Genre Dynamics and Network Patterns: Unveiling Relationships Between Books' and Authors' Networks and Discovering Popularity Trends on Goodreads (Midterm Progress Report)

Reza Abbaszadeh Darban*
York University
Toronto, Ontario, Canada
rezaab@yorku.ca

Hamed Taherkhani
York University
Toronto, Ontario, Canada
hamedth@yorku.ca

Mahdiye Khalajiolyaie
York University
Toronto, Ontario, Canada
mahdiye7@yorku.ca

Abdullah Tauqeer
York University
Toronto, Ontario, Canada
abdullatauqir@gmail.com

KEYWORDS

Data analytics, Recommendation system, Goodreads, Network analysis, User behaviour, Clustering, Genre Analysis

ACM Reference Format:
Reza Abbaszadeh Darban, Hamed Taherkhani, Mahdiye Khalajiolyaie, and Abdullah Tauqeer. 2024. Genre Dynamics and Network Patterns: Unveiling Relationships Between Books' and Authors' Networks and Discovering Popularity Trends on Goodreads (Midterm Progress Report). In . ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Data analysis plays a pivotal role across various domains, enabling informed decision-making and societal progress. It involves systematically exploring, transforming, and interpreting data to uncover meaningful patterns and trends. Graphs are a special type of data structure that consists of edges and nodes, with edges connecting nodes. One of the most relevant features of graphs representing real systems is community structure, or clustering, i.e. the organization of vertices in clusters, with many edges joining vertices of the same cluster and comparatively few edges joining vertices of different clusters.[1].

1.1 Motivation

Books play a pivotal role in society. There is an enormous amount of data on books available on the internet. This data is spread across many websites accessible to all users worldwide. Goodreads is among the most popular websites, hosting millions of books and users. This vast dataset presents new challenges and opportunities for both the market and users. There are numerous patterns to be discovered within this dataset. The literacy market can grow by enhancing user experience through discovering these patterns and employing them as a recommender system for its users. There exists a network of authors that can be discovered and used to identify similar authors, suggesting users follow recommended authors. Another network revolves around books, where related titles often fall within particular genres. Certain genres might experience a decline in popularity while others see a surge in interest. Understanding shifts in user preferences would benefit both authors and stakeholders in the literacy industry. Some other questions arise in the networks of book authors and books. For example, which authors and books are most influential in the literacy market? How do authors' popularity change over time?

Users can also review and post comments on books. By leveraging these comments and reviews, we can discern evolving trends in genre popularity and illuminate the influential roles of various authors. This analysis unlocks deeper insights into the reading habits and preferences of users, which are essential for understanding the dynamics of the literacy market.

Furthermore, this analysis is significant for offering valuable insights to publishers for crafting effective marketing strategies and to authors for understanding reader reception. In addition, the analysis holds immense potential in enhancing recommendation algorithms, thereby refining user experience on literary platforms.

1.2 Dataset

The "Goodreads Book Reviews" dataset on Kaggle is an expansive resource that provides a comprehensive look into the literary world as reflected on the Goodreads platform. It encompasses a wide range of data, including Books, Authors, Genres, Series, and reviews. All the data are in JSON format. The overall size of this dataset is more than 30GB of text files, which are separated into multiple files. The most important datasets are as follows:

- The author's dataset: This dataset includes fields such as *average_rating*, *name*, *author_id*, *rating_count*, etc. The field names are self-explanatory.

*all authors contributed equally to this research.

Conference'17, July 2017, Washington, DC, USA

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM
https://doi.org/10.1145/nnnnnnn.nnnnnnn

- Books dataset: This dataset has many features that will be part of our analysis, such as *text_review_count*, *publisher*, *authors*, *publication_year*, *language*, *description*, etc. There are more than 1.5 million books in the dataset.
- Book reviews: This consists of 15 million records of book reviews. Each review contains fields like *user_id*, *book_id*, *date_added*, *number_of_votes*, *number_of_comments*, *review_text*, *rating*. This dataset contains about 15 million reviews of about 2 million books, posted by almost 500 thousand users.
- Genre Dataset: This dataset has two fields: *book_id* and *genres*. Each book may have multiple genres.

We are going to base the network of our books on the user reviews. Thus, we will have an edge connecting two books if they have the same reviewer. We are also going to build another network of authors, where nodes connect if they co-author at least one paper. Our recommendation system will rely on user reviews, providing users with similar books ranked by book popularity. Hence, it will be a collaborative-based recommendation system. This dataset is too large to be analyzed in a graph network. So we are going to trim the dataset to be practical to be implemented on graphs. For example, we are going to drop nodes that have at least N number of edges. By doing so, we keep a core graph preserving the main characteristics of the graph with limited computational and memory constraints.

1.3 Research question

Research Question 1: What are the overall characteristics of our books and authors networks? Do they follow random graphs or small-world graphs?
Research Question 2: How do genres and books fluctuate in popularity over time? How do patterns in users' interests change over time?
Research Question 3: How effective is the recommendation system that is based on the reviews of our users?
Research Question 4: How similar are the books and authors networks when we cluster these two networks? Is there any relationship between the genres identified in the books network and the identified clusters in the authors network?
Research Question 5: Can we find important authors based on the PageRank algorithm?
Research Question 6: Can we find related books based on the reviews?

2 METHODOLOGY

The dataset enables us to construct different networks based on the different entities available including books, authors, and users. Three possible network architectures are as follows:

2.1 Books Network

In this particular network, books are considered to be nodes of the graph. Two books share an edge if there is a user that has submitted a rating or review for both of them. This network is expected to have high clustering and small diameter. Having constructed the network for specific periods of time, we can show how the users' interest have changed over time. It can be visualized that how trends are formed and how books gained or lost popularity over time.

2.2 Users Network

Users form another network in which two users are connected to each other if there is a book both of them have read or for which both of them have submitted a review or rating. It makes it possible to find users with similar interests and use the extracted knowledge to build a recommendation system that introduces books to users based on the interests of users with similar behavior. We are going to use a link prediction algorithm such as adamic/adar.

2.3 Authors Network

Authors form the nodes of this graph. A book that is co-authored by two or more authors creates a connection between the nodes of this network. Using this network, it can be examined whether books, which are proven similar from the users' interactions, are written by authors that are connected to each other in this network or whether these two approaches of finding similarity are not correlated.

In this work, different network analyses should be employed to find the specific characteristics of each network. One approach would be to calculate graphs such as degree distribution, path length, and clustering coefficient and compare the results with a null model to ascertain what aspects of the network stand out and provide meaningful information.

3 EVALUATION METRICS

To evaluate our recommendation system, we are going to split our dataset. We use 80% of our data for building our graph and another 20% to evaluate our system. We split them based on time so evaluation data comes later. We are going to report Mean Absolute Error(MAE), recall, and precision metrics for this part.

4 RESULTS

4.1 Statistical Analysis

We have used four different datasets in to build our graph. In this section, we provide some statistical analysis of these datasets. In the following table, we showed the important features of the book's dataset. The total number of books is 1512654.

Book's dataset				
Feature name	mean	max	Q1	Q3
Average Rating	3.85	5	3.62	4.14
Publication Year	2007	2018	2004	2011
Text Reviews Count	34	142645	2	13
Number of Pages	263	945077	148	344
Ratings count	505	4899965	7	860

In this table it is clear that the average number of ratings is a lot. 25 percentage of books have less than 8 reviews and 25 percentage of them have more than 860 reviews. So the number of reviews is not random and hence the generated graph based on the reviews will not be likely a random graph.

The next dataset is the genre's dataset. This is a simple json dataset in which each book_id corresponds to a unique book. Each book can have multiple genres. The following figure describes different genres in the dataset and number of books in the corresponding

genre. The fiction genre has the most number of corresponding books and the poetry has the lowest number.

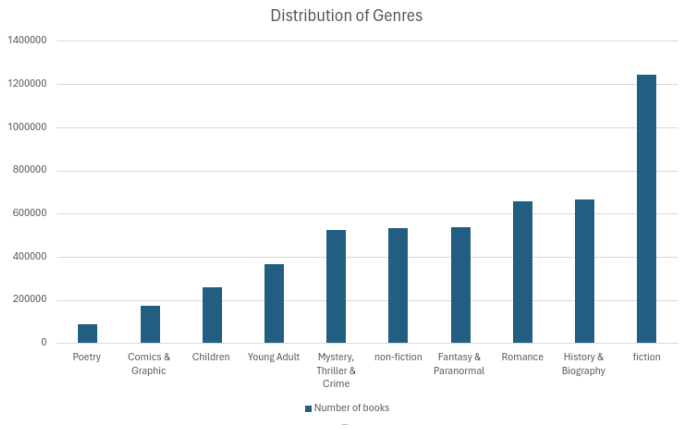

Figure 1: Distribution of Genres

Reviews dataset consists of almost 16 million reviews submitted from 2001 to 2017 by 465,000 users for 2 million books. The most active user of the platform had submitted 21,811 reviews. Fig 2 demonstrated how the Goodreads platform got more popular over the time.

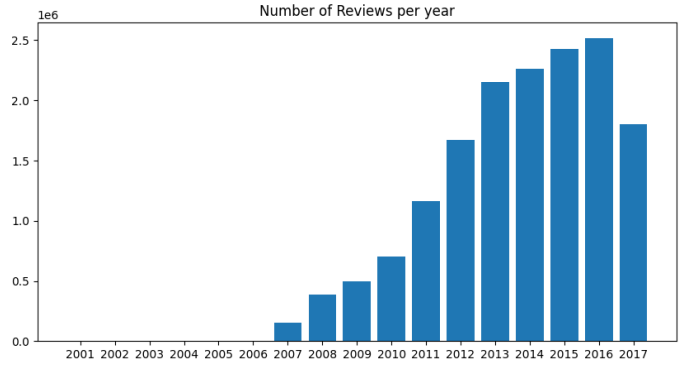

Figure 2: Year Distribution

Finally, the authors dataset consists of more than 800 thousand of authors. Their average rating is 3.84 and there are on average 1595 number of ratings per author in the goodreads dataset.

4.2 Book's network

In the book's network, the nodes are the books and two books are connected with each other if there is at least one user who has a review on both books. Because of the very large dataset, to construct our graph, we selected a subset of books and extracted the reviews on these subset of books. Initially, We chose 1000 books in each year from 1980 to 2018. The reason was that we wanted to pick books on different years so that they could be a valid representation of all books in the dataset. Totally we selected 37000 books and 254197 reviews which range from from 2001 to 2017. Building our books

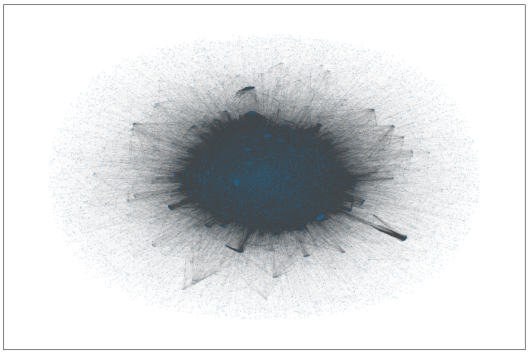

Figure 3: Book's graph visualisation

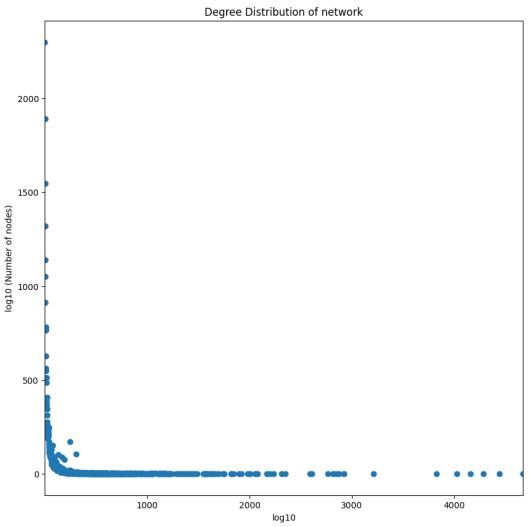

Figure 4: Distribution of edges on book's graph

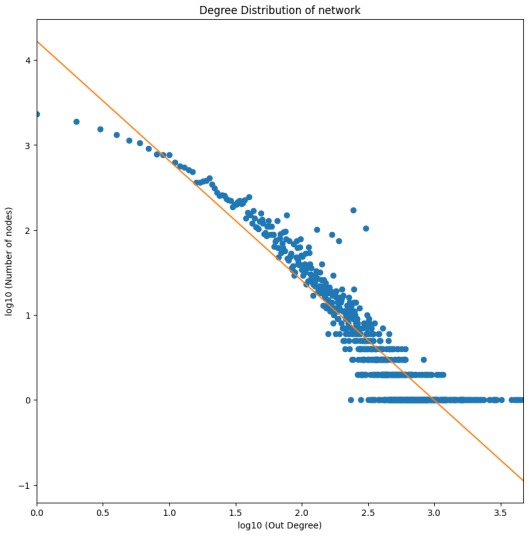

Figure 5: Log Distribution of edges on book's graph

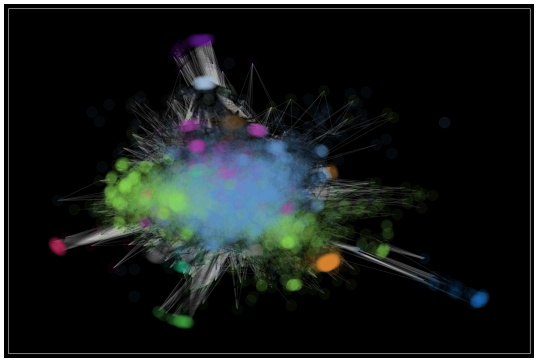

Figure 6: Community detection using greedy modularity

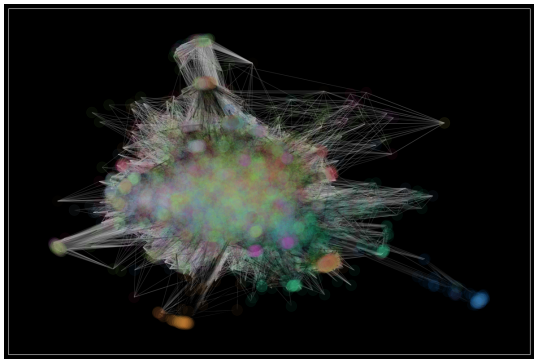

Figure 7: Books separated by different genres

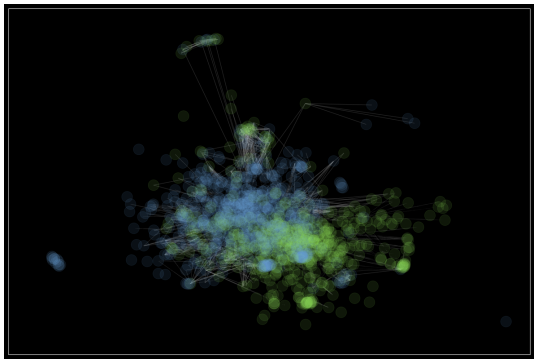

Figure 8: Thriller vs Romance

graph, we had a graph of 30778 nodes and 856080 edges. The density of this graph is 0.0018, the average degree of nodes is 27.8 and the average clustering coefficient is 0.70. In order to compare our graph to a random graph, we created an erdos_renyi_graph with the same number of nodes and edges of our book's graph(the probability of edge creation is 0.0018). The average clustering coefficient of this random graph is 0.0018 while this feature in the book graph was 0.7. This is a behaviour that we expect on small world graphs. The average shortest path in the books graph is not measurable because it is not a connected graph. While the average shortest path in the corresponding random graph is 2.9. Figure 3 shows the visualisation of the graph.

The degree distribution and the log degree distribution of our network is depicted in figure 4 and 5 respectively. The measured alpha for this distribution is 1.41.

Moving on, we want to do a community detection analysis on this graph and find out their relationship with different genres. We want to find out that whether different genres make different communities in the graph or not. First, we tried girvan newman algorithm. But since it takes too long and still does not end, we tried louvain, async_fluid, and greedy modularity and opted the greedy modularity at the end. For better representation of the communities we dropped every node with less than 20 degrees in the graph. The community detection using greedy modularity algorithm is depicted in figure 6. The question is whether this communities represent different genres of books or not. To answer this question we showed books of different genres with different colors in figure 7. Comparing figures 6 and 7 indicates that different genres are not visually recognizable in the network. But, if we build a network of books with two number of genres, they will be visually differentiable using greedy modularity algorithm as shown in figure 8. Choosing some pairs of genres like adult vs children also produces the same results, while choosing some other pairs, we will not be able to visually differentiate between two genres. This means that some genres have much overlaps with each other and some others don't have less overlaps so we can differentiate between them. This makes sense, because books of one genre may be of other related genres as well. On the other hand, for example, a thriller books is less likely to be romance as well. To measure the differences between the communities of different genres, we used conductance measure. This feature measures how good is a partition of two communities. The results are shown in the following table:

–	hi	fi	fa	my	po	ro	no	ch	yo	co
history	0.079	0.29	0.075	0.093	0.077	0.050	0.149	0.076	0.029	0.036
fiction	0.291	0.192	0.292	0.309	0.252	0.151	0.32	0.209	0.259	0.263
fantasy	0.075	0.29	0.096	0.090	0.062	0.137	0.085	0.077	0.117	0.108
mystery	0.093	0.30	0.090	0.118	0.053	0.076	0.104	0.057	0.069	0.062
poetry	0.077	0.25	0.062	0.053	0.126	0.011	0.113	0.079	0.024	0.030
romance	0.050	0.151	0.137	0.076	0.011	0.19	0.045	0.043	0.067	0.024
non-fiction	0.149	0.32	0.085	0.104	0.113	0.045	0.087	0.105	0.084	0.096
children	0.076	0.20	0.077	0.057	0.079	0.043	0.105	0.172	0.131	0.102
young-adult	0.029	0.25	0.117	0.069	0.024	0.067	0.084	0.131	0.092	0.052
comics	0.036	0.26	0.108	0.062	0.030	0.024	0.096	0.102	0.052	0.103

This conductance table indicates that for example history and fiction genres are more discernible than other genres. As you may notice, the fiction genre is the most discernible genre, meaning that fiction books are less likely to overlap with other genres.

In the next analysis phase of the graph, we want to find the most important nodes on the graph. For this, we tried pagerank, closeness centrality, degree centrality and betweenness centrality and ranked them based on the measurements. Interestingly pagerank, closeness and degree centrality produce almost same nodes for the fist 20 most important nodes. But for the betweenness centrality measure,

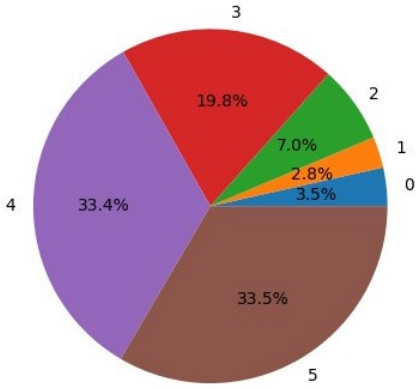

Figure 9: Distribution of Ratings

the result is much more different than others. The following table shows the most 5 important nodes based on pagerank algorithm.

- The Girl Who Played with Fire (Millennium, #2) - 2009
- Unbroken: A World War II Story of Survival, Resilience, and Redemption - 2010
- Shiver (The Wolves of Mercy Falls, #1) - 2009
- The Name of the Wind (The Kingkiller Chronicle, #1 - 2007
- Good Omens: The Nice and Accurate Prophecies of Agnes Nutter, Witch - 2006

4.3 Users' network

Users' network consists of users as nodes. An edge would connect two users in the network if there is a book for which these two users have submitted a rating over 3 stars. In total, there are almost 16 million reviews submitted by 465,000 users for 2,000,000 books.

4.3.1 Data Reduction.

So, the main challenge regarding network construction is the high volume of data. Some heuristics are employed to reduce the size of the network. Figure 9 represents the proportion of each rating value in the dataset while rating 0 means no rating is submitted. to make connection between users that share the same interest we only consider ratings of 4 and 5 stars to make links and ignore the lowers ones which accommodate one third of the data.

We have conducted our analysis on reviews before 2017, and separated the data associated to 2017 for our evaluation step, particularly for the evaluation of the recommendation system. It is also worth mentioning that the number of reviews which are submitted before 2007 are not significant, so they are not included in our analysis.

All books with less than 3 reviews were also removed, so as to create a strongly connected network without completely isolated islands in the network.

After all the reductions, there were still 8 million reviews remaining. Therefore, we decided to choose users who had at least one interaction in the last year of our analysis (2016), since it is more likely for them to be active in 2017 as well. Finally, 160,000 users nominated, and among them 5,000 users where selected randomly to create a network that fits in the memory as is computationally reasonable.

Users Network	
Number of Nodes	4513
Number of Edges	627333
Average Degree of Nodes	278
Diameter	5
Average Path Length	2.16
Clustering Coefficient	0.64

4.3.2 Network Properties.

The table shows that the constructed network of users is noticeably dense, due to high clustering coefficient and low diameter. So we expect a small-word model behavior from this network. Degree Distribution of the users network is shown in Figure 10 can verify this assumption.

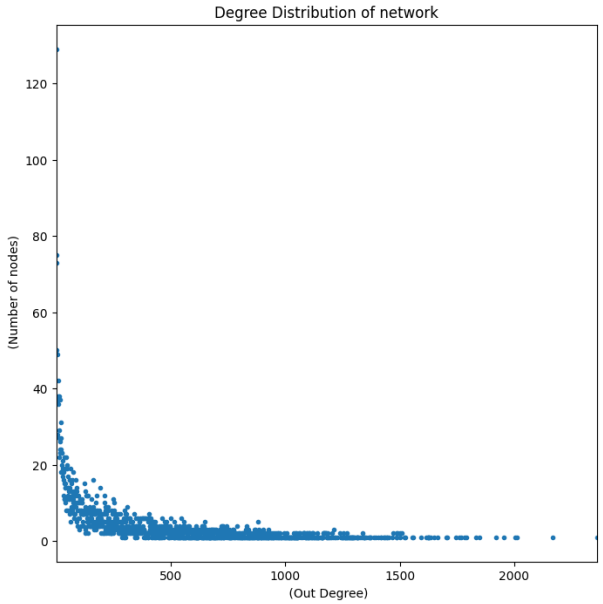

Figure 10: Users Network Degree Distribution

Furthermore, it can be seen in Figure 11 that the node degree distribution of the graph is an exponential function of k (the fraction of nodes that has degree k) and is a power-law. The estimated value for the Power-law degree exponent (α) = 1. This observation states that as the degree increases, the proportion of nodes with that degree decreases but in a very slow pace compared to famous graphs discussed in the lectures. As a result, this network has many nodes with high degree.

4.3.3 Community Structures.

To distinguish communities in users network, we used **Louvain communities** algorithm and tried to maximize the modularity as a metric of how well this network is partitioned into communities. The best modularity that we could have achieved was 0.2 which shows that although it is not considered a significant community

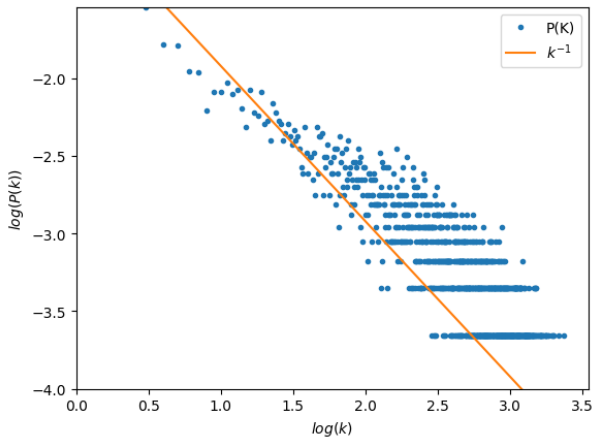

Figure 11: Users Network Power Law

structure, the number of edges within groups exceeds the expected number. Communities are of lengths 75, 1090, 2028, 261, 655, 291, and 113. Fig 12 depicts the seven identified communities. 336,119 edges form the communities (black edges), while the rest 291,214 edges (red edges) connect communities to each other.

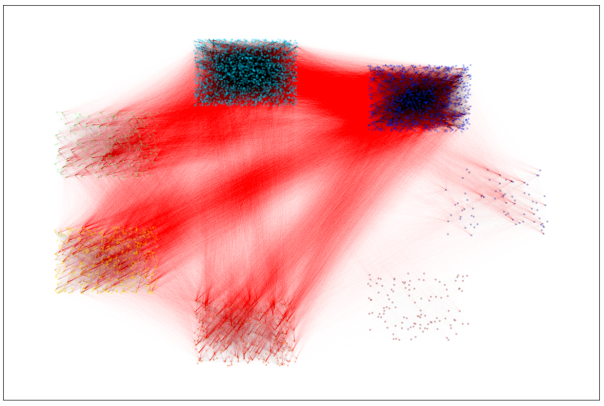

Figure 12: Users Clustering

4.3.4 Recommendation System.

This part is aimed to be done for the final report, however we have conducted some preliminary analysis using link prediction methods introduced in the class, namely jaccard coefficient, adamic adar, preferential attachment, common neighbors, simrank. Top 10,000 predictions of each algorithm is taken and is evaluated with the network containing reviews submitted in 2017. The performance of the mentioned algorithms was not satisfactory per se as shown in the following table. Therefore, we are trying to combine these algorithms together. A possible approach would be to feed a Machine Learning model with the results of this algorithms and find users with similar behaviors that are not connected at the moment in the network. Then, we can recommend their favorite books to each other.

Algorithm	Accuracy
jaccard coefficient	11.89%
adamic adar	11.91%
preferential attachment	10.47%
common neighbors	13.68%
simrank	6.54%

The superiority of common neighbors algorithms could be justified considering the fact that it gives higher score to edges within the same community compared to edges between different communities.

4.4 Author's network

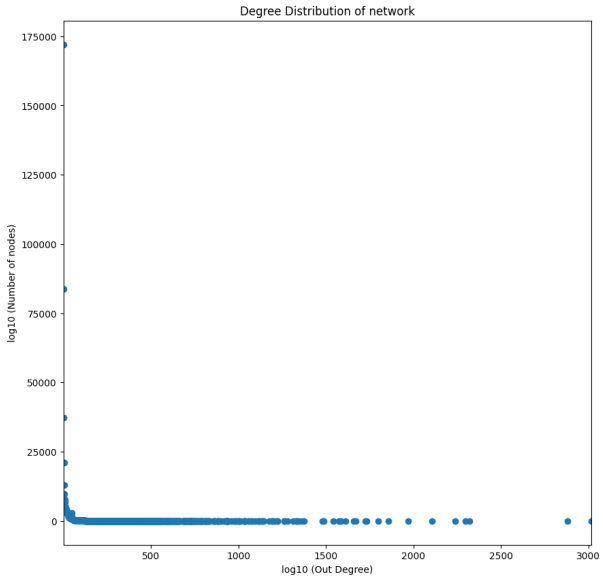

Figure 13: Degree distribution of authors graph

To build the network of authors, we used the authors as nodes and two authors connect if they have participated as co-authors in at least one book. This network has 470275 nodes and 2628626 edges with the average degree of 11.17. The average clustering coefficient of this graph is 0.46, indicating that the graph is not a random graph. The degree distribution of this graph is depicted in 13. We conducted pagerank algorithm on this graph to find the most important authors in the graph.

- William Shakespeare
- Edgar Allan Poe
- Charles Dickens
- Stephen King
- Neil Gaiman
- Arthur Conan Doyle
- Mark Twain
- Ray Bradbury
- Agatha Christie

- Fyodor Dostoyevsky

We are going to add the rest of our analysis on this graph like community detection in the final results.

REFERENCES

[1] Santo Fortunato. 2010. Community detection in graphs. *Physics Reports* 486, 3 (2010), 75–174. https://doi.org/10.1016/j.physrep.2009.11.002