Information Processing and Management 56 (2019) 479–494

Contents lists available at ScienceDirect

Information Processing and Management

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

Paragraph-based representation of texts: A complex networks approach

Henrique F. de Arruda*[,a], Vanessa Q. Marinho[a], Luciano da F. Costa[b], Diego R. Amancio[a,c]

[a] Institute of Mathematics and Computer Science, University of São Paulo, São Carlos, SP, Brazil
[b] São Carlos Institute of Physics, University of São Paulo, São Carlos, SP, Brazil
[c] School of Informatics, Computing and Engineering, Indiana University, Bloomington, IN 47408, USA

ARTICLE INFO

Keywords:
Network science
Complex networks
Language networks
Document classification
Text networks

ABSTRACT

An interesting model to represent texts as a graph (also called network) is the word adjacency (co-occurrence) representation, which is known to capture mainly syntactical features of texts. In this study, we propose a novel network model, which is based on the similarity between the content of the paragraphs of the text. By considering this representation, we characterized the networks with respect to measurements developed in the network science area. We characterized these measurements according to their properties regarding their ability to discriminate between real and shuffled texts, and to capture information regarding the content similarity of chunks of text. In order to compare the results with a more sophisticated approach, we employed a methodology based on *word2vec*. When comparing real and shuffled texts, the results revealed that real texts tend to have a more well-defined community structure. This characteristic can be related to the organization of subjects in real texts. The network-based measurements that were found to be able to discriminate real from shuffled texts were used as features in a classifier. As a result, the obtained accuracy was 98.72%. In order to compare with a different methodology, we used *doc2vec*-based features in the classifier, yielding an accuracy rate of 70.8%. The proposed network-based features were employed to analyze the Voynich manuscript, which was found to be compatible with real texts according to the considered characteristics.

1. Introduction

Due to the ever-increasing number of available online texts, many machine learning techniques have been developed to treat this kind of information (Belbachir & Boughanem, 2018; Hsu, Lee, Chang, & Sung, 2018; Kim & Kang, 2018; Sicilia, Giudice, Pei, Pechenizkiy, & Soda, 2018; Symeonidis, Effrosynidis, & Arampatzis, 2018; Xiong, Wang, Zhang, & Ma, 2018). Among many statistical methods, network-based approaches have also been proposed to address several natural language processing problems, including writing style analysis (Amancio, 2015c), authorship attribution (Posadas-Durán et al., 2017) and sentiment analysis (Giatsoglou et al., 2017). Several graph-based approaches hinge on the topological information of the obtained networks to perform some type of classification (Amancio, 2015b; Angelova & Weikum, 2006; Cancho & Solé, 2001; Erkan & Radev, 2004; Jin & Srihari, 2007; Liu & Cong, 2013; Yu, Wang, Zhang, Zhang, & Liu, 2017).

A well-known representation of texts as complex networks is the co-occurrence model (Amancio, 2015b; Cancho & Solé, 2001; Liu

* Corresponding author.
E-mail addresses: hfarruda@icmc.usp.br (H. F. de Arruda), diego.amancio@usp.br (D. R. Amancio).

https://doi.org/10.1016/j.ipm.2018.12.008
Received 1 July 2018; Received in revised form 21 December 2018; Accepted 30 December 2018
Available online 21 January 2019

& Cong, 2013; Wachs-Lopes & Rodrigues, 2016). This model represents words as nodes, and edges are established for every pair of adjacent words. Recently, this representation was found to capture mainly syntax features (Amancio, Altmann, Rybski, Oliveira Jr, & Costa, 2013; Masucci & Rodgers, 2006), which has been confirmed by numerous works using co-occurrence networks to study language styles (Amancio, 2015a; de Arruda, Costa, & Amancio, 2016b; Cong & Liu, 2014; Mehri, Darooneh, & Shariati, 2012; Segarra, Eisen, & Ribeiro, 2015). In order to grasp features that go beyond syntax, other models have been proposed. In de Arruda, Costa, and Amancio (2016a), the authors still consider words as nodes, but the connections are created using a larger window, rather than only consecutive words. Upon applying community detection methods, this approach was successfully employed to detect topics. In addition to representing words as networks, larger scales have also been considered to construct text networks de Arruda, Silva, Marinho, Amancio, and Costa (2018b).

In this work, we propose a novel paragraph-based network. Edges are created based on the similarity between texts, which is created with basis on similarities by employing *tf-idf* (term frequency-inverse document frequency) weighting (Manning & Schütze, 1999) together with cosine similarity. Differently from previous approaches (Salton, Singhal, Mitra, & Buckley, 1997), the paragraph-based networks considered here are analyzed concerning their topological and dynamical properties. The properties of the adopted network representation were probed by considering two different criteria. To test the informativeness of the networks, we investigated whether paragraph-based networks are able to discriminate real from shuffled texts. In the second test, we analyzed if the networks can capture syntax and, most importantly, to reflect the content of the texts. Our results showed that the modularity played an important role in distinguishing real and shuffled texts since the presence of communities turned out to be a characteristic inherent of real texts. We also found that particular measurements are able to capture content features of texts, a characteristic that has not been observed in most co-occurrence networks modeling texts (Amancio et al., 2013), since co-occurrence networks capture mostly syntactical (stylistic) text features.

In addition to the analysis aimed at better understanding the statistical properties of paragraph-based networks, we probed the statistical properties of an unknown text – the Voynich manuscript – using the framework based on the network representation proposed here. The Voynich manuscript is an undeciphered text with uncertain origins (Reddy & Knight, 2011). It is written with a set of unknown characters and, consequently, the approached subject is unknown. Researchers have been studying the characteristics of this mysterious text, which include textual analysis (Reddy & Knight, 2011) and also network science-related tools (Amancio et al., 2013; Montemurro & Zanette, 2013). In order to deal with this text, it is necessary to convert the Voynich characters into an intelligible representation. In this study, we employed the European Voynich Alphabet (Znadbergen, 2018), in which the characters were manually translated into European characters.

In contrast with other approaches, we did not assume that pages are organized in any specific order. It is an important feature because a recent study revealed that the traditionally assumed pages ordering might be unreliable (Reddy & Knight, 2011). Observe that here we employed a method that is unable to capture synonyms (tf-idf) since the proposed network representation is intended to analyze any arbitrary sequence of symbols (e.g., unknown languages). So, it is difficult to use a more sophisticated method because there are no other texts that can be used as a training set (this is the case when analyzing the Voynich manuscript). However, the results obtained for real and artificial texts suggest that the proposed methodology was able to analyze the nature of unknown documents. Interestingly, our results indicate that the Voynich manuscript is compatible with natural languages and incompatible with shuffled texts. These conclusions were mostly corroborated by observing the community structure arising from the manuscript.

The remainder of this paper is organized as follow. Section 2 presents the related studies and Section 3 describes the goals and research questions of this study. In Section 4, we present the employed datasets, the proposed methodology, and the used complex network measurements. Section 5 presents an analysis of the paragraph-network properties by comparing real documents with two versions of shuffled texts. Furthermore, in the same section, we present a case study where we analyze the Voynich manuscript. Finally, in Section 6, we conclude the study with perspectives for further works.

2. Related works

Many efforts have been made to develop approaches devoted to understanding unstructured data using computational tools. Popular research topics in the text mining area include topic modeling (Dai, Olah, & Le, 2015), disambiguation (Henry, Cuffy, & McInnes, 2017), sentiment analysis (Giatsoglou et al., 2017), among others (Amancio, 2015c; Posadas-Durán et al., 2017). Recent advances in language processing and deep learning have enabled analyzing texts with word embeddings (Mikolov, Sutskever, Chen, Corrado, & Dean, 2013b). This methodology consists in creating a distributional representation of words, with some interesting characteristics. For example, if the vectors representing the words “king” and “woman” are summed, the expected result is very close to the vector representing the word “queen”. Another methodology that also has similar characteristics is the method based on document embeddings (Le & Mikolov, 2014), referred to as *doc2vec* (Le & Mikolov, 2014). In this case, the vector representation aims at representing the subject of the text. This methodology has been employed in the task of topic modeling (Dai et al., 2015). Observe that the training step in embedding-based techniques normally requires extensive datasets to work properly (Dai et al., 2015; Mikolov, Chen, Corrado, & Dean, 2013a).

Another type of text-mining approach is that based on networked/graph representations. Complex networks have been used to model several language problems over the last years (Metzler & Croft, 2004; Zhao, Mao, & Lu, 2018). The structure of the co-occurrence networks – formed by linking adjacent words – was found to display several characteristics of other real systems, including the small-world effect, characterized by typical low distances and high-clustering coefficient (Cancho & Solé, 2001). Interestingly, when more accurate links (syntactic links) are considered, the same patterns were observed (Amancio, Nunes, Oliveira Jr, & Costa, 2012a; Cancho, Solé, & Köhler, 2004). In addition to the small-world effect, word networks were found to exhibit the power-law distribution, which is a consequence of the well-known Zipf's Law (Newman, 2005). The network structure has also been used to explain some linguistic patterns (Cancho & Solé, 2003).

Word networks have been used to tackle stylistic problems by using structural and dynamical properties of complex networks. The analysis of language complexity was analyzed in Wu, Zhang, and Ren (2017), where the authors found that there is a strong relationship between the complexity of the network structure and textual cognitive complexity. In addition, word networks have also been used to tackle the word sense disambiguation problem. In Agirre and Soroa (2009), the authors constructed graphs from natural language documents and then studied how ambiguous words in these documents can be connected to semantic concepts in the WordNet (Miller, 1995). They found that it is possible to identify the sense being conveyed by a word when the PageRank algorithm is used to identify which is the most central synset in a given context. The same problem was addressed using a structural approach in Silva and Amancio (2012).

It has been shown that word co-occurrence networks are able to model structural features of languages since such representation captures most of the syntactical links (Cancho et al., 2004). This finding explains, in a certain way, the widespread use of word networks in structural natural language processing tasks. More recently, language networks have been studied using a larger textual scale, by modeling texts where nodes are sentences, paragraphs or larger chunks. A network based on the similarity of large chunks was proposed by de Arruda et al. (2018b). This methodology was found to be useful to understand and visualize the unfolding of stories (de Arruda, Marinho, Lima, Amancio, & Costa, 2018a; Marinho, de Arruda, Lima, Costa, & Amancio, 2017). In the summarization context, another approach that also took into consideration larger chunks of texts is the network of connected paragraphs (Salton et al., 1997). Networked representations of texts have also been used to distinguish real from shuffled texts (de Arruda et al., 2018b; Margan, Martinčić-Ipšić, & Meštrović, 2014a; Margan, Mestrovic, & Martincic-Ipsic, 2014b; Masucci & Rodgers, 2009). In these studies, the authors found that some topological patterns are occurring only in real texts. Differently from these previous works, here we focus on the analysis of the statistical properties of *paragraph-based* networks with application to authenticity verification.

3. Research questions

The objective of this study encompasses mainly two research questions. The first regards the use of a network model that is mostly related to the content (i.e., the meaning) of the text. In other words, here we investigate the characteristics of the paragraph-based networks, including its ability to discriminate between real and shuffled texts. Note that this approach contrasts with the traditional networked models that primarily reflects syntactical characteristics (Amancio et al., 2013). The second research question regards the Voynich manuscript, which has an undiscovered subject written in an unknown language. More specifically, here we aim at answering the following question: "Does the Voynich manuscript bear any meaning?". This issue has drawn the attention of many scholars from various disciplines (Belfield, 2007). More specifically, we compare the characteristics of a network created from Voynich with real and shuffled documents, which were written in different languages. These questions are addressed by representing texts as paragraph-networks, which are then analyzed via structural and dynamical measurements combined with traditional machine learning techniques.

4. Materials and methods

This section describes the employed datasets, the approach devised to create paragraph-based networks and the measurements extracted from the text networks.

4.1. Dataset

We employed two datasets. The first one, henceforth referred to as the Holy Bible dataset, was used to represent the variation of syntax *across different languages* when the text/content is the same. It comprises three books from the New Testament of the Holy Bible: Matthew, Mark, and Luke. 16 different languages were considered: Arabic, Basque, English, Esperanto, German, Greek, Hebrew, Hungarian, Korean, Latin, Maori, Portuguese, Russian, Swahili, Vietnamese, and Xhosa. The three books were concatenated into a single document to obtain a larger text, as our method is more reliable when larger pieces of texts are used to construct the network. This same procedure has been applied in similar studies (Amancio et al., 2013). For all considered languages, the paragraphs comprise the same verses. In total, 658 paragraphs were manually identified.

The second dataset, henceforth referred to as Books dataset, comprises 53 books in different languages, namely English, French, German, Italian and Portuguese. This dataset was used to analyze how the network structure varies *across different documents* in the same language. The list of books is presented in Appendix A.

4.2. Paragraph-based networks

In this work, texts are modeled as complex networks. A network (or graph) can be defined as a set $V = \{v_1, v_2, \dots, v_n\}$ of nodes and a set $E = \{e_1, e_2, \dots, e_m\}$ of edges. In an unweighted network, the element a_{ij} of the adjacency matrix **A** is equal to 1 if node i is connected to node j; otherwise, $a_{ij} = 0$. In weighted networks, the element a_{ij} corresponds to the weight of the link between nodes i and j.

The main objective of the adopted network model is to represent how short contexts (i.e., paragraphs) relate to each other in a textual document. To create a paragraph-based network the raw text is divided into paragraphs. Each paragraph is considered as a

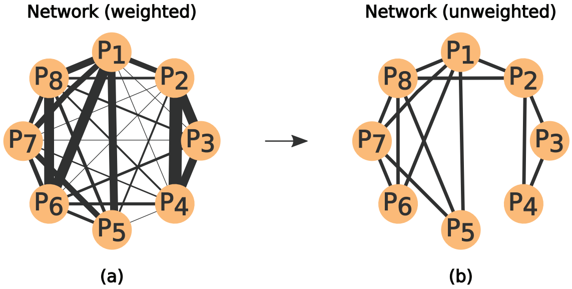

Fig. 1. Example of thresholding paragraph-based networks. In the weighted version (a), all nodes (paragraphs) are connected among themselves and the weight of each edge is given by the textual (content) similarity between the nodes. The unweighted version (b) is obtained by removing all edges with weights below a given threshold T.

network node, as illustrated in Fig. 1(a)). In order to establish links between paragraphs, each node is considered as a document d in the set D of documents. The *tf-idf* (term frequency-inverse document frequency) weighting map (Manning & Schütze, 1999) is then computed to quantify the relevance of each word $w \in d$:

$$\text{tf-idf}(w, d, D) = \frac{f_{w,d}}{n} \times \log\left(\frac{|D|}{d_w}\right), \tag{1}$$

where $f_{w,\ d}$ is the frequency of $w \in d,\ n$ is the total number of words in d and d_w is the number of documents (paragraphs) in which w appears. Observe that, to compute tf-idf, we consider each paragraph as a single document. Due to the differences of pre-processing of texts written in different languages, we opted for keeping all words of the text, even the stop words (i.e., the words conveying low semantic content). Because stop words tend to be frequent and topic independent, the tf-idf weights computed for such words also tend to be low. For each paragraph, a vector containing the tf-idf weights for the words is created, and the edge weights are computed by using the cosine similarity for all pairs of paragraphs (nodes). Note that this methodology creates a fully connected, weighted graph, as illustrated in Fig. 1(a). Because many of the complex networks measurements are defined only for unweighted networks, we converted the network into their unweighted versions. For that, we removed the weakest edges which have values lower than a threshold T, and the remaining edges were considered as unweighted. For the considered networks, we chose a threshold for each network in order to keep all networks with the same size and density $E = 5\%$. It is an important step in the pre-processing phase because several network measurements are known to be very sensitive to both size and density (Amancio, 2015c; Costa, Rodrigues, Travieso, & Villas Boas, 2007). In preliminary experiments, we found that a perturbation in T does not alter the conclusions reported here, as preliminary tests confirmed it. The effect of thresholding the weighted network is illustrated in Fig. 1(b).

The proposed methodology is similar to the mesoscopic networks approach (de Arruda et al., 2018b) regarding the network edge weights. Actually, paragraph networks can be understood as a specific case of mesoscopic networks in which each chunk of text is a single paragraph with no forced overlap between adjacent chunks. An advantage of the present study is that here we can analyze documents in which the order of the pages and paragraphs are unknown. Note that other techniques of word representation, such as word embeddings, were not considered to compute the content similarity because the proposed method was developed to be applied even in texts whose language is unknown.

4.3. Network variations

In order to compare real and shuffled texts, three types of networks were considered. The paragraph-based network – denoted as *real texts* (RT) – is obtained from the pre-processed texts of the considered datasets, as described in Section 4. The other networks are obtained from shuffled versions of the original text. The versions are created by shuffling words (*SW*) or sentences (*SS*). It is important to highlight that both shuffled versions (SW and SS) recover the paragraph sizes of the original text. More specifically, the process of shuffling is applied to the entire text, and the new paragraphs are defined according to the original number of words of the respective text. The versions obtained from an extract of the book *The Adventures of Sherlock Holmes*, by Arthur Conan Doyle are:

1. *Real Text (RT) version*: ”Quite so,” he answered, lighting a cigarette, and throwing himself down into an armchair. ”You see, but you do not observe. The distinction is clear.
2. *Shuffled words (SW) version*: ”Quite a into do distinction armchair. but lighting and answered, The observe. himself down you so,” not throwing he see, cigarette, is clear. ”You an
3. *Shuffled Sentences (SS) version*: ”You see, but you do not observe. ”Quite so,” he answered, lighting a cigarette, and throwing himself down into an armchair. The distinction is clear.

4.4. Network characterization

The following network measurements were used to characterize the paragraph-based networks:

1. **Degree (k):** This measurement quantifies the number of immediate neighbors of a node i (Costa et al., 2007) and it is obtained as $k_i = \sum_j A_{ij}$.
2. **Betweenness (B):**This measurement quantifies the relevance of a node (or edge) in terms of the number of shortest paths including that node (or edge) (Freeman, 1977). The betweenness centrality of a given node i is calculated as

$$B_i = \sum_{s,t} \frac{g_{s,t}^i}{g_{s,t}}, \tag{2}$$

   where $g_{s,t}^i$ is the number of shortest paths connecting nodes s and t that include node i, and $g_{s,\ t}$ is the number of shortest paths connecting s and t, for all pairs s and t. In text networks, this measurement has been applied to identify if a concept/node is semantically related to one or more topological communities (Amancio, Altmann, Oliveira Jr., & Costa, 2011).
3. **Clustering coefficient (cc):** The clustering coefficient represents the probability of two neighbors of a given node being connected with each other (Strogatz, 2001). Locally, the clustering coefficient is calculated as $cc_i = 2e_i/(k_i^2 - k_i)$. In text analysis, the clustering coefficient has also been used to identify if a concept appears in generic or specific contexts. Differently, from the betweenness, only local information is considered.
4. **Neighborhood (N):** this measurement quantifies the amount of nodes in the h-th concentric level around node i (Newman, 2010). In this study, we used $h = 3$.
5. **Eccentricity (Ecc):** the eccentricity of a node i is a centrality index equal to the maximum length of all the shortest paths from i to the other nodes in the network (Harary, 1969).
6. **Eigenvector centrality (EC):** the eigenvector centrality assigns a value to a given node i proportional to the sum of the eigenvector centrality values of the nodes connected to i. By doing so, the centrality value of a node increases when it is connected to nodes with high eigenvector centrality (Bonacich, 1987).
7. **Closeness centrality (C):** this measurement is given by the inverse of the average distance from a node to the other nodes in the network (Newman, 2010). It is obtained as $C_i = l_i^{-1} = n/\sum_j d_{ij}$, where l_i is the average distance from node i to all the other nodes, and d_{ij} is the length of a geodesic path connecting nodes i and j.
8. **Accessibility ($\alpha^{(h)}$):** This measurement quantifies the number of accessible nodes at the h-th concentric level centered at node i (Travenolo & Costa, 2008) (we used $h = \{2, 3\}$). This analysis accounts for the accessibility of a node taking into account the probability $p_{i,j}^{(h)}$ of a random walker to reach a given node j departing from i, in h steps. The equation that describes this measurement is based on the Shannon entropy, as follows

$$\alpha_i^{(h)} = \exp\left(-\sum p_{i,j}^{(h)} \log p_{i,j}^{(h)}\right). \tag{3}$$

   In language networks, the accessibility (and its variations) has been used as an important feature to identify the relevance of words in the context of structural/stylistic analysis (Amancio, 2015b; 2015c).
9. **Generalized Accessibility ($\alpha^{(\infty)}$):** The generalized accessibility does not depend on the parameter h. In contrast with the previous measurement, generalized accessibility uses a modified random walk, called accessibility random walk, which assigns higher weights to the shortest paths and penalizes the longest ones (de Arruda et al., 2014). Mathematically, the measurement is defined as

$$\alpha_i^{(\infty)} = \exp(-\sum P_{i,j} \log P_{i,j}), \tag{4}$$

   where P is computed as the probability transition of all the pairs of nodes i and j. More details can be found in de Arruda et al. (2014).
10. **Symmetry ($S^{(h)}$):** As another variation of accessibility, this measurement quantifies the symmetry of the topology around a given node i, by considering its neighborhood (h) (Silva et al., 2016b). $S^{(h)}$ is defined in a two-fold manner: (i) the *backbone* ($Sb^{(h)}$), in which the connections between nodes in the same hierarchical level (h) are removed and (ii) *merged* ($Sm^{(h)}$), where the nodes that are connected and belong to the same hierarchical level (h) are merged into a single node. The measurement is computed as

$$S_i^{(h)} = \frac{\exp\left(-\sum p_{i,j}^{(h)} \log p_{i,j}^{(h)}\right)}{|H_h(i)| + \sum_{r=0}^{h-1} \eta_r}, \tag{5}$$

   where $H_h(i)$ is the set of all nodes in the $h - th$ hierarchic level of node i, $|H_h(i)|$ is the number of nodes in $H_h(i)$, and by considering a given hierarchic level r, η_r is the number of nodes without edges connecting to the next hierarchical level. In this study, we employed $h = \{2, 3, 4\}$. In text networks, the symmetry has been useful to identify the authorship of texts (Amancio, Silva, & Costa, 2015).
11. **Modularity (Q):** proposed by Newman and Girvan (2004), the modularity measures the quality of a given network partitioning in terms of its communities. It can be obtained as:

$$Q = \frac{1}{2m} \sum_{i=1}^{n} \sum_{j=1}^{n} \left[a_{ij} - \frac{k_i k_j}{2m} \right] \delta(c_i, c_j), \tag{6}$$

where m is the number of edges, n is the number of nodes, $\delta(c_i, c_j) = 1$ if the nodes i and j are from the same class (community) and $\delta(c_i, c_j) = 0$, otherwise. This measurement ranges from $-\infty \leq Q < 1$. For $Q > 0$, the number of edges inside the communities is greater than the expected in a equivalent random network. In other words, a positive value of modularity is an indicative that the network is organized in communities.

Apart from the modularity, all of the measurements mentioned above are locally defined, i.e., each node has a specific value. To summarize the values obtained for a measurement X across all nodes of the network, we took the average ($\langle X \rangle$) and the standard deviation ($\sigma(X)$). Note that this approach has already been adopted in similar works (de Arruda et al., 2016b; Marinho et al., 2017).

An important issue arising from the characterization and classification of networks concerns the comparison of networks with different sizes. Since several network measurements may depend on the total number of nodes, we decide to construct the networks so as the total number of nodes ($|V|$) is constant. This number corresponds, by construction, to the total number of paragraphs, which was set as $|V| = 723$ in order to match the length of the Voynich manuscript. So, for all of the books, we considered only the first 723 paragraphs.

4.5. Informativeness analysis

In the adopted network representation, we define as informative the measurements whose values obtained from real books and the respective shuffled versions are significantly different. Measures complying with this condition are therefore able of discriminating between real and random manuscripts. Note that an informative measurement is useful to verify if an unknown manuscript is compatible with a known textual structure (e.g., the structure observed in documents written in natural languages).

Two criteria were used to test the informativeness of the networks:

1. *Criterion A:* this criterion is aimed at verifying if the values obtained from the set of all shuffled texts of the dataset can be discriminated from the values obtained for all real texts. Let N_{RT} and N_{S} be the total number of books in the RT dataset and the number of shuffled versions generated for each book in RT, respectively. Here, we perform a comparison of N_{RT} values in RT with $N_{\mathrm{RT}} \cdot N_{\mathrm{S}}$ values obtained from shuffled texts.
2. *Criterion B:* it consists in comparing the value obtained for the real (RT) text with the values obtained in the corresponding shuffled versions of the same text. For a given measurement X, the distance between a real text and the respective shuffled versions is obtained by computing the *z-score* (i.e. the standard score):

$$z(X) = \frac{x - \langle X^{(R)} \rangle}{\sigma(X^{(R)})}, \tag{7}$$

where x is the value obtained in the real text, $X^{(R)}$ is the set of values obtained from the N_{S} shuffled versions (SW or SS); and $\langle \dots \rangle$ and σ represent the mean and standard deviation of the distribution, respectively.

In our tests, for each real text, we created $N_{\mathrm{S}} = 30$ samples for each SW and SS versions.

4.6. Dependency with language and contents

An important property to be verified in a text network is the ability of the extracted measurements to capture syntactical and/or content-based features of the represented texts (Amancio et al., 2013). In order to study the dependency of the measurements on syntax and content, the measurements are extracted in two classes of datasets. For a given measurement, $X_{t=\pi,l}$ represents the set of values obtained for X in a dataset comprising the same book $t = \pi$ in different languages (l). In a similar fashion, $X_{t,l=\lambda}$ represents the set of values obtained for X in a dataset comprising different texts (t) written in the same language $l = \lambda$. If a given network measurement depends more on the language (i.e. the syntax) than on the approached subject (i.e. the content), one expects that variability of the distribution of $X_{t=\pi,l}$ will be larger than the variability of $X_{t,l=\lambda}$. Conversely, if X is more dependent on content, one expects that the variability of $X_{t,l=\lambda}$ will be larger than the variability of $X_{t=\pi,l}$ (Amancio et al., 2013). Here, the variability of the distributions is computed by using the coefficient of variation (CV) of the distribution, i.e.

$$\mathrm{CV}(X) = \frac{\sigma(X)}{\langle X \rangle}. \tag{8}$$

5. Results and discussion

In this section, we analyze the properties of the metrics extracted from the proposed network representation. Here we focus on two main properties: informativeness and the ability of the metrics to capture syntactical and/or content features. The applicability of the adopted representation is then illustrated in the analysis of an unknown text: the Voynich manuscript.

5.1. Informativeness

In this study, we used distinct ways to quantify informativeness (Amancio et al., 2013). In the first approach, we consider a

Table 1
Measurements obtained for the different network types (RT, SS, and SW) by considering the Holy Bible dataset, the English part of the book dataset, and the entire book dataset (see Appendix Appendix A). Note that all the presented data is standardized to be possible to compare different measurements.

	Holy Bible dataset			Books dataset (all)		
X	RT	SS	SW	RT	SS	SW
$\langle k \rangle$	− 0.12 ± 1.14	+ 0.13 ± 0.87	− 0.01 ± 0.96	− 0.18 ± 1.02	+ 0.05 ± 0.98	+ 0.13 ± 0.98
$\sigma(k)$	− 0.59 ± 0.52	+ 0.56 ± 1.15	+ 0.02 ± 0.86	− 0.78 ± 0.65	+ 0.47 ± 0.92	+ 0.31 ± 0.90
$\langle B \rangle$	+ 0.48 ± 0.90	− 0.23 ± 1.02	− 0.24 ± 0.90	− 0.05 ± 1.07	+ 0.05 ± 1.03	+ 0.00 ± 0.89
$\sigma(B)$	− 0.22 ± 0.89	+ 0.23 ± 1.11	− 0.01 ± 0.93	− 0.66 ± 0.68	+ 0.52 ± 0.94	+ 0.13 ± 0.97
$\langle cc \rangle$	+ 0.74 ± 0.60	− 0.58 ± 1.09	− 0.16 ± 0.74	+ 0.31 ± 0.90	− 0.17 ± 1.06	− 0.15 ± 0.96
$\sigma(cc)$	+ 0.52 ± 0.45	− 0.32 ± 1.37	− 0.20 ± 0.70	− 0.28 ± 0.64	+ 0.27 ± 1.21	+ 0.01 ± 0.98
$\langle N \rangle$	− 0.14 ± 0.74	− 0.09 ± 1.33	+ 0.23 ± 0.76	+ 0.31 ± 0.83	− 0.13 ± 1.14	− 0.18 ± 0.92
$\sigma(N)$	+ 0.39 ± 0.76	− 0.18 ± 1.28	− 0.21 ± 0.75	− 0.28 ± 0.85	+ 0.08 ± 1.13	+ 0.20 ± 0.94
$\langle Ecc \rangle$	+ 0.64 ± 1.04	− 0.42 ± 0.91	− 0.22 ± 0.67	+ 0.01 ± 1.08	− 0.09 ± 1.03	+ 0.07 ± 0.87
$\sigma(Ecc)$	+ 0.11 ± 1.02	+ 0.01 ± 1.09	− 0.13 ± 0.86	− 0.38 ± 1.04	+ 0.18 ± 0.93	+ 0.20 ± 0.92
$\langle EC \rangle$	+ 0.73 ± 1.22	− 0.54 ± 0.49	− 0.19 ± 0.64	− 0.18 ± 1.13	+ 0.06 ± 0.92	+ 0.12 ± 0.91
$\sigma(EC)$	+ 0.33 ± 1.09	− 0.13 ± 0.96	− 0.20 ± 0.85	− 0.53 ± 0.82	+ 0.32 ± 1.01	+ 0.21 ± 0.93
$\langle C \rangle$	− 0.64 ± 0.79	+ 0.32 ± 1.02	+ 0.33 ± 0.84	− 0.02 ± 1.08	+ 0.00 ± 1.03	+ 0.02 ± 0.89
$\sigma(C)$	− 0.42 ± 0.55	+ 0.43 ± 1.28	− 0.01 ± 0.83	− 0.87 ± 0.59	+ 0.44 ± 0.89	+ 0.43 ± 0.84
$\langle Sb^{(2)} \rangle$	− 0.09 ± 0.67	− 0.09 ± 1.26	+ 0.18 ± 0.96	+ 0.48 ± 0.87	− 0.40 ± 0.97	− 0.09 ± 0.96
$\sigma(Sb^{(2)})$	+ 0.13 ± 0.53	+ 0.06 ± 1.39	− 0.19 ± 0.84	− 0.45 ± 0.69	+ 0.29 ± 1.13	+ 0.16 ± 0.96
$\langle Sm^{(2)} \rangle$	+ 0.12 ± 0.37	− 0.33 ± 1.52	+ 0.21 ± 0.63	+ 0.55 ± 0.83	− 0.34 ± 1.06	− 0.20 ± 0.85
$\sigma(Sm^{(2)})$	+ 0.18 ± 0.60	− 0.02 ± 1.20	− 0.16 ± 1.08	+ 0.37 ± 0.62	− 0.43 ± 1.24	+ 0.06 ± 0.87
$\langle Sb^{(3)} \rangle$	+ 0.13 ± 0.65	− 0.20 ± 1.29	+ 0.06 ± 0.93	+ 0.41 ± 0.90	− 0.26 ± 1.02	− 0.14 ± 0.94
$\sigma(Sb^{(3)})$	− 1.09 ± 0.68	+ 0.61 ± 0.57	+ 0.48 ± 0.65	− 0.52 ± 0.96	+ 0.21 ± 1.00	+ 0.31 ± 0.82
$\langle Sm^{(3)} \rangle$	− 0.48 ± 0.72	+ 0.51 ± 1.16	− 0.03 ± 0.81	− 0.38 ± 1.01	+ 0.15 ± 1.01	+ 0.23 ± 0.85
$\sigma(Sm^{(3)})$	− 0.66 ± 0.96	+ 0.53 ± 0.89	+ 0.13 ± 0.74	− 0.46 ± 1.05	+ 0.15 ± 0.94	+ 0.30 ± 0.83
$\langle Sb^{(4)} \rangle$	+ 0.30 ± 0.61	− 0.02 ± 1.39	− 0.28 ± 0.72	− 0.28 ± 0.90	+ 0.11 ± 1.10	+ 0.17 ± 0.93
$\sigma(Sb^{(4)})$	+ 0.42 ± 0.62	− 0.25 ± 1.29	− 0.18 ± 0.83	− 0.30 ± 0.91	− 0.01 ± 1.11	+ 0.31 ± 0.87
$\langle Sm^{(4)} \rangle$	+ 1.00 ± 0.71	− 0.69 ± 0.60	− 0.30 ± 0.76	+ 0.35 ± 1.10	− 0.28 ± 0.93	− 0.07 ± 0.84
$\sigma(Sm^{(4)})$	− 0.23 ± 0.95	+ 0.24 ± 1.05	− 0.01 ± 0.94	− 0.57 ± 1.07	+ 0.29 ± 0.83	+ 0.28 ± 0.82
$\langle \alpha^{(\infty)} \rangle$	− 0.22 ± 1.10	− 0.08 ± 1.02	+ 0.30 ± 0.78	+ 0.31 ± 0.95	− 0.25 ± 1.04	− 0.06 ± 0.93
$\sigma(\alpha^{(\infty)})$	− 0.20 ± 1.00	− 0.25 ± 0.86	+ 0.45 ± 0.98	− 0.86 ± 0.71	+ 0.37 ± 0.85	+ 0.49 ± 0.82
$\langle \alpha^{(2)} \rangle$	− 0.50 ± 0.52	+ 0.12 ± 1.28	+ 0.38 ± 0.82	+ 0.19 ± 0.96	− 0.17 ± 1.05	− 0.02 ± 0.96
$\sigma(\alpha^{(2)})$	− 0.04 ± 0.81	− 0.15 ± 1.16	+ 0.19 ± 0.97	− 0.79 ± 1.05	+ 0.20 ± 0.65	+ 0.60 ± 0.68
$\langle \alpha^{(3)} \rangle$	+ 0.02 ± 0.51	− 0.24 ± 1.38	+ 0.22 ± 0.86	+ 0.48 ± 0.78	− 0.33 ± 1.05	− 0.15 ± 0.96
$\sigma(\alpha^{(3)})$	+ 1.03 ± 0.56	− 0.78 ± 0.68	− 0.26 ± 0.69	+ 0.77 ± 1.01	− 0.52 ± 0.78	− 0.25 ± 0.68
Q	+ 1.32 ± 0.33	− 0.91 ± 0.22	− 0.41 ± 0.31	+ 1.24 ± 0.65	− 0.64 ± 0.38	− 0.60 ± 0.37

measurement X as informative if the value obtained for X in a real (RT) text differs from the values of X obtained in *any other* shuffled (SW and SS) text of the considered dataset (see *Condition A* described in the methodology). The results obtained for this type of analysis are shown in Table 1. To facilitate the comparison of measurements taking values in distinct intervals a normalization was applied. For each measurement, and for each of the datasets (Holy Bible and Books), the results are standardized considering all three types of texts (RT, SS, and SW). As such, the average value of each normalized measurement in Table 1 is zero (i.e. RT + SS + SW = 0) and the standard deviation is one.

Considering the Holy Bible dataset, the modularity (Q) was the measurement that best discriminated real from shuffled texts. The modularity in real networks differs 2.24 and 1.73 (standardized values) from the SW and SS versions, respectively. This result suggests that the community structure is much more apparent in real networks, which might be a consequence of the bursty topical textual structure present in real texts (de Arruda et al., 2016a). In addition to the modularity, other measurements were also found to be informative. When comparing RT and SW, the largest differences of values were found for the accessibility ($\sigma(\alpha^{(3)})$), symmetry ($\sigma(Sb^{(3)})$ and $\langle Sm^{(4)} \rangle$) and the clustering coefficient ($\langle cc \rangle$). The best discrimination between RT and SS was found for the symmetry ($\sigma(Sb^{(3)})$ and $\langle Sm^{(4)} \rangle$), accessibility ($\sigma(\alpha^{(3)})$) and closeness ($\langle C \rangle$). Interestingly, several of the measurements were able to distinguish between real and shuffled texts, regardless of the considered shuffling process.

By considering the Books dataset, the modularity also turned out to be the measurement that best discriminated real from shuffled texts. Once again, real texts frequently displayed a clearer community structure. It means that the informativeness achieved by the modularity is a characteristic that seems to depend neither on syntax nor content features. Apart from Q, the following measurements were also found to discriminate real from both shuffled networks: clustering coefficient ($\sigma(C)$), degree ($\sigma(k)$) and accessibility ($\alpha^{(\infty)}$, and $\sigma(\alpha^{(2)})$ and $\sigma(\alpha^{(3)})$).

As a complementary test, for each measurement, we used the *z-score* (see Eq. (7)) to compare a real text and its corresponding shuffled versions (informativeness test based on *Condition B*). Note that this is a less strict informativeness test because, differently from the previous case, we do not compare a real text with shuffled versions from *all texts* of the dataset. Here, we rather compare a real text and the shuffled versions generated only from the *same book*. In Table 2, we show the percentage of documents in which we observed a significant difference between real and shuffled texts – according to the z-score defined in Eq. (7).

Table 2
Percentage of documents in each dataset where the difference between a real text (English part) and the corresponding shuffled version was found to be significant. Apart from the modularity, the informativeness seems to depend on the type of dataset used.

Measurements		SW		SS	
		Holy Bible	Books	Holy Bible	Books
Degree	$\langle k\rangle$	31.25%	30.30%	18.75%	36.36%
	$\sigma(k)$	56.25%	78.79%	31.25%	72.73%
Betweenness	$\langle B\rangle$	37.50%	48.48%	50.00%	57.58%
	$\sigma(B)$	6.25%	66.67%	6.25%	60.61%
Clustering	$\langle cc\rangle$	50.00%	27.27%	50.00%	39.39%
	$\sigma(cc)$	12.50%	30.30%	62.50%	30.30%
Neighborhood	$\langle N\rangle$	6.25%	42.42%	6.25%	63.64%
	$\sigma(N)$	25.00%	39.39%	37.50%	60.61%
Eccentricity	$\langle Ecc\rangle$	37.50%	30.30%	43.75%	45.45%
	$\sigma(Ecc)$	37.50%	51.52%	56.25%	54.55%
Eigenvector	$\langle EC\rangle$	62.50%	57.58%	62.50%	57.58%
	$\sigma(EC)$	31.25%	78.79%	37.50%	66.67%
Closeness	$\langle C\rangle$	43.75%	39.39%	62.50%	51.52%
	$\sigma(C)$	12.50%	90.91%	25.00%	93.94%
Symmetry	$\langle Sb^{(2)}\rangle$	6.25%	54.55%	25.00%	51.52%
	$\sigma(Sb^{(2)})$	6.25%	42.42%	12.50%	42.42%
	$\langle Sm^{(2)}\rangle$	0.00%	57.58%	6.25%	63.64%
	$\sigma(Sm^{(2)})$	6.25%	30.30%	12.50%	42.42%
	$\langle Sb^{(3)}\rangle$	18.75%	45.45%	18.75%	48.48%
	$\sigma(Sb^{(3)})$	93.75%	36.36%	93.75%	48.48%
	$\langle Sm^{(3)}\rangle$	56.25%	57.58%	25.00%	66.67%
	$\sigma(Sm^{(3)})$	75.00%	57.58%	68.75%	57.58%
	$\langle Sb^{(4)}\rangle$	0.00%	54.55%	37.50%	66.67%
	$\sigma(Sb^{(4)})$	18.75%	63.64%	37.50%	66.67%
	$\langle Sm^{(4)}\rangle$	93.75%	54.55%	87.50%	54.55%
	$\sigma(Sm^{(4)})$	31.25%	54.55%	31.25%	66.67%
Accessibility	$\langle \alpha^{(\infty)}\rangle$	0.00%	54.55%	37.50%	54.55%
	$\sigma(\alpha^{(\infty)})$	12.50%	93.94%	18.75%	93.94%
	$\langle \alpha^{(2)}\rangle$	12.50%	42.42%	50.00%	42.42%
	$\sigma(\alpha^{(2)})$	18.75%	51.52%	12.50%	81.82%
	$\langle \alpha^{(3)}\rangle$	0.00%	54.55%	25.00%	51.52%
	$\sigma(\alpha^{(3)})$	100.00%	66.67%	87.50%	63.64%
Modularity	Q	100.00%	100.00%	100.00%	100.00%

As found in the first test, Q is the most critical measurement for both of the considered datasets, reaching 100% of informativeness. Other measurements had similar results for both datasets, eg., $\sigma(k)$, $\langle Sm^{(3)}\rangle$, and $\langle cc\rangle$, which we found to be informative for approximately 50% of the samples. However, for many other measurements, the level of informativeness varied according to the dataset. For example, $\sigma(Sb^{(3)})$ and $\sigma(\alpha^{(3)})$ were found to be more informative in the Holy Bible dataset. Conversely, $\sigma(C)$ and $\langle Sm^{(2)}\rangle$ seemed to be more informative in the Books dataset.

All in all, the results obtained here suggest that, apart from the modularity, it is important to analyze the characteristics of the dataset to decide if network measurements extracted from paragraph networks can be classified as informative – even if a less strict definition of informativeness is taken into account. Interestingly, the results obtained here indicate that paragraph networks are less informative than other types of text networks, as supported by Amancio (2015c). In the case of word adjacency networks, most of the measurements were found to be informative, independently of the characteristics of the considered datasets.

5.2. Dependency on syntax and content

In this section, we evaluate the dependency of the measurements by considering their variability in two distinct scenarios: in datasets where (i) the content (text) is constant and the language (syntax) is varied; and (ii) the language is constant and the content varies. To represent (i), we used the Holy Bible dataset. The dataset employed in the second scenario was created by selecting only the Books in English from the Books dataset. We decided to use the English language because, in the considered dataset, a larger number of books written in this language is available.

In the first analysis, we identified the measurements that were able to capture syntax/language subtleties. The measurements that were found to display significant variability in this scenario (i.e. in the Holy Bible dataset) were: accessibility ($\langle \alpha^3\rangle$ and $\sigma(\alpha^2)$), degree ($\langle k\rangle$), eccentricity ($\sigma(Ecc)$), symmetry ($\langle Sb^2\rangle$, $\langle Sb^3\rangle$, $\sigma(Sm^4)$ and $\sigma(Sb^2)$), neighborhood ($\langle N\rangle$) and betweenness ($\sigma(B)$).

We also identified the measurements that are sensitive to changes in content. The measurements taking the highest coefficients of variation in the English Books datasets were: eccentricity ($\langle Ecc\rangle$), closeness ($\langle C\rangle$), symmetry ($\langle Sm^4\rangle$ and $\langle Sb^4\rangle$), betweenness ($\langle B\rangle$), degree ($\langle k\rangle$), eigenvector centrality ($\langle EC\rangle$), accessibility ($\langle \alpha^2\rangle$ and $\langle \alpha^\infty\rangle$) and neighborhood ($\langle N\rangle$). Note that some measurements might depend on both syntax and content. This is the case of $\langle N\rangle$. Interestingly, for both symmetry and accessibility, the ability to

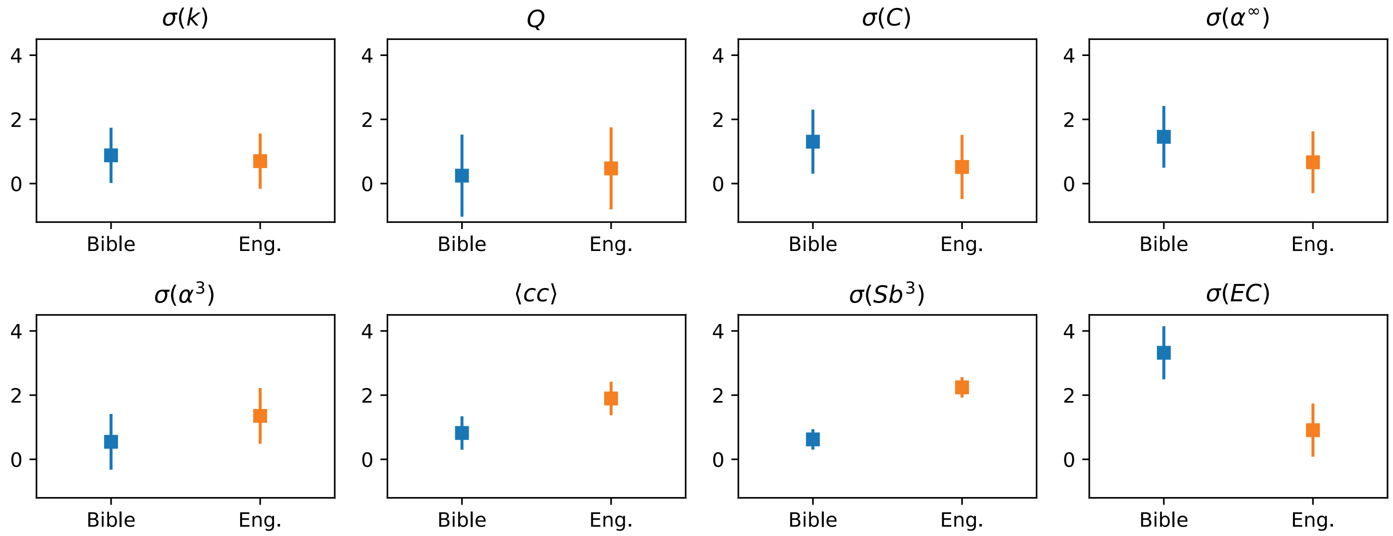

Fig. 2. Comparisons among some of the measurements that provided lower values of CV. The error bars represent the confidence interval of a mean for the interval of 95% of confidence.

capture syntax or content subtleties depends on the hierarchical level being analyzed.

In addition to the tests as mentioned above, we probed, for each measurement, which of the two phenomena is more prevalent: (i) the ability to detect changes in syntax; and (ii) the ability to detect changes in content. This prevalence analysis was conducted by comparing the coefficient of variation in the considered datasets, as described in the methodology. The obtained results are shown in Fig. 2. The top sub-panels illustrate the results obtained for $\sigma(k)$, Q, $\sigma(C)$ and $\sigma(\alpha^{\infty})$. In most of these cases, while the variability across languages (Bible dataset) or topics (English dataset) is high, there is no significant difference between these values. This means that, for these measurements, both syntax and content are captured.

A different behavior can be observed for the measurements depicted in the bottom sub-panels of Fig. 2. For both $\langle cc \rangle$ and $\sigma(Sb^{(3)})$, a significant difference of coefficients of variation was found. This result is an interesting finding in text networks since measurements extracted from other texts networks (such as co-occurrence networks) are mostly dependent on syntax (Amancio et al., 2013). By considering that co-occurrence networks represent distinct information, this result suggests that paragraph-based networks can be used to complement the analysis based on traditional co-occurrence networks when both syntax and content are relevant for the problem being addressed.

5.3. Classification tests

To illustrate the applicability of the paragraph-based network in classification tasks, some classification problems were tackled using paragraph-based networks. In the first example, we considered the problem of deciding whether a manuscript has a structure compatible with a shuffled, meaningless document. In the second classification problem, we probed whether an unknown text – the Voynich manuscript – can be considered compatible with real texts.

5.3.1. Discriminating between real and shuffled texts

We applied our method to distinguish real from shuffled texts in order to illustrate the capabilities of paragraph-based networks to characterize texts regarding a real application. For each book presented in Appendix A, the three paragraph-based networks were created, RT, SW, and SS. After that, the network measurements described previously were extracted, the values were standardized, and those values were used as classification features. To select the features for this task, we considered the most informative measurements obtained from Table 1. More specifically, for each pair of real vs. shuffled texts (i.e., RT vs. SS and RT vs. SW) we identified the top 10 measurements that provide the best discrimination. Then, we selected those measurements appearing in both top 10 lists. The measurements selected as feature vectors are: Q, $\sigma(C)$, $\sigma(\alpha^{(\infty)})$, $\sigma(k)$, $\sigma(B)$, $\sigma(\alpha^{(3)})$, $\sigma(\alpha^{(2)})$, $\langle Sm^{(2)} \rangle$, and $\sigma(EC)$.

The classification was evaluated by the live-one-out cross-validation and the SMO classifier algorithm, which is an SVM implementation available in Weka (Hall et al., 2009; Witten, Frank, Hall, & Pal, 2016). The parameters were chosen according to the procedure defined in (Amancio et al., 2014). When considering three classes (RT, SW, and SS), the accuracy was 76.08%. However, the true positive rate was 0.98 for the RT samples and 0.71 and 0.50 for SW and SS, respectively. The false negative rates were 0.02, 0.24, and 0.14 for RT, SW, and SS, respectively. These results mean that the proposed framework can easily differentiate between real and shuffled texts. Conversely, the discrimination between the two classes of shuffled documents represents a more challenging task.

A variation of the same classification problem considered both shuffled versions as having the same class. In this case, SVM reached 98.72% of accuracy. The false positive rate of the RT networks was 0, and the only two classification mistakes were made by real texts classified as shuffled texts. Fig. 3 illustrates the separation between the two classes by considering the projection into a

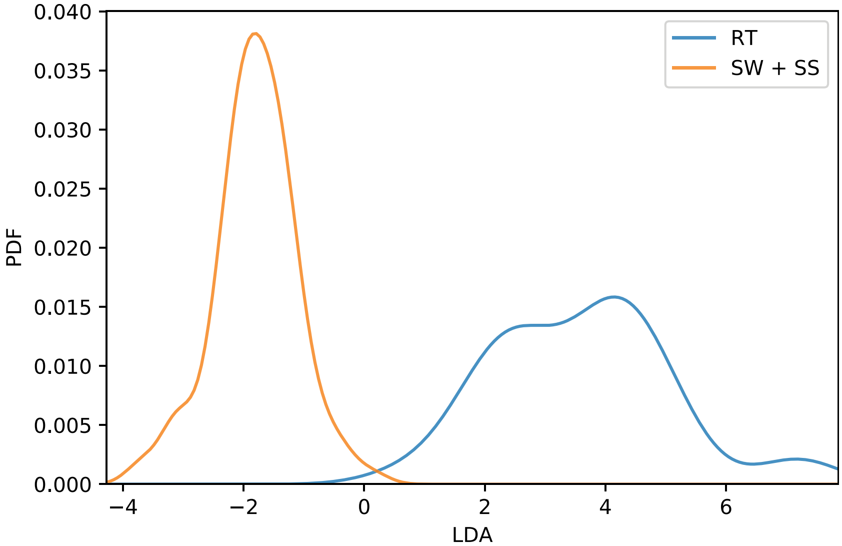

Fig. 3. Probability density function (pdf) of the linear discriminant analysis projection obtained from the selected features in the classification of texts in two classes: real vs. shuffled (SW and SS) texts.

single dimension obtained via linear discriminant analysis (Friedman, Hastie, & Tibshirani, 2001). Given the importance stressed by the modularity in the informativeness analysis, we also evaluated the performance when only this measurement is used for the classification. In this case, the accuracy rate reached 96.79%, which confirms the importance of the modularity in discriminating real and shuffled texts.

We also compared our methodology to a *word2vec* based approach. Here we considered the version implemented in *Gensim* library[1], which is based on an embedding model (Le & Mikolov, 2014), and negative sampling method (Mikolov et al., 2013b). Also, we adopted the distributed memory version, and we employed the mean of the context word vectors (Dai et al., 2015). The size of the vectors was set as 200 elements, the window size (maximum distance between the words that represent the current and predicted positions in a sentence) was 8. We disregarded words with an occurring less than 19 times (this value was set to provide a similar vocabulary size to the methodology presented by (Dai et al., 2015)). The model number of iterations was set to 10. Due to the characteristics of this type of method, we considered only the English part of our dataset. Because our dataset consists of a small set of books, we pre-trained our *word2vec* by using the entire English version of Wikipedia[2]. Besides, in order to increase the number of English documents to this test, we included the following books: The Book of Snobs (by William Makepeace Thackeray), The Unbearable Bassington (by Saki), The Works of Edgar Allan Poe – Volumes 1, 3 and 5 (by Edgar Allan Poe), Through the Magic Door (by Arthur Conan Doyle), and When William Came (by Saki). Observe that these books were not considered in the other tests presented here because they contain a fewer number of paragraphs than Voynich manuscript.

Fig. 4shows a PCA (Principal Component Analysis) (Jolliffe, 2002) projection of the *word2vec* vectors computed from the three versions of the books (RT, SW, and SS). By considering the three classes and the SVM classifier, with the same characteristics as employed before, we found the accuracy of 70.8%. Interestingly, all samples of SW were correctly classified. Note that this methodology was able to classify texts that do not follow the English grammar (SW) with 100% of the samples being right classified. However, in the case of RT, many texts were classified as SS, which are either without meaning. More specifically, in 57.5% of the RT samples were classified as SS. In order to compare the *word2vec* based results to the previous cases, we classified only between two classes, Real (RT) and the shuffled versions (SW and SS). For this test, we found the accuracy of 63, 3%, which is a low value when we consider only two classes. Observe that 67.5% of the RT samples were incorrectly classified as a shuffled text (SW + SS group). From these results, we conclude that the employed methodology is not able to distinguish between RT and SS.

5.3.2. *Case example: Voynich manuscript*

The Voynich manuscript is known to be a mysterious text, and many of its aspects have been studied for several years (Reddy & Knight, 2011). Some studies have relied on textual analysis (Reddy & Knight, 2011), while others have used complex networks tools to study its properties (Amancio et al., 2013; Montemurro & Zanette, 2013). In order to handle the manuscript – originally written in an unknown alphabet –, it is necessary to translate its characters into a known set of symbols. Here we used the European Voynich Alphabet (EVA) (Znadbergen, 2018), which provides the original characters manually translated into European characters. To provide a better quality translation, for each line of the text, different translations are available. Here we considered the voting of the most recurring character for all different translations of the same line. Additionally, because our approach relies on text paragraphs, we detected paragraphs by visually inspecting the original manuscript. The paragraphs were identified by a single person, by following the criteria of the distance between chunks of texts. We believe that the possible mistakes are not significative to the final results since the average size of the Voynich paragraphs is compatible with the books dataset statistics, as shown in Fig. 5. When comparing the Voynich manuscript with shuffled texts, we disregarded the SS versions because there is no trivial way to detect sentences in the Voynich manuscript.

First, we analyzed if the Voynich manuscript, when characterized by the metrics extracted from paragraph-based networks, is compatible with real texts and not compatible with gibberish, shuffled texts. It is a long-standing question about the manuscript since several scholars have questioned the existence of a meaningful textual structure in this mysterious text (Belfield, 2007). An illustration comparing the structure of the Voynich manuscript and a shuffled network is shown in Fig. 6. It is clear from the visualizations that the Voynich manuscript presents a well-defined community structure, with two dominant groups. The communities seem to capture the topical organization of the manuscript in some degree: the extract about plants seems to be separated in a specific community. The equivalent shuffled network, shown in Fig. 6(b) reveals no apparent community structure. Since the modularity was found to be informative in the previous analysis, the organization in communities in Fig. 6(a) suggests, at the paragraph level, that the Voynich manuscript is not compatible with shuffled texts. Observe that the same conclusion has been reported when different types of networks are used to represent the manuscript (Amancio et al., 2013; Montemurro & Zanette, 2013; Reddy & Knight, 2011).

As a complementary analysis, taking into account the community structure of the networks, we analyzed the modularity Q, which is much higher for the set of real texts (RT) when compared to the set of shuffled texts (SS and SW), as shown in Fig. 7. The modularity obtained for the Voynich (represented with a blue arrow in the figure) is not compatible with any of the two distributions obtained for shuffled texts. On the other hand, the modularity of the manuscript is compatible with the modularity extracted from real texts.

In order to analyze the Voynich manuscript, we employed the same classifier as in the previous section. As a result, the document was classified as real text. A set of 30 SW networks of Voynich were also classified, and the accuracy of 100% was found. This perfect classification can also be seen in Fig. 7, which shows that the generated SW networks of Voynich (orange arrows) are mostly

[1] https://radimrehurek.com/gensim/

[2] https://www.wikipedia.org

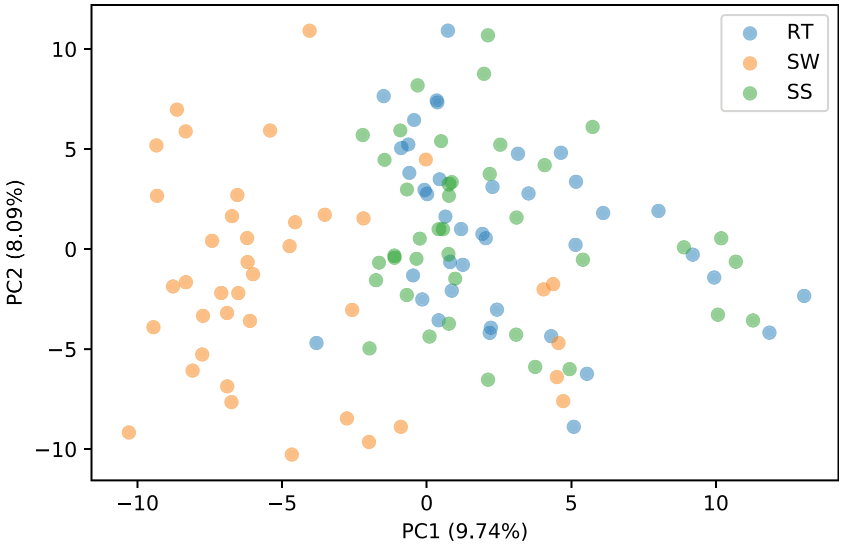

Fig. 4. PCA projection computed from *doc2vec* features. Observe that there are two groups, in which the first represents SW, and the second group represents RT and SS.

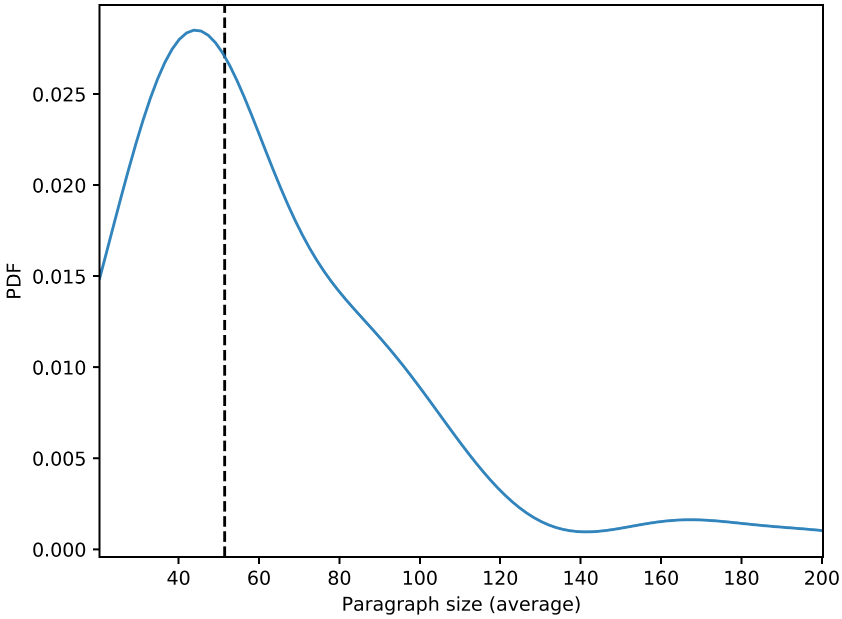

Fig. 5. Probability density function (pdf) of the average number of words of the books paragraphs obtained from the entire dataset. The dashed line represents the average computed for Voynich manuscript.

compatible with the distributions obtained for shuffled texts.

6. Conclusions

In the current study, we probed the properties of a paragraph-based networked representation of texts. Interestingly, we found that the most informative measurement is the modularity since artificial, shuffled texts, are not organized in well-defined communities. Our results also revealed that several measurements are able to capture content. It is an important feature since the well-known word adjacency (co-occurrence) networks are only able to capture syntax features. Our findings suggest that both co-occurrence and paragraph-based networks can be used in a complementary way when both syntax and content features are important for a natural language processing task.

The adopted network representation was used to analyze the statistical nature of the Voynich manuscript. Previous studies hinging on word networks showed that the Voynich syntax is coherent with natural languages (Amancio et al., 2013; Montemurro &

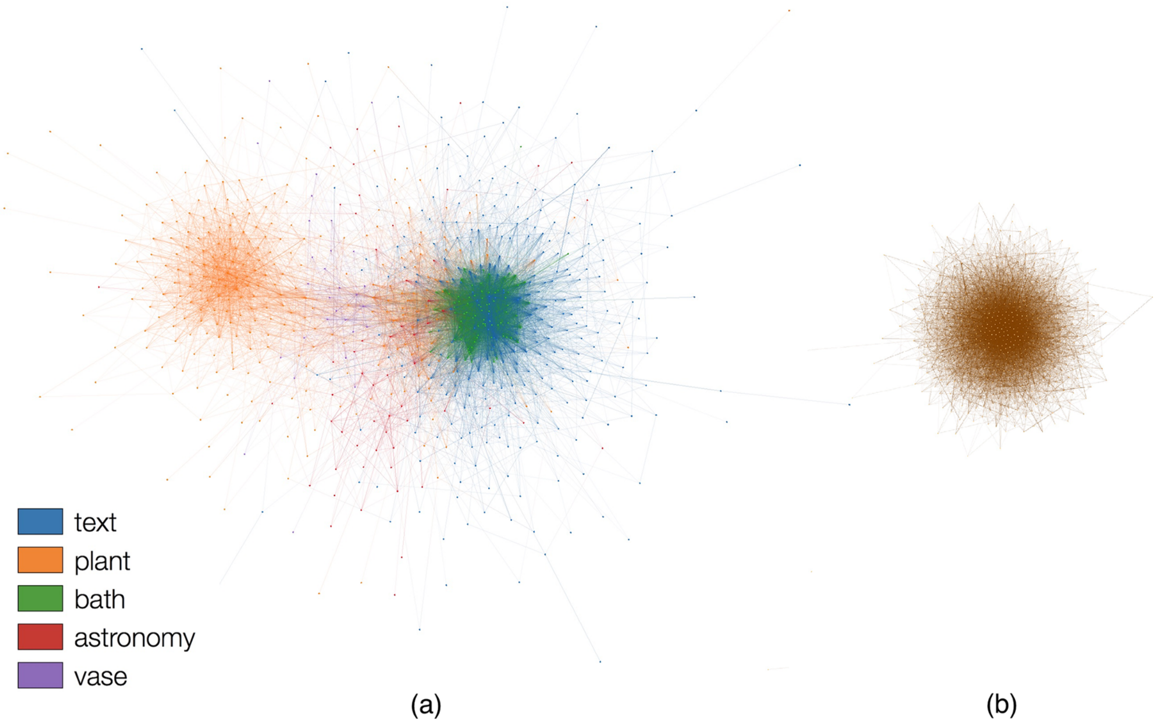

Fig. 6. Network visualizations of the two version of paragraph- based networks of Voynich manuscript. In (a) the paragraphs were labelled by considering the figures in the corresponding pages. The topics considered were: (i) *text*, when no images are available; (ii) *plants*; (iii) *bath*, with figures of women and bath-like shape; (iv) *astronomy*, with spatial-like figures; and (v) *vase*. In (b) the network visualization represents the RW version of the Voynich manuscript. The visualization was provided by the software implemented by Silva, Amancio, Bardosova, Costa, and Oliveira Jr (2016a).

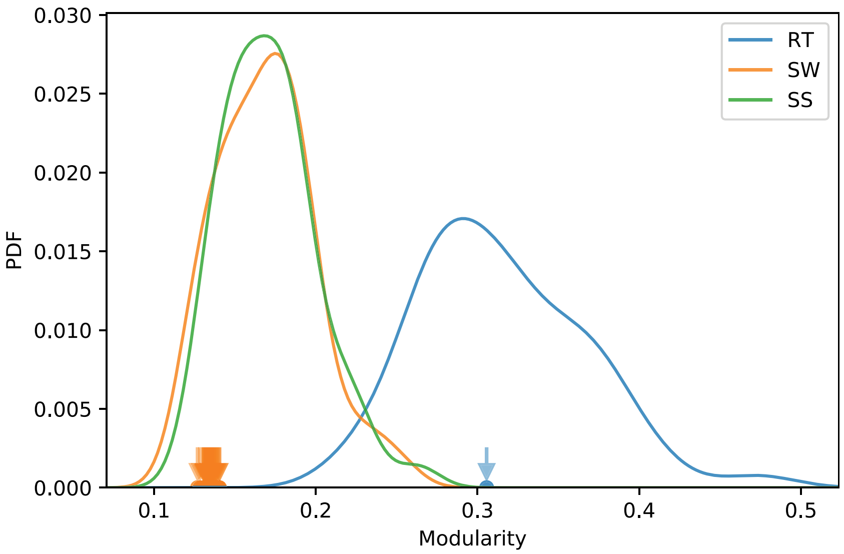

Fig. 7. Probability density function (pdf) of the modularity measurement for three types of networks (RT, SW, and SS). The modularity of the two versions of the Voynich manuscript RT and SW are represented by the blue and orange arrows, respectively. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Zanette, 2013). Recently, an extensive analysis using several natural languages argued that Hebrew is the most probable language of the manuscript (Hauer & Kondrak, 2016). Here, we proposed a different analysis, by focusing on the organization in paragraphs. Our analysis revealed that the Voynich manuscript is compatible with natural languages at the paragraph level. This finding was confirmed by analyzing the organization of the text into well-defined communities: similarly to several natural languages, the Voynich manuscript also displays a clear community structure organization. Furthermore, we applied our classification approach, and the Voynich manuscript was classified as real text. As a complement, the accuracy of 100% was found when we classified 30 samples of the shuffled version of the Voynich manuscript.

Because we could identify the features that are more dependent either on language or content, we paved the way to new applications regarding the proposed features. Here, we have found promising results for the task of discriminating between real and shuffled texts. Furthermore, the proposed features were employed to analyze the Voynich manuscript, which was found to be compatible with natural languages at the paragraph level. This result suggests that the attempts to decipher the manuscript are not worthless. Apart from the presented results, many other tasks could employ the same methodology, or use paragraph-based features as a complement to the standard features. Some examples of tasks are authorship and plagiarism identification (Amancio, 2015a), automatic identification of literary movements (Amancio, Oliveira Jr, & da Fontoura Costa, 2012b), fraud detection (Wang & Xu, 2018), among others (Salloum, Al-Emran, Monem, & Shaalan, 2017).

Acknowledgments

Henrique F. de Arruda Henrique F. de Arruda acknowledges the Coordena£o de Aperfeioamento de Pessoal de Nvel Superior - Brasil (CAPES) - Finance Code 001. Vanessa Q. Marinho thanks FAPESP (grant no. 2015/05676-8) for financial support. Luciano da F. Costa thanks CNPq (grant no. 307333/2013-2) and NAP-PRP-USP for sponsorship. Diego R. Amancio acknowledges FAPESP (grant nos. 16/19069-9 and 17/13464-6) for financial support. This work has been supported also by FAPESP grants 11/50761-2 and 2015/22308-2. The authors acknowledge Filipi Nascimento Silva for fruitful conversations.

Appendix A. Dataset

The list of books used to analyze how the network structure varies across different documents in the *same language* is shown below. Five different languages were considered: English, French, German, Italian and Portuguese. The list of books is organized by language. The author of each book is listed between parentheses after each title. The books were obtained from the Project Gutenberg[3].

1. **English:** The Adventures of Sherlock Holmes (Arthur Conan Doyle) – 2523 paragraphs, The Tragedy of the Korosko (Arthur Conan Doyle) – 871 paragraphs, The Valley of Fear (Arthur Conan Doyle) – 1523 paragraphs, Uncle Bernac - A Memory of the Empire (Arthur Conan Doyle) – 1211 paragraphs, Dracula's Guest (Bram Stoker) – 739 paragraphs, The Lair of the White Worm (Bram Stoker) – 871 paragraphs, The Jewel Of Seven Stars (Bram Stoker) – 1194 paragraphs, The Man (Bram Stoker) – 1834 paragraphs, The Mystery of the sea (Bram Stoker) – 1625 paragraphs, A Tale of Two Cities (Charles Dickens) – 3268 paragraphs, Barnaby Rudge: A Tale of the Riots of Eighty (Charles Dickens) – 4598 paragraphs, American Notes (Charles Dickens) – 1022 paragraphs, Great Expectations (Charles Dickens) – 3835 paragraphs, Hard Times (Charles Dickens) – 2217 paragraphs, The Works of Edgar Allan Poe – Volume 2 – 906 paragraphs, The Works of Edgar Allan Poe – Volume 4 (Edgar Allan Poe) – 961 paragraphs, Beasts and Super-Beasts (Hector H. Munro) – 1319 paragraphs, The Chronicles of Clovis (Hector H. Munro) – 1006 paragraphs, The Toys of Peace (Hector H. Munro) – 1077 paragraphs, The Girl on the Boat (P. G. Wodehouse) – 2425 paragraphs, My Man Jeeves (P. G. Wodehouse) – 1943 paragraphs, Something New (P. G. Wodehouse) – 2259 paragraphs, The Adventures of Sally – 2357 paragraphs, The Clicking of Cuthbert (P. G. Wodehouse) – 1876 paragraphs, A Pair of Blue Eyes (Thomas Hardy) – 3666 paragraphs, Far from the Madding Crowd (Thomas Hardy) – 3407 paragraphs, Jude the Obscure (Thomas Hardy) – 2990 paragraphs, The Mayor of Casterbridge (Thomas Hardy) – 2117 paragraphs, The Hand of Ethelberta (Thomas Hardy) – 3112 paragraphs, Barry Lyndon (William M. Thackeray) – 1169 paragraphs, The History of Pendennis – 924 paragraphs, The Virginians (William M. Thackeray) – 1410 paragraphs, and Vanity Fair (William M. Thackeray) – 2589 paragraphs;
2. **French:** Le fils du Soleil (Gustave Aimard) – 2337 paragraphs, Face au Drapeau (Jules Verne) – 1487 paragraphs, Pierre de Villerglé (Louis Amédée Achard) – 1106 paragraphs, Les Idoles d'argile (Louis Reybaud) – 1083 paragraphs, and Han d'Islande (Victor Hugo) – 3983 paragraphs;
3. **German:** Die Wahlverwandtschaften (Goethe) – 3503 paragraphs, Der Moloch (Jakob Wassermann) – 1483 paragraphs, Königliche Hoheit – 1208 paragraphs, and Lichtenstein (Wilhelm Hauff) – 1770 paragraphs;
4. **Italian:** Il Peccato di Loreta (Alberto Boccardi) – 1592 paragraphs, La Montanara (Anton Giulio Barrili) – 2481 paragraphs, Alla Finestra (Enrico Castelnuovo) – 2069 paragraphs, Sciogli la treccia, Maria Maddalena (Guido da Verona) – 1834 paragraphs and La Pergamena Distrutta (Virginia Mulazzi) – 5861 paragraphs;
5. **Portuguese:** Amor de Perdi£o (Camilo Castelo Branco) – 1612 paragraphs, A Cidade e as Serras (Eça de Queirós) – 1453 paragraphs, Os Bravos do Mindello (Faustino da Fonseca) – 2115 paragraphs, Transviado (Jaime de Magalhães Lima) – 1947 paragraphs, and Uma Família Inglesa (Júlio Dinis) – 5241 paragraphs.

References

Agirre, E., & Soroa, A. (2009). *Personalizing pagerank for word sense disambiguation. Proceedings of the 12th conference of the european chapter of the association for*

[3] http://www.gutenberg.org

computational linguistics. Association for Computational Linguistics33–41.

Amancio, D. R. (2015a). Comparing the topological properties of real and artificially generated scientific manuscripts. *Scientometrics, 105*(3), 1763–1779.

Amancio, D. R. (2015b). A complex network approach to stylometry. *PLoS One, 10*(8), e0136076.

Amancio, D. R. (2015c). Probing the topological properties of complex networks modeling short written texts. *PLoS ONE, 10*(2), e0118394.

Amancio, D. R., Altmann, E. G., Oliveira Jr., O. N., & Costa, L. F. (2011). Comparing intermittency and network measurements of words and their dependence on authorship. *New Journal of Physics, 13*(12), 123024.

Amancio, D. R., Altmann, E. G., Rybski, D., Oliveira Jr, O. N., & Costa, L. F. (2013). Probing the statistical properties of unknown texts: application to the voynich manuscript. *PLoS ONE, 8*(7), e67310.

Amancio, D. R., Comin, C. H., Casanova, D., Travieso, G., Bruno, O. M., Rodrigues, F. A., & Costa, L. F. (2014). A systematic comparison of supervised classifiers. *PLoS One, 9*(4), e94137.

Amancio, D. R., Nunes, M. G., Oliveira Jr, O. N., & Costa, L.d. F. (Nunes, Oliveira Jr, Costa, 2012a). Extractive summarization using complex networks and syntactic dependency. *Physica A: Statistical Mechanics and its Applications, 391*(4), 1855–1864.

Amancio, D. R., Oliveira Jr, O. N., & da Fontoura Costa, L. (Oliveira Jr, da Fontoura Costa, 2012b). Identification of literary movements using complex networks to represent texts. *New Journal of Physics, 14*(4), 043029.

Amancio, D. R., Silva, F. N., & Costa, L. F. (2015). Concentric network symmetry grasps authors' styles in word adjacency networks. *EPL (Europhysics Letters), 110*(6), 68001.

Angelova, R., & Weikum, G. (2006). *Graph-based text classification: learn from your neighbors. Proceedings of the 29th annual international ACMSIGIR conference on research and development in information retrieval*. ACM485–492.

de Arruda, G. F., Barbieri, A. L., Rodríguez, P. M., Rodrigues, F. A., Moreno, Y., & Costa, L. F. (2014). Role of centrality for the identification of influential spreaders in complex networks. *Physical Review E, 90*, 032812.

de Arruda, H. F., Costa, L. F., & Amancio, D. R. (Costa, Amancio, 2016a). Topic segmentation via community detection in complex networks. *Chaos: An Interdisciplinary Journal of Nonlinear Science, 26*(6), 063120.

de Arruda, H. F., Costa, L. F., & Amancio, D. R. (Costa, Amancio, 2016b). Using complex networks for text classification: discriminating informative and imaginative documents. *EPL (Europhysics Letters), 113*(2), 28007.

de Arruda, H. F., Marinho, V. Q., Lima, T. S., Amancio, D. R., & Costa, L. F. (Marinho, Lima, Amancio, Costa, 2018a). An image analysis approach to text analytics based on complex networks. *Physica A: Statistical Mechanics and its Applications, 510*, 110–120.

de Arruda, H. F., Silva, F. N., Marinho, V. Q., Amancio, D. R., & Costa, L. F. (Silva, Marinho, Amancio, Costa, 2018b). Representation of texts as complex networks: A mesoscopic approach. *Journal of Complex Networks, 6*(1), 125–144.

Belbachir, F., & Boughanem, M. (2018). Using language models to improve opinion detection. *Information Processing & Management, 54*(6), 958–968.

Belfield, R. (2007). *The six unsolved ciphers: Inside the mysterious codes that have confounded the World's greatest cryptographers.* Ulysses Press.

Bonacich, P. (1987). Power and centrality: A family of measures. *American Journal of Sociology,* 1170–1182.

Cancho, R. F., & Solé, R. V. (2001). The small world of human language. *Proceedings of the Royal Society of London B: Biological Sciences, 268*(1482), 2261–2265.

Cancho, R. F., & Solé, R. V. (2003). Least effort and the origins of scaling in human language. *Proceedings of the National Academy of Sciences, 100*(3), 788–791.

Cancho, R. F., Solé, R. V., & Köhler, R. (2004). Patterns in syntactic dependency networks. *Physical Review E, 69*(5), 051915.

Cong, J., & Liu, H. (2014). Approaching human language with complex networks. *Physics of Life Reviews, 11*(4), 598–618.

Costa, L. F., Rodrigues, F. A., Travieso, G., & Villas Boas, P. R. (2007). Characterization of complex networks: A survey of measurements. *Advances in Physics, 56*(1), 167–242.

Dai, A. M., Olah, C., & Le, Q. V. (2015). Document embedding with paragraph vectors. arXiv:1507.07998.

Erkan, G., & Radev, D. R. (2004). Lexrank: Graph-based lexical centrality as salience in text summarization. *Journal of Artificial Intelligence Research, 22*, 457–479.

Freeman, L. (1977). A set of measures of centrality based on betweenness. *Sociometry, 40*, 35–41.

Friedman, J., Hastie, T., & Tibshirani, R. (2001). *The elements of statistical learning. 1*. Springer series in statistics New York.

Giatsoglou, M., Vozalis, M. G., Diamantaras, K., Vakali, A., Sarigiannidis, G., & Chatzisavvas, K. C. (2017). Sentiment analysis leveraging emotions and word embeddings. *Expert Systems with Applications, 69*, 214–224.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., & Witten, I. H. (2009). The weka data mining software: An update. *ACM SIGKDD Explorations Newsletter, 11*(1), 10–18.

Harary, F. (1969). *Graph theory. Addison-Wesley Series in Mathematics*Addison Wesley.

Hauer, B., & Kondrak, G. (2016). Decoding anagrammed texts written in an unknown language and script. *Transactions of the Association for Computational Linguistics, 4*, 75–86.

Henry, S., Cuffy, C., & McInnes, B. (2017). Evaluating feature extraction methods for knowledge-based biomedical word sense disambiguation. *BioNLP,* 272–281.

Hsu, F.-Y., Lee, H.-M., Chang, T.-H., & Sung, Y.-T. (2018). Automated estimation of item difficulty for multiple-choice tests: An application of word embedding techniques. *Information Processing & Management, 54*(6), 969–984.

Jin, W., & Srihari, R. K. (2007). *Graph-based text representation and knowledge discovery. Proceedings of the 2007 ACM symposium on applied computing.* ACM807–811.

Jolliffe, I. (2002). *Principal component analysis.* Wiley Online Library.

Kim, S. G., & Kang, J. (2018). Analyzing the discriminative attributes of products using text mining focused on cosmetic reviews. *Information Processing & Management, 54*(6), 938–957.

Le, Q., & Mikolov, T. (2014). *Distributed representations of sentences and documents. International conference on machine learning*1188–1196.

Liu, H., & Cong, J. (2013). Language clustering with word co-occurrence networks based on parallel texts. *Chinese Science Bulletin, 58*(10), 1139–1144.

Manning, C. D., & Schütze, H. (1999). *Foundations of statistical natural language processing.* Cambridge, MA, USA: MIT Press.

Margan, D., Martinčić-Ipšić, S., & Meštrović, A. (Martinčić-Ipšić, Meštrović, 2014a). *Network differences between normal and shuffled texts: Case of croatian. Complex networks V.springer*. Springer275–283.

Margan, D., Mestrovic, A., & Martincic-Ipsic, S. (Mestrovic, Martincic-Ipsic, 2014b). *Complex networks measures for differentiation between normal and shuffled croatian texts. Information and communication technology, electronics and microelectronics (MIPRO), 2014 37th international convention on*. IEEE1598–1602.

Marinho, V. Q., de Arruda, H. F., Lima, T. S., Costa, L. F., & Amancio, D. R. (2017). *On the "calligraphy" of books. Textgraphs.* Association for Computational Linguistics1–10.

Masucci, A. P., & Rodgers, G. J. (2006). Network properties of written human language. *Physical Review E, 74*(2), 026102.

Masucci, A. P., & Rodgers, G. J. (2009). Differences between normal and shuffled texts: structural properties of weighted networks. *Advances in Complex Systems, 12*(01), 113–129.

Mehri, A., Darooneh, A. H., & Shariati, A. (2012). The complex networks approach for authorship attribution of books. *Physica A: Statistical Mechanics and its Applications, 391*(7), 2429–2437.

Metzler, D., & Croft, W. B. (2004). Combining the language model and inference network approaches to retrieval. *Information Processing & Management, 40*(5), 735–750.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013a). Efficient estimation of word representations in vector space. arXiv:1301.3781.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (Sutskever, Chen, Corrado, Dean, 2013b). *Distributed representations of words and phrases and their compositionality. Advances in neural information processing systems*3111–3119.

Miller, G. A. (1995). Wordnet: A lexical database for english. *Communications of the ACM, 38*(11), 39–41.

Montemurro, M. A., & Zanette, D. H. (2013). Keywords and co-occurrence patterns in the voynich manuscript: An information-theoretic analysis. *PLoS One, 8*(6), e66344.

Newman, M. (2010). *Networks: An introduction.* New York, NY, USA: Oxford University Press, Inc.

Newman, M. E. (2005). Power laws, pareto distributions and zipf's law. *Contemporary Physics, 46*(5), 323–351.

Newman, M. E. J., & Girvan, M. (2004). Finding and evaluating community structure in networks. *Physical Review E, 69*(026113).

Posadas-Durán, J.-P., Gómez-Adorno, H., Sidorov, G., Batyrshin, I., Pinto, D., & Chanona-Hernández, L. (2017). Application of the distributed document representation in the authorship attribution task for small corpora. *Soft Computing, 21*(3), 627–639.

Reddy, S., & Knight, K. (2011). *What we know about the voynich manuscript. Proceedings of the 5th ACL-HLT workshop on language technology for cultural heritage, social sciences, and humanities*. Association for Computational Linguistics78–86.

Salloum, S. A., Al-Emran, M., Monem, A. A., & Shaalan, K. (2017). A survey of text mining in social media: Facebook and twitter perspectives. *Advancesin Scienceand Technologyand Engineering Systems Journal, 2*(1), 127–133.

Salton, G., Singhal, A., Mitra, M., & Buckley, C. (1997). Automatic text structuring and summarization. *Information Processing & Management, 33*(2), 193–207.

Segarra, S., Eisen, M., & Ribeiro, A. (2015). Authorship attribution through function word adjacency networks. *IEEE Transactions on Signal Processing, 63*(20), 5464–5478.

Sicilia, R., Giudice, S. L., Pei, Y., Pechenizkiy, M., & Soda, P. (2018). Twitter rumour detection in the health domain. *Expert Systems with Applications*.

Silva, F. N., Amancio, D. R., Bardosova, M., Costa, L. F., & Oliveira Jr, O. N. (Amancio, Bardosova, Costa, Oliveira Jr, 2016a). Using network science and text analytics to produce surveys in a scientific topic. *Journal of Informetrics, 10*(2), 487–502.

Silva, F. N., Comin, C. H., Peron, T. K., Rodrigues, F. A., Ye, C., Wilson, R. C., et al. (Comin, Peron, Rodrigues, Ye, Wilson, Hancock, Costa, 2016b). Concentric network symmetry. *Information Science, 333*, 61–80.

Silva, T. C., & Amancio, D. R. (2012). Word sense disambiguation via high order of learning in complex networks. *EPL (Europhysics Letters), 98*(5), 58001.

Strogatz, S. H. (2001). Exploring complex networks. *Nature, 410*(6825), 268–276. https://doi.org/10.1038/35065725.

Symeonidis, S., Effrosynidis, D., & Arampatzis, A. (2018). A comparative evaluation of pre-processing techniques and their interactions for twitter sentiment analysis. *Expert Systems with Applications*.

Travenolo, B., & Costa, L. F. (2008). Accessibility in complex networks. *Physics Letters A, 373*(1), 89–95.

Wachs-Lopes, G. A., & Rodrigues, P. S. (2016). Analyzing natural human language from the point of view of dynamic of a complex network. *Expert Systems with Applications, 45*, 8–22.

Wang, Y., & Xu, W. (2018). Leveraging deep learning with lda-based text analytics to detect automobile insurance fraud. *Decision Support Systems, 105*, 87–95.

Witten, I. H., Frank, E., Hall, M. A., & Pal, C. J. (2016). *The WEKA workbench. online appendix for “data mining: Practical machine learning tools and Techniques”*. Morgan Kaufmann.

Wu, J., Zhang, G., & Ren, Y. (2017). A balanced modularity maximization link prediction model in social networks. *Information Processing & Management, 53*(1), 295–307.

Xiong, R., Wang, J., Zhang, N., & Ma, Y. (2018). Deep hybrid collaborative filtering for web service recommendation. *Expert Systems with Applications*.

Yu, D., Wang, W., Zhang, S., Zhang, W., & Liu, R. (2017). Hybrid self-optimized clustering model based on citation links and textual features to detect research topics. *PLoS One, 12*(10), e0187164.

Zhao, W., Mao, J., & Lu, K. (2018). Ranking themes on co-word networks: Exploring the relationships among different metrics. *Information Processing & Management, 54*(2), 203–218.

Znadbergen, R. (2018). Text analysis – transcription of the text. http://www.voynich.nu/transcr.html.