Genre Galaxy: A Network Graph of Literary Genres

Nadia Choudhury
choudn2@rpi.edu

Afsana Bhuiyan
bhuiya@rpi.edu

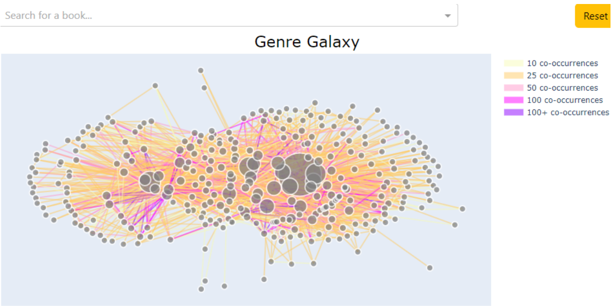

Figure 1: Visual representation of the Genre Galaxy showcasing the complex web of connections between literary genres. Nodes represent individual genres, with node size reflecting genre popularity. Edge colors indicate the strength of co-occurrence, with the legend explaining the range from sparse (light yellow) to dense (dark purple) connections.

Abstract

"Genre Galaxy" is an interactive data visualization project designed to map and analyze the complex relationships between literary genres. Utilizing a comprehensive dataset of books sourced from Kaggle, this project applies data processing techniques using Python and visualizes the data through an interactive network graph constructed with Plotly and Dash. The visualization illuminates the interconnectedness of literary genres, revealing how frequently genres co-occur within books and highlighting significant thematic overlaps. By quantifying these relationships, the "Genre Galaxy" provides a novel tool for academic researchers, publishers, and literature enthusiasts to explore genre dynamics. The insights gained from this project offer a new perspective on genre classification and its influence on literary and publishing trends.

1 Introduction

The vast expanse of literature comprises numerous genres and subgenres, each carrying unique thematic elements and cultural significance. With the increasing digitization of literary content, the potential to analyze and visualize these genres through advanced data techniques presents a compelling opportunity to explore how different literary forms interconnect and influence each other. This project, "Genre Galaxy," as shown in figure 1, utilizes an interactive network graph to visually map the relationships between various literary genres based on their co-occurrence in books.

1.1 Motivation and Audience

Understanding the intricate relationships between literary genres not only enriches the academic study of literature but also provides valuable insights for publishers and writers aiming to grasp current trends and reader preferences. The motivation for this project stems from the desire to create a tool that visually simplifies these complex relationships, making it accessible not only to academics but also to a broader audience including publishers, writers, and avid readers. By presenting literary data in an interactive format, the "Genre Galaxy" aims to foster a deeper understanding and appreciation of literary genres, facilitating discoveries that text-based analysis might overlook.

1.2 Design Evolution

Initially, this project was conceived to explore the success metrics of book-to-film adaptations across various genres, hypothesizing that certain genres translate more effectively to film than others. However, after receiving feedback on the proposal and engaging with the available datasets, we realized the need to pivot our focus. The revised project direction, now centered on the "Genre Galaxy," emerged as a more feasible and insightful endeavor within the given timeframe. This transition allowed for a broader exploration of literary data, moving away from the narrow focus on adaptation success to a more expansive view of genre relationships.

1.3 Research Question and Hypothesis

Our investigation is driven by the research question: What patterns of connectivity and influence among literary genres can be discerned from their co-occurrence in a comprehensive dataset of books? This question aims to use network visualization to analyze the relationships between genres, focusing on identifying significant clusters and central nodes within the network.

We hypothesize that the structure of the literary genre network, depicted through our "Genre Galaxy" visualization, will reveal discernible clusters and central genres that play a pivotal role in the literary field. These clusters are expected to illustrate not only the prevalence of certain genres but also their potential influence over others, reflecting broader trends and preferences in literary production and consumption.

2 Related Work

2.1 Literary Data Visualization

Advances in literary data visualization have provided new methodologies for examining large text collections and extracting meaningful insights from complex literary data. One pivotal work in this domain by Jänicke et al. (2017) introduces sophisticated techniques for visual text analysis in digital humanities, showcasing how thematic shifts within literary works can be visually traced and analyzed over time. This approach highlights the significant potential of visualization tools to elucidate patterns in literary themes that might otherwise remain obscured in traditional literary analysis.

2.2 Network Analysis in Literary Studies

Network analysis has proven to be a valuable method in the exploration of narrative structures and relationships within literary works. Elson, Dames, and McKeown (2010) have made substantial contributions to this field with their method for extracting and analyzing social networks from literary fiction. By mapping character interactions as networks, their research provides a quantitative approach to understanding character relationships and narrative structures across different genres.

Hachmann (2023) enhances our understanding of network analysis in literature, examining its impact on our views of creativity and agency in the arts. This study suggests that network analysis can challenge traditional ideas about who controls the narrative and how stories are developed. It reveals that network analysis not only uncovers how characters or themes are connected but also offers fresh insights into the creative processes behind literature and art.

3 Methodology

3.1 Data Collection and Processing

The dataset for the "Genre Galaxy" project was sourced from Kaggle, specifically from the collection titled "Best Books 10K Multi-Genre Data". This dataset, originally compiled from Goodreads, was selected for its comprehensive range of books and associated genres. It includes data on thousands of books, which is ideal for analyzing patterns across a wide spectrum of literary genres.

The data cleaning and preparation phase was critical in ensuring the dataset was optimized for net-

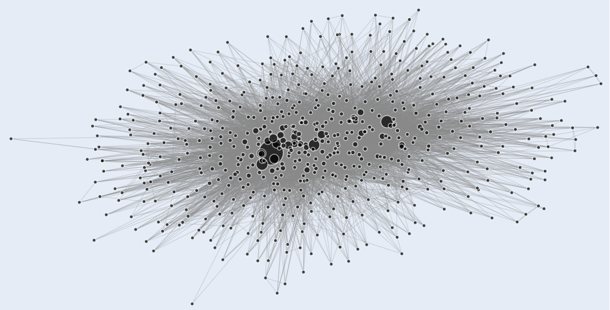

Figure 2: Initial prototype of Network Graph with genre node sizes indicating popularity of genre.

work analysis. Initially, unnecessary columns such as 'URL', 'Description', and any unnamed residual columns were removed to streamline the dataset. Subsequently, the 'Genres' column, which was presented as a string of lists, was transformed into actual Python list objects using the ast.literal_eval method, facilitating more straightforward manipulation of genre data. We also carefully removed any rows that lacked complete genre information to maintain data integrity. To explore the interconnections between genres, we generated pairs of genres from each book's genre list using the Python itertools.combinations method. This step produced a comprehensive set of genre pairs, each indicating a relationship between genres. Both the cleaned data and the genre pairs were then exported into CSV files, to be used in subsequent analysis stages.

3.2 Tools and Technologies Used

Our project was developed using the Dash framework by Plotly, which facilitated the creation of an interactive web-based visualization application. The core visualization logic was implemented using Python, with significant reliance on libraries such as Pandas for data manipulation, NetworkX for network graph construction and analysis, and Plotly for rendering interactive graphical outputs. The network graph was initialized using NetworkX, and genre relationships were depicted as nodes and edges in the graph, with edges colored based on the strength of genre co-occurrences and nodes sized proportionally to the genre popularity. Interactivity was enhanced through Dash components like dropdowns and clickable graph elements, enabling users to explore different genres and their connections dynamically. The layout and node positioning were managed using the Kamada-Kawai layout algorithm, which is effective for large graphs, ensuring that the visualization remains comprehensible and visually appealing.

4 Visualization Design and Implementation

4.1 Initial Design and Prototyping

The initial prototype of "Genre Galaxy", as shown in figure 2, was crafted using Plotly due to its robust capabilities for generating interactive, web-based visualizations. The graph's structure was managed with NetworkX, which facilitated the dynamic manipulation of nodes and edges, representing literary genres and their connections. The Kamada-Kawai layout was chosen to minimize edge crossings and improve the visual appeal of the network, particularly suitable for our expansive dataset.

4.2 Iterative Design Improvements

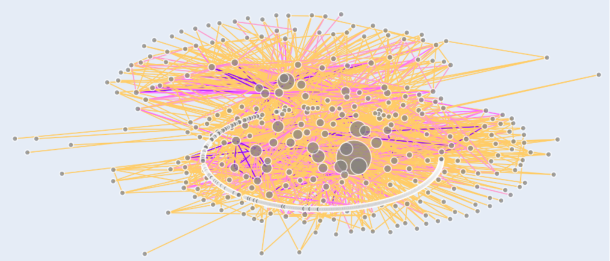

Figure 3: Enhanced prototype of Network Graph with edge colors indicating strength of connection between genre nodes.

As the visualization evolved, significant enhancements were made to its functionality and aesthetics. Early in the development, we decided to adjust the node sizes within the graph to represent the popularity of each genre, as shown in figure 2. This was achieved by analyzing the dataset to count the occurrences of each genre and using these counts to set the node sizes—larger nodes for more prevalent genres, thereby making the graph's representation of data both clearer and more visually engaging.

Simultaneously, we refined the way connections between genres were visualized. Initially, the plan was to use line thickness to denote the strength of these connections. However, as the complexity of the data became apparent, a color-coded system was implemented instead, as shown in figure 3. This adjustment involved a significant redesign of how edges were displayed. Each connection, or edge, between genres was assigned a unique color based on the frequency of their co-occurrence, with a gradient from light yellow to deep purple, inspired by selections from the Color Brewer tool. This method greatly enhanced the readability of the network and helped convey the strength of relationships more intuitively.

The implementation of this color scheme posed a challenge, particularly because Plotly did not support legends for line colors in scatter plots. To address this, we devised custom solutions by deploying multiple scatter traces within Plotly to display edges in unique colors, effectively distinguishing the connection strengths between genres.

4.3 User Feedback

Throughout the development of "Genre Galaxy," we actively engaged with users to gather feedback, which significantly influenced the evolution of our interactive features. One significant piece of feedback was the need for more interactive features that would allow users to engage more deeply with the data. Users expressed a desire for the ability to isolate specific information and explore the literary network based on their interests.

A suggestion was the implementation of a search functionality that would enable users to find specific books within the network. Another common request was for the ability to click on a genre node and see all its direct connections, facilitating a focused exploration of relationships within the graph. Users also proposed that the visualization include a way to visually represent the strength of connections, suggesting color-coded edges to differentiate between weak and strong relationships.

Additionally, there was feedback about enhancing the educational aspect of the visualization. Users recommended providing more context about the connections, such as displaying a list of related genres and their connection strengths when a genre node is clicked. This would not only aid in understanding the graph but also enrich the user's knowledge about genre interrelationships.

4.4 Final Implementation

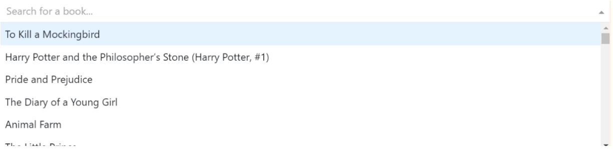

Figure 4: Dropdown menu for the Book Search feature, allowing users to select a book and highlight its genres on the network graph.

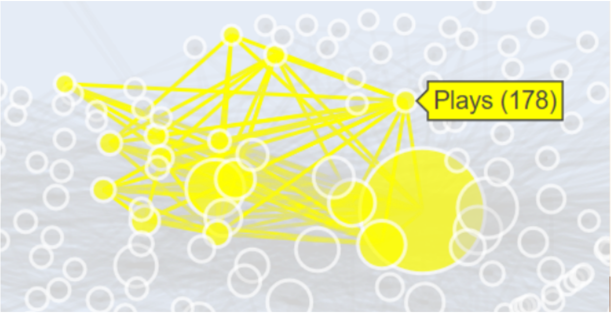

Figure 5: Visualization of click functionality: When the 'plays' node is clicked, the network graph highlights relevant genres and nodes and dims non-relevant ones by increasing their transparency.

In response to the feedback, we incorporated several new features into "Genre Galaxy" to enhance its interactivity and educational value. The implementation of Dash later in the development phase allowed us to introduce a sophisticated search functionality. Users can now select a book from a dropdown menu, as shown in figure 4, which triggers the visualization to highlight the genres related to the selected book, while dimming unrelated genres. This selective highlighting significantly improves the user's ability to dissect and understand the genre relationships specific to their interests.

Furthermore, clicking on a genre node now highlights all its direct connections, as shown in figure 5, and the connected genres are displayed beneath the graph, ranked by the strength of their relationships based on co-occurrence data. This not only focuses the user's attention on specific relationships but also provides a structured path for exploring the network. The color coding of edges, based on the strength of connections, was also implemented, enhancing the visual distinction between different levels of relationships and making the network easier to interpret at a glance.

These user-driven enhancements have transformed "Genre Galaxy" into a more dynamic and interactive tool, facilitating a deeper engagement with the dataset and making complex relationships between literary genres more accessible and understandable to a broad audience. Each advanced feature was carefully designed to respond to user feedback, significantly enhancing both the usability and educational potential of the visualization.

5 Team Contributions

Afsana Bhuiyan took the lead on data processing and cleaning, utilizing Python to process the initial "goodreads_dataset" sourced from Kaggle. Beyond data cleaning, Afsana worked on generating pairs of genres from each book's genre list, creating another crucial dataset that mapped the relationships between genres. She also developed the initial network graph using NetworkX, where she coded the visualization to reflect genre popularity by adjusting the size of nodes based on how frequently genres appeared in the dataset. Additionally, Afsana implemented interactive features such as the click functionality, which upon user interaction, highlights a genre and its connections, and displays a ranked list of related genres and their co-occurrence counts beneath the graph. She also integrated a reset button to allow users to return the visualization to its original state.

Nadia Choudhury focused on enhancing the visualization's ability to display the strength of relationships between genres. Initially, Nadia experimented with using edge thickness to denote connection strength but opted to shift to a color-coded scheme for greater clarity and visual impact. Moreover, Nadia developed the legend that explains the color scheme used for the edges, making the graph more intuitive and informative. She also implemented a hover feature over the nodes, which displays the genre name and the number of books in which the genre appears. Nadia's contributions extended to improving user interactivity by implementing a search function that enables users to find and highlight genres associated with specific books.

6 Results and Discussion

6.1 Visualization Outcomes

The "Genre Galaxy" visualization successfully mapped the complex network of literary genres based on their co-occurrence in books. One of the most notable results was that the largest nodes, representing genres with the highest number of associated books, such as Fiction, Nonfiction, and Classics, also had the most connections. This indicates a strong prevalence of these genres across a wide array of literary works, underscoring their foundational role in literature. These genres serve as pivotal hubs within the network, linking to numerous other genres and subgenres, suggesting their influence and versatility in thematic and narrative connections.

6.2 Analysis of Genre Relationships

Fiction emerged as a central node with significant connections to genres like Classics, Contemporary, and Young Adult, suggesting a versatile appeal that spans traditional literary forms to modern narrative styles. This genre's ability to intertwine with various themes such as Fantasy and Romance also illustrates its capacity to blend different narrative elements, attracting a broad reader base.

Nonfiction showed strong ties to Biography, History, and Memoir, indicating a concentrated interest in factual and autobiographical content. The connection to Audiobooks highlights a trend towards consuming non-fiction in audio format, reflecting modern consumption habits and the genre's adaptability to new media.

Classics were predominantly linked with Fiction, underscoring their foundational influence on literary storytelling. Connections to genres like Historical Fiction and Fantasy suggest that classics often incorporate rich historical details and elements of the fantastical, resonating through time and influencing a range of literary categories.

6.3 Discussion of Hypothesis

The structure of the literary genre network, as depicted through our visualization, confirmed our hypothesis by clearly identifying discernible clusters with central genres that play pivotal roles in the literary field. These observed clusters not only showcased the prevalence of certain genres but also their extensive influence over others. This aligns with broader trends in literary production and consumption, where central genres evolve and adapt, influencing a wide array of subgenres and thematic areas.

7 Conclusions

The "Genre Galaxy" project successfully visualized the complex network of literary genres, revealing significant relationships and central nodes within the literary landscape. By implementing interactive features, the project not only facilitated a deeper understanding of these relationships but also made the exploration of literary genres accessible and engaging to a broad audience. The visualization confirmed our hypothesis by identifying key clusters and central genres that play pivotal roles in the literary field, reflecting broader trends in literary production and consumption.

8 Future Work

Looking to the future, the "Genre Galaxy" has several planned enhancements to further enrich the user experience and expand the project's capabilities. Firstly, we aim to provide more detailed information about specific books when selected, allowing users to delve deeper into the books' content and understand their relationships to prevailing genre trends. This feature will enable a richer exploration of thematic and narrative elements. Additionally, we intend to introduce a functionality that lists the most popular books for each genre. This will not only enhance the dataset but also help users discover seminal works within each genre, potentially guiding their reading choices. Another significant enhancement will involve the implementation of advanced filtering options, such as filters for time periods, authors, and ratings. This will allow users to customize their exploration of the literary genre network, enabling them to view connections through various lenses that align with their specific interests or research needs. These enhancements are designed to deepen user engagement and enhance the educational utility of the "Genre Galaxy," ensuring its continued relevance to a wide audience, from avid readers to academic researchers.

References

[1] Elson, D.K., Dames, N., and McKeown, K. (2010). Extracting Social Networks from Literary Fiction. In Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, pages 138–147, Uppsala, Sweden: Association for Computational Linguistics.

[2] Hachmann, G. (2023). Network Analysis in Literature and the Arts: Rethinking Agency and Creativity. Journal of Literary Theory, vol. 17, no. 2, pp. 221-240. https://doi.org/10.1515/jlt-2023-2010

[3] Jänicke, S., Franzini, G., Cheema, M.F., and Scheuermann, G. (2017). Visual Text Analysis in

Digital Humanities. Computer Graphics Forum, 36: 226-250. https://doi.org/10.1111/cgf.12873

[4] Johari, I. (2023). Best Books 10K: Multi-Genre Data [Data set]. Kaggle. https://www.kaggle.com/datasets/ishikajohari/best-books-10k-multi-genre-data