PageRank-Based Link Analysis on Amazon Books

Aleksandr Dudakov (32312A)

2025-06-10

Abstract

This study investigates the application of standard PageRank with taxation and topic-sensitive PageRank algorithms for ranking books based on reviewer interactions within the Amazon Books Reviews dataset. Using a massive graph constructed by linking books reviewed by at least two common users, standard and genre-specific PageRank variants are computed using PySpark to ensure scalability and efficiency. Results demonstrate the capability of PageRank to identify influential titles, with the topic-sensitive variant effectively highlighting central works within specific genres, exemplified by the Business & Economics category. Empirical analyses confirm that both methodologies robustly pinpoint prominent literature, offering practical insights for personalized recommendation systems and targeted bibliometric evaluations. Scalability experiments show linear runtime growth with the number of edges, underscoring the suitability of this implementation for analyzing extensive reviewer networks.

University of Milan
Department of Economics, Management and Quantitative Methods (DEMM)
Algorithms for massive data
Prof. Dario Malchiodi

UNIVERSITÀ
DEGLI STUDI
DI MILANO

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

With the rapid growth of online platforms, it has become crucial to understand and leverage massive datasets in order to extract meaningful insights. One of the most powerful tools for analysing extensive relational data is the PageRank algorithm, which was initially developed to assess the relevance of web pages by exploiting their underlying hyperlink structure. Beyond its original application, PageRank has proven to be remarkably effective in a variety of contexts, including bibliometrics, social networks and recommendation systems.

This project implements and evaluates standard PageRank with taxation and topic-sensitive PageRank methodologies to rank books within a large-scale Amazon book review network. The network is constructed by defining edges between books reviewed by the same (at least two distinct) users, enabling influential titles to be identified based on reviewer interactions rather than explicit rating metrics. Furthermore, genre-specific teleportation distributions are introduced to facilitate topic-sensitive ranking that emphasises domain relevance within specific genres. In this study, the Business & Economics genre is used as an example.

The topic-sensitive approach offers significant practical benefits, such as enabling targeted recommendations that closely align with users' interests, thereby enhancing personalised discovery experiences. For example, bookshops and online platforms can use these rankings to suggest influential titles in particular genres, thereby improving user engagement and satisfaction. Similarly, researchers and publishers can identify influential works in specific fields to inform acquisitions, promotional strategies or further bibliometric studies.

Implemented using PySpark, the approach is designed to ensure computational scalability while maintaining efficiency, even when applied to millions of edges. Empirical analyses confirm the robustness of PageRank-based rankings and the effectiveness of topic-sensitive variants in highlighting influential literature within specific genres. These results highlight the usefulness of link-based ranking algorithms in revealing the subtle structures within large datasets, providing a robust foundation for personalised recommendations and bibliometric analysis.

2 Data

All experiments are based on the Amazon Books Reviews corpus. The datasets are downloaded dynamically via the Kaggle API to ensure full reproducibility. Two datasets are described in the table below:

File	Size / Rows	Attributes used in this study
`Books_rating.csv`	2.9 GB / 3M	`User_id` (reviewer), `Id` (book), `Title`
`books_data.csv`	181 MB / 212K	`Title`, `categories` (raw genre list)

Only the book identifier (`Id`), and user identifier (`User_id`) are required to construct the undirected graph where books are linked by shared reviewers. The `categories` field, retrieved from external metadata, is used to assign a coarse genre label to each book for topic-sensitive teleportation. `Title` is used for interpretation of the results, and also for joining the two tables. All other fields (e.g., rating scores, review text, prices) are excluded, as they play no role in defining the graph structure or influencing PageRank dynamics.

2.1 Preprocessing

Data transformations are implemented in **PySpark**. The pipeline is controlled by a global `sample` flag, which enables seamless scaling from small samples to the full dataset.

Step 1. Filtering: Rows with missing `User_id` or `Id` are discarded. This results in 2 437 750 valid reviews with well-defined book–user relationships.

Step 2. Sampling: An optional fraction `sample` $\in (0, 1]$ can be applied to dataset for faster development and scalability analysis. All final experiments are conducted with the full dataset (`sample` $= 1.0$).

Step 3. Edge construction: An undirected edge is created between books b_1 and b_2 if they have been reviewed by at least two distinct users. Formally, books are connected if $|\{u : R(u, b_1) \wedge R(u, b_2)\}| \geq 2$, where $R(u, b)$ denotes that user u reviewed book b. This threshold filters out noisy links caused by a single user reviewing many unrelated titles.

Step 4. Vertex indexing: Book identifiers are mapped to consecutive integer indices $0, \ldots, |V| - 1$. These indices are used internally for efficient PageRank computation. The original book IDs and titles are stored locally for labeling the output.

2.2 Resulting graph

Reviews processed	Unique books $\lvert V\rvert$	Undirected edges $\lvert E\rvert$
2 437 750	91 009	6 931 860

The graph is constructed by linking books that have been co-reviewed by at least two distinct users. Each such pair is represented by two directed edges, one in each direction. This results in $|E| = 6\,931\,860$ undirected edges and $|V| = 91\,009$ nodes. Although the graph is conceptually undirected, each edge is stored as a pair of symmetric directed links to support matrix-based computation.

2.3 Genre annotation

Genre metadata, extracted from `books_data.csv`, is used to define topic-specific teleportation distributions in the topic-sensitive PageRank variant.

Step 1. Cleaning: The raw `categories` field contains comma-separated genre tags enclosed in brackets and quotes. These delimiters are stripped.

Step 2. Primary genre selection: The cleaned list is split on commas, and the first non-empty token is retained as the primary genre label. This heuristic typically yields a representative category.

Step 3. Join by title: Since `books_data.csv` lacks Amazon book identifiers, genre labels are joined to the main review dataset using the `Title` field. Titles are de-duplicated before joining. Unmatched entries are assigned the placeholder label `<genre unknown>`.

The five most frequent genre labels assigned to books in the review graph are: *Fiction* (20 020 books), *Religion* (7 923), *History* (7 892), *Business & Economics* (4 663), and *Computers* (3 659).

The resulting mapping $i \mapsto$ `genre`, where $i \in \{0, \ldots, |V| - 1\}$, is saved and used to construct topic-conditioned teleport vectors.

3 Theoretical Background

Ranking methods based purely on keyword frequency cannot distinguish genuinely useful content from pages engineered for *term spam*. PageRank overcomes this limitation by

exploiting the underlying *link structure*: a vertex is considered important if many other important vertices link to it. In our setting, vertices represent *books*, and an undirected edge connects two books if they share at least two common reviewers. Thus, PageRank allows us to rank books according to their centrality within the reviewer network rather than relying solely on potentially noisy or manipulated review ratings.

Random-surfer model. Let $n = |V|$, and let d_j be the out-degree of vertex j, defined as the number of vertices directly linked from vertex j. We define the column-stochastic transition matrix $M \in R^{n \times n}$ as:

$$M_{ij} = \begin{cases} \frac{1}{d_j}, & \text{if } (j \to i) \in E, \\ 0, & \text{otherwise.} \end{cases}$$

Starting from an arbitrary probability vector $v^{(0)}$, the position of a *random surfer* evolves according to the update rule:

$$v^{(k+1)} = Mv^{(k)}.$$

In an idealized graph (strongly connected and without dead ends), repeated multiplication converges to the principal eigenvector of M. This limiting vector represents the long-term stationary probabilities of the surfer visiting each vertex.

Taxation. Real-world graphs typically contain *dead ends* (vertices with $d_j = 0$) that absorb probability mass, as well as *spider traps*, which are subgraphs from which probability mass cannot escape. PageRank addresses these issues using a mechanism known as *taxation*: at each step, with probability $\beta \approx 0.85$, the surfer follows an outgoing link; otherwise, it teleports to a uniformly random vertex. Formally, let the uniform teleportation vector be:

$$\nu(i) = \frac{1}{n}, \quad i = 1, \ldots, n.$$

The resulting damped power iteration takes the form:

$$v^{(k+1)} = \beta M v^{(k)} + (1 - \beta)\nu, \tag{1}$$

which converges to a unique fixed point π satisfying:

$$\pi = \beta M \pi + (1 - \beta)\nu.$$

The *PageRank score* of vertex i is defined as the stationary probability π_i.

Topic-sensitive teleportation. Uniform teleportation treats all vertices equally, ignoring user preferences or topical context. Search quality can be improved by biasing teleportation towards vertices related to specific topics of interest. Given a topic represented by a teleport set $S \subseteq V$ (e.g., books labeled with the genre *Business & Economics*), we define a topic-sensitive teleportation vector:

$$\nu_S(i) = \frac{1}{|S|}\mathbf{1}_{\{i \in S\}}.$$

Replacing ν with ν_S in (1) gives the *topic-sensitive PageRank*:

$$\pi_S = \beta M \pi_S + (1-\beta)\nu_S,$$

whose probability mass is concentrated on vertices closely linked to the set S. Computing such vectors $\{\pi_S\}$ for various topics allows personalized ranking.

4 Implementation

The PageRank algorithm is implemented in Python 3.9 using PySpark, facilitating distributed computation and scalability. The implementation relies on matrix-vector multiplication expressed through resilient distributed datasets (RDDs) to leverage parallel execution.

Construction of Transition Matrix The stochastic transition matrix M is constructed as an RDD of tuples (i, j, p), representing the probability p of transitioning from vertex j to vertex i. Each vertex's outgoing probability is uniformly distributed across its neighbors. The adjacency list is computed once and materialized as a cached RDD of transition triples, enabling efficient repeated access during iteration. This precomputation ensures rapid iteration, crucial for scalability to large datasets.

PageRank Iteration Procedure The damped power-iteration method is implemented with a teleportation factor set to the standard $\beta = 0.85$. The iterative update proceeds as follows:

1. Initialize the rank vector uniformly as $v_i^{(0)} = \frac{1}{n}$ for all vertices i.

2. Compute successive iterations using the formula:

$$v^{(k+1)} = \beta M v^{(k)} + (1 - \beta)\nu,$$

where ν denotes the teleportation distribution.

3. Convergence is assessed by the L_2 norm of the difference between successive rank vectors:

$$\|v^{(k+1)} - v^{(k)}\|_2 < \text{tolerance}.$$

The convergence tolerance is set to 10^{-6}, balancing computational efficiency and numerical accuracy, and the maximum number of iterations is capped at 100 to prevent excessive run-time in case of slow convergence.

Implementation Details The iterative step is executed via PySpark RDD operations, specifically utilizing `map` and `reduceByKey` transformations. Here, each rank update step is parallelized and aggregated efficiently, exploiting PySpark's lazy evaluation:

```
links_rdd.map(lambda t: (t[0], t[2] * pagerank[t[1]]))
         .reduceByKey(lambda a, b: a + b)
```

The resultant vector is normalized to enforce numerical stability and ensure that the ranks form a valid probability distribution.

Topic-Sensitive PageRank To implement topic-sensitive ranking, I define a teleportation vector biased toward books of a particular genre (in this study experiments focus on genre *Business & Economics*). The genre-specific teleportation enables context-aware ranking.

Scalability Considerations The algorithm scales effectively due to PySpark's distributed operations and memory management. At each iteration, computation time grows proportionally to the number of edges $|E|$, since one update is performed per edge. The transition matrix is stored as a list of non-zero entries (one tuple per edge) while the rank vector is an array of length $|V|$. Overall memory therefore grows with $|E| + |V|$, which ensures that the implementation remains efficient and scalable even on large graphs. All experiments were run on a local Spark driver with 8 GB of memory.

5 Results

I present results from standard and topic-sensitive PageRank computations applied to the Amazon Books Review graph.

Table 1: Top 20 Books by Standard PageRank

Rank	Title	Genre	Score
1	*Roy Gardner: My Story - Hellcatraz*	Robbers and Outlaws	1.103×10^{-3}
2	*Harry Potter and the Sorcerer's Stone*	Juvenile Fiction	6.059×10^{-4}
3	*The Hobbit*	Juvenile Fiction	5.985×10^{-4}
4	*Dune*	Art	5.461×10^{-4}
5	*The Two Towers*	Fiction	5.340×10^{-4}
6	*The Two Towers: Part II of The Lord of the Rings*	Young Adult Fiction	5.340×10^{-4}
7	*The Two Towers*	Fiction	5.340×10^{-4}
8	*Advanced Programming in the UNIX Environment (2nd Edition)*	Computers	5.146×10^{-4}
9	*User Stories Applied*	`<genre unknown>`	5.104×10^{-4}
10	*The Catcher in the Rye*	`<genre unknown>`	4.565×10^{-4}
11	*The Catcher in the Rye [Audiobook] [CD] [Unabridged]*	Young Adult Fiction	4.521×10^{-4}
12	*The Catcher in the Rye*	Fiction	4.518×10^{-4}
13	*The Hobbit; Or, There and Back Again*	`<genre unknown>`	4.513×10^{-4}
14	*Foundation*	Education	4.325×10^{-4}
15	*Foundation*	Education	4.325×10^{-4}
16	*Foundation*	Education	4.325×10^{-4}
17	*Foundation and Empire*	Religion	4.325×10^{-4}
18	*Foundation and Empire*	Religion	4.325×10^{-4}
19	*The Martian Way*	Fiction	4.325×10^{-4}
20	*Foundation and Empire*	Religion	4.325×10^{-4}

5.1 Standard PageRank Results

The standard PageRank algorithm with taxation ($\beta = 0.85$) achieved convergence after 37 iterations (1147 s). Convergence behavior, measured using the L_2 norm between consecutive rank vectors, exhibited exponential decay (see Figure 2).

Table 1 presents the top 20 books ranked by their standard PageRank scores. The highest-ranked title, *Roy Gardner: My Story – Hellcatraz*, attained a notably higher PageRank score (1.103×10^{-3}) compared to subsequent entries, indicating its central role within the review network. Popular books like *Harry Potter and the Sorcerer's Stone* and *The Hobbit* followed closely, reflecting their widespread readership and extensive review coverage. Notably, multiple editions or formats of certain titles, such as *The Two Towers*

and *Foundation*, consistently appeared among top ranks, underscoring robust reviewer overlap across these variants.

Table 2: Top 20 Books by Topic-Sensitive PageRank (Business & Economics)

Rank	Title	Genre	Score
1	*The Tipping Point: How Little Things Can Make a Big Difference*	Reference	1.741×10^{-3}
2	*Rich Dad, Poor Dad*	Business & Economics	1.254×10^{-3}
3	*The Innovator's Dilemma: When New Technologies Cause Great Firms to Fail*	`<genre unknown>`	1.205×10^{-3}
4	*Before Beveridge – Welfare Before the Welfare State (Choice in Welfare)*	Great Britain	1.172×10^{-3}
5	*Roy Gardner: My Story – Hellcatraz*	Robbers and Outlaws	1.139×10^{-3}
6	*Who Moved My Cheese? An Amazing Way to Deal with Change in Your Work and in Your Life*	`<genre unknown>`	8.974×10^{-4}
7	*Profitable Growth Is Everyone's Business: 10 Tools You Can Use Monday Morning*	Business & Economics	8.704×10^{-4}
8	*Advanced Programming in the UNIX Environment (2nd Edition)*	Computers	8.667×10^{-4}
9	*User Stories Applied*	`<genre unknown>`	8.663×10^{-4}
10	*Working With Emotional Intelligence*	Business & Economics	8.531×10^{-4}
11	*Jack: Straight from the Gut*	`<genre unknown>`	8.425×10^{-4}
12	*The 7 Habits of Highly Effective People: Wisdom and Insights*	Character	8.354×10^{-4}
13	*The Art of Deception: Controlling the Human Element of Security*	`<genre unknown>`	8.109×10^{-4}
14	*The Learning Paradox*	`<genre unknown>`	8.074×10^{-4}
15	*The Mystery of Capital: Why Capitalism Triumphs in the West and Fails Everywhere Else*	`<genre unknown>`	7.923×10^{-4}
16	*The Four Obsessions of an Extraordinary Executive: A Leadership Fable*	Business & Economics	7.873×10^{-4}
17	*Leadership Secrets of Attila the Hun*	Business & Economics	7.747×10^{-4}
18	*The Two Percent Solution: Fixing America's Problems in Ways Liberals and Conservatives Can Love*	Political Science	7.618×10^{-4}
19	*Power of Six Sigma*	Business & Economics	7.523×10^{-4}
20	*Backfire: Carly Fiorina's High-Stakes Battle for the Soul of Hewlett-Packard*	Business & Economics	7.515×10^{-4}

5.2 Topic-Sensitive PageRank Results

The topic-sensitive PageRank analysis, biased toward the *Business & Economics* genre, converged more slowly (see Figure 3), reaching stability after 45 iterations (1361 s).

Table 2 highlights the top-ranked books from this genre-specific analysis. *The Tipping*

Point secured the highest score (1.741×10^{-3}), signifying its prominent influence within the *Business & Economics* context, despite being categorized as *Reference*. Other highly ranked titles, such as *Before Beveridge* and *The Innovator's Dilemma*, while classified under different or unknown genres, are substantively aligned with *Business & Economics*, further demonstrating the sensitivity of this PageRank variant. Canonical texts in the domain, including *Rich Dad, Poor Dad* and *Profitable Growth Is Everyone's Business*, similarly exhibited strong centrality.

Notably, the highest-ranked book from the standard analysis, *Roy Gardner: My Story – Hellcatraz*, retained significant prominence in this genre-specific analysis. Such cross-genre presence underscores substantial interdisciplinary reviewer overlap and connectivity within the review network.

Interpretation The leading titles identified by both PageRank variants predominantly represent influential works, confirming the effectiveness of PageRank in identifying books that are central to the reviewer network. The topic-sensitive variant improves the interpretability of rankings by emphasising influential texts within specific genres, making it particularly useful for personalised recommendations and targeted bibliometric analysis. Therefore, topic-sensitive PageRank amplifies genre-specific influence while retaining globally connected titles, revealing nuanced domain centrality.

5.3 Scalability Analysis

To evaluate the scalability of the proposed PageRank implementation, runtime experiments were conducted on progressively larger subgraphs of the full Amazon Books Review network. Each subgraph corresponds to a randomly sampled subset of reviews, controlled by a sampling parameter `sample`. The final notebook executes on the full dataset (`sample` = 1.0), but all results can be reproduced using smaller sample sizes by modifying the `sample` variable and rerunning the notebook.

Table 3 summarizes key performance metrics for different sample sizes.

As illustrated in Figure 1, the runtime per iteration scales linearly with the number of edges ($|E|$). An ordinary least squares (OLS) regression was conducted to formally assess this relationship, yielding an excellent fit with $R^2 = 0.998$. This confirms that the implementation achieves near-linear scalability with respect to the edge count, as theoretically expected for PageRank algorithm.

Table 3: Scalability of PageRank Computation

Sample	Nodes ($\|V\|$)	Edges ($\|E\|$)	Runtime per iteration (s)
0.05	1,913	6,316	1.65
0.10	5,676	29,554	1.86
0.15	10,624	71,683	2.09
0.25	20,544	237,071	2.96
0.40	37,019	766,827	5.41
0.50	47,635	1,314,828	7.13
0.75	70,290	3,526,660	15.34
1.00	91,009	6,931,860	31.00

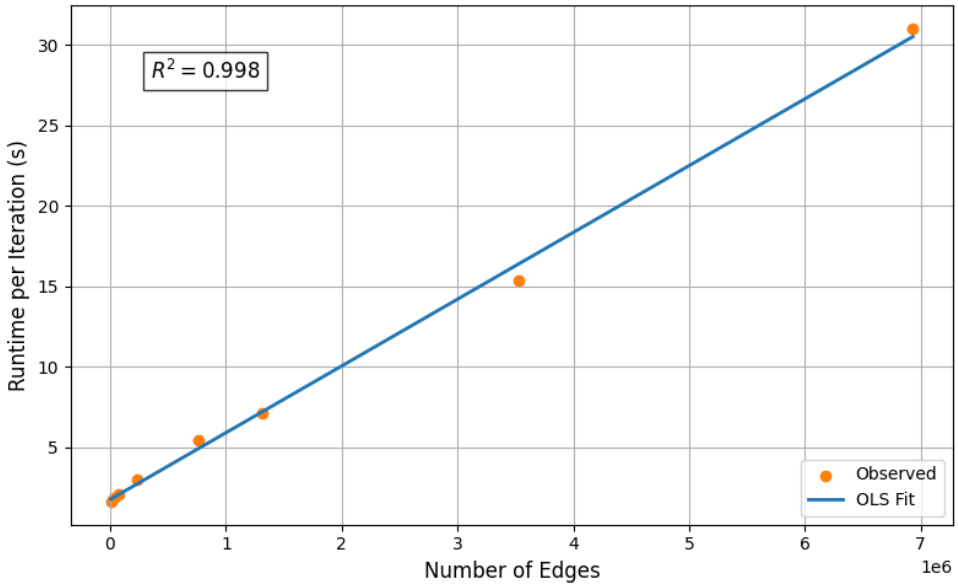

Figure 1: Linear relationship between runtime per iteration and graph size.

These results validate the suitability of the proposed approach for analyzing large-scale networks efficiently, highlighting its capacity to handle extensive datasets within reasonable computational limits.

6 Conclusion

This study presented the effective implementation and rigorous evaluation of standard and topic-sensitive PageRank algorithms on a large-scale Amazon Books review network. By constructing a relational graph in which edges represented significant reviewer overlap, the methodology successfully captured the centrality and influence of books, independently of any explicit rating information.

The empirical results emphasised the robustness of standard PageRank, clearly identifying works that are universally influential based solely on the relational structure of

reviewer interactions. Furthermore, the topic-sensitive variant demonstrated enhanced interpretability and practical relevance by accurately highlighting influential literature within specific genres, as exemplified by the *Business & Economics* domain. This targeted ranking provides valuable insights for personalised recommendations and bibliometric analyses.

Scalability assessments verified that the proposed PySpark implementation achieves near-linear computational complexity in relation to the number of edges in the network. This makes it suitable for the massive datasets commonly encountered in contemporary digital environments.

Future research could improve the analysis by incorporating more precise genre classification and weighting transition probabilities according to the number of shared reviews between books. These improvements could enhance the interpretability and practical relevance of the resulting rankings.

A Appendix

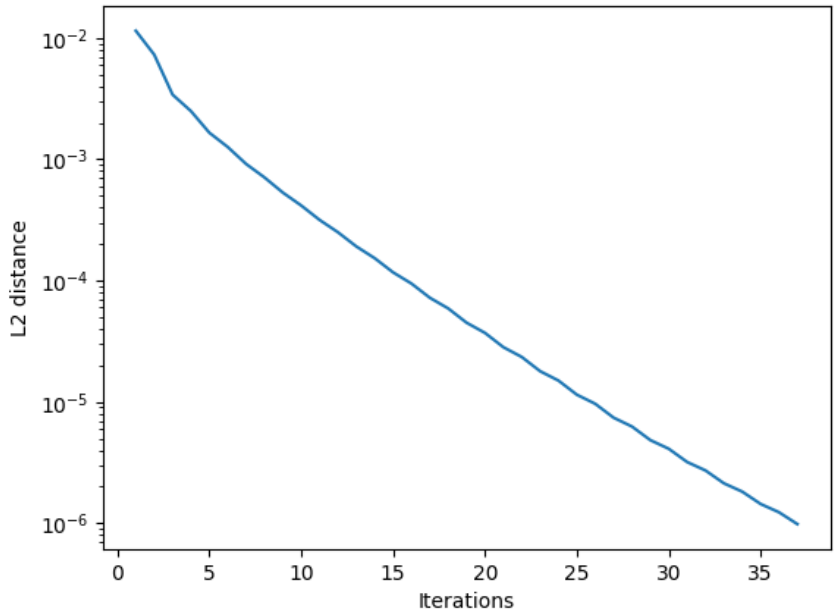

Figure 2: Convergence of standard PageRank with taxation ($\beta = 0.85$). The L_2 norm between successive rank vectors decreases exponentially, indicating stable convergence within 37 iterations.

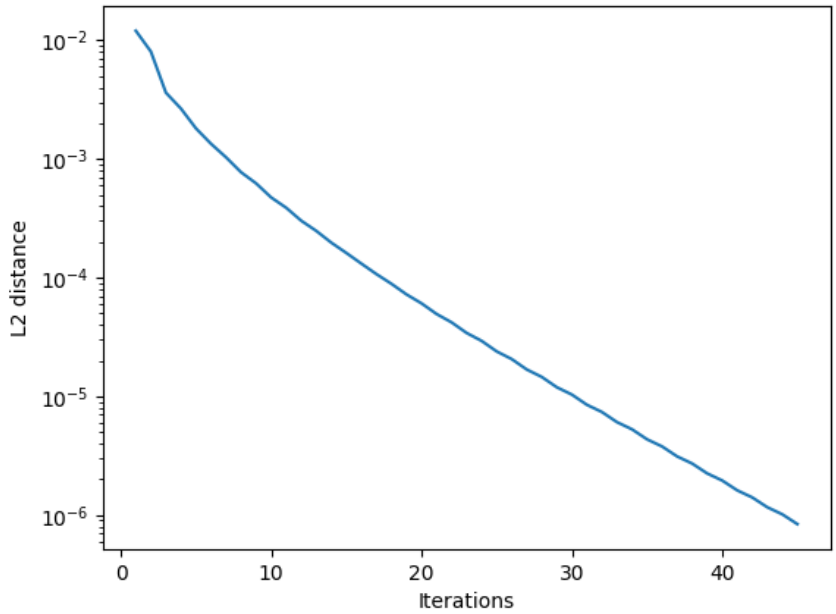

Figure 3: Convergence of topic-sensitive PageRank with taxation (genre: *Business & Economics*, $\beta = 0.85$). The L_2 norm between successive rank vectors decreases exponentially, indicating stable convergence within 45 iterations.

References

Chambers, B., & Zaharia, M. (2018). *Spark: The definitive guide: Big data processing made simple*. " O'Reilly Media, Inc."

Leskovec, J., Rajaraman, A., & Ullman, J. D. (2020). *Mining of massive data sets*. Cambridge university press.