Physica A 557 (2020) 124895

Contents lists available at ScienceDirect

Physica A

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

Semantic flow in language networks discriminates texts by genre and publication date

Edilson A. Corrêa Jr., Vanessa Q. Marinho, Diego R. Amancio *

Institute of Mathematics and Computer Science, University of São Paulo, São Carlos, São Paulo, Brazil

ARTICLE INFO

Article history:
Received 7 January 2020
Received in revised form 6 April 2020
Available online 30 June 2020

Keywords:
Complex networks
Semantic networks
Co-occurrence networks
Language networks
Word embeddings
Network embeddings

ABSTRACT

We propose a framework to characterize documents based on their semantic flow. The proposed framework encompasses a network-based model that connected sentences based on their semantic similarity. Semantic fields are detected using standard community detection methods. As the story unfolds, transitions between semantic fields are represented in Markov networks, which in turn are characterized via network motifs (subgraphs). Here we show that different book characteristics (such as genre and publication date) are discriminated by the adopted semantic flow representation. Remarkably, even without a systematic optimization of parameters, philosophy and investigative books were discriminated with an accuracy rate of 92.5%. While the objective of this study is not to create a text classification method, we believe that semantic flow features could be used in traditional network-based models of texts that capture only syntactical/stylistic information to improve the characterization of texts.

1. Introduction

In the last few years, several interesting findings have been reported by studies using network science to model language [1–10]. Network-based models have been used e.g. to address the authorship recognition problem, where the structure of the networks can provide valuable language-independent features. Other relevant applications relying on network science include the word sense disambiguation task [11,12], the analysis of text veracity and complexity [13,14]; and scientometric studies [15].

Whilst most of the network-based language research have been carried out at the word level [16,17], only a limited amount of studies have been performed based on mesoscopic structures (sentences or paragraphs) [18]. In addition, most of the studies have analyzed language networks in a static way [19,20]. In other words, once they are obtained, the order in which nodes (words, sentences, paragraphs) appear is disregarded. Here we probe the efficiency of sentence-based language networks in particular classification problems. Most importantly, differently from previous works hinging on network structure characterization [16,17], we investigate whether the semantic flow along the narrative is an important feature for textual characterization in the considered classification tasks.

During the construction of a textual narrative, oftentimes authors follow a structured flow of ideas (introduction, narrative unfolding and conclusion). Even in books displaying a non-linear, complex narrative unfolding, one expects that an underlying linear semantic flow exists in authors' mind. In other words, even though narrative events might not organize themselves in a trivial linear form, the linearity imposed by written texts requires some type of linearization (e.g. by performing a walk through the network). This idea is illustrated in Fig. 1.

* Corresponding author.
E-mail address: diego@icmc.usp.br (D.R. Amancio).

https://doi.org/10.1016/j.physa.2020.124895
0378-4371/

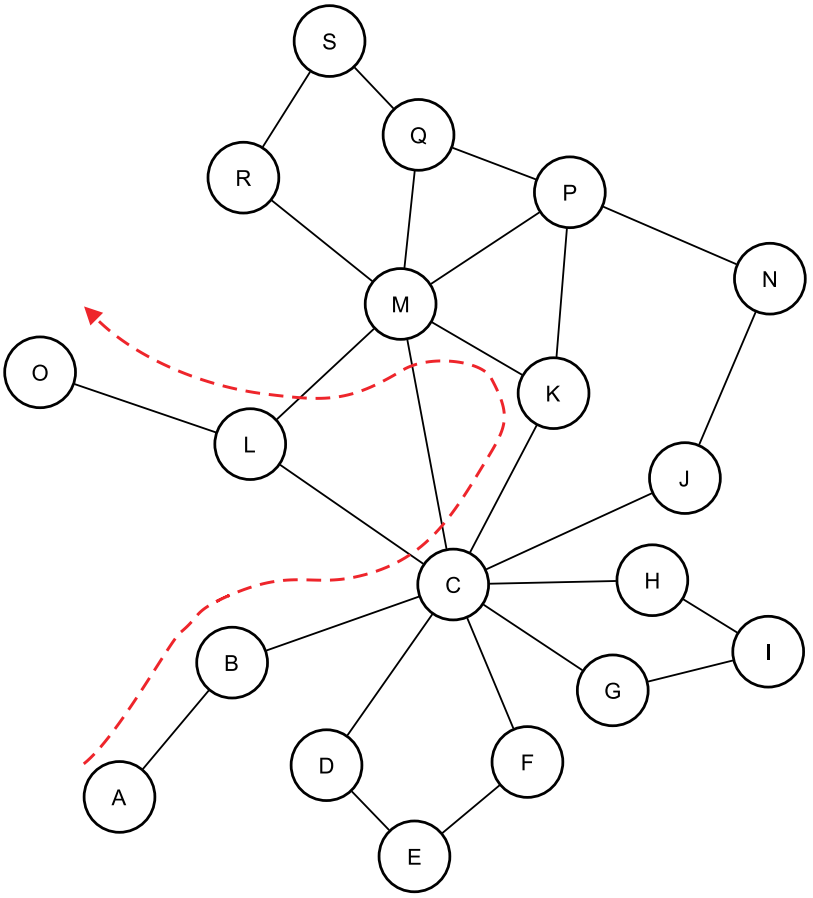

Fig. 1. High-level scheme illustrating the process of creating a text from a network of ideas. Each node represents an idea in the text. Usually, in semantic networks, nodes can represent words, sentences, paragraphs or even a sequence of paragraphs [12,18]. Edges represent the relationship (similarity) between ideas. In this paper we model each sentence as a distinct node in the network. A written text can be seen as a walk on this network (see e.g. [21]). In the example, the following sequence of ideas is produced: "A, B, C, K, M, L, O".

The ideas conveyed by a text can be represented as a complex network, where nodes represent semantic blocks (e.g. sentences, paragraphs), and edges are established according to semantic similarities. To map such a conceptual network into a text, authors perform a linearization process, where nodes (concepts, ideas) are linearly chosen and then transformed into a linear narrative (see Fig. 1). Such a projection of a multidimensional space of ideas into a linear representation has been object of studies both on network theory and language research. A consequence of such a linearization in texts is the presence of long-range correlations at several linguistic levels, a property that has been extensively explored along the last years [22–25].

While complex semantic networks have been used in previous works to represent the relationship between ideas and concepts, only a minor interest has been devoted to the analysis of how authors navigate the high-dimensional semantic relationships to generate a linear stream of words, sentences or paragraphs. In [26], a mesoscopic representation of networks was proposed. The authors used as a semantic, meaningful block a set of consecutive paragraphs. The semantic blocks were connected according to a lexical similarity index. The model aimed at combining a networked representation with an idea of semantic sequence obtained when reading a document. Even though some interesting patterns were found, the concept of semantic fields were not clear, as no semantic community structure arises from mesoscopic networks. The problem of linearization of a network structure was studied in [21]. A systematic analysis of the efficiency of several random walks in different topologies was probed. The efficiency was probed in a twofold manner: (i) the efficiency in transmitting the projected network; and (ii) the efficiency in recovering the original network. In [27], the authors explored the efficiency of navigating an idea space, by varying network topologies and exploration strategies.

In the current paper, we take the view that authors write documents by applying a linearization process to the original network of ideas, as shown in the procedure illustrated in Fig. 1. Upon analyzing the flow of ideas with the adopted network-based framework, we show that features extracted from the networks can be employed to characterize and classify texts. More specifically, we defined the network of ideas as a network of sentences linked by semantic similarity. *Semantic fields* of similar sentences (nodes) were identified via network community detection. These fields (network

communities) were then used to characterize the dynamics of authors' choices in moving from field to field as the story unfolds. Using a stochastic Markov model to represent the dynamics of choices of semantic fields performed by the author along the text, we showed, as a proof of principle, that the adopted representation can retrieve textual features including style (publication epoch) and complexity.

2. Research questions

The main objective is to answer the following research questions: is there any patterns of semantic flow in stories? Are these patterns related to textual characteristics? To address these questions, we use sentence networks to represent the semantic flow of ideas in texts. Such networks are summarized using a high-level representation based on the relationship between communities extracted from the sentence networks. Using this representation, we show that motifs extracted from such a high-level representation can be used to classify texts according to the style in which authors unfolds their stories. We are not proposing a novel text classification method, but investigating whether semantic flow is a feature that depends on text genre and publication date. We argue that the obtained results suggest that the proposed high-level view of a text network could be further probed in other Natural Language Processing classification tasks.

3. Materials and methods

This study can be divided in two parts. In the first step, we identify the semantic clusters (fields) of the story. Differently from the analysis of short texts, where semantic groups can be identified mostly by identifying paragraphs, in long texts – the focus of this study – the identification of semantic clusters is more challenging because semantic topics might not be organized in consecutive sentences/paragraphs owing to the linearization process illustrated in Fig. 1. In other words, the process of obtaining semantic clusters can be understood as the reverse operation depicted in Fig. 1.

In order to identify semantic clusters from the text, we first create a network of sentences for each document, where sentences are linked if the similarity between them is above a given threshold. The obtained network is then analyzed via community detection methods, where groups of densely connected sentences are identified and considered as semantic clusters. A qualitative analysis of the obtained communities suggested that most of the largest communities are in fact related to a specific subtopic approached in the text. This idea relating semantic fields and network communities has also been used to construct automatic summarization systems [28].

In the second step of this study, we investigate the semantic flow of ideas developed by authors while unfolding their stories. We consider each community found as a semantic cluster, and as the story unfolds (one sentence after another), we analyze the community labels of the adjacent sentences to create a Markov chain, where each state represents a community and transitions are given by the text dynamics. Once the Markov chain representing the transitions of semantic clusters is obtained, the text is characterized by finding and counting different chain motifs. Such a characterization is then used to classify texts according to the semantic flow as revealed by sentences membership to different network communities.

The main objective of this work is to provide a framework to analyze and verify whether the semantic flow in texts can be used to characterize documents. Because the framework encompasses some steps, several alternatives could be probed in each step. We decided not to conduct a systematic analysis of combination of methods (and parameters) owing to the complexity of such analysis. A systematic study of the parameters and methods optimizing the proposed framework is intended to be conducted as a future work.

In Fig. 2 we show a representation of the framework proposed to analyze stories. In the next section, we detail each of steps used in this framework.

3.1. Word and sentence embeddings

Usually any vector representation of words is known as a *word embedding*. However, since the creation of *neural word embeddings* [29], the term is mostly used to name those approaches based on neural network representations. The *word embedding* model proposed in [29] aimed at classifying texts based on raw text input. Thus, the classification does not require that textual features as input. Typically, *word embeddings* are dense vectors that are learned for a specific vocabulary, with the objective of addressing some task.

A typical task addressed with word embeddings is the language modeling problem, which aims at learning a probability function describing the sequence of words in a language. More recently, this same vector representation has been used in more complex models, with the objective of addressing several Natural Language Processing tasks simultaneously, including POS tagging, name entity recognition, semantic role labeling and others [30,31]. Despite its relative success in the above mentioned tasks, the adopted embeddings could not be used in general purpose applications [30,31]. In order to allow the use of embeddings in wider contexts, the Word2Vec representation was proposed [32,33].

The Word2Vec is a neural model proposed to learn a dense, high-quality representation that is able to capture both syntactical and semantical language properties. As a consequence, vectors representing words conveying the same meaning are close in the considered space. An interesting property of the Word2Vec technique is the *compositionality*, which allows that large information blocks (e.g. sentences) can be represented by combining the representation of the

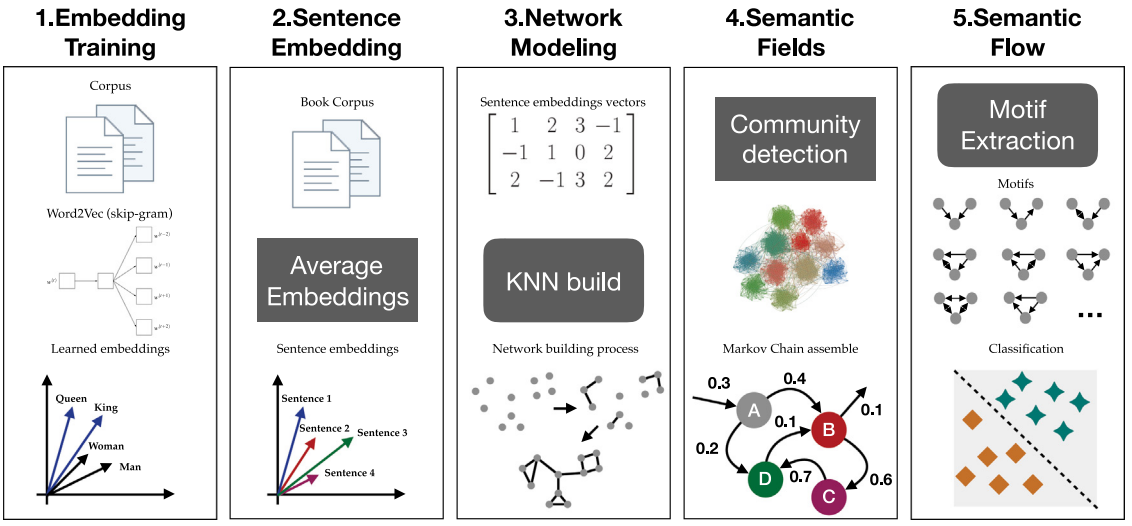

Fig. 2. Sequence of steps employed to characterize documents using the proposed framework: (1) word embedding generation; (2) sentence embeddings generation from word embeddings; (3) a sentence similarity network is generated based on the similarity of sentence embeddings; (4) network communities are detected and a Markov chain is built based on the story unfolding (semantic flow); and (5) motifs are identified in the Markov chain representing the semantic flow. These motifs are then used as features in a classification method.

vector representing the words in the sentence. Other interesting property is the ability to combine embeddings in a intuitive fashion [32,34]. For example, using the Word2Vec technique, the following relationship can be obtained:

$$\text{vector}(\text{“King”}) - \text{vector}(\text{“Man”}) + \text{vector}(\text{“Woman”}) \simeq \text{vector}(\text{“Queen”}). \tag{1}$$

The Word2Vec model is a robust, general-purpose neural representation that has been widely used in several Natural Language Processing tasks, including machine translation [35], summarization [28,36], sentiment analysis [37] and others. Given the success of the this model and the possibility of composition in different scenarios (sentiment analysis and sense disambiguation) [37,38], in the current study we used a representation of sentences based on the Word2Vec. We have chosen the Word2Vec as embedding method to illustrate the proposed framework for the aforementioned advantages. However, we note that other embeddings techniques exist [39,40]. A comparison of techniques, however, revealed no significant difference in performance.

More specifically, here the embedding $\mathbf{s}$ of a sentence s is represented by the average embedding of the words in s:

$$\mathbf{s} = \frac{1}{\omega(s)} \sum_{i=0}^{\omega} \mathbf{w}_i. \tag{2}$$

where $\mathbf{w}_i$ is the embedding of the ith word in s and $\omega(s)$ is the total number of words in s.

The word embedding technique used here was obtained with the Word2Vec method (skip-gram). The training phase used the Google News corpus [32,33]. According to [32,33], the parameters of the method are optimized in the context of semantical similarity task. The combination of embeddings to represent a sentence in Eq. (2) could also be performed by summing individual embeddings. However, it has been shown that there is no significant difference when sentence embeddings are used to construct a network of sentence similarity [41]. We note that some words are removed from this analysis. This includes *stopwords* (e.g. articles and prepositions) and words with no embeddings in the Google News corpus. Thus, whenever a sentence contains only words with no available embeddings, it is removed from the analysis.

3.2. Modeling sentence embeddings into complex networks

This step corresponds to the reverse process illustrated in Fig. 1. In other words, a network representing the relationship between ideas is created from the text. The construction of networks from vector structures has been explored in recent works. In [42], the authors present such a transformation as a framework in complex systems analysis. The transformation of vector structures into networks has also been used in the context of text analysis [41,43]. The creation of a complex network from Word2Vec was proposed using a twofold approach. The d-proximity technique links all nodes whose distance from the reference node is lower than d. The second technique is the k-NN approach, which links all k nearest nodes to the reference node. In the same line, [41] created a network based on word embeddings. However, the authors aimed at creating a network that takes into account the sense of words to solve ambiguities. Each occurrence of an ambiguous words was modeled as a node in the network. Nodes were represented by a vector combining the embeddings

of the words in the context. Two occurrences of an ambiguous were then connected whenever the respective embeddings were similar. In other words, two ambiguous words were linked if they appeared in similar contexts.

In the current study, sentences were connected according to the k-NN technique, as suggested by other works [43]. Each sentence is represented as a vector according to Eq. (2). The value of k in the main experiments were chosen to allow that each network is composed of a single connected component. In particular, the lowest k allowing the creation of a connected network was used for each book.

3.3. Community detection

The next step in the proposed framework concerns the detection of semantic fields, i.e. the communities in the network of sentences. A recurrent phenomena in several complex networks is the existence of communities, i.e. groups of strongly connected nodes. Similarly to other network measurements, the detection of communities gives important information regarding the organization of networks. Communities are present in different networks including in biological, social and information networks [44].

A well-known measure to quantify the quality of partitions in complex networks is the modularity [45,46]. This measure compares the obtained partition with a null model, i.e. a network with similar properties but with no community structure. Several algorithms have been proposed to address the community detection problem via optimization of the modularity. In the main experiments we used the Louvain method [47] to identify communities. The main advantage of this method is its computational efficiency, which has allowed its use in several contexts [43,48]. Another advantage associated to this algorithm is that no additional parameters are required to optimize the modularity. In additional experiments, we also probed the effect of other community detection methods on the performance of the framework in the considered classification tasks. In addition to the Louvain method, we also used the following three methods: (i) fastgreedy, (ii) eigenvector, and (iii) walktrap. An introduction to these methods can be found in [44,49].

In the proposed network representation, communities represent groups of interconnected sentences about a given topic. Because the k-NN construction allows nodes to be connected to other close nodes and, considering the Word2Vec an efficient semantic representation, the linking strategy allows the creation of dense clusters of semantically related sentences. This idea of semantic clusters has also been explored via community detection in similar works [41,43,50]. For example, using networks built at the word level, the groups detected in [43] were found to represent large cities, professions and others topics. In [41], the obtained groups were found to represent words conveying the same sense.

In order to illustrate the process of obtaining semantic communities, we performed an analysis of the obtained communities in the book "Alice's Adventures in Wonderland", by Lewis Carroll (see Fig. 3). We summarize below the main topics approaches in some of the communities obtained by the Louvain algorithm:

1. *Community A*: this community includes sentences mentioning animals (e.g. "pet", "cat", "mouse" and "dog"). This community also includes dialogs between Alice and animals. "Cat" is the main character in this community.
2. *Community B*: this community includes words sentiment words expressed via speeches. Some of the words in this community are "passionate", "melancholy", "angrily", "shouted" and "screamed".
3. *Community C*: this community includes several adverbs related to Alice's actions.
4. *Community D*: this community includes words related to sentiments such as anger, tranquility and peacefulness.
5. *Community E*: this community is most related to the word "soup".
6. *Community F*: this community is related to geographical locations, including countries and cities (Australia, Rome and New Zealand). Interestingly, this community also included the word "Cricket", a prominent sport in Australia.
7. *Community G*: this community included mostly sentences referring to "Dormouse", one of the main characters in the plot.

While most of the obtained communities are informative, a few communities were found to be more dispersed, approaching more than one topic. This might occur given the limitations of the embeddings model, since some words might not be available in the considered model. Despite these limitations, we show that the flow of information (from sentence to sentence) in the obtained semantic communities can be used to characterize texts.

3.4. Markov chains

In order to capture how authors move from community to community (semantic field) as their story unfolds, we create a representation of community transitions. The idea of studying language via Markov process is not recent. One of the first uses of this model is the study of letters sequences [51]. Since then, Markov chains are used as a statistical tool in several natural language processing problems, including language modeling, machine translation and speech recognition [52].

Here we represented the transitions between semantic fields (network communities) as a first order Markov chain. In this representation, each community becomes a state. Note that this approach of representing communities as a single unit has also been used in other contexts [15]. The probabilities of transition are considered according to the frequency of transitions observed in adjacent sentences. As we shall show, using this model, it is possible to detect patterns of how authors change topics in their stories. As a proof of principle, these patterns are used to characterize texts in distinct classification tasks.

Fig. 3. Example of sentence network obtained from the book "Alice's Adventures in Wonderland", by Lewis Carroll. Colors represent community labels obtained with the Louvain method. The visualization was obtained with the method described in [15].

The process of creating a Markov chain from a network divided into communities is shown in Fig. 4. In the previous phase, communities are identified to represent distinct semantic field of the story (see left graph in Fig. 4). Because each sentence belongs to just one community, the text can be regarded as discrete time series, where each element corresponds to the membership (community label) of each sentence. Using this sequence of community labels, it is possible to create a Markov chain representing all transitions between communities (see graph on the top left of Fig. 4). Transitions weights are proportional to the frequency in which they occur and normalized so as to represent a probability. This representation is akin to a Markov chain used in other works addressing the language modeling problem [53]. The main difference here is that we are not interested in the use of particular words, but in semantic fields [54]. Once the Markov chain is obtained, we characterize this structure using *network motifs*. Note that the obtained Markov chain is equivalent to a weighted and directed complex network. Thus, traditional network tools can be used to identify network motifs [55,56].

3.5. Motifs

Network motifs are used to analyze a wide range of complex systems, including in biological, social and information networks [57]. Motifs can be defined as small subgraphs (see Fig. 5) occurring in real systems in a significant way. To quantify the significance, in general, one assumes an equivalent random network as null model. In text analysis, motifs have been used to analyze word adjacency networks in applications focusing on the syntax and style of texts [58]. More recently, an approach based on labeled motifs showed that authors tend to use words in combination with particular motifs [59]. Examples of considered motifs are represented in Fig. 5. Mathematically, the frequency of the motif with nodes "i", "j" and "k" in Fig. 5 can be computed as:

$$f_m = \sum_i \sum_j \sum_k a_{ki} a_{kj} a_{ji}, \tag{3}$$

where a_{ij} is an element of the adjacency matrix (i.e. $a_{ij} = 1$ if there is an edge from i to j and $a_{ij} = 0$, otherwise). A similar equation can be used to compute the frequency of all possible motifs comprising three nodes. In very large networks, efficient strategies for motif discovery have been proposed [60,61].

While the structure of the Markov Chains could be analyzed using traditional network measurements, we decided not to use these measurements owing to the limited size of these structures. As suggested in related works, a characterization

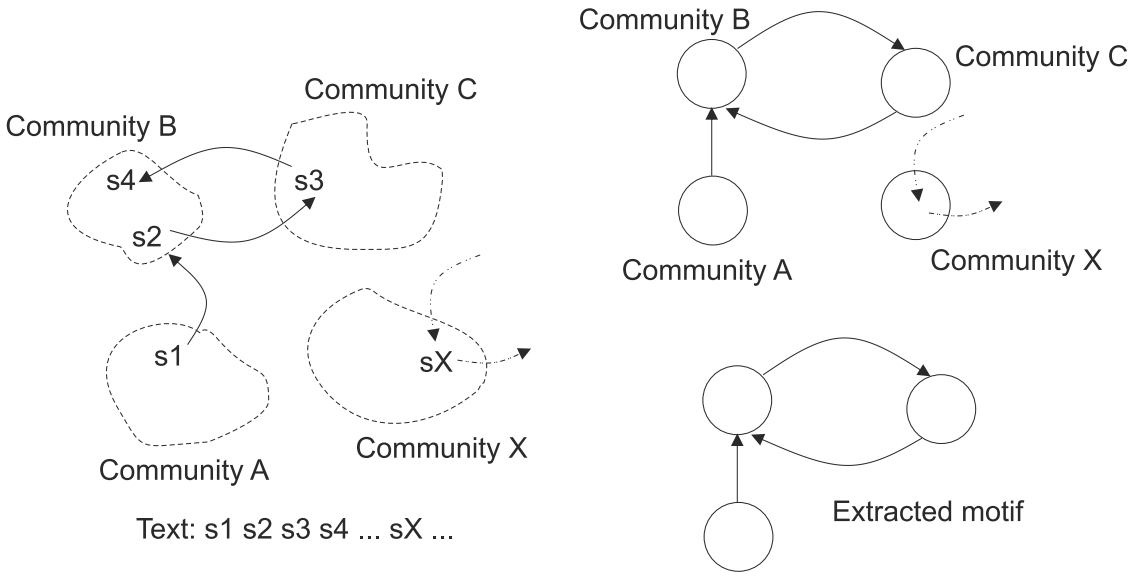

Fig. 4. Example of extraction of motifs from the network. As the text unfolds according to a given order of sentences $(s1, s2, s3, s4 \ldots sX \ldots)$ a sequence of communities is generated (Community A, Community B, Community C, Community B). This sequence is used to create a Markov chain. Finally, the Markov chain is characterized by counting different patterns (motifs) of community transitions.

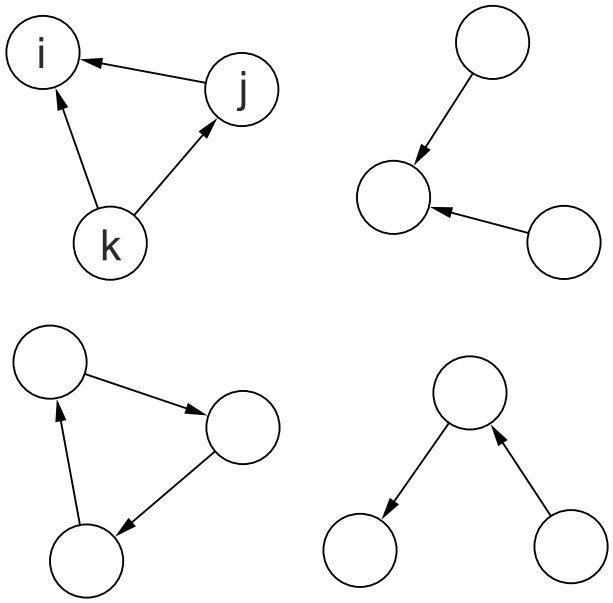

Fig. 5. Example of representative motifs comprising three nodes. The frequency of occurrence of the motif in the left upper corner can be computed using Eq. (3).

based on network metrics in small networks might not be informative [20,62,63]. As we shall show in the results, this is a simple, yet useful approach to classify small Markov Chains. In the results section, we only show the results obtained for three-node connected motifs. In preliminary experiments, no significant gain was observed when considering network motifs comprising four nodes.

The following approaches were considered to extract motifs from Markov networks. We discriminated strategies according to the use of weights to count motifs (unweighted vs weighted). If a thresholding is applied before extracting motifs, the strategy is referred to as "simplified" (see strategies 2–4 below).

1. *Unweighted strategy*: no thresholding is applied. All weights are disregarded. Every time a motif is detected, its frequency is increased by one.

Table 1
Accuracy rate and p-value obtained for the classification subtasks. Only the best results are shown among all considered classifiers. We considered the *unweighted* version of the Markov networks to extract motifs.

Subtask	Acc.	p-value
Children × investigative	50.8%	5.56×10^{-1}
Children × philosophy	71.6%	1.30×10^{-3}
Investigative × philosophy	70.8%	3.30×10^{-3}
Children × investigative × philosophy	43.8%	3.50×10^{-2}

2. *Simplified unweighted strategy*: this approach is the same as the *unweighted* strategy. However, before counting motifs, the weakest edges are removed according to a given threshold.
3. *Simplified weighted strategy*: before counting motifs, the weakest edges are removed according to a given threshold. Then motifs are identified disregarding edges weights (e.g. using Eq. (3)). Edge weights are then considered to update the frequency associated to that motif. Every time a given motif is found, the respective "frequency" of that motif is increased by the sum of the weights of its edges.
4. *Simplified weighted strategy with local thresholding*: this technique is similar to the *simplified weighted strategy*. However, here a different approach is used to threshold the networks. We consider here a local threshold, which is established for each edge. A local thresholding strategy is important for some network applications because a simple global threshold value might overlook important network structures, such as network communities [64–66].
The local thresholding strategy proposed in [66] evaluates the relevance of an edge by computing a p-value determined in terms of a null model. The null model computes the likelihood of a node v having an edge with a specific weight by taking into account the other edges connected to v. In practice, the relevance α_{ij} of an edge e_{ij} connecting nodes i and j is computed as:

$$\alpha_{ij} = 1 - (k_i - 1)\int_0^{\pi_{ij}} (1-x)^{k_i-2} dx, \tag{4}$$

$$\pi_{ij} = w_{ij}\left(\sum_{ik \in E} w_{ik}\right)^{-1}, \tag{5}$$

where w_{ij} is the weight of e_{ij} and $k_i = \sum_j a_{ij}$. All edges with $\alpha_{ij} < \mathcal{A}$ are removed from the network, where $\mathcal{A}$ is the adopted local threshold.

3.6. Classification

The extracted motifs from the Markov Chains are used as input (features) to the classification systems. The following methods were used in the experiments: Decision Tree (CART), kNN, SVM (linear) and Naive Bayes [67]. The evaluation was performed using a 10-fold cross-validation approach. As suggested in related works, all classifiers were trained with their default configuration of parameters [68].

4. Results and discussion

Here we probed whether the dynamics of changes in semantic groups in books can be used to characterize stories. The proposed methodology was applied in two distinct classification tasks. In the first task, we aimed at distinguishing three different thematic classes: (i) children books; (ii) investigative; and (iii) philosophy books. The second aimed at discriminating books according to their publication dates. All books (and their respective classes) were obtained from the Gutenberg repository. The list of books and respective authors are listed in the Supplementary Information. The corpora size is compatible with other works in the literature [18,63,69,70].

In the first experiment, we evaluated if patterns of semantic changes are able to distinguish between children, philosophy or investigative books. We considered problems with two or three classes. The obtained results are shown in Table 1. In this case, weights were disregarded after the construction of the Markov networks (*unweighted* version). Considering subtasks encompassing only two classes, only the distinction between children and investigative texts were not significant, with a low accuracy rate. The distinction philosophy books and the other two classes, however, yielded a much better discrimination. These results were found to be significant. When all three classes are discriminated, a low accuracy rate was found (43.8%), even though this still represents a significant result. The low accuracy rate found using the proposed approach is a consequence of a regular behavior found in the Markov chains. In other words, in most of the books, all communities were found to be connected to each other, hampering thus the discriminability of different types of books.

Table 2
Accuracy rate and p-value obtained for the classification subtasks. Only the best results are shown among all considered classifiers and thresholds. We considered the *simplified unweighted* version of the Markov networks to extract motifs.

Subtask	Acc.	Threshold	p-value
Children × investigative	65.8%	0.060	1.64×10^{-2}
Children × philosophy	81.0%	0.190	1.19×10^{-5}
Investigative × philosophy	91.6%	0.075	2.23×10^{-10}
Children × investigative × philosophy	62.2%	0.075	2.00×10^{-7}

Table 3
Accuracy rate and p-value obtained for the classification subtasks. Only the best results are shown among all considered classifiers and thresholds. We considered the *simplified weighted* version of the Markov networks to extract motifs.

Subtask	Acc.	Threshold	p-value
Children × investigative	70.8%	0.075	3.30×10^{-3}
Children × philosophy	89.0%	0.145	1.62×10^{-8}
Investigative × philosophy	92.5%	0.120	2.23×10^{-10}
Children × investigative × philosophy	62.7%	0.075	2.00×10^{-7}

Table 4
Accuracy rate obtained for the classification subtasks, considering distinct community detection methods: Louvain, walktrap, eigenvector and fastgreedy [44]. Only the best results are shown among all considered classifiers, thresholds and community detection methods. We considered the *simplified weighted* version of the Markov networks to extract motifs. The following parameters were used in the classifiers: SVM (linear kernel and penalty parameter of the error term = 1.0), CART (criterion to measure the quality of a split = gini, minimum number of samples required to split an internal node = 2, minimum number of samples required to be at a leaf node = 1), Naive Bayes (GaussianNB) and kNN (k = 1, Euclidean distance).

Subtask	Acc.	Threshold	Method	Classifier
children × investigative	76.7%	0.045	Eigenvector	SVM
children × philosophy	90.8%	0.07	Walktrap	SVM
investigative × philosophy	97.5%	0.185	FastGreedy	CART
children × investigative × philosophy	70.0%	0.17	FastGreedy	kNN

Given the low accuracy rates obtained with the *unweighted* strategy, we analyzed if the *simplified* unweighted version was able to provide a better characterization. In this case, the weakest edges were removed before the extraction of motifs. We considered the thresholding ranging between 0.01 and 0.20. The main idea here is to remove less important links between communities. The obtained results are shown in Table 2. All obtained results turned out to be significant. All previous accuracy rates were improved. Interestingly, a high discrimination rate (91.6%) was obtained when discriminating investigative and philosophy books. These results suggest that the threshold is an important pre-processing step here, given that it can boost the performance of the classification by a large margin.

When combining thresholding and edges weights in the *simplified weighted* version, the results obtained in Table 3 were further improved. The highest gain in performance was observed when discriminating children from philosophy books: the performance improved from 81.0% to 89.0%. Only a minor improvement was observed when all three classes were discriminated. Overall, this results suggest that both thresholding and the use of edges weights might be useful to characterize Markov networks. Most importantly, all three methods showed that, in fact, there is a correlation between the thematic approached and the way in which authors approaches semantic groups in texts.

While the main focus of this manuscript is not to provide the best combination to optimize the performance of a classification task, it is still interesting to probe how the classification based on the concept of semantic flow can benefit from different partitions (semantic clusters) extracted from different community detection methods. The best results obtained by comparing 4 distinct methods are summarized in Table 4. Interestingly, note that an impressive 97.5% accuracy rate was observed when discriminating investigative and philosophy books. In this case, a feature relevance analysis revealed that two particular motifs are responsible for most of the discriminative power (see analysis in Fig. 6). Additional results considering the strategy based on *unweighted* motifs strategy are shown in Table S1 of the Supplementary Information.

The discriminative power of the obtained networks was also investigated using a local strategy to threshold the network. In other words, the relevance of an edge in the *simplified weighted strategy with local thresholding* depends on the weights of its neighboring edges (see Section 3.5) [66]. We show in Table 5 the results obtained with this technique when adopting as local threshold the value $\mathcal{A} = 0.95$. We found that, for this particular technique, the results are not improved, even when other values for $\mathcal{A}$ are considered (additional results are shown in the Supplementary Information). For this reason, we did not consider the *simplified weighted strategy with local thresholding* in the next results.

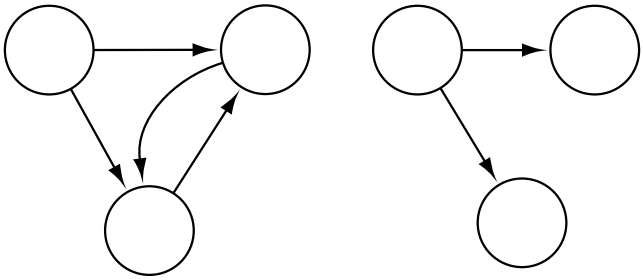

Fig. 6. The above motifs are responsible for most of the discriminative power considering the task "investigative x philosophy" in the configuration of methods and parameters used in Table 4. In order to compute relevance, we used the Gini index [71] (Γ) in the CART classification method. The obtained relevance values for the motifs on the left and right sides of the figure are, respectively, $\Gamma = 0.809$ and $\Gamma = 0.147$. While the relevance of features (i.e. motifs) may differ in other classification tasks, the above motifs were found to be among the most discriminative ones in the classification tasks studied in this paper (result not shown).

Table 5
Accuracy rate and p-value obtained for the classification subtasks. We considered the *simplified weighted strategy with local thresholding* to extract motifs. All edges with $\alpha_{ij} < \mathcal{A}$ were removed (see Eq. (4)). The obtained performance is no better than the one obtained with the global thresholding approach..

Subtask	Acc.	p-value
Children × investigative	63.3%	3.25×10^{-2}
Children × philosophy	66.3%	1.60×10^{-2}
Investigative × philosophy	71.3%	3.30×10^{-3}
Children × investigative × philosophy	48.0%	5.94×10^{-3}

Table 6
Performance of the proposed method using the *simplified unweighted* motif characterization of Markov networks. For each subtask, only the best threshold obtained for the best classifier is shown.

Subtask	Acc.	Threshold	p-value
1700–1799 × 1800–1899	70.0%	0.195	1.34×10^{-3}
1700–1799 × 1900 or later	75.0%	0.060	6.73×10^{-5}
1800–1899 × 1900 or later	70.0%	0.160	1.34×10^{-3}
1700–1850 × 1851 or later	66.0%	0.010	6.74×10^{-3}
1700–1799 × 1800–1899 × 1900 or later	55.0%	0.025	1.22×10^{-5}

In order to investigate the dependency of the classification results on the word embeddings model, we also considered embeddings obtained from the BERT model [39]. While there is only a minor improvement in particular cases, the results obtained with the Word2vec model (see Table 4) provides most of the best results. The results obtained with the BERT model are summarized in Tables S2 and S3 of the Supplementary Information.

We also investigated if the patterns of semantic flow varies with the publication date. For this reason, we selected a dataset with books in different periods. The following classes were considered, according to the range of publication dates:

1. Books published between 1700 and 1799.
2. Books published between 1800 and 1899.
3. Books published after 1900.
4. Books published between 1700 and 1850.
5. Books published after 1851.

The results obtained in the classification for different subtasks is shown in Table 6. We only show here the results obtained for the simplified unweighted characterization because it yielded the best results. Overall, all classification results are significant, confirming thus that there are statistically significant differences of semantic flow patterns for books published in different epochs. However, the results obtained here are worse than the ones obtained in the dataset with books about different themes (see Table 3). Therefore, patterns of semantic flow seems to be less affected by the year of publication, while being more sensitive to the subject/topic approached by the text.

5. Conclusion

In this paper we investigate whether patterns of semantic flow arises for different classes of texts. To represent the relationship between ideas in texts, we used a sentence network representation, where sentences (nodes) are

connected based on their semantic similarity. Semantic clusters were identified via community detection and high-level representation of each book was created based on the transition between communities as the story unfolds. Finally, motifs were extracted to characterize the patterns of transition between semantic groups (communities). When applied in two distinct tasks, interesting results were found. In the task aiming at classifying books according to the approached themes, we found an high accuracy rate (92.5%) when discriminating investigative and philosophy books. A significant performance in the classification was also obtained when discriminating books published in distinct epochs. However, the discriminability for this task was not as high as the ones obtained when discriminating investigative, philosophy and children books.

Given the complexity of the components in the proposed framework, we decided not to optimize each step of the process. Even without a rigorous optimization process, we were able to identify semantic flow patterns that were able to discriminate distinct classes of texts. As future works, we intend to perform a systematic analysis on how to optimize the process. For example, during the construction of the networks, different approaches could be used to create embeddings and link similar sentences [72]. In a similar fashion, different strategies to identify communities could also be used in the analysis. Finally, we could also investigate additional approaches to characterize the obtained Markov networks.

The proposed framework identified clusters of ideas being conveyed in texts. We basically measured, for each story, how authors move from one semantic cluster to another while the story is being told. This gives the sense of "semantic flow" measured in terms of network motifs. Our results suggest, therefore, that different classes of stories have distinct semantic flow patterns. For example, in the classification of children and philosophical books, one should expect that the dynamics of changing topics in children books should be much less complex than the semantic flow observed in books about philosophy. Such a difference could be related to the cognitive efforts required to the reader to understand different patterns of semantic flow. Concerning the classification based on publication dates, the high discriminability could be related to the fact that each century is characterized by a different style. These are hypothesis that should be evaluated in future works by using potential available datasets.

Our results suggest that semantic flow motifs could play an important role in other NLP tasks. For example, in the authorship recognition task, patterns extracted from a semantic flow analysis could be combined with other techniques to improve the characterization of authors [73,74]. In fact, the use of motifs in the microscopic level has already provided a good characterization of authors [59]. A similar idea could also be applied to the analysis of other stylometric tasks. In addition, we suggest that the semantics of the texts could be combined with the concept of semantic flow by using "labeled motifs", as proposed in our previous work [59]. Since semantic networks have been studied in cognitive sciences, we believe that the adopted network representation could be adapted and used – as an auxiliary tool – to study complex brain and cognitive processes that could assist the diagnosis of cognitive disorders via text analysis [73,75].

CRediT authorship contribution statement

Edilson A. Corrêa Jr.: Methodology, Software, Validation, Formal analysis, Investigation, Writing - original draft, Visualization. **Vanessa Q. Marinho:** Software, Validation, Investigation. **Diego R. Amancio:** Conceptualization, Methodology, Validation, Formal analysis, Writing - original draft, Writing - review & editing, Supervision, Project administration, Funding acquisition.

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.

Acknowledgments

E.A.C. Jr. and D.R.A. acknowledge financial support from Google (Google Research Awards in Latin America grant). V.Q.M. acknowledges financial support from São Paulo Research Foundation (FAPESP) (Grant no. 15/05676-8). D.R.A. also thanks FAPESP (Grant no. 16/19069-9) and CNPq-Brazil (Grant no. 304026/2018-2) for support.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.physa.2020.124895.

References

[1] W. Jin, R.K. Srihari, Graph-based text representation and knowledge discovery, in: Proceedings of the 2007 ACM Symposium on Applied Computing, ACM, 2007, pp. 807–811.
[2] R.F. Cancho, R.V. Solé, R. Köhler, Patterns in syntactic dependency networks, Phys. Rev. E 69 (5) (2004) 051915.
[3] M.A. Montemurro, D.H. Zanette, Keywords and co-occurrence patterns in the voynich manuscript: An information-theoretic analysis, PLoS One 8 (6) (2013) e66344.
[4] K. Ban, M. Perc, Z. Levnajić, Robust clustering of languages across wikipedia growth, R. Soc. Open Sci. 4 (10) (2017) 171217.

[5] B. Tadić, M. Andjelković, B.M. Boshkoska, Z. Levnajić, Algebraic topology of multi-brain connectivity networks reveals dissimilarity in functional patterns during spoken communications, PLoS One 11 (11) (2016) e0166787.
[6] D.R. Amancio, F.N. Silva, L.F. Costa, Concentric network symmetry grasps authors' styles in word adjacency networks, Europhys. Lett. 110 (6) (2015) 68001.
[7] M. Stella, A. Zaytseva, Forma mentis networks map how nursing and engineering students enhance their mindsets about innovation and health during professional growth, PeerJ Comput. Sci. 6 (2020) e255.
[8] N. Castro, M. Stella, The multiplex structure of the mental lexicon influences picture naming in people with aphasia, J. Complex Netw. 7 (6) (2019) 913–931.
[9] M. Stella, Modelling early word acquisition through multiplex lexical networks and machine learning, Big Data Cogn. Comput. 3 (1) (2019) 10.
[10] M. Stella, S. De Nigris, A. Aloric, C.S. Siew, Forma mentis networks quantify crucial differences in STEM perception between students and experts, PLoS One 14 (10) (2019).
[11] E. Agirre, A. Soroa, Personalizing pagerank for word sense disambiguation, in: Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics, Association for Computational Linguistics, 2009, pp. 33–41.
[12] D.R. Amancio, O.N. Oliveira Jr., L.F. Costa, Unveiling the relationship between complex networks metrics and word senses, Europhys. Lett. 98 (1) (2012) 18002.
[13] D.R. Amancio, O.N. Oliveira Jr, L.F. Costa, Identification of literary movements using complex networks to represent texts, New J. Phys. 14 (4) (2012) 043029.
[14] D.R. Amancio, S.M. Aluisio, O.N. Oliveira Jr, L.F. Costa, Complex networks analysis of language complexity, Europhys. Lett. 100 (5) (2012) 58002.
[15] F.N. Silva, D.R. Amancio, M. Bardosova, L.d.F. Costa, O.N. Oliveira, Using network science and text analytics to produce surveys in a scientific topic, J. Informetr. 10 (2) (2016) 487–502.
[16] R.F. Cancho, R.V. Solé, The small world of human language, Proc. R. Soc. Lond. Ser. B: Biol. Sci. 268 (1482) (2001) 2261–2265.
[17] H. Liu, J. Cong, Language clustering with word co-occurrence networks based on parallel texts, Chin. Sci. Bull. 58 (10) (2013) 1139–1144.
[18] H.F. Arruda, V.Q. Marinho, L.F. Costa, D.R. Amancio, Paragraph-based representation of texts: a complex networks approach, Inf. Process. Manage. 56 (2019) 479–494.
[19] H.F. Arruda, L.F. Costa, D.R. Amancio, Using complex networks for text classification: Discriminating informative and imaginative documents, Europhys. Lett. 113 (2) (2016) 28007.
[20] D.R. Amancio, Probing the topological properties of complex networks modeling short written texts, PLoS One 10 (2) (2015) e0118394.
[21] H.F. de Arruda, F.N. Silva, C.H. Comin, D.R. Amancio, L.F. Costa, Connecting network science and information theory, Physica A 515 (2019) 641–648.
[22] W. Ebeling, A. Neiman, Long-range correlations between letters and sentences in texts, Physica A 215 (3) (1995) 233–241.
[23] A. Schenkel, J. Zhang, Y.-C. Zhang, Long range correlation in human writings, Fractals 1 (01) (1993) 47–57.
[24] E. Alvarez-Lacalle, B. Dorow, J.-P. Eckmann, E. Moses, Hierarchical structures induce long-range dynamical correlations in written texts, Proc. Natl. Acad. Sci. 103 (21) (2006) 7956–7961.
[25] M. Amit, Y. Shmerler, E. Eisenberg, M. Abraham, N. Shnerb, Language and codification dependence of long-range correlations in texts, Fractals 2 (01) (1994) 7–13.
[26] H.F. Arruda, F.N. Silva, V.Q. Marinho, D.R. Amancio, L.F. Costa, Representation of texts as complex networks: a mesoscopic approach, J. Complex Netw. 6 (1) (2018) 125–144.
[27] H.F. Arruda, F.N. Silva, L.F. Costa, D.R. Amancio, Knowledge acquisition: A complex networks approach, Inform. Sci. 421 (2017) 154–166.
[28] J.V. Tohalino, D.R. Amancio, Extractive multi-document summarization using multilayer networks, Physica A 503 (2018) 526–539.
[29] Y. Bengio, R. Ducharme, P. Vincent, C. Jauvin, A neural probabilistic language model, J. Mach. Learn. Res. 3 (Feb) (2003) 1137–1155.
[30] R. Collobert, J. Weston, A unified architecture for natural language processing: Deep neural networks with multitask learning, in: Proceedings of the 25th International Conference on Machine Learning, ACM, 2008, pp. 160–167.
[31] R. Collobert, J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu, P. Kuksa, Natural language processing (almost) from scratch, J. Mach. Learn. Res. 12 (Aug) (2011) 2493–2537.
[32] T. Mikolov, K. Chen, G. Corrado, J. Dean, Efficient estimation of word representations in vector space, 2013, arXiv preprint arXiv:1301.3781.
[33] T. Mikolov, I. Sutskever, K. Chen, G.S. Corrado, J. Dean, Distributed representations of words and phrases and their compositionality, in: Advances in Neural Information Processing Systems, 2013, pp. 3111–3119.
[34] T. Mikolov, W.-t. Yih, G. Zweig, Linguistic regularities in continuous space word representations, in: Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2013, pp. 746–751.
[35] I. Sutskever, O. Vinyals, Q.V. Le, Sequence to sequence learning with neural networks, in: Advances in Neural Information Processing Systems, 2014, pp. 3104–3112.
[36] A.M. Rush, S. Chopra, J. Weston, A neural attention model for abstractive sentence summarization, in: Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, 2015, pp. 379–389.
[37] R. Socher, A. Perelygin, J. Wu, J. Chuang, C.D. Manning, A. Ng, C. Potts, Recursive deep models for semantic compositionality over a sentiment treebank, in: Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, 2013, pp. 1631–1642.
[38] I. Iacobacci, M.T. Pilehvar, R. Navigli, Embeddings for word sense disambiguation: An evaluation study, in: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Berlin, Germany, 2016, pp. 897–907.
[39] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, 2018, arXiv:1810.04805.
[40] J. Camacho-Collados, M.T. Pilehvar, From word to sense embeddings: A survey on vector representations of meaning, J. Artif. Int. Res. (ISSN: 1076-9757) 63 (1) (2018) 743–788.
[41] E.A. Corrêa Jr, D.R. Amancio, Word sense induction using word embeddings and community detection in complex networks, Physica A 523 (2019) 180–190.
[42] 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. (2020).
[43] B. Perozzi, R. Al-Rfou, V. Kulkarni, S. Skiena, Inducing language networks from continuous space word representations, in: Complex Networks V, Springer, 2014, pp. 261–273.
[44] S. Fortunato, Community detection in graphs, Phys. Rep. 486 (3) (2010) 75–174.
[45] M.E. Newman, M. Girvan, Finding and evaluating community structure in networks, Phys. Rev. E 69 (2) (2004) 026113.
[46] M.E. Newman, Modularity and community structure in networks, Proc. Natl. Acad. Sci. 103 (23) (2006) 8577–8582.
[47] V.D. Blondel, J.-L. Guillaume, R. Lambiotte, E. Lefebvre, Fast unfolding of communities in large networks, J. Statist. Mech.: Theory Exp. 2008 (10) (2008) P10008.
[48] E. Corrêa Jr., A.A. Lopes, D.R. Amancio, Word sense disambiguation: A complex network approach, Inform. Sci. 442–443 (2018) 103–113.
[49] M. Newman, Networks: An Introduction, Oxford University Press, Inc., New York, NY, USA, 2010.

[50] L. Antiqueira, O.N. Oliveira Jr, L. da Fontoura Costa, M.d.G.V. Nunes, A complex network approach to text summarization, Inform. Sci. 179 (5) (2009) 584–599.
[51] A.A. Markov, An example of statistical investigation of the text eugene onegin concerning the connection of samples in chains, Proc. Bibliogr. Acad. Sci. 7 (6) (1913) 153–162.
[52] C.D. Manning, H. Schütze, Foundations of Statistical Natural Language Processing, MIT Press, Cambridge, MA, USA, ISBN: 0-262-13360-1, 1999.
[53] J.M. Ponte, W. Croft, A language modeling approach to information retrieval, in: Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, ACM, 1998, pp. 275–281.
[54] F. Li, T. Dong, Text categorization based on semantic cluster-hidden markov models, in: International Conference in Swarm Intelligence, Springer, 2013, pp. 200–207.
[55] S. Wernicke, F. Rasche, FANMOD: a tool for fast network motif detection, Bioinformatics 22 (9) (2006) 1152–1153.
[56] S. Omidi, F. Schreiber, A. Masoudi-Nejad, MODA: an efficient algorithm for network motif discovery in biological networks, Genes Genet. Syst. 84 (5) (2009) 385–395.
[57] R. Milo, S. Shen-Orr, S. Itzkovitz, N. Kashtan, D. Chklovskii, U. Alon, Network motifs: simple building blocks of complex networks, Science 298 (5594) (2002) 824–827.
[58] D.R. Amancio, E.G. Altmann, D. Rybski, O.N. Oliveira Jr, L.F. Costa, Probing the statistical properties of unknown texts: application to the voynich manuscript, PLoS One 8 (7) (2013) e67310.
[59] V.Q. Marinho, G. Hirst, D.R. Amancio, Labelled network subgraphs reveal stylistic subtleties in written texts, J. Complex Netw. 6 (4) (2018) 620–638.
[60] Y. Kavurucu, A comparative study on network motif discovery algorithms, Int. J. Data Min. Bioinform. 11 (2) (2015) 180—204.
[61] S. Wernicke, Efficient detection of network motifs, IEEE/ACM Trans. Comput. Biol. Bioinform. 3 (4) (2006) 347–359.
[62] B.C. Van Wijk, C.J. Stam, A. Daffertshofer, Comparing brain networks of different size and connectivity density using graph theory, PLoS One 5 (10) (2010) e13701.
[63] D.R. Amancio, Comparing the topological properties of real and artificially generated scientific manuscripts, Scientometrics 105 (3) (2015) 1763–1779.
[64] M. Tumminello, S. Micciche, F. Lillo, J. Piilo, R.N. Mantegna, Statistically validated networks in bipartite complex systems, PLoS One 6 (3) (2011).
[65] F. Radicchi, J.J. Ramasco, S. Fortunato, Information filtering in complex weighted networks, Phys. Rev. E 83 (4) (2011) 046101.
[66] M.Á. Serrano, M. Boguná, A. Vespignani, Extracting the multiscale backbone of complex weighted networks, Proc. Natl. Acad. Sci. 106 (16) (2009) 6483–6488.
[67] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, et al., Scikit-learn: Machine learning in python, J. Mach. Learn. Res. 12 (Oct) (2011) 2825–2830.
[68] M.Z. Rodriguez, C.H. Comin, D. Casanova, O.M. Bruno, D.R. Amancio, L.d.F. Costa, F.A. Rodrigues, Clustering algorithms: A comparative approach, PLoS One 14 (1) (2019).
[69] S. Segarra, M. Eisen, A. Ribeiro, Authorship attribution through function word adjacency networks, IEEE Trans. Signal Process. 63 (20) (2015) 5464–5478.
[70] S. Segarra, M. Eisen, G. Egan, A. Ribeiro, Attributing the authorship of the henry VI plays by word adjacency, Shakespear. Quart. 67 (2) (2016) 232–256.
[71] W. Shang, H. Huang, H. Zhu, Y. Lin, Y. Qu, Z. Wang, A novel feature selection algorithm for text categorization, Expert Syst. Appl. (ISSN: 0957-4174) 33 (1) (2007) 1–5.
[72] R. Kiros, Y. Zhu, R.R. Salakhutdinov, R. Zemel, R. Urtasun, A. Torralba, S. Fidler, Skip-thought vectors, in: Advances in Neural Information Processing Systems, 2015, pp. 3294–3302.
[73] A. Baronchelli, R. Ferrer-i Cancho, R. Pastor-Satorras, N. Chater, M.H. Christiansen, Networks in cognitive science, Trends Cogn. Sci. 17 (7) (2013) 348–360.
[74] E.A. Corrêa Jr, F.N. Silva, L.F. Costa, D.R. Amancio, Patterns of authors contribution in scientific manuscripts, J. Informetr. 11 (2) (2017) 498–510.
[75] C.T. Kello, G.D. Brown, R. Ferrer-i Cancho, J.G. Holden, K. Linkenkaer-Hansen, T. Rhodes, G.C. Van Orden, Scaling laws in cognitive sciences, Trends Cogn. Sci. 14 (5) (2010) 223–232.