COMP0123 Proposal

Ahamed Samir Sarker

December 2024

1 Dataset & Network Details

1.1 Dataset

- Dataset is Based on books properties extracted from goodreadbooks, it has details such as Title, Average Rating, Author, Genres and etc. (Data about 10352 books)
- Source: kaggle (`https://www.kaggle.com/datasets/middlelight/goodreadsbookswithgenres`)
- File Format: .csv

1.2 Network

- Nodes: Book genres

  Total nodes found: 890

- Edges(Weighted): co-occurring genres in books. E.g. if there are 2 books with genres Romance,action then there is a link from romance to action with weight 2.

  Total links found 27051

  6.84% possible edges exist.

2 Research Questions

2.1 Are Certain Genres Frequently Paired Together, and Do Prominent Genre Pairings Differ Across Book Ratings?

- Goal: Investigate whether specific genre pairs are more prominent in highly-rated books compared to lower-rated ones, identifying unique patterns across rating groups.
- Techniques: Calculate edge weights for genre pairs by grouping books based on their ratings, creating separate sub-networks for high-rated and low-rated books to compare connection patterns.

2.2 Which genres serve as bridges between distinct thematic clusters?

- Goal: Identify genres that act as connectors between otherwise distinct clusters, showcasing their role in bridging diverse themes and highlighting cross-thematic influence.
- Techniques: Perform community detection to identify distinct clusters of genres, and calculate betweenness centrality to find key genres that facilitate connections across these clusters

2.2.1 Subquestion: What are the most central genres in the network

- Use degree centrality