Analyzing Character Networks in Portuguese-language Literary Works

Mariana O. Silva[1], **Gabriel P. Oliveira**[1], **Mirella M. Moro**[1]

[1]Department of Computer Science
Universidade Federal de Minas Gerais (UFMG) – Belo Horizonte, MG – Brazil

{mariana.santos,gabrielpoliveira,mirella}@dcc.ufmg.br

Abstract. *Literary works are complex narratives with multifaceted character relationships. Studying these relationships can reveal important insights into the story's structure and each character's contribution to the plot development. This research investigates character networks in Portuguese-language literature using two main analytical approaches: structural network analysis and character importance metrics. Our analyses emphasize the significance of character networks in understanding the narrative structure of literary works and reveal the intricate interplay between characters in Portuguese-language literature. These findings deepen our comprehension of literary works' fundamental structure and the characters' pivotal role in shaping the story.*

1. Introduction

Literary works are rich sources of information about human behavior and society, with complex narratives of relationships between characters. Such connections enlighten the base plot structure and the characters' roles in driving it forward. For example, in Machado de Assis' "*Dom Casmurro*", the protagonist's relationships with his childhood friend and love interest, *Capitu*, and his best friend, *Escobar*, are key to grasping the motivations and actions that shape the plot. Diving into these relationships through network analysis allows understanding of literary works' themes and social dynamics.

In recent years, network analysis has emerged as a powerful tool for studying these relationships and understanding the narrative structure of literary works due to its ability to visually represent the complex relationships between characters in a story [Aires et al. 2017, Li et al. 2019, Labatut and Bost 2019]. By analyzing these relationships, researchers can identify the central characters, explore their roles and relationships, and uncover the underlying themes and motifs of the work. This approach is particularly useful for studying the behavior and interactions of individuals within a society, portrayed through the lens of fictional characters.

Although network analysis has become a valuable tool for studying character relationships, there is a significant lack of research on character network analysis in Portuguese-written literature. Indeed, extracting and analyzing text from Portuguese literary works using automated methods is a complex task. Compared to English, supported by many natural language processing and text analysis tools, resources for analyzing lesser-spoken languages (such as Portuguese) are limited. Consequently, researchers often resort to translating works into English, which can be error-prone, particularly when translating from a language as rich and complex as Portuguese. For example, in "*Dom Casmurro*", the female character *Capitu* asks "Você jura?" ("Do you swear?"), to which

Bentinho answers "Juro" ("I swear"). Then, an automatic process translates *Juro* to *interest*; i.e., it confuses the verb *jurar* (to swear) in its first singular person with the noun *juro* (interest), and the original dialogue becomes nonsense.

However, while analyzing character relationships in Portuguese literature poses challenges, there is still an opportunity to investigate character relationships using semi-automated approaches (e.g., text extraction tools and manual coding of character interactions) to build character networks. We employ two main approaches to fill this research gap: structural analysis of networks and importance identification of characters. The former aims to reveal the underlying properties of the character network, including degree distribution, centrality measures, and global clustering coefficient. The latter is accomplished through node centrality metrics. Our analysis of character networks not only contributes to a better understanding of the underlying properties of character networks but also advances social network analysis by applying its techniques to uncharted territories, Portuguese-language literature, and its complex characters.

2. Related Work

Social networks (SNs) are powerful tools for unveiling relationships and patterns in large volumes of data [Fonseca et al. 2021]. One of the most relevant and challenging tasks in the literature context is the extraction of character networks, which represent the relationships between the characters of a story. Indeed, there are distinct approaches for extracting such relationships, and recent studies focus on developing automated pipelines for such a task [Shahsavari et al. 2020, Yang 2022]. Furthermore, analyzing character networks allows several applications, including summarization, classification of fictional novels, and role detection [Aires et al. 2017, Li et al. 2019, Labatut and Bost 2019].

However, most studies on character networks go over English-written literature (including translations). Processing content in other languages is extra challenging since the most used Natural Language Processing (NLP) methods are still designed for English. Even so, the number of works considering content in other languages is increasing. The Portuguese language is widely spoken in the world, with over 230 million native speakers,[1] and NLP methods have been used for several applications using Portuguese-language content, including regional reading preferences discovery [Silva et al. 2021a], detection of loanwords [Muhongo et al. 2022], and genre classification [Scofield et al. 2022].

To the best of our knowledge, our study is the first to build a social network of characters from Portuguese-language literary works. We use a structured dataset with public domain books in Portuguese and apply a specific methodology for character extraction and recognition. In addition, we build and characterize networks for all such works, which allows us to uncover their existing patterns. Therefore, we believe this work enriches the existing knowledge about Portuguese-language literature and contributes to further complex analyses of such works.

3. Methodology

To analyze character networks in literary works, we followed a generic methodology proposed by Labatut and Bost, which involves three main steps, as illustrated in Figure 1:

[1]Ethnologue: `https://www.ethnologue.com/statistics/`

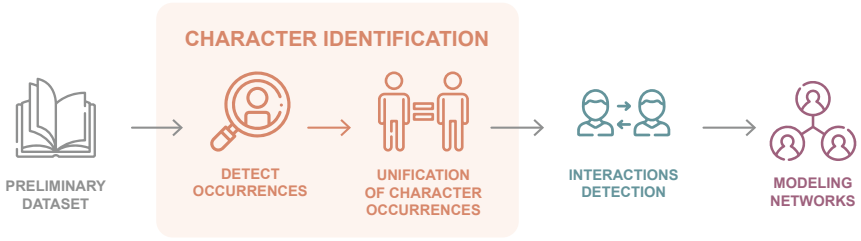

Figure 1. Overview of the generic character network extraction methodology.

(*i*) identifying characters, (*ii*) capturing their interactions, (*iii*) and extracting the network. Each step depends on the considered narrative's nature, the character network's planned usage, and methodological considerations [Labatut and Bost 2019]. Therefore, in this section, after we describe the dataset used (Section 3.1) and the text extraction processing (Section 3.2), we outline how we deal with each of the three methodological steps (Sections 3.3 – 3.5) and present the social network metrics considered to analyze the resulting character networks (Section 3.6).

3.1. Data

We use *PPORTAL*, a cross-collection dataset with metadata related to public domain Portuguese-language works [Silva et al. 2021b]. composed of three digital libraries for public domain works, mainly from Brazil and Portugal: Domínio Público,[2] Projecto Adamastor,[3] and Biblioteca Digital de Literatura de Países Lusófonos (BLPL).[4] To obtain a representative sample of works, we use a list of the top 100 literary books in Portuguese, ranked through voting by Goodreads users.[5] After cross-referencing this list with the books available on *PPORTAL*, we identified which books had a download link, resulting in 60 works that met such criteria. We also extract the list of characters for each book from external sources. To do so, we create a crawler to extract character lists from two different sources: Wikipedia and *Todo Estudo* website.[6] As a result, we extract character lists for 34 works out of the 60 downloaded books. The comprehensive list of the final set of works is available in the project repository,[7] including information such as title, author, number of identified characters, and external source.

3.2. Text Extraction Processing

We download all works in PDF format and use the *pdfplumber* Python library[8] to extract their text. Although *pdfplumber* is a powerful extraction tool, it is limited in identifying text paragraphs because, in PDF format, the text is often broken down into lines rather than paragraphs. Consequently, we employ additional text processing techniques to segment the text into separate paragraphs for further analysis. We break down the literary works into chapters, as they are a common way to segment books. However, not all works are explicitly divided. For instance, the work "O Alquimista" (by Paulo Coelho)

[2]Domínio Público: `https://www.dominiopublico.gov.br`
[3]Projecto Adamastor: `https://projectoadamastor.org`
[4]BLPL: `https://www.literaturabrasileira.ufsc.br`
[5]Top 100: `https://www.goodreads.com/list/show/366.Best_Literature_in_Portuguese`
[6]*Todo Estudos* provides summaries and information on literary books.
[7]`https://marianaossilva.github.io/DSW2021/`
[8]*pdfplumber*: `https://github.com/jsvine/pdfplumber`

has no type of textual structure to segment the chapters. In such cases, we divide the text into chunks of 100 lines, serving as a proxy for a chapter.

3.3. Character Identification

The first step in the character network extraction methodology is character identification, which consists of detecting characters that appear in the narrative and when exactly they appear. This stage is a critical process that heavily depends on the work's narrative structure, as it begins with the work of fiction itself. According to [Labatut and Bost 2019], character identification involves two substeps: to detect occurrences of characters in the narrative and to unify these occurrences (i.e., to determine which ones correspond to the same character). We further describe both steps as follows.

Detect Occurrences. Automating character identification is challenging, as characters can appear under three different forms: *proper nouns* (e.g., "Capitu"), *pronouns* (e.g., "she"), and *nominals*, which are anaphoric noun phrases that refer to characters (e.g., "Bentinho's wife"). While existing methods effectively handle the first form, detecting the latter two forms is often more complex [Labatut and Bost 2019]. To address this challenge, one simple approach is to use a pre-defined list of character names and perform exact matching [Aires et al. 2017]. Such a list may be constituted manually by the researchers themselves or through an external source such as the work's Wikipedia page.

As aforementioned, we adopt the second approach, but in a semi-automatic way. Using two external sources (Wikipedia and *Todo Estudo*), we implemented a crawler to extract the list of characters from each literary work. However, not all Wikipedia pages follow the same formatting. Therefore, a manual processing step was required to clean up and structure the character name lists. The output of this first step is a list for each work containing both names and descriptions of the characters. For example, the character Capitu from "*Dom Casmurro*" is described as *"Capitolina: chamada de Capitu. É o grande amor e esposa de Bentinho. Diferentemente do marido, vem de família pobre e se mostra inteligente e à frente de seu tempo."*[9]

Unification of Character Occurrences. The second step in character identification is unifying their occurrences, which is challenging as characters may appear in three different forms in text: proper nouns, nominals, and pronouns. Unifying such occurrences requires tackling the coreference resolution problem, where *coreference chains* representing the same character must be identified. Overall, most studies use some form of name clustering to perform alias resolution, where two factors are generally considered to determine that two aliases point at the same character: string similarity and gender compatibility. Another common approach is to use linguistic resources to associate close synonym nominals with the same character.

Our methodology addresses this unification step by extracting linguistic features from character definitions and generating a list of proper nouns associated with each character using the Python library *spaCy*.[10] *spaCy* allows us to extract linguistic annotations, including universal POS tags, which provide information about a word's lexical and grammatical properties, such as adjectives, nouns, pronouns, and proper nouns. Using this

[9] "Capitolina: called Capitu. She is Bentinho's great love and wife. Unlike her husband, she comes from a poor family and is smart and ahead of her time." (as automatically translated by translate.google.com)

[10] *spaCy:* `https://spacy.io/`

information, we create a list of proper nouns for each character based on their descriptions. We then compare these proper nouns with the names of other characters, removing any matches to obtain a list of nominal synonyms that are used to unify character occurrences. For instance, the nominal synonyms for the character Capitu include "Capitolina" and "Capitu". With the nominal synonyms, we can disambiguate and consolidate the different occurrences of each character in the text.

3.4. Interactions Detection

After identifying the occurrences of each character, the following extraction step involves detecting all interactions between each character pair in the narrative. These interactions can be either explicitly described or inferred from the narrative, depending on the definition of interaction. Co-occurrence is the most widely used due to its simplicity. This method breaks down the text into smaller narrative units and assumes two characters interact when they appear together within the same unit. Although simple, this approach has some limitations, mainly due to its imprecise nature. For example, two characters can appear together without interacting at all, such as when they are both spectators or one is mentioned in the absence of the other. Therefore, such a simple approach can be used only as a proxy for actual character interactions.

Regarding the text partitioning into narrative units, existing approaches rely solely on physical aspects, which can lead to arbitrary splits such as chapters, paragraphs, or even sentences. Hence, essential co-occurrences may be missed. One potential solution is to use smaller narrative units (i.e., one or more sentences) that are more natural and avoid such arbitrary divisions. Still, the length of sentences and paragraphs can vary greatly among writers, to the extent that one writer's paragraph could be shorter than another. Here, to detect interactions between characters, we partitioned each chapter into narrative units consisting of three sentences. For each of these units, we search for co-occurrences between pairs of characters, evaluating their names and nominal synonyms. We consider the combination of character pairs, not their permutation, to avoid duplicate results.

3.5. Modeling Networks

After identifying the characters and their interactions, the final step models the character network. This step requires making two methodological choices: defining the nodes and edges. Typically, characters are represented as individual nodes. Nonetheless, in some cases, one can group certain characters together and represent them as a single vertice. Such an approach makes sense when groups of people are mentioned indistinguishable, such as *Menores abandonados* in "*Capitães da Areia*", by Jorge Amado. However, dealing with such collective nodes can be challenging when their composition changes over time or when the same character appears as an individual and as part of a group.

Regarding edges, their definition is more complex, as one has to consider several aspects of the interactions, which must be translated into graph-related concepts: their laterality, score, polarity, and temporality. The general approach is to represent unilateral interactions by directed edges and bilateral ones by undirected edges or pairs of reciprocal directed edges. The scores computed to measure interaction intensity are modeled by edge weights, which can be signed to represent the polarity of the interaction (friendly vs. hostile). This results in graphs that can be (un)directed, (un)weighted, and (un)signed.

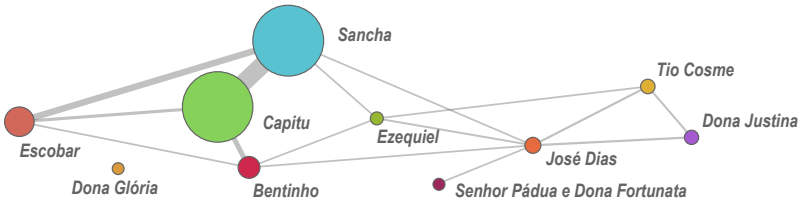

Figure 2. Character Network of "*Dom Casmurro*", by Machado de Assis

For the sake of simplicity, we model the character networks as unsigned, undirected, weighted networks. Specifically, each network is defined as a weighted graph $G(N, E)$, where N is the nodes set, and E is the edges. A node represents individual characters or groups, whereas the edges represent one or more co-occurrences between nodes. The edges are weighted to reflect the strength of the co-occurrence relationship between the two nodes they connect. This modeling provides a straightforward yet effective way to represent a story's relationships between characters and groups. For example, Figure 2 shows the character network of "*Dom Casmurro*".

3.6. Social Network Analysis

In the final step of the methodology, we use social network analysis metrics to characterize the networks of characters generated for each literary work. Such characterization allows an indirect bottom-up approach to reveal patterns and narrative aspects. Specifically, we conduct two analyses: *structural analysis* (Section 3.6.1) and *character importance* (Section 3.6.2). Each analysis relies on topological metrics extracted from the networks.[11] By examining these metrics, we can gain insight into the structure of the networks and the relative importance of each character within them.

3.6.1. Structural Analysis

In this analysis, we use the networks' structure to investigate which books have more interconnected character networks, i.e., more interactions between their characters. Further, the structural analysis may also reveal whether the books present different patterns of relationships. Therefore, we consider the following graph-related measures commonly used to describe the properties of a network.

Size. The number of network nodes, i.e., the number of characters of each book.

Density. The proportion between existing and all possible network connections (i.e., edges). The higher the density, the more interconnected the network. In other words, a network with high density means that the book's characters interact with many others.

Maximum Degree. The maximum degree informs the character with the most connections in the book, which possibly plays the main role in it.

Average Path Length. The average value of all shortest paths in the network.

Assortativity. It measures the tendency of nodes to connect to similar nodes with respect to their degree. In other words, a node with a high degree and high assortativity connects

[11] For formal definitions and formulas, see [Newman 2010, Barabási 2016].

mostly with other nodes with a high degree.

Global Clustering Coefficient. The average value for the nodes' clustering coefficients. The clustering coefficient is the tendency of a node's neighbors to connect to each other based on transitive relations. The higher its value, the more interconnected are the node's neighbors. Here, a node (character) with a high cluster coefficient has a high probability that its neighbors (i.e., the characters they interact with) also interact.

3.6.2. Character Importance

To assess the character importance, we compare the node centrality measures of the characters across the different networks to see if certain types of characters tend to be more important in one story than others.

Betweenness Centrality (BC). It measures node centrality based on the number of shortest paths that pass through it. A character with high betweenness has a high information flow and may be relevant to connect distinct parts of the network (i.e., character groups).

Eigenvector Centrality (EC). Measures node influence in a network based on the incoming edges. In other words, the eigenvector centrality for a node depends on the number of influential nodes linking to it. Therefore, if a character has a high eigenvector value in our networks, other relevant characters interact with them.

Closeness Centrality (CC). The intuition behind this metric is that central nodes are close to all other nodes. Thus, it is based on the distance to all other reachable nodes. Here, a node with high closeness is a character with connections to most of the story cores.

Degree Centrality (DC). The degree of a node is the number of nodes it connects with, i.e., the number of characters with whom they interact. This value is normalized by dividing it by the maximum possible degree in the network.

4. Results

This section showcases the outcomes of the two proposed analyses that aim to reveal novel insights and recognize shared patterns in Portuguese-language literature. We structure our presentation around each social network analysis, specifically the *structural analysis* (Section 4.1) and *character importance* (Section 4.2). The former aims to uncover the underlying structural properties of the character network, such as the degree distribution, centrality measures, and the global clustering coefficient, whereas the latter seeks to identify the most influential and critical characters in the story, using centrality metrics.

4.1. Structural Analysis

This section presents the structural analysis results, considering six network-based topological measures: network size, density, maximum degree, average path length, assortativity, and clustering coefficient. These measures may offer valuable insights into the underlying structural properties of the networks and help to better understand the complex relationships between characters in these narratives. To get an overview of the results, we plot the distribution of each metric in Figure 3. To distinguish potential outliers, we use the DBSCAN clustering algorithm [Ester et al. 1996], a density-based method that groups

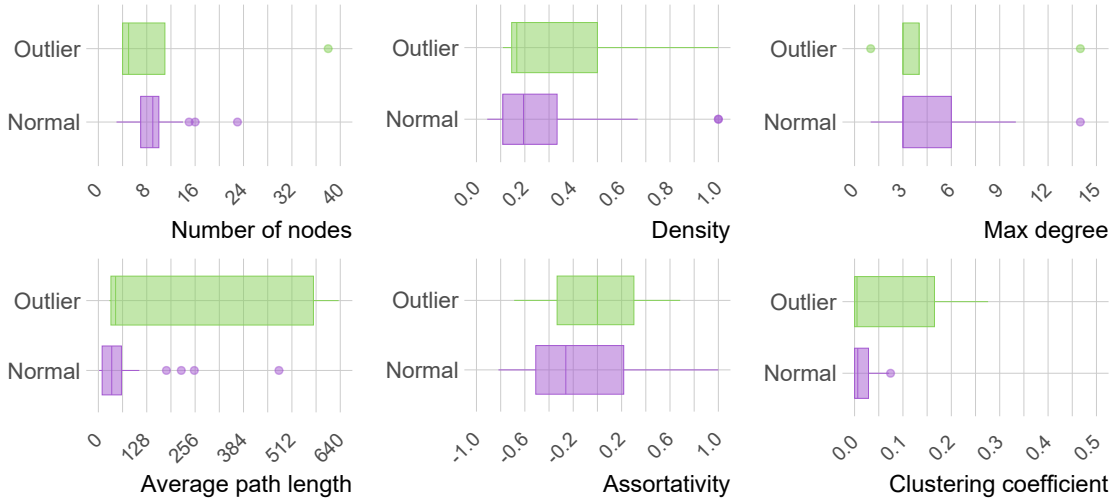

Figure 3. Distribution of the graph-related metrics for all considered works, divided into normal (bottom, purple) and outlier (top, green) groups.

Table 1. Examples of works from the normal and outlier groups. Colored cells indicate outliers values for the correspondent metric.

#	Title	Author	Lit. Mov.	#N	D	MD	APL	A	GCC	Group
1	O Guarani	José de Alencar	Romanticism	11	0.15	3	636	0.68	0.00	Outlier
2	Os Maias	Eça de Queirós	Realism	38	0.11	14	46	0.30	0.01	Outlier
3	O Crime do Padre Amaro	Eça de Queirós	Realism	5	0.50	4	29	-0.69	0.17	Outlier
4	O Ano da Morte de Ricardo Reis	José Saramago	Modernism	4	0.17	1	569	0.00	0.00	Outlier
5	A Morgadinha dos Canaviais	Júlio Dinis	Romanticism	4	1.00	3	33	-0.33	0.28	Outlier
6	Dom Casmurro	Machado de Assis	Realism	10	0.33	6	9	0.22	0.03	Normal
7	Capitães da Areia	Jorge Amado	Modernism	23	0.13	14	76	-0.33	0.01	Normal
8	O Alquimista	Paulo Coelho	Postmodernism	8	0.25	5	51	-0.51	0.04	Normal
9	Memórias Póstumas de Brás Cubas	Machado de Assis	Realism	16	0.07	4	30	0.30	0.00	Normal
10	Iracema	José de Alencar	Romanticism	11	0.11	3	478	0.02	0.00	Normal

Lit. Mov. = literary movement; #N = network size; D = density; MD = maximum degree; APL = average path length; A = assortativity; GCC = great clustering coefficient

data points that are closely packed and identifies any points that lie alone in low-density regions as outliers. Table 1 shows ten examples of normal and outlier works.

Network size (#N). Based on the total *number of nodes*, our analysis shows the network size varies little between literary works, with an average of nine characters. This slight variation may reflect authors' tendency to focus on a small group of characters to tell their stories rather than creating large character networks. It also suggests authors may prioritize in-depth description and development of a few characters rather than thinly spreading their attention across a large cast. On the other hand, the outlier group presented a much more significant variation in size, with an average of 12 characters and one work containing over 30 characters (Table 1, work #**2**), which suggests some authors may choose to provide a broader perspective or to illustrate the interconnectedness of a community or society. Yet, it is worth noting that these outlier works are the minority (14.71%), and the trend towards smaller networks is still prevalent in the considered literary works.

Density (D). Based on the proportion of possible edges present in the network, the density

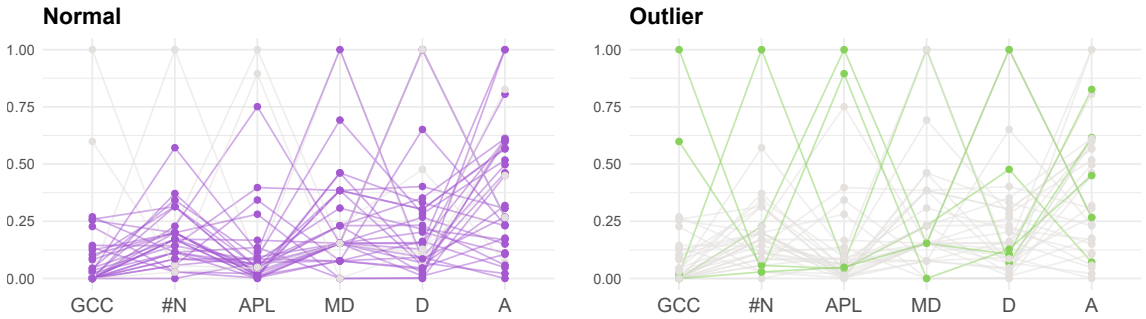

Figure 4. Summarized comparison of normal (left, purple) and outlier (right, green) works, based on six topological metrics.

of the networks is generally low for both normal and outlier groups. Such results indicate only a small fraction of possible interactions between characters are represented in the narrative. This finding suggests authors selectively include interactions between the most relevant characters in the story. One exception is the work **#5** (Table 1), whose character network has a density of 1, indicating that all characters are connected to each other.

Maximum Degree (MD). Based on the highest number of edges connected to a single node in the network, our analysis shows the MD varies across literary works for both groups. Some networks exhibit few highly connected nodes, known as hubs, while others have a more uniform distribution of edges between the nodes, indicating a more egalitarian network structure. On average, the networks have a maximum of two to four interactions between characters, but this can vary significantly depending on the complexity of the work and the number of characters involved. Both groups have outlier networks, with a MD over 10 (works **#2** and **#7**, Table 1), indicating a highly centralized network structure with one or a few dominant characters interacting with many other characters.

Average Path Length (APL). Based on the average shortest path between any two nodes in the network, our analysis shows the average path length is generally short, indicating that characters in the narrative are connected to each other through a few intermediaries. This finding is consistent with the small-world nature of social networks. By contrast, for the outlier group, we found a wider range of APLs, with some networks having a similar short APL as the normal group, while others have significantly longer APLs, indicating a more fragmented network structure (e.g., Table 1, work **#10**).

Assortativity (A). Based on the tendency of nodes to connect to nodes with similar characteristics, our analysis shows the networks are generally disassortative, i.e., highly connected nodes tend to be connected to poorly connected nodes. This finding suggests that characters are connected to characters with different degree characteristics, reflecting the diversity of social interactions in real-world networks.

Global Clustering Coefficient (GCC). Based on the tendency of nodes to form clusters or tightly knit groups, our analysis found that the normal group networks have a low clustering coefficient, indicating an absence of tightly knit clusters of characters within the s. This finding contrasts real-world social networks, which tend to exhibit a high clustering coefficient, suggesting that the structure of literary networks differs from that of social networks. In contrast, the outlier group networks exhibit a wide range of clustering

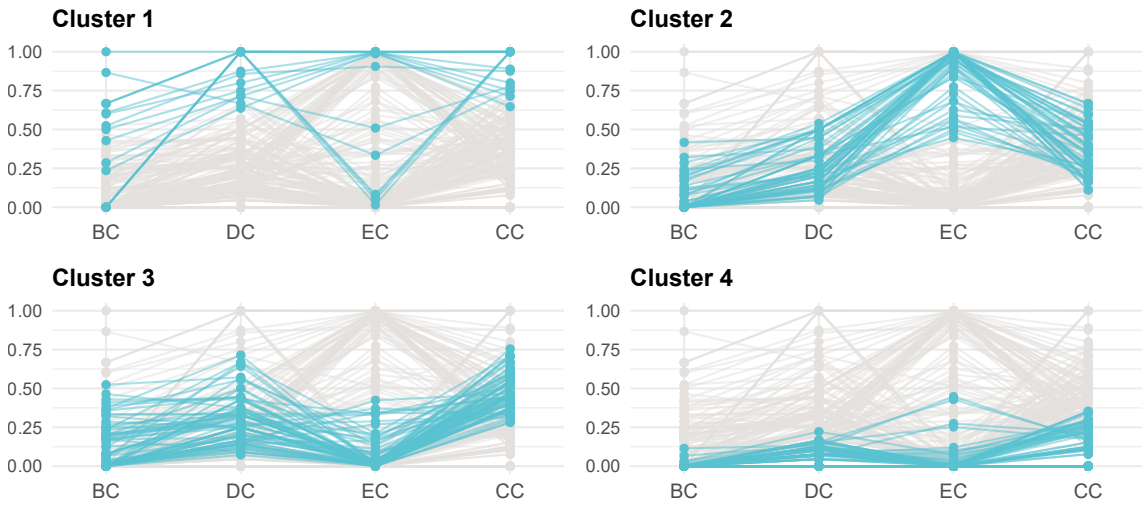

Figure 5. Summarized comparison between the identified clusters, based on four node-related measures.

coefficients, with some networks having a high clustering coefficient while others have a lower clustering coefficient, suggesting a more scattered network structure.

Overall discussion. As summarized by Figure 4, the structural analysis shows Portuguese-language literary works tend to have small and sparsely connected character networks, with an average of nine characters and selective inclusion of interactions between the most relevant ones. Additionally, the networks are highly centralized and disassortative and present a short average path length. Finally, most networks exhibit a low clustering coefficient, indicating a lack of tightly-knit clusters of characters.

4.2. Character Importance

In this section, we present the results of the second analysis, which focuses on character importance in literary works. To assess character importance, we consider four node centrality measures: betweenness, eigenvector, closeness, and degree centralities. We discard the outliers identified in the previous analysis to obtain more reliable results. We then cluster the characters of the remaining literary works based on the four topological metrics using the KMeans algorithm. We identify four clusters, as summarized in Figure 5 and discussed next.

Cluster 1: contains characters with medium betweenness values and high eigenvector, closeness, and degree centralities. These characters are highly connected and influential within the network and have significant personal power. They are likely to be central figures in the story's conflicts and struggles and may significantly impact the overall outcome of the narrative. Examples of characters in this cluster may include charismatic leaders, powerful wizards or sorceresses, or highly skilled warriors or spies. Two examples of characters in this cluster are *Alquimista*, in "*O Alquimista*", and *Professor*, in "*Capitães da Areia*" (Table 1, works **#8** and **#7**). *Alquimista* is a wise and mysterious character who guides the protagonist, *Santiago*, on his quest to follow a legend. Similarly, the *Professor* is a charismatic and influential character who can connect with the street children and help them survive in the harsh realities of their world.

Cluster 2: contains characters with low values of betweenness and degree centralities,

high values of eigenvector centrality, and medium values of closeness centrality. These characters are not well-connected to other characters in the network but are highly respected and influential within their own social circles. They may play important roles as advisors, mentors, or guides to the main characters. Examples of characters include *Capitu* in "*Dom Casmurro*" and *Sabina* in "*Memórias Póstumas de Brás Cubas*" (Table 1, works **#6** and **#9**). Capitu is the protagonist's love interest and plays a significant role in shaping the protagonist's perceptions and decisions, making her an influential figure within the story's social world. Similarly, despite not being a central character herself, Sabina serves as a guide and mentor to the protagonist.

Cluster 3: contains characters with low betweenness, eigenvector, and degree centralities but medium closeness centrality values. These characters have relatively few direct connections to other characters in the network but can quickly and easily communicate with others. They may serve as connectors or mediators between other characters who are more central to the plot. Examples of characters in this cluster could include sidekicks, assistants, or advisors to more powerful or influential characters and characters who occupy a neutral or independent position within the story's social hierarchy. Two examples of characters in this cluster are *Escobar*, in "*Dom Casmurro*", and *Poti*, in "*Iracema*" (Table 1, works **#6** and **#10**).

Cluster 4: contains characters with low values of betweenness, eigenvector, closeness, and degree centralities, indicating that they have relatively few connections to other characters within the network and who are not particularly influential or central within the network. These characters may not play major roles in the plot or may be secondary to the main actions of the story. Examples include *Dona Glória* in "*Dom Casmurro*" and *Fátima* in "*O Alquimista*" (Table 1, works **#6** and **#8**). *Dona Glória* is the protagonist's mother, but her role is mainly to provide background information about the family history and support her son's actions. Likewise, *Fátima* is a minor character who appears briefly and provides insight into the protagonist's motivations and beliefs.

Overall discussion. Our findings reveal significant variations in character importance, as identified by four clusters, each containing characters with specific characteristics and roles. While some characters were found to impact the plot directly, others had an indirect impact through their connections to well-connected characters. Our analysis shows each centrality measure provides complementary perspectives on character importance, enriching our understanding of literature's complex network of characters.

5. Conclusion

In this work, we analyzed networks of characters from Portuguese-written literary works through two main approaches: structural analysis of networks and importance of characters. Structural analysis results indicate networks are sparsely connected with selective inclusion of interactions between the most relevant characters. The literary works also show highly centralized networks, with an average path length generally short, indicating a small-world social network. The networks also tend to be disassortative, with a low clustering coefficient. Regarding the second analysis, our results show significant variations in character importance. We identified four clusters, each with specific characteristics and roles. The centrality measures provided complementary perspectives on character importance, highlighting the complexity of character networks in literature. Overall, this study

contributes to a better understanding of the underlying properties of character networks and advances social network analysis by applying its techniques to uncharted territories, such as Portuguese-language literature and its complex characters.

Limitations and Future Work. There are some limitations to our methodology. The difficulty in analyzing and processing Portuguese text may impact the accuracy of our results, and incomplete character lists from external sources may limit network analysis. Our generating nominal synonyms process may introduce some noise and inaccuracies when assigning proper names. Moreover, the co-occurrence approach used to identify character interactions has an imprecise nature by default. Still, we have worked with modern solutions for all such issues. Indeed, future work includes addressing these limitations by developing or applying more robust text-processing approaches to improve the quality and depth of our analyses.

Acknowledgments. The work is supported by CNPq, CAPES, and FAPEMIG, Brazil.

References

Aires, V. P. et al. (2017). Construção e análise das redes sociais de personagens dos filmes da franquia o senhor dos anéis. In *BraSNAM*, Porto Alegre, RS, Brasil. SBC.

Barabási, A.-L. (2016). *Network science*. Cambridge University Press.

Ester, M. et al. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. In *Procs. of KDD*, pages 226–231, Portland, USA.

Fonseca, B. et al. (2021). Social network analysis and mining: challenges and applications. In *BraSNAM*, pages 287–294, Porto Alegre, RS, Brasil. SBC.

Labatut, V. and Bost, X. (2019). Extraction and analysis of fictional character networks: A survey. *ACM Comput. Surv.*, 52(5):89:1–89:40.

Li, J. et al. (2019). Complex networks of characters in fictional novels. In *ICIS*, pages 417–420. IEEE.

Muhongo, T. et al. (2022). Detection of loanwords in angolan portuguese: A text mining approach. *Inteligencia Artif.*, 25(69):87–106.

Newman, M. E. J. (2010). *Networks: An Introduction*. Oxford University Press.

Scofield, C. et al. (2022). Book genre classification based on reviews of portuguese-language literature. In *PROPOR*, volume 13208 of *Lecture Notes in Computer Science*, pages 188–197. Springer.

Shahsavari, S. et al. (2020). An automated pipeline for character and relationship extraction from readers literary book reviews on goodreads.com. In *WebSci*, pages 277–286. ACM.

Silva, M. O. et al. (2021a). Exploring brazilian cultural identity through reading preferences. In *BraSNAM*, pages 115–126, Porto Alegre, RS, Brasil. SBC.

Silva, M. O. et al. (2021b). PPORTAL: Public domain Portuguese-language literature Dataset. In *SBBD DSW*, pages 77–88, Rio de Janeiro, Brazil. SBC.

Yang, F. (2022). An extraction and representation pipeline for literary characters. In *AAAI*, pages 13146–13147. AAAI Press.