A Network Analysis of Censored Books

Identifying Genre and Author Trends in Banned and Challenged Books

Jeremy Piperni
Computer Science
McGill University
Montreal, QC, Canada
jeremy.piperni@mail.mcgill.ca

ABSTRACT

This progress report aims to update the reader on the progress made on the network analysis of censored books. The project aims to identify genre and author trends in banned and challenged books. By pursuing this project, we could identify if some genres and/or authors are more likely to be censored. These trends could lead to understanding if political or ideological factors play a role in the rise of book bans.

KEYWORDS

Networks, Censorship, Books, Banned Books, Challenged Books, Genre Analysis, Author Analysis, Education, Politics

1 INTRODUCTION AND MOTIVATION

Censorship and freedom of speech have been at the forefront of politics and public discourse in North America. While freedom of speech tends to be important to a lot of individuals, the banning of books has surged in the US, with an increase of 65% in 2023 compared to 2022 [1]. In the past, certain books were associated with events like assassinations, which lead to justifying certain book bans. Famously, the man who killed John Lennon was found with "The Catcher in the Rye" on his person [2]. The book is often criticized for its vulgarity, dishonesty, and use of alcohol [3], which led to it being heavily censored. The assassination reinforced the beliefs of many, that the book could lead to violence, leading to them justifying the ban. Nowadays, book censorship is more frequent, although justifications and reasonings are often lacking.

This prompts several questions: What/Who is leading this increase in censorship? What political and ideological factors are maybe driving this censorship? Some articles have been written with the goal of answering these questions, but none have utilized network science as their methodology. This project aims to identify genre and author trends in banned books by utilizing network models. By analyzing these networks, some trends might lead us to answers on the possible agendas of the individuals or groups pushing for these book bans. A network analysis approach can also uncover hidden relationships and patterns, by revealing clusters of nodes that might be easily missed with other approaches.

2 RELATED WORK

There is a lack of research for network analyses that focuses on censored books. This project aims to bridge this gap and provide some insights to future work. The following articles provide relevant information on both network approaches, and book censorship context.

2.1 Detecting Network-based Internet Censorship via Latent Feature Representation Learning [4]

The article written by Shawn P. Duncan and Hui Chen designed a classification model that detects network-based internet censorship. The authors propose a sequence-to-sequence autoencoder to capture the structure data [4]. They then utilize a densely connected multi-layer neural network model to estimate the probability of censorship events [4]. The authors also created a second model, that uses network reachability data for an image-based classification model [4]. Both models were able to successfully detect network-based internet censorship [4]. While this article focuses on internet censorship and not book censorship, it can provide interesting network-based approaches that can be applied in this project.

2.2 Book bans in political context: Evidence from US schools [5]

This article written by Langrock et al. focuses on the rise in book bans in the US. Using data from PEN America of 2,532 book bans during the 2021-2022 school year, the authors analyze the types of banned books and authors, the socio-political environments of book bans, and the interest in book bans. The authors discover that banned books feature disproportionately characters of color in children's books, and that banned books are disproportionately written by people of color [5]. They also discovered that counties that are right-leaning but have become less conservative are more likely to ban books [5]. This article provides a useful analysis of both genres and authors of banned books, which we can compare to the results of this project. However, the authors take a more direct approach to their research, and do not utilize network models.

2.3 A History of Censorship in the United States [6]

This essay written by Jennifer Elaine Steele provides a history of censorship in the US. The author delves into the censorship in public libraries and the censorship in schools. They speak about many types of censorship, like the censorship of religion, comic books, communist texts, and many more. This essay provides a necessary context of censorship, which can be used to explain some of the censorship trends that we will find in this project.

3 PROBLEM DEFINITION

The censorship of books can often appear to be subjective, where political and ideological factors often drive these decisions. This project aims to bring an understanding of whether certain genres or authors are targeted disproportionately to censorship. Several network models will be created using the dataset mentioned in section 4. These models will be analyzed, with the goal of answering multiple censorship questions like the following: Do specific authors face a higher likelihood of censorship? Do specific genres tend to be censored more often? Do other hidden censorship patterns emerge from analyzing thematic connections between banned books? Answering these questions would provide insights into who and what is censored more often, and if there could be political or ideological factors responsible.

4 DATASET

A dataset compiled by Chieler Li [7], gathered book information during the DotData Hackathon in 2025. The dataset contains the title, author, book description, and genre of around 17,000 books, which of these around 7,700 books are labelled as censored, challenged, or banned. All descriptions and genres were gathered from Goodreads, and all banned/challenged books were gathered using ALA, Wiki, and PEN America's Index. All uncensored books will be ignored, as they do not pertain to this project.

5 METHODOLOGY

For the network analyses, multiple networks will be created. This section will explain the network construction, and graph analysis techniques for each network.

5.1 Genre-Book Bipartite Network

5.1.1 Network Construction. The network will be a bipartite graph, where nodes are divided into two distinct sets, and edges can only exist between sets. The first set of nodes will consist of the genres, while the second set of nodes will consist of the books. Edges will connect books to their corresponding genres.

5.1.2 Graph Analysis. The bipartite network provides the opportunity of conducting centrality measure analyses. Degree Centrality will be performed on the genre nodes. This will provide information on which genres are banned often. A simple degree distribution will be conducted on the book nodes, to provide insight on if multi-genre books are more likely to be banned. Eigenvector Centrality on the genre nodes will measure a nodes influence based on the node's connections to other nodes in the network, identifying which genres tend to be linked to the most "influential" books.

5.2 Author-Book Bipartite Network

5.2.1 Network Construction. The network will be a bipartite graph where the first set of nodes will consist of the authors, while the second set of nodes will consist of the books. Edges will connect books to their corresponding authors.

5.2.2 Graph Analysis. The bipartite network provides the opportunity of conducting centrality measure analyses. Degree Centrality will be performed on the author nodes. This will provide information on which authors are censored the most often. Eigenvector Centrality on the author nodes

will provide insights on which authors tend to be linked to the most “influential” books.

5.3 Genre Co-Occurrence Network

5.3.1 Network Construction. The nodes of the network are varying genres. Edges will be added between each node, when a banned book belongs to both genres. The graph is weighted, therefore when a book shares genres, a weight of 1 will be added to the weight of the edge.

5.3.2 Graph Analysis. Clusters will be created between genres that are often banned together. This graph structure permits community detection analysis and centrality measure analysis. For community detection, the Louvain method is utilized; Where communities will be created, demonstrating genre clusters that are frequently challenged together. Certain centrality measures will also provide some insights regarding the relationship between genres. Firstly, Degree Centrality will be used to measure the number of direct connections for a genre. It will identify the most frequent co-occurrent banned genres. Betweenness Centrality will also provide valuable insight on identifying key genres that link other genres, by measuring how often a genre lies on the shortest path between two other genres. Lastly, Closeness Centrality will be used to identify the genres that are at the core of censored books.

5.4 Author-Genre Bipartite Network

5.4.1 Network Construction. The network will be a bipartite graph where the first set of nodes will consist of the authors, while the second set of nodes will consist of the genres. Edges will connect genres to their corresponding authors.

5.4.2 Graph Analysis. Both Degree and Eigenvector Centrality measures will be used. Degree Centrality on the author nodes will measure how many genres an author’s banned books belong to. Degree Centrality on the genre nodes will provide insight on which genres have the highest number of censored authors. Eigenvector Centrality will be utilized for the genre nodes, which will identify the most “influential” banned genres. Community detection algorithms such as the Louvain method will be used to find both author and genre clusters. The method will detect if censorship patterns group authors in distinct communities. Lastly the genre clusters can help us confirm the results from the community detection from network 5.3.

5.5 Book Similarity Network

5.5.1 Network Construction. Genres don’t always capture the similarities between books. This network will aim to provide a different approach to clustering. The nodes of this network will be the book titles. Edges will be constructed when two books share a high similarity. The similarity measure will be calculated by using TD-IDF and Cosine Similarity on the descriptions of the books provided.

5.5.2 Graph Analysis. Clusters will be created between books that share similarities, even if they belong to different genres. The Louvain method will be utilized for community detection, where books with similar themes will be clustered. These books will then be analyzed for their genres, to see if hidden thematic connections were found that the genre analysis was missing.

6 EXPERIMENT SETUP

6.1 Data Extraction / Cleaning

To prepare the network data, the Numpy and Pandas libraries for Python3 were utilized. The raw book csv data was extracted, and all non-censored books were removed. An exploratory analysis of genres was conducted, where 430 unique genres appeared. After an examination of these genres, redundancy was apparent. A mapping file was created to remove some of these redundancies within reason. For example, mapping “nonfiction” to “non-fiction”, mapping "humorous” to “humor”. Furthermore, children’s books had dozen of genres: “children’s literature”, “children’s non-fiction”, etc… These were all mapped to “childrens”. After the genre cleanup, with entailed turning every character into lowercase, removing out-of-place values, and mapping certain genres, 306 unique genres remained.

The author and description columns were also cleaned; All characters were formatted to lowercase and all non-ascii characters were removed. The author column in particular needed special attention, as author name convention was not uniform. For example, both “cast p. c.” and “p. c. cast” appeared, but both represented the same author. Regex and special rules were applied to fix these issues.

6.2 Network Creation

To facilitate network creation, 4 .csv files were created from the original file with only the necessary columns needed for each network. The Numpy, Pandas, and NetworkX libraries

for Python3 were utilized for creating the networks. The Matplotlib library was used to draw the networks.

6.2.1 Genre-Book Bipartite Network. The genre_book_network csv file was used to create this network. The network contains 2869 book nodes, and 306 genre nodes. 17976 edges connect the book and genre nodes together. Figure 1 shows the Genre-Book Bipartite Network. Note that a small unconnected subgraph of 1 book and 1 genre node is removed from the visualization. Book nodes are colored light-blue, and genre nodes in red.

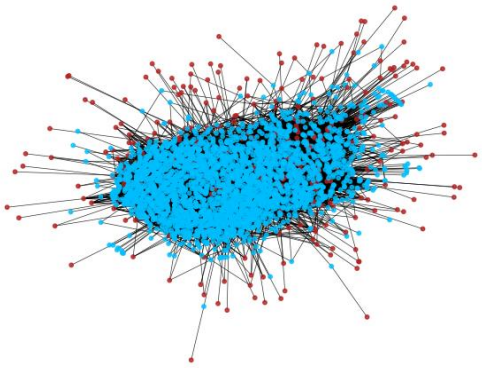

Figure 1: Genre-Book Bipartite Network

6.2.2 Author-Book Bipartite Network. The author_book_network csv file was used to create this network. The network consists of 2876 book nodes, and 2334 author nodes. 3600 edges connect the book and author nodes. Figure 2 shows the Author-Book Bipartite Network. Book nodes are colored light-blue, and author nodes in green.

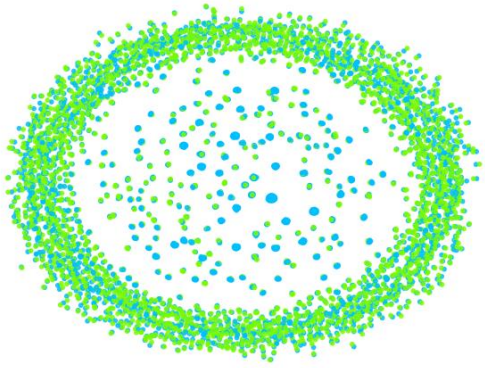

Figure 2: Author-Book Bipartite Network

6.2.3 Genre Co-Occurrence Network. The genre_book_network csv file was used to create this network. The network consists of 306 nodes representing genres. The network holds 4355 weighted edges between genres, representing books that have shared genres. Figure 3 shows the Genre Co-Occurrence Network.

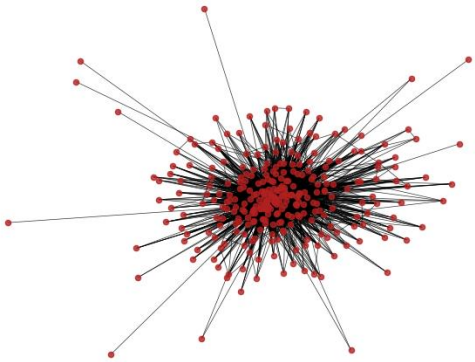

Figure 3: Genre Co-Occurrence Network

6.2.4 Author-Genre Bipartite Network. The author_genre_network csv file was used to create this network. The network contains 2327 author nodes, and 305 genre nodes. The 16245 edges connect the author and genre nodes. Figure 4 shows the Author-Genre Bipartite Network. Note that a small unconnected subgraph of 1 author and 1 genre node is removed from the visualization. Author nodes are colored green, and genre nodes in red.

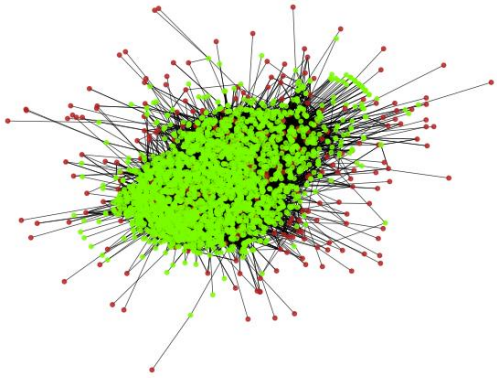

Figure 4: Author-Genre Bipartite Network

6.2.5 Book Similarity Network. The book_description_network csv file will be used to create this network. The nodes will be represented by nodes, and edges will be created with TD-IDF and cosine similarity. This network is the most complex of the 5 networks and will be added for the final project submission.

7 PRELIMINARY RESULTS

Using the Genre-Book bipartite network, degree centrality was conducted on both genre and book nodes. For the genre nodes, a higher degree centrality score demonstrates that the genre was more heavily censored than other genres. Table 1 shows these results. The degrees of the book nodes were transformed to a degree distribution graph, as can be seen in Figure 6. Lastly, eigenvector centrality values were calculated on the genre nodes; results can be seen in Table 2.

Table 1: Highest Degree Centrality Scores of Genres

Genre	Score	Genre	Score
Fiction	0.707	Audiobook	0.197
Young Adult	0.526	Children's	0.197
Contemporary	0.325	Historical	0.187
Romance	0.325	Queer	0.151
Fantasy	0.232	Non-Fiction	0.133
LGBT	0.226	Mystery	0.125
Realistic Fiction	0.226	Picture Books	0.110

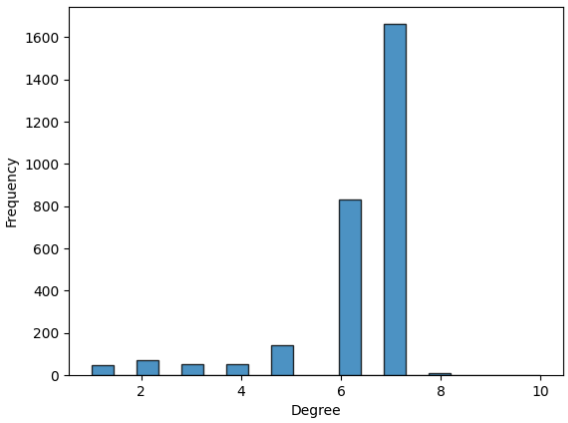

Figure 6: Degree Distribution of Book Nodes (Genres per Banned Book)

Table 2: Highest Eigenvector Centrality Scores of Genres

Genre	Score	Genre	Score
Fiction	0.194	LGBT	0.148
Historical	0.165	Realistic Fiction	0.145
Audiobook	0.162	Non-Fiction	0.145
Young Adult	0.161	Fantasy	0.135
Children's	0.158	Classics	0.132
Contemporary	0.158	Middle Grade	0.130
Romance	0.155	Mystery	0.130

REFERENCES

[1] *American Library Association*. 2024. Book Ban Data. Retrieved March 18, 2025, from https://www.ala.org/bbooks/book-ban-data.

[2] Carroll, S. 2020. The role of Catcher in the Rye in John Lennon's assassination. Medium. Retrieved March 28, 2025, from https://medium.com/@sec220/the-role-of-catcher-in-the-rye-in-john-lennons-assassination-4216033a75ed

[3] Frangedis, H. (1988). Dealing with the Controversial Elements in "The Catcher in the Rye". The English Journal, Vol. 77, No.7, 72-75, https://doi.org/10.2307/818945

[4] Duncan, S.P., Chen, H. (2023). Detecting Network-based Internet Censorship via Latent Feature Representation Learning. Computers & Security, Vol. 128, https://doi.org/10.48550/arXiv.2209.05152

[5] Langrock, I., LaViolette, J., Goncalves, M., Sargsyan, K., Carter, A. M., & Sherrill-Mix, S. (2024). Book bans in political context: Evidence from US schools. PNAS Nexus, 3(6), pgae197. https://doi.org/10.1093/pnasnexus/pgae197

[6] Steele, J.E. (2020). A History of Censorship in the United States. Journal of Intellectual Freedom and Privacy, 5(1), 6–19. https://doi.org/10.5860/jifp.v5i1.7208

[7] Chieler Li. 2024. *Banned Book Dataset*. Kaggle, from https://www.kaggle.com/datasets/chielerli/banned-book-dataset