Università degli Studi di Milano
Master in Data Science for Economics

PageRank Link Analysis on Prado Museum Paintings

Emile Rahal
Student ID: 10744A

Algorithms for Massive Data Module
December 2024

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.

Abstract

This project explores the application of network analysis to artwork datasets by employing three variations of the **PageRank** algorithm—**Unweighted, Weighted, and Adjusted Weighted PageRank**—to rank paintings from the Prado Museum dataset.

The dataset, sourced from **Kaggle**, consists of metadata for **13,487 artworks**, including information on authors, titles, and descriptive tags. By constructing a **graph** where nodes represent paintings and edges represent shared tags and authorship, the algorithms highlight the importance and connectivity of paintings within the dataset.

The implementation leverages **PySpark** for scalability, enabling efficient processing of the large network. The results reveal patterns of artistic and thematic relationships, offering insights into the interconnectedness of the **Prado Museum**'s collection.

The complete implementation and supplementary materials for this analysis are available in the public **GitHub repository**[4], check it out!

Contents

1 Introduction

1.1 Background and Motivation

With the rise of big data, network analysis has become an important tool for finding hidden patterns and relationships in large datasets. Museums, for example, have large collections of artworks, each with historical and thematic significance. By analyzing these collections using network science, we can gain insights into art movements, connections between themes, and the influence of individual artists.

The Prado Museum, a leading art gallery worldwide, provides a large dataset of European paintings. By using link analysis techniques, especially the PageRank algorithm, we can measure the importance of each painting in the collection based on how it connects to others. This helps in deciding which artworks to highlight and improves the educational and interpretive experience for visitors and researchers.

1.2 Objective

This project aims to apply three variations of the PageRank algorithm—Unweighted, Weighted, and Adjusted Weighted—to rank paintings in the Prado Museum dataset. The primary goal is to develop an effective ranking system that identifies key paintings and influential artists based on their connectivity within the dataset. Additionally, the project seeks to explore thematic and artistic relationships through shared tags and authorship, providing deeper insights into the connectivity of the artworks.

By using distributed computing tools like PySpark, the project also aims to demonstrate the scalability of graph-based algorithms, ensuring that the proposed solution can efficiently handle large-scale data. Ultimately, the integration of these objectives will culminate in a robust ranking system that highlights significant patterns and influences within the Prado Museum's collection.

1.3 Overview of the PageRank Algorithm

Originally developed by Sergey Brin and Larry Page for ranking web pages in search engine results, the PageRank algorithm assigns a numerical weight to each node in a graph, signifying its relative importance within the network. The algorithm operates on the principle that a node is deemed important if it is linked to by other important nodes.

In the context of this project, paintings are represented as nodes, and edges represent connections based on shared attributes. By applying PageRank, we can rank paintings not just by direct attributes but also by their relationships within the entire collection.

1.4 Report Structure

This report is organized as follows:

- **Section 2: Dataset Description** - Details the dataset used and the specific components considered.
- **Section 3: Data Organization and Pre-processing** - Describes data loading, cleaning, and pre-processing steps.
- **Section 4: Graph Construction** - Explains how the graph was built and the different strategies for defining edges.
- **Section 5: Algorithm Implementation** - Discusses the PageRank algorithm variations and their implementation.
- **Section 6: Experiments and Results** - Presents the findings from the PageRank computations and analyses.
- **Section 7: Discussion** - Provides insights into the results and discusses the impact of certain factors.
- **Section 8: Conclusion** - Summarizes the work and suggests future directions.
- **References** - Lists the sources and materials referenced.

2 Dataset Description

2.1 Chosen Dataset

The dataset utilized in this project is the *Prado Museum Pictures* dataset, available on **Kaggle**[1]. This dataset comprises metadata for 13,487 artworks housed in the Prado Museum, including details such as titles, authors, descriptive tags, and various technical attributes. It serves as a comprehensive resource for analyzing the interconnectedness of paintings based on thematic content and authorship.

Once acquired, the dataset was loaded into a **Spark DataFrame** to leverage **PySpark**'s distributed computing capabilities, which are essential for handling large-scale data efficiently. The loading process involved initializing a Spark session with appropriate memory configurations and reading the `prado.csv` file using PySpark's CSV reader with headers enabled. This ensured that the data was structured appropriately for subsequent analysis.

2.2 Dataset Exploration

A preliminary exploration of the dataset was conducted to assess its quality and understand its structure. The **dataset comprises 13,487 rows**, each representing a unique painting in the Prado Museum collection. An analysis of **missing values** revealed that key columns such as `work_tags`, `work_title`, and `author` had no missing values, ensuring reliable analysis in these areas. However, significant missing data was identified in technical detail columns (e.g., `technical_sheet_autores`, `author_bio`, and `author_url`), leading to their exclusion from further analysis.

The dataset includes **2,560 unique authors**, highlighting a diverse range of artists contributing to the collection. Notably, the term `Anónimo` (Anonymous) appears frequently, accounting for **2,698 paintings**, which indicates a substantial presence of anonymous works. After cleaning, there are **3,547 unique tags** in the `work_tags` column. The most common tags include: `Óleo` (Oil) with 4,177 occurrences, `Lienzo` (Canvas) with 3,225 occurrences, and `Pluma` (Feather/Pen) with 1,923 occurrences.

2.3 Data Cleaning and Preparation

To prepare the dataset for analysis, several cleaning and preprocessing steps were undertaken.

First, columns with significant missing data, such as various technical details, were excluded from the analysis to maintain data integrity. This decision was based on the observation that incomplete technical information could skew the analysis and lead to unreliable results.

Second, the `work_tags` column contained semicolon-separated strings with some artifacts like the '+' symbol, resulting from scraping errors. These artifacts were identified and removed to ensure clean and meaningful tag data. Specifically, the '+' symbol was eliminated from the tags array to prevent it from being treated as a valid tag. The column was split into an array using the semicolon (';') as a separator, facilitating the analysis of individual tags and their relationships.

After that, each painting was assigned a unique, non-sequential ID using PySpark's `monotonically_increasing_id()` function. This ID serves as a unique identifier for each node in the graph, facilitating efficient mapping and analysis in the subsequent PageRank computation.

Finally, for the purpose of link analysis and PageRank computation, the following key columns were selected:

- **ID**: Unique numeric identifier for each painting.
- **Work Title**: Title of the painting.
- **Author**: Name of the artist, critical for adjusted edge weights in the graph.
- **Work Tags**: Array of descriptive attributes forming the basis of thematic connections between paintings.

These columns provide sufficient information to establish connections between paintings based on thematic content and authorship, which are essential for constructing the graph used in the PageRank algorithm.

To illustrate the structure of the dataset after cleaning and preparation, Table 1 presents a sample of five paintings with their respective IDs, titles, authors, and tags.

ID	Work Title	Author	Tags Array
0	Cabeza de Gigante	Bayeu y Subías, Francisco	[Serie de dibujos para el fresco la Caída de los Gigantes en el Palacio Real de Madrid, Lápiz negro, Papel verdoso, Estudio de cabeza, Mitología, 1764]
1	Susana y los viejos	Crespi, Giuseppe Maria	[Sanguina, Papel agarbanzado, Desnudo femenino, Profetas de Israel, Susana, 1701]
2	Dama con una flor amarilla	Cronenburch, Adriaen van	[Óleo, Tabla, Retrato civil, Caléndula / Maravilla (Calendula officinalis), Sombrero, 1562]

Table 1: Example of the Prado Museum dataset after cleaning.

With this structure in place, we can now proceed to create edges and assign weights, which are essential for building the graph used in the PageRank algorithm.

3 Graph Construction

3.1 Defining Links Between Paintings

The main strategy for connecting paintings in the dataset is based on **shared tags**. This method uses two types of connections, or **edges**, to represent the relationships between paintings.

We started with the **Unweighted Edges** where a simple connection is made between two paintings if they share at least one common tag. This shows that the paintings have a thematic similarity without indicating how strong that similarity is.

Later on, we generated **Weighted Edges** where connections are given a weight based on the number of shared tags between paintings. The more tags two paintings share, the stronger their connection.

Additionally, to highlight the influence of the artist we created something called **Adjusted Weighted Edges** having the condition If two paintings are created by the same artist, a constant weight (e.g., +2) is added to their existing connection. This emphasizes the artistic relationship between works by the same creator, making these connections more significant in the **PageRank** analysis.

3.2 Edge Weight Calculation and Normalization

After establishing the connections, the next step is to **calculate and normalize** the weights of these edges to prepare the graph for **PageRank** analysis. Each shared tag between two paintings adds a weight of **1** to their connecting edge, and if two paintings share multiple tags, their connection's weight increases accordingly. For example, if Paintings 1 and 2 share two tags, the weight of their edge becomes **2**. To ensure that each painting's influence is properly scaled, the weights of **outgoing** connections from each painting are **normalized** so that they sum up to **1**. This means that the total weight of all connections going out from a painting equals **1**, maintaining the probabilistic nature required for the PageRank algorithm.

3.3 Incorporating Author-Based Edge Weighting

As we said before, to add an extra layer of meaning to the graph, **author-based weighting** is incorporated. If two paintings are created by the same artist, an additional constant weight (e.g., **+2**) is added to their existing connection. This adjustment not only reflects their shared thematic elements but also highlights the artistic influence of the creator across their works. This combined approach ensures that the graph accurately represents both the thematic and artistic relationships between paintings, providing a robust foundation for the PageRank analysis.

As illustrated in Table 2, the edge generation process has created a dense network comprising **13,478 nodes** and **36,145,728 edges**.

The **degree distribution** of the graph, depicted in Figure 1, is **right-skewed**, indicating that a small number of paintings possess very high connectivity, while most paintings have degrees clustered around the mean. Specifically, the **mean degree** is **2,680.04**, with a **standard deviation** of **1,439.37**, highlighting significant variation in node connectivity. This distribution showcases the presence of highly connected central hubs within the network, which can substantially influence the outcomes of the **PageRank** analysis by disproportionately affecting the ranking of connected nodes.

Statistic	**Value**
Number of Nodes	13,478
Total Number of Edges	36,145,728
Minimum Degree of a Painting	11
Maximum Degree of a Painting	6,514

Table 2: Graph Statistics

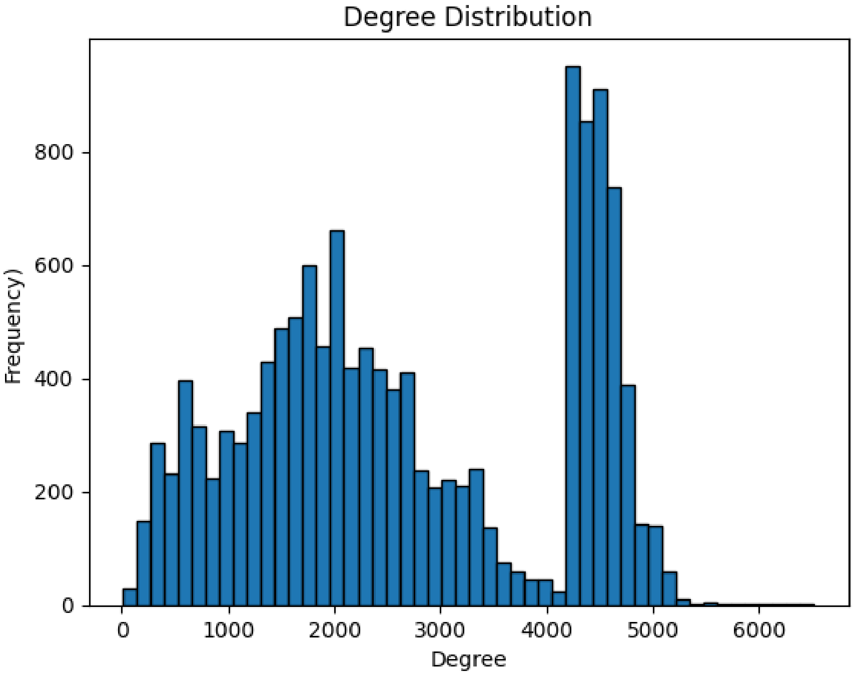

Figure 1: Degree distribution of the graph.

Overall, the large number of connections and the variety in how paintings are linked in the graph accurately show the complex relationships between paintings in the Prado Museum collection. This **structural complexity** not only highlights the many different connections in the data but also provides a strong base for the upcoming **PageRank** analysis. By using these detailed connections, the **PageRank** algorithm can effectively identify and rank the most important paintings, offering deeper insights into the **significance** of the artworks in the collection.

4 Algorithm Implementation

4.1 PageRank Algorithm Overview

The **PageRank** algorithm is a fundamental tool for evaluating the importance of nodes within a graph by analyzing the structure of incoming links. Originally developed to enhance web search engine results, PageRank assigns a numerical weight to each node, reflecting its relative significance in the network.

PageRank operates based on the **random surfer model**, where a hypothetical user navigates through the graph by randomly following outgoing links from one node to another. The probability distribution of the surfer's location converges to the PageRank values of the nodes after numerous iterations. This iterative process ensures that nodes with more incoming connections, especially from other important nodes, receive higher PageRank scores.

The PageRank of a node i is calculated using the following formula for the **unweighted** PageRank, where each outgoing link from a node distributes the PageRank score equally among its neighbors.

$$PR(i) = \frac{1-d}{N} + d \sum_{j \in M_i} \frac{PR(j)}{L(j)}$$

Where:

- d is the **damping factor** (typically set to 0.85), representing the probability that the random surfer continues following links.
- N is the total number of nodes in the graph.
- M_i represents the set of nodes linking to node i.
- $L(j)$ is the number of outbound links from node j.

For the **weighted** PageRank, where the PageRank score is distributed based on the normalized weights of the outgoing links, allowing for a more nuanced evaluation of node importance, the formula is adjusted to account for the strength or significance of each connection:

$$PR(i) = \frac{1-d}{N} + d \sum_{j \in M_i} PR(j) \times W(j,i)$$

Where $W(j,i)$ = Normalized weight of the link from node j to node i

4.2 Implementation Details

A custom PageRank function, `run_pagerank()`, was developed to process both unweighted and weighted graphs. The implementation encompasses the following key components:

4.2.1 Adjacency List Construction

To efficiently represent the graph, **adjacency lists** were utilized; Unweighted Adjacency List is used for equal score distribution in unweighted PageRank. Each node's outgoing edges are grouped into a list of neighbors. For example, if node 1 links to nodes 2 and 3, its adjacency list entry is: **1: [2, 3]**

On the other hand, **Weighted Adjacency List** is employed for proportional score distribution in weighted PageRank. Each node is paired with its neighbors and their normalized edge weights. For instance, node 1 might have: **1: [(2, 0.667), (3, 0.333)]**}

4.2.2 Graph Assumptions Verification

Before executing the **PageRank** algorithm, it is essential to verify that the graph meets necessary assumptions to ensure accurate results. At first, we examined the presence of **dead ends**—nodes with no outgoing edges that can trap PageRank scores and disrupt the algorithm's flow. Our analysis confirmed that there are **no dead ends** in the graph. Additionally, we intended to assess the graph's **connectivity** to verify that it is strongly connected, meaning there is a path between any pair of nodes. However, due to the graph's substantial size, performing a comprehensive connectivity check proved to be computationally expensive and was therefore not feasible within our resource constraints. Maybe in the future!

4.2.3 Initialization

The algorithm's parameters and initial conditions were set as follows:

- **Parameters**:
  - **Iterations**: Maximum of 50 iterations to limit computation time.
  - **Damping Factor** (d): Set to 0.85, representing the probability of following a link.
  - **Tolerance**: 1×10^{-6} as the convergence threshold.
- **Initial PageRank Values**: Each node was assigned an initial PageRank value of $\frac{1}{N}$, where $N = 13,487$. This results in an initial PageRank of approximately 0.000074 for each node.

4.2.4 Iterative Computation Process

The PageRank algorithm was executed iteratively, following these steps in each iteration:

1. **Contribution Calculation**:
   - **Unweighted Graph**: Each node's PageRank is equally divided among its neighbors. For example, if node i has 3 neighbors, each neighbor receives $\frac{PR(i)}{3}$.
   - **Weighted Graph**: Contributions are scaled based on the normalized edge weights. For instance, a neighbor linked with a weight of 0.667 receives $PR(i) \times 0.667$.
2. **PageRank Update**: The new PageRank for each node is computed using the contributions from incoming links and the damping factor:

   $$PR_{\text{new}}(i) = d \times \sum_{j \in M_i} \text{Contribution from } j + \frac{1-d}{N}$$

3. **Convergence Check**: The **Euclidean** distance between the old and new PageRank vectors is calculated:

   $$\text{Distance} = \sqrt{\sum_{i=1}^{N} (PR_{\text{new}}(i) - PR_{\text{old}}(i))^2}$$

   If the distance is **less** than the specified tolerance, the algorithm has converged and terminated. Otherwise, it proceeds to the next iteration.

4.2.5 Function `run_pagerank()`

The `run_pagerank()` function was implemented to execute the PageRank algorithm efficiently. Below is an overview of its functionality:

- **Inputs**:
  - **Adjacency List**: Represents the graph structure.
  - **Weighted Flag**: Determines if we need to use the weighted contributions.
  - **Parameters**:
    - Damping factor $d = 0.85$.
    - Maximum iterations: 50.
    - Tolerance for convergence: 1×10^{-6}.
    - Total noode count $N = 13478$.

- **Process**:

  Here is how it goes, the adjacency list and initial PageRank RDDs are **partitioned and cached** to optimize performance. Next, we choose if the edges are **weighted or unweighted**. Later, for each iteration, the function calculates **contributions**, updates **PageRank** values, and checks for **convergence**. After each iteration, the function **logs** the current iteration number, **distance metric**, top-ranking nodes, and the sum of PageRank values, which should be equal to 1, to monitor progress.

  The algorithm returns the **final PageRank** scores for all nodes **after convergence**.

- **Execution and Results**:

  – The algorithm typically **converged** within **30 iterations**

    ∗ **Unweighted PageRank** converged in 22 iterations.

    ∗ **Weighted PageRank** converged in 24 iterations.

    ∗ **Adjusted Weighted PageRank** converged in 24 iterations.

    , where the **Euclidean** distance between consecutive PageRank vectors fell **below the tolerance threshold**.

  – The algorithm successfully identified **the most influential** paintings based on their connectivity, providing a good ranking system for paintings.

5 Massive Data and Scalability

The PageRank implementation for the Prado Museum dataset required handling a large graph of over 13k nodes and 36 million edges, making scalability a critical challenge. To address this, we utilized PySpark's distributed computing capabilities, leveraging its **Resilient Distributed Datasets (RDDs)** and DataFrames for parallel processing. This enabled efficient graph construction, edge normalization, and iterative computations.

Repartitioning: The graph data was carefully partitioned across 400 nodes to ensure balanced workloads and avoid computational bottlenecks, optimizing the resource utilization of the Spark cluster.

Caching and Broadcast Variables: To improve efficiency, frequently accessed data, such as adjacency lists and intermediate PageRank results, were **cached** in memory. Additionally, static datasets like the ID-to-author mapping were distributed using **broadcast variables**, reducing network communication overhead.

Fault Tolerance: PySpark's fault-tolerant design allowed automatic recovery of lost partitions during iterative PageRank updates, ensuring reliability for the large dataset.

By combining PySpark's capabilities with efficient graph representation and optimized resource usage, we successfully scaled the PageRank algorithm to analyze thematic and artistic relationships in the Prado Museum dataset. This approach highlights PySpark's effectiveness in managing massive data and complex graph-based computations.

6 Results

After computation, PageRank scores were mapped back to paintings using the unique IDs. This allowed for ranking paintings and analyzing the results in the context of their metadata.

6.1 Unweighted PageRank Results

The top 10 paintings based on unweighted PageRank are presented in Table 3.

Painting Title	Author	PageRank Score
Nin y Tudó, José	Retrato mortuorio del periodista Pedro Avial Taracena	0.000143
Anónimo	San Joaquín, Santa Ana y la Virgen Niña	0.000140
Sánchez Perrier, Emilio	Berlina tirada por un caballo en Sevilla	0.000138
Reynolds, Sir Joshua	Retrato de James Bourdieu	0.0.000137
Anónimo (El Greco)	San Francisco en oración	0.0.000136
Creti, Donato ?	Estudio de guerrero romano	0.0.000136
Hernández Nájera, Miguel	Víspera del Dos de Mayo	0.000136
Jover y Casanova, Francisco	Tratado de Cambray	0.000135
Oliva y Rodrigo, Eugenio	Cervantes, en sus últimos días	0.000135
Goya y Lucientes, Francisco de	Academia, figura velada, de perfil..	0.000135

Table 3: Top 10 paintings by Unweighted PageRank score.

The leading painting, *Retrato mortuorio del periodista Pedro Avial Taracena,* stands out due to its profound thematic depth and intricate composition, which resonate strongly with other artworks in the collection. Paintings such as *San Joaquín, Santa Ana y la Virgen Niña* and *Berlina tirada por un caballo en Sevilla* exhibit rich elements/tags and stylistic features that create numerous interconnections, enhancing their prominence in the network.

6.2 Weighted PageRank Results

The top 10 paintings based on weighted PageRank are shown in Table 4.

Author	Painting Title	PageRank Score
Nin y Tudó, José	Retrato mortuorio del periodista Pedro Avial Taracena	0.000143
Reynolds, Sir Joshua	Retrato de James Bourdieu	0.000140
Anónimo	Francisco de Ocáriz y Ochoa	0.000138
Ribera y Fieve, Carlos Luis de	Busto de Rodrigo Calderón	0.000137
Goya y Lucientes, Francisco de	Anotación sobre la boda del artista. Referenci...	0.000136
Goya y Lucientes, Francisco de	El sueño de san José o La muerte de san Franci...	0.000136
Goya y Lucientes, Francisco de	Virgen con el Niño sentado en su regazo, enmar...	0.000136
Goya y Lucientes, Francisco de	Y no hai remedio	0.000135
Goya y Lucientes, Francisco de	Estragos de la guerra	0.000135
Anónimo	Proyecto de decoración arquitectónica	0.000135

Table 4: Top 10 paintings by Weighted PageRank score.

Retrato mortuorio del periodista Pedro Avial Taracena maintains its top position, reflecting its central role and strong thematic ties to other key pieces. Paintings like *Cervantes, en sus últimos días* and *Estragos de la guerra* demonstrate significant thematic relevance and stylistic impact, which amplify their weighted PageRank scores. The recurrence of artworks with complex narratives and compelling visual elements, such as *Academia, figura velada, de perfil,* highlights how their enhanced connections contribute to their elevated status within the collection.

6.3 Adjusted Weighted PageRank Results

The top 10 paintings based on adjusted weighted PageRank are listed in Table 5.

Author	Painting Title	PageRank Score
Anónimo	Retrato de caballero	0.000147
Anónimo	Luis Vives	0.000143
Goya y Lucientes, Francisco de	Y no hai remedio	0.000143
Anónimo	Un hijo de Francisco Ramos del Manzano	0.000143
Goya y Lucientes, Francisco de	Estragos de la guerra	0.000142
Goya y Lucientes, Francisco de	Anotación sobre la boda del artista	0.000141
Goya y Lucientes, Francisco de	El sueño de san José o La muerte de san Franci	0.000141
Goya y Lucientes, Francisco de	Virgen con el Niño sentado en su regazo, enmar...	0.000141
Anónimo	Orfeo y los animales	0.000140
Anónimo	Estudio de figura femenina arrodillada	0.000140

Table 5: Top 10 paintings by Adjusted Weighted PageRank score.

The highest-ranked painting, *Retrato de caballero,* an anonymous work, indicates significant thematic connections and influence across the collection. Notably, several works by Francisco de Goya y Lucientes appear prominently, such as *Y no hai remedio* and *Estragos de la guerra,* reflecting their strong interconnections and pivotal roles in the network. The presence of multiple anonymous paintings, including *Un hijo de Francisco Ramos del Manzano* and *Orfeo y los animales,* shows their extensive linkage and thematic relevance despite the lack of attributed authorship.

6.4 PageRank Score Distributions

Histograms of the PageRank score distributions for each variation are shown in Figure 2.

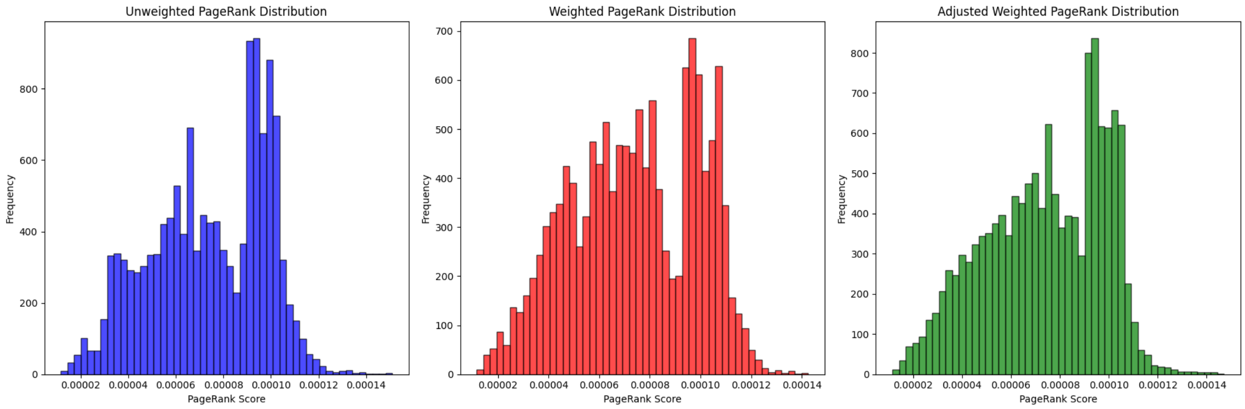

Figure 2: PageRank score distributions for each variation.

The **Unweighted PageRank** shows a slightly right-skewed distribution, highlighting nodes with many connections. The **Weighted PageRank** refines this by prioritizing nodes with stronger connections, leading to a more balanced spread of scores and highlighting the significance of edge weights. Finally, the **Adjusted Weighted PageRank** further enhances differentiation by capturing thematic and structural nuances, as reflected in sharper peaks. These distinctions demonstrate how weighting and adjustments improve the relevance and accuracy of PageRank scores in identifying influential nodes within the network.

6.5 Author Rankings

The cumulative PageRank scores for authors were calculated by summing the PageRank scores of their paintings. The bar chart, shown in Figure 3, highlights the top 10 authors by their total PageRank scores across unweighted, weighted, and adjusted weighted metrics.

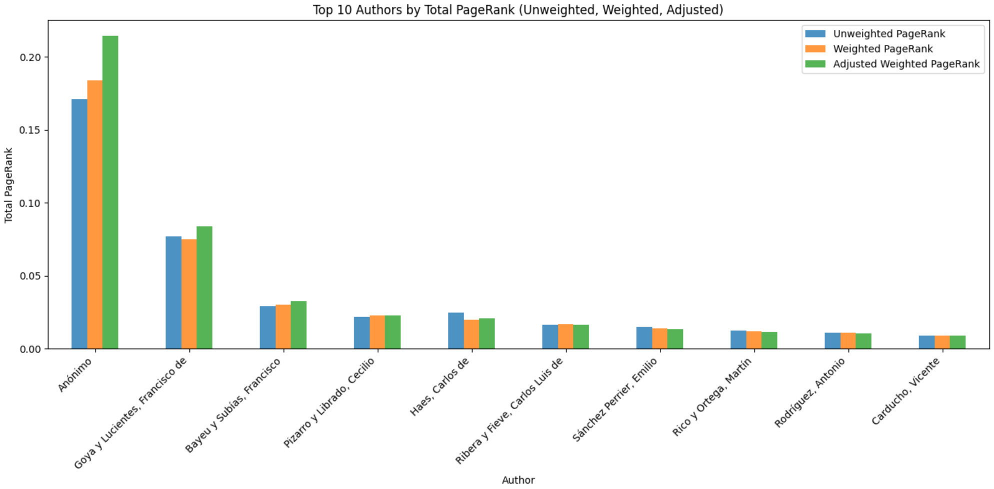

Figure 3: Top 10 authors by cumulative PageRank scores.

Anónimo consistently ranks the highest, which is because of the significant presence of works attributed to anonymous artists within the dataset. **Goya y Lucientes, Francisco de** stands out as the most prominent named artist, reflecting the enduring influence and centrality of his works. The progression from unweighted to adjusted weighted PageRank reveals a slight shift in rankings, emphasizing the importance of thematic and structural adjustments in evaluating artistic connections. Other authors, such as **Bayeu y Subías, Francisco** and **Pizarro y Librado, Cecilio**, maintain relatively consistent positions, indicating their balanced representation in both connectivity and thematic weight.

7 Conclusion

This PageRank-based analysis provided a robust framework for ranking paintings in the Prado Museum collection by leveraging **connectivity**, **thematic depth**, and **authorship**. By integrating these methods, the system highlights both **artistic value** and **thematic importance**, making it a valuable tool for museums, researchers, and art enthusiasts.

7.1 Key Highlights

Best Painting Overall: *"Retrato mortuorio del periodista Pedro Avial Taracena"* by José Nin y Tudó, which ranked **1st** in both **Unweighted** and **Weighted PageRank**, demonstrating broad connectivity and thematic importance but dropped to **23rd** in **Adjusted Weighted PageRank** due to limited shared authorship.

Best Artist Overall: *Francisco de Goya y Lucientes* who dominates the **Weighted** and **Adjusted Weighted PageRank** rankings with works such as *Y no hai remedio* and *Estragos de la guerra* therefore his thematic and artistic influence consistently emerges as a focal point in the analysis.

7.2 Insights from Each Ranking

1. Unweighted PageRank: Focused on **broad connectivity**, emphasizing paintings with numerous links. For example, *San Joaquín, Santa Ana y la Virgen Niña* ranked **2nd** but dropped in other rankings due to weaker thematic connections.

2. Weighted PageRank: Highlighted **thematic depth**, prioritizing paintings with strong shared tags, like *Francisco de Ocáriz y Ochoa* by *Anónimo* ranked **3rd**, showcasing its rich tag-based connections.

3. Adjusted Weighted PageRank: Combined **themes and authorship**, amplifying influence from artists. Some paintings like *Retrato de caballero* (by *Anónimo*) ranked **1st**, while Goya's works had dominance due to their strong thematic and artistic qualities.

8 Future Improvements

Balancing "Anónimo": The prevalence of `Anónimo` (anonymous artists), with 2,698 paintings, created an imbalance, overshadowing named artists. Adjusting its weight could reduce bias and showcase artists like Goya or Reynolds better. While this adjustment was not applied here for simplicity, it remains a priority for future enhancements.

Dynamic Weight Adjustments: A more sophisticated weighting mechanism could incorporate factors such as shared authorship and tag significance. For instance:

$$\textbf{Adjusted Weight} = \textbf{Base Weight} + (\textbf{Authorship Weight} \times \textbf{Influence Score})$$

Where the *Influence Score* reflects the number of works by an artist or their thematic importance.

Further Analytical Directions: Use **Community Detection** to identify clusters of paintings based on shared styles or themes. Add the **Temporal Evolution** of the arts, by analyzing artistic trends over time to map the evolution of movements. Adding **Scalable Extensions** to apply these methods to larger datasets from multiple museums to uncover broader artistic patterns.

By addressing these areas, future research could refine the insights provided by this PageRank-based system, advancing our understanding of the cultural and thematic interconnections within art collections.

References

[1] Prado Museum Pictures Dataset. Kaggle. Retrieved from `https://www.kaggle.com/datasets/maparla/prado-museum-pictures`

[2] Brin, S., & Page, L. (1998). The anatomy of a large-scale hypertextual web search engine. *Computer Networks and ISDN Systems*, 30(1-7), 107-117.

[3] Apache Spark. (2023). Apache Spark Documentation. Retrieved from `https://spark.apache.org/docs/latest/api/python/`

[4] Rahal, E. (2024). Prado_PageRank: PageRank Analysis on Prado Museum Dataset. GitHub Repository. Retrieved from `https://github.com/eeemile777/Prado_PageRank`