Algorithm for Massive Data – Link Analysis

Zhanat Nurlayeva – 30664A

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.

1 Introduction

This project applies link analysis techniques to the Amazon Books Review dataset with the objective of constructing a ranking system based on the PageRank index. Books are modelled as nodes in a co-review graph, and two books are connected when they have been reviewed by at least two distinct users. This representation captures how reading behaviour implicitly links titles through shared readership, enabling the study of structural importance inside the network.

User reviews naturally form a relational structure: when multiple users review the same pairs of books, they create connections that reveal common reading trajectories. Analysing this network allows us to identify books that occupy central positions in the collective behaviour of readers. Such a graph can also serve as the basis for recommendation approaches—if two books share many reviewers, they are likely to appeal to similar audiences—but in this project the focus remains on link-based ranking through PageRank.

PageRank provides a principled way to measure the global influence of each book. The algorithm evaluates a node not only by counting its connections, but by weighting those connections according to the importance of the neighbouring nodes. This follows the random-surfer interpretation: a hypothetical reader moves across the network by following co-review links, occasionally jumping to a random book through the teleportation mechanism. Books that consistently attract flows from influential neighbours receive higher PageRank scores. In this sense, a book becomes important not merely because it is widely reviewed, but because it serves as a structural hub connecting different reading communities.

The analysis proceeds as follows. After cleaning and reorganising the dataset—including title canonicalisation and reviewer-based filtering—a co-review graph is constructed and its structural properties examined. PageRank is then computed on the largest connected component, and its output is compared with traditional metrics such as degree, review count, and average rating. A temporal extension of PageRank is also explored to assess how recent user activity modifies the ranking. Finally, a subgraph of the top-20 PageRank books is visualised to illustrate the local connectivity patterns among the most influential titles.

2 Dataset Description

The analysis is based on the *Amazon Books Reviews* dataset (Kaggle, 2013), which contains large-scale user–book interactions collected from the Amazon product platform. Each record corresponds to a single review written by a user for a specific book and includes the following relevant fields: `User_id` (anonymous reviewer identifier), `Title` (raw book title as displayed on Amazon), `review/score` (numerical rating), and `review/time` (UNIX timestamp of the review). These variables form the behavioural core of the dataset; they enable the construction of a co-review network where books are linked when they are reviewed by the same users.

2.1 Dataset Characteristics

The raw dataset contains approximately 3,000,000 reviews written by about 1,000,000 unique users and includes 212,403 distinct title strings, spanning the period 1995–2013. The distribution of activity is highly skewed: the majority of users interact with only one or two books, and 90% of all books receive fewer than ten reviews. In particular, 75% of users contribute exactly one or two reviews. This sparsity strongly influences the structure of the co-review graph, since only a relatively small subset of titles has enough overlapping reviewers to form meaningful edges.

This extreme sparsity has important implications:

- most books do not have enough overlapping reviewers to produce meaningful edges;
- a small subset of titles contains dense co-review information;
- preprocessing must reduce the dataset to a structurally reliable subset before graph construction.

These patterns motivate the row-reduction and filtering strategies applied later in Chapter 3.

2.2 Noise and Inconsistencies in the Raw Titles

A substantial portion of the dataset exhibits inconsistencies typical of user-generated content: missing or invalid user identifiers, duplicated titles referring to the same book, inconsistent formatting (punctuation, capitalisation, subtitles, edition strings), multiple editions of identical works treated as distinct titles, and playlist-related variations such as audiobook or Kindle versions embedded directly into the title string. These issues do not affect individual reviews but have major consequences for network analysis, as they artificially fragment the same work into several nodes. This motivates the title canonicalisation pipeline described in Chapter 3.

3 Data Preprocessing

The preprocessing phase ensures that the dataset is structurally consistent and suitable for link-analysis techniques. The Amazon Books Review dataset contains several forms of noise typical of user-generated content, and careful preparation is required before constructing the co-review graph.

3.1 Basic Cleaning

The initial cleaning step addressed standard data-quality issues. Three operations were applied:

1. Removal of all reviews lacking essential identifiers (`User_id`, `Id`, or `Title`). This eliminated 561,982 invalid rows.
2. Replacement of missing category labels with a neutral placeholder (`Unknown`).
3. Removal of duplicate user–book pairs, keeping only the first occurrence.

After these operations, the dataset was reduced from approximately 3,000,000 rows to 2,397,419 unique user–book interactions. This cleaned table forms the basis for the subsequent reduction and normalisation steps.

3.2 Row Reduction Strategy

The distribution of interactions is extremely sparse, with most users reviewing very few books and most titles receiving limited engagement. To obtain a structurally meaningful network and to reduce computational cost, a targeted filtering strategy was adopted.

Two thresholds were applied:

- titles were kept only if they had at least 20 reviews;
- users were kept only if they had reviewed at least 3 different books.

This approach preserves the densest portion of the interaction graph while discarding entries that cannot contribute meaningful co-review information. Applying these filters resulted in a core dataset of 1,106,297 interactions, covering 18,421 titles and 156,762 users.

This reduced table (`df_core_raw`) ensures that subsequent steps operate on a stable and informative subset of the data.

3.3 Title Normalisation and Canonicalisation

A central difficulty of this dataset is that the same book appears under many slightly different title strings due to punctuation differences, edition or format markers (“paperback”, “Kindle edition”, “audiobook”), subtitles, or typographical noise. If left untreated, these variations fragment a single logical book across many nodes, distorting degree counts and PageRank.

To obtain a one–to–one mapping between logical books and graph nodes, an advanced multi–stage canonicalisation pipeline was implemented.

(i) Basic text cleaning. Each title undergoes lowercasing, removal of punctuation, normalisation of whitespace, and filtering of non-informative tokens such as *paperback*, *hardcover*, *edition*, *audiobook*, and similar format indicators. This step removes superficial noise and prepares strings for comparison, but it does not yet resolve deeper structural inconsistencies across title variants.

(ii) Extraction of a simplified core representation. A secondary “core” version of each title is produced by stripping subtitles and parenthetical fragments, removing residual punctuation, and keeping only alphanumeric tokens. This reduces many remaining formal differences, creating a more stable basis for recognising duplicates; however, titles can still differ in wording while referring to the same book, motivating an additional similarity-based procedure.

(iii)) String-similarity clustering. To identify variant titles that refer to the same book, a similarity score is computed between titles using both Jaro–Winkler and Levenshtein distances, applied to the cleaned and core text forms. Variants exceeding a similarity threshold are grouped into clusters. While this captures a large portion of duplicated titles, string-based similarity alone cannot reliably detect semantically equivalent titles with substantial textual variation.

(iv) Selection of a canonical representative. For each cluster, all variants are mapped to a single canonical title chosen through a scoring rule combining:

- popularity (number of occurrences in the dataset),
- textual brevity,
- absence of format markers,
- reduced punctuation.

This ensures that frequently used and semantically neutral titles are preferred. Although this step standardises most duplicate groups, some high-profile books still appear across heterogenous title forms that cannot be safely merged through string similarity alone.

(5) Popularity-aware semantic refinement (top 5,000 titles). To further reduce residual duplication, a semantic refinement step was applied to the most-reviewed titles. Sentence-transformer embeddings (`all-MiniLM-L6-v2`) were computed for the 5,000 most frequent canonical candidates, and cosine similarities were used to identify additional clusters of semantically equivalent titles (similarity ≥ 0.85). This final step resolves duplicates that string-based methods cannot capture (e.g. paraphrased subtitles or alternate long-form titles), while ensuring scalability by restricting semantic matching to the most influential portion of the catalogue and avoiding over-merging. .

Outcome. After all stages, each original title is mapped deterministically to a stable canonical form. In the core dataset, the number of distinct raw titles is reduced from 18,421 to 16,160 canonical titles. Duplicate user–title interactions created by merging were removed to re-establish a one-to-one user–book relation.

This canonical interaction table (`df_core`) is the definitive dataset used for reviewer-set construction and graph generation.

4 Network Construction

This chapter describes how the book–book co-review network is built from the cleaned canonical interaction table `df_core`. The target output is an undirected weighted graph $G = (V, E)$ where nodes are canonical book titles and edges represent shared reviewers.

4.1 Graph Definition

Let V be the set of canonical titles (`Title_canon`). For two books $b_i, b_j \in V$, we add an undirected edge $(b_i, b_j) \in E$ if the two titles share at least k distinct reviewers, with $k = 2$ (Project 3 requirement). The edge weight is the exact number of common reviewers:

$$w(b_i, b_j) = \left| U(b_i) \cap U(b_j) \right|,$$

where $U(b)$ is the set of users who reviewed book b. This produces a weighted co-review network in which large weights correspond to strong audience overlap.

4.2 Reviewer-Set Construction

The interaction table `df_core` is inverted into reviewer sets:

$$U(b) = \{u : (u, b) \text{ appears in } \texttt{df\_core}\}.$$

In the final canonical core, this step yields 16,160 titles with at least one reviewer. The reviewer-set sizes are highly skewed: the median book has 16 reviewers, the 90th percentile has 67, and the maximum reaches 3,570 reviewers. This skew motivates using a scalable pair-generation strategy rather than checking all book pairs.

4.3 LSH-Based Candidate Generation via MinHash

Computing overlaps for all book pairs would require $O(|V|^2)$ comparisons, which is infeasible at $|V| = 16{,}160$. To scale candidate generation, each reviewer set $U(b)$ is compressed into a MinHash signature of length $p = 128$. MinHash preserves Jaccard similarity approximately:

$$J(U(b_i), U(b_j)) = \frac{|U(b_i) \cap U(b_j)|}{|U(b_i) \cup U(b_j)|}.$$

The signatures are indexed in an LSH structure with threshold $\tau = 0.1$. For each title b, the LSH query returns only a reduced list of candidate neighbours whose Jaccard similarity is potentially non-negligible. This converts the quadratic pair enumeration into a much smaller set of plausible pairs.

4.4 Exact Overlap and Edge Construction

For each candidate pair (b_i, b_j) returned by LSH, the pipeline computes the exact intersection size $|U(b_i) \cap U(b_j)|$ and adds an edge only if it satisfies the project threshold $k = 2$. This ensures correctness of the final graph: LSH is used only for *candidate generation*, while edge inclusion is decided by exact overlap.

With $k = 2$, the final graph contains:

$$|V| = 16{,}160, \qquad |E| = 40{,}813.$$

Edge weights are strongly long-tailed: the median edge has weight 2, the 75th percentile is 4, the 90th percentile is 9, and the maximum edge weight reaches 1,389 shared reviewers.

4.5 Network Summary and Largest Connected Component

The resulting co-review graph is very sparse:

$$\text{density}(G) \approx 0.000313.$$

Degree statistics show a typical heavy-tailed pattern (projection graph): the mean degree is 5.05, the median is 1, the maximum is 104, and many titles remain isolated.

The graph decomposes into 6,509 connected components. The largest connected component (LCC) contains 8,687 nodes, corresponding to 53.76% of all titles. Since PageRank is meaningful as a global importance measure only within a connected (or strongly connected) structure, all ranking experiments are performed on the LCC subgraph G_{LCC}.

5 Baseline PageRank

PageRank is used to rank books according to their structural importance within the co-review network. It models a random walk on the graph and assigns higher scores to nodes that are connected to other important nodes.

Given the weighted graph $G = (V, E)$, PageRank is defined as the stationary distribution of a Markov chain. For a node $v \in V$, its PageRank score $\mathrm{PR}(v)$ satisfies:

$$\mathrm{PR}(v) = \frac{1-\alpha}{|V|} + \alpha \sum_{u \in N(v)} \frac{w_{uv}}{\sum_{k \in N(u)} w_{uk}} \mathrm{PR}(u),$$

where $\alpha \in (0,1)$ is the damping factor, $N(v)$ denotes the neighbours of v, and w_{uv} is the edge weight between nodes u and v. Throughout the experiments, the standard value $\alpha = 0.85$ is used.

5.1 Computation on the Largest Connected Component

As shown in Chapter 4, the co-review graph is highly fragmented. Since PageRank requires a connected structure to propagate importance globally, the algorithm is applied only to the largest connected component G_{LCC}.

In this dataset, G_{LCC} contains 8,687 nodes and 39,661 edges, covering approximately 54% of all canonical titles. Nodes outside the LCC are excluded from the ranking and implicitly receive zero PageRank.

5.2 Weighted Power Iteration

PageRank is computed using power iteration on the weighted adjacency structure of G_{LCC}. Edge weights are incorporated into the transition probabilities, so that transitions along edges supported by many shared reviewers are more likely.

Starting from a uniform initial distribution, the PageRank vector is iteratively updated until convergence to the stationary distribution. Convergence is reached when the ℓ_1-difference between successive iterations falls below a fixed tolerance.

This procedure yields a stable ranking that reflects both local connectivity (degree) and global position within the co-review network.

5.3 Temporal Decay PageRank

From an algorithmic perspective, the temporal variant is implemented as a *personalized PageRank* computation. Rather than modifying edge weights directly, temporal information is incorporated through a non-uniform teleportation distribution. Each book is assigned a recency score based on the timestamps of its reviews, and these scores define the personalization vector used in the PageRank iteration.

Formally, the PageRank update becomes

$$\pi = \alpha \mathbf{P}^{\top} \pi + (1-\alpha) v,$$

where $\mathbf{P}$ is the transition matrix derived from the co-review graph, α is the damping factor, and v is a probability distribution biased toward recently reviewed books. This formulation preserves the standard power-iteration algorithm and scalability properties discussed in the course, while shifting probability mass toward nodes with higher recent activity.

As a result, temporal PageRank highlights *current influence* in the network, in contrast to the baseline PageRank, which captures long-term structural prestige.

6 Results and Analysis

This section presents the main results obtained from the PageRank-based analysis of the Amazon Books co-review network.

Table 1 reports the top 20 books ranked by PageRank score. In addition to PageRank, the table shows each book's degree in the LCC and its total number of reviews after canonicalisation.

Tabell 1: Top 20 books by PageRank score

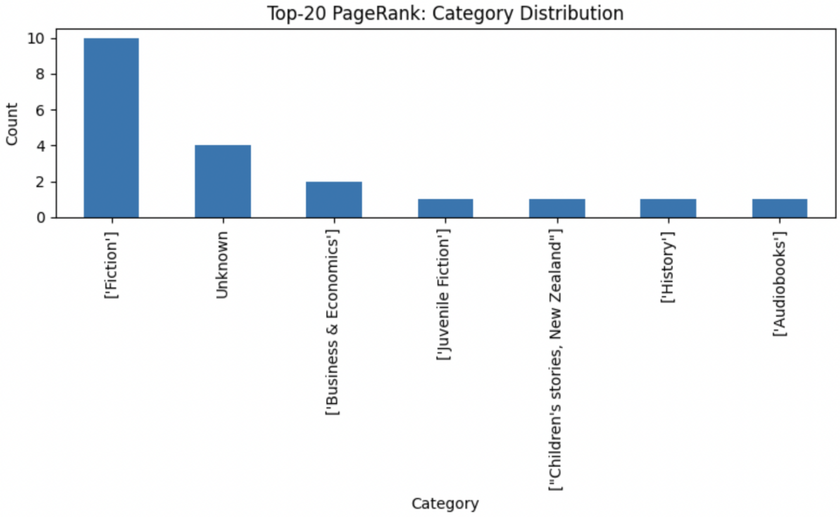
Rank	Title	PR	Deg	Rev
1	City of Bones	0.00113	68	178
2	1st to Die: A Novel	0.00087	68	311
3	Gilead	0.00077	87	224
4	Echo Park (SIGNED)	0.00074	70	191
5	Persuader (Jack Reacher, No. 7)	0.00073	62	151
6	The Summons	0.00070	36	298
7	The Bridge of San Luis Rey	0.00066	83	73
8	The Big Bad Wolf	0.00065	52	153
9	Girl with a Pearl Earring	0.00064	26	423
10	Prince of Fire	0.00062	91	68
11	Moral Intelligence	0.00060	43	16
12	Equal Rites	0.00059	55	93
13	One Shot (Jack Reacher, No. 9)	0.00059	26	212
14	Mayflower	0.00058	62	224
15	Why Great Leaders Don't Take Yes	0.00057	38	17
16	Sharpe's Triumph	0.00056	51	107
17	Bet Me	0.00056	67	213
18	Sharpe's Havoc (Book VII)	0.00056	90	31
19	Capital Liberalization in Italy	0.00056	90	31
20	Sharpe's Havoc (Campaign in N. Po)	0.00056	90	31

The top-ranked titles are not simply the most reviewed books. Several books with moderate review counts achieve high PageRank due to strong structural connectivity, acting as bridges between different reader groups. This confirms that PageRank captures network position rather than raw popularity.

6.1 Genre Composition of Top-Ranked Books

Figure 1 reports the category distribution of the top 20 books ranked by PageRank. The ranking is dominated by *Fiction*, which accounts for half of the top positions, while other genres such as *Business & Economics*, *History*, and *Audiobooks* appear more sparsely.

This confirms that high PageRank scores are primarily associated with books that connect large fiction-oriented reader communities, while non-fiction titles tend to occupy more peripheral positions in the reviewer-overlap network.



Figur 1: Category distribution of the top 20 books by PageRank

6.2 Network Structure and Inequality of Influence

The PageRank distribution exhibits strong concentration. Approximately 35.9% of books account for 80% of the total PageRank mass, indicating a pronounced inequality of influence. This pattern is consistent with heavy-tailed degree distributions commonly observed in projection graphs derived from bipartite data.

6.3 Correlation with Traditional Popularity Measures

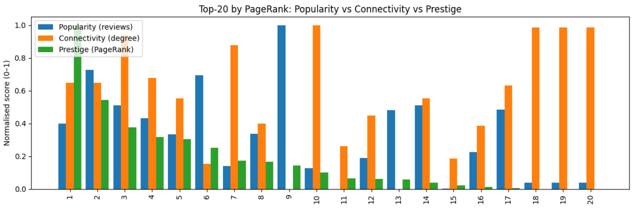
To assess whether PageRank reproduces simple popularity metrics, correlations were computed between PageRank and traditional indicators.

Tabell 2: Correlation between PageRank and traditional metrics

Metric Pair	Correlation
PageRank vs Degree	0.741
PageRank vs Number of Reviews	0.236
PageRank vs Average Rating	0.015

PageRank is strongly correlated with degree, confirming its structural nature. The weak correlation with review count and the negligible correlation with average rating indicate that PageRank captures network importance rather than perceived quality or sheer volume of reviews.

The figure highlights that highly ranked books under PageRank are not necessarily the most reviewed ones. While degree closely tracks PageRank, popularity alone fails to explain prestige. Average rating is excluded from the visualization, as it reflects perceived quality



Figur 2: Normalized comparison of popularity (number of reviews), connectivity (degree), and prestige (PageRank) for the top–20 PageRank titles.

rather than structural position in the co-review network and is therefore analysed separately via correlation.

6.4 Comparison with Other Centrality Measures

To complement PageRank, additional centrality measures were computed on the largest connected component (LCC), namely degree centrality, closeness centrality, and betweenness centrality. Each measure captures a different notion of importance within the co-review network.

Tabell 3: Top books by different centrality measures

(a) Degree centrality

Rank	Title	Value
1	Heat Wave	0.0120
2	I’ll Be Watching You	0.0117
3	Some Sunday	0.0112
4	Don’t Want No Sugar	0.0108
5	Prince of Fire	0.0105

(b) Closeness centrality

Rank	Title	Value
1	Executive Power	0.2564
2	Capital Liberalization in Italy	0.2557
3	Sharpe’s Havoc (Campaign in N. Po)	0.2557
4	Sharpe’s Havoc (Book VII)	0.2557
5	Edge of Danger	0.2548

(c) Betweenness centrality

Rank	Title	Value
1	Moral Intelligence	0.0314
2	Arthur & George (Sound Library)	0.0224
3	Heat Wave	0.0218
4	How to Change the World	0.0216
5	Battlestar Galactica	0.0215

Degree centrality highlights books that are co-reviewed with a large number of other titles, identifying local hubs in the reader overlap network.

Closeness centrality identifies books that are, on average, at short distance from all others, indicating efficient access to the entire network.

Betweenness centrality reveals books that act as bridges between otherwise weakly connected parts of the network, playing a key role in information flow across communities.

The rankings differ substantially across centrality measures. While PageRank emphasises global structural importance, degree centrality highlights local hubs, closeness identifies globally accessible nodes, and betweenness reveals inter-community connectors. This confirms that influence in the co-review network is multi-dimensional and cannot be captured by a single metric.

6.5 Community Structure

Community detection was performed on the largest connected component using the Louvain algorithm. The method identifies groups of books that are densely connected through shared reviewers, revealing latent thematic structures in the co-review network.

A total of 77 communities were detected. Table 4 reports the ten largest communities, together with their size and dominant genre.

Tabell 4: Largest communities detected in the co-review graph

Community ID	Number of Books	Total PageRank	Dominant Genre
15	969	0.1176	Fiction
74	625	0.0679	Business & Economics
53	613	0.0716	Fiction
37	555	0.0551	Fiction
6	504	0.0544	Fiction
67	488	0.0498	History
5	395	0.0500	Fiction
17	357	0.0485	Fiction
27	335	0.0431	Fiction
0	285	0.0339	Fiction

The results show a strong modular organisation of the network. Most large communities are dominated by *Fiction*, indicating that reader overlap is largely driven by genre-specific consumption. Smaller but structurally relevant communities emerge around *Business & Economics* and *History*, reflecting more specialised reading audiences.

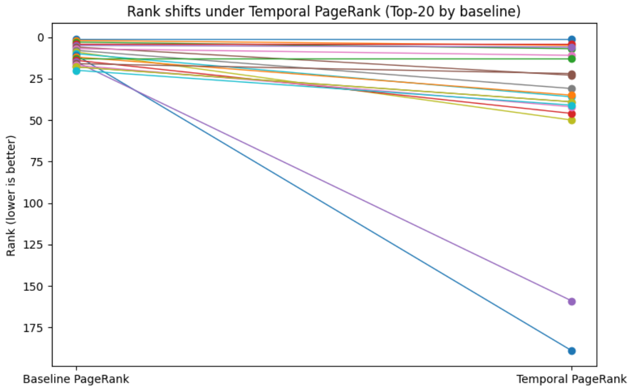
This confirms that the co-review graph is not random but organised into well-defined clusters, with books acting as hubs within their respective thematic communities.

6.6 Temporal PageRank Dynamics

A temporally weighted PageRank variant is computed to emphasise recent reviewing activity by incorporating an exponential decay on review timestamps. This modification shifts the interpretation of PageRank from long-term structural prestige to *current influence*.

Figure 3 compares baseline PageRank ranks with temporal PageRank ranks for the top–20 books according to the baseline ranking. Each line represents a book, connecting its baseline rank to its temporal rank; upward movements indicate a loss of influence under temporal weighting, while downward movements indicate gains driven by recent engagement.

Most titles remain close to their baseline positions, indicating that PageRank is generally stable over time. However, a small number of books exhibit pronounced rank shifts, either improving substantially due to recent bursts of attention or declining as their influence is mainly supported by older reviews. This pattern shows that temporal PageRank is able to distinguish between *structural classics* and books with *current momentum*, a distinction that is not visible in the static network alone.



Figur 3: Rank shifts under temporal PageRank for the top–20 books by baseline PageRank. Lines connect baseline ranks to temporal ranks; lower values correspond to higher influence.

6.7 Local Co-Review Structure of Top-Ranked Books

To complement the quantitative ranking analysis, we inspect the local co-review structure of the highest-ranked books. Figure 4 shows the induced subgraph formed by the top–20 titles according to PageRank, extracted from the largest connected component.

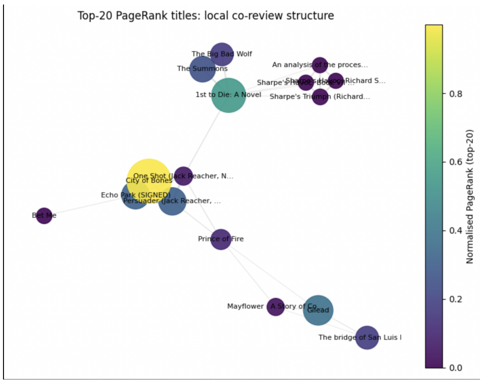
In this visualisation, nodes correspond to books and edges represent co-review relationships (i.e. at least two shared reviewers). Node size and colour reflect the relative PageRank score within the top–20 set, with darker and larger nodes indicating higher prestige.

The figure highlights that top-ranked books are not isolated hubs. Instead, they tend to occupy structurally central positions that connect multiple neighbourhoods in the co-review network. Several titles act as bridges between otherwise weakly connected clusters, which explains why they achieve high PageRank despite having only moderate numbers of reviews.

This local network perspective provides a qualitative confirmation that PageRank captures structural importance arising from audience overlap, rather than simple popularity alone.

7 Discussion and Limitations

The PageRank-based ranking reveals clear structural patterns in reader behaviour. Top-ranked books are not necessarily those with the largest number of reviews, but rather those occupying central positions in the co-review network. In particular, several titles act as bridges between



Figur 4: Local co-review structure of the top–20 PageRank titles. Nodes represent books and edges indicate shared reviewers. Node size and colour are proportional to the normalised PageRank score within the top–20 set.

different reader communities, indicating broad appeal across audiences. This confirms that PageRank captures *network influence* rather than raw popularity.

Despite these insights, several limitations should be noted. First, PageRank is computed only on the largest connected component, which contains approximately 54% of all canonical titles. Books outside this component receive zero PageRank, even if they are locally important within smaller communities. Second, edges are defined using a minimum threshold of shared reviewers, which abstracts away finer distinctions in connection strength. Finally, although a temporal variant of PageRank was explored, the network remains an aggregate representation and does not fully capture the dynamic evolution of reading behaviour over time.

8 Conclusions and Future Work

This project applied link analysis techniques to the Amazon Books dataset in order to identify influential titles through PageRank. By modelling books as nodes connected via shared reviewers, the analysis uncovered structural importance that is not visible through traditional popularity measures alone.

From a methodological perspective, the project demonstrates the effectiveness of: (i) a scalable title canonicalisation pipeline, (ii) efficient construction of a sparse co-review graph, and (iii) PageRank-based ranking on large networks.

Overall, the results show that network-based metrics can complement existing recommendation signals by highlighting books that connect multiple reader communities. Future work could extend this framework by incorporating richer edge weights, fully dynamic temporal models, or hybrid approaches combining network structure with content-based features.