Amazon Books Reviews

Network Science and Recommender Systems
applied to the retail market

Riccardo Tenuta
Matr n° **26940A**

Table of contents

01 Introduction

Introduction of the dataset and data pre-processing

02 Community detection

Computation of Louvain algorithm and community extraction

03 Network analysis

Analysis of the principal characteristics and metrics of the network

04 Reccomender System

Graph embedding with Node2Vec for similarity prediction

01

Introduction

The dataset and data pre-processing

Introduction

The dataset describe and contains data related to books that have been purchased and reviewed on Amazon. The data are split into two different datasets.

- **books_rating**: contains information about the reviews such as the score, the user id and the title of the book
- **books_data**: contains specifics of the book such as the category, the author and the decription

The goal of the project, starting from the two dataset, is to create a network generated with nodes representing the books more liked by the customers, so with a review score greater than 4, and then to study the relationships betweeen the links generated between the nodes. A link between two books exists if the books have been reviewed by at least a common customer, and the weight is the number of customers that have reviewd both.

The datasets

Main atributtes used for the analysis

books_rating.csv

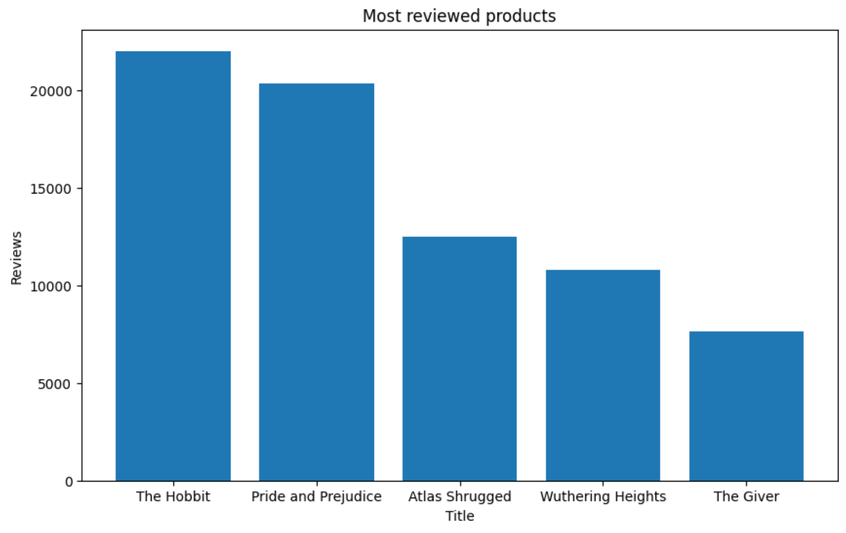
- Id
- Title
- User_id
- Review score
- Review summary

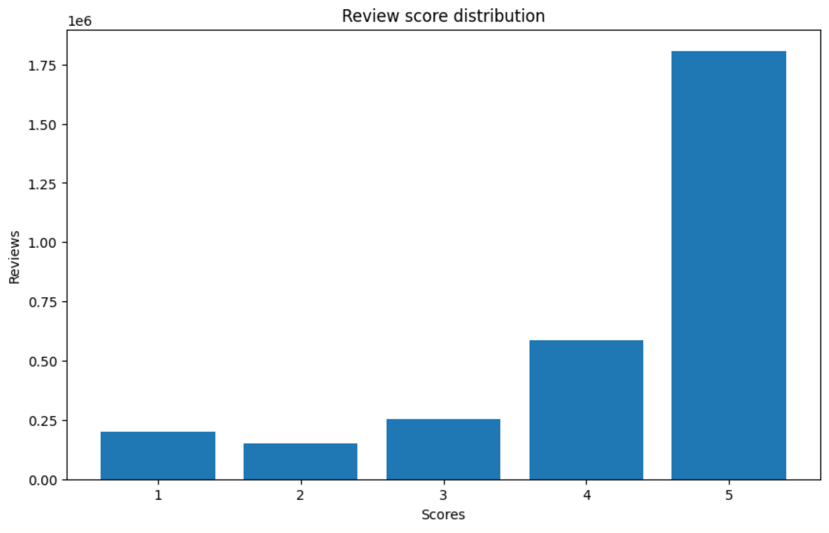
books_data

- Title
- Description
- Author/s
- Link
- Categories

Data exploration

Some plots about the final dataset





Top 5 most reviewd books

Review score distribution

02

Network Analysis

Analysis of the principal characteristics
and metrics of the network

The Graph

The graph has been built considering the books as nodes and the edges as the number of customers that reviewed both the books.

Only the best reviews has been taken into acocunt in order to build a better suggestion network. So only reviews with score greater than **4** have been considered.

The graph sampled the 1% of the total reviews

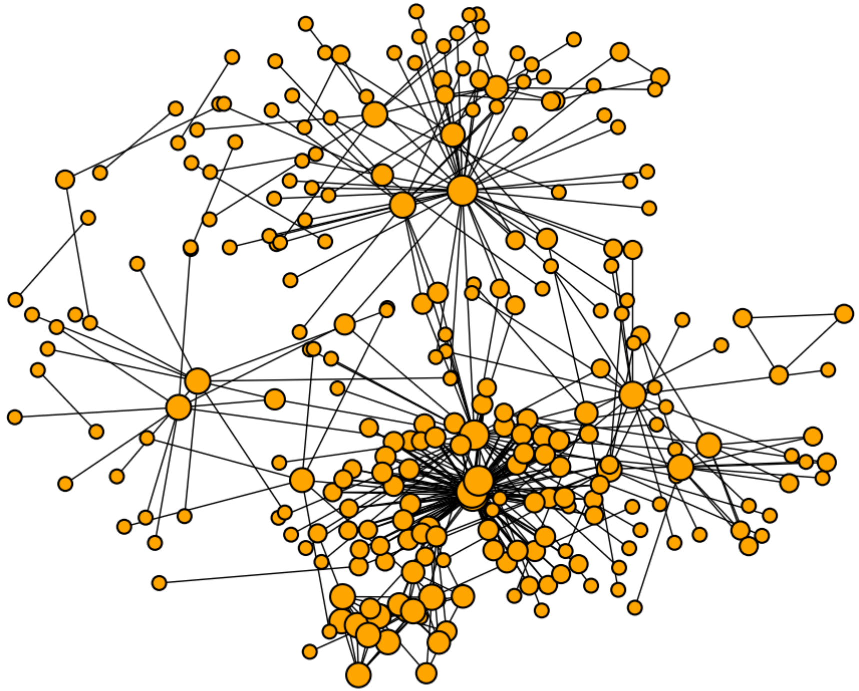
The Graph

N° nodes: 2217
N° edges: 5929

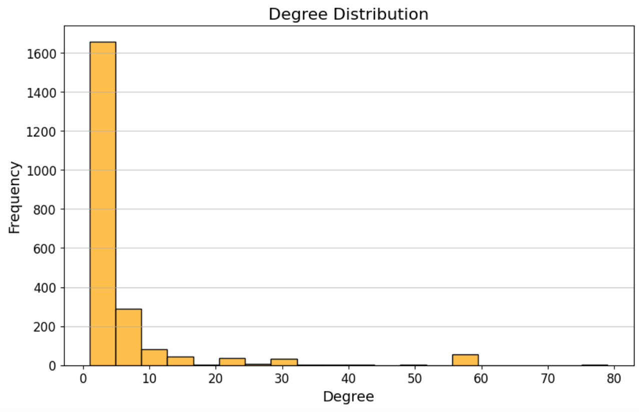
Density: 0.002

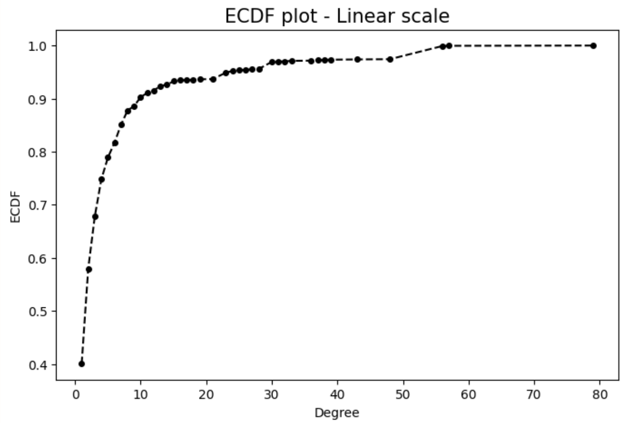
Average degree: 5.35
Median degree: 2
Max degree: 79
Min degree: 1

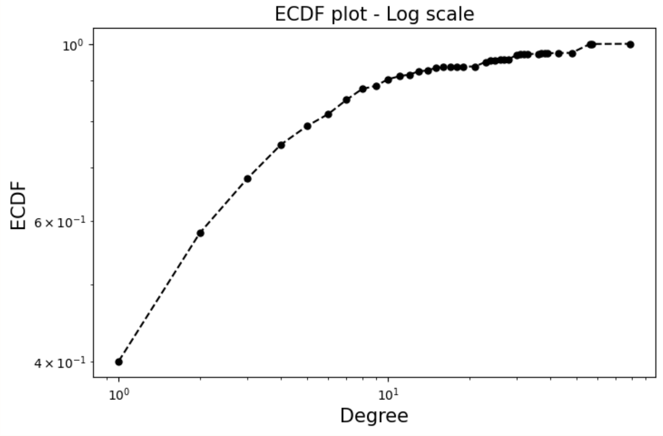
Std. Deviation: 10.02

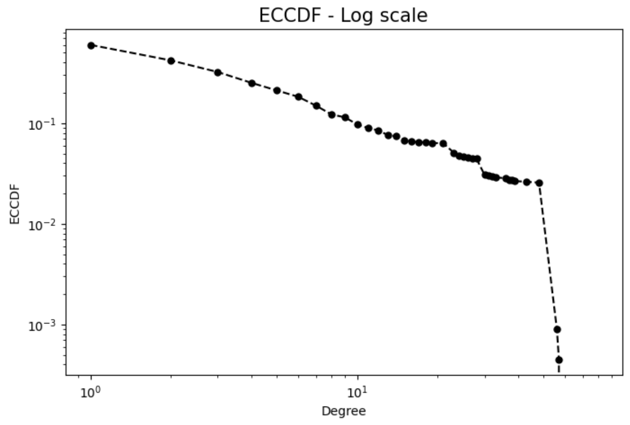


The Degree distribution

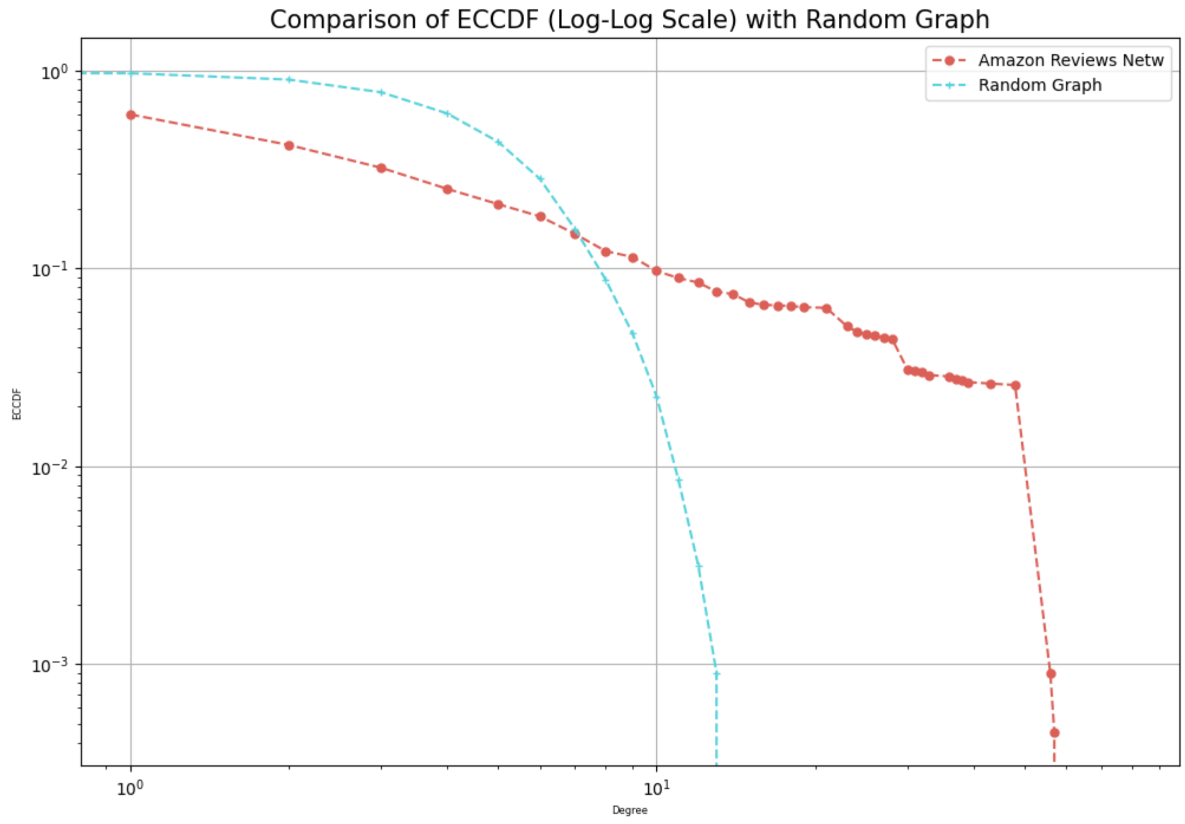








The network vs. a random network



Centrality measures

Degree

Anansi Boys, 0.035
Alice's Adventures in Wonderland, 0.025
Stories of Hope and Spirit, 0.025

Closeness

The Picture of Dorian Gray, 0.117
Pride and Prejudice, 0.115
Wuthering Heights, 0.109

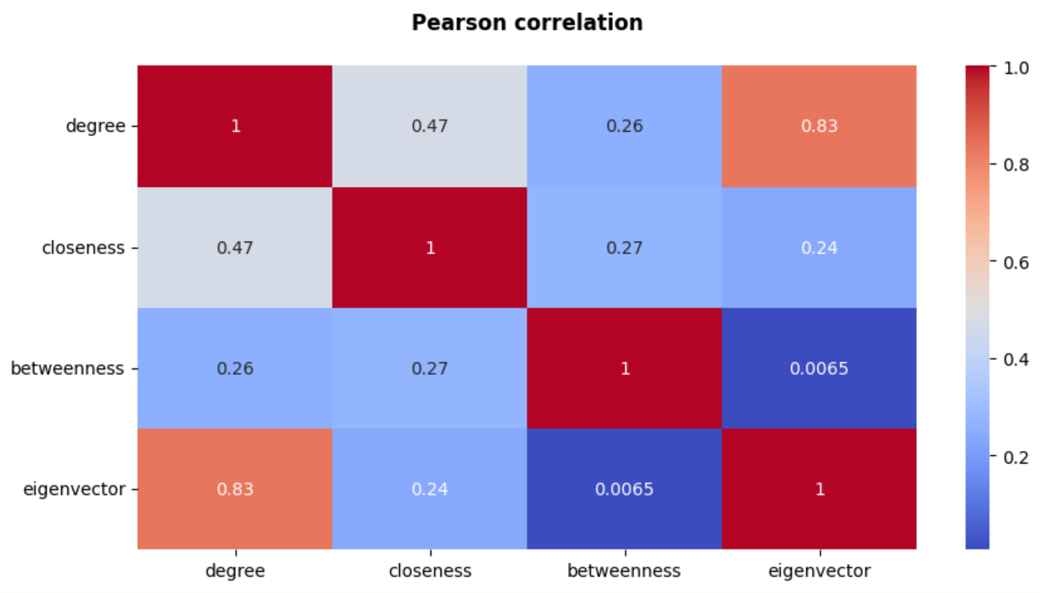
Betweenness

The Picture of Dorian Gray, 0.043
Pride and Prejudice, 0.042
Wuthering Heights, 0.027

Eigenvector

Anansi Boys, 0.134
Alice's Adventures in Wonderland, 0.132
Stories of Hope and Spirit, 0.132

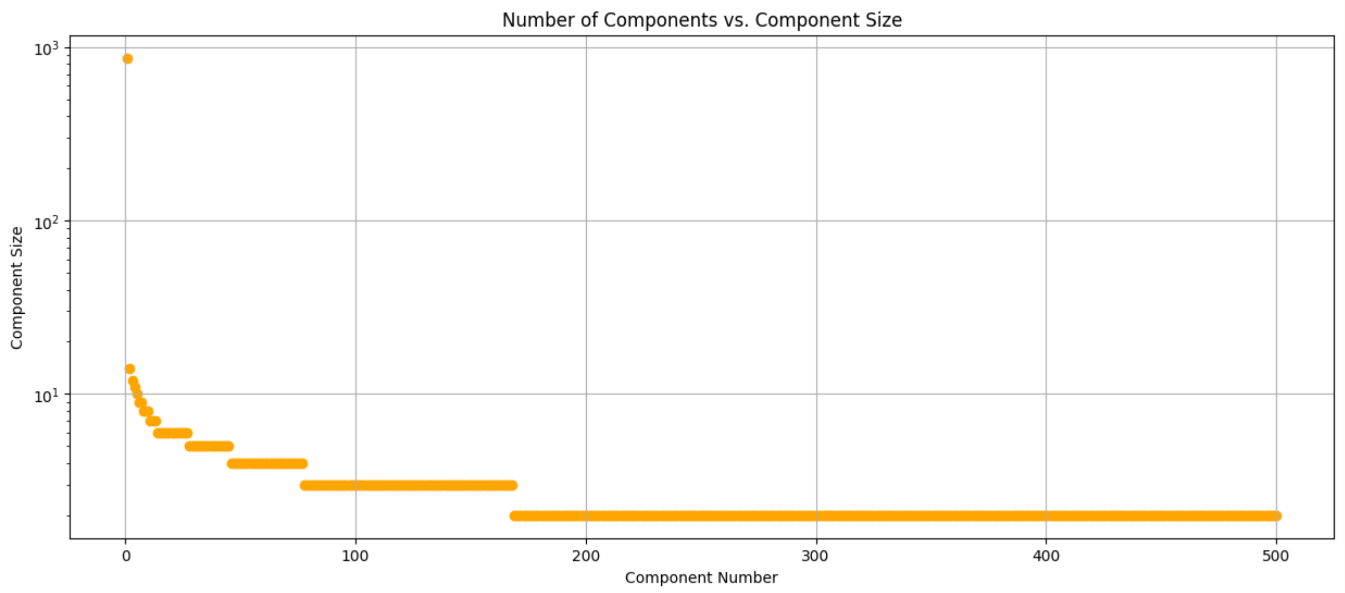
Pearson correlation matrix



Pearson correlation

	degree	closeness	betweenness	eigenvector
degree	1	0.47	0.26	0.83
closeness	0.47	1	0.27	0.24
betweenness	0.26	0.27	1	0.0065
eigenvector	0.83	0.24	0.0065	1

Connected components

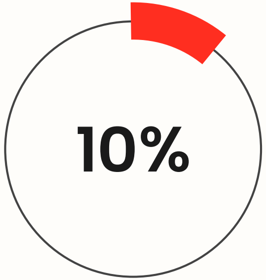


N° of components: 500

Larger components: 868

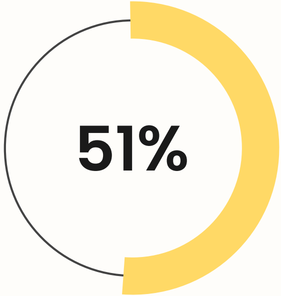
Smaller components: 2

Others network's metrics!



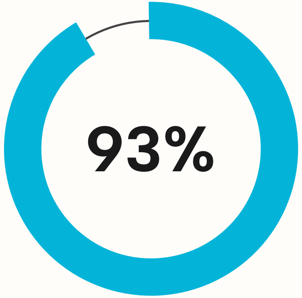
Bridges

10.2% of bridges, 603 nodes



Clustering coeff.

51.6% of global clustering coefficients



Assortativity

93% of assortativity, how much probable is the connection between two nodes with the same degree

03

Community Detection

Community detection using the Louvain algorithm

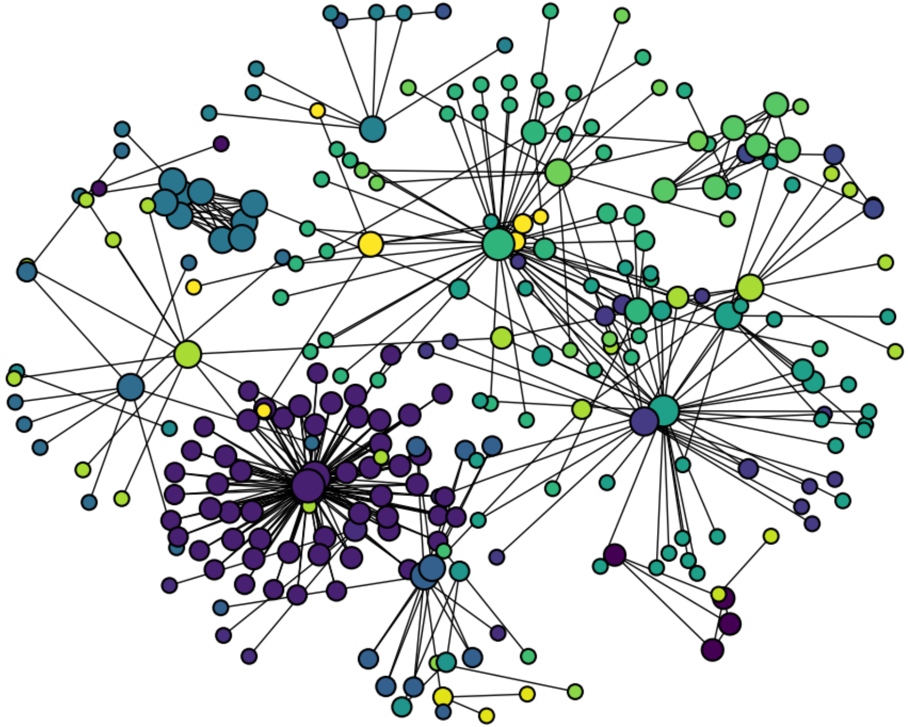
The communities

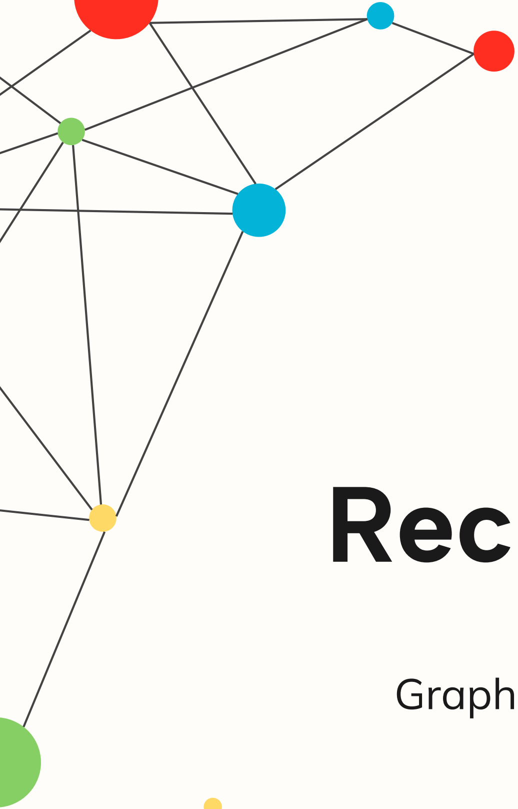
N° communities: **24**

Modularity of the partitions **0.79**

Top 3 communities:

1) 57 nodes (14%)
2) 41 nodes (10%)
3) 37 nodes (9%)

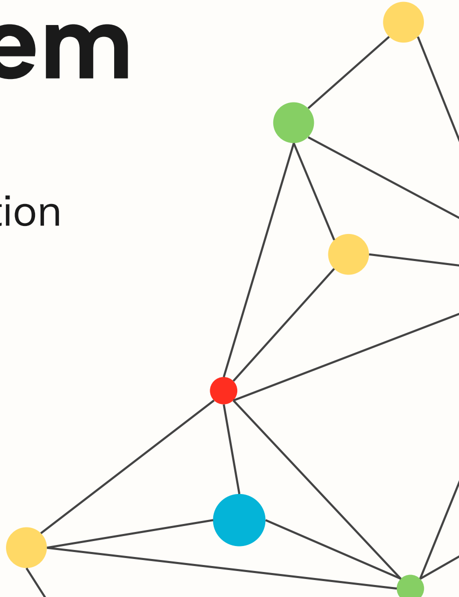




04

Recommendation System

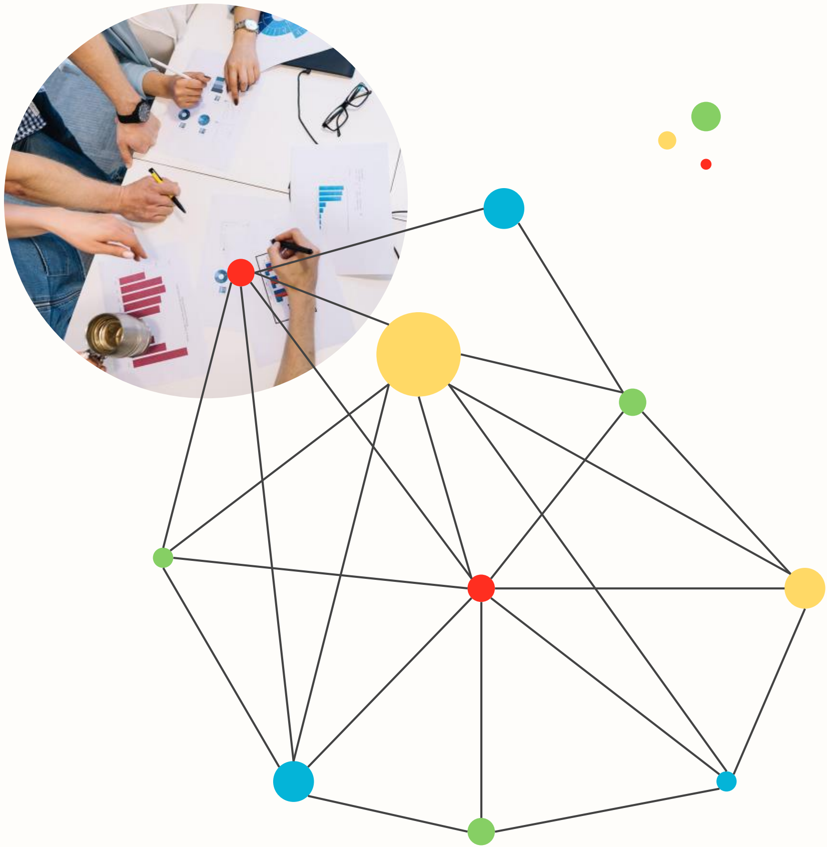
Graph embedding with Node2Vec for similarity prediction



RecSys with Node2Vec

Node2Vec is a graph embedding algorithm that, turning nodes of the graph into a vector strcuture, optimize the neighborhood of the nodes through a biased random walk. This allow to find similar nodes and recommend in this case new books to buy.

Choosing the parameters **p** and **q** has been optimized the similarity algorithm and the neighborhood exploration.



Some recommendations

After having fitted the model, the algorithm has been tested with some examples

1 **recommend_book_from(**'George Orwell 1984’)

- Cat's cradle (A Dell book)
- Little men : life at Plumfield with Jo's boys

2 **recommend_book_from**('The Picture of Dorian Gray (Classic Collection (Brilliance Audio))’)

- Hamlet (The Shakespeare Folios)
- The Berlin Stories: The Last of Mr. Norris

3 **recommend_book_from**('Pride and Prejudice’)

- Dragonwyck
- Emma (Penguin Readers, Level 4)

As we noticed, book similarity is very related to the category

Conclusions

In this case the Node2Vec algorithm is very precise since the fact that a customer bought and positively reviewed different books is a very good recommendation for others to buy them.

Genres and categories helps into identifying the most similar books, specially for those in the same clique with a high clustering coefficient.

Thanks for the attention!