Link Analysis

Sofia Campregher - DSE

June 2025

1 Absract

This project presents a link analysis on user-user. The Dataset “Amazon Books Reviews” which is obtained from Kaggle, is composed of over 3 million instances. The project considers only an extraction of a random sample which is processed to handle and encode missing values. The user–user graph represents users who read the same book, which are connected by edges. Two different graph structures are implemented to better analyze the relationships of data; the first is a static representation and the second is a dynamic and interactive visualization.

A particular emphasis is placed on the use of PageRank algorithm which is used to identify the central core of the model. The graph is used to define relationships between users based on shared book reviews. In other words, it identifies an influential group and analyzes their behavior.

2 Introduction

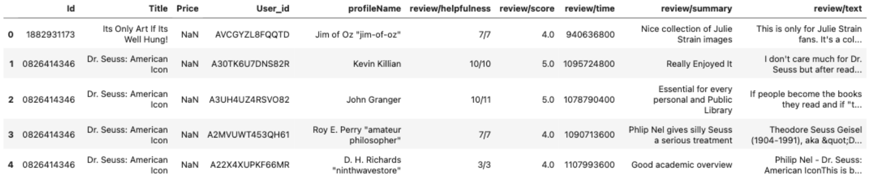
The study applies link analysis to a small subset that is obtained from a larger data set collected from Amazon. The subset explores user reviews to offer insight into consumer behavior and preferences.

The core of the model regards the implementation of a user–user graph structure used to connect the influential group who read and reviewed the same books.

3 Dataset

The dataset 'Amazon Books Reviews' is obtained from Kaggle and is downloaded directly from the repository identified by the URL suffix mohamedbakhet/ amazon-books-reviews.

The Dataset is made up of a collection of two distinct files containing over 3.2 million instances in total. The project focuses exclusively on one of the two files, Books Rating, which alone count approximately 3 million rows x 10 columns. The data set contains information on 212,404 unique books such as their IDs, prices, and titles. Furthermore, the data set contains data on users who have contributed reviews, including user IDs and profile names, as well as the textual content of the reviews themselves and several associated attributes.



	Id	Title	Price	User_id	profileName	review/helpfulness	review/score	review/time	review/summary	review/text
0	1882931173	Its Only Art If Its Well Hung!	NaN	AVCGYZL8FQQTD	Jim of Oz "jim-of-oz"	7/7	4.0	940636800	Nice collection of Julie Strain images	This is only for Julie Strain fans. It's a col...
1	0826414346	Dr. Seuss: American Icon	NaN	A30TK6U7DNS82R	Kevin Killian	10/10	5.0	1095724800	Really Enjoyed It	I don't care much for Dr. Seuss but after read...
2	0826414346	Dr. Seuss: American Icon	NaN	A3UH4UZ4RSVO82	John Granger	10/11	5.0	1078790400	Essential for every personal and Public Library	If people become the books they read and if "t...
3	0826414346	Dr. Seuss: American Icon	NaN	A2MVUWT453QH61	Roy E. Perry "amateur philosopher"	7/7	4.0	1090713600	Phlip Nel gives silly Seuss a serious treatment	Theodore Seuss Geisel (1904-1991), aka &quot;D...
4	0826414346	Dr. Seuss: American Icon	NaN	A22X4XUPKF66MR	D. H. Richards "ninthwavestore"	3/3	4.0	1107993600	Good academic overview	Philip Nel - Dr. Seuss: American IconThis is b...

Figure 1: Dataset sample from Amazon Book Reviews

The project analysis only considered two of the ten attributes in the graph implementation; User id for users, and Title for books.
In order to obtain a representative and manageable dataset, the project considers only an extraction of a random sample of 15,000 rows using a fixed random seed to ensure reproducibility.
The first step, to guarantee a clean dataset, is the detection and removal of potential missing value. The analysis reveals that only three attributes includes NaN values: Price, User id, and ProfileName. The results are summarized in the following table:

Column	Number of NaN
Price	12,663
User id	2,787
profileName	2,788

Table 1: Number of missing values (NaN)

All rows that present missing values in the User id field are removed, as this attribute is essential for the analysis. The other two attributes with missing values are excluded from the cleaning process because they do not significantly impact the analysis: the project does not focus on product prices, and ProfileName is a non-essential descriptive attribute. Missing values in the ProfileName column are replaced with the placeholder value "Unnamed", as indicated by the table above.
Lastly, the dataset is examined to detect any possible duplicate entries based on the combination of User id and Title. The analysis identified 44 identical reviews submitted for the same book by the same user.

4 Metodologhy

Pagerank is an algorithm that assigns a weight to each interconnected element in a set. The algorithm is widely used by Google Search to rank web pages. It is chosen because of its ability to capture the quantity and quality of a nodes's connection in a way not easy to fool. The innovation brought by Google evaluates the importance of web pages using a function that assigns a real number to each page in the web.
The project applied the above concept into the User Connection scenario which is used to define relationships between users based on shared book reviews.

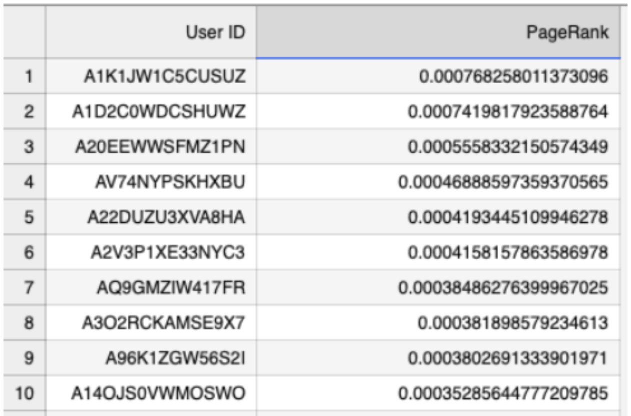
The methodology is applied to a fixed size sample of 15,000 but it can also be extended to larger datasets since the model uses library suggested for large size databases. Indeed, the project reveals a high degree of scalability in computational analysis. The approach applied makes the analysis suitable for scaling to millions of connections.

4.1 User Connection

The approach implements a user-user graph based on a Kaggle Dataset and applies the PageRank algorithm in order to identify the most influential users within the network.

The initial step of the code converts the column User ID and Title into integer indices and adds them to the dataset. Then, a sparse binary matrix A (users x books) is implemented; rows represents users, columns represents books and values indicate whether an user reviewed a specific book. This step is crucial because is it possible to memorize millions of data using the lowest amount of memory. The User-User Matrix is multiplied by its transpose $A \cdot A^{\top}$, in order to compute a symmetric user–user co-occurrence matrix, where each element indicates the number of co-reviewed books by a pair of users. Zero-valued and diagonal entries are removed to obtain a clean edges list.

Moreover, the code computes the PageRank Scores in order to identify central users: to qualify the influence within the network. Values are stored in a Dataframe and is exported to a CSV, as follow:



	User ID	PageRank
1	A1K1JW1C5CUSUZ	0.000768258011373096
2	A1D2C0WDCSHUWZ	0.0007419817923588764
3	A20EEWWSFMZ1PN	0.0005558332150574349
4	AV74NYPSKHXBU	0.00046888597359370565
5	A22DUZU3XVA8HA	0.0004193445109946278
6	A2V3P1XE33NYC3	0.0004158157863586978
7	AQ9GMZIW417FR	0.00038486276399967025
8	A3O2RCKAMSE9X7	0.000381898579234613
9	A96K1ZGW56S2I	0.0003802691333901971
10	A14OJS0VWMOSWO	0.00035285644777209785

Figure 2: User-User Graph showing the top users by PageRank score

The algorithm identifies the most important central users. As the table above illustrates, the highest-ranked user achieving a PageRank score of 0.000768 which is closely followed by another user. Both users reviewed books that are also reviewed by others, meaning they are connected to the network. The remaining users show a moderate drop, down to 0.000468 and 0.0000245.

The figure represents the user-user network graph. Each blue node represents a unique user and the grey edges represent the connection between users who have reviewed at least one book in common. This particular illustration positions nodes with more connections closer to the center and nodes with no connection to the border. The subset of data - containing the top 300 users - illustrates only a small percentage of nodes collocated in the central core; which are the highly interconnected users. All the other users, which are not interconnected with the influencing group have probably reviewed fewer or more unpopular books. The analysis highlights user's behaviors; users with similar taste.

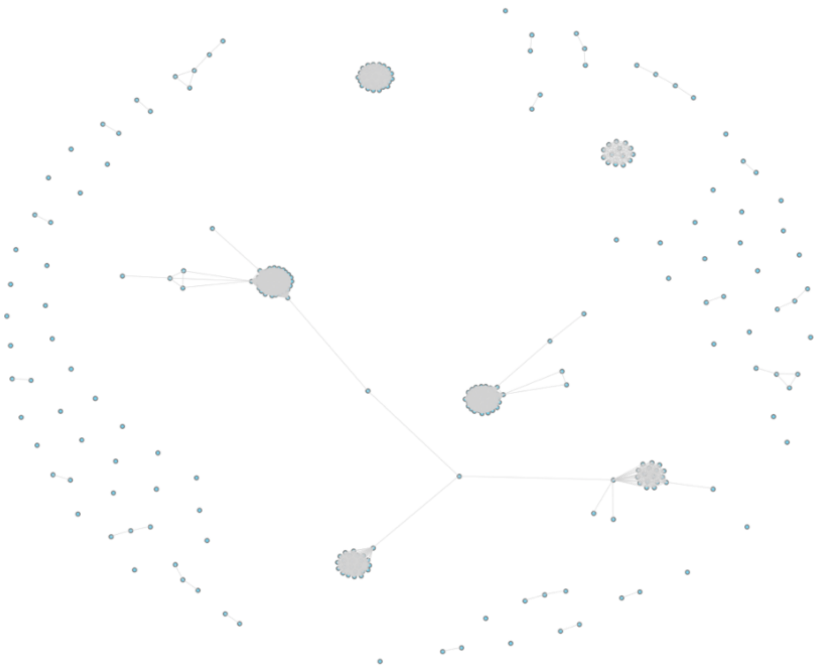


Figure 3: User-User Graph

The project offers a statistical overview to better understand the user–user network structure. The graph consists of 11,445 nodes and 46,636 edges with a low density of 0.00071, confirming a highly sparse network. The Dataset presents only 7,583 connections meaning that most users only review a few books. The following scatterplot is used to highlight the relationship between a user's PageRank score and their number of reviews: The scatterplot reveals a linear and weakly correlated pattern, where some users achieve high PageRank with few reviews, meaning that higher PageRank values do not depend on volume, rather by strategic connections.

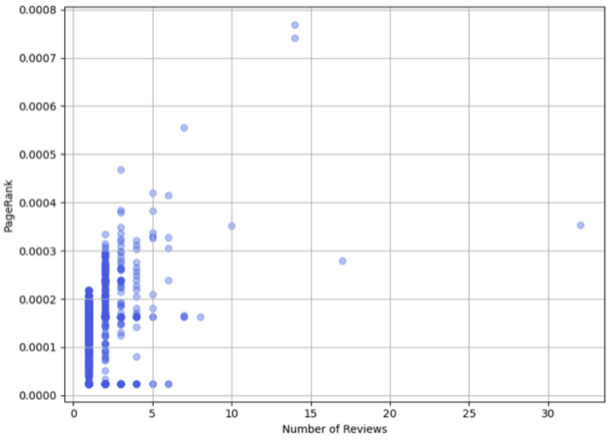
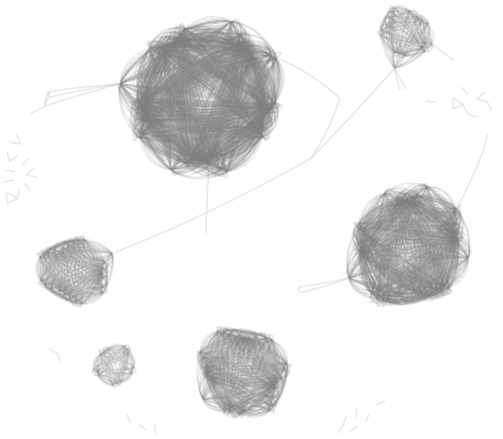
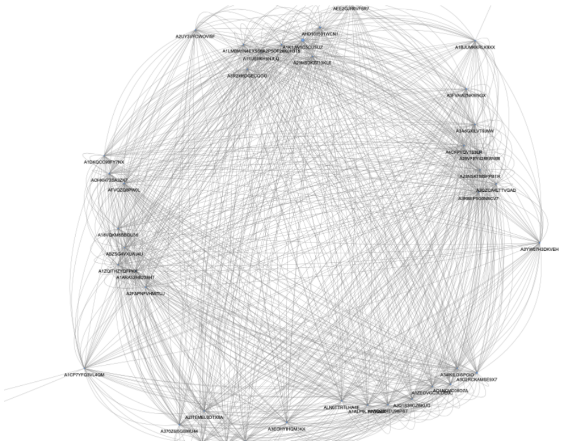


Figure 4: Pagerank vs Numer of Reviews

The scatterplot reveals a linear and weakly correlated pattern, where some users achieve high PageRank with few reviews, meaning that higher PageRank values do not depend on volume, rather by strategic connections.

The last section offers an interactive and navigable visualization of the PageRank subgraph using the Pyvis library. The graph is based on the previous user-user network implementation and it focuses on the top 300 users identified by the PageRank algorithm. Each node represents a user, and each edges represents a co-reviewd book. Finally, the dynamic graph is saved as an HTML file, allowing users to explore the graph visually.





The first image highlights the overall density connection among top users. The interaction is complex and sparse; it is possible to count six different community of influential users. The second image focuses on a single community, revealing the underlying structure of connection. The presence of few inter-cluster links highlight specific groups of preferences among readers, which means shared interests.

5 Conclusion

The project identifies the most influential group of users analyzing their interactions within a book review network. The implementation of the user-user graph and the application of PageRank algorithm reveals that the most interconnected users are not those who left the highest number of review but rather those with the most strategically positioned ones.

References

[1] Kaggle.
`https://www.kaggle.com/datasets/mohamedbakhet/amazon-books-reviews`

[2] Course Lecture Slides.

[3] Wikipedia.
`https://en.wikipedia.org/wiki/Link_analysis`

[4] Esri ArcGIS.
`https://doc.arcgis.com/en/insights/latest/analyze/link-analysis.htm`

The report must also contain the following declaration: “I/We declare that this material, which I/We now submit for assessment, is entirely my/our own work and has not been taken from the work of others, save and to the extent that such work has been cited and acknowledged within the text of my/our work, and including any code produced using generative AI systems. I/We understand that plagiarism, collusion, and copying are grave and serious offences in the university and accept the penalties that would be imposed should I engage in plagiarism, collusion or copying. This assignment, or any part of it, has not been previously submitted by me/us or any other person for assessment on this or any other course of study.