

Uncovering Book Popularity via Networks and Text Analysis

Team 5: Emma, Jae, Farhad



Agenda

Problem/Research Questions



Problem

Book ratings alone cannot explain why some books become popular or how they are interconnected through shared readers.

Aim: To understand what drives book popularity and reader engagement.

Approach:

- Build a 2 different types of networks
- Apply text analytics to review texts

Research Questions

1. Which genres have the strongest communities of readers based on the books they read?
2. What topics are talked about more positively vs negatively within each genre?
3. Which books serve as “bridge” between communities?



01

Motivation



Why this idea matters



Reveals hidden
book
communities



Connects
network
patterns with
review content



Supports
real-world
applications

02

Data



Goodreads book review

- Collected in 2017
- 8 genres, ~15m reviews, ~2m books, 465k users
- Link to data source:
<https://cseweb.ucsd.edu/~jmcauley/datasets/goodreads.html>

Columns Used

User ID	Anonymous ID for the reviewer.
Book ID	ID for the book being reviewed.
Review ID	ID for this specific review.
Rating	User's numeric score (0-5) for the book.
Review Text	Written comments about the book.



Romance

A book focused on a love story with happy ending

of Rows: 3,565,378
of unique book ID: 334,957
of unique user ID: 198,141



Fantasy

A book set in an imaginative world with magical or supernatural elements.

of Rows: 3,424,641
of unique book ID: 258,212
of unique user ID: 256,088



History/ Biography

Books that tell real stories about people or past events.

of Rows: 2,066,193
of unique book ID: 302,346
of unique user ID: 238,450



Preprocessing

- No missing values
- Selecting the first 10k
- Removing punctuation, stopwords, lowercasing, non-English

03

Social Network





Genre	Nodes and Edges
Romance	Book to book: 7675 Nodes 4056 Edges
	User to User (edge weight = 2): 386 Nodes 729 Edges
Fantasy	Book to book: 6465 Nodes 9112 Edges
	User to User (edge weight = 3): 177 Nodes 1149 Edges
History/ Biography	Book to book: 7485 Nodes 3196 Edges
	User to User (edge weight = 3): 116 Nodes 377 Edges

Book to Book Network

Nodes = books

Edge between two books = the same user reviewed both

Edge weight = number of reviewers ($n=2$)

Due to having so many nodes for our graphs we used the first random 300 nodes for each graph

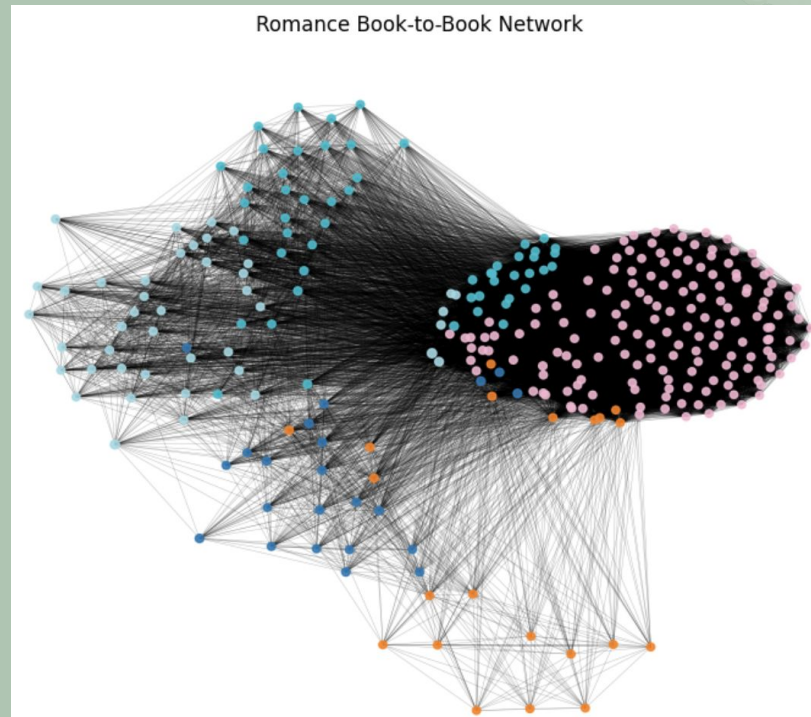
Example

Book A	Todd: A, B	A \leftrightarrow B (read by Todd)
Book B	Alex: A	B \leftrightarrow C (read by Jordan)
Book C	Jordan: B, C	A \rightarrow/C (no shared reader)
	Taylor: C	



Book to Book Romance

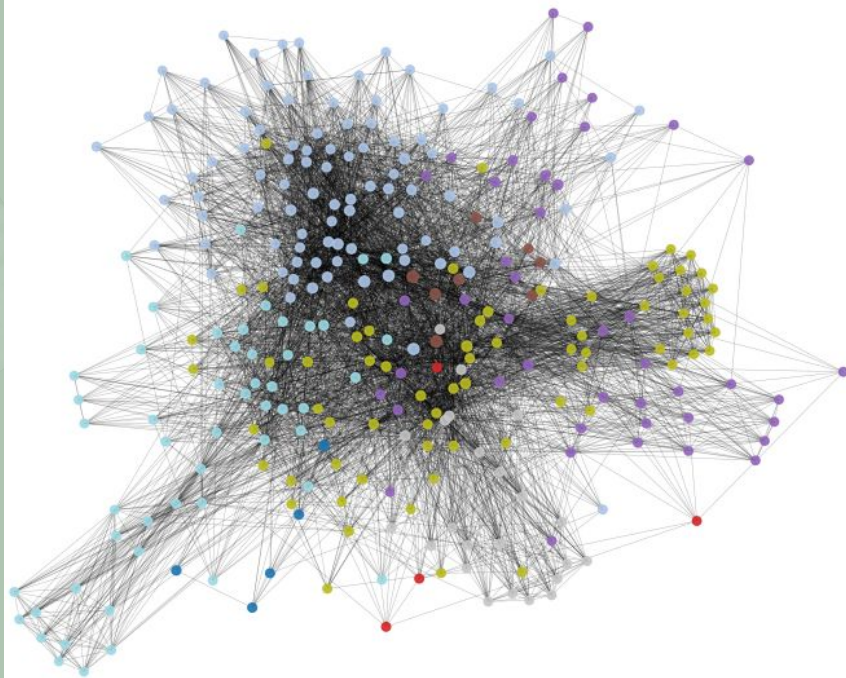
- **Very sparse** (density = 0.0014), meaning few books share the same readers.
- Each book connects to **~10** others on average, showing moderate overlap.
- A **clustering coefficient of 0.12** indicates small, tightly connected sub-groups.
- Community detection shows many small groups, with the largest cluster containing **319 books**.





Book to Book Fantasy

Fantasy Book-to-Book Network



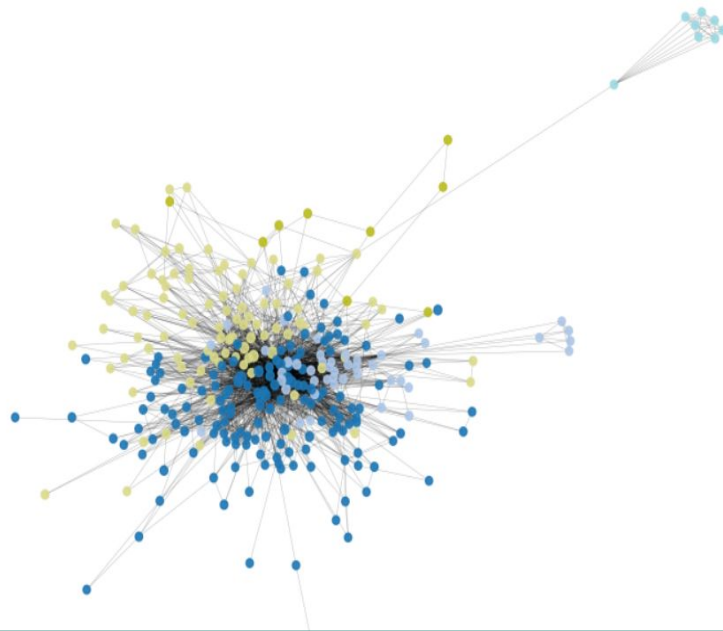
- **Sparser** than Romance (density = 0.00044)
- Each fantasy book connects to **2-3** others on average
- **Clustering coefficient = 0.098**, sub-groups are less tightly interconnected
- Largest fantasy community has **204** books
- Visualization looks less organized and **lacking a dominant core blob**.

Book to Book History/ Biography



- **Very sparse network** (density = 0.00013), even sparser than Fantasy.
- On average, each book connects to **less than 1 other book** (avg degree = 0.89).
- Clustering is **very low** (0.042), meaning most books do not form tight groups.
- The **largest connected group** is only 125 books — much smaller than Romance or Fantasy.
- Visualization shows a **small dense center** with many books sitting on the outskirts.

History Book-to-Book Network



Book to Book Comparison:

There's a **clear gradient**:
Romance → Fantasy → History
becomes **less connected, less cohesive, and less centralized.**

The largest cluster size shrinking from **319 → 204 → 125** visually illustrates how genre communities break down as the audience becomes more fragmented.



This suggests Romance readers often overlap heavily, Fantasy readers overlap moderately, and History readers read in narrower, more niche patterns..

1. Which genre has the strongest communities of readers based on the books they read?



3. Which books serve as “bridge” between communities?

	book_id	betweenness	avg_rating
1744	15717943	0.001310	4.666667
2035	16073738	0.000904	4.333333
1655	13644052	0.000771	4.700000
4756	22573348	0.000744	4.000000
1204	11505797	0.000712	4.466667
2086	16102004	0.000668	3.571429
1751	15724654	0.000667	3.333333
3198	18143950	0.000648	4.545455
1794	15760001	0.000607	4.700000
2828	17828418	0.000533	4.000000

Romance

	book_id	betweenness	avg_rating
0	256683	0.001857	2.968750
1	11235712	0.001232	4.066667
2	16096824	0.001161	3.391304
3	20821111	0.000330	3.600000
4	6644117	0.000422	3.916667
5	9460487	0.001435	3.933333
6	15839984	0.000481	3.300000
7	345627	0.000473	3.833333
8	42899	0.000433	3.666667
9	13206828	0.000469	4.000000

Fantasy

	book_id	betweenness	avg_rating
3261	4667024	0.000610	4.314286
279	19063	0.000515	4.277778
44	2657	0.000465	4.583333
2987	2728527	0.000436	4.166667
181	10964	0.000331	3.909091
138	7445	0.000253	3.950000
3805	7148256	0.000202	3.000000
4120	8664353	0.000201	4.642857
89	5043	0.000173	3.176471
6255	21853621	0.000171	4.666667

History/
Biography

User to User Network

Nodes = Users

Edge between two books = users who reviewed the same books

Edge weight = how many books they have in common
(Different between genres either ≥ 2 or 3)

Example

Book A

Book B

Book C

Todd: A, B

Alex: A

Jordan: B, C

Taylor: C

Todd ↔ Alex (both read Book A)

Todd ↔ Jordan (both read Book B)

Jordan ↔ Taylor (both read Book C)

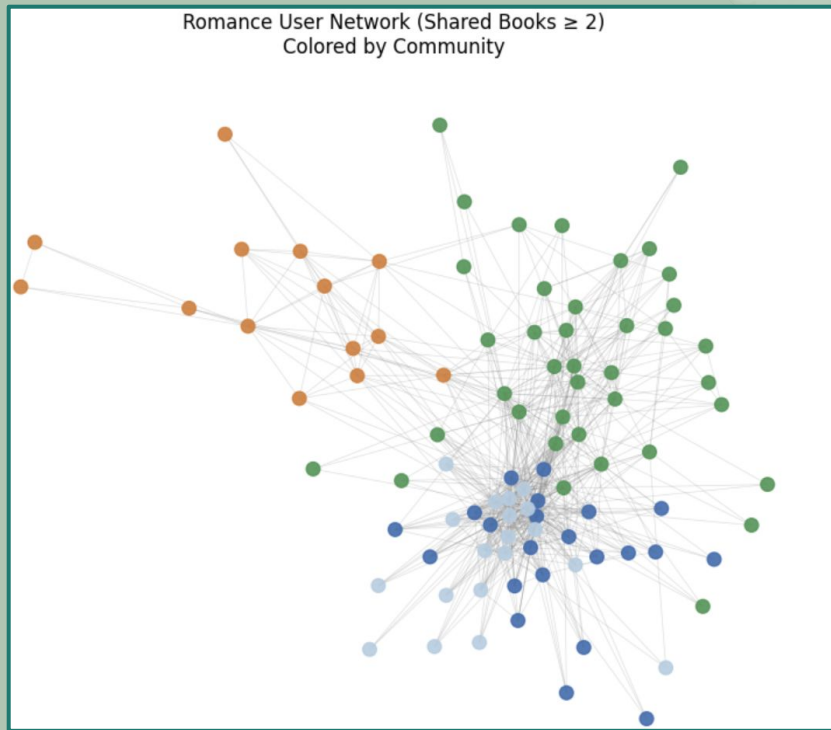
Alex →/ Jordan (only share one book, but not enough)

Alex →/ Taylor (share nothing)



User to User Romance

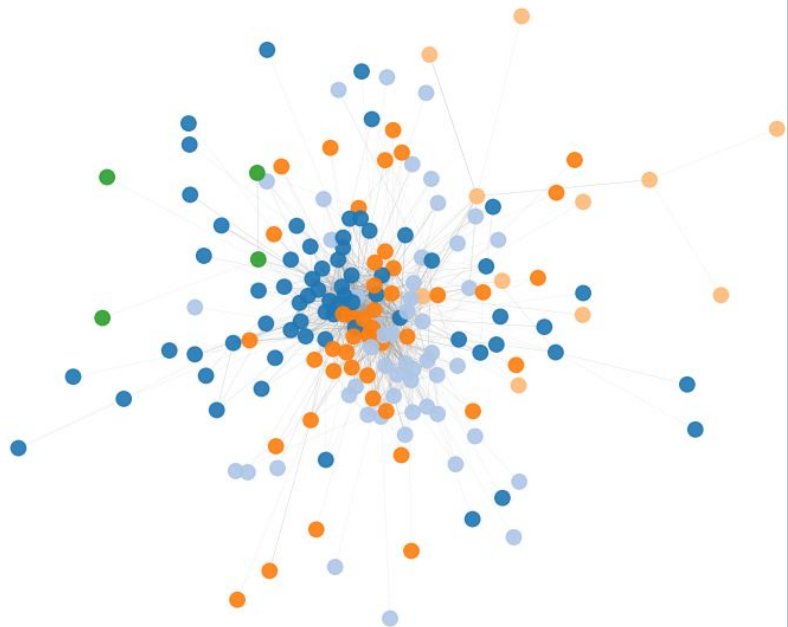
- **386** Romance users connected (≥ 2 shared reviews).
- **Very sparse interaction network** (density 0.0098) with low cohesion (avg clustering 0.18)
- Forming **262 micro-communities**; largest group has 60 users.
- **Average degree ~3.8**: Romance readers share few books with each other





User to User Fantasy

User Network (shared books ≥ 3)
Colored by Community



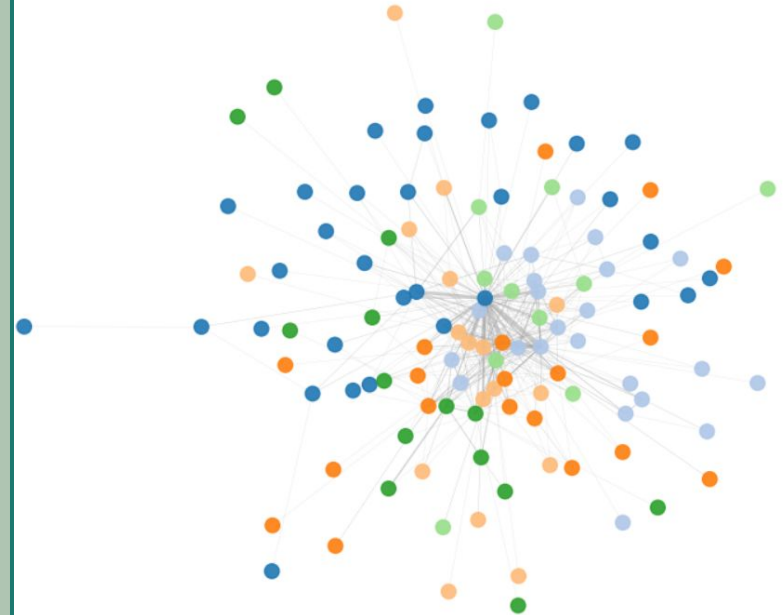
- Total **177 users** connected (≥ 3 shared reviews).
- **Sparse interaction network** (density 0.074) with moderate social cohesion (avg clustering 0.52).
- Users form **5 communities**, with the largest group containing 67 users.
- **Average degree ~13**: Fantasy readers tend to overlap heavily in what they read



User to User History/ Biography

- Total **116 users** connected (≥ 3 shared reviews).
- **Sparse interaction network** (density 0.057) with moderate social cohesion (avg clustering 0.48).
- Users form **6 communities**, with the largest group containing 32 users.
- **Average degree 6.5**: History readers tend to overlap strongly in what they read

User Network (shared books ≥ 3)
Colored by Community



User to User Comparison

1

Fantasy Has the Most Overlap

Fantasy readers **share the most books**, creating the strongest user connections and the densest core network.

2

Romance Has the Most Users but Weak Ties

Romance has the **largest user base**, but readers overlap very little, forming many tiny, scattered micro-communities.

3

Fantasy Readers Form Fewer but Larger Communities

Fantasy readers cluster into big shared-interest groups rather than many tiny circles.

1. Which genre has the strongest communities of readers based on the books they read?



04

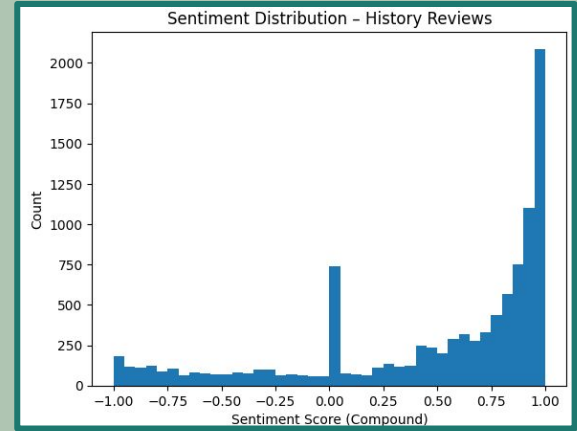
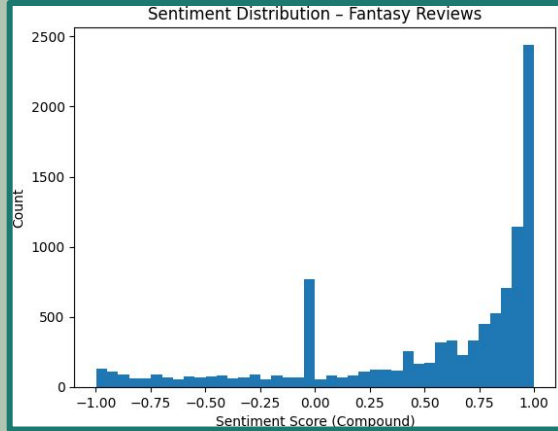
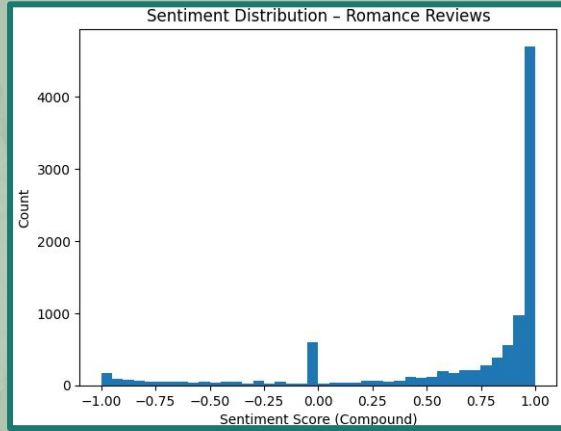
Text Analytics



Top 10 Keywords Across Genres

Romance	Fantasy	History/Biography
Book Read Love Story Like Loved Series Stars Really Characters	Book Read Story Series Really Like Love Characters Good Books	Book Read Story Good Like Really Love Time Life Characters

Sentiment Distributions Across Genres



Question 2:
**What topics are talked about
more positively vs negatively
within each genre?**



Romance Reviews

- **Most positive topics**
 - *Romance themes, storylines, characters, and relationships* (avg sentiment 0.80).
 - Readers also speak positively about topics involving *personal, emotional, or life themes* (avg sentiment 0.69).
- **The *largest review cluster* focuses on reading experience**
 - *"Book, Read, Like, Love"*
 - Moderately positive (avg sentiment 0.65).
- **Most negative topic but small cluster**
 - Specific *author or character names* (avg sentiment 0.03)
 - Indicating mixed or critical reactions.



Fantasy Reviews

- **Most positive topics**

- General enjoyment of the reading experience (avg sentiment 0.59): *"Book, Story, Read, Characters"*
- Romantic & adventure driven fantasy (0.55): *"New, Love, Romance, Fun, Paranormal"*

- **Mixed or negative topics**

- Mixed reactions to royalty/kingdom tropes (0.24): *"Kingdom, Queen, Prince"*
- A character-focused topic with negative opinions (-0.01): *"Percy, Katy, Demon, Juliette"*



History/ Biography Reviews

- **Most positive topics**

- Praising the work (avg sentiment 0.56): *"Book, Read, Story, Characters"*
- Major history theme (0.43): *"Life, War, People, History"*

- **Mixed or negative topics**

- Political theme – close to neutral sentiment (0.29): *"President, Political, Lincoln"*
- "Golden, hours, hornblower, radio" – negative sentiment (-0.34), low instance count, topic seems obscure, likely about a certain historical event.

Reviews Comparison



Romance
readers are
the most
positive

The most negative
topics come from
very small, niche
clusters.

Readers respond
most positively to
topics about the
overall reading
experience — not the
genre-specific
content.

Key Practical Insights



Launch a Romance/Fantasy “Community Favorites” badge. Use high-sentiment, high-degree books to create a label like “#1 in Romance Community”.



Use high-betweenness “bridge books” for “Customers Also Liked” recommendations.



Highlight books with strong “overall reading experience” sentiment in promotions, since these topics consistently drive the most positive reactions across all genres.

Limitations and Directions for Future Work

No book title available

Hard to interpret results without titles.



Request book titles or author information in future datasets.

Users across genres

Did not track users reading across multiple genres.



Build a unified user graph.

Updated data

Dataset is outdated (collected in 2017).



Collect recent reviews for trend analysis

Conclusion



Romance has the
strongest
community
structure.

High potential for targeted
marketing and bundled
sales strategies.



Bridge books can
improve
recommendations.

Future datasets need
titles for meaningful use.



Future Analysis on
all genres

Look Deeper into
connections between all
genres.



Thanks!
Questions?