Information Sciences 641 (2023) 119124

Contents lists available at ScienceDirect

Information Sciences

journal homepage: www.elsevier.com/locate/ins

Text characterization based on recurrence networks

Bárbara C. e Souza [a], Filipi N. Silva [b], Henrique F. de Arruda [c,d,e], Giovana D. da Silva [a], Luciano da F. Costa [d], Diego R. Amancio [a,*]

[a] Institute of Mathematics and Computer Sciences, University of São Paulo, PO Box 369, 13560-970, São Carlos, SP, Brazil
[b] Indiana University Network Science Institute, Bloomington, IN, USA
[c] CENTAI Institute, Corso Inghilterra, 3, 10138, Turin, Italy
[d] São Carlos Institute of Physics, University of São Paulo, PO Box 369, 13560-970, São Carlos, SP, Brazil
[e] ISI Foundation, Via Chisola 5, 10126, Turin, Italy

ABSTRACT

Several complex systems are characterized by exhibiting intricate properties that occur at multiple scales. These multi-scale characterizations are used in various applications. In particular, texts can be characterized by a hierarchical structure, which can be approached by using multi-scale concepts and methods. Here, we adopt an extension of the multi-scale, mesoscopic approach – hereafter referred to as a recurrence network – to represent text narratives, in which only the recurrent relationships among tagged parts of speech (subject, verb and direct object) are considered to establish connections among sequential pieces of text. The characterization of the texts was then achieved by considering scale-dependent complementary methods: accessibility and symmetry. To evaluate the potential of these concepts, we approached the problem of distinguishing between meaningful and meaningless texts and different literary genres (namely, fiction and non-fiction). A set of 300 books was considered and compared by using the above approaches. The recurrence network characterization was able to discriminate to some extent between real and meaningless and between the two genres assessed. Thus, our results indicate that recurrence networks are able to capture subtleties in book plots, suggesting that a similar methodology can be used in related networked applications.

1. Introduction

Several structures and dynamics in the natural, as well as in artificial, worlds involve several *scales* regarding space, time, etc. [3,50]. The successive partitioning of more generic, abstract concepts into new concepts and subgroups establishes a *hierarchy* of representations. As studied systematically in areas such as pattern recognition and artificial intelligence, hierarchies are inherently associated with specific scales. Hence, we have that human language is an intrinsically multi-scale system, in which the levels of generality, detail, abstraction, and specificity vary according to specific situations and demands while consolidating knowledge into written text, or while orally communicating between individuals.

While texts have been frequently studied from the perspective of word adjacency or proximity (e.g., [30,40]), the scale-dependent, hierarchical aspects of human language have motivated more systematic approaches capable of taking into account not only smaller scales (e.g., related to the composition of words) but also mesoscopic and macroscopic scales [7]. Thus, in addition to considering the more local interrelationship between words (e.g., by adjacency or proximity), it becomes important to systematically approach texts in terms of sentences, paragraphs, sections, chapters, and even book collections, and whole libraries related to specific themes or epochs.

[*] Corresponding author.
E-mail address: diegoraphael@gmail.com (D.R. Amancio).

https://doi.org/10.1016/j.ins.2023.119124
Received 19 April 2022; Received in revised form 3 May 2023; Accepted 5 May 2023

Available online 10 May 2023

In [4], the authors used a networked representation formed by the co-occurrence of words to address the problem of identifying authors' styles. Differently from the traditional approach based on the frequency of words, the authors introduced a hybrid approach taking into account two main factors: (i) the frequency of words; and (ii) the topological measures of complex networks. Interestingly, it was found that frequency- and topological-based approaches complement each other since both strategies yielded useful, complementary information to identify authors' styles. Similar word co-occurrence networks have also been used in related contexts [19]. While this approach captures some of the temporal/spatial narratives, the word-level approach can only extract syntactical and stylistic features of texts [13]. In addition, the traditional co-occurrence approach does not link similar words. Alternatively, in [44], the information about punctuation was included in the co-occurrence network, which allowed distinguishing different linguistic styles among authors. Also, statistics derived from punctuation have shown that it can play an important role in characterizing text and can be used for future applications of machine learning [24]. Furthermore, punctuation patterns have been found to vary across Western languages [45], emphasizing the relevance of punctuation in the analysis of textual data.

In [11], the semantic flow of texts was studied. In the proposed approach, nodes are sentences, and edges are established by taking into account the semantic similarity of the respective nodes. The authors found that the transition between semantical clusters can be used to discriminate between distinct styles. In particular, the semantical flow allowed discriminating philosophy from investigative books with an accuracy larger than 92%. In [13], the authors analyzed networks formed via paragraph semantic similarity. The authors found that this type of representation complements traditional word co-occurrence networks because paragraph networks can grasp the semantic features of texts. While the proposed paragraph network can go beyond syntax/style, the narrative temporal aspect is not taken into account since paragraph order is not taken into account while creating the network.

Recently, methods based on Neural Networks (NN) have been used to obtain dense representations to classify texts and documents [17,2,49]. For example, techniques such as word2vec [33] and doc2vec [25] can capture short-range relationships between words in a small window, which are used to classify sentences or whole documents. More sophisticated methods, such as BERT [14] and sentence-BERT [38], can capture larger-scale relationships across sentences in the text. However, they are still limited to a few hundred tokens, thus insufficient to capture the full content of a book. In addition, new NN methods have been developed to embed networks and their elements, such as node2vec [21] and graph attention networks [47]. When combined with a non-linear dimension reduction method, such as UMAP [32], they are used to obtain visualizations of large networks [10,8]. The internals of such approaches are intrinsically related to force-directed layouts, and methods that explicitly combine the two concepts have been proposed in the literature [36].

In the present work, we aim at studying texts, more specifically books from the Gutenberg project [20], from (i) the mesoscopic perspective of linear sequences along the text, as well as (ii) the accessibility and symmetry [43,46] of texts when represented as recurrence networks.

The first above-mentioned approach involves treating the text as a linear sequence of paragraphs while identifying tf-idf [27] cosine similarity between the obtained paragraphs along the whole sequence. A way to capture mesoscopic relationships in complex systems is through the use of recurrence networks [15]. In texts, such networks can be constructed from short and long-range relationships derived from the text narrative, e.g., by adjacency and content similarity. These relationships, as gauged by the co-occurrence of words, are likely to indicate recurrent situations in space, time, or subject along the narrative. For instance, a location may recur along the text, giving rise to several respective longer-range links. This same effect can occur with characters.

Different from other approaches that focus on local (or co-occurrence) similarity (e.g., [48,5]), our approach can capture long-range references in the text narrative. The limitation with approaches based on word adjacency concerns the fact that single words are typically not enough to characterize a well-defined context. By employing paragraphs (or even larger portions of the text), it becomes possible to establish correspondences between paragraphs referring to the same context, such as a situation, place, character, etc. Given that these paragraphs can be far away from one another (e.g., in different chapters), *long-range* connections can be obtained.

The second approach integrated into the current analysis concerns the estimation of the accessibility and symmetry of each of the nodes in paragraph-based networks. The accessibility was proposed as a means to quantify how effectively, according to specific dynamics taking place on a network, other nodes can be accessed by a given network node [46]. Interestingly, this measure intrinsically incorporates means for investigating multi-scale relationships in the analyzed data, which is achieved by varying the order of the considered neighborhood [46]. The accessibility approach depends on preliminarily established transition probabilities between the network nodes, such as defined by random walks. Being closely related to accessibility, the symmetry approach [43,46] aims at identifying topological symmetries or connections established along successive neighborhoods of each node. The symmetry concept involves two related components, namely the backbone and the merged. In the former case, the interconnections between nodes in the same neighborhood of a node are not taken into account, while in the latter the nodes belonging to the same neighborhood are subsumed.

Several interesting results are reported and discussed in the present approach, including the identification of the potential of the accessibility and symmetry for distinguishing between real and meaningless texts and between distinct literary genres, suggesting that these scale-dependent measures are capable of quantifying the heterogeneity of the narrative.

The current work starts by presenting the basic concepts and methods adopted and proceeds by presenting the application of the described measures extracted from the proposed recurrence network respectively to discriminate between real and meaningless texts and between literary genres. Finally, we compare the results yielded by the proposed with other known approaches from the literature. The source code is publicly available at: https://github.com/giovanadanieles/recurrenceNetworks.

Table 1
Processing example of one paragraph of the book “*The Arabian Nights*”.

Input text	Output text
The stories in the Fairy Books have generally been such as old women in country places tell to their grandchildren. Nobody knows how old they are, or who told them first. The children of Ham, Shem and Japhet may have listened to them in the Ark, on wet days. Hector’s little boy may have heard them in Troy Town, for it is certain that Homer knew them, and that some of them were written down in Egypt about the time of Moses.	*listen tell boy be it know who homer woman grandchild child hear story nobody place*

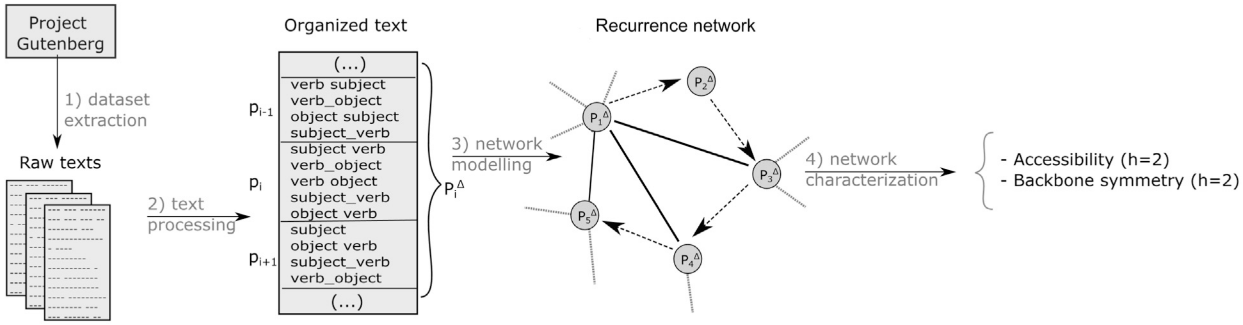

Fig. 1. Diagram of the execution pipeline of the methodology proposed. At first, the dataset is extracted from Project Gutenberg. The pre-processing of the raw text yields the organized text (O), which is then used to generate the recurrence network. Finally, several measures are extracted from the network to characterize it.

2. Materials and methods

This section describes the procedure employed to obtain (and characterize) networks from any text, which includes books and any other documents structured with paragraphs. The adopted pipeline is illustrated in Fig. 1. The procedure is described as follows:

1. *Dataset extraction*: at first, 300 books are extracted from Project Gutenberg in raw text format;
2. *Text processing*: in this step, we are interested in analyzing the relationship between specific words. For example, we aim to identify recurrent behavior by the same subject. For this reason, syntactic parsing is applied in order to identify such relevant words. In this step, words conveying low semantic meaning will be disregarded, as we are only interested in subject, verb and direct object, the most relevant ones to characterize the narrative semantic flow, or, in other words, the story constructed within the text. An example of this processing for the book “*The Arabian Nights*” is shown at Table 1;
3. *Network modeling*: here, a semantic representation of the narrative flow is created, where nodes represent a sequence of paragraphs and edges are established according to the semantic similarity between the nodes. The text is split into paragraphs based on the presence of two line breaks (a standard format in the Gutenberg raw text files). Then, to create the recurrence network for that book, the text is separated into paragraph windows, and each of them is mapped into a node in that network.

The computational expenses associated with the proposed methodology are detailed in Appendix A. Additional aspects of the adopted procedure are described as follows.

2.1. Dataset and genre classification

The dataset used in this project consisted of 300 different books. They were all retrieved from a random selection of Project Gutenberg [1]. The selection considered only books written in the English language written between 1000 and 2000 paragraphs. This selection was not specific to any other aspect of the books, i.e., author, publication date, literary genre, etc.

The Gutenberg Project provides open access to books written in several different languages. For each book, in addition to the full content, additional metadata are provided. This includes author, illustrator, title, language and associated subjects (which will be referred to as literary genres in this document). In this study, we retrieved, for each book its content in raw text format together with its set of genres.

An important aspect of this dataset is the lack of a well-defined classification for a book’s genre. Since Gutenberg Project provides a set of genres without any distinction of importance, there is no straightforward way of assigning one specific genre to each book. This issue is also emphasized by the fact that there is not a single granularity for the book’s genres other than a few general ones such as *PR (Language and Literature: English literature)* or *PZ (Language and Literature: Juvenile belles lettres)*, as well as particularly specific ones, such as *Scarecrow Fictitious character from Baum* and *National Research and Education Network Computer network*. Another problem is that a single book can be labeled with several of these non-informative labels.

So that the reader can further understand this scenario, in Fig. 2 we show the distribution of the 40 most frequent genre labels that appear in the dataset and their number of occurrences among the 300 books in question. It becomes clear how non-informative

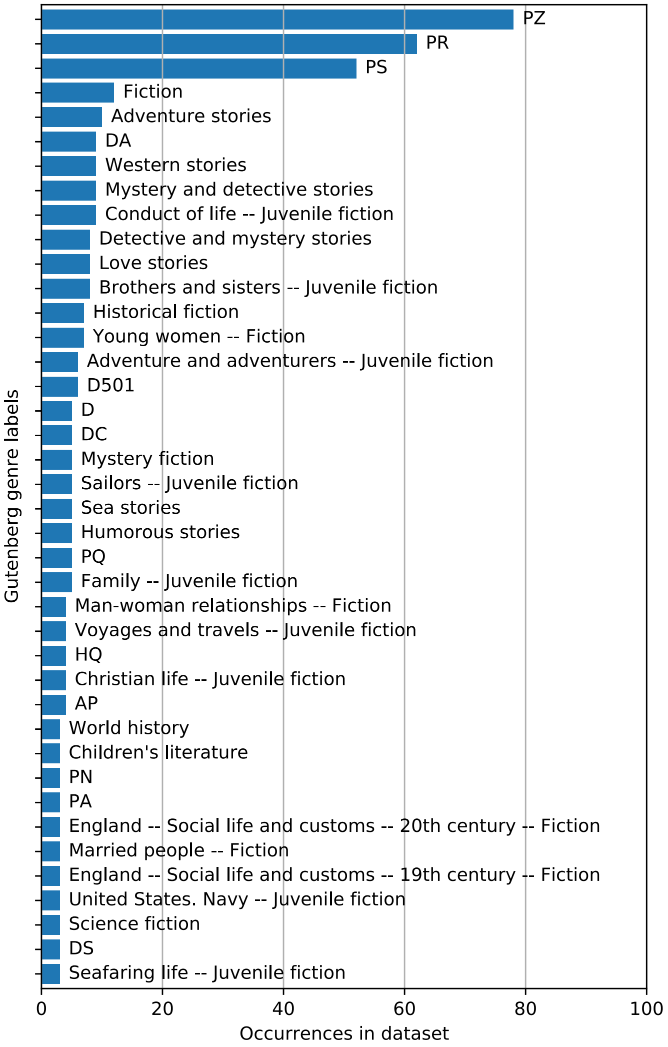

Fig. 2. Distribution of the top 40 genres in the dataset. The y- and x-axis show, respectively, the genre label and the number of times the label appears in the dataset.

the genre labels can be and how most of them are too specific (less than 5 occurrences overall). For example, the book "*Salute to adventurers*", a historical adventure novel, has the following genre labels: '*Jamestown (Va.) – Fiction*', '*Glasgow (Scotland) – Fiction*', '*Virginia – History – Colonial period, ca. 1600-1775 – Fiction*', '*Historical fiction*', '*Scottish Americans – Virginia – Fiction*' and '*PR*'.

2.2. From texts to networks

When tackling any sort of problem from a real-life system, it is essential to consider an appropriate scale to develop a solution that fits the phenomena being studied. Consider, for illustration purposes, any natural environment: there are numerous problems that demand an overall view of the macro aspects of that system. Take weather forecast as an example: to analyze and predict the weather of that specific system, one must consider features such as location, vegetation, humidity, and others, while the micro aspects such as each individual animal in the fauna or plant in the flora are not as important for the problem in question. It is very common to observe problems that focus either on the macro or the micro scale of the system. However, there is still a broad scope of intermediate scales between these two extremes that are yet to be explored.

In the context of text analysis, this issue is also present: while there are many different works that focus on the microscopic scale of the texts (e.g., word adjacency networks) or on the macroscopic scale (e.g., citation networks), there are still substantially fewer works that are placed anywhere between the two extremes.

Only more recently, there has been an increasing interest in systematically applying techniques to form networks from documents, while taking into account their mesoscopic structure [13,11]. For example, in [12], the authors employed image analysis techniques to network visualizations to extract topological features of mesoscopic networks for authorship recognition. Similarly, in [31], the authors leveraged the mesoscopic network's ability to capture a narrative's story flow and the author's "calligraphy" to address

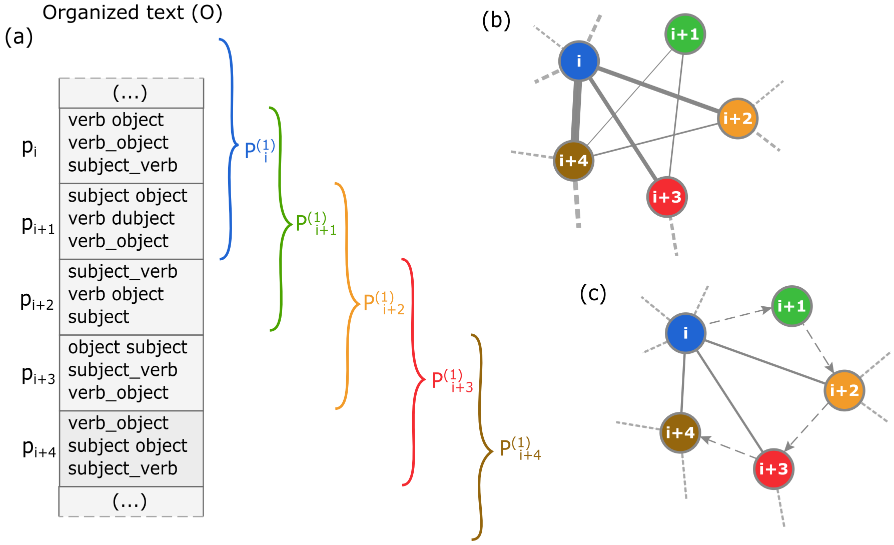

Fig. 3. Illustration of the presented methodology on how to construct the recurrence network based on the organized text O (a). Initially, there is an almost fully connected weighted network (with the exception of the edges between nodes representing adjacent paragraph windows), where the edge weights are the cosine similarity calculated between both nodes and are illustrated in (b) by the width of the lines. Next, only the $|V| \times T$ strongest connections are maintained to ensure that the average degree of the graph is now equal to T. Finally, the sequence edges are added between consecutive vertices, as illustrated by the dashed arrows in (c).

the authorship recognition problem. Lastly, in [7], mesoscopic networks were used to analyze and study texts, while taking into consideration the overall narrative and its temporal evolution.

As an extension of the concepts presented in the previously mentioned works, here we focus on these mesoscopic aspects of the texts, often overlooked by more traditional approaches, to propose a new technique to model networks from texts, namely the *recurrence networks*. By looking at texts from a mesoscopic perspective, it is possible to grasp the semantic context of a narrative, which can potentially be used to tackle a variety of different problems, such as genre classification and authorship recognition. This methodology, summarized in the pipeline shown in Fig. 1, will be explained in detail hereon.

Differently from other typical approaches, here we do not start with punctuation or stopwords removal, since those are important for the employment and performance of the syntactical dependency analysis method. However, it is important to note that stopwords will eventually be removed as well, but at the end of the text processing pipeline, after syntactical analysis. Hence, the only cleanup performed at this point is the removal of underscores ("_") and chapter markers, since they can introduce noise to the syntactic analysis that will be performed next. With the same goal, a co-reference resolution technique is applied [29]. Given that co-references are expressions that refer to a previously mentioned entity in the text (e.g., pronouns), this technique ensures that there is a minimum occurrence of multiple terms referring to the same entity.

After this pre-processing step, syntactic analysis is employed, where each paragraph is reduced to a set of tokens with specific syntactic roles: either *subject*, *verb* or *direct object*. This role selection was chosen with the intention of getting only the tokens that provide the most contextual meaning to the sentences where they are found while excluding any possible noise. This strategy was based on the fact that, in order to understand the semantics of something, it is usually necessary to answer the questions *what is happening* and *what or who is doing it*. To obtain an answer to the latter, we retrieve the subject in the sentence, which is the entity responsible for the activity happening. Secondly, when recovering the verb and its direct object, we are referring to the action occurring and, therefore, answering the first inquiry. This approach is also supported by the fact that the incorporation of linguistic knowledge can contribute to text summarization [34], and, therefore, to capture the contextual meaning of texts by enhancing informativeness. In this work, we used the CoreNLP parser to handle this processing [29].

At last, each of these tokens is normalized to its canonical form by using a lemmatization technique [28], resulting in the disregard of inflections in verbal tense, number, case or gender. Therefore, the sentence "*thought Alice to herself*", for example, will first be transformed to "*thought Alice to Alice*", then to "*thought Alice*" and, finally, to "*think Alice*". By that, the semantics of the sentence is actually summarized in the final set of tokens, where it is possible to answer *what* is happening (by the verb *think*) and *who* is taking that action (by the subject *Alice*).

After all this processing, the resulting organized text is denoted as O. O is composed of a sequence of paragraphs $(p_0, p_1, p_2, \ldots)$, with each paragraph comprising a sequence of words $(w_{i0}, w_{i1}, w_{i2}, \ldots)$. To construct the recurrence network, a sequence of paragraphs $(p_{i-\Delta}, p_{i-\Delta+1}, \ldots, p_i, p_{i+1}, \ldots, p_{i+\Delta})$ is mapped into a node i, resulting in $P_i^{(\Delta)}$. The process of obtaining paragraph windows $P_i^{(\Delta)}$ for their corresponding nodes in the recurrence network is illustrated in Fig. 3 for the case where $\Delta = 1$.

Next, to create the edges, we employ the techniques *bag-of-words* followed by the $\mathrm{tf{-}idf}$. At first, we use the bag-of-words to represent the text from now on, since only the tokens are important now and not the actual text structure. Then, we apply the tf-idf technique considering a text (O) as a collection of paragraphs and calculating the score for each of the paragraph windows, $P_A^{(\Delta)}$, which represents node A in the network representation. Finally, cosine similarity is used to calculate $sim(P_A, P_B)$ from the tf-idf

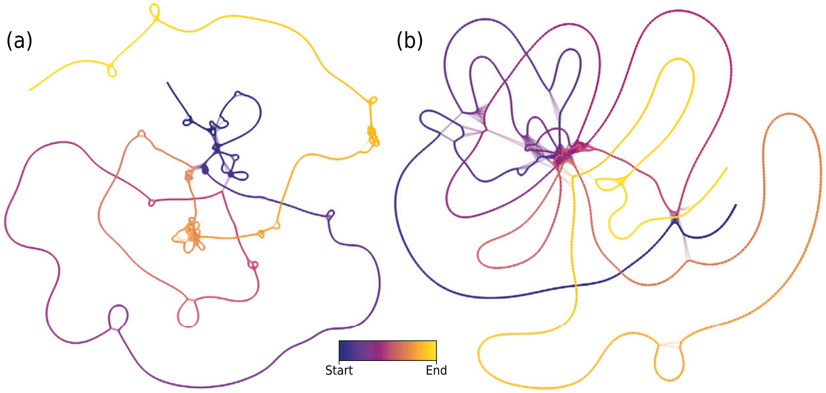

Fig. 4. Visualizations of recurrence networks obtained for the books "*The Arabian Nights Entertainments*" (a) and "*Salute to Adventurers*" (b). The colors indicate the sequence of nodes and the positions of the nodes were generated by the FR layout algorithm [18].

scores, between nodes of indexes A and B, where $|A - B| > \Delta$. This second restriction is due to the fact that, otherwise, the two paragraph windows would share at least one paragraph and, therefore, the similarity computed between them would be biased.

As a result, a weighted network is created (see Fig. 3(b)), in which the edge weights correspond to the similarity $sim(P_A, P_B)$ normalized between 0 and 1 among each pair of nodes. The final recurrence network is obtained by pruning the weakest connections until the average degree of the network reaches a specified threshold T. After this procedure, edge weights are ignored, resulting in an unweighted network (see Fig. 3(c)) with a fixed average degree greater than 0. All of these edges will be referred to as *similarity edges* hereafter. In addition, $|O| - 1$ edges are inserted linking nodes that represent adjacent paragraphs, that is, P_1^Δ will be linked to P_2^Δ, P_2^Δ to P_3^Δ and so on, as illustrated in Fig. 3(c) by the dashed edges. These edges are marked as *sequence edges*.

Such edges will guarantee that the network is a connected component. In addition, sequence edges provide a temporal narrative perspective, as it happens in traditional word adjacency networks [6,23]. With this, it is also possible to visualize the network as a sequential, continuous line, representing the text's narrative, as shown in Fig. 4, with the visualizations for the networks obtained from the books "*The Arabian Nights Entertainments*" and "*Salute to Adventurers*". The colors indicate the order of the nodes along the book and their positions are determined by applying the FR algorithm [18]. To render the network, we employed the Helios-Web software [41]. We opted to use the FR algorithm as it is known to preserve the topological structure of simple networks. Techniques based on neural network embedding could also be used in the case of larger networks.

2.3. Network topology measures

To understand and discriminate a network's topology, it is essential to be able to extract measures from the graph that reflect the desired properties. The accessibility measure [46], for example, was proposed to quantify the number of effectively accessible nodes given an established distance while respecting specific dynamics. In that manner, the accessibility can measure how peripheral or central a node actually is, which can be useful to reflect the network topology as a whole when considering all of its nodes individually. In addition, accessibility-based measures can also identify relevant words in texts, being thus useful to detect keywords and discriminate authors [6].

The definition of this measure is based on the concept of random walks (or any other dynamics) and concentric levels. In a random walk, there is an agent that moves between the nodes of a network through its edges. One variation of this definition is the self-avoiding random walk, in which the agent cannot go through the same node more than once. This kind of walk is the one used to define the accessibility measure. Besides, the concentric level h of a node i is defined as the set of nodes that are at a distance h when departing from i. Moreover, it is possible to define the probability vector $p_i(h) = \{p_1^{(h)}, p_1^{(h)}, \ldots, p_{N_i(h)}^{(h)}\}$ of reaching each one of the $N_i(h)$ neighbors of i in its concentric level h, when considering a self-avoiding random walk. Hence, the accessibility value k for a node i and a concentric level h can be calculated as:

$$k_i(h) = \exp\left(-\sum_j p_j^{(h)} \log p_j^{(h)}\right) \tag{1}$$

Unlike the node degree, the accessibility measure considers how many nodes can be effectively accessed, given the probability vector $p_i(h)$. In Fig. 5, we show two examples of the accessibility calculation for the node i, in red. Considering the first concentric level $h = 1$, in green, the calculation is trivial, since, by definition, it is actually the concentric degree. However, for the second level $h = 2$, it is already possible to note a difference between the two values. In Fig. 5(a), the probability vector of reaching each neighbor of i is more uniform, which reflects in the high value of accessibility obtained, approximately 9.9. The magnitude of this value can be confirmed by considering that the theoretical maximum by definition is 10, the number of nodes in that concentric level. In contrast, for Fig. 5(b), one can see that the accessibility value for $h = 2$ is considerably smaller. That can also be explained by the probability vector of reaching the neighbors of i, which shows a greater discrepancy between each of its values.

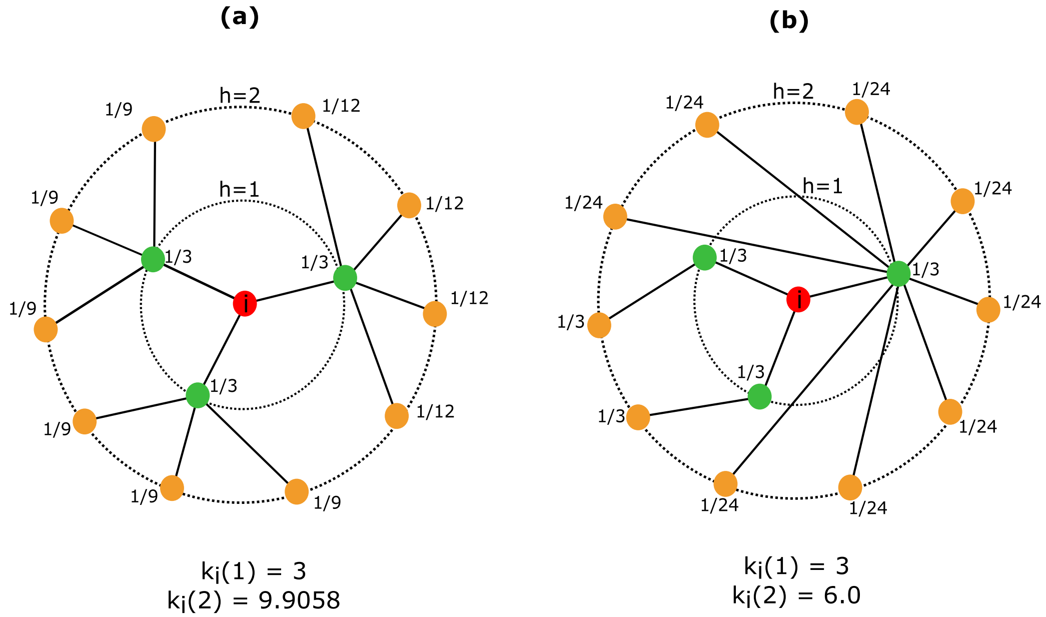

Fig. 5. Examples of the accessibility calculation for node i (in red). By definition, for $h = 1$, the accessibility value is the actual concentric degree and is trivially obtained for both networks. In (a), given the network topology, the nodes in the second concentric level have nearly the same probability of being reached, hence the high value obtained for the accessibility measure (considering that the maximum would be 10). In (b), since the discrepancy of the probabilities is greater, the value obtained for accessibility is smaller.

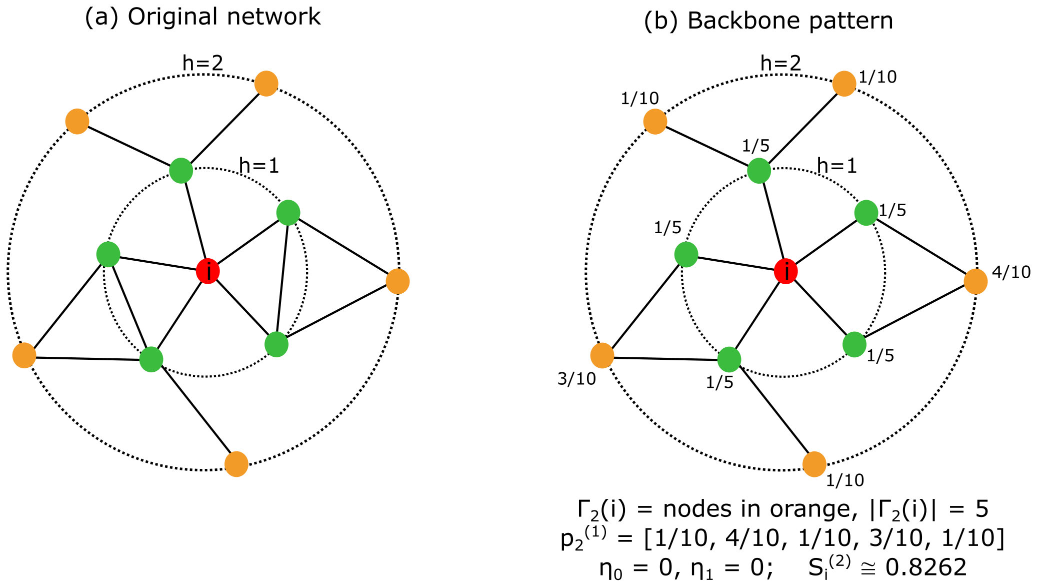
(a) Original network
h=2
h=1
i
(b) Backbone pattern
h=2
1/10
1/10
1/5
h=1
1/5
1/5
4/10
i
3/10
1/5
1/5
1/10
$\Gamma_2(i)$ = nodes in orange, $|\Gamma_2(i)| = 5$
$p_2^{(1)} = [1/10, 4/10, 1/10, 3/10, 1/10]$
$\eta_0 = 0, \eta_1 = 0; \quad S_i^{(2)} \cong 0.8262$

Fig. 6. Example of the symmetry calculation for node i (in red) and level $h = 2$. In (a), the original network is shown. In (b), one can see the backbone pattern for the original network, the one that will be used to calculate the symmetry value. We also have the parameter values used for the calculation and the final symmetry value.

From the concept of accessibility, concentric levels and probability vectors when considering random walks, the authors in [43] also derived the definition of concentric symmetry, which will also be useful for this work. The *backbone pattern of the concentric symmetry*, which is the one considered hereafter, is created by removing the edges between nodes in the same concentric level, as shown in Fig. 6. In their work, the authors also follow up by defining the merged pattern, which consists of merging together the nodes that were connected by these removed edges. However, the usage of the merged pattern for our classifications yielded no relevant improvement when compared to the backbone pattern, and, therefore, we will only consider the latter for this work for simplicity's sake.

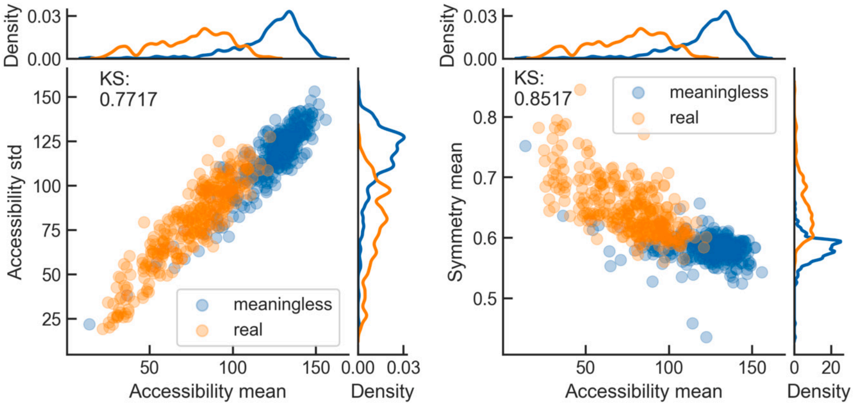

Fig. 7. On the left, discriminating between real and meaningless texts using accessibility's mean and standard deviation extracted from the recurrence network, with a high KS, equals to 0.7717. On the right, the same discrimination task, but now using accessibility and backbone symmetry means, also with a high KS, equals to 0.8517.

The symmetry value can be calculated using the equation:

$$S_i^{(h)} = \frac{\exp\left(-\sum_{j\in\Gamma_h(i)}(p_{ij}^{(h)} \times \log p_{ij}^{(h)})\right)}{|\Gamma_h(i)| + \sum_{r=0}^{h-1} \eta_r}, \tag{2}$$

where η_r is the number of nodes that are not connected to the next concentric level $(h+1)$, $\Gamma_h(i)$ is the set of neighbors of i that also belong to level h, and $p_i^{(h)}$ is the probability vector of node i considering the concentric level h. The specifications of these parameters can also be found in Fig. 6, alongside the symmetry value calculated for the network in question.

Accessibility and symmetry were chosen due to their effectiveness in characterizing networks across various applications. These measures are able to capture the heterogeneity of the network by considering the existence of *jumps* from one region of the network to another. Moreover, they are considered multi-scale, that is, it is possible to adapt the path length considered to reach further or closer regions of the network, which is also important for the text characterization problem in question. In addition, such measures capture mesoscopic-scale characteristics while remaining independent of network size [31]. Other local properties, such as node degree and clustering coefficient, are unsuitable due to the former being controlled and the latter being heavily influenced by adjacency-created links. Global measures based on shortest path lengths are dependent on size, which introduces further complexity. Nonetheless, size-independent alternatives, such as degree assortativity may offer potential improvements and merit exploration in future work.

3. Results and discussion

In this section, we present the results obtained from our study. Those results are divided into three main experiments: discriminating between real and meaningless texts, distinguishing between different genres, and a comparison of the proposed methodology with other approaches present in the literature. In the following, there is a subsection dedicated to each of these experiments.

3.1. Discrimination between real and meaningless texts

The first experiment performed was meant to evaluate how well the proposed methodology was able to grasp the sense of narrative from a text. For that, we extended the dataset by considering its shuffled version for each book. The shuffled version of a book is obtained by randomizing the order of the paragraphs of the text without modifying anything inside any of the paragraphs. Therefore, even though we maintain each sentence's syntactical and semantic structure unharmed, the sense of narrative for the whole text is lost since there is no meaningful storyline anymore. Then, to discriminate between real and meaningless texts, we considered three different measures extracted from the recurrence network:

- mean of the accessibility for the second concentric level of every node: $mean_k = mean(\{k_1(2), k_2(2), ..., k_n(2)\})$
- standard deviation of the accessibility for the second concentric level of every node: $std_k = std(\{k_1(2), k_2(2), ..., k_n(2)\})$
- mean of the backbone symmetry for the second concentric level of every node: $mean_S = mean(\{S_1(2), S_2(2), ..., S_n(2)\})$

Since the network's topological characteristics are similar to the small-world model, considering too high concentric level values would cause an overall sweep of the network at once, and it would not be possible to properly discriminate the nodes within the network. Therefore, smaller concentric levels are a better fit for the problem and, after empirical tests, $h = 2$ yielded the most satisfactory results in the experiments.

Fig. 7, on the left, displays the obtained results when considering $mean_k$ and std_k to discriminate between real (in orange) and meaningless (in blue) texts. It is visually noticeable that the two categories are separately clustered in the plot, indicating that our approach can successfully discriminate between them. We also consider the Kolmogorov–Smirnov score (KS) to compare the

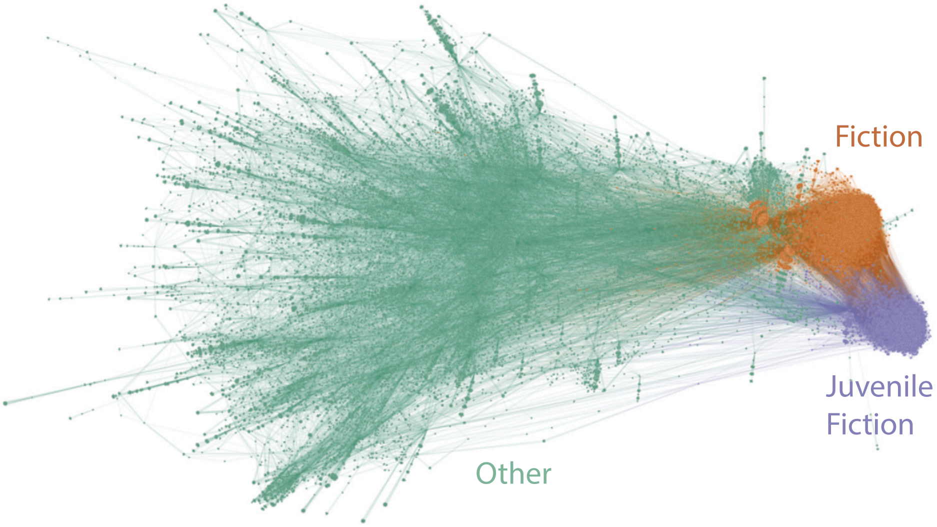

Fig. 8. Visualization of the projection network onto genres for the whole Gutenberg dataset, colored by the community detected. Generated using the *Networks3d* software [42].

two distributions [16]. In this scenario, we want the KS score to be as high as possible, indicating that each of the distributions is sufficiently different.

Additionally, the plot on the right of Fig. 7 displays the obtained results when considering $mean_k$ and $mean_S$ to discriminate between real (in orange) and meaningless (in blue) texts. Again, by observing the plot and evaluating the KS score obtained for the aforementioned measures, it is possible to conclude that this set of measures can also successfully discriminate between real and meaningless texts, therefore indicating that the measures extracted from the recurrence network can, indeed, capture the narrative of the text.

3.2. *Genre discrimination*

The second experiment performed aimed at verifying how much the proposed methodology was capable of discriminating books respectively to literary genres. To make that possible, first, we had to establish a strategy to create the genre labels for the dataset, given the non-trivial nature of the dataset, previously discussed in section 2.1.

To gain a deeper insight into the relationship between literary genres and labels offered by the Gutenberg Project, we created a bipartite network that establishes links between books and their respective genres. This bipartite graph was then projected onto the genres, resulting in a network in which each node represents a distinct literary genre, and a connection is established between two genres whenever there is at least one book containing both. This method led to the creation of a network comprising around $40,000$ nodes. We present a visualization of this network in Fig. 8.

Since it was impracticable to deal with a set of that many labels, we needed to increase the granularity level of the genre set. In order to cluster the set of genres into these broader labels, the Louvain community detection algorithm [9] was applied. Three main communities were found. After an empirical analysis, we found that the first one, in orange, comprises mostly genres related to adult fiction, while the second one, in purple, consisted mostly of juvenile fiction. Lastly, the third and most sparse one, in green, contains all the other possible genres in the Gutenberg Project, including history, art, bibliographies, etc.

Finally, we can refer to these communities in order to define a label for a given book. Since every literary genre can now be associated with a single community in the network, it can also translate to a single label. Therefore, one can assign a label to any book by choosing the community that comprises the majority of that book's genres.

Given that there are three main communities in the genre network: *Adult Fiction*, *Juvenile Fiction* and *Others*. Here, we make use of these communities and the genres set provided by the Gutenberg Project to define one specific label for each book, regarding its literary genre. For a book to be labeled into the *fiction* group, the majority of its genres listed in the Gutenberg dataset must belong to either of the fiction groups. Otherwise, the book will be labeled *others*. The two fiction communities have been merged in order to provide a more balanced and representative set. For this task, we also considered the same three measures mentioned earlier: $mean_k$, std_k and $mean_s$.

The chart on the left of Fig. 9 displays the obtained results when considering $mean_k$ and std_k to discriminate between *fiction* (in orange) and *others* (in blue). Even though the separation of the groups is not trivially observed in the chart as before, the KS values obtained are sufficiently high to indicate that the measures are considerably discriminating between the two groups. Furthermore, density plots for each of the measures also point to a significant distribution difference between them.

Additionally, the chart on the right of Fig. 9 displays the obtained results when considering $mean_k$ and $mean_S$ to discriminate between *fiction* (in orange) and *others* (in blue). Again, by observing the plots and evaluating the KS values obtained, it is also possible to notice a tendency to distinguish the two literary genres in question. This fact indicates that the genre label is somehow reflected

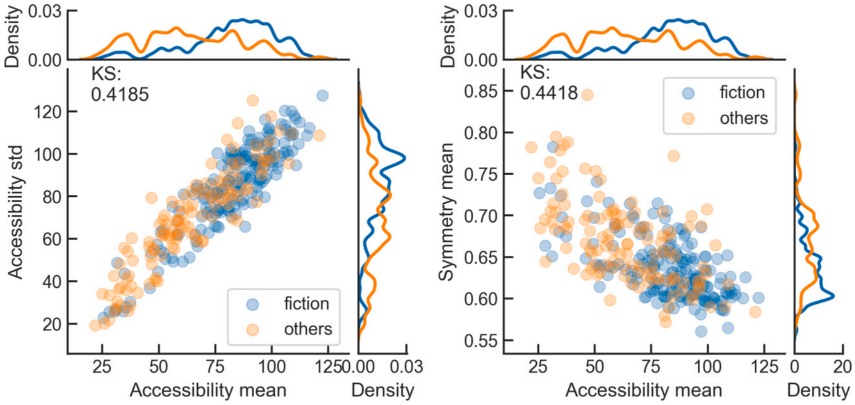

Fig. 9. Discriminating literary genres using accessibility's mean and standard deviation extracted from the recurrence network, on the left, with KS = 0.4185. On the right, the same discrimination task, but now using accessibility and backbone symmetry means, with KS = 0.4418.

in the narrative constructed in a book. More specifically, texts within the *fiction* group are written in a similar way, with a similar story flow, whereas the books belonging to the *others* group are slightly different from them, according to our methodology.

3.3. Comparing the proposed methodology with distinct approaches

Our goal in this work is not to develop the best method to accomplish the two afore-explained tasks, but rather to propose a new approach to characterize texts, focusing on their narrative semantic structure. Nevertheless, we perceive the importance of comparing our methodology with other well-established methods in the literature. In this manner, this subsection promotes a comparison between the recurrence networks, co-occurrence networks, and doc2vec modeling. With the obtained outcomes, we hope to illustrate that our method performs reasonably well in solving the two proposed tasks.

Co-occurrence networks connect adjacent words, reflecting the linear sequence of events in a narrative [4,26]. In this network representation, each node corresponds to a word, and two nodes share an edge if they appear together in the text at some moment. For example, given the phrase "In there stepped a stately Raven", we would link the preposition "in" to the adverb "there", the adverb "there" to the verb "stepped", and so on. It is worth noting that co-occurrence networks can be directed or undirected, and weighted or unweighted. In our case, we opted to use the undirected and unweighted version, as it provides the most straightforward arrangement.

To allow a forthright comparison between the recurrence and the co-occurrence networks, the accessibility and symmetry measures needed to be extracted for the latter modeling. As the books belonging to the dataset have different numbers of words, each network had a distinctive number of nodes. Such a characteristic makes it difficult to compare the same measure between networks with contrasting sizes, as the metrics' scale can vary with a such dissimilarity. In this sense, the number of words used to build the networks has been limited. As the shortest book in the dataset has 8306 vocables, this became the size of all texts before the construction of co-occurrence networks. In other words, we cut the textual content of the books so that all instances would share a similar quantity of nodes and edges and their accessibility and symmetry values could be compared.

It is worth mentioning, nonetheless, that the performed workaround leads to an easing in discriminating between meaningful and meaningless texts by the co-occurrence approach. Suppose we are working with two books, A and B, the first having one paragraph with 30 words and the second having two segments with 30 terms. When we perform the shuffling, book A will remain the same, while book B will have the order of its two paragraphs inverted. Then, the cut is performed, and all four samples (original and shuffled) now have 30 words. So when we compare the original instance of B with its random counterpart, B_{real} and B_{shuffled} will not share any content since B_{real} is currently composed of the book's first paragraph, and B_{shuffled} is composed of the second. Consequently, the original text and its random counterpart are distinct, making the task, at some level, more effortless.

The point is, even if the task has been made simpler, it is still possible to observe in Fig. 10 (a) and (c) the inability of the co-occurrence approach to capture discrepancies between meaningful and meaningless texts. In (a), when assessing the accessibility's mean and the accessibility's standard deviation, one can see several overlaps between the two considered classes. The same can be verified in (c) when examining the accessibility and symmetry means, although, in this case, the data are more disseminated along the plot. In fact, the KS values are near zero in both cases.

As aforementioned, we also modeled the problem using the doc2vec representation. Proposed by Mikolov and Le, this method seeks to learn a fixed-length feature representation from variable-length pieces of texts (in our case, full texts) based on an unsupervised algorithm [25]. To implement this, we used the open-source Python library Gensim [37]. The parameters selected to learn the representations were: *vector size* equals 128, *window* equals 5, *minimum count* equals 1, and 40 *epochs*. The *window size* and *minimum count* values are the standard ones in the used library. The number of *epochs*, in turn, was selected to guarantee the methods' convergence. As the doc2vec output was a dense representation vector, the UMAP technique was used for dimensionality reduction to enable the visualization of the obtained results [32].

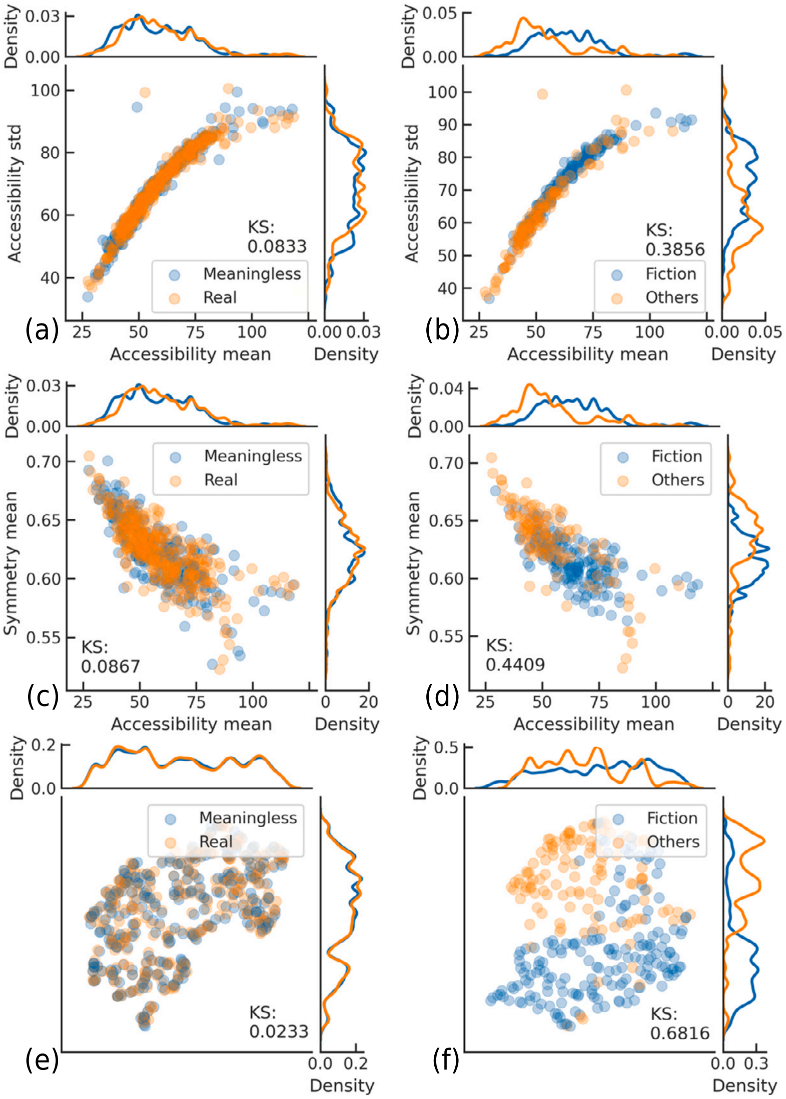

Fig. 10. On the left, discriminating between meaningful and meaningless texts with distinct approaches. On the right, distinguishing between textual genres. The subplots (a), (b), (c), and (d) refer to the co-occurrence network modeling. Subplots (e) and (f) depict the doc2vec representation after applying the UMAP dimensionality reduction technique.

The obtained result of discriminating between meaningful and meaningless texts using the doc2vec approach is depicted in Fig. 10 (e). Similar to the already-obtained results for the co-occurrence networks, a substantial overlay of classes can be observed in the plot. Besides that, once more the KS value presented a score very close to zero, meaning that this representation lacks the capability of distinguishing between the two analyzed classes. Notwithstanding, it is worth saying that this was an expected result since the method's purpose is to learn representations from texts, regardless of their paragraph appearance order in the narrative.

Regarding distinguishing between fiction and other genres task, the co-occurrence approach offered results as shown in Fig. 10 (b) and (d). In (b), we still have some overlapping classes, although now they are separable to some extent. The same behavior occurs in (d), albeit data are more distributed along the graph. Concerning KS values, both cases offer fair-to-middling results, with scores near 0.4 and 0.5, respectively. In the doc2vec approach, on the other hand, a suitable split is obtained, as shown in Fig. 10 (f). With very few plot overlaps, its KS value is close to 0.7.

Table 2 offers a summary comparison of the three explored approaches in terms of their KS values. As this score is a quantitative measure, it is the simplest way to compare the obtained results. As one can see, the recurrence network was the one that best performed the task of discerning meaningful and meaningless texts, with the highest KS score of 0.85, by using accessibility and symmetry information. Conversely, doc2vec was the one with the poorest result, with a score of 0.02. Notwithstanding, in terms of the textual genres task, doc2vec reached the highest discrimination capacity, offering a KS score of 0.68. Our method, nonetheless, got higher values than co-occurrence by using only accessibility information and proximate results using accessibility and symmetry data.

Table 2
KS values obtained for different representations. For the recurrence and co-occurrence networks, each respective column refers to the y-axis of the plot, with the accessibility's mean corresponding to the x-axis. For example, respectively to column *recurrence → symmetry mean*, the value 0.8517 refers to the KS score obtained by comparing the accessibility's and symmetry's mean using a recurrence network modeling approach (which can alternatively be consulted on the right of Fig. 7). In bold are highlighted the best-obtained result for each of the two assessed tasks.

	Recurrence		Co-occurrence		Doc2vec
	Accessibility std	Symmetry mean	Accessibility std	Symmetry mean	
Real & Meaningless	0.7717	**0.8517**	0.0833	0.0867	0.0233
Fiction & Others	0.4185	0.4418	0.3856	0.4409	**0.6816**

Table 3
ROC AUC values obtained for different representations. For the recurrence and co-occurrence networks, each respective column refers to the second feature considered, the first being accessibility's mean in all cases. For example, respectively to column *recurrence → symmetry mean*, the value 0.74 ± 0.06 refers to the AUC obtained by using the accessibility's and symmetry's mean of the recurrence networks as features in the SVC. In bold are highlighted the best-obtained result for each of the two assessed tasks.

	Recurrence		Co-occurrence		Doc2vec
	Accessibility std	Symmetry mean	Accessibility std	Symmetry mean	
Real & Meaningless	$\mathbf{0.95 \pm 0.02}$	$\mathbf{0.95 \pm 0.02}$	0.48 ± 0.06	0.46 ± 0.06	0.50 ± 0.01
Fiction & Others	0.75 ± 0.05	0.74 ± 0.06	0.74 ± 0.09	0.71 ± 0.07	$\mathbf{0.96 \pm 0.01}$

To check if the same information was being codified by the accessibility and symmetry from the two network modeling approaches, we calculated the Pearson (ρ_p) and Spearman (ρ_s) correlation indices between recurrence and co-occurrence accessibility's mean, recurrence and co-occurrence accessibility's standard deviation, and recurrence and co-occurrence symmetry's mean, all acquired from the textual genre networks (fiction and others). We chose to analyze these two classes as they presented similar values in Table 2. For the accessibility's mean, we obtained $\rho_p = 0.129$ and $\rho_s = 0.184$. For the accessibility's standard deviation, $\rho_p = 0.087$ and $\rho_s = 0.115$. Finally, for the symmetry's mean, $\rho_p = 0.284$ and $\rho_s = 0.274$. Hence, it was possible to infer that although the two methods share similar KS scores, different information is grasped by the distinct methodologies.

Alternatively, we obtained the Receiver Operating Characteristic's area under the curve (ROC AUC) to test the suitability of the recurrence method if compared to the co-occurrence modeling. The ROC AUC was preferred instead of Precision and Recall's area under the curve (PR AUC) as we have equivalent interest in the two pairs of two classes studied, despite one of the pairs presenting data imbalance. This choice strategy is discussed in depth in [39].

For the aforestated purpose, we ran an SVC classifier with 6-fold cross-validation, using as features (i) accessibility's mean and accessibility's standard deviation and (ii) accessibility's mean and symmetry's mean. The test assessed both the *real and meaningless* and *fiction and others* tasks. Six was chosen as the value for the k-fold to ensure at least 20 instances in each fold in all analyzed scenarios. Results are exhibited in Table 3. For the shuffling task, the recurrence method showed results far superior to those of doc2vec and co-occurrence, the last two being unable to surpass even the random expectation (known to be 0.5 for AUC, regardless of the class proportion). For the genres differentiation task, the results obtained using the accessibility's mean and standard deviation were quite similar. Nonetheless, the combination of accessibility's and symmetry's mean presented higher – and thus better – outcomes for the recurrence approach. Doc2vec remained the best result for the task.

The above-mentioned outcomes emphasize the potential of the recurrence networks once they can discriminate between imaginary and real texts and distinct textual genres in a way comparable to well-established methods in the literature. We emphasize, however, that the objective of the here described technique is not to obtain results that surpass the state-of-the-art approaches, but rather to describe an alternative modeling method that captures the narrative semantic structure.

4. Conclusions

Several interesting problems involving complex systems and structures are characterized by presenting properties along several scales. In the case of literary texts, their properties extend from more microscopic characteristics such as the vowels and consonants composing words to more macroscopic organization of the text along successive chapters. The present work aimed at developing further some previous approaches focusing on *mesoscopic* characterization of text [7,12]. Here, we aimed to investigate to what extent two types of measures: (a) accessibility [46]; and (b) symmetry [43], when calculated from the recurrence network proposed in the present work, can discriminate between real and meaningless texts and between different literary genres. Finally, to evaluate how well our methodology is able to perform at the given tasks, we compare our results with the results yielded from other, well-known methods in the literature, namely co-occurrence networks and doc2vec modeling.

We considered 300 books from the Gutenberg project [20], organized into two major genres. First, the texts were pre-processed and segmented into paragraphs, and the co-reference resolution technique was applied. Next, we employ a syntactic analysis to select the most contextual meaning of the sentences. More specifically, the paragraphs were converted into specific syntactic parts (subject, verb, or direct object). Finally, the recurrence network was created by considering the cosine similarity between the paragraphs. Two main types of experiments were performed, comparing real texts with meaningless texts, and comparing fiction and other genres.

From the first set of experiments, we conclude that both accessibility and symmetry calculated over the recurrence network are adequate to differentiate between real and meaningless texts, corroborating the hypothesis that the semantic aspects of the book's narratives are being captured by the proposed methodology. Moreover, the obtained results indicate that the chosen measures allowed reasonable separation in both the first and second experiments, indicating that this measure was capable of capturing, to some extent, the heterogeneity along the narrative. Corroborating the importance of scale in the analysis of texts, it has been verified that the best separations between the literary genres were obtained while choosing $h = 2$ as the topological scale for the accessibility and symmetry measures.

Finally, in our third experiment, by comparing the results yielded from our proposed methodology with other well-established approaches from the literature we are able to numerically reinforce the discrimination capability of the recurrence network. By the KS score values obtained, the approaches using the recurrence network performed better than the co-occurrence network for both discrimination tasks, between real and imaginary texts and between fiction and others. Furthermore, even though the doc2vec approach performed slightly better at differentiating between literary genres, our methodology with the recurrence network performed significantly better for the first experiment. Thus, the results obtained corroborate the hypothesis that our method is able to successfully grasp the semantic narrative of texts.

Among the related future developments, we have the consideration of additional books and genres, as well as other complementary measures including the BERT-related approaches [14,22]. In particular, it would be interesting to incorporate methods capable of capturing semantic aspects of the narrative. Additionally, in the future, we would like to employ our methodology in other, larger datasets, considering other literary genres (both with a narrative structure and without one) to further evaluate our method's ability to grasp the semantic context and the narrative story flow of texts.

Furthermore, the proposed network representation can be used for more practical applications. For instance, an option would be to use it to improve the quality of recommendation systems, in which features obtained from these networks feed these systems. The recurrence approach can be adapted to other types of texts, such as lyrics [35]. Another possible application is the area of science. For instance, nodes can represent the papers authored by researchers connected by similarity and by adjacency to model the dynamics of their research interests over time.

CRediT authorship contribution statement

Bárbara C. e Souza: Conceptualization, Formal analysis, Investigation, Methodology, Project administration, Supervision, Writing – original draft, Writing – review & editing. **Filipi N. Silva:** Conceptualization, Formal analysis, Investigation, Methodology, Project administration, Supervision, Writing – original draft, Writing – review & editing. **Henrique F. de Arruda:** Conceptualization, Formal analysis, Investigation, Methodology, Project administration, Supervision, Writing – original draft, Writing – review & editing. **Giovana D. da Silva:** Conceptualization, Formal analysis, Investigation, Methodology, Project administration, Supervision, Writing – original draft, Writing – review & editing. **Luciano da F. Costa:** Conceptualization, Formal analysis, Investigation, Methodology, Project administration, Supervision, Writing – original draft, Writing – review & editing. **Diego R. Amancio:** Conceptualization, Formal analysis, Investigation, Methodology, Project administration, Supervision, Writing – original draft, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgements

B.C. e Souza acknowledges CAPES for sponsorship - Brasil (CAPES) - Finance Code 001. H.F. de Arruda acknowledges FAPESP for sponsorship (grant no. 2018/10489-0, from 1st February 2019 until 31st May 2021) and Soremartec S.A. and Soremartec Italia, Ferrero Group, for partial financial support (from 1st July 2021). His funders had no role in study design, data collection and analysis, decision to publish, or manuscript preparation. L. da F. Costa thanks CNPq (grant no. 307085/2018-0) and FAPESP (grant no. 15/22308-2). D.R. Amancio thanks FAPESP (grant no. 20/06271-0) and CNPq (grant no. 304026/2018-2 and 311074/2021-9). This work has also been supported by the FAPESP (grant no. 19/07665-4 and 21/01744-0).

Appendix A. Computational expenses

In order to have a better idea of the computational expenses involved in the adopted methodology, we separated the tasks into three main groups: (a) Text processing (including syntactic analysis and text cleanup); (b) Network modeling (including the calculation of the edge weights and their respective pruning); and (c) Network characterization (including the calculation and extraction of the accessibility and symmetry measures).

The whole procedure was executed ten times respectively to the longest book, and the obtained average and standard deviation of the execution times were, in seconds: (a) 84.8 ± 5.5; (b) 151.4 ± 20.8; and (c) 0.5 ± 0.2, resulting in a total time of 236.7 ± 26.1 seconds. These results refer to execution in an x86-64 i7 1.8 GHz CPU, with the applications being implemented in Python.

The longest execution time was observed for the network modeling tasks, followed by text processing, with the network characterization corresponding to a fraction of the other times. Network modeling required the longest execution time because it requires the calculation of the cosine distance between pairwise combinations of nodes (paragraph windows). The network characterization tasks resulted noticeably fast, as it requires a relatively small hierarchical neighborhood.

References

[1] Project gutenberg, https://www.gutenberg.org.

[2] H.I. Abdalla, A.A. Amer, On the integration of similarity measures with machine learning models to enhance text classification performance, Inf. Sci. 614 (2022) 263–288.

[3] Y.-Y. Ahn, J.P. Bagrow, S. Lehmann, Link communities reveal multiscale complexity in networks, Nature 466 (7307) (2010) 761–764.

[4] D.R. Amancio, A complex network approach to stylometry, PLoS ONE 10 (8) (2015) e0136076.

[5] D.R. Amancio, Network analysis of named entity co-occurrences in written texts, Europhys. Lett. 114 (5) (2016) 58005.

[6] D.R. Amancio, F.N. Silva, L. da, F. Costa, Concentric network symmetry grasps authors' styles in word adjacency networks, Europhys. Lett. 110 (6) (2015) 68001.

[7] H.F. Arruda, F.N. Silva, V.Q. Marinho, D.R. Amancio, L. da F. Costa, Representation of texts as complex networks: a mesoscopic approach, J. Complex Netw. 6 (1) (2018) 125–144.

[8] A. Benatti, H.F. de Arruda, F.N. Silva, C.H. Comin, L. da Fontoura Costa, On the stability of citation networks, Phys. A, Stat. Mech. Appl. (ISSN 0378-4371) (2022) 128399, https://doi.org/10.1016/j.physa.2022.128399, https://www.sciencedirect.com/science/article/pii/S0378437122009578.

[9] V. Blondel, J.-L. Guillaume, R. Lambiotte, E. Lefebvre, Fast unfolding of communities in large networks, J. Stat. Mech. Theory Exp. 2008 (2008), https://doi.org/10.1088/1742-5468/2008/10/P10008.

[10] C.H. Comin, T. Peron, F.N. Silva, D.R. Amancio, F.A. Rodrigues, L.d.F. Costa, Complex systems: features, similarity and connectivity, Phys. Rep. 861 (2020) 1–41.

[11] E.A. Corrêa Jr, V.Q. Marinho, D.R. Amancio, Semantic flow in language networks discriminates texts by genre and publication date, Phys. A, Stat. Mech. Appl. 557 (2020) 124895.

[12] H.F. de Arruda, V.Q. Marinho, T.S. Lima, D.R. Amancio, L. da F. Costa, An image analysis approach to text analytics based on complex networks, Phys. A, Stat. Mech. Appl. 510 (2018) 110–120.

[13] H.F. de Arruda, V.Q. Marinho, L. da F. Costa, D.R. Amancio, Paragraph-based representation of texts: a complex networks approach, Inf. Process. Manag. 56 (3) (2019) 479–494.

[14] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: pre-training of deep bidirectional transformers for language understanding, arXiv preprint, arXiv:1810.04805, 2018.

[15] R.V. Donner, Y. Zou, J.F. Donges, N. Marwan, J. Kurths, Recurrence networks—a novel paradigm for nonlinear time series analysis, New J. Phys. 12 (3) (2010) 033025.

[16] G. Fasano, A. Franceschini, A multidimensional version of the Kolmogorov–Smirnov test, Mon. Not. R. Astron. Soc. (ISSN 0035-8711) 225 (1) (1987) 155–170, https://doi.org/10.1093/mnras/225.1.155.

[17] J. Feng, Z. Zhang, C. Ding, Y. Rao, H. Xie, F.L. Wang, Context reinforced neural topic modeling over short texts, Inf. Sci. 607 (2022) 79–91.

[18] T.M. Fruchterman, E.M. Reingold, Graph drawing by force-directed placement, Softw. Pract. Exp. 21 (11) (1991) 1129–1164.

[19] M. Garg, M. Kumar, Identifying influential segments from word co-occurrence networks using AHP, Cogn. Syst. Res. 47 (2018) 28–41.

[20] M. Gerlach, F. Font-Clos, A standardized Project Gutenberg corpus for statistical analysis of natural language and quantitative linguistics, Entropy 22 (1) (2020) 126.

[21] A. Grover, J. Leskovec, node2vec: scalable feature learning for networks, in: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016, pp. 855–864.

[22] S. Han, L. Shi, R. Richie, F.R. Tsui, Building Siamese attention-augmented recurrent convolutional neural networks for document similarity scoring, Inf. Sci. 615 (2022) 90–102.

[23] A. Kulig, S. Drożdż, J. Kwapień, P. Oświęcimka, Modeling the average shortest-path length in growth of word-adjacency networks, Phys. Rev. E 91 (3) (2015) 032810.

[24] A. Kulig, J. Kwapień, T. Stanisz, S. Drożdż, In narrative texts punctuation marks obey the same statistics as words, Inf. Sci. 375 (2017) 98–113.

[25] Q. Le, T. Mikolov, Distributed representations of sentences and documents, in: International Conference on Machine Learning, PMLR, 2014, pp. 1188–1196.

[26] J. Machicao, E.A. Corrêa Jr, G.H. Miranda, D.R. Amancio, O.M. Bruno, Authorship attribution based on life-like network automata, PLoS ONE 13 (3) (2018) e0193703.

[27] C. Manning, H. Schutze, Foundations of Statistical Natural Language Processing, MIT Press, 1999.

[28] C.D. Manning, H. Schütze, Foundations of Statistical Natural Language Processing, MIT Press, Cambridge, MA, USA, ISBN 0262133601, 1999.

[29] C.D. Manning, M. Surdeanu, J. Bauer, J.R. Finkel, S. Bethard, D. McClosky, The Stanford CoreNLP natural language processing toolkit, in: Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations, 2014, pp. 55–60.

[30] V.Q. Marinho, G. Hirst, D.R. Amancio, Authorship attribution via network motifs identification, in: 2016 5th Brazilian Conference on Intelligent Systems (BRACIS), IEEE, 2016, pp. 355–360.

[31] V.Q. Marinho, H.F. de Arruda, T. Sinelli, L. da, F. Costa, D.R. Amancio, On the "calligraphy" of books, in: Proceedings of TextGraphs-11: the Workshop on Graph-Based Methods for Natural Language Processing, Association for Computational Linguistics, Vancouver, Canada, 2017, pp. 1–10.

[32] L. McInnes, J. Healy, J. Melville, UMAP: uniform manifold approximation and projection for dimension reduction, 2018.

[33] T. Mikolov, K. Chen, G. Corrado, J. Dean, Efficient estimation of word representations in vector space, arXiv preprint, arXiv:1301.3781, 2013.

[34] B. Mutlu, E.A. Sezer, M.A. Akcayol, Candidate sentence selection for extractive text summarization, Inf. Process. Manag. 57 (6) (2020) 102359.

[35] B.G. Patra, D. Das, S. Bandyopadhyay, Retrieving similar lyrics for music recommendation system, in: Proceedings of the 14th International Conference on Natural Language Processing (ICON-2017), 2017, pp. 290–297.

[36] M.K. Rahman, M.H. Sujon, A. Azad, Force2vec: parallel force-directed graph embedding, in: 2020 IEEE International Conference on Data Mining (ICDM), IEEE, 2020, pp. 442–451.

[37] R. Řehůřek, P. Sojka, Software framework for topic modelling with large corpora, in: Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks, Valletta, Malta, ELRA, May 2010, pp. 45–50.

[38] N. Reimers, I. Gurevych, Sentence-BERT: sentence embeddings using Siamese BERT-networks, arXiv preprint, arXiv:1908.10084, 2019.

[39] T. Saito, M. Rehmsmeier, The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets, PLoS ONE 10 (3) (2015) 1–21, https://doi.org/10.1371/journal.pone.0118432.

[40] L. Santos, E.A. Corrêa Júnior, O. Oliveira Jr, D. Amancio, L. Mansur, S. Aluísio, Enriching complex networks with word embeddings for detecting mild cognitive impairment from speech transcripts, in: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Vancouver, Canada, July 2017, pp. 1284–1296.

[41] F. Silva, Helios-web, https://github.com/filipinascimento/helios-web, 2022.

[42] F.N. Silva, D.R. Amancio, M. Bardosova, L. da F. Costa, O.N. Oliveira Jr., Using network science and text analytics to produce surveys in a scientific topic, J. Informetr. 10 (2) (2016) 487–502.

[43] F.N. Silva, C.H. Comin, T.K. Peron, F.A. Rodrigues, C. Ye, R.C. Wilson, E.R. Hancock, L. da F. Costa, Concentric network symmetry, Inf. Sci. 333 (2016) 61–80.

[44] T. Stanisz, J. Kwapień, S. Drożdż, Linguistic data mining with complex networks: a stylometric-oriented approach, Inf. Sci. 482 (2019) 301–320.

[45] T. Stanisz, S. Drożdż, J. Kwapień, Universal versus system-specific features of punctuation usage patterns in major western languages, Chaos Solitons Fractals 168 (2023) 113183.

[46] B. Travençolo, L. da F. Costa, Accessibility in complex networks, Phys. Lett. A 373 (1) (2008) 89–95.

[47] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Lio, Y. Bengio, Graph attention networks, arXiv preprint, arXiv:1710.10903, 2017.

[48] M.C. Waumans, T. Nicodème, H. Bersini, Topology analysis of social networks extracted from literature, PLoS ONE 10 (6) (2015) e0126470.

[49] X. Yang, Y. Li, D. Meng, Y. Yang, D. Liu, T. Li, Three-way multi-granularity learning towards open topic classification, Inf. Sci. 585 (2022) 41–57.

[50] F. Zheng, S. Zhang, C. Churas, D. Pratt, I. Bahar, T. Ideker, HiDeF: identifying persistent structures in multiscale ‘omics data, Genome Biol. 22 (1) (2021) 1–15.