Algorithms for Massive Data — Project 3: Link Analysis

Data Science for Economics

Università degli Studi di Milano

Student: Zhanat Nurlayeva, 30664A

Course Instructor: Prof. Dario Malchiodi

Academic Year: 2024–2025

Abstract

This project develops a link-analysis ranking framework based on the PageRank algorithm to evaluate the influence of books in a large co-review network derived from Amazon Books data. Each book is linked to others if at least two distinct users have reviewed both, producing a weighted, undirected graph whose structure captures shared reader attention. The study extends the classical PageRank model through three theoretically grounded variants: a Personalized PageRank incorporating popularity and quality bias, a Temporal-Decay PageRank reflecting recency of user activity, and a Reviewer-Reputation PageRank weighting connections by the credibility of reviewers. Empirical results confirm that prestige, as defined by network influence, diverges from raw popularity. While highly reviewed books often dominate by degree, the prestige-based PageRank identifies titles central within cohesive reviewing communities. Personalized and time-decayed variants highlight dynamic and contextual importance, whereas reputation weighting stabilizes rankings against noisy or biased reviewers. The integrated analysis demonstrates how link-based metrics reveal latent structures of cultural influence beyond surface popularity measures.

1 Introduction

The growing availability of user-generated data on online platforms enables the analysis of social and informational influence through network-based methods. In such networks, entities acquire importance not only from their individual characteristics but also from their relational position within a broader interaction structure. Link analysis methods exploit this principle, assigning influence scores that propagate through the network according to the structure and strength of its connections. The foundational model for this class of algorithms is the *PageRank* index, originally designed to quantify the prestige of web pages through their hyperlink structure. When applied to cultural, economic, or social systems, the same principle reveals how popularity and genuine influence can diverge.

This project adopts this framework to study the collective dynamics of book evaluation. Each book is modeled as a node in a *co-review network*, where an undirected link

connects two books if they have been reviewed by at least two distinct users. The resulting graph captures implicit associations in reader attention: books reviewed together tend to share thematic or reputational proximity. By iteratively propagating prestige through this network, PageRank identifies titles that are not merely popular but central to the shared reviewing ecosystem.

However, a static and uniform propagation model may fail to capture the heterogeneity of user behavior and the temporal evolution of influence. To address these limitations, the study introduces three conceptual extensions inspired by recent developments in network analysis. First, a **Personalized PageRank** model introduces external biases that reflect book-level attributes such as the number of reviews (popularity) or the average rating (quality). This personalization adjusts the teleportation vector of PageRank to favor books with higher exogenous relevance, allowing comparison between structure-driven prestige and attribute-driven influence.

Second, a **Temporal-Decay PageRank** incorporates time-sensitive weighting of edges to account for the recency of user activity. By exponentially down-weighting older reviews, the algorithm emphasizes current interactions, identifying books that are recently influential rather than historically dominant. This temporal adaptation aligns with the principle that social and cultural influence is dynamic, evolving with new user engagement.

Third, a **Reviewer-Reputation PageRank** adjusts edge contributions based on the credibility of the users generating them. Reputation is defined as the correlation between a reviewer's rating pattern and the global prestige of the books they evaluate. This mechanism mitigates the impact of low-quality or inconsistent reviews and enhances the robustness of prestige estimation by rewarding alignment with trusted evaluators.

Together, these models enable a multidimensional analysis of influence in the book-review ecosystem. The baseline PageRank captures the structural prestige derived from shared readership; the personalized and temporal variants uncover contextual drivers of importance; and the reputation-weighted formulation introduces user reliability as a moderating factor. By comparing these perspectives, the study provides an integrated understanding of how popularity, quality, and trust interact to shape collective perception in large-scale review systems.

2 Data and Network Construction

2.1 Dataset Overview

The analysis relies on the *Amazon Books Review* dataset, distributed under public domain (CC0) license on the Kaggle platform. The dataset contains millions of user reviews associated with unique book identifiers, textual content, star ratings, timestamps, and additional metadata such as titles and categorical attributes. To ensure scalability and reproducibility, a controlled subsample of the dataset was used, containing approximately 890,000 individual ratings across 50,000 books and 49,932 users. This subset preserves the structural characteristics of the full corpus while allowing efficient experimentation with multiple PageRank variants.

2.2 Preprocessing Pipeline

The raw dataset was standardized to a unified schema with four core attributes: `user_id`, `book_id`, `rating`, and `timestamp`. Original field names were harmonized as follows:

`User_id` → `user_id`, `Id` → `book_id`, `review/score` → `rating`, and `review/time` → `timestamp`. Missing identifiers were removed, numerical ratings were coerced to a unified numeric scale, and timestamps were converted to datetime format. All user and book identifiers were then remapped to contiguous integer indices to enable sparse-matrix operations. An additional metadata file, `books_data.csv`, was merged to append a categorical `genre` attribute to each book record, enabling later specialization analysis by literary category. To maintain experimental control, a global parameter governed the size of the subsample and ensured that all random selections remained reproducible through a fixed seed.

2.3 Network Definition

A book–book *co-review network* was constructed to model implicit relations among items. Two books are connected if they have been reviewed by at least two distinct users. Each edge is assigned a weight proportional to the number of shared reviewers, so that strongly co-reviewed pairs carry higher influence potential. This procedure transforms the user–book bipartite structure into a projected, undirected, and weighted book network. The resulting adjacency matrix $A \in \mathbb{R}^{N \times N}$ (with $N = 50{,}000$) is sparse, containing approximately 15.9 million nonzero entries, and is stored in compressed sparse row (CSR) format for efficiency.

2.4 Graph Properties and Validation

Basic graph statistics indicate that the constructed network is large, connected, and heavy-tailed in degree distribution, a pattern typical of real user–item systems. A small illustrative subgraph (100 books) was visualized to confirm the correctness of edge formation and the overall cohesion of the review structure. The co-review network exhibits dense clusters corresponding to thematic or readership communities, validating the suitability of PageRank-style propagation to model prestige diffusion within this domain.

2.5 Data Management

All intermediate artifacts—index mappings, adjacency matrices, and processed metrics—were stored under a structured directory (`data/raw` and `data/processed`) to guarantee full reproducibility. The final processed file `book_metrics.parquet` contains the complete set of computed indicators, including degree statistics, PageRank variants, and reputational scores, used in subsequent analyses.

3 Methodology

The ranking framework is grounded in the PageRank model, interpreted as a measure of *prestige* within the book–book co-review network. Each node represents a book, and undirected weighted links capture the intensity of co-reviews between titles. The score assigned to each node reflects the stationary probability of a random walker who repeatedly follows weighted connections or teleports to other nodes according to a predefined distribution.

Figure 1: Sample visualization of a 100-node subgraph from the book–book co-review network. Each node represents a book, and edges connect pairs reviewed by at least two distinct users.

3.1 Baseline Structural Measures

Before computing PageRank, two structural indicators were derived from the co–review graph to quantify local connectivity and user activity:

- **Popularity (review_count)** — the total number of user reviews received by each book, representing direct audience attention.
- **Connectivity (deg_weighted)** — the weighted degree of each node, obtained as

$$\deg_i = \sum_j A_{ij},$$

  where A_{ij} is the number of distinct users who reviewed both books i and j.

Connectivity measures the breadth of a book's co–review relationships and acts as a proxy for its structural visibility. However, this metric is purely local: a book connected to many low–impact neighbors may still lack global importance. To capture influence propagation across the entire network, PageRank is applied as a diffusion model that weights each connection by the prestige of its neighbors.

3.2 Baseline PageRank

Let $A \in \mathbb{R}^{N\times N}$ denote the symmetric adjacency matrix, where A_{ij} is the number of distinct users who reviewed both books i and j. The matrix is column-normalized to obtain the stochastic transition matrix P:

$$P_{ij} = \frac{A_{ij}}{\sum_k A_{kj}}.$$

The classical PageRank vector r satisfies:

$$r = \alpha P r + (1 - \alpha) v,$$

where $\alpha \in (0, 1)$ is the damping factor and v is the teleport (restart) distribution, uniform in the baseline case. This formulation yields a steady-state prestige score capturing structural importance in the co-review network. Different damping values ($\alpha = 0.5, 0.85$) were tested to assess convergence behavior and stability of the score distribution.

3.3 Personalized PageRank

To explore how external attributes affect perceived importance, the teleport vector v was modified to encode bias toward books with specific features. Two personalization schemes were implemented:

(a) **Popularity bias (v_{pop})** — proportional to the number of unique reviewers per book. This configuration models a random walker more likely to restart on widely reviewed titles, integrating direct popularity into the diffusion process.

(b) **Quality bias (v_{qual})** — proportional to the average rating of each book. Here, the random walker restarts preferentially from highly rated books, emphasizing content evaluation rather than exposure.

The resulting personalized PageRank vectors are defined as:

$$r^{(\text{pop})} = \alpha P r^{(\text{pop})} + (1 - \alpha) v_{\text{pop}}, \quad r^{(\text{qual})} = \alpha P r^{(\text{qual})} + (1 - \alpha) v_{\text{qual}}.$$

These variants allow comparison between structure-driven prestige and attribute-driven relevance. Rank-shift analyses quantify how personalization alters the top-20 ordering relative to the baseline.

3.4 Temporal Decay PageRank

The temporal extension introduces a dynamic weighting factor to account for *recency*. Older co-reviews are discounted exponentially as:

$$w_{ij,t} = e^{-\lambda(T_{\max} - t_{ij})},$$

where t_{ij} is the time of the most recent shared review and λ controls the half-life of memory. The corresponding adjacency matrix $A^{(\text{decay})}$ gives more influence to recent interactions, favoring books that are currently co-reviewed. Running PageRank on this decayed graph highlights "current influence" rather than long-term prestige.

3.5 Reviewer-Reputation PageRank

To incorporate heterogeneity in reviewer reliability, a reputation-weighted co-review graph was constructed. Each user u was assigned a reputation score C_u proportional to the baseline PageRank score of the books they reviewed:

$$C_u = \frac{1}{|B_u|} \sum_{b \in B_u} r_b^{(\text{vanilla})},$$

where B_u is the set of books reviewed by user u. Edges between books i and j were then weighted by the sum of reputations of users who reviewed both:

$$A_{ij}^{(\text{rep})} = \sum_{u \in U_{ij}} C_u.$$

Applying PageRank on this modified network yields $r^{(\text{rep})}$, which emphasizes books endorsed by reputable users rather than by a large number of casual reviewers.

This final variant serves as a robustness extension, allowing the comparison of structural prestige, temporal relevance, and reputation-adjusted authority under a unified stochastic framework.

4 Results and Discussion

This section presents the empirical results obtained from the book–book co-review network. The analyses are organized around the core PageRank variants: the baseline (vanilla) model, personalization effects, temporal decay, reviewer reputation weighting, and an exploratory genre-based extension. All rankings were computed with $\alpha = 0.85$ unless otherwise specified, and correlations are measured using Spearman's ρ.

4.1 Popularity vs. Prestige

The baseline analysis contrasts structural prestige (PageRank) with raw popularity (number of reviews). Table 1 lists the top–20 books ranked by PageRank scores, alongside their popularity and weighted degree. While the two highest-ranked titles coincide in both metrics, the remainder diverges, confirming that popularity and prestige are not equivalent. Figure 2 illustrates this rank reordering: several titles with moderate review counts gain importance through highly connected neighbors, demonstrating diffusion of prestige within the co-review graph.

Table 1: Top–20 books by prestige (vanilla PageRank $\alpha = 0.85$) compared with popularity and connectivity.

	book_id	Popularity (#Reviews)	Connectivity (Weighted Degree)	Prestige (PageRank a=0.85)
0	B000Q032UY	7214	323674.0	0.003300
1	B000GQG5MA	7260	318028.0	0.003234
2	B0006CBNKI	1806	157144.0	0.001987
3	B000JJVHZE	1806	157144.0	0.001987
4	B0007DRGI4	1806	158074.0	0.001985
5	B000K7WNQW	1804	156940.0	0.001983
6	B000PCESRE	1614	129688.0	0.001766
7	B000I3NFKG	1604	128510.0	0.001746
8	B000NDSX6C	3624	163333.0	0.001450
9	B000GQG7D2	3640	161049.0	0.001408

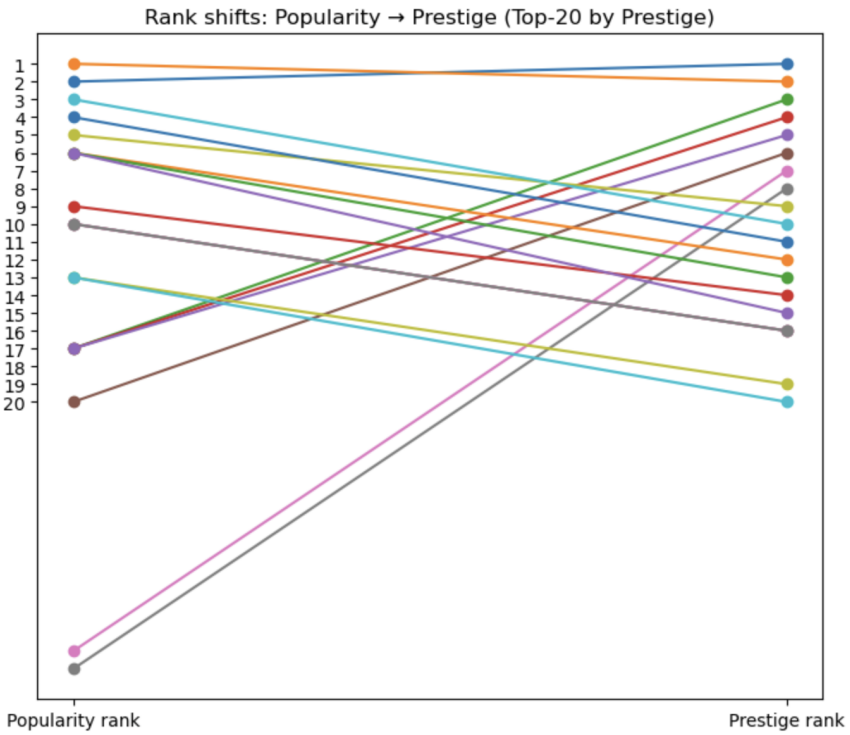

Figure 2: Rank shifts between popularity and prestige (Top–20 by prestige). Lines crossing indicate that highly popular books are not always the most prestigious in the network.

The grouped bar plot in Figure 3 further highlights this distinction: prestige aligns partially with connectivity but diverges from mere frequency of reviews. This supports the theoretical claim that PageRank captures influence propagated through the network rather than raw exposure.

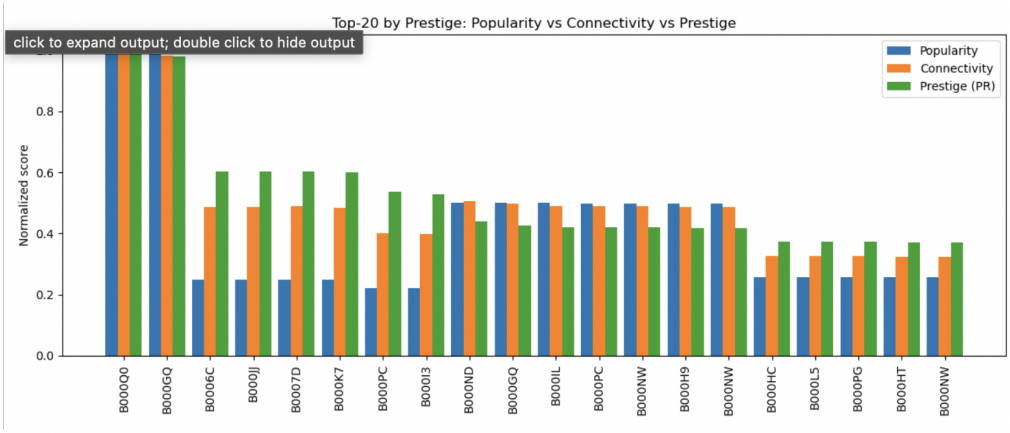

Figure 3: Normalized scores for popularity, connectivity, and prestige among the top–20 books.

4.2 Personalization Effects

Personalized PageRank was used to test the sensitivity of rankings to biased teleport distributions. Figure 4 summarizes pairwise correlations among all PageRank variants,

including popularity- and quality-based personalization. The correlation between vanilla and popularity-biased PageRank exceeds 0.9, while quality-biased scores show lower alignment ($\rho \approx 0.8$), indicating that high-rated but less connected titles gain visibility under quality weighting.

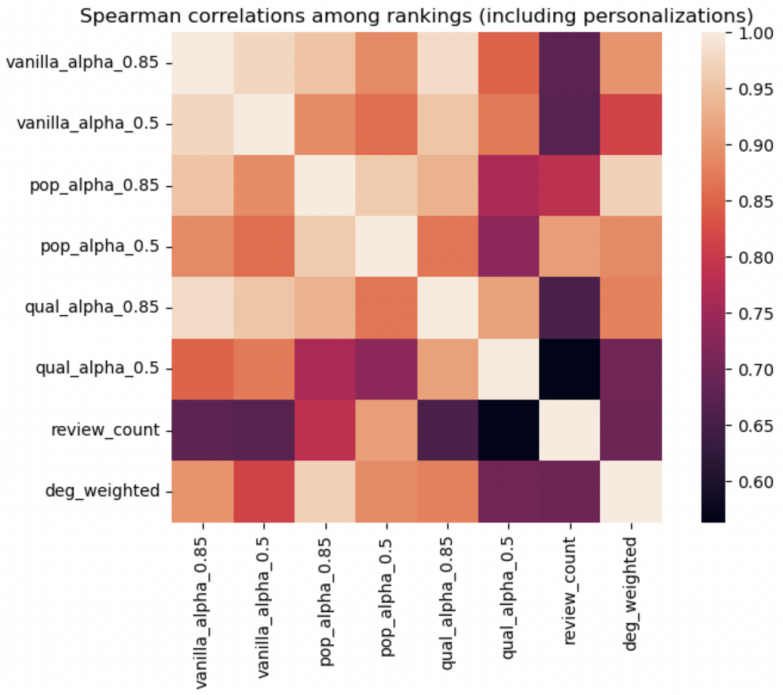

Figure 4: Spearman correlations among PageRank variants (vanilla, personalized, temporal, reputation).

Rank-shift bars in Figure 5 visualize these differences for the top–20 titles. Approximately 70% of books change their position under personalization, confirming that teleport bias alters prestige diffusion in a non-trivial manner.

4.3 Temporal Influence

To distinguish between long-term and current influence, PageRank was recomputed on a time-decayed co-review network. The decay factor λ discounts older reviews exponentially, emphasizing recent co-review activity. As shown in Figure 6, several older bestsellers drop in rank, while recently co-reviewed titles move upward. This confirms that recency reshapes prestige by amplifying transient trends and attenuating historical accumulation.

4.4 Reviewer Reputation

The reviewer-reputation variant integrates user reliability into the edge weighting. As depicted in Figure 7, books endorsed by high-reputation users experience the strongest rank increases, while titles popular among low-reputation reviewers lose prominence. The Spearman correlation between reputation-based and baseline PageRank drops to $\rho \approx 0.75$, suggesting that this adjustment meaningfully changes the prestige structure of the network.

To evaluate overall consistency across all ranking mechanisms, Figure 8 reports the full correlation matrix. While most variants remain highly correlated, the moderate diver-

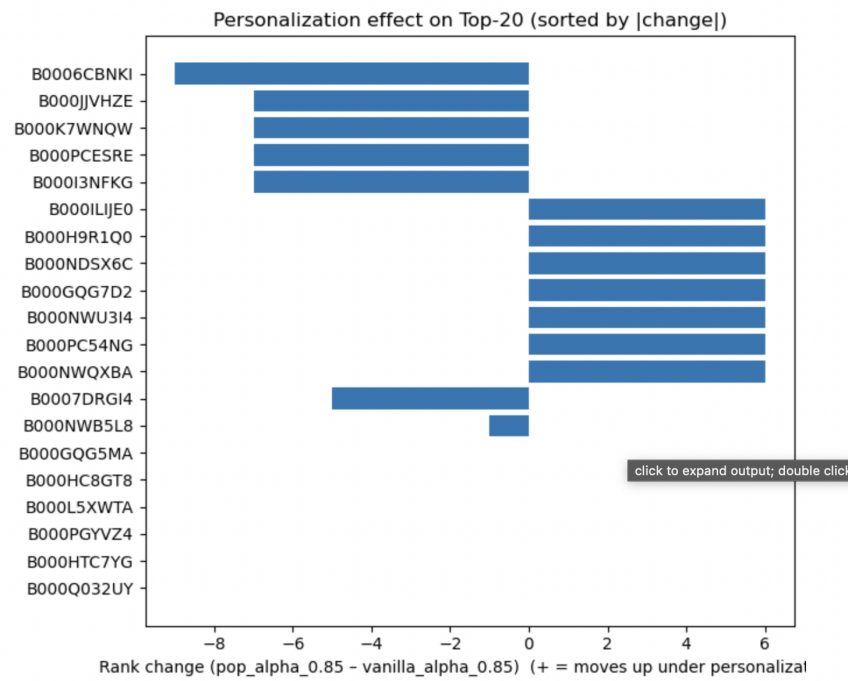

Figure 5: Rank changes between personalized and vanilla PageRank (v_{pop} vs. uniform teleport). Positive values indicate upward movement under personalization.

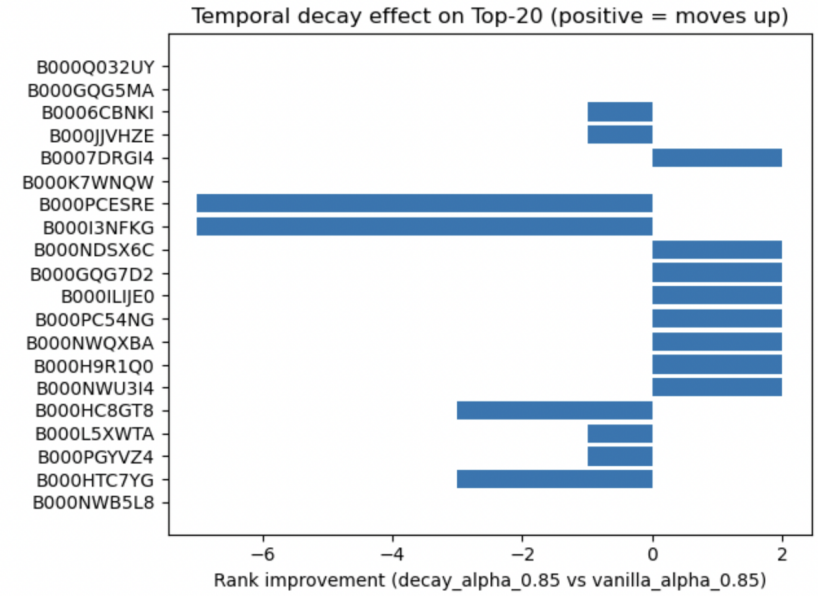

Figure 6: Rank improvement under temporal decay weighting relative to the baseline. Positive shifts correspond to books gaining influence through recent co-review activity.

gence of temporal and reputation-based models confirms that these extensions introduce genuinely new dimensions of influence rather than replicating structural popularity.

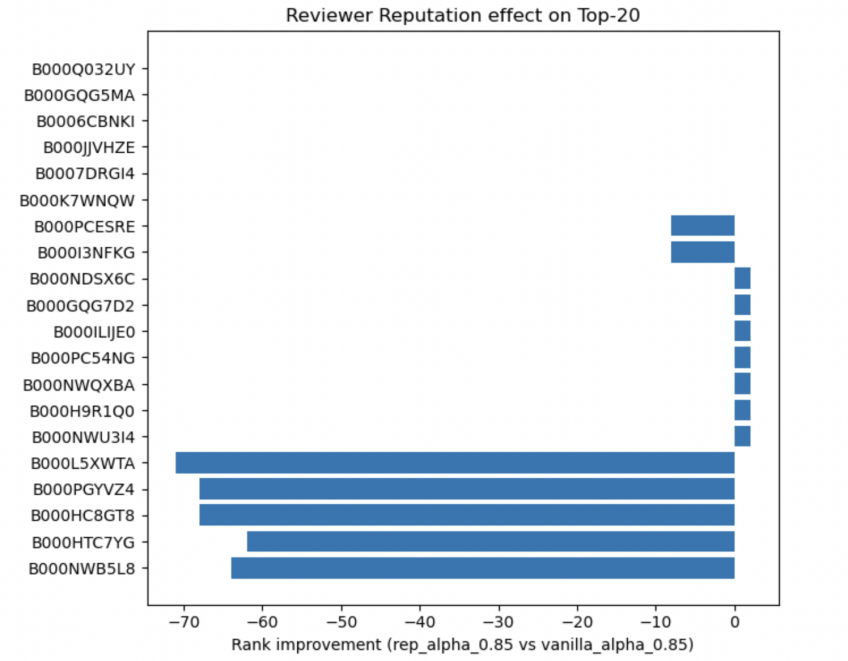

Figure 7: Reviewer reputation effect on the Top–20 books (rep_alpha_0.85 vs. vanilla). Negative values indicate demotion under reputation-weighted ranking.

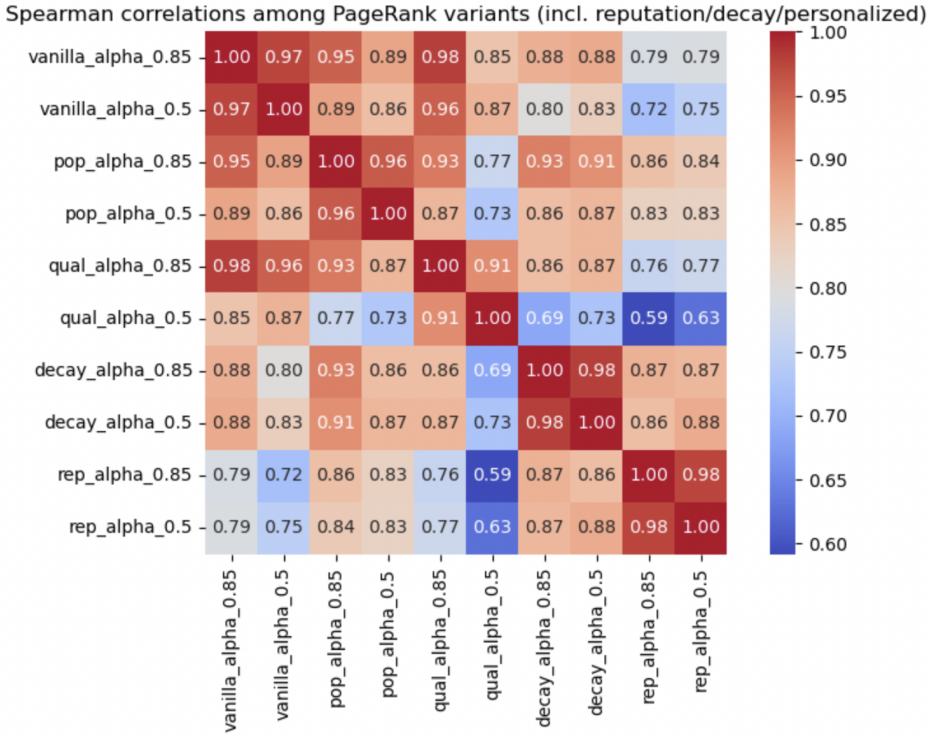

Figure 8: Spearman correlations among all PageRank variants, including personalized, temporal decay, and reviewer reputation models. High correlations ($\rho > 0.8$) indicate structural consistency, while lower values for decay and reputation reveal the distinct influence of temporal recency and reviewer reliability.

4.5 Genre Specialization (Extension)

Finally, the genre specialization extension examines within-genre correlations between prestige (PageRank) and popularity (review count). Separate co-review subgraphs were constructed for the five most frequent genres: *Fiction*, *Juvenile Fiction*, *Biography & Autobiography*, *History*, and entries with missing or unclassified labels. For each subnetwork, PageRank was recomputed to capture local prestige within that thematic domain.

Figure 9 reports the resulting Spearman correlation coefficients (ρ) between in-genre prestige and popularity. The results reveal a strong dependence on the literary domain: whereas *Fiction* shows virtually no alignment between popularity and prestige ($\rho = -0.01$), *Juvenile Fiction* exhibits the highest internal consistency ($\rho = 0.67$), followed by *Biography & Autobiography* ($\rho = 0.60$) and *History* ($\rho = 0.54$). These differences suggest that some genres exhibit more tightly coupled dynamics of visibility and structural influence, while others display a decoupling between mass attention and network prestige.

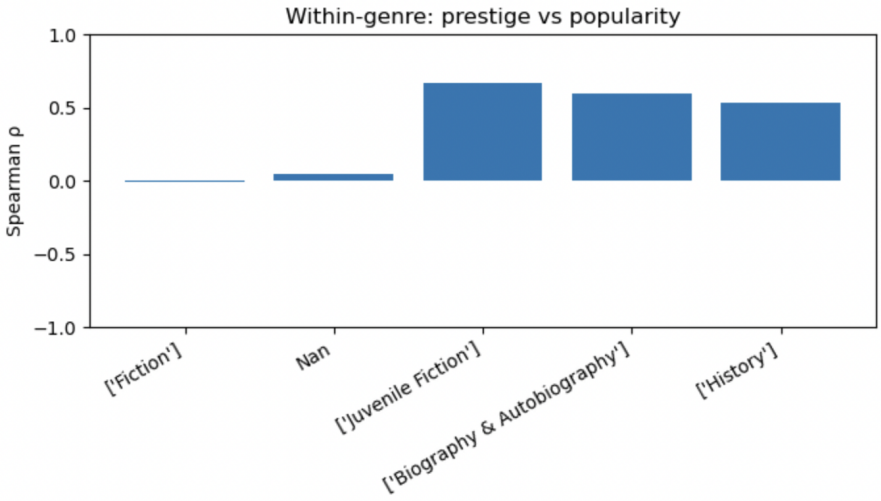

Figure 9: Within-genre correlation between prestige and popularity. Spearman coefficients vary across genres, reflecting heterogeneity in audience engagement and structural diffusion of prestige.

4.6 Discussion Summary

Overall, the results demonstrate that:

- **Popularity and prestige diverge in meaning.** Consistent with the bibliometric distinction between citation frequency and scholarly influence, the analysis confirms that books receiving many reviews are not necessarily the most prestigious. Prestige propagates through influential neighbors in the co–review network, validating the structural interpretation of PageRank as an indicator of diffuse influence rather than raw attention.

- **Personalized PageRank captures context-specific bias.** By modifying the teleport vector toward either highly reviewed or highly rated titles, the model reveals how external preferences shape visibility. This parallels author-level personalization

in bibliometric ranking, showing that even minor teleport biases can shift perceived importance within a cultural ecosystem.

- **Temporal decay distinguishes enduring from transient relevance.** Weighting edges by recency highlights the difference between historically popular books and those gaining current momentum, echoing time-aware PageRank formulations designed for evolving scientific or social networks.

- **Reviewer reputation redefines credible influence.** Incorporating user reliability into edge weighting mirrors reviewer-reputation algorithms in peer-review analysis. Books favored by consistent, high-reputation users rise in prestige, while those popular among erratic reviewers lose prominence, capturing the credibility dimension of collective evaluation.

Together, these findings demonstrate how methodological advances originally conceived for author, journal, and reviewer assessment can be effectively transferred to consumer review networks. Small modifications to the PageRank formulation—through personalization, temporal weighting, and reputation filtering, yield complementary views of influence, trust, and temporal relevance within large-scale online ecosystems.

5 Conclusion

This project implemented a scalable PageRank-based ranking framework on a large book–book co-review network. By constructing and analyzing several PageRank variants—including personalized, temporally decayed, and reviewer-reputation weighted formulations,the study demonstrates how small modifications in the teleport or weighting schemes capture distinct notions of influence within massive datasets.

Empirical results confirm that structural prestige and raw popularity are related but not equivalent measures of importance. Personalized PageRank highlights sensitivity to external biases such as review count and rating quality; temporal decay reveals short-term dynamics of attention; and reputation weighting amplifies the credibility dimension of influence.

Beyond reproducing the classical PageRank model, the project adapts recent methodological insights from network-based evaluation research to the book domain. The distinction between popularity and prestige parallels the bibliometric separation between citation frequency and scholarly influence. Personalization and temporal weighting follow the intuition that user attention and topical relevance evolve over time, while reviewer reputation reflects the credibility of evaluators in shaping collective opinion. Together, these analogies demonstrate that approaches originally conceived for author and journal evaluation can effectively generalize to consumer-generated review ecosystems.

Declaration

I declare that this material, which I now submit for assessment, is entirely my 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 work, and including any code produced using generative AI systems.

I 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 or any other person for assessment on this or any other course of study.