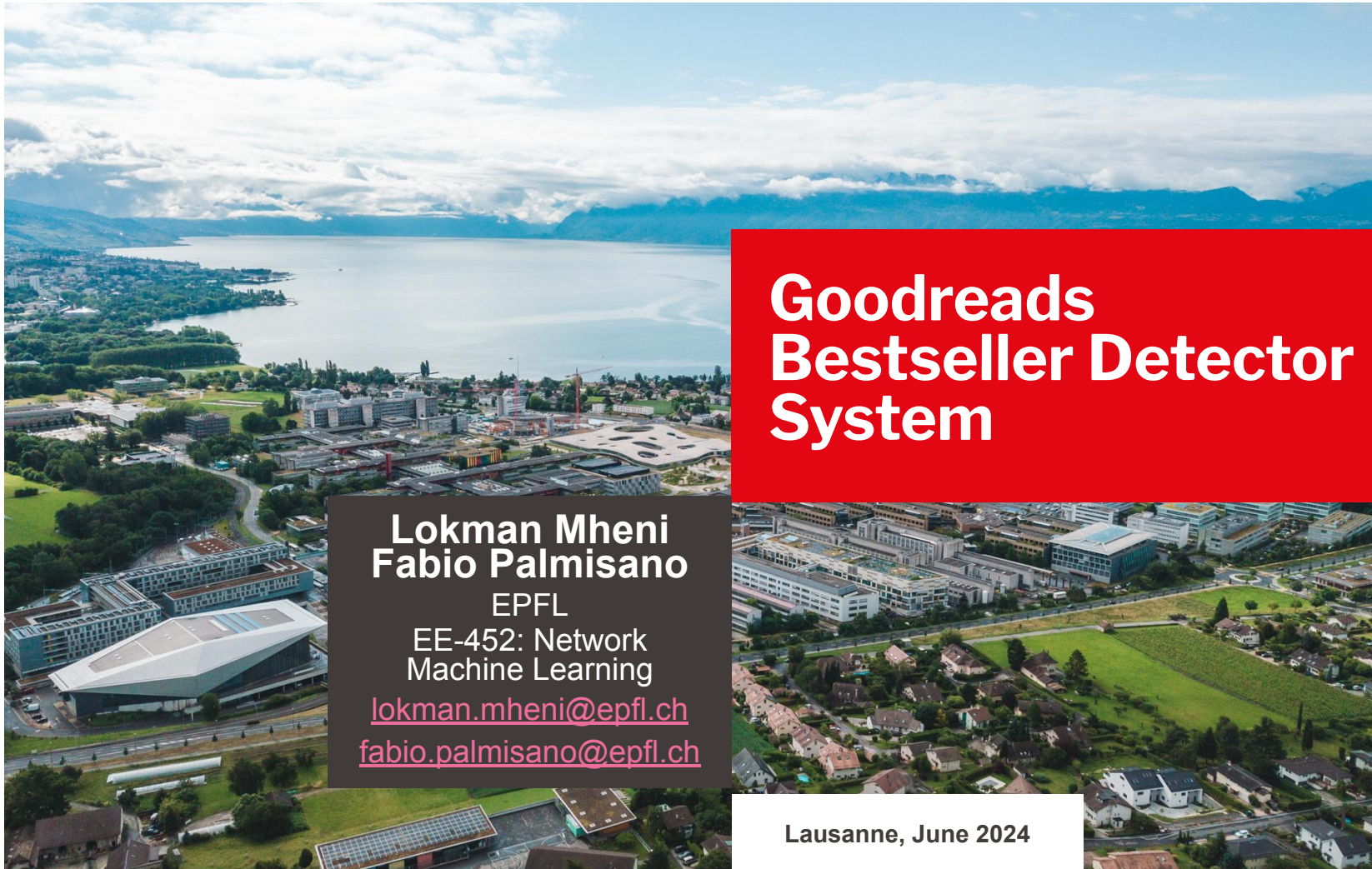
Goodreads Bestseller Detector System



Lokman Mheni
Fabio Palmisano
EPFL
EE-452: Network Machine Learning
lokman.mheni@epfl.ch
fabio.palmisano@epfl.ch

Lausanne, June 2024

École polytechnique fédérale de Lausanne

Table of Content

Introduction

Theoretical Elements

“To read” book-to-book

“Ratings” book-to-book

Critical Analysis

Conclusion

Introduction - Project Overview

Lokman Mheni

- This project investigates the application of graph network analysis to determine if a book is a **bestseller or not**.
- Using the **goodbooks-10k dataset**, multiple graphs will be constructed.
- Big datasets are considered so sampling needs to be done.
- Machine learning techniques, including linear models and Graph Neural Networks, will be used to refine the determination of blockbusters.

Introduction - Which data are used ?

- Using the **goodbooks-10k dataset**
  - To Read : Users labelled books as “to read” (912’705)
  - Ratings : All the ratings done by users (~ 6 millions ratings, 53’424 users)
  - Books : Author, Year, Number of Ratings (10’000 books)
  - Tags : Not used
  - Books Tag : Not used

- Due to computational cost, data need to be reduced (sampling)

Introduction - What is a Bestseller?

- Subjective considerations had to be made to determine what qualifies as a **blockbuster** and what does not.

- For a book to be designated as a Bestseller, it must be marked as "to read" by more than 50 users

Introduction

Theoretical Elements

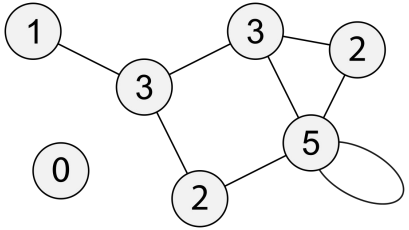
- Main Graph properties

“To read” book-to-book
“Ratings” book-to-book
Models
Results
Critical Analysis
Conclusion

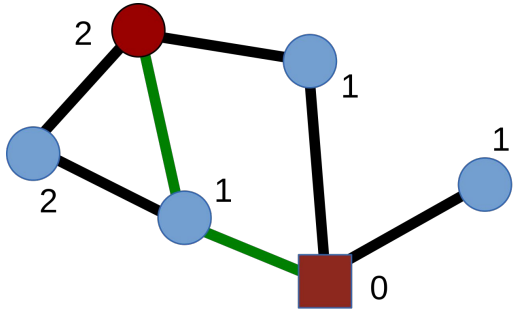
Lokman Mheni

Main Graph Properties

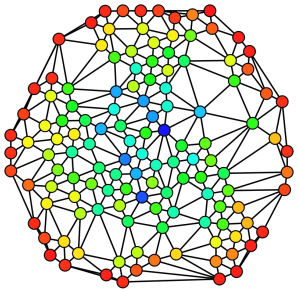
- **Node degree**



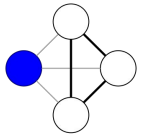
- **Closeness centrality**



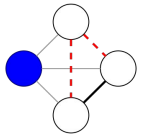
- **Betweenness centrality**



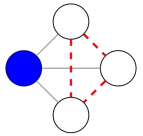
- **Clustering coefficient**



c = 1



c = 1/3



c = 0

Introduction
Theoretical Elements

“To read” book-to-book

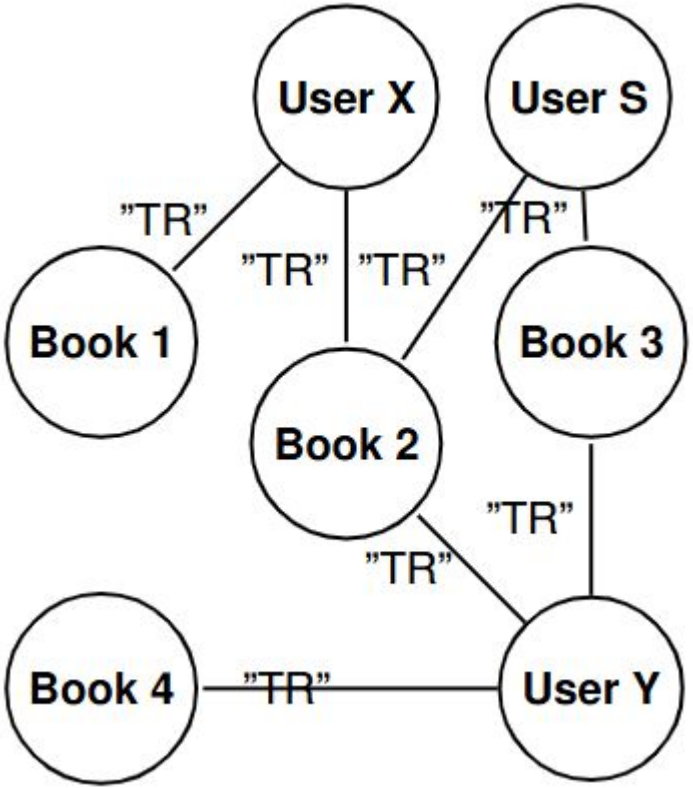
- Data Analysis
- Graph Construction
- Models
- Results

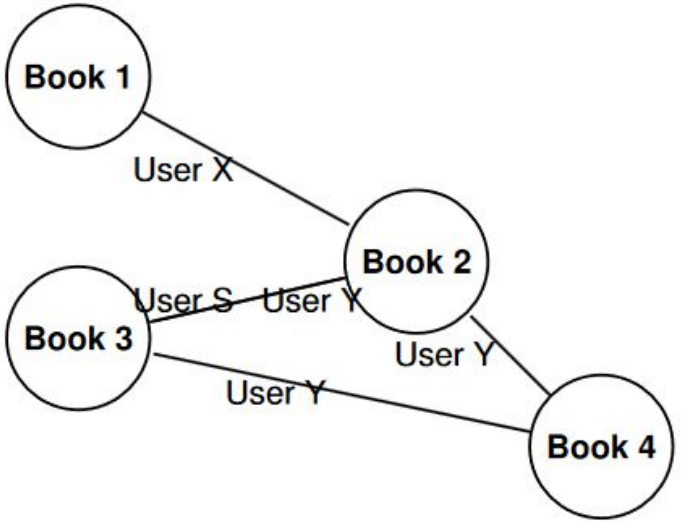
“Ratings” book-to-book
Critical Analysis
Conclusion

Data Analysis

- **Strong correlation** between frequency of “to read” and number of ratings
- **No correlation** between percentage of “to read” tags and number of ratings.
- **No correlation** between number of “to read” tags and average rating grades.
- **20%** of raters give a “To Read” tag
- Very **unbalanced** number of “to read” tags per book

Book-Book Graph with “to read”





Models

- Used graph properties as features
- Binary labelling (1 to read, 0 not to read)
- Only linear models trained, due to “triviality” of the task

TABLE I: Graph one Book-to-Book Trivial Case

Model	Features	Accuracy	F1-score
Logistic Regression	All	0.744	0.803
Logistic Regression	Selecting the Best	0.716	0.813
Random Forest	Best Features	0.739	0.805
Random Forest	All	0.748	0.814
SVM	All	0.718	0.807

Introduction
Theoretical Elements
“To read” book-to-book

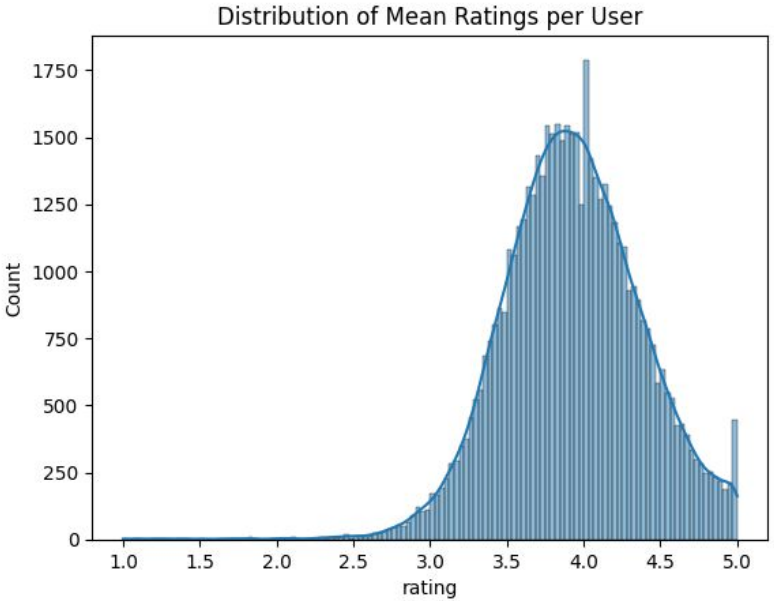
“Ratings” book-to-book

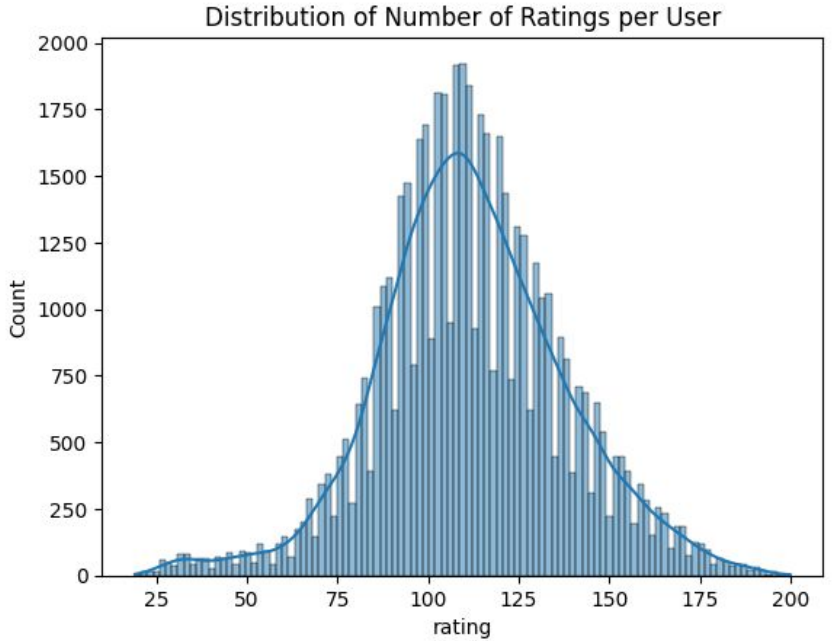
- Data Analysis
- Graph Construction
- Models
- Results

Critical Analysis
Conclusion

Data Analysis

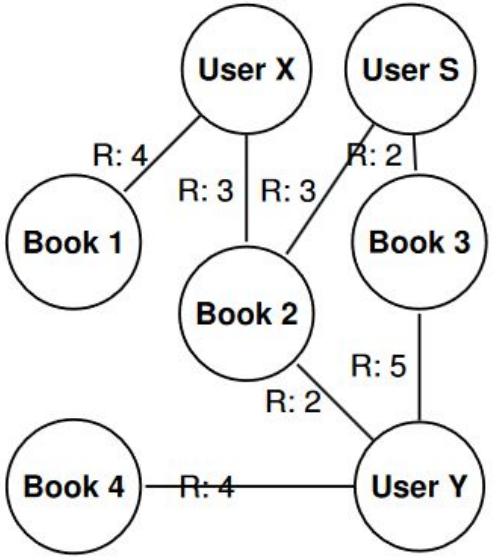
- Need to sample the data not in random way
- Use the Gaussian distribution to extract relevant data
- Use finally the 40th and 60th percentile (computational cost)

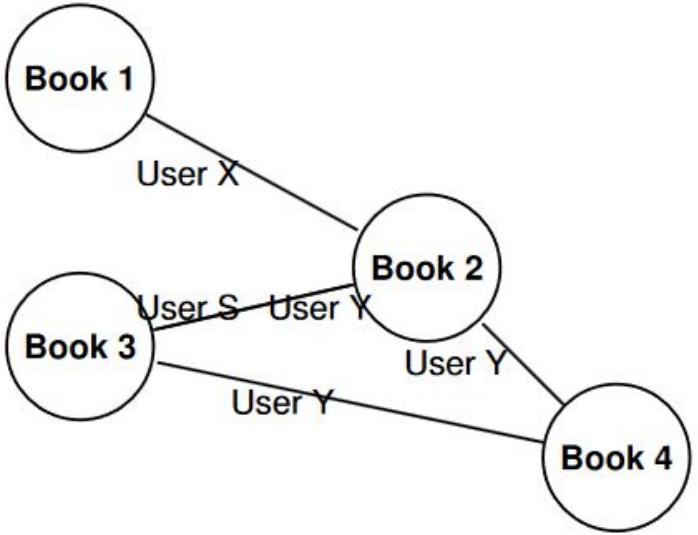




Fabio Palmisano

Graph Creation





- 2’578 Users (4%)
- 9’823 Books (98%)
- 6 millions of edges
- Weights : Number of Users ratings same books

Models

- Used graph properties as features
- Binary labelling (1 to read, 0 not to read)

TABLE II: Graph Book-to-Book Only with graph properties

Model	Features	Accuracy	F1-score
Logistic Regression	Some properties	0.746	0.683
Logistic Regression	All properties	0.746	0.683
Logistic Regression	Best features	0.754	0.692
Simpler Neural Model	Best features (2 L.)	0.747	0.65
Simpler Neural Model	Best features (3 L.)	0.754	0.69

Models

- Add book statistics to features
- Binary labelling (1 to read, 0 not to read)
- Linear models results improved
- GCN predicts almost only 1s

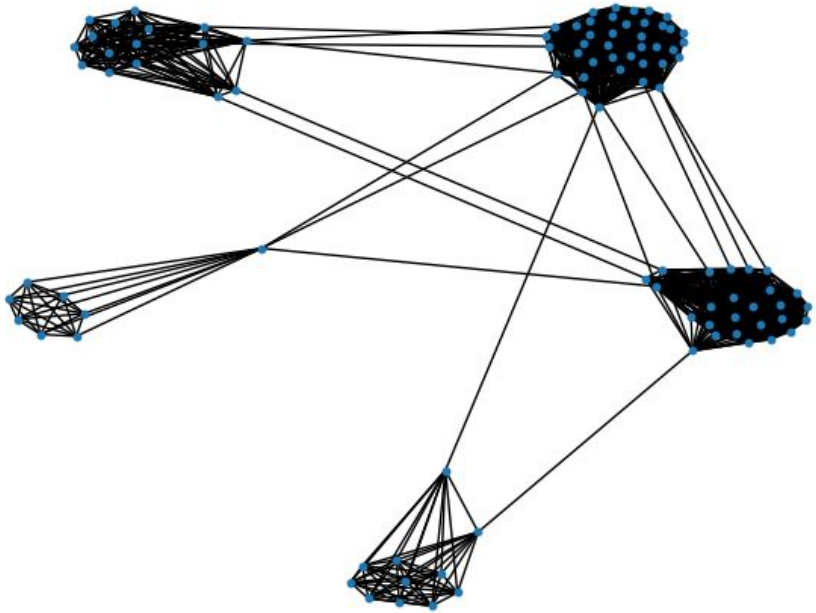
TABLE III: Graph Book-to-Book adding books statistics

Model	Features	Accuracy	F1-score
Logistic Regression	All Features	0.809	0.775
Logistic Regression	Best Features	0.821	0.787
Random Forest	Best Features	0.829	0.800
Random Forest	Grid Search	0.830	0.799
Random Forest	Best + Standardize	0.833	0.803
MLP	All + Standardize	0.755	0.687
GCN	All	0.443	0.613

Models

- Tests to overcome GCN results issue
  - test simpler models to check overfitting issues
  - exploration of different learning rates and optimizers
  - early stopping
  - data preprocessing variations

User-to-User Graph



Introduction
Theoretical Elements
Data Analysis
Graphs
Models
Results

Critical Analysis

Conclusion

Key Findings

- The use of logistic regression and random forest models yielded reasonable accuracy and F1-scores.
- Adding statistical features of books improved model performance.
- Graph-based features provided useful insights but were limited by the size and complexity of the dataset.
- Neural network models, including GNNs, struggled with prediction consistency, indicating potential areas for further investigation and improvement.

Possible Improvements

Several areas for improvement were identified :

- Refining threshold
- Enhancing edge feature management
- Using the complete dataset
- Perform some **meaningful** results on GNN

Thank you

Lokman Mheni
Fabio Palmisano
EPFL, Lausanne
School of Engineering
Institute of Electrical Engineering