

### **Table of Contents**

Problem Statement and Approach

Topic Modeling
Analysis

05 Insights

O2 Goodreads Book Data O4 Bi-partite Network Modeling

06 Final Remarks

# Problem Statement

To infer the **competitive market characteristics** of the book publisher
industry by leveraging the vast repository of **book data** available through Goodreads



Lorem ipsum dolor sit amet consectetur adipiscing elit massa sociis mollis faculis, arcu fermentum per ullamcorper id at tempor rpis volutpat curae, quisque vitae montes egna sodales purus fringilla mattis sem nunc. em rhoncus consequat mollis turpis porta, vivamus bibendum porttitor suscipit.

fristique curae mollis porta sed fusce tempor lacus est mus magna iaculis, eu ultricies aliquet aptent diam accumsan varius auctor non Dapibus purus varius potenti sociosqu pulvinar quis ut sagittis, per turpis fermentum litora Magnis class tincidunt, faucibus auctor tristique inceptos erat vehicula eu. Platea metus auctor blandit id cubilia dapibus tristique bibendum praesent laoreet nulla Ornare pharetra sociosqu penatibus, gravida mi aliquam malesuada condimentum eu imperdiet at

Tristique curae mollis porta sed fusce tempor lacus est mus magna iaculis, eu ultricies aliquet aptent diam accumsan varius auctor non. Dapibus purus varius potenti sociosqu pulvinar quis ut sagittis, per turpis fermentum litora magnis class tincidunt, faucibus auctor tristique inceptos erat vehicula eu. Platea metus auctor blandit id cubilia dapibus parturient dictum risus aliquet aptent diam accussan varius suctor non, Copfus purus
varius potenti sociosus pulvinar quis ut
varius potenti sociosus pulvinar tora sagitits, per turpis ferentum ticales tincidunt, faucisus suctor tristique
class tincidunt, faucisus suctor tristique
potential pulvinar polatea netus auctor inceptos erat vehicula eu. Platea metus auctor llandit id cubilia dapibus parturient dictum nimment lagretyolutpat, rhoncus tristique bibendum prosent lacret mulla ormare phinetra sociosqu penatbus, gravida ni aliqua nalesuada condimentum eu imperdiet at nam. Lorem ipsum dolor sit amet consectetur adipiscing elit massa sociis mollis iaculis, arcu fermentum per ullamcorper id at tempor turpis volutpat curae, quisque vitae montes magna sodales purus fringilla mattis sem

turpie volutpat curie, qu'illa matria sen pur magna sodates purus frincip neuts hac l'el lementum tempor scale purus frincip neuts hac l'el lementum tempor scale all'is turpis port sen rhonoux consequat editis turpis port sen rhonoux consequat editis turpis port vivamus bibendum portitire suscipit.

Vivamus bibendum porta sed fusca (con l'acceptant de la legis est una nagna sicultis, eu ultricles l'agus est una nagna sicultis l'agus est una n magnis class tincidunt, faucibus auctor tristique inceptos erat vehicula eu. Plater

- 1. Connect publishers to the topics their books are about using LDA topic modeling on book descriptions
- 2. Perform bi-partite **network analysis** to infer competitive network characteristics

### **Goodreads Book Data**



#### **Tools**

- BeautifulSoup : Get Book
   IDs from genre list
- BetterReads: Get information on the books



### **Information**

- Title
- Author
- Publisher
- Language
- Description



### Cleaning Data

- Removed foreign books
- Removed any books with missing information

### **Goodreads Book Data**



#### **Young Adult**

**Books:** 3341

Publishers: 1321

Authors: 2301

#### **Top Publishers**

- 1. HarperTeen (74)
- 2. HarperCollins (70)
- 3. Simon Pulse (69)



# Science Fiction and Fantasy

**Books:** 3096

**Publishers:** 1210

Authors: 1592

#### **Top Publishers**

- 1. Del Rey (98)
- 2. Tor Books (91)
- 3. Ace (76)



### **Crime and Mystery**

**Books:** 3099

Publishers: 1032

Authors: 1577

#### **Top Publishers**

- 1. Minotaur Books (65)
- 2. Bantam (62)
- B. Grand Central Publishing (60)

### **Topic Modeling**



1. LDA topic modeling based on synopsis



2. Topics for each book within a genre
(Book Topics = Topic Weights > 0.4)



3. Aggregate by publisher and topic



4. Topics for each genre

## Topic Modeling - Genres & Topics

#### **Young Adult**

- Adventure
- Love
- Mysterious
- Life
- High School
- Transition
- Tragedy
- Friendship
- Family
- Drama

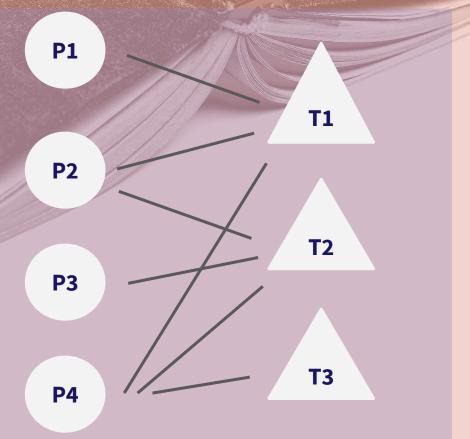
#### Sci-Fi

- Adventure
- Mythical
- Alien
- Medieval
- Magical
- Love
- Space
- Heroic
- War
- Historical

#### **Mystery**

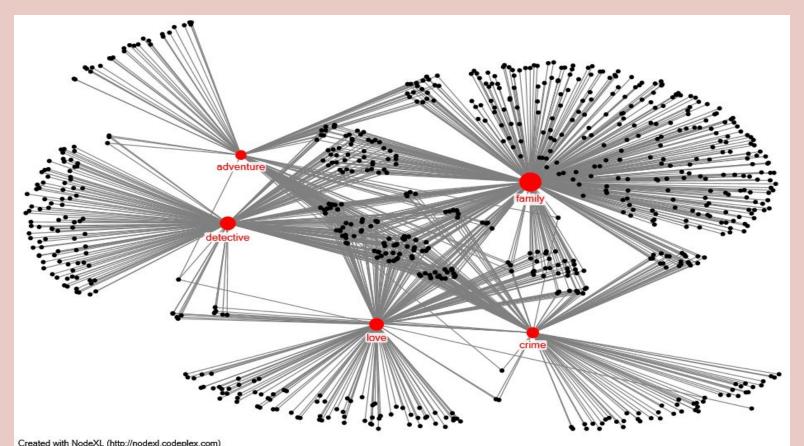
- Crime
- Love
- Detective
- Adventure
- Family

# Bi-partite Network Modeling

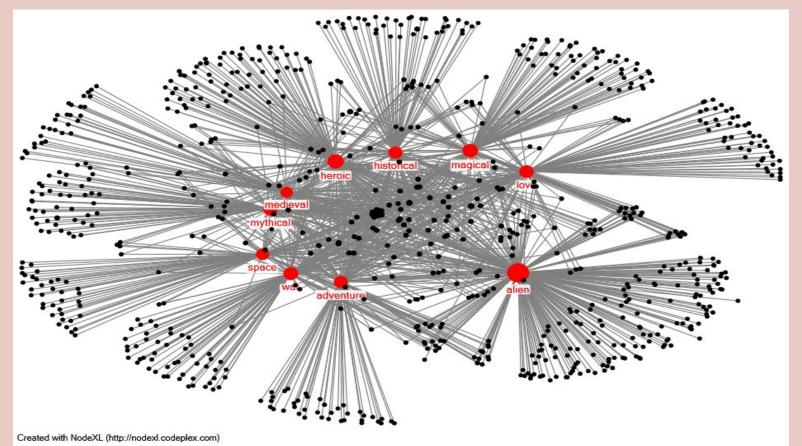


- Created a bi-partite network of Publishers and Topics for each genre
- Computed a matrix with the publishers on each axis with cell values as the number of topics in common
- De-constructed the bi-partite network into one of publisher-to-publisher based on number of topics in common

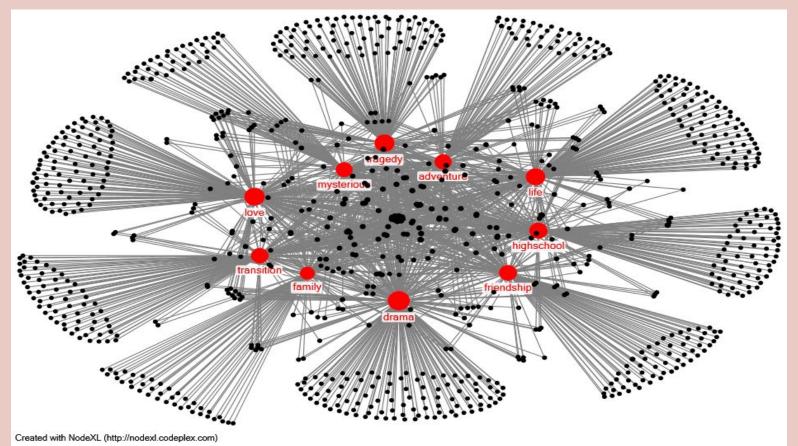
# Bi-partite Network: Mystery



# Bi-partite Network: SciFi



# Bi-partite Network: Young Adult



## **Network Analysis**

Genre	Num publishers with more than 2 topics	Num publishers in the main community
Young Adult	189	188
SciFi	123	122
Mystery	121	112

- Created a publisher-to-publisher network with an edge between publishers with > 2 topics in common
- Applied Girvan-Newman to find communities within the network
- Publisher-to-publisher network did not yield very interesting results

## **Insights**

#### For Publishers

- Understand crowding of book topics so if a publisher wants to deviate from the norm, it can publish on something "niche"
- By the same token, understand the potential of a book based on what it is about if topic is popular

#### **For Authors**

- Know which publishers to go to based on the topics they publish about
- Write on successful but niche topics

### Final Remarks

- There is value in understanding "niche" publishers with only 1 associated Topic
- A better approach would be to use Girvan-Newman on the bi-partite networks
- An extension would include looking at the network across genres

# **Thanks**

Does anyone have any questions?

CREDITS: This presentation template was created by Slidesgo, including icons by Flaticon, and infographics & images by Freepik.

Please keep this slide for attribution.

