Universiteit
Leiden
The Netherlands

Data Science & Artificial Intelligence

How much alike are we to one another?

Analyzed with online footprint

Seyidali Bulut

Supervisors:
Dr. A. Saxena
Dr. A. Zohrehvand

BACHELOR THESIS

Leiden Institute of Advanced Computer Science (LIACS)
www.liacs.leidenuniv.nl

28/08/2024

Abstract

Understanding co-interest networks can be highly valuable as they provide insights into human behaviour. By analysing how users behave and connect with each other through shared interests, we can gain a deeper understanding of the dynamics of online community formation, the reasons behind community formation and user engagement. Companies can use these insights to improve user engagement, strengthen community connections, and to provide better recommendations on various online platforms. In this thesis, I aim to obtain these insights by applying network analysis techniques to different co-interest network datasets acquired from online platforms. To be specific, this thesis will focus on detecting naturally forming communities in Twitch and Amazon book review datasets. Twitch is an online streaming platform where anyone can stream and earn money through viewer donations. We create the Twitch co-interest network where two streamers are connected if they share common viewers and the edge-weight denotes the total number of common viewers. To create Amazon book co-interest network, we use Amazon book review dataset to link books that share readers. After creating both of these networks, we analyse the macro and meso-scale characteristics of both datasets to better understand these networks. By examining meso-scale characteristics, we identify communities within a network and understand their evolution. My results show that both datasets exhibit a community based structure where communities are defined not only by language difference but also by differences in interests. The Amazon dataset network showed strong interconnectivity, indicating specialised areas of interest among readers. In the Twitch network, communities are influenced not only by the type of content but also by the spoken language, which leads to a more modular network structure than the Amazon network.

Contents

1 Introduction **4**
- 1.1 Co-interest network . . . 4
- 1.2 Research question . . . 5
- 1.3 Thesis overview . . . 5

2 Background **6**
- 2.1 Definitions . . . 6
- 2.2 Measures . . . 6
- 2.3 Theoretical Background . . . 7
- 2.4 Related Work . . . 8

3 Datasets **10**
- 3.1 Amazon Books Reviews . . . 10
- 3.2 Twitch Community Analysis . . . 11

4 Methodology **12**
- 4.1 Amazon Dataset . . . 12
  - 4.1.1 Data Simplification . . . 12
  - 4.1.2 Data Analysis and Visualisation . . . 13
- 4.2 Twitch Dataset . . . 13
  - 4.2.1 Data Cleaning . . . 14
  - 4.2.2 Data Visualisation . . . 14

5 Results **15**
- 5.1 Meso-Scale Characteristics . . . 16
  - 5.1.1 Amazon . . . 16
  - 5.1.2 Twitch . . . 20
- 5.2 Macro-scale Properties . . . 24
- 5.3 Discussion . . . 30
  - 5.3.1 Twitch Data . . . 30
  - 5.3.2 Amazon Data . . . 31
  - 5.3.3 Highlighted Differences in Network Structures . . . 32

6 Conclusion and Further Research **34**
- 6.1 Further Research . . . 35

References **37**

1 Introduction

Every 5 days, we create as much information as we did in 2010 and this is mostly due to increase in digitisation [1]. A hundred years ago, our footprint consisted mainly of physical products made by us or under our name. This could include our family, but also a simple product such as a home is also considered footprint. Furthermore, the knowledge that we left behind is also part of our footprint. However, back then, these footprints were similar for most people and were not substantial. The reach of footprint that someone left behind did not extend beyond their close network because it was much more challenging to preserve actions or knowledge. Most knowledge was shared orally and it was lost after a couple of generations.

Now, it is much easier to leave a larger information footprint and this is largely due to the rise of the internet. Every action you take on the Internet like the messages you send, the places you have visited, the apps you opened today and how long you have slept are saved. All of this information is stored on the Internet. Consequently, the term 'big data' has emerged. Big data refers to the structured and unstructured data that overwhelm businesses and organisations daily. This data can be analysed and the insights derived can be used to understand more about the users.

There are various ways to analyse big data and the method chosen can vary depending on the goals for the data. Some of these methods include predictive analysis, prescriptive analysis, diagnostic analysis and descriptive analysis. Each of these methods contains multiple approaches to extract information from the data. The method that will be used in this paper is network analysis [2] for understanding co-interest networks.

1.1 Co-interest network

Networks are a system of entities that are interconnected with each other that represent relationships of interaction between entities. Those entities can represent anything from people in a university to cells in a human body. In all networks, entities are represented as nodes and the interconnections between the nodes are called edges or links.

In a co-interest network, nodes indicate the entities of interest (e.g. words, concepts, items or people) based on the context under research. However, these nodes are not the main target group. The main target group that we want to analyse is the people who engage in an action with those nodes. In this network, an edge between two nodes indicates that some individuals in the target group share an interest in both of the entities (nodes). The strength of this connection tells us the size of the people from target group who are interested in both nodes and this can be further explained by the frequency of how often the co-interest occurs. This is also called the weight of the connection. Theoretical background of this thesis will be explained in more detail in section 2.3.

An example for such a network in e-commerce analysis is a network that shows how clients are connected to each other through a similarity in behaviour. In e-commerce analysis, we can look at how likely particular products are bought together by a client. This will help reveal which two products are more likely to be bought together, a cluster of products that form their own client

base and trends that occur in a shop. It can also be used in different fields, such as biology and text analysis. In this thesis we will use it to analyse social behavior to investigate how entities are connected based on common interest of individuals in an online environment.

The above mentioned cases of co-interest network usages are primarily used to help analyse complex relationships. This is done by using network analysis methods on complex datasets that are difficult to convert to graphical representation so that the identification of patterns can be detected and this can be used to better understand our data. This is why, co-interest network analysis can be very useful in handling and analysing large datasets. We analyze macro and meso-scale properties of these networks [3]. The results are quite helpful in uncovering relationships and patterns that traditional analysis methods do not reveal.

1.2 Research question

The following research question will be investigated in this thesis:
Which conclusions can we derive from these datasets about their network structures after analysing the macro and meso-scale characteristics.
We can break down this research question into sub-questions:

- RQ1: How the macro-scale properties, such as degree distribution, clustering coefficient distribution, network diameter, betweenness centrality distribution, closeness centrality distribution, harmonic closeness centrality distribution, eccentricity distribution, density, PageRank distribution, k-core distribution and clustering coefficient distribution look like?
- RQ2: Do these networks have any community structure?
- RQ3: What are the main characteristics of these communities that differentiate them from one another?

1.3 Thesis overview

The structure of the remainder of this thesis is organised as follows: Section 2 will explain the background of this thesis; Section 3 will explain how the datasets are acquired and analysed; Section 4 will explain which methodologies are used to analyse the datasets; Section 5 will show the results and explain what they mean; Section 6 will describe the conclusions and possible future research.

2 Background

This section will provide insight into the related work done on the topic of this thesis. Next to that, additional information about the definitions and terminology used in this thesis will be explained.

2.1 Definitions

To understand the analyses in this thesis, several key terms and concepts needs to be explained:

Undirected Graph: A graph where edges have no direction.This means that the connection between nodes has no direction. In both the Amazon Books Reviews and Twitch datasets, the graphs are undirected because the relationships (shared users or viewers) are mutual.

Weighted Graph: A type of graph in which the connections between the nodes have a weight assigned to them [4]. These weights can have different meanings, but most of the time they indicate the strength of the relationship between the two nodes. For example, a higher weight on an edge between two streamers indicates a higher number of shared viewers.

Community Detection: The process of identifying groups of nodes that are more densely connected to each other than to the rest of the network. This helps in understanding the underlying structure of the network, such as identifying groups of users with similar interests.

2.2 Measures

The network structures of complex networks can be measured using various network measures. The following measures are used in this thesis:

Network Properties:

- **Network Diameter:** The longest shortest path between any two nodes in the network. This measure gives an indication of the maximum distance between nodes in the network.
- **Density:** The proportion of actual connections between nodes in a network relative to the maximum possible number of connections. High density means that a network has almost all the connections that it possibly can have.
- **Clustering Coefficient Distribution:** This measure indicates how nodes within a network tend to cluster together. It shows the distribution of the frequency of the clustering coefficient values for all nodes. This provides insights into the local exclusivity of the network.

Centrality Measures: We analyze the distribution of different centrality measures [5] to better understand the network structure.

- **Degree Distribution:** Degree centrality measure shows the distribution of the frequency by the number of times a node is connected to other nodes [6]. This provides information on the overall connectivity and identifying the presence of highly connected nodes.
- **Betweenness Centrality Distribution**: This measure shows the frequency of betweenness centrality values across all nodes. This distribution indicates how often nodes act as bridges along the shortest paths in the network.
- **Closeness Centrality Distribution**: This measure shows the frequency of the centrality values of the proximity for all nodes. This indicates how close a node is to all other nodes in the network. We also analyze the closeness centrality pattern as done in [7].
- **Harmonic Closeness Centrality Distribution**: This is similar to closeness centrality but it is calculated using the harmonic mean of distances. This provides better results for disconnected networks.
- **Eccentricity Distribution**: This measure shows the frequency distribution of the eccentricity values. Eccentricity values represent the greatest distance from a node to any other node in the network.
- **PageRank Distribution:** This measure shows the frequency of PageRank values across all nodes. PageRank value for each node indicates the relative importance of the nodes based on their connections and the importance of their neighbours.
- **K-Core Distribution:** A k-core is a part of a network where each member is connected to at least k other nodes, highlighting tightly connected groups. The higher the k-core, the more central and influential the members are [8]. This helps identify the core of a network and the most connected participants.

These measures provide a deep understanding of both the global and local properties of the network, which is critical for analysing the social networks in this thesis.

2.3 Theoretical Background

In this thesis, we will analyse multiple databases in which users can be connected to each other through their history of consuming different products. The most critical information is that which we can connect back to the user such as book purchase history or favourite films. Consider a database that contains information on the favourite film genres of users.

For example, person A has watched films the Batman and Joker while person B is more into movies and have watches the Batman, Joker and Oppenheimer. We can connect those movies to each other on a network and give weight to the edges because they share viewers with each other. The Batman and Joker films will have a connection of weight 2 with each other because they share two viewers that have watched both films. Both of these films will also have a connection to Oppenheimer but with a weight of 1 because they share only one viewer that has watched both films.

We can thus create a network of nodes and connections called a co-interest network. In graph theory terms, we represent films as nodes $\{b, j, o\}$ and amount of shared viewers with each other as edges.

These edges will have weights and their values will vary depending on the shared viewers.

This social network can be depicted as a graph $G = (V, E)$, where G represents the graph, V represents the nodes (films) and E represents the edges (shared viewers between films). An edge is denoted as $\{b, j\}$ for $b, j \in V$, representing the edge between nodes b and j. Edges can be directed/undirected and weighted/unweighted [9]. In this work, we will primarily focus on undirected weighted graphs. An example of an undirected and weighted graph of the above mentioned example is shown in Figure 1.

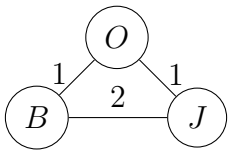


Figure 1: Simple undirected and weighted graph

2.4 Related Work

During a network analysis the feature and characteristics of a network are gathered in detail to understand the relationships between the entities within the network. This can be done in different ways but in a social network analysis the relationships between entities becomes the first priority, rather than focusing on properties of the entities. This is why network analysis looks mainly at the broader picture to understand how the network functions as a whole [10]. Individual characteristics of the entities are used in second priority to fully understand social phenomena.

The attention on complex systems is only increasing as Stephen Hawking has pointed in his quote "I think the next century will be the century of complexity" [11]. Behind complex systems there is also a complex network that encodes the interactions between the components of the system [12]. By analysing the underlying network of a complex system we can understand the structure and function of those systems. This can be done by identifying the patterns and trends in the network and how these influence the behaviour of the system as a whole. Complex network analysis has been used to understand different complex systems, ranging over different types of social systems [13, 14], financial systems [15], criminal networks [16], chemical systems [17], biological systems [18], and so on.

Different insights from network analysis can be used together to derive a conclusion and attach meaning to the network. However, the conclusion is only an interpretation of the insights obtained and different conclusions can be drawn from a single network by different data scientists. For example, one of the first works on network theory by Moreno [19] would have different results if the data were analysed with the current algorithms. This might also be the case with the algorithms that we use and one scientists interpretation might be different from another scientists.

The insights are used to understand the network and detect key influencers whose position in the network is important for the structure [20, 21]. Furthermore, identifying how information is spread through the network is another important insight [22] and in the spread of information, bottlenecks

and vulnerabilities can be detected. Lastly, analysing the structure and understanding the patterns of connectivity will show the communities within a network which are difficult the notice without analysing [23].

The study of network analysis has seen significant advancements in recent years. This is because there are more and more large datasets available and the need to efficiently understanding and visualisation also increases. There are different community detection algorithms that can be used in network analysis to detect communities within a network. These communities can be defined as densely connected cluster of nodes than other groups. There are different community detection algorithms that can be divided into different groups that differ in their methodologies and characteristics.

The Girvan-Newman algorithm [24] is a hierarchical and divisive method that identifies communities by iteratively removing edges with the highest betweennees centrality (2.2). However, this method requires a lot of computational power and this is not advantageous for large datasets that are analysed in this thesis. This is why I used the modularity-based algorithm called Louvain [25] method instead. A modularity based algorithm aim to maximise the modularity measure. This modularity measure can represent the strength of the division of the network by comparing the density of edges inside a community to the density expected if the edges were distributed randomly. Because of its efficiency, scalability, and being able to produce high quality results, I have chosen to use modularity algorithm to detect communities. The used modularity method is developed on a Belgian mobile phone network of 2 million customers and with 118 million nodes.

3 Datasets

This chapter will dive into the two datasets that are used in this thesis and explain what the contents of these datasets are.

3.1 Amazon Books Reviews

In this paper, we will analyse multiple datasets. The first dataset is the Amazon Books Reviews [26], which contains book review information from May 1996 to July 2014. It includes data on 3 million users and 142.8 million reviews.

After the data cleaning and simplification, this dataset contained 20,523 nodes and 10,111,177 edges. The simplification process will be explained further in the methodology (4.1.1) section.

The Amazon Books Reviews dataset contains the following features:

- id: The ID of the book
- Title: Book title
- Price: The price of the book
- User_id: ID of the user who rates the book
- profileName: Name of the user who rates the book
- review/helpfulness: Helpfulness rating of the review
- review/score: Rating from 0 to 5 for the book
- review/time: Time of the given review
- review/summary: Summary of the text review
- review/text: Full text of the review

Not every feature was necessary to build the desired network visualisation. The following features were selected:

- id: Used to identify every unique node.
- Title: Used to recognize different nodes during the analysis.
- User_id: Used to recognize the users and detect if the same user has read multiple books.

These selected features together provided a comprehensive dataset for creating a co-interest network analysis based on users book reviews.

3.2 Twitch Community Analysis

Twitch is one of the most popular streaming websites, where anyone can stream themselves engaging in different activities. While Twitch started as a platform mainly for gaming, it has evolved over time to include other types of streams such as 'Just Chatting' and real-life vlogging.

Twitch is different than most other streaming services because it makes data about its streamers and viewers available for free. This data can be accessed through the Twitch API [27]. Using this API, the Twitch dataset was created. The dataset contains multiple snapshots of the top 100 streamers and their viewers. In total, the dataset includes 2,200 streamers and 82,275 edges.

In the network created using this dataset, nodes represent streamers and edges between these nodes represent shared viewers between the streamers. A thicker edge between two nodes indicates more overlap.

4 Methodology

In this chapter, we will begin with the Amazon dataset, explaining how it was collected, simplified, and analysed, along with its visualisation. Afterwards, we will proceed with the Twitch dataset, explaining how it was collected, cleaned, analysed, and visualised.

4.1 Amazon Dataset

For the following steps of data cleaning, I used the Python Panda library. The methodology used for this dataset changed during the analysis period because, while working with this dataset, its deficiencies became clear. First, some of the books contained more than one ID because the titles were different from each other. For example, *The Lord of the Rings* trilogy contained more than one ID because titles written in small or capital letters caused them to be seen as new books which resulted in the same books automatically receiving different book IDs. To solve this, I started by changing the title of every book to all lowercase letters. Furthermore, titles containing the symbols “”, !, or ? were also changed to letters only to prevent double entries of same books.

After solving this problem, I counted how many times a single book ID was read and created a new column where the amount of read values was saved as ‘Count.’ This made it possible to create a Books.csv file with Book ID, Title and Count columns. However, there were still books with the same title but different book IDs. To solve this problem, the books with the same title were grouped together and they all received the same book ID. To prevent any data loss with this merge function, the count values of the book IDs that were going to be merged were summed up into the one selected book ID that would be used for all the same titles. The created Books.csv will later be used as a dictionary for all books.

Books.csv alone was not enough to create the necessary files for the analysis. This is why a second dataframe with the columns Book ID and User ID was needed. This file contained every review. However, this dataframe still contained book IDs that we had changed in the previous step. Therefore, the book IDs in the new dataframe were also changed according to Books.csv. After this step, I created the necessary two files, Books.csv and Books_rating_filtered.csv to finish the cleaning process.

4.1.1 Data Simplification

Our goal after analysing the dataset is to see the relationship between the books that have been read by the same users. This only requires the inclusion of users who have read more than one book for data analysis. Therefore, I removed every row that contained a user who had only entered one review. This reduced the number of rows by 700,000, but the dataframe was still too large to begin the analysis.

Another important consideration is that this dataframe contained too many books. This meant that books with no popularity were also included in the dataframe and these books are not relevant for the analysis. Including these books causes the network to be very scattered because they have proportionally a very low degree and are located at the edges of the network. Therefore, I decided

to add the criterion that books need to have at least 100 reviews on Amazon to earn the status of 'popular' book and be included in the dataframe. This reduced the number of nodes to 20,523. This number of nodes was manageable for analysis and visualisation, so I decided to create the edges between the nodes.

At this point, I decided to create the network using the NetworkX library in Python. Using this library, I grouped each user with all the books they had read. After this, I connected the books to each other to create the necessary edges. Each connection increased the weight of the edge by one. After completion of the graph, the number of edges was 12,232,231. This number of edges was too large for the available computational power, so I decided to keep only the proportionally strong edges. I implemented this by keeping only the edges with a weight value higher than 10. This means that there is a significant connection between two books and the connection is not random if at least 10 users have read both books. This step has lowered the amount of edges to 10,111,177.

Lastly, I updated the books.csv by removing the book_id's that no longer had connection to other books because any connection of weight lower than 10 is now removed. At this point, I had all the necessary files to start analysing with Gephi 0.10.1 [28].

4.1.2 Data Analysis and Visualisation

Gephi is an open-source, interactive network exploration and manipulation software package. It is essentially a Photoshop for graph data, where one can explore and unleash the hidden complexity of relationships within large amounts of data. In addition, users can work with millions of nodes and edges, do force-directed layout and uncover patterns in the arrangement of connections.

Using Gephi, I imported the data and started creating communities with the modularity report function which used Louvian alghorithm to detect communities. This way, the nodes were divided into different colours, with each colour representing a different community. I also changed the appearance of the network using the ForceAtlas2 and Fruchterman-Reingold algorithms to improve the visualisation of the network.

4.2 Twitch Dataset

The Twitch dataset is created mainly using the Twitch API. The first step with this API is to request the names of the top 100 streamers at a specific moment. Note that this is the current top 100 streamers and as expected, streamers who are not streaming at that specific moment will not be added to the dataframe

The second step is to get the list of viewers for the top 100 streamers. The two different datasets will then be combined into a large dictionary where each streamer's name is connected with the names of their viewers at that moment. As expected, with the first run, none of the viewers will be listed in the dictionaries of two different streams because most viewers watch a single stream at a time. To be able to see the common viewers between the streamers, the code should be run multiple times at different times to catch the viewers watching different streams each time.

The obtained viewers and streamers are added to an existing TwitchData.csv file. With each run, the dataset grows larger and the current dataset has been created by executing the code over 5 different days.

4.2.1 Data Cleaning

After collecting the large TwitchData.csv file, the dataset is read as a dictionary without any NaN values. Then, a new dictionary is created where each key is the name of a streamer and each value is a list of other streamers with whom the key streamer shares viewers, along with the number of shared viewers. However, this file still needs to be reformed to be used by Gephi. Therefore, a new dataframe with the columns 'Source,' 'Target,' and 'Weight' is created and these values are imported from the earlier created dictionary. Here, 'Weight' represents the number of shared viewers between the streamers. This file serves as the edge list. To create the node list, we run a final function that generates a node list with the columns 'ID,' 'Label,' and 'Count.' In this list, the streamers are represented by 'ID' and 'Label,' and 'Count' represents the unique viewers they have

4.2.2 Data Visualisation

As it is mentioned with the Amazon dataset, I used gephi to visualisaze and analyse the dataset.

5 Results

The results chapter is divided into two sections. In section 5.1, we will go through the meso-scale characteristics of both datasets. In this section, we will look at the visualisations of both networks, which will show different detected communities, each with a different colour than neighbouring communities. We will analyse these to understand the main characteristics of these communities that differentiate them from each other. This way, we will find an answer to our second and third research questions.

The section 5.2 in this chapter will focus on the macroscopic properties of both datasets. These properties will also be plotted side by side to make it easier to compare both graphs. In this section, we will answer the first research question.

5.1 Meso-Scale Characteristics

5.1.1 Amazon

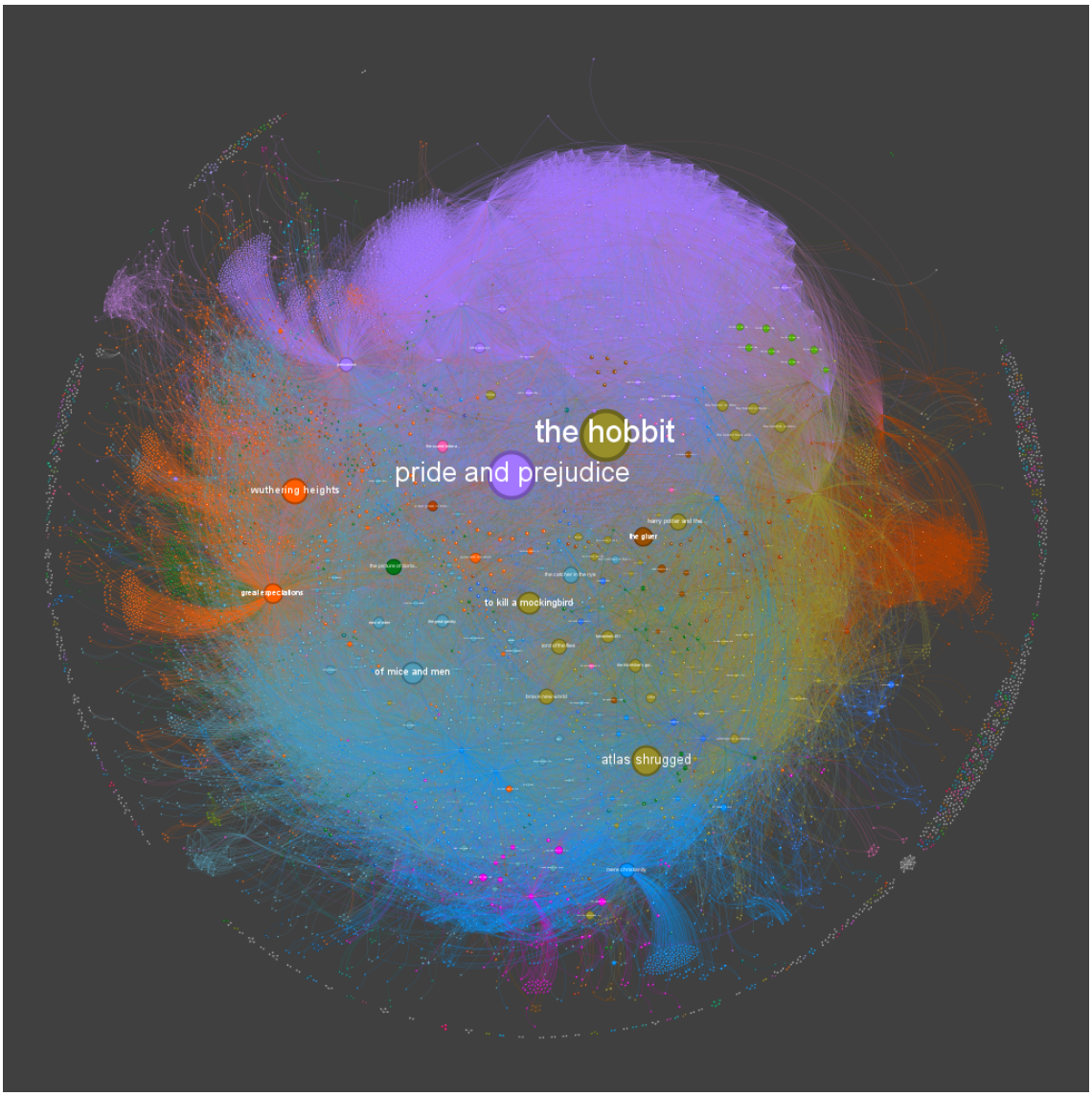


Figure 2: Amazon Dataset Visualisation

As we can see in Figure 2 the Amazon network is divided into communities that can be easily detected with the Louvain algorithm [25]. In this visualisation, the size of the nodes is determined by the number of times a book is read. The more popular a book is, the bigger the size of its node. We see this clearly with the books like *The Hobbit* and *Pride and Prejudice* which have sold more than ten million.

In this visualisation, we see that not every big node is located near the community that shares its colour. For example, *The Hobbit* is positioned in the middle and is surrounded by nodes of other colours. This is also the case with other larger nodes. This shows that these books are so popular

that they become outliers in their own communities. We can see this very clearly with *The Hobbit* because it is the book everyone thinks of when someone mentions fiction books. These books also appeal to readers of other genres, making them key recommendations, which has led to these books becoming trendsetters.

Another reason these communities are highly intertwined is that most of the books belong to more than one genre. This means that while some books can be defined by their primary genre, they also get associated with other books that share their secondary genres. This is clearly the case in Figure 3. Here, we see that the books in brown, such as *The Giver* and *Alice's Adventures in Wonderland* are in the same community because their defining genre is children's literature. Books like Harry Potter can also be defined as children's literature, but their main genre is fiction. This is why the Harry Potter books have a dark yellow (olive tone) color and not brown. Although they have the colour of a different community, they are located near books like *Alice's Adventures in Wonderland* because they share many common readers.

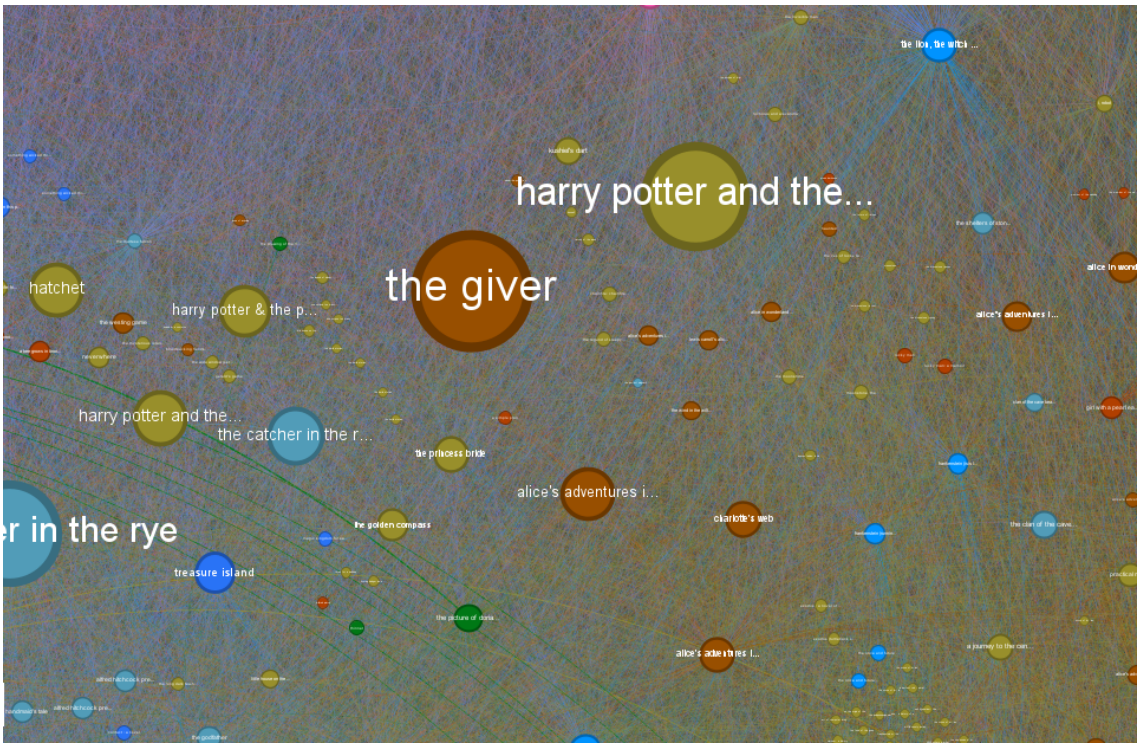


Figure 3: Books read by children located together

Lastly, the frequency with which a book is read is not evenly distributed across all genres. Some books within a specific genre become hits and are read proportionally more than others in the same genre. This causes them to act as outliers and be located more centrally in their community and serve as bridges from their community to others, as seen in Figure 4. *Wuthering Heights* and *Great Expectations* are examples of such bridges. They both belong to the orange community classified as Gothic Fiction, but they are positioned more towards the centre of the map compared to their community members because their popularity connects them to other books. This is also the case

with *Mere Christianity*, which forms a bridge to other blue coloured religious books, as seen in Figure 5.

Figure 4: Books that form a bridge

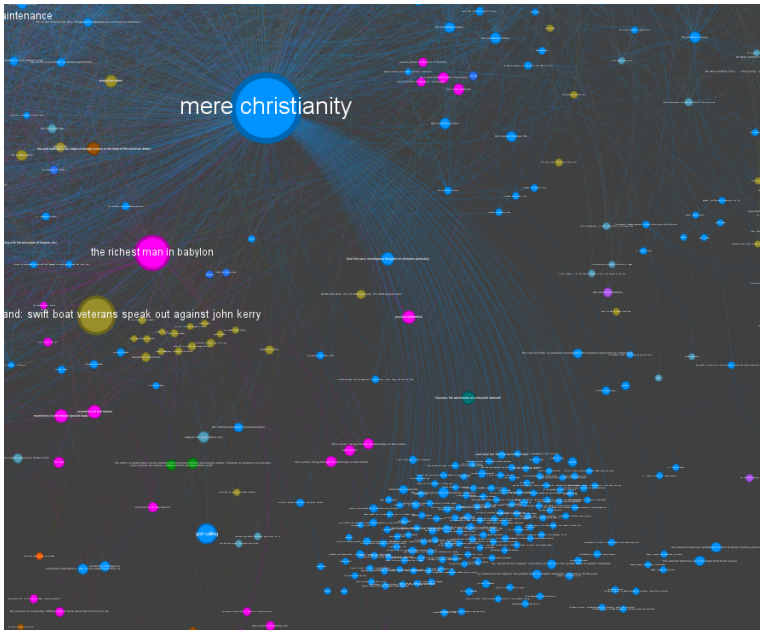


Figure 5: Mere Christianity forming bridge for other religious books

Book Genres

First of all, it is important to note that not every colour on the Amazon network represents a genre of a book. Some of the books are sequels, which makes them proportionally more connected to their sequels compared to other books. This leads the Louvain algorithm [25] to consider these books as their own community. Additionally, different books with different titles but the same content are

also highly connected with each other and can be considered their own community despite being more or less the same book. We can clearly see this in Figure 6. Here, different versions of *The Jefferson Bible* are connected with each other and form their own community. Furthermore, we see that the book *Lust For Life* is recorded three different times, each with a different title, which is why these books form their own community while being the same book. This phenomenon can also be observed with the green coloured Lord of The Rings (LOTR) books in Figure 7. The reason these books share readers with each other is that they are sold on the same page on Amazon.com as different versions of the same product, each with a different product id. On such a page, if a reader leaves a comment on one version of the book, it gets registered multiple times under every book ID listed on that page. This means that those books actually do not share readers but share comments which is registered multiple times.

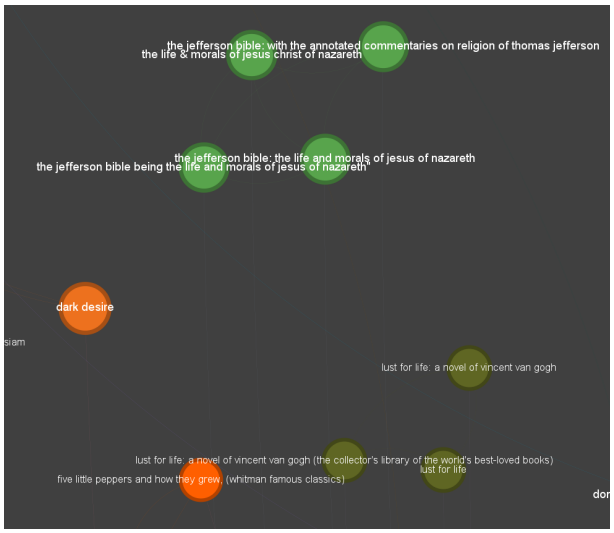


Figure 6: Same books with different titles forming their own community

Figure 7: LOTR books having different titles

However, this is not the case with most of the communities. Most of the colours represent a specific genre and an overview of the genres can be found in Table 1.

Color	Community Name
	Love Novel
	Fiction
	Gothic Fiction
	Tragedy
	Child Literature
	Melodrama
	Christian Books
	Personal Finance
	The LOTR Books
	Drama

Table 1: Overview of communities in Amazon

5.1.2 Twitch

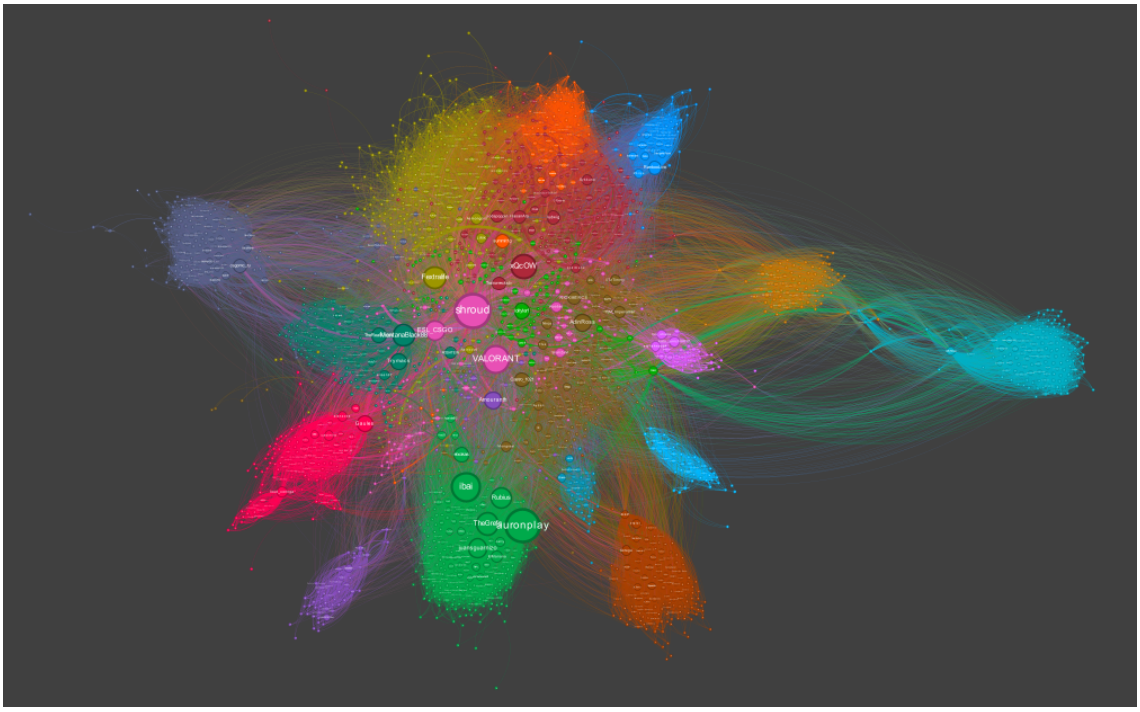


Figure 8: Twitch Network Visualisation

As we can see in Figure 8, the Twitch network is divided into communities that can be easily detected with the Louvain method [25]. When we zoom into one of the communities, as shown in the figures below, we see that this community is separated from the others because of its language barrier. The community in Figure 9 consists of streamers who stream in Turkish. In a country where only 17 percent of the population speaks English, the community focuses on Turkish-speaking streamers. This is also the case with the Spanish-speaking community, as seen in Figure 10 and with Korean streamers, as seen in Figure 11. In these countries, the effect of language is greater

than other differences and this is what separates them from the general community.

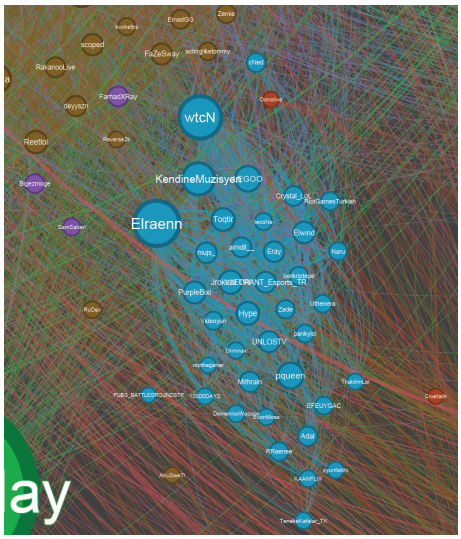


Figure 9: Turkish Streamers

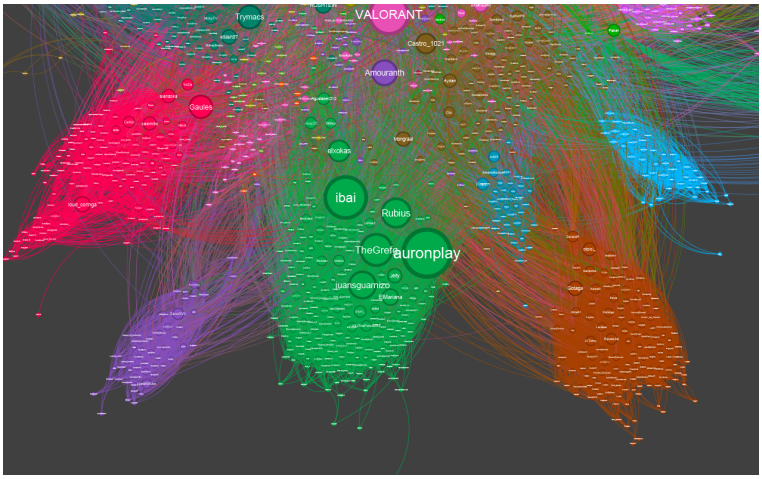


Figure 10: Spanish Streamers in Green

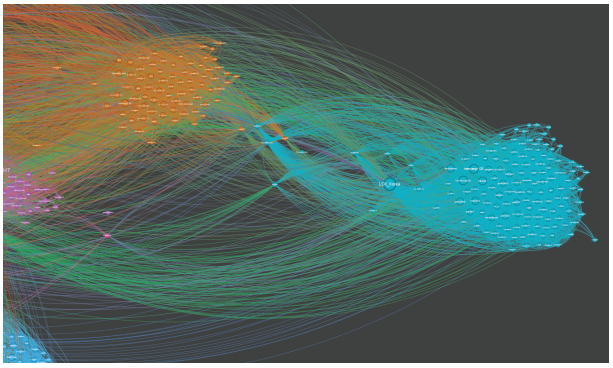


Figure 11: Korean Streamers in Light Blue

However, this is not the case in every community. When we look at the other detected communities, we see that the streamers within consist mainly of English-speaking people, as shown in Figure 12. This shows that those streamers are not together in the same community because they speak the same language, but are divided in different communities due to the content of their streams.

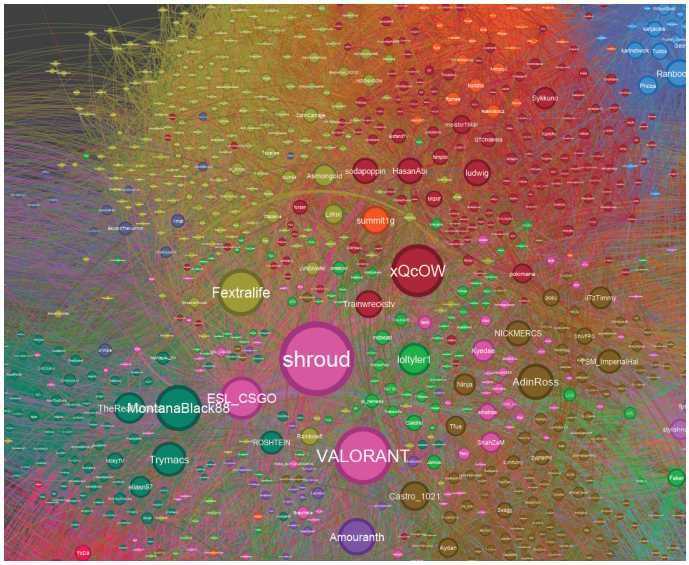


Figure 12: Mix of english speaking streamers in different communities

Twitch stream types

There are different types of streamers on Twitch and a brief explanation of the types of stream will help to understand the different communities.

- Variety: Streamer who streams different kind of games/streamer types. There is almost no consistency in kind of stream of games the streamer plays.
- First Person Shooter: Shooter games that are from first person point of view.
- In Real Life: Streamers who stream while being outside in real life.

Furthermore, there are streamers that stream one game only the whole time and this factor becomes the defining factor that differs them from the other communities. The viewers of those streamers watch only streamers of their favourite game and not other streamers, which causes them to create their own communities. Big enough games that create their own communities are for example CS:GO, FIFA, League of Legends, Dota2, Minecraft, Apex Legends, GTA V and Rocket League.

An overview of the different communities in the Twitch dataset can be seen in the table 2.

Color	Community Name
	First Person Shooter
	Variety
	CS:GO
	GTA V
	Turkish
	Spanish
	Korean
	German
	Third Person Games
	Portugese
	Italian
	French
	Apex Legends
	League of Legends
	Minecraft
	Japaneese

Table 2: Overview of communities in Twitch

These visual representations of the networks show us that these communities have a community structure. Some communities are more intertwined with each other, as seen in Figure 3 with the children's literature and fiction communities, while some communities are more apart from each other and more visible, as shown in Figure 11 with the Korean-speaking community lying further from other communities with different languages. These observations answer our second research question posed in this thesis and show that both networks contain a community structure.

The detected communities can be analysed to explain why certain communities books/streamers within them share specific characteristics that attract people to watch different streamers or read different books in that specific community. This similarity between books/streamers within a community enables them to share users with each other.

However, some books/streamers in those communities might have more than one characteristics with multiple communities. In such cases, the streamer/book is assigned to the community with which it shares the most similarity. We see this in Figure 3, as explained in section 5.1.1. The overview of the communities and their defining characteristics is shown in Table 1 and Table 2.

These tables answer our third research question posed in this thesis and show that some main characteristics of each community differentiate it from another community.

5.2 Macro-scale Properties

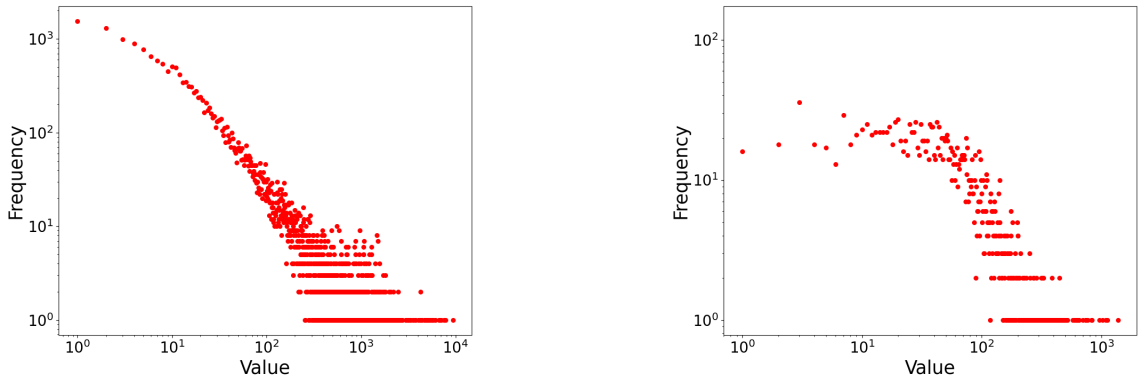
Table 3 shows an overview of macro-scale properties that will be analysed in this section.

	Amazon	Twitch
Nodes	20,523	2,200
Edges	10,111,177	82,275
Diameter	8	6
Density	0.005	0.034
Avg. Path Length	2.4	2.3
Avg. Degree	98.541	74.795
Avg. Clustering Coefficient	0.750	0.760
Total Triangles	73,211,661	2,114,518

Table 3: Overview of macro-scale properties

Degree Distribution
The Amazon graph shows that most of the books have a relatively low degree, indicating that most books are connected to a few other books through shared reviewers. The distribution has a long tail, with a few books having a very high degree, up to 9,000 connections. The average degree of the network is 98.541, suggesting that on average each book is connected to about 99 other books.

The Twitch graph of the Degree Distribution illustrates that most nodes have a low degree. As the degree increases, the frequency of the nodes decreases. The distribution shows that, while some nodes have degrees as high as 1,000, these occurrences are rare. The average degree of the nodes in this network is 74.795, indicating that on average, each streamer is connected to about 75 other nodes.



(a) Amazon Dataset (b) Twitch Dataset

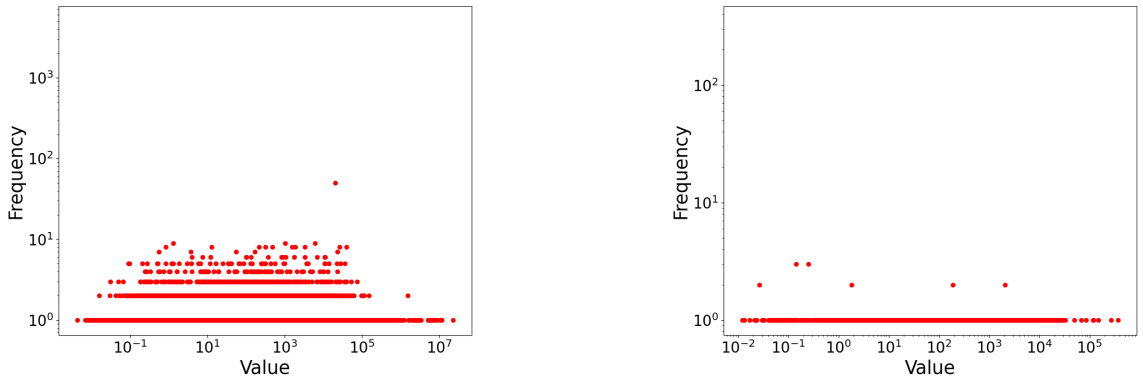
Figure 13: Degree Distributions

Network Diameter
The Amazon Books Reviews network has a diameter of 8. The average path length is 2.4, suggesting that, on average, any two books are separated by approximately 2 to 3 steps.

The Twitch Dataset has a diameter of 6. This means that the maximum distance between the most distant nodes in the Twitch network is 6 steps. Additionally, the average path length is 2.3.

Betweenness Centrality Distribution
The Betweenness Centrality Distribution for the Amazon Books Reviews dataset shows that the majority of nodes have average betweenness centrality values, indicating that most books act as bridges along the shortest paths between other books.

The Twitch graph shows that most nodes have low betweenness centrality values. There are a few nodes with high values, up to 350,000, but these are rare.

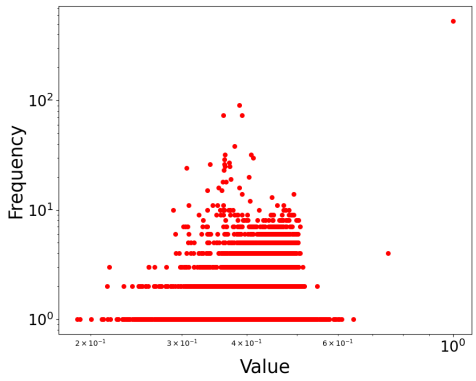


(a) Amazon Dataset (b) Twitch Dataset

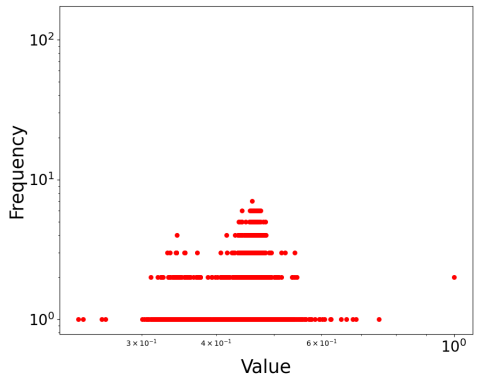
Figure 14: Betweenness Centrality Distributions

Closeness Centrality Distribution
The Amazon Closeness Centrality Distribution shows that most nodes have low closeness centrality values, clustered around a range around 0.5. There are a few outliers with higher values, suggesting that some books are more centrally located within the network.

The graph of the Closeness Centrality Distribution for Twitch shows that most nodes have low closeness centrality values, clustered around a range of 0.5. However, there are some outliers with higher closeness centrality values, suggesting that some streamers are more centrally located within the network.



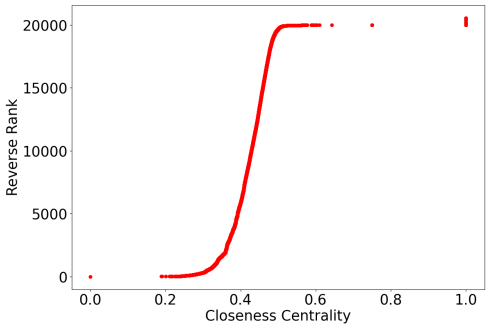
(a) Amazon Dataset



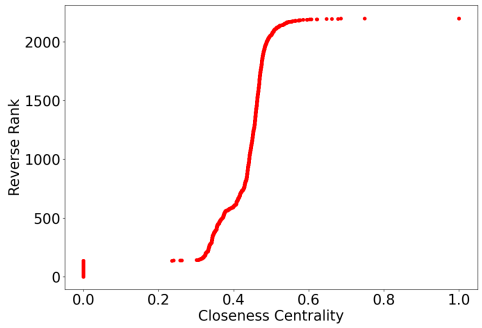
(b) Twitch Dataset

Figure 15: Closeness Centrality Distribution

To gain more insight into the distribution of the closeness centrality values, I have plotted closeness centrality values in reverse rank order. This plot shows that the distribution of these values are similar in both networks.



(a) Amazon Dataset

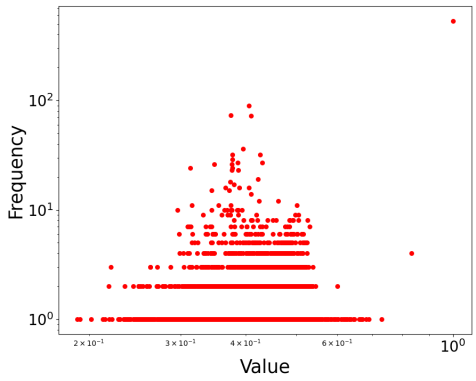


(b) Twitch Dataset

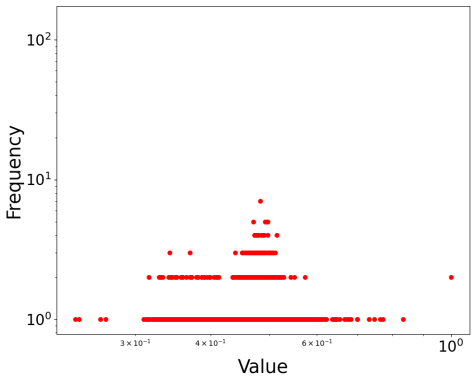
Figure 16: Reverse Rank versus Closeness Centrality

Harmonic Closeness Centrality Distribution

For both harmonic closeness centrality distributions we get more or less the same plots as the closeness centrality distributions.



(a) Amazon Dataset



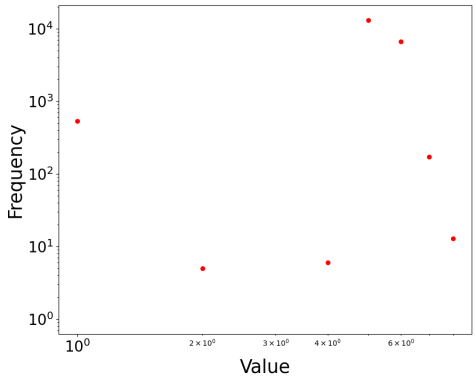
(b) Twitch Dataset

Figure 17: Harmonic Closeness Centrality Distribution

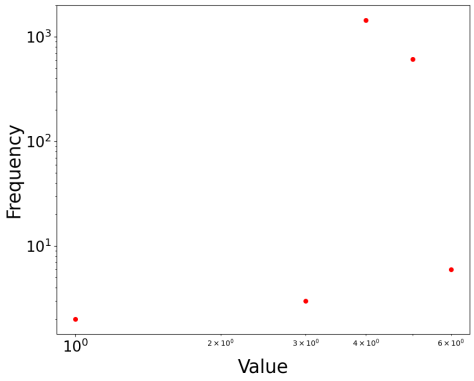
Eccentricity Distribution

The Eccentricity Distribution for Amazon shows that most nodes have an eccentricity value of 5 and 6, indicating that the greatest distance from these nodes to any other node in the network is 5 steps. There are also nodes with eccentricity values of 1, 2, 4, 7 and 8, but these are less common.

The graph of the Twitch Eccentricity Distribution shows that the majority of nodes have an eccentricity value of 4 and 5 but there are also nodes with eccentricity values of 3, 5 and 6 but these are less common.



(a) Amazon Dataset



(b) Twitch Dataset

Figure 18: Eccentricity Distribution

Density

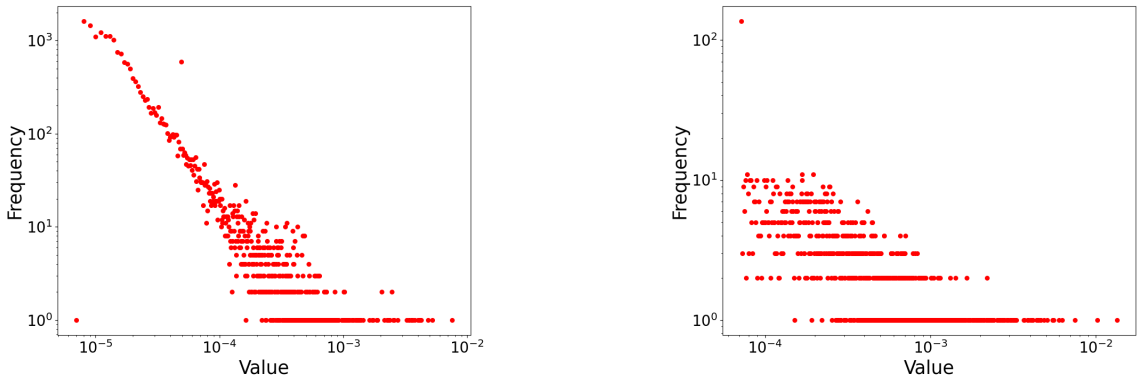
The density of Amazon is 0.005. This very low density indicates that only 0.5% of all possible connections between nodes (books) are actually present.

The density of the Twitch network is 0.034. This means that only 3.4% of the possible connections between nodes are actually present in the network.

PageRank Distribution

The PageRank Distribution graph for Amazon shows that the majority of nodes (books) have very low PageRank scores. This indicates that most books are of low relative importance within the network.

The graph of the Twitch PageRank Distribution shows that the majority of nodes have very low PageRank scores, and a very few nodes have high pagerank values.



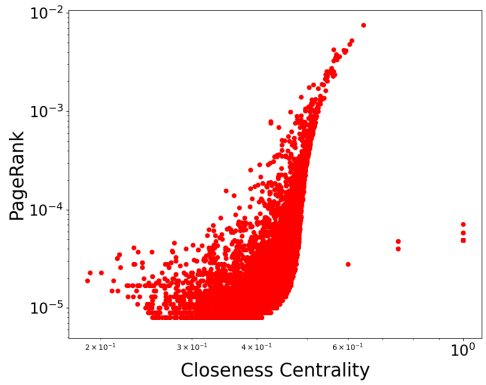
(a) Amazon Dataset

(b) Twitch Dataset

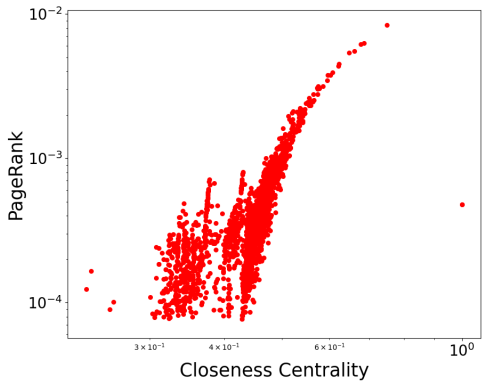
Figure 19: PageRank Distribution

PageRank vs Closeness Centrality

In both plots that are displayed in Figure 19, we see that nodes with higher closeness centrality have also a higher PageRank value. This shows that centrally located nodes have a greater influence on the network. Both plots also contain outliers that have high closeness centrality value and low pageRank value but they are very few in number.



(a) Amazon Dataset



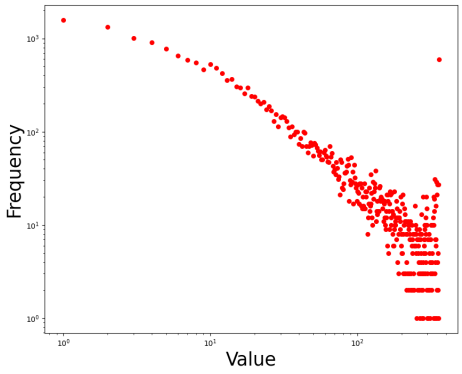
(b) Twitch Dataset

Figure 20: PageRank vs Closeness Centrality

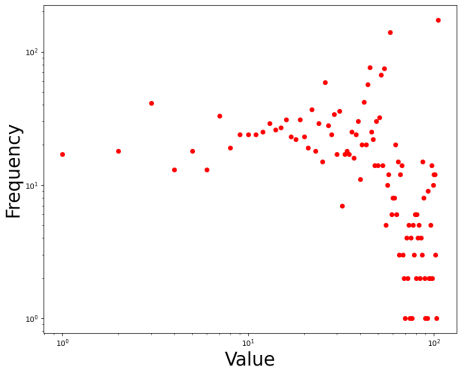
K-Core Distribution

The k-core distribution graph for Amazon shows that the network has hierarchical structure. The plot shows that a large number of the books have smaller k-core value, while core group of books have values between 100 and 1000. This indicates that a big group of books are of low relative importance within the network while a smaller core group acquire the top position and connecting multiple readers based on common interest.

The graph of the Twitch k-core distribution shows that the hierarchical structure is not as prominent in this network as it is in the Amazon network. In Twitch network, we observe that the frequency of nodes sharply decreases for higher k-core values.



(a) Amazon Dataset



(b) Twitch Dataset

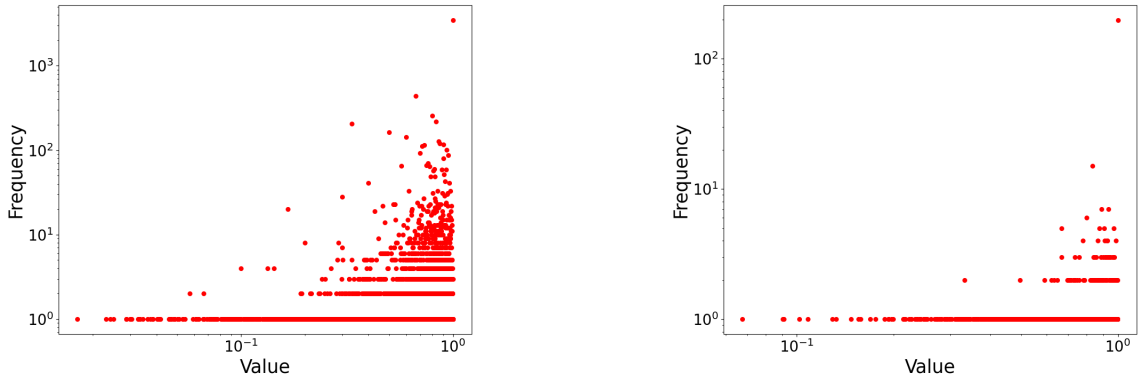
Figure 21: K-Core Distribution

Clustering Coefficient Distribution

The Amazon Clustering Coefficient Distribution shows that most nodes have coefficients clustered around 0.60 to 1.0, indicating a high tendency for books to form tightly knit groups or triangles

with other books.

The graph of the Twitch Clustering Coefficient Distribution shows that the majority of nodes have clustering coefficients clustered around 0.75 to 1.0. There are some nodes with lower clustering coefficients, but these are less common.



(a) Amazon Dataset

(b) Twitch Dataset

Figure 22: Clustering Coefficient Distribution

In this section, we have used the plots of the macro-scale characteristics to understand how the distribution of each property looks. When comparing the plots of both datasets, we see that the plots roughly have the same form but differ in values, which can be attributed to the different sizes of the datasets. This answers the first research question posed in this thesis.

5.3 Discussion

In this section, we will analyse both datasets one by one and then look at the differences and explain why these differences exist.

5.3.1 Twitch Data

When considering the degree distribution, we can conclude that the twitch network of streamers is closely connected to each other because of a low diameter of 6 and an average path length of 2.3. This means that viewers can easily find commonly watched streamers with other viewers and navigate through the streaming platform.

The network has a density of 0.034, which indicates that only a small fraction of all possible connections are made. This can be explained by the language barrier, which prevents streamers from sharing many viewers from different languages. Without looking at the general picture, we see that streamers in a specific community are highly connected with each other, and this is also shown by the average clustering coefficient of 0.760.

The connection between these closely knit communities and other communities is facilitated by dominant streamers acting as bridges. This is shown in the PageRank distribution, where streamers act as critical bridges within the network. PageRank distribution indicates that while most streamers have low relative importance, a smaller group hold significant influence. These streamers are very important because they function as links between different parts of the network. Dominant English-speaking streamers like 'XQC' and 'Shroud' are great examples of streamers who act like bridges.

However, the difference in language causes for the creation of hubs which are further away and far from the bigger central group in the network. We see this in the low eccentricity values compared to the Amazon network, which has a closely connected network. These hubs form their own communities based on language and are highly connected to each other, as we can see in clustering coefficient plots. However, these communities are not totally disconnected from the main group because of the dominant streamers mentioned above that act as bridges. No difference in the Closeness centrality distribution and harmonic closeness centrality distribution further proves this.

When we analyse the visualisation of the network, we observe that communities within Twitch are often influenced by the language barrier. This is evident as Turkish, Spanish, and Korean streamers cluster together with other same language speaking streamers and form distinct communities. This is in accordance with the macro-measurement results. In contrast, English-speaking communities are more diverse and often defined by content rather than language, where streamers specialise in a specific way of streaming and create communities with other similar streamers. This shows that the structure of the Twitch network reflects its diverse and dynamic viewership. The combination of highly connected hubs, strong local clustering and efficient overall connectivity through dominant streamers supports a vibrant community.

5.3.2 Amazon Data

The distribution of degrees suggests that the network exhibits a highly connected. It makes sense because of the average degree of 98.541, which means that on average each book has around a hundred connections to other books. The average clustering coefficient of 0.750 shows strong local clustering. It implies that books are often in their own tight clusters, frequently within the same genre and share reviewers.

As we can see from the k-core distribution and PageRank distribution, most books have relatively low importance while a smaller popular books play a central role in the network. These books act as pivots around suggestions, trendsetting and connecting diverse groups while reinforcing the connectivity of the network. These books also cause a low diameter of 8 and an average path length of 2.4 which indicates that books are fairly well-connected. These bridging books appear also in the betweenness centrality distribution where the few books have high betweenness centrality values and act as critical bridges node between clusters.

The slightly right-skewed closeness and harmonic closeness centrality distributions also confirm that the books are quite close to each other in terms of path length, further showing the close structure of the network. However, there are a few books with higher centrality values, which means that

they occupy central positions in the network. This is also shown in the eccentricity distribution where most books fall on a value of 5, 6 and 1 which implies a dense network where even two far apart nodes are not too distant.

When we analyse the visualisation of the network, we see that the size of the nodes are determined by the number of times a book is read. This explains the large size of the centrally located nodes, which are read a lot and, because of this, are highly connected to other books. The high connectivity of these nodes makes them outliers in their own community due to their broad appeal across different reader groups. We can see this because those nodes are not in the middle of similarly colored nodes but rather in the middle of the map.

Furthermore, books that are part of more than one genre can exhibit similar behaviour to highly connected nodes. These books are not located strictly within their own communities but are intermixed with other books because they share readers. For example, *The Giver* and *Alice's Adventures in Wonderland* are primarily classified as children's literature and share the same community. However, this is not the case with the Harry Potter books. Despite being children's literature, Harry Potter books are mainly categorised as fiction. This causes them to be part of different communities but located near each other due to shared readers.

Moreover, some communities have their own unique readers, which can be observed by their high connectivity within the community and their location at the edges of the map. However, these communities can have a 'bestseller' book that is proportionally more read, making it larger and closer to the central nodes. These nodes act as bridges between their own community and other communities. *Wuthering Heights* and *Great Expectations* are examples of such bridges within the Gothic Fiction community, while *Mere Christianity* forms a bridge for religious books.

Additionally, not every formed community represents a book genre that contains different books. Readers of sequels of the same book are highly connected and some small communities are formed based on those sequels. Some books are not even sequels, but are different versions of the same book, such as papercover and hardcover that might have different book-id. These books are sold on the same page in Amazon.com and when a reader leaves a review for one of the books, it gets registered under each book-id. This causes them to act as if they share readers in the network and form their own community, as we can see with the book *Lust For Life* but they share same reviews instead of same readers.

5.3.3 Highlighted Differences in Network Structures

When we examine the ratio $(avg.degree/nodes) * 100 = Density$, we see that the Twitch network has a higher density (3.4) compared to Amazon (0.48), despite Twitch having a lower average degree than Amazon. This indicates that the Twitch network is more connected than the Amazon network. The lower diameter and average path length in the Twitch network could also support this conclusion, which may be influenced by the smaller data size of the Twitch network.

Another difference is observed in the betweenness centrality distribution. Here, we see that the Amazon network has a higher maximum value and more values at the higher end, indicating that

the Amazon network has a more circular structure, with fewer disconnected hubs from the main body compared to the Twitch network. The language differences that create hubs further away from the main body might explain this phenomenon. Furthermore, this trend is also evident in the closeness centrality distribution, where Amazon has a higher number of nodes with a value of 1, compared to the less frequent nodes with a value of 1 in the Twitch network. Finally, the eccentricity distribution further suggests this conclusion, where Amazon has more nodes with value 1 than Twitch.

A difference that can be explained by different methods of data cleaning is evident in the PageRank distribution, where the Twitch network has a large number of nodes in the far left upper corner that are disconnected from the rest of the red dots. In contrast, the Amazon network shows a more linear log-log plot, which is due to the removal of unpopular books during the data simplification process—a step that was not applied to the Twitch dataset, as mentioned in Section 3.

Lastly, the difference in the maximum values of clustering coefficients can be explained with the nature of books, which often form smaller hubs due to sequels or different titles of the same work. This characteristic causes more smaller hubs in Amazon network, unlike in streaming services, where it is unlikely that two small streamers share high amount of viewers.

6 Conclusion and Further Research

The objective of this thesis was identifying the existence of different online communities through network analysis based on data sets from two online platforms. We analyze networks with an aim to reveal structures and relationships within the co-interest network datasets.

The analysis of both networks reveals the presence of a complex, highly connected community structure, which answers the second research question. The presence of hubs in the clustering coefficient distribution plots and the visualisation of the network shows that both networks have highly connected communities. This is also evident in the relatively short average path length and small network diameter, which suggest that it is easy to move from one book/streamer to another. This results in a positive user experience overall, providing diverse content in the Twitch network and indicating that readers tend to have diverse book genre preferences.

The communities within these networks differ not only in terms of content preference but also in their ability to consume certain types of content. In the Amazon network, the communities are characterised by the genre they represent. A book may belong to more than one genre, but it is assigned to the community with which it shares the most readers. However, this differs in the Twitch network, where communities are characterised not only by stream type and game type but also by the spoken language. Turkish, Spanish, and Korean streamers actively form distinct clusters due to language differences, whereas English speaking communities are more content driven. With this characterisation of communities, the third research question is answered.

When we compare the macro-scale properties of both networks, we observe a few differences. First of all, the Twitch network is in general more interconnected than the Amazon network. This can be explained by the fact that watching multiple streams during a week is much less time-consuming than reading multiple books and leaving reviews on Amazon. This makes it easier for streamers to share viewers compared to books sharing readers.

Secondly, macro measurements like betweenness centrality distribution and the closeness centrality distribution show that the Amazon network has a higher maximum value and more values at the higher than Twitch network. The reason for this difference is the language barrier in Twitch network, which prevents large non-English speaking communities from being highly connected to centrally located English speaking communities. This causes the Twitch network to have a more dispersed structure, as we also observe in the visualisation of the network as part of the meso-scale characteristics.Furthermore, another reason why Amazon network does not have this structure is because a book can have multiple genre and share readers from different genres and this causes the communities to be more intertwined. With this, we answer the first research question.

Lastly, the clustering coefficient distribution shows that the Amazon network contains more highly locally connected nodes, which might lead to tightly knit communities compared to the Twitch network. This is due to the fact that some of the communities in the Amazon network consist of book sequels or books with different titles but the same content. These communities are small and highly connected. This is less likely to occur in the Twitch network.

6.1 Further Research

While this thesis provides a comprehensive analysis of two specific datasets, this opens up several questions for further research:

- **Improved Data Cleaning:** As we have seen, the Amazon Books Review dataset can be further cleaned by adding books together that are sequels or are same books with different titles, for example book title + illustrated edition.
- **Application to Other Datasets:** Extending the methodology to other types of datasets, such as social media interactions, e-commerce transactions or academic citations, could provide additional insight into community structures.
- **Dynamic Network Analysis**: Investigating the evolution of networks to understand how community structures change over time would be a valuable addition. This could involve tracking the formation, growth and dissolution of communities.
- **User Behaviour Analysis:** A deeper investigation of micro-measurements and individual user behaviours and their impact on the overall structure of the network could provide more information. This could include studying user motivations, preferences and interactions in greater detail.

In summary, this thesis highlights the potential of co-interest network analysis in uncovering communities and understanding user behaviour within large datasets. The insights gained from this research can help companies improve user engagement and increase community connections across various online platforms.

References

[1] F. Duarte, "Amount of data created daily (2024)," tech. rep., explodingtopics, June 13, 2024.

[2] P. Gibson, *Types of Data Analysis*.

[3] A. Saxena and S. Iyengar, "Evolving models for meso-scale structures," in *2016 8th international conference on communication systems and networks (COMSNETS)*, pp. 1–8, IEEE, 2016.

[4] A. Saxena, "Evolving models for dynamic weighted complex networks," *Principles of social networking: the new horizon and emerging challenges*, pp. 177–208, 2022.

[5] A. Saxena and S. Iyengar, "Centrality measures in complex networks: A survey," *arXiv preprint arXiv:2011.07190*, 2020.

[6] A. Saxena, R. Gera, and S. Iyengar, "Estimating degree rank in complex networks," *Social Network Analysis and Mining*, vol. 8, no. 1, p. 42, 2018.

[7] A. Saxena, R. Gera, and S. Iyengar, "A heuristic approach to estimate nodes' closeness rank using the properties of real world networks," *Social Network Analysis and Mining*, vol. 9, no. 1, p. 3, 2019.

[8] A. Saxena and S. Iyengar, "K-shell rank analysis using local information," in *International Conference on Computational Social Networks*, pp. 198–210, Springer, 2018.

[9] R. J. Wilson and J. M. Aldous, *Graphs and Applications An Introductory Approach*. Springer London, 2000.

[10] E. Ottel and R. Rousseau, "Social network analysis: a powerfulstrategy, also for the informationsciences," *Journal of Information Science*, 2 April 2002.

[11] A. N. Gorban and G. S. Yablonsky, "Grasping complexity," *Computers and Mathematics with Applications*, 18 March 2013.

[12] A.-L. Barabási, *Network Science*. Cambridge University Press, 2016.

[13] A. Marin and B. Wellman, "Social network analysis: An introduction," *The SAGE handbook of social network analysis*, pp. 11–25, 2011.

[14] A. Saxena, P. Saxena, H. Reddy, and R. Gera, "A survey on studying the social networks of students," *arXiv preprint arXiv:1909.05079*, 2019.

[15] A. Saxena, Y. Pei, J. Veldsink, W. van Ipenburg, G. Fletcher, and M. Pechenizkiy, "The banking transactions dataset and its comparative analysis with scale-free networks," in *Proceedings of the 2021 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, pp. 283–296, 2021.

[16] R. Gera, R. Miller, A. Saxena, M. MirandaLopez, and S. Warnke, "Three is the answer: Combining relationships to analyze multilayered terrorist networks," in *Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2017*, pp. 868–875, 2017.

[17] A. Ozkanlar and A. E. Clark, "Chemnetworks: A complex network analysis tool for chemical systems," *Journal of computational chemistry*, vol. 35, no. 6, pp. 495–505, 2014.

[18] G. A. Pavlopoulos, M. Secrier, C. N. Moschopoulos, T. G. Soldatos, S. Kossida, J. Aerts, R. Schneider, and P. G. Bagos, "Using graph theory to analyze biological networks," *BioData mining*, vol. 4, pp. 1–27, 2011.

[19] J. Moreno, *Who Shall Survive?: A New Approach to the Problem of Human Interrelationsh*. The Journal of Nervous and Mental Disease, 1934.

[20] R. Bian, Y. S. Koh, G. Dobbie, and A. Divoli, "Identifying top-k nodes in social networks: a survey," *ACM Computing Surveys (CSUR)*, vol. 52, no. 1, pp. 1–33, 2019.

[21] I. Bermudez, D. Cleven, R. Gera, E. T. Kiser, T. Newlin, and A. Saxena, "Twitter response to munich july 2016 attack: Network analysis of influence," *Frontiers in big Data*, vol. 2, p. 17, 2019.

[22] A. Saxena, S. Iyengar, and Y. Gupta, "Understanding spreading patterns on social networks based on network topology," in *Proceedings of the 2015 IEEE/ACM international conference on advances in social networks analysis and mining 2015*, pp. 1616–1617, 2015.

[23] G. Learning, "What is network analysis – an overview," tech. rep., Great Learning, 2023.

[24] M. Girvan and M. E. J. Newman, "Community structure in social and biological networks," *PNAS office*, 2002.

[25] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre, "Fast unfolding of communities in large networks," *J. Stat. Mech. (2008) P10008*, 2008.

[26] M. Bekheet, "Amazon books reviews," 2022. Last accessed 26 March 2024.

[27] Twitch, "Twitch api documentation," 2017. Last accessed 26 March 2024.

[28] 2024. Last accessed 23 July 2024.