Social Network Analysis of *The Witcher* series

Maddalena Cavallo, Applied Physics - Physics, 0001058371
Alka Rana, Digital Humanities and Digital Knowledge, 0001072304
Sara Vorabbi, Artificial Intelligence, 0001026226

1 Introduction

Literature is a vast domain, that is the expression of human complexities and encompasses numerous sub-fields. Usually, poetry gains the most attention, being perceived as the noblest form, but prose or narrative books also hold significant value and deserve their place as important contributors to this field. This project is situated within this broad context of literature, in particular in the sub-context of novels. It delves into the intricate world of character interactions and relationships, using the acclaimed "The Witcher" series as a case study. By applying network analysis techniques to this literary work, the project aims to uncover new insights and deepen our understanding of the narrative structure and character dynamics. This innovative approach allows us to explore literature from a fresh perspective, demonstrating the potential of interdisciplinary research in enriching our appreciation of literary texts.

This report presents a social network analysis of the collection of seven books in “The Witcher” series, a popular fantasy saga written by the polish author Andrzej Sapkowski. The series revolves around Geralt of Rivia, a monster hunter with supernatural abilities known in the series as Witcher, who navigates a world filled with political intrigue and magical creatures. Throughout the span of the seven books, Geralt finds himself with the goal of protecting Ciri, a young princess and granddaughter of Queen Calanthe of Cintra, who has a great prophecy hanging over her head. As the plot progresses, Geralt meets various and numerous characters. Among the most important ones, we can cite Yennefer, a powerful sorceress that plays a crucial role in his quest to protect Ciri, the bard Dandelion, and Triss, a sorceress and Geralt’s friend that is a member of the Lodge of Sorceresses (an organization of female mages). The plot of "The Witcher" deals with nuanced character dynamics and unexpected events that keep the plot of the series progressing. In addition to this, the presence of many different subplots with many different secondary characters adds complexity to the story as it progresses through the seven books.

In this project, we embark on a journey into the heart of these stories, utilizing social network analysis as a powerful tool to dissect the evolution of character interactions and identify the key figures that shape the saga’s essence in all the books. This can be seen as a form of literary analysis, as we gain insights of the structure and evolution of the storylines.

2 Problem and Motivation

This project aims to perform a social network study of "The Witcher" series. In particular, it tries to study the interactions between characters within each book and to understand how these relationships evolve over the course of the seven books. Having that “The Witcher” saga is known for its complexity, this type of analysis could be fundamental if someone want to do a literary analysis study on it.

One of the key objectives is to identify the main characters in the story, computing, for example, centrality measures on the graphs. This can be useful in offering insights into their roles and their importance at the plot level.

In addition to this, groups of character can be found in the network. This type of analysis can have a double purpose. Firstly, it helps to better understand characters dynamics and alliances inside the story. Secondly, the groups found from this analysis can be compared with the ones that actually exist in the story. If they correspond, this could help to increase the trust in the validity of the dataset.

Such project assumes a theoretical importance since the study of fictional social networks allows the analysis of its evolution in a controlled environment. In addition, this work can help to show the applicability of social network analysis techniques to literature, and their importance as they provide additional information on the plot that can’t be obtained with a usual literature study.

There are several ways one could use to analyse and better understand the Witcher networks in our study. In this work, graphs were initially compared using the visual representation provided by the Gephi tool [1]. We chose to perform this visual analysis to provide an intuitive glance on the difference in characters interactions among the books. However, visual representation can’t be seen as a proper analysis. For this reason, a further analysis can be found in this project. Quantitative measures were computed in order to have a more objective understanding of character interactions and to compare results of different books e.g. the change of nodes’ importance or the development of clusters as the story evolves. Operating in this way, we aim to provide an objective and reproducible analysis of the book series.

3 Datasets

The Witcher Network dataset [2] is a free public dataset, published on the online platform Kaggle. The dataset contains the entire cast of the story narrated in the novel "The Witcher" and aims to represent all the interactions the characters have throughout the span of seven books.

The data were collected by searching characters names in each line of the books; once a name is encountered it's saved as a `Source` character. A window of 5 successive lines was defined in such a way that if another character, the `Target`, was listed within those 5 lines, there would be an edge connecting the two characters with a corresponding weight of 1. In the case the edge already existed, a $+1$ is added to the edge's weight. Another important information stored is the book in which the interaction takes place. These means that the dataset has 4 relevant columns: `Source, Target, Weight, Book`.

Starting from this dataset, it was necessary to perform some pre-processing operations to prepare the data that would later be used to create the graph containing the networks. A critical issue was the presence of rows belonging to the same book that have inverted `Source` / `Target` and different weights. This problem is the result of the way in which data were collected.

$$Source = Geralt, Target = Foltest, Book = 1, Weight = 5$$

$$Source = Foltest, Target = Geralt, Book = 1, Weight = 4$$

To solve this issue, it was used the function `group_by` provided by the Pandas library [3]. Rows with same `Source`, `Target` and `Book` are aggregated and the values in the `Weight` column - representing the number of interactions between characters - are summed up for each row.

To implement the network structure, and proceeding with the subsequent analysis, the library Network X [4] was used. The result was that seven different networks were obtained, each corresponding to one of the seven books. All the networks are weighted, undirected, one-mode graphs. Each node corresponds to one character and the existence of the edge between two nodes represents the interaction between the two characters. There are neither self-edges nor multi-edges.

It was decided that the graph including all seven books would also be studied. A total of 224 nodes and 1267 edges were found. Tab. 1 shows the distribution of these nodes and arcs among the books.

	Book 1	*Book 2*	*Book 3*	*Book 4*	*Book 5*	*Book 6*	*Book 7*	*All Books*
Nodes	64	36	56	83	76	72	117	224
Edges	177	94	196	305	314	232	452	1267

Table 1: *Number of nodes and edges for each book*

Gephi [1], an open-source and free software, was used to facilitate the visualization of the graphs. Fig. 1 shows the Gephi render for the graph corresponding to Book 1. From visualization only, one could notice how Geralt is at the center of Book 1 interactions.

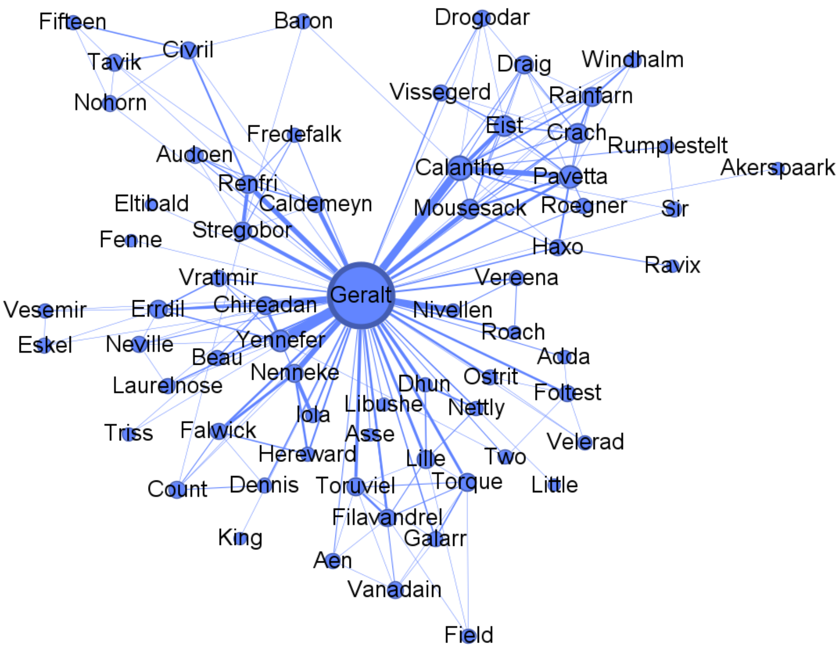

Figure 1: *Graph from Book 1. Nodes dimensions depend on their degree*

4 Validity and Reliability

Validity indicate how accurately the dataset can represent the fictional reality of the world of The Witcher. Since the dataset was created by a Kaggle user, there can be no absolute certainty when it comes to validity. On one side, the way data was collected can be critiqued. Validity could be influenced by the subjective interpretation of what constitutes an 'interaction' between characters: it is not necessarily true that if two names appear five lines apart then the characters are interacting.

On the other hand, there are two factors that suggest the dataset may be valid. One is Kaggle's usability rating as 10.0/10. Since the dataset is sourced from a platform known for its reliable and peer-reviewed datasets, a certain level of accuracy can be assumed. The second and most important factor is that from our findings during the graph analysis, a high level of consistency was found between the relationships established by the characters in the books and those represented in the networks. Two examples can be given. The most important nodes of the networks turned out to actually be the main characters in the books. Nodes representing the characters of Geralt, Ciri, and Yennefer were shown to have centrality values that reflected their leading roles. A second example are the clusters and cliques found in the course of the analysis. For example, in Book 1 characters related to Queen Calanthe's court were recognized as belonging to the same cluster. This leads one to think that the dataset can be used with an high level of validity.

The reliability (i.e. the reproducibility of our study) can be considered high, since it was used a publicly available database, loaded on the Kaggle platform, easy to download and well organized. We followed standard data pre-processing steps and other researchers should be able to reproduce the study with the same dataset, when following those steps explained in Section 3. In particular, the pre-processing choice of adding weights when joining the rows is essential to correctly reproduce our data. The code developed for the study of the network will be available on GitHub for further reproducibility.

5 Measures and Results

5.1 Measures

Simply inspecting the network by visualization could only give a first glimpse on the network properties but cannot serve as a complete analysis. Thus, it is necessary to define some mathematical measures that capture interesting features of network structure quantitatively.
Firstly, we computed some centrality measures on the nodes: Degree centrality, Eigenvector centrality, Betweenness centrality and Closeness centrality. Then we tried to identify group of nodes, finding Cliques, k-Cores and Local Clustering Coefficients, and to find similarities through the Assortative Mixing calculated by degree. Lastly, we studied the whole network measuring its cohesion through the Density and studying its structure through the distribution of centrality measures.

Degree Centrality

The simplest measure one can compute on a node is the degree. The degree of a node is defined as the number of edges connected to it and can be seen as a centrality measure. Having a node i, it can be computed as reported in Eq. 1, with A_{ij} being the element on the i row and the j column of the adjacency matrix associated to the graph. Degree centrality values can be normalized to make comparisons among different networks.

$$\deg(i) = \sum_{j=1}^{n} A_{ij} \tag{1}$$

In our case, each node represent a character and its degree represents the total number of different interactions that character has in the book (i.e. with how many other characters he/she interacts). Thus, by analysing these values, one can identify main characters (larger degree nodes) and marginal characters (low degree nodes) of the plot. We expect Geralt's node to be the one with largest degree centrality.
Scale-free networks are defined as networks whose degree distribution follows a power-law behaviour. These are highly robust networks, that can survive the failure of a sensible number of their nodes. Power law distributions tend to follow a straight line behaviour when plotted in log-log scale. This means that if we take the logarithm of the values, these are linked by a linear relationship: $ln\, p_d = -\alpha\, ln\, d + c$, with p_d the fraction of nodes with degree d. α represents the angular coefficient of the line in log plot and the power of the distribution in the usual scale. One purpose of our analysis will be to check if our networks belong to this particular category of graphs.

Eigenvector Centrality

Another way to measure a node importance is by studying if it is connected to other nodes that are themselves important. This provides a different way to find main characters in the plot and thus we expect to find different names.
The Eigenvector Centrality of a node x_i is computed as the following:

$$x_i = \sum_{j \in \text{neighbours(i)}} x_j \tag{2}$$

The recursivity of this formula can be solved by computing the eigenvectors of the adjacency matrix. Assuming the centrality values to be all positive, one can use the Perron-Frobenius

theorem, that provides the existance of a unique largest eigenvalue **k**. The corresponding eigenvector **x** contains what is defined as the eigenvector centrality value x_i for each node. Finally, the eigenvector centrality values need to be normalized to make the values comparable among different networks.
Eigenvector centrality often has a right-skewed distribution, similar to that of the degree: in log scale the cumulative distribution should be a straight line.

Betweenness Centrality

Betweenness centrality estimates the extent to which a node lies on shortest paths between other nodes. This can help to identify key characters that are gateways in the interactions among the others. Being Geralt a key node in all the networks, we expect him to have a large betweenness centrality, i.e. to be the "bridge" between other characters. On the contrary, having that most of the nodes in our network are in the periphery, usually with few interactions (low degree), we expect them to have very low values of betweenness.
Betweenness centrality is computed as reported in Eq. 3, where n^i_{sd} is the number of shortest paths from s to d that pass through node i and g_{sd} is the total number of shortest paths from s to d. Dividing by the total number of node pairs n^2, the betweenness centrality turns out to be a number between 0 and 1.

$$x_i = \frac{1}{n^2} \sum_{sd} \frac{n^i_{sd}}{g_{sd}} \tag{3}$$

Closeness Centrality

Closeness centrality measures the mean distance from a node to all the other nodes. In particular, it is defined as the inverse of the mean distance l_i as reported in Eq. 4, where d_{ij} is the shortest distance between node i and node j.

$$C_i = \frac{1}{n-1} \sum_{j(\neq i)} \frac{1}{d_{ij}} \tag{4}$$

The larger the closeness centrality the better, as it means that the node is separated from others by only a short distance on average. In our case this could mean that they are characters that might have more direct influence on others. Again, we expect Geralt to have a large value.

Cliques

A clique is a set of nodes such that every member of the set is connected by an edge to every other member. This is an indication of a highly cohesive subgroup. We aim to find cliques in all our networks, in order to provide interesting insights into the relationship between characters. Of particular interest may be both group structures and the appearance of the same character in several different cliques of one network.

K-Cores and Core Periphery Structure

A k-core is a connected set of nodes where each is joined to at least k of the others. It's a less stringent notion of grouping than cliques, but it could be used for the same purpose. This type of analysis provides a way to identify groups of characters that frequently interact with each other. The higher the value of k, the more closely knit the group of characters is.

In addition to this, finding the k-cores with the largest value of k in a network can be used as a measure of core-periphery structures: nodes in this group are "core" nodes within the network, all the others are "peripheral" nodes. Thus, we used this measure to find dichotomised core-periphery structures in the network.

Local Clustering Coefficient

The local clustering coefficient C_i quantifies the fraction of pairs of neighbours of node i that are themselves neighbours, as reported in Eq. 5. This can give an interesting insight of the structure of our network, in particular of the transitivity of interactions between characters.

$$C_i = \frac{\text{\# connected pairs of neighbours of i}}{\text{\# of pairs of neighbours of i}} \tag{5}$$

Local clustering can be seen as a type of centrality measure, where the smaller the values, the more "powerful" the node. We hypothesize that Geralt would have a low clustering coefficient, as he his the center of the story (i.e. of the social network) and thus he will interact with characters not linked one to each other. Additionally, it has been generally shown that nodes with high degree tend to have low local clustering. In Section 5.2 we will check if our networks behave in this common way.

Assortative Mixing by Degree

Assortative mixing (or homophily) reports the tendency of nodes in a network to draw ties with other nodes that are similar to them. If a significant fraction of the edges in the network runs between nodes of the same type, a network is called to be assortative. In our social network, we wanted to check if assortative mixing by degree exists. If this holds, high degree nodes will be preferentially connected to other high degree nodes and conversely. Normalizing the values, the assortativity coefficient ranges from -1 to 1, from perfectly dis-assortative to perfectly assortative network. A positive value indicates that nodes tend to connect with others with similar degree, conversely for negative values.

Density

The density is easily computed as the ratio between the number of edges in the network with respect to their possible total number, as shown in Eq. 6. Its value goes from 0 (no edges) to 1 (complete graph).

$$d = \frac{2m}{n(n-1)} \tag{6}$$

We decided to compute this as an easy way to provide information on the cohesion of our networks i.e. the likelihood of nodes to be connected to each other.

5.2 Results

5.2.1 Degree Centrality

Our first aim was to understand which nodes were the ones with larger degree centrality in each book (i.e. main characters). An example of the values we obtained is reported in Fig. 2, where Geralt clearly stands out above all the others, as expected. To find the other main characters we put a threshold on degree centrality values in each book. Having that most nodes have very low centrality values, we decided to use a threshold equal to 0.35, finding the following principal characters: *Geralt* (all books), *Dandelion* (Books 2 and 5), *Ciri* (Books 3, 4, 6, 7), *Yennefer* (Book 3), *Emhyr* (book 4) and *Milva* (book 5). In the following sections, we will check if the same characters are found as crucial in the networks.

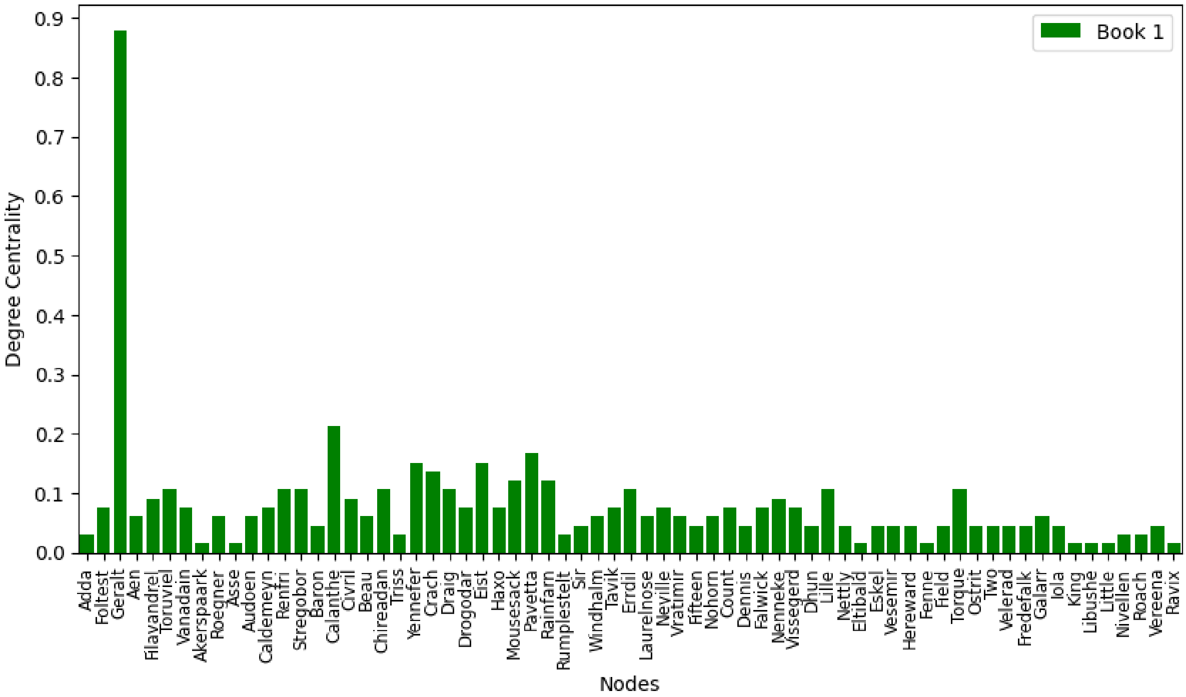

Figure 2: Degree centrality values of nodes in Book 1. Most nodes have very low values

To check if our networks are scale-invariant, we plotted the distributions of the values in a log-log plot, computing the linear regression and linked statistics. The linear fit was performed excluding Geralt's values. Given his role as the hero of the story, his degree centrality values significantly deviated from the others, making him an outlier of the statistic.
Fig. 3 shows two examples of the linear fit for Book 1 and for the graph that includes all books data. Quantitative results of the fit for all graphs are shown in Tab. 2

	Book 1	*Book 2*	*Book 3*	*Book 4*	*Book 5*	*Book 6*	*Book 7*	*All Books*
Pearson's R	-0.86	-0.75	-0.76	-0.81	-0.88	-0.80	-0.88	-0.97
p-value	1.4e-3	1.2e-2	5.9e-4	2.2e-5	1.1e-7	2.3e-4	8.2e-9	5.7e-12

Table 2: Statistics of the linear fit on degree centrality distributions for all the graphs

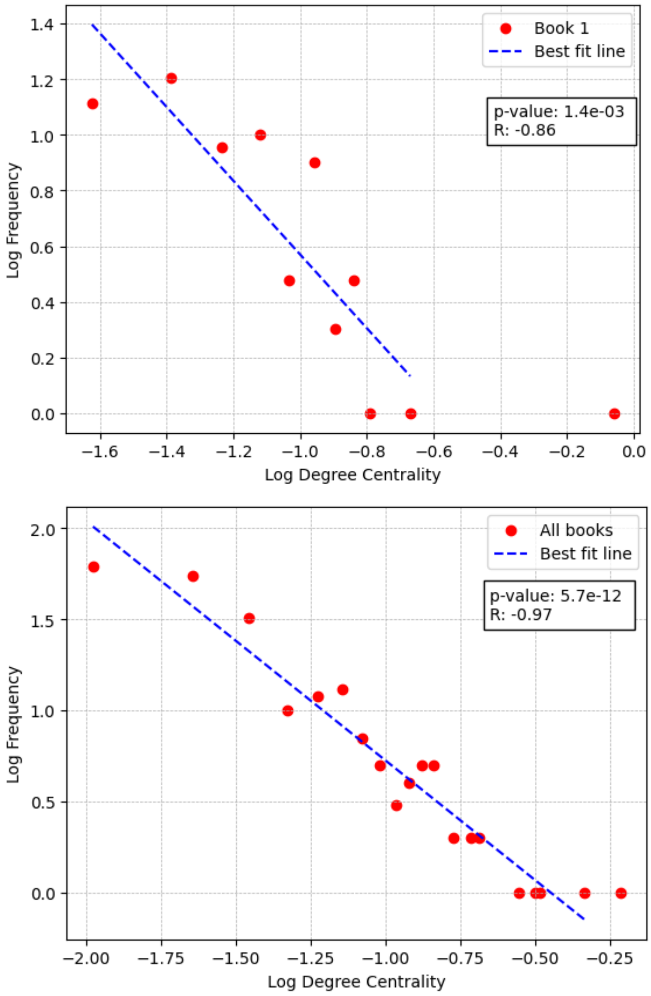

Figure 3: *Log-log plot of the degree distribution for Book 1 and for all books together*

The graph including all the books yielded the best results ($R = -0.97$, $p - value = 5.7 \times 10^{-12}$), primarily due to its larger node count compared to single book graphs. Regarding single books, best values were obtained for Book 7 ($p - value = 8.2 \times 10^{-9}$). Also in this case we can assume that this result is due to its large sample number, as shown in Tab. 1. Generally speaking, all outcomes suggest our network to be scale-free. Indeed, Pearson's correlation coefficient $R = [-0.97, -0.75]$ suggest a strong linear anti-correlation. The p-value range $[10^{-12}, 10^{-2}]$ indicates that we can reject the null hypothesis, as also the worst value 1.2×10^{-2} is larger than the common used threshold = 0.05. Thus we can conclude that our networks are likely to be a scale-free. The large distribution of order of magnitude of the p-value reflects the different structures of the networks of each book. As expected, the best (i.e. lowest) p-value was observed for the largest graph that includes all nodes and all edges. This is because statistical measures generally improve with an increase in sample size. This could also be the reason for the worst statistical scores in Book 2 (lowest number of nodes).

The cumulative distribution of degree centrality scores was also computed, as shown in Fig. 4 for the total graph. The plot is an increasing curve that proceeds by steps, as it can be easily seen in logarithmic scales. This is reasonable as the degrees are normalized and thus they are computed as ratio of integer numbers, providing discrete values. This means that increasing the values on the x-axis, the cumulative will show an horizontal line for values that don't exist and then an abrupt increase on allowed discrete values.

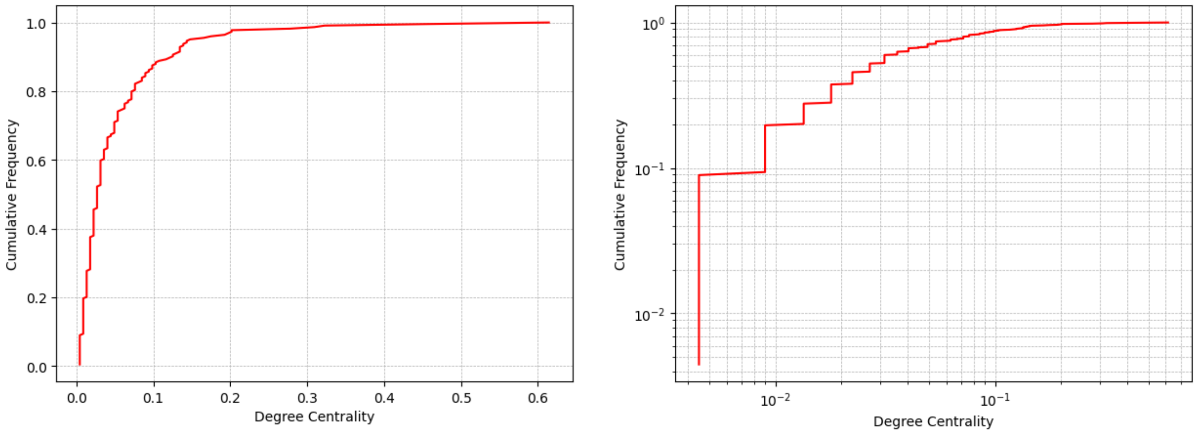

Figure 4: *Cumulative distribution of the degree centrality in normal scale and in log plot for the whole network*

5.2.2 Eigenvector Centrality

Eigenvector centrality gives another way to find important characters, giving more importance to nodes that are linked to other important nodes. Results obtained differ from the ones for degree centrality, as can be observed qualitatively in Fig. 5. With eigenvector centrality, some nodes acquire an additional "importance" while others loose it.

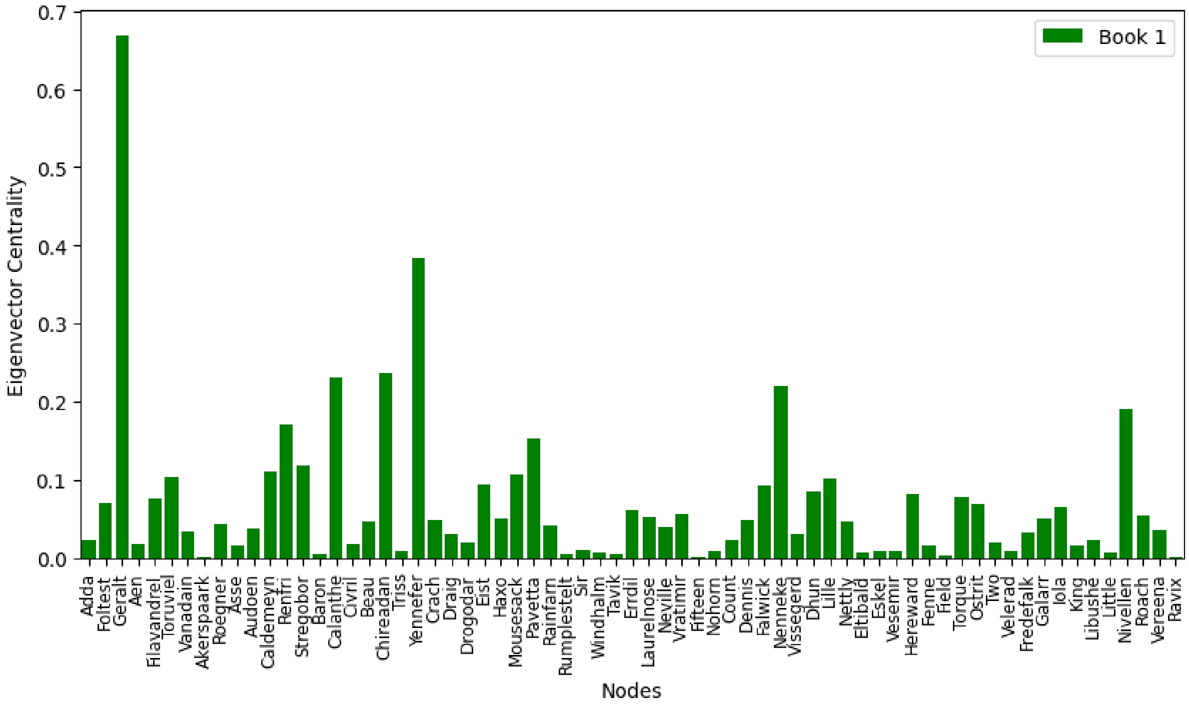

Figure 5: *Eigenvector centrality values of nodes in Book 1. Different distribution of values compared to the degree centrality*

More specifically, using the same threshold as before, we found the following characters: *Geralt* (all books), *Yennefer* (books 1, 3, 4, 6, 7), *Dandelion* (books 2 and 5), *Ciri* (books 3, 4, 6, 7), *Triss* (book 3), *Milva* (book 5), *Zoltan* (book 5), *Cahir* (book 6).

We could observe that Geralt, Ciri and Dandelion keep the same importance as in the previ-

ous section. But we can appreciate, for example, that Yennefer has increased her importance. Lastly, we can observe as other characters as Cahir or Triss emerge as "important" using the previously cited measure. This again means they have few but important connections.
Fig. 6 shows the eigenvector centrality distribution as a scatter plot and its cumulative in log scale for Book 1. Qualitatively, one can observe that the scatter plot resembles a right-skewed distribution, while the cumulative a straight line, as expected.

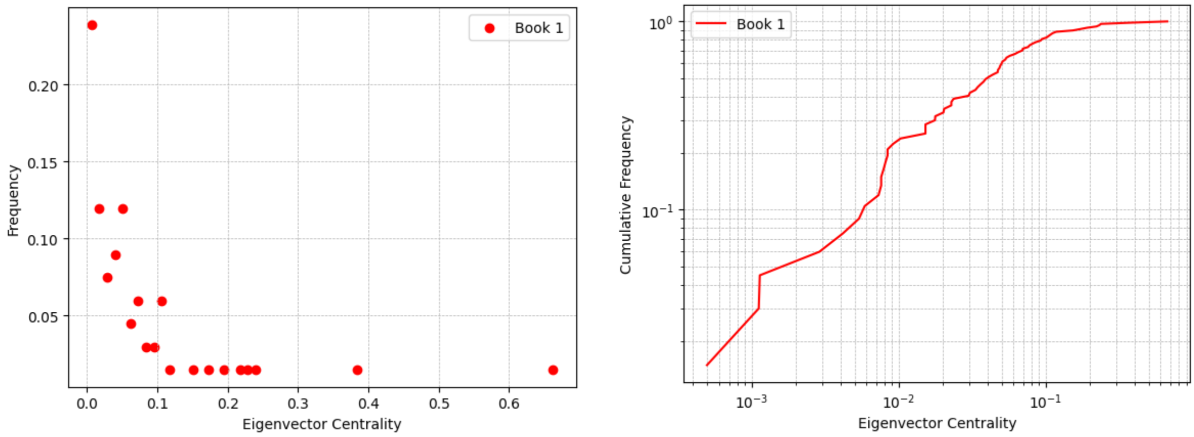

Figure 6: *Scatter plot and cumulative for eigenvector centrality values in Book 1.*

5.2.3 Betweenness Centrality

We computed the betweenness centrality to find nodes that are "bridges" in interactions between others. Most nodes were found to have zero or close to zero scores, meaning they are outside the major part of the shortest paths in the graphs. As an example, results for Book 1 are shown in Fig. 7.
Concerning the comparison among the books, surprisingly, Geralt wasn't the node with largest betweenness centrality in all of them. In particular, using a threshold of 0.2 on the values, we found the results reported in Tab. 3.
Different nodes were found compared to the ones found in Sections 5.2.1 and 5.2.2. New characters were found (Aen and Braen), meaning that they are not important characters concerning their interactions but they are in the relationship among their neighbours. Some characters disappeared: Yennefer among all of them. This shows how different centrality measures characterise the importance of the characters in different ways. Finally, one could notice that both graphs for Book 7 and for all books don't have any node with betweenness centrality larger than the threshold. This means that there aren't nodes that serve as bridges more than others, reducing the reliance of the graph on a single node.

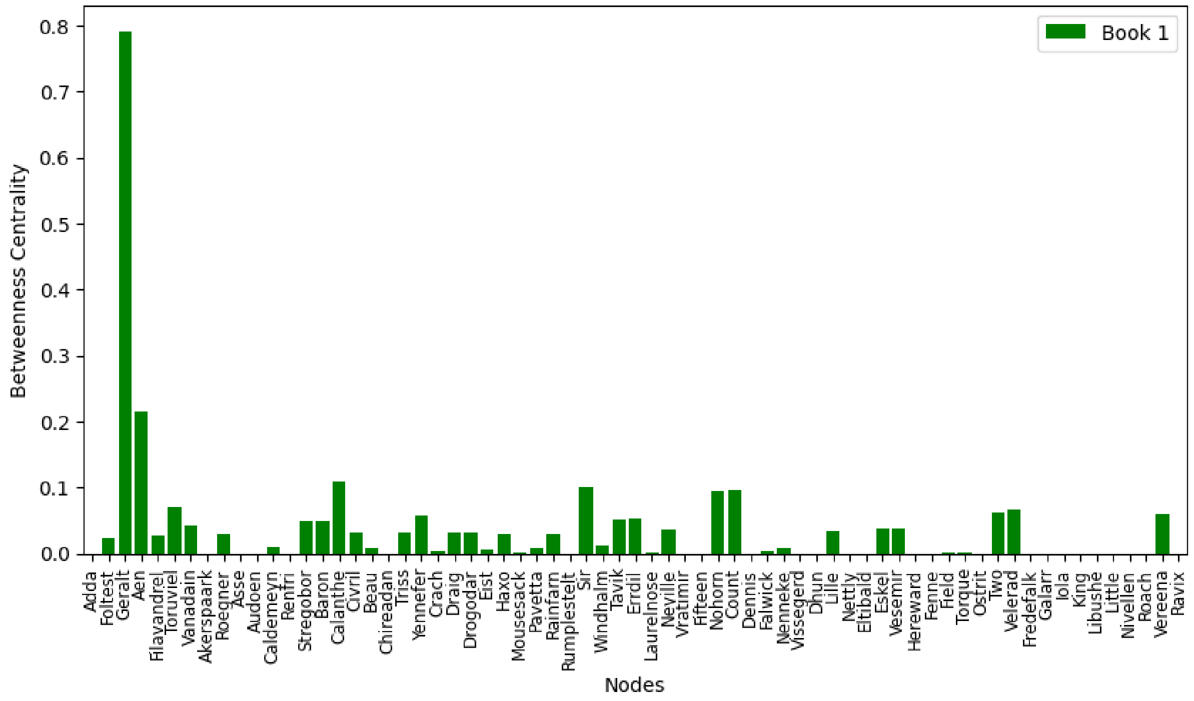

Figure 7: *Betweenness centrality values of nodes in book 1. Most nodes have value zero or near zero*

	Book 1	*Book 2*	*Book 3*	*Book 4*	*Book 5*	*Book 6*	*Book 7*	*All Books*
Geralt	✓	✓	✓		✓	✓		
Aen	✓							
Braen		✓						
Dandelion		✓						
Ciri		✓	✓			✓		
Emhyr				✓		✓		

Table 3: *Betweenness centrality values above 0.2 in all the different books*

5.2.4 Closeness Centrality

As expected, Geralt was found to have the largest closeness centrality value in all the networks (e.g. nearly 0.9 in book 1, Fig. 8). This means that he is maximally close to most of the nodes in the network (indeed, he also has largest degree).

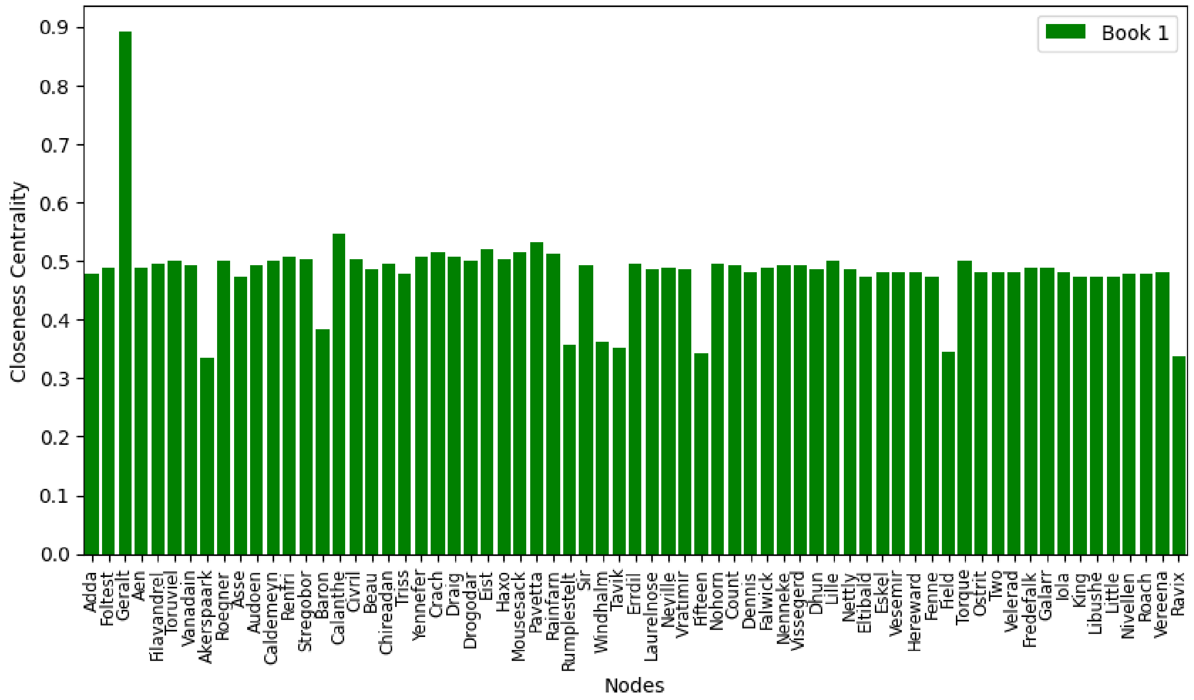

Figure 8: *Closeness centrality values of nodes in book 1. Geralt has by far the largest value.*

Concerning the distribution of values in a single book, it was generally found that most nodes scores of closeness centrality are in a small range (excluding Geralt, see Fig. 8). This could be better appreciated, for instance, through the distribution plots for Book 1 and all books as shown in Fig. 9. In order to quantitatively find where each histogram was peaked, mean values were computed in the distributions of each book through a weighted mean. Results are shown in Tab. 4. All different books show similar values of the mean of the closeness centrality, with the lowest being in Book 4 (0.41) and the largest in Book 2 (0.51).

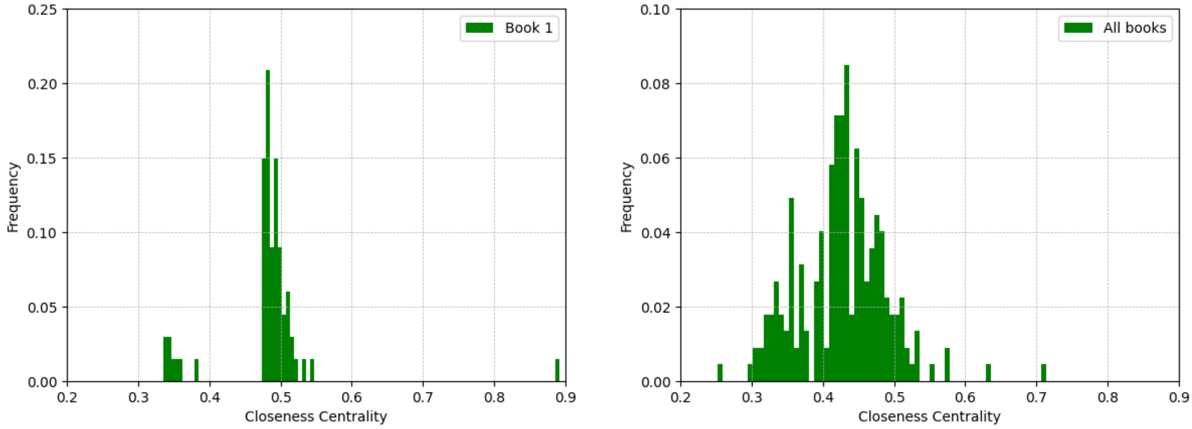

Figure 9: *Closeness centrality distributions for graphs related to Book 1 and all books.*

	Book 1	*Book 2*	*Book 3*	*Book 4*	*Book 5*	*Book 6*	*Book 7*	*All Books*
Mean value	0.48	0.51	0.45	0.41	0.43	0.42	0.39	0.42

Table 4: *Mean Closeness centrality values in each book*

5.2.5 Local Clustering Coefficient

Local clustering coefficients for nodes in Book 1 are shown in Fig. 10 as an example. As expected, Geralt has a low value (< 0.1), as his connections are unlikely to interact among each other. Many nodes have score one but this is reasonable. Indeed, taking Adda as an example, she is only linked to Geralt and Foltest, that are neighbours itself. This means they are a small clique and this makes Adda to have local clustering coefficient equal to one. On the contrary some characters have score 0: this happens when they are linked to one character only, e.g. in Akespaark case.

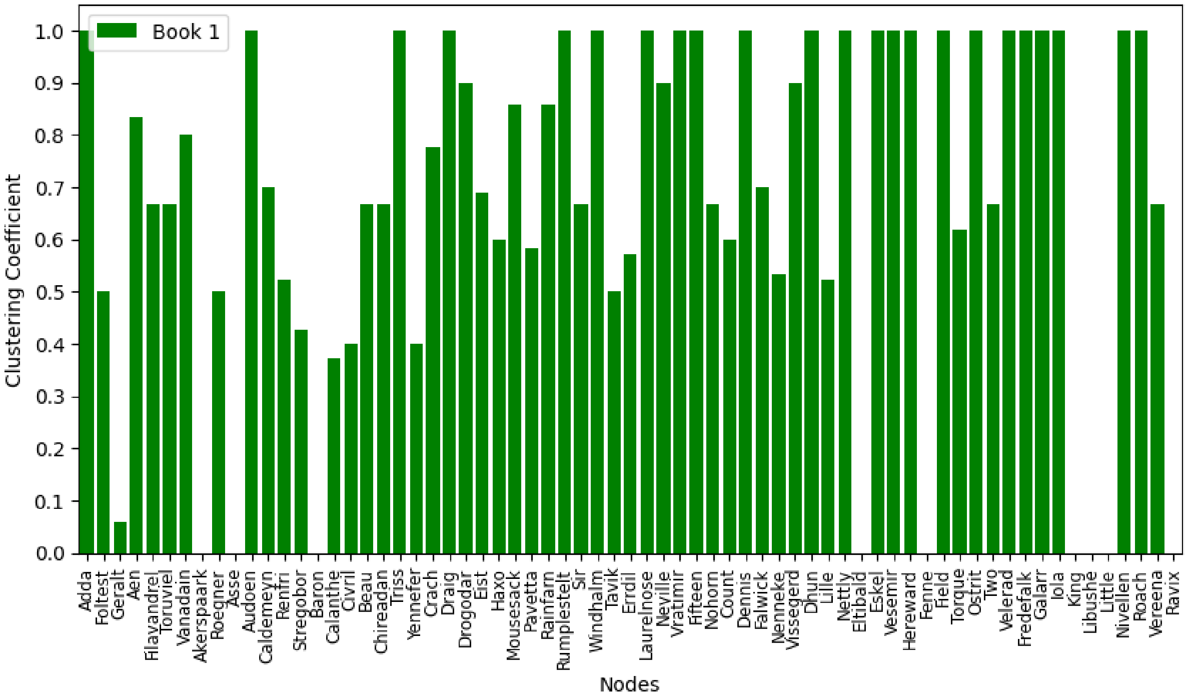

Figure 10: *Local clustering coefficient for nodes in Book 1*

Plots of the local clustering coefficient versus the degree centrality are shown in Fig. 11 and 12. In all books we observed an inverse relation between values on x and y axis: while increasing the degree centrality of a node, its local clustering coefficient starts decreasing. This phenomenon can be explained considering that most of the characters tend to belong to small but well-connected groups. This leads to a low degree centrality value but an high local clustering coefficient. Viceversa, for main characters. Taking Geralt as an example, he interacts with almost all nodes (high degree) but these nodes belong to different groups. This means that his neighbours are poorly connected among each other and this leads to a lower local clustering coefficient.

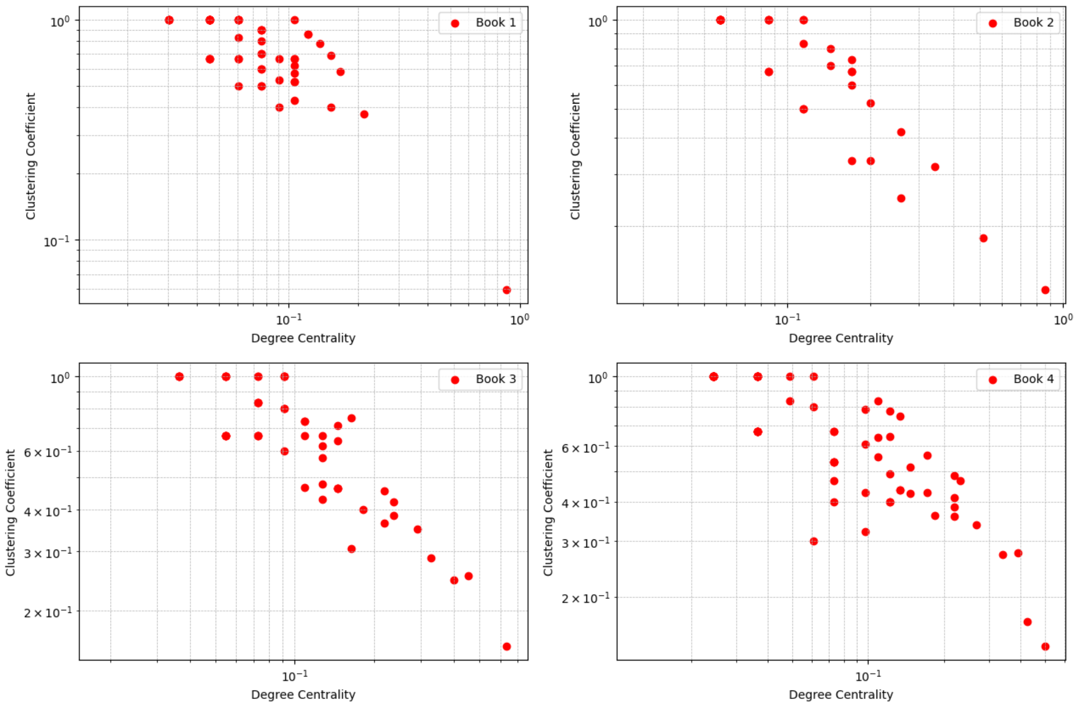

Figure 11: *Local clustering versus degree centrality plot, from Book 1 to Book 4. The local clustering coefficient decreases while increasing the degree centrality of a node.*

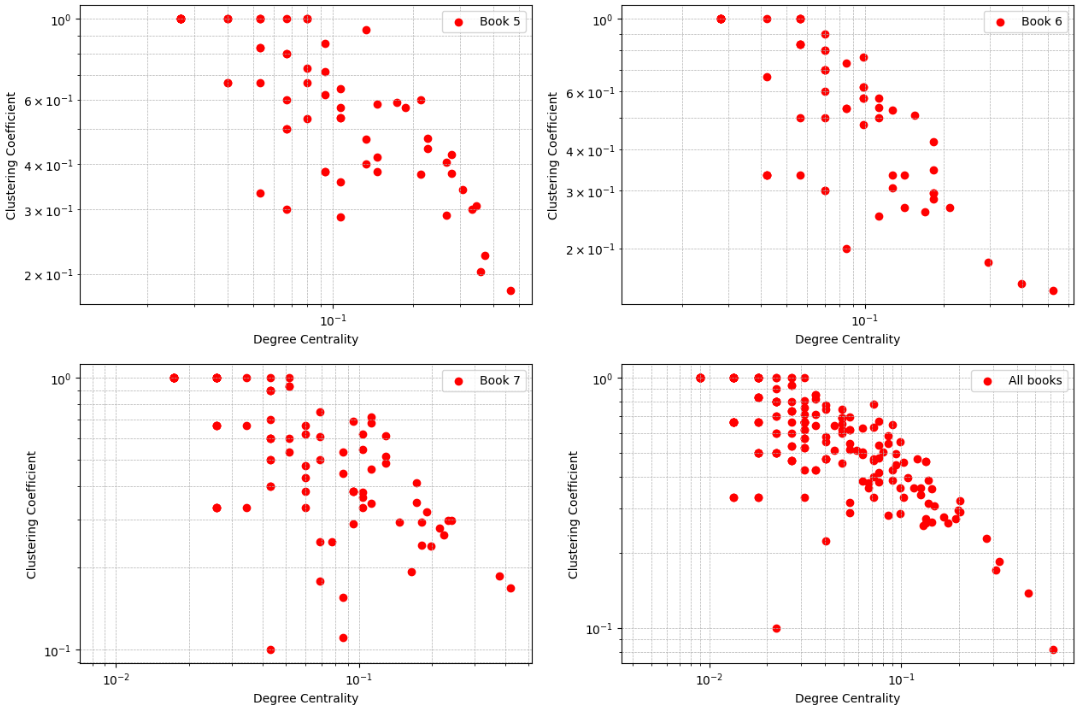

Figure 12: *Local clustering versus degree centrality plot, from Book 5 to Book 7 and all books. The local clustering coefficient decreases while increasing the degree centrality of a node.*

5.2.6 Cliques

Studying cliques in the network is another way to understand character relationships. In order to focus on the most interesting cliques, only the ones with a number of nodes larger than 6 were included. No cliques of this type were found in Books 2 and 6, revealing the presence of small groups only in these graphs.

A first interesting example of a clique appears in the first book. This clique is shown in Fig. 13 and consists of 8 nodes: *Geralt*, *Calanthe*, *Pavetta*, *Mousesack*, *Draig*, *Rainfarn*, *Crach*, *Eist*. This clique is noteworthy, as all characters within it are associated with Queen Calanthe's court, a crucial location in Book 1.

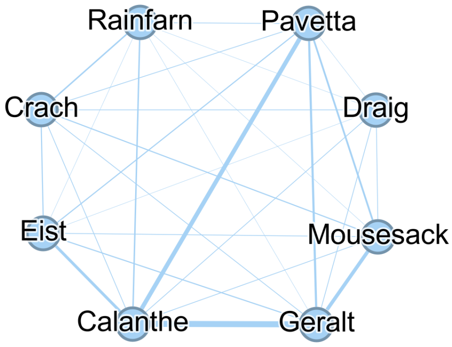

***Figure 13:** Largest cliques from Book 1. The edges are weighted*

Surprisingly, in the graph associated with Book 4, eleven cliques of size 7 were found. All of them share three main characters: *Geralt*, *Yennefer* and *Tissaia*. This means that these characters are key figures in the interactions of these cliques and can be seen as "connectors" between them. This suggest that they play a significant role in the plot of this book, being involved in many different subplots inside the story. A similar pattern emerges in the fifth book. In this network, we discovered that three mage characters - *Sabrina*, *Philippa*, and *Yennefer* - are present in all the ten cliques of the maximum dimension of 8 that were identified. Given that an important portion of the plot revolves around the Lodge of Sorceresses, to which these mages belong, this observation highlight the validity of our dataset.

Two cliques of 8 elements were found in Book 7 (Fig. 14). The peculiarity is that seven out of eight characters are present in both of the cliques. These seven characters, indeed, play a central role in the narrative of the last book (Geralt, Yennefer and Ciri among all). Furthermore, the unique character in each clique, Fringilla and Zoltan, could be seen as a bridge that connects different parts of the narrative or different groups of characters. Fringilla, for example, is the gate of the clique to the mages inside the Lodge of Sourceress, that otherwise would be poorly or even not connected at all to the other nodes in the clique.

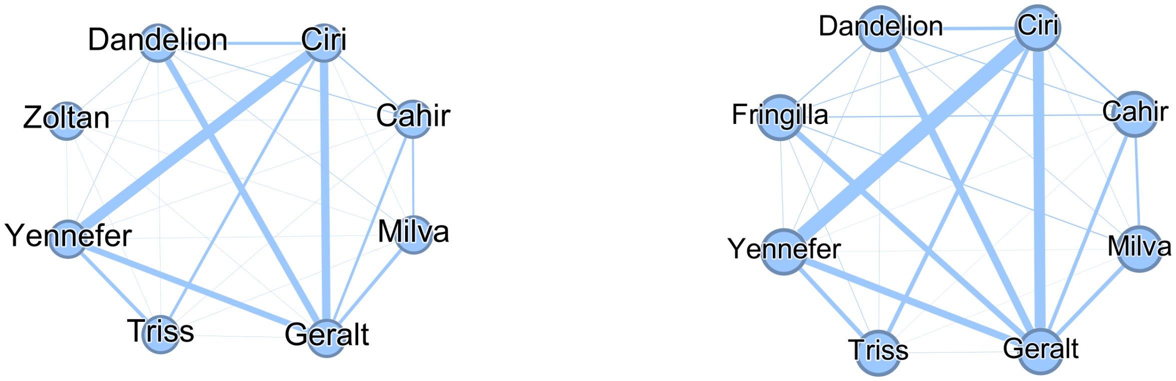

Figure 14: *Largest cliques in Book 7. They only differ for one node (Zoltan and Fringilla)*

5.2.7 K-Cores and Core-Periphery Structure

As we said in Sec. 5.1, k-cores are a less stringent notion of grouping so we expect to find more characters in them with respect to cliques.

Generally speaking, we found that the most frequent characters in the k-cores are the ones cited in the previous measures. *Geralt* appears in the 4-, 5-, 6-, and 7-cores of every book (where such cores exist), corroborating his leading role in the plot. Other characters like *Ciri*, *Yennefer*, *Dandelion*, *Calanthe*, and *Vilgefortz* also frequently appear in the 4-cores, and together with *Emhyr*, *Philippa*, and *Triss* they appear frequently also in the 5-cores and 6-cores.

Interestingly, in Book 1, the characters *Calanthe*, *Pavetta*, *Draig*, *Vissegerd*, *Crach*, *Mousesack*, *Eist*, *Rainfarn*, and *Drogodar* form a 5-core together with *Geralt*. This again identify the presence of a strong relationship between characters associated to the court of Queen Calanthe, as we found in Section 5.2.6 studying the 8-clique for this book. In particular, all the characters in the 8-clique also belong to the 5-core.

Increasing k, fewer characters were found in the k-cores, indicating a tighter group. In particular, each book has a different maximum value of k, after which no k-cores are found. Book 2 and Book 6 are the ones with lower values of the maximum k, 4 and 5 respectively. These maximum k-cores also have less characters in them compared to same k-value cores of other books. This is in accordance with what we found in Sec. 5.2.6, where no large cliques were found in these books. This support the hypothesis that these books have small groups of characters only.

Among all books, the largest value of k for which a k-core exist was found in Book 5, with $k = 9$ (Fig. 15). Significant characters form this 9-core, as most of them are mages belonging to the Lodge of Sorceresses. Once again, this finding corresponds with what was found in the cliques in Section 5.2.6.

Lastly, the largest-k cores for each graph were computed. This allowed to find core-periphery structures in all of them. For example, the 9-core in Fig. 15 is the core of Book 5. That is, eleven over seventy-six nodes in the graph belong to what can be called core of the network, while the other sixty-five constitute the periphery. Remarkably, no character appears in all the maximum k-value cores of the books. This happens also for Geralt, who is absent from the 9-core of Book 5, as can be seen in the previously mentioned figure.

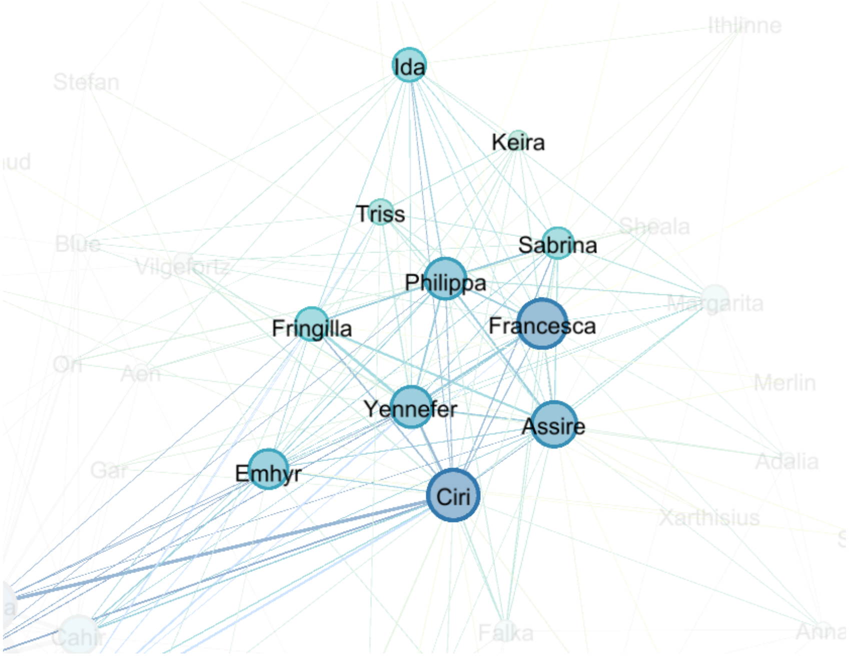

Figure 15: *K-Core with k=9 from Book 5*

5.2.8 Assortative Mixing by Degree

All the assortativity coefficients found are negative, as shown in Tab. 5. It can then be concluded that all the networks from the book series are disassortative, even if this is not a strong tendency as values are not very close to -1. This means that characters with a large number of interactions (high degree nodes) tend to interact mostly with low degree characters, but not exclusively.

	Book 1	*Book 2*	*Book 3*	*Book 4*	*Book 5*	*Book 6*	*Book 7*
Assortativity coeff.	-0.25	-0.39	-0.24	-0.25	-0.25	-0.19	-0.14

Table 5: *Assortative Mixing by degree for each book*

Examining the values, we could notice an abrupt increase from Book 1 to Book 2, then a constant value up to the ending of the story, where it decreases (absolute values). This pattern can be attributed to the evolution of character interactions throughout the books and can be explained as the following. Book 1 is an introductory book, mainly focused on Geralt (only character with high degree, see Sec. 5.2.1). All other characters are low degree ones and Geralt interacts with some of them. Book 2 marks the real start of the story, where new important characters are presented (e.g. Ciri) and others considerably increase their degree (e.g. Dandelion). Each of them is initially presented separately in his/her own group (e.g. among his/her friends or family) and rarely interact with each other (large negative assortativity coefficient). As the story progresses, these main characters become more interconnected, In particular, they will become more and more linked when going to the end of the story. This means that, especially in the last books, they start to interact slightly less with secondary characters and slightly more between them. This behaviour leads to a better balance in the interactions between low and high degree characters, causing the assortativity coefficient to move to values closer to 0.

5.2.9 Density

As shown in Tab. 6, the densities computed are generally low: $[0.07 - 0.15]$. This indicates that our networks are not very cohesive i.e. the characters are not very connected one to each other. Among the books, Book 2 exhibits the largest value, possibly suggesting a stronger and more complex relationship between its characters. The lowest density values were found instead in Book 1 and Book 7. In Sec. 6 an hypothesis for the pattern of these values among the books will be presented.

	Book 1	*Book 2*	*Book 3*	*Book 4*	*Book 5*	*Book 6*	*Book 7*
Density	0.08	0.15	0.13	0.09	0.11	0.09	0.07

Table 6: *Density scores for each book*

One important note should be done here. It is important to remember that the density doesn't take into account the weight of the edges. Therefore, networks with the same density could potentially have very different total numbers of interactions in reality.

6 Conclusions

This work aimed to provide a valuable insight into the story of "The Witcher" book series, a collection of fantasy novels. This was done computing some quantitative measures on the graphs of the seven books and on the graph that includes all nodes and edges of the story. The main purpose of this analysis was to study the interactions of characters inside each book and the evolution of it throughout the series, together with the identification of main characters and characteristic groups inside the networks.

Different centrality measures were computed in order to compare which characters were found as the most important ones. Regarding degree centrality, *Geralt* was found to have the largest values in all the books, in agreement with his leading role in the plot. The finding that he is the only "important" character in Book 1 is reasonable, as the story of this book only focuses on him. Other characters start to appear in the other books as the story progresses. Putting a threshold on degree centrality in all books, the other important characters were found to be: *Ciri*, *Yennefer*, *Dandelion*, *Emhyr* and *Milva*. These characters can be seen as the main ones concerning the number of nodes they are connected to.

The above mentioned characters were compared with the ones found with the same procedure applied on eigenvector centrality values. We found that *Geralt*, *Ciri* and *Dandelion* keep the same importance as in degree centrality section. *Yennefer*, instead, increase a lot her importance. This makes sense as in the plot she interacts mainly with other important characters (Geralt, Ciri and Dandelion), and this increases her eigenvector centrality scores.

This result can be linked with what was found in Sec. 5.2.3, where Yennefer wasn't found as an important character regarding betweenness centrality. This shows how different centrality measures differently characterise characters' importance. Yennefer is one of the main characters in the plot but she don't serve as a bridge to connect other characters. Indeed, in the story she is mainly an isolated character, without a family or a group of friends except from Geralt and Ciri. This leads to high eigenvector centrality values but low betweenness ones. Among characters with the highest betweenness centrality, one could notice that new characters were found (*Aen* and *Braen*) and that Geralt has a value above the threshold in all book except the fourth one. Most nodes were found to have 0 or close to 0 values, meaning that the network's shortest paths mainly rely on few nodes. Only in Book 7 and in the whole graph no node with betweenness centrality larger that the threshold were found. This means that there aren't nodes that serve as bridges more than others, reducing the reliance of the graph on a single node. We can hypothesize this to be due to the larger node number of these two graphs compared to the others (see Tab. 1), that could lead to a more even distribution of shortest paths across different nodes.

Results from closeness centrality confirm the leading role of Geralt, as his value clearly emerge from the others as seen in Fig. 8. This is reasonable as he is the hero of the story, that interacts with many characters (highest degree, see Sec. 5.2.1) and thus it is directly linked to most nodes in the network. In terms of closeness, this means that he is maximally close to most of the nodes in the network, resulting in a large closeness centrality value. Other characters' scores were found to belong to a small range both in the same book and among books. The mean closeness value for the books go from the lowest value in Book 4 (0.41) to the largest in Book 2 (0.51).

Analyzing the cliques, as seen in the Sec 5.2.6, we were able to find some interesting groups of characters. In Book 4, for example, all the eleven 7-cliques found share three of the most

important characters of the story, namely *Geralt*, *Yennefer* and *Tissaia*. We pinpoint these as "connectors", characters that are involved in more subplots that are linkers between cliques. In support of this hypothesis, we found that these characters have closeness centrality values among the highest of Book 4. As mentioned in the appropriate section, in Book 7 two cliques that only differ by one node, *Zoltan* and *Fringilla*, were found. Comparing the values of betweenness centrality, we found that Zoltan node's value is the highest of Book 4 while Fringilla's is among the lowest. We can explain this fact using the visual representation of Book 4 graph provided by Gephi, and pointing out how Fringilla is a connection point exclusive to the nodes that correspond to the Lodge of Sorceresses, while Zoltan is a link to a multitude of characters from different subplots.
The analysis of cliques proceeded hand in hand with the analysis of k-cores. Taking up what we found in Sec. 5.2.6, the conclusions from the k-cores found are similar to the ones from the cliques. For example, in Book 1, all the characters found in a 8-clique also belong to the 5-core. We also noticed that all these characters belong to Queen Calanthe's court. This finding further emphasizes the power of social network analysis: even without any prior knowledge of the narrative, one could still infer the existence of a tightly interwoven subplot involving these characters. Studying core-periphery structures, no characters were found in the core of all graphs. This can be seen as a reflection of the complexity and intricacy of "The Witcher" plot, which features numerous subplots throughout the whole story with many different characters (224 in total, see Tab. 1).
Local clustering coefficient was computed in each node for each graph. Being Geralt the center of interactions in Book 1, we found him to have low local clustering coefficient value (< 0.1), as it's unlikely that also his neighbours are linked. Many values equal to 0 and 1 were found and coherently explained in Sec. 5.2.5, by looking for hints from the Gephi visualization. Nodes with null local clustering coefficient, for example, were found to be the ones only linked to another node. Finally, an inverse relationship was found between the local clustering coefficient and the degree centrality of a node.

All the networks were found to be disassortative, as all the assortativity coefficients were negative (even if this was not a strong tendency). This means that characters with a large number of interactions (high degree nodes) tend to interact mostly with low degree characters, but not exclusively. The pattern of the evolution of the assortativity coefficients among books could be explained with the hypothesis previously made. In particular, we found Book 1 to be an introductory book focused on Geralt, Book 2 representing the "real" start of the story, with new characters initially presented in their own groups that then begin to interact among each other as the story progresses in the following books.

The network was also found to be poorly cohesive, as density values were generally low: $[0.07 - 0.15]$. It's interesting to notice that the lowest density values were found in Book 1 and Book 7. For the first book, this can come from the hypothesis previously made in Sec. 5.2.8: this introductory book is mainly focused on Geralt and its interactions, leaving little room for other characters to interact. As for Book 7, instead, the low density could be due to the larger number of nodes in this graph. Indeed, increasing the number of nodes, it becomes less likely for all nodes to be well connected. Among the books, Book 2 exhibits the largest density value, possibly suggesting a stronger and more complex relationship between its characters. This hypothesis would be supported by results in Tab. 4, where Book 2 graph also have the largest mean closeness value. However, one should also notice from Tab. 1 that Book 2 is also the one with the lowest number of nodes, and this could be the underlying reason for which its nodes are better connected. Indeed, it's more likely for graphs with few nodes to have larger density values.

Performing a linear regression on the degree centrality distribution in each book, we found that our graphs are likely to be scale-invariant. Indeed, all p-values and R-values supported a linear relationship between values in the log-log distribution. In particular, the Pearson's correlation coefficient $R = [-0.97, -0.75]$ is close to -1, suggesting a strong linear anti-correlation. The p-value $[10^{-12}, 10^{-2}]$ is lower than the common used threshold < 0.05 used to reject the null hypothesis H_0, the hypothesis of our data being not correlated. Thus, we concluded that our networks are likely to be scale-free. This also means that they are highly robust i.e. they can survive the failure of a sensible number of their nodes. Applied to our context, this means that, apart from the specific Geralt case and other high-degree nodes, other characters can die (or disappear from the plot) without heavily affecting the progression of the story.

7 Critique

The outcomes achieved from the analysis of the seven graphs do correspond to what can be found studying the plot of "The Witcher" series. This increases the validity of the dataset. However, further controls could have been implemented. For example, we could have cross-verified a sample of interactions from the dataset with the actual part of text from the books. Furthermore, the method of data collection could lead to some perplexities, as a strong assumption about what constitutes an interaction was made. Indeed, it cannot be assured that if two characters appear in close lines, they actually interact.
To sum up, the way data were collected have some potential limitations, but our results increase our confidence in the validity of the dataset.

As far as our work is concerned, we tried to follow standard pre-processing and processing steps, using commonly computed metrics to extract information from our graphs. New insights on the plot, characters dynamics and books structures were found, demonstrating the potential of interdisciplinary research.
Among the measures computed, one observation can be made about the density. We compared the densities between our graphs assuming they have a comparable number of nodes. However, for better precision, one could have assessed the graph cohesion by computing the average degree of the network. This would allow to include also the entire graph (the one with all the data) in the comparison, as it has a significantly larger number of nodes than the others.
In addition to the measures computed, Structural Equivalence could also have been computed in addition to Assortative Mixing, in order to better study network structure in terms of similarity.
Lastly, we found in Sec. 5.2.1 our networks to be robust, as a consequence of being scale-free. Further studies can verify if the networks can truly survive the death (i.e. the removal of the node from the graph) of randomly chosen characters (being scale free, it is unlikely to pick up and remove high-degree nodes).

References

[1] *Gephi - The Open Graph Viz Platform*. `https://gephi.org/`.

[2] Ava Sadasivan. *Witcher Network*. `https://www.kaggle.com/datasets/avasadasivan/witcher-network`.

[3] *pandas - Python Data Analysis Library*. `https://pandas.pydata.org/`.

[4] *NetworkX*. `https://networkx.org/`.