIN,26
AEMZRE6QYVQBS,25
A3GA09FYFKL4EY,24
A3R7MXVQRGGIQ9,22
A30H2335OM7RD6,22
A1RPTVW5VEOSI,21
AKMEY1BSHSDG7,21
AVF9FV7AMRP5C,20
A2H3JURQZOHVMB,20
A2O6SU5YDVSDN4,19
A3FVAWZNKW9GX,19
AA14AMM03HMXW,19
```

Fig. 6. List of reviewers with maximum reviews.

Topic modeling was performed on the user (A3JPFWKS83R49V) with the most reviews. We filtered the dataset to capture the reviews written by the most active reviewer. We used Latent Dirichlet allocation (LDA) [6] method for fitting a topic model. This technique treats each document as a mixture of topics, and each topic as a mixture of words. The function parameters to the LDA model - the number of topics and number of words were set to four. Fig. 7 depicts the outcome of the topic modeling. Words like home, food, travel, and life appear in the first topic, implying that the user reviewed more about food and travel. The words business and income appeared in the second topic, writing product appeared in the third, and technology and education appeared in the fourth. This implies that the reviewer is a voracious reader and actively evaluates magazines about these subjects.



```
[(0, '0.031*"home" + 0.018*"travel" + 0.014*"food" + 0.012*"life"'),
 (1, '0.050*"business" + 0.038*"ads" + 0.033*"monthly" + 0.023*"small"'),
 (2, '0.035*"writers" + 0.032*"writing" + 0.024*"tips" + 0.018*"products"'),
 (3,
  '0.031*"technology" + 0.027*"popular" + 0.019*"educational" + 0.019*"reviews"')]
```

Fig. 7. Result of topic modeling on user with maximum reviews.

E. Sentiment Analysis

Out of 88318 edges or reviews, we only got 2567 (2.8 %) (see Fig. 8) negative reviews, the rest were either positive or neutral . In our case, an edge with weight as the sentiment score represents the relation between the reviewer and the product. An edge with positive sentiment score should indicate that the user praised the product and other users having common properties to that user should get recommendation of the product.

However, as observed from the Fig. 8, the number of positive, negative, and neutral reviews are very skewed. This indicates that, when it comes to magazine subscriptions, people do not tend to write negatively about them. Neutral edges can also be considered harmless to link prediction as these are most likely stating facts that summarize what is published in the magazine [18]. Considering these findings, sentiment analysis was effective in deciding not to use the sentiment scores for weighing the graph due to the skewed distribution and high complexity of bipartite weighted graph.

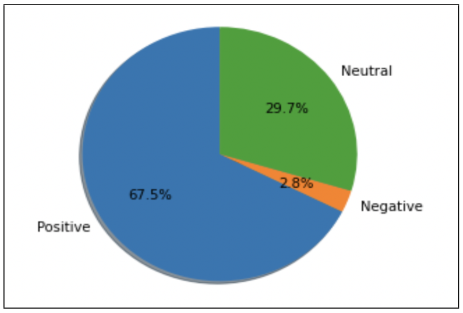
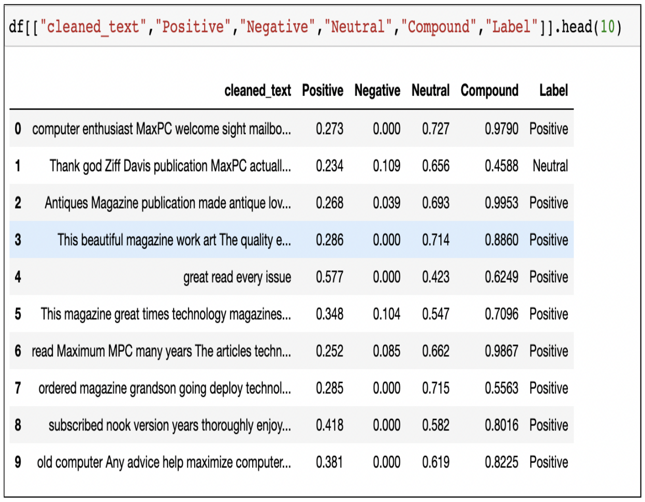


Fig. 8. Sentiment analysis results.



```
df[["cleaned_text","Positive","Negative","Neutral","Compound","Label"]].head(10)
```

	cleaned_text	Positive	Negative	Neutral	Compound	Label
0	computer enthusiast MaxPC welcome sight mailbo...	0.273	0.000	0.727	0.9790	Positive
1	Thank god Ziff Davis publication MaxPC actuall...	0.234	0.109	0.656	0.4588	Neutral
2	Antiques Magazine publication made antique lov...	0.268	0.039	0.693	0.9953	Positive
3	This beautiful magazine work art The quality e...	0.286	0.000	0.714	0.8860	Positive
4	great read every issue	0.577	0.000	0.423	0.6249	Positive
5	This magazine great times technology magazines...	0.348	0.104	0.547	0.7096	Positive
6	read Maximum MPC many years The articles techn...	0.252	0.085	0.662	0.9867	Positive
7	ordered magazine grandson going deploy technol...	0.285	0.000	0.715	0.5563	Positive
8	subscribed nook version years thoroughly enjoy...	0.418	0.000	0.582	0.8016	Positive
9	old computer Any advice help maximize computer...	0.381	0.000	0.619	0.8225	Positive

Fig. 9. Sentiment analysis scores and labels.

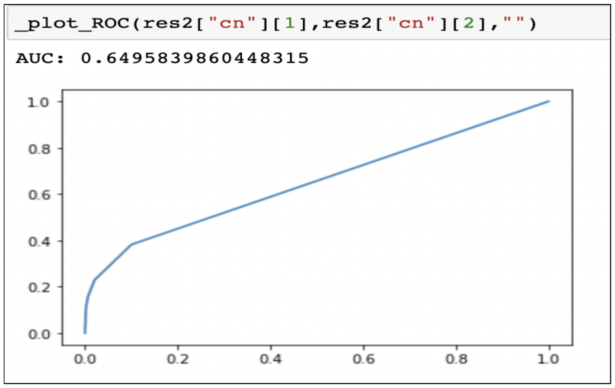
F. Link Prediction

In the realm of graphs and networks, link prediction is one of the most important research areas. The goal of link prediction is to find node pairs that will establish a link in the future. The simplest link prediction metrics are similarity-based metrics, which produce a similarity score for each pair x and y. The structural or node properties of the considered pair are used to calculate the score S(x,y). The non-observed linkages are given scores based on how similar they are. The anticipated link between two nodes is represented by the pair with the highest score. In the study [19], the authors details that the structural attribute of the network is one of the properties that may be used to determine the similarity measures between each pair. Local and global scores, node-dependent and path-dependent scores, parameter-dependent and parameter-free scores, and so on can all be categorized using this feature.

1) Common neighbors: The size of common neighbors [11] for a given pair of nodes is determined as the intersection of the two node neighborhoods in a particular network or graph.

$$S(x, y) = |\Gamma(x) \cap \Gamma(y)|$$

where, $\Gamma(x), \Gamma(y)$ are neighbors of the node x and y respectively. With the number of shared neighbors between them, the likelihood of a link between x and y increases. Newman estimated this quantity in a collaboration network and showed that the likelihood of collaboration between two nodes is determined by their shared neighbors. In a large social network, Kossinets and Watts [7] found that two students who have a lot of similar friends are more likely to be friends. The common neighbor method has been found to outperform other complex methods on most real-world networks. On applying this technique to the amazon dataset, we found the Area under the curve to be 0.65. Fig. 10 shows the ROC curve for the same.



```
_plot_ROC(res2["cn"][1],res2["cn"][2],"")
```

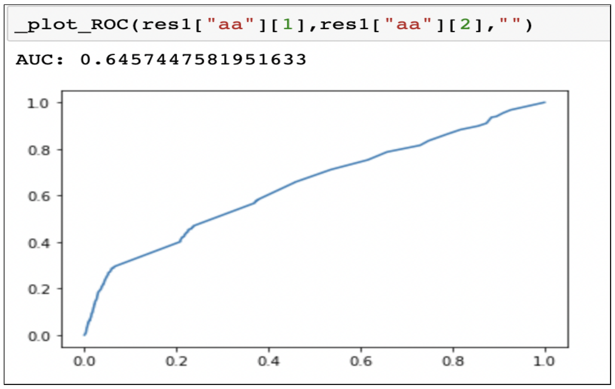
```
AUC: 0.6495839860448315
```

Fig. 10. ROC curve for common neighbors algorithm.

2) Adamic-Adar: Adamic and Adar[1] proposed a measure for calculating a similarity score between two web pages based on shared traits, which Liben-Nowell et al. modified and employed in link prediction.

$$S(x,y) = \sum 1/log(k_z)z \in \Gamma(x) \cap \Gamma(y)$$

where, k_z denotes the node z's degree. The equation clearly shows that the common neighbors with lower degrees are given more weight. This is also intuitive in the real world; for example, a person with a larger number of friends spends less time or resource on a single friend than someone with fewer friends[16]. On applying this technique to the dataset, we found the area under the curve to be 0.65, which is comparable to the results of Common Neighbors. Fig. 11 shows the ROC curve for the same.



```
_plot_ROC(res1["aa"][1],res1["aa"][2],"")
```

```
AUC: 0.6457447581951633
```

Fig. 11. ROC curve for Adamic-Adar algorithm.

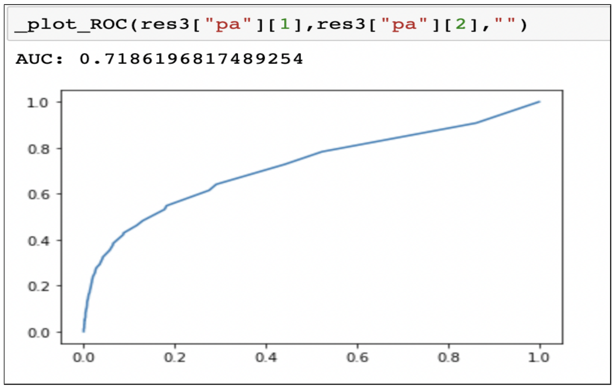
3) Preferential attachment: In this study [2], to create a growing scale-free network, the concept of preferential attachment is used. The phrase "growing" refers to the progressive nature of nodes in a network over time. The likelihood of adding a new link to a node x is related to k_x, the node's degree. The preferential attachment score between two nodes, x and y can be calculated using the formula:

$$S(x,y) = k_x.k_y$$

The key advantages of this metric are its simplicity (since it takes the least amount of information for score calculation) and computing time. It can also be employed in a non-local setting because it simply requires degree as information rather than common neighbors. The PA's effectiveness improves in assortative networks, while it deteriorates in disassortative networks. In other words, if greater degree nodes are densely connected and smaller degree nodes are rarely connected, PA produces superior outcomes. Hasan et al. [5] demonstrated that aggregation functions (e.g., sum, multiplication, etc.) over feature values of vertices might be used to compute link feature value in a supervised learning framework. Fig. 12 presents the ROC curve for Preferential Attachment algorithm.

Summation can be used instead of multiplication as an aggregate function in the preceding equation, and it has been proven to be highly effective. In [5], the authors show that preferential attachment combined with the aggregate function "sum" works effectively for link prediction in co-authorship networks.

Compared to other methods, we got better results for Preferential Attachment model because of the nature of the graph. Degree distribution (Fig. 14) shows that more than half of the nodes are having degree equal to two. Additionally, as observed, very few number of nodes have high degree. Fig. 13 shows that there are some nodes in the network to which many nodes having lesser degree are attached. This situation favors the Preferential Attachment model yielding better results than other models.



```
_plot_ROC(res3["pa"][1],res3["pa"][2],"")
```

```
AUC: 0.7186196817489254
```

Fig. 12. ROC curve for preferential attachment algorithm.

4) Jaccard Coefficient: This [12] metric is similar to the common neighbor. Additionally, it normalizes the above score, as given below.

$$S(x,y) = |\Gamma(x) \cap \Gamma(y)| \div \Gamma(x) \cup \Gamma(y)|$$

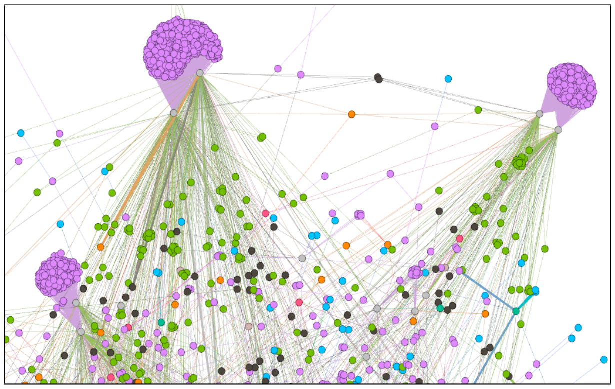


Fig. 13. Gephi visualisation - graph component where the reviewer has written 2 or more reviews.

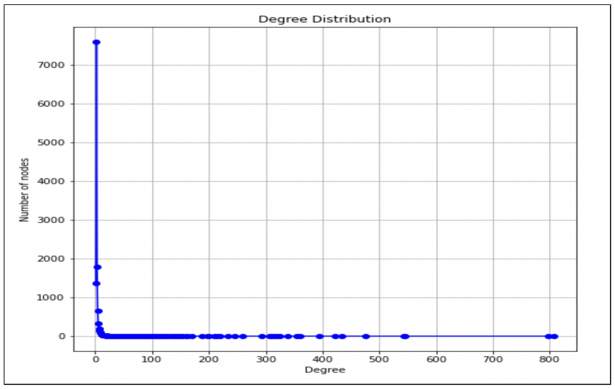
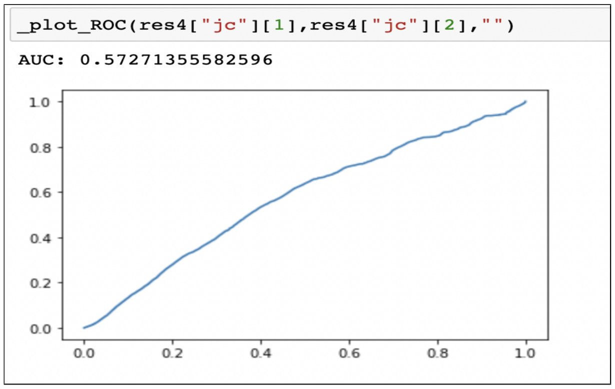


Fig. 14. Degree distribution of the graph.

The probability of choosing common neighbors of paired vertices from all the neighbors of either vertex is described by the Jaccard coefficient. The number of shared neighbors between the two vertices being analyzed improves the pairwise Jaccard score. This similarity metric performs worse than Common Neighbors, as proved by Liben-Nowell et al.[8]. Jaccard Coefficient algorithm performs the worst amongst all algorithms. As observed in Fig. 15, the area under the curve for this technique is 0.57.

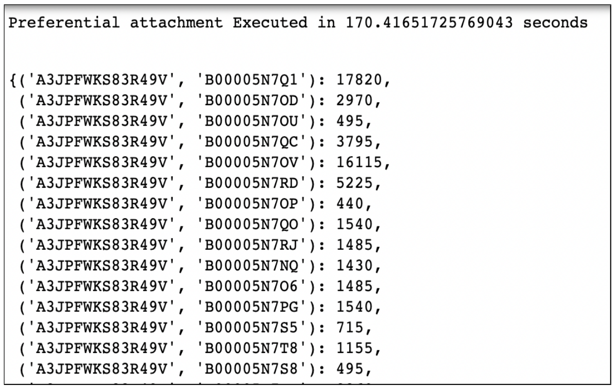


```
_plot_ROC(res4["jc"][1],res4["jc"][2],"")
```

```
AUC: 0.57271355582596
```

Fig. 15. ROC curve for jaccard coefficient algorithm.

The results in Fig. 16 shows the links obtained using the Preferential Attachment algorithm. As observed, the results indicate the possible links between the reviewer and the productID along with the Preferential attachment score for reviewer and the product. The score indicates how likely it is to establish a link between the two.



```
Preferential attachment Executed in 170.41651725769043 seconds


{('A3JPFWKS83R49V', 'B00005N7Q1'): 17820,
 ('A3JPFWKS83R49V', 'B00005N7OD'): 2970,
 ('A3JPFWKS83R49V', 'B00005N7OU'): 495,
 ('A3JPFWKS83R49V', 'B00005N7QC'): 3795,
 ('A3JPFWKS83R49V', 'B00005N7OV'): 16115,
 ('A3JPFWKS83R49V', 'B00005N7RD'): 5225,
 ('A3JPFWKS83R49V', 'B00005N7OP'): 440,
 ('A3JPFWKS83R49V', 'B00005N7QO'): 1540,
 ('A3JPFWKS83R49V', 'B00005N7RJ'): 1485,
 ('A3JPFWKS83R49V', 'B00005N7NQ'): 1430,
 ('A3JPFWKS83R49V', 'B00005N7O6'): 1485,
 ('A3JPFWKS83R49V', 'B00005N7PG'): 1540,
 ('A3JPFWKS83R49V', 'B00005N7S5'): 715,
 ('A3JPFWKS83R49V', 'B00005N7T8'): 1155,
 ('A3JPFWKS83R49V', 'B00005N7S8'): 495,
```

Fig. 16. Predictions for preferential attachment algorithm.

5) Other Methods: We tried link prediction with the graph distance method, which uses the shortest path length between nodes without an edge. The shortest path lengths are then sorted in ascending order, with the top "K" edges designated as future links. With the help of sentiment scores, we changed our bipartite network into a weighted graph. The edges are weighted using the compound score indicated in the Fig. 9, are used as weights of the edges. To prevent negative weights, these scores were transformed to a range of 0 to 2. Then, to find out the edges which do not exist, it was required to iterate the network of nearly 28000 edges in a quadratic time complexity. Every iteration also included time-consuming processes like identifying the shortest path length. The anticipated time it would take to run this algorithm on the network was roughly 100 days [20]. As a result, the weighted graph and shortest path length method was avoided.

Approaches like user-based collaborative filtering and item-based collaborative filtering were also investigated. Users or items are both considered vectors in these procedures. The user is then recommended an item based on the similarity between these vectors. Due to the significant sparsity of data and edges, the majority of the entries would be 0. Both algorithms will become ineffective as a consequence of this. Hence, these methods were avoided.

V. Discussion

The results from our study reveal that the Preferential Attachment model consistently outperformed other link prediction algorithms such as Common Neighbors, Adamic-Adar, and Jaccard Coefficient. This outcome can be attributed to the inherent skewed degree distribution within the reviewer-product bipartite graph, where a small number of users have disproportionately high activity levels. These highly connected reviewers are more likely to form new links, aligning with the assumptions of the Preferential Attachment model. An important insight gained from the sentiment analysis was the overwhelming positivity or neutrality in user reviews—only 2.8% of the reviews were classified as negative. While sentiment analysis provided valuable context for understanding edge importance, the sentiment imbalance ultimately made it impractical to use these scores as edge weights for graph modeling. This highlights a broader challenge in applying sentiment-weighted graphs in domains where user feedback is rarely negative. The use of topic modeling also proved insightful in identifying the thematic diversity of active users. For instance, the top reviewer in our dataset wrote about a wide range of topics such as travel, food, business, and education. This reinforces the importance of capturing thematic preferences in designing personalized recommender systems. However, choosing the optimal number of topics remains a subjective decision, and while coherence scores guided our choice of four topics, more sophisticated tuning strategies could further improve interpretability. From a broader perspective, this study demonstrates the value of integrating semantic and behavioral insights into graph-based recommender systems. By combining network structure with textual features, we move closer to personalized and context-aware recommendations. However, computational complexity remains a concern, especially for scaling methods like shortest path-based prediction on weighted graphs. Future work could explore more efficient graph embedding techniques that incorporate both structure and sentiment while maintaining interpretability.

VI. Conclusion

In this study, we constructed a bipartite network connecting Amazon reviewers to magazine subscriptions using a real-world dataset. To extract deeper behavioral and thematic insights, we applied sentiment analysis, topic modeling, and network visualization techniques. These insights were then leveraged to perform link prediction, aiming to recommend relevant magazine subscriptions to users.

We evaluated several link prediction algorithms such as Common Neighbors, Adamic-Adar, Jaccard Coefficient, and Preferential Attachment—using K-fold cross-validation and ROC-AUC scores. Among these, the Preferential Attachment model yielded the most promising results, largely due to the skewed degree distribution observed in the graph, where a small number of highly active users dominated the network structure.

Although sentiment analysis offered valuable context, the overwhelmingly positive and neutral nature of the reviews limited its effectiveness as a weighting factor. As a result, link prediction was performed on an unweighted graph to avoid introducing bias from sentiment imbalance.

Looking forward, future research could explore methods to balance or normalize sentiment distributions to support weighted graph modeling. Additionally, incorporating graph embedding techniques or temporal dynamics could further enhance recommendation accuracy and scalability. Overall, this work highlights the potential of combining network structure with natural language insights to improve recommendation systems in content-rich domains like magazine subscriptions.

References

[1] Lada A Adamic and Eytan Adar. 2003. Friends and neighbors on the Web. Social Networks 25, 3 (2003), 211–230. hNps://doi.org/10.1016/S0378-8733(03)00009-1

[2] A.L Barabási, H Jeong, Z Néda, E Ravasz, A Schubert, and T Vicsek. 2002. Evolution of the social network of scientific collaborations. Physica A: Statistical Mechanics and its Applications 311, 3 (2002), 590– 614. hNps://doi.org/10.1016/S0378-4371(02)00736-7

[3] J. Kleinberg D. Liben-Nowell. 2003. The Link Prediction Problem for Social Networks. CIKM (2003).

[4] Evan Darke, Zhou Zhuang, and Ziyue Wang. 2017. Applying Link Prediction to Recommendation Systems for Amazon Products.

[5] M.A. Hasan, V. Chaoji, S. Salem, and M. Zaki. 2006. Link prediction using supervised learning. Proc. of SDM 06 Workshop on Link Analysis, Counterterrorism and Security (04 2006).

[6] Hamed Jelodar, Yongli Wang, Chi Yuan, and Xia Feng. 2017. Latent Dirichlet Allocation (LDA) and Topic modeling: models, applications, a survey. (11 2017).

[7] Gueorgi Kossinets and Duncan J. Watts. 2009. Origins of Homophily in an Evolving Social Network. Amer. J. Sociology 115, 2 (2009), 405–450. hNps://doi.org/10.1086/599247 arXiv:https://doi.org/10.1086/599247

[8] David Liben-Nowell and Jon Kleinberg. 2003. The Link Prediction Problem for Social Networks. In Proceedings of the Twelfth International Conference on Information and Knowledge Management (New Orleans, LA, USA) (CIKM '03). Association for Computing Machinery, New York, NY, USA, 556–559. hNps://doi.org/10.1145/956863.956972

[9] Yilun Liu, Chunhao Wu, and Xiaohui Tong. 2021. Prediction of Co-purchasing Products. (02 2021).

[10] NetworkX. 2022. https://networkx.org/. (04 2022). hNps://networkx.org

[11] M. E. J. Newman. 2001. Clustering and preferential attachment in growing networks. Phys. Rev. E 64 (Jul 2001), 025102. Issue 2. hNps://doi.org/10.1103/PhysRevE.64.025102

[12] Jaccard P. 2013. Distribution de la flore alpine dans le bassin des Dranses et dans quelques régions voisines. (06 2013).

[13] Sushil Khairnar, "Application of Blockchain Frameworks for Decentralized Identity and Access Management of IoT Devices" International Journal of Advanced Computer Science and Applications(IJACSA), 16(6), 2025. http://dx.doi.org/10.14569/IJACSA.2025.0160604

[14] Zhang, Y., & Chen, X. (2020) Explainable recommendation: A survey and new perspectives. Foundations and Trends in Information Retrieval, 14(1), 1–101. [https://doi.org/10.1561/1500000056]

[15] Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet Allocation. Journal of Machine Learning Research, 3, 993–1022.

[16] McAuley, J., Targett, C., Shi, Q., & van den Hengel, A. (2015). Image-based recommendations on styles and substitutes. Proceedings of the 38th International ACM SIGIR Conference, 43–52.

[17] D. Bodra and S. Khairnar, "Comparative Performance Analysis of Modern NoSQL Data Technologies: Redis, Aerospike, and Dragonfly," J. Res. Innov. Technol., vol. 4, no. 2, pp. 193-200, 2025. https://doi.org/10.57017/jorit.v4.2(8).05 [https://doi.org/10.1145/2766462.2767755]

[18] Friedman, J., Hastie, T., & Tibshirani, R. (2001). The Elements of Statistical Learning. Springer Series in Statistics.

[19] Rajaraman, A., & Ullman, J. D. (2011). Mining of Massive Datasets. Cambridge University Press.

[20] Pak, A., & Paroubek, P. (2010). Twitter as a corpus for sentiment analysis and opinion mining. LREC 2010.