Physica A 510 (2018) 110–120

Contents lists available at ScienceDirect

Physica A

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

An image analysis approach to text analytics based on complex networks

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

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

HIGHLIGHTS

- A method to analyze the unfolding of texts is proposed.
- The method can be used to complement document classification systems.
- The obtained visualizations can be used to analyze similar complex systems.

ARTICLE INFO

Article history:
Received 23 January 2018
Received in revised form 21 May 2018
Available online 23 June 2018

Keywords:
Complex networks
Mesoscopic networks
Authorship
Text mining
Computer vision

ABSTRACT

Text network analysis has received increasing attention as a consequence of its wide range of applications. In this study, we extend a previous work founded on the study of topological features of mesoscopic networks. Here, the geometrical properties of visualized networks are quantified by using several image analysis techniques. Such properties are used to probe the networks characteristics in terms of authorship. It was found that the visual features account for performance similar to that achieved by using topological measurements. Also, we combined and compared the two types of features, topological and geometrical, and the results suggest that the information provided by network topology and image features are complementary.

1. Introduction

With the ever-increasing availability of machine-readable text, the interest in automatic textual analysis tools has grown substantially given its potential for several important applications. One area of special interest is authorship attribution, which has been studied from different perspectives. In authorship attribution, we are interested in assigning an author to a given text [1,2]. Traditional approaches use statistical analysis of lexical and syntactical features [1,3,4], while other methods focus on text modeling using complex networks [5–11]. The latter is capable of representing structural as well as semantic characteristics [12–16], complementing traditional methods.

Recently, a network method for text modeling was proposed [17]. This method uses mesoscopic networks to capture the flow of narratives. In [18], this approach was used for authorship attribution. By *mesoscopic* it is meant that the derived networks can reflect text relationships at a topological scale larger than usually approached by using, for instance, word

[*] Corresponding author at: Institute of Mathematics and Computer Science, University of São Paulo, Brazil.
E-mail addresses: h.f.arruda@gmail.com (H.F. de Arruda), diego@icmc.usp.br (D.R. Amancio).

https://doi.org/10.1016/j.physa.2018.06.110

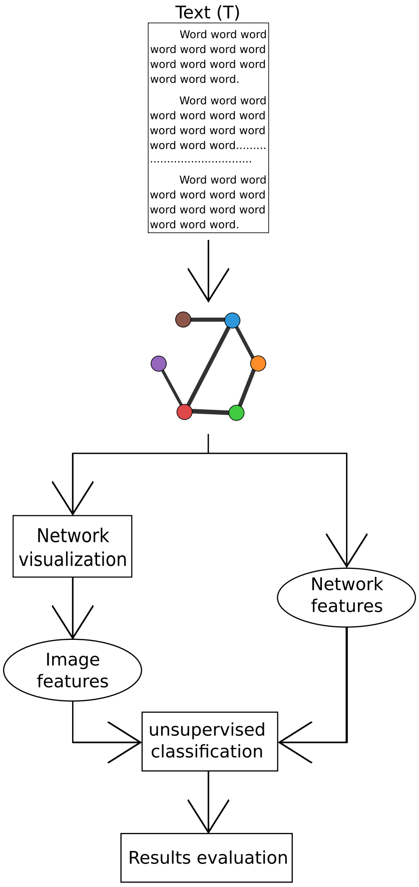

Text (T)
Word word word word word word word word word word word word word word.
Word word word word word word word word word word word word word word.........
....
Word word word word word word word word word word word word word word.
Network visualization
Network features
Image features
unsupervised classification
Results evaluation

Fig. 1. Scheme illustrating the mapping of texts into networks and the classification process. Initially, documents are mapped into networks. Then, two types of features are calculated: network features and attributes derived from network visualizations. The classification can then be performed using both network and image features.

adjacency. It has been suggested that such model can provide insights into the discursive structure of texts [18], an issue which is explored further in the current work.

The work presented in [18] focuses on network feature extraction, while the present incorporates features from the visualizations of these networks. This has been done to explore how the visual features of the networks, which was called calligraphy in [18], can contribute to the accuracy of authorship attribution. The framework used in this work is illustrated in Fig. 1. The obtained results suggest that these network visual features are as capable as topological measurements to characterize authors' styles. Furthermore, when the visual and topological features were combined, a better result is achieved.

The remaining of this paper is organized as follows: Section 2 defines the authorship attribution problem and presents some traditional and network-based approaches. The proposed pipeline, which includes the process of obtaining mesoscopic networks and the selected image and network measurements, is presented in Section 3. Then, Section 4 reports the obtained results using the image and network features. Finally, our conclusions are drawn in Section 5.

2. Related work

Given a text of unknown or disputed authorship and a set of candidate authors, the goal of an authorship attribution is to identify the correct unknown author [1]. One of the most important findings in authorship studies was reported by Mosteller and Wallace [19]. They investigated the authorship of several political documents and discovered that the frequencies of common words – such as some pronouns, prepositions, and articles – are useful to characterize the authorship of texts.

Traditionally, works in authorship attribution use simple, yet useful features to characterize writing styles, such as statistics extracted from words and characters (e.g., keywords, frequency of the *n*-grams of words or characters, and the frequency of punctuation marks) [1,3,4,20,21]. Syntactic and semantic features can also be used for the task, such as the frequency of the constituent parts of a sentence and information about words synonyms [1]. In recent years, authorship attribution has been studied from a different perspective. Some works have taken advantage of dense and real-valued vectorial representations and deep neural networks [22–24], such as their application to train language models for each author or to learn continuous vectors for n-grams of words.

Complex network-based approaches have also been used to assign authorship. In most of these methods, co-occurrence networks are created from the texts [5,8–10]. In such approaches, co-occurrence networks are created by connecting adjacent words. Then, several topological measurements – such as the clustering coefficient, degree, frequency of motifs – are extracted, and their respective values are used as features in machine learning algorithms. These previous works have shown that structure plays a prominent role in characterizing authors' styles. Interestingly, even if only stopwords are used, the structure of the remaining structure in the networks is still essential for the accurate identification of styles [8].

Another possibility for text representation is to map texts as mesoscopic networks, thus providing a model for identifying the relationship between larger chunks of texts [17]. Interestingly, Marinho et al. [18] presented mesoscopic networks created from a dataset and the provided visualizations revealed characteristics of the authors, such as the preference for short stories over novels, as well as characteristics of each text, such as the similarities between the beginning and end of a particular book. These mesoscopic networks also characterize the unfolding of the stories.

3. Materials and methods

Our authorship attribution method is based on the mesoscopic modeling approach [17]. The first step to create mesoscopic networks is the preprocessing of the text. In this study, we remove stopwords (such as articles and prepositions), and the remaining words are lemmatized, which means that the inflected words were grouped into a single lemma. For example, a given word and its respective plural form shall be represented by the same lemma (i.e. the canonical form of a word). For nouns and verbs, the canonical forms are their singular and bare infinitive versions, respectively. Many methods have been proposed to lemmatize texts, here we employed the WordNet Lemmatizer, which uses the WordNet dataset to find lemmas [25]. The preprocessed text is then partitioned into a set of paragraphs $T = \{p_1, p_2 \ldots, p_n\}$, where n is the total number of paragraphs, and p_i is the set of words belonging to the same paragraph (see Fig. 2(a)). Regarding the paragraphs order, all possible sets of Δ consecutive paragraphs are grouped, as shown in Fig. 2(b). Each set $W_i^{\Delta} = \{p_i, p_{i+1} \ldots, p_{i+\Delta-1}\}$ represents a network node. From these nodes, a weighted and undirected network is created, in which all nodes i and j are connected, and the weights are computed as the cosine similarity between the vector containing the values of tf–idf statistics associated to each node [26], as illustrated in Fig. 2(c). In the current study, we employed the following tf–idf mapping

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

where $f_{w,d}$ is the frequency of a given word w occurring in document d comprising n words, $|D|$ is the total number of documents and d_w is the number of documents in which the word w occurs. Note that, in this approach, each set of consecutive paragraphs W, i.e., each node, is considered a document.

In order to obtain unweighted networks, we remove the edges with the lowest weights until a certain criterion is met, as can be seen in Fig. 2(d). In this paper, the selected criterion requires that all networks have the same network average degree $\langle k \rangle = 2E/N$, where E is the number of edges and N is the number of nodes. The selected value for $\langle k \rangle$ was 40. Note that such an approach was previously used in [18].

3.1. Dataset

We used a dataset of 50 English texts written by ten authors. These books are available from the Project Gutenberg repository.[1] The list of the 50 texts is presented in Table 1.

3.2. Network analysis

After the networks have been created, topological measurements can be extracted. Here, we measure accessibility (for $h = \{2, 3\}$) [27], degree [28], backbone and merged symmetry (for $h = \{2, 3, 4\}$) [29,30], assortativity [31], average degree of neighbors [32] and clustering coefficient [28]. Because most of the selected measurements apply to a single node, some statistics – such as the average, standard deviation, and skewness – were extracted from each distribution and used as features for the classification algorithms.

3.3. Image analysis

According to Marinho et al. [18], the visualization of mesoscopic networks can provide information about the writing style of authors. In the present study, we used image processing analysis to extract characteristics from these visualizations, which are used to characterize authors. First, the complex networks are visualized, using a force directed algorithm, which is based on attraction force between connected nodes and repulsion force between all pairs of nodes [33]. To that end, we

[1] Project Gutenberg - https://www.gutenberg.org/

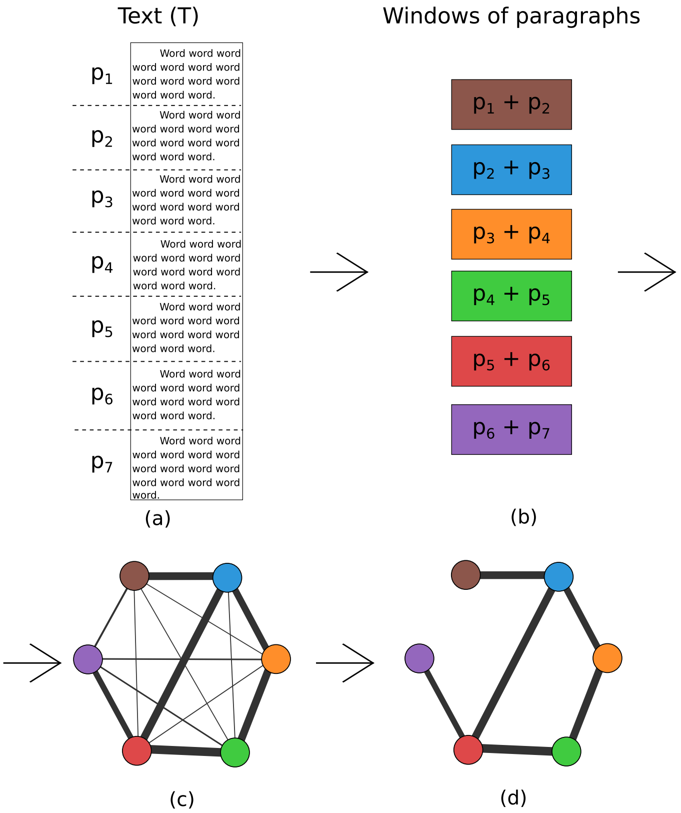


Fig. 2. Mesoscopic network. In (a), the document is divided into paragraphs. The nodes of the network are defined as adjacent paragraphs in (b). The links are established according to the textual similarity of the nodes in (c). A threshold is applied in (d) to remove the weakest links.

Table 1
List of 50 texts used in this work. The list comprises books from ten authors.

Author: Texts
Herman Melville: Moby Dick, Or, The Whale; The Confidence-Man: His Masquerade; The Piazza Tales; Typee: A Romance of the South Seas; White Jacket, Or, The World on a Man-of-War
B. M. Bower: Cabin Fever; Lonesome Land; The Long Shadow; The Lookout Man; The Trail of the White Mule
Jane Austen: Emma; Mansfield Park; Persuasion; Pride and Prejudice; Sense and Sensibility
Mark Twain: A Connecticut Yankee in King Arthur's Court; Adventures of Huckleberry Finn; The Adventures of Tom Sawyer; The Prince and the Pauper; Roughing It
Charles Darwin: Coral Reefs; Geological Observations on South America; The Different Forms of Flowers on Plants of the Same Species; The Expression of the Emotions in Man and Animals; Volcanic Islands
Charles Dickens: American Notes; A Tale of Two Cities; Barnaby Rudge: A Tale of the Riots of Eighty; Great Expectations; Hard Times
Edgar Allan Poe: The Works of Edgar Allan Poe (Volume 1 - 5)
Hector H. Munro (Saki): Beasts and Super Beasts; The Chronicles of Clovis; The Toys of Peace; The Unbearable Bassington; When William Came
Thomas Hardy: A Changed Man and Other Tales; A Pair of Blue Eyes; Far from the Madding Crowd; Jude the Obscure; The Hand of Ethelberta
Henry James: The Ambassadors; The American; The Portrait of a Lady—Volume 1; The Real Thing and Other Tales; The Turn of the Screw

employed the software provided by Silva et al. [34] with fixed parameters. 2D visualizations were used to remove the need to define the projection angle. Furthermore, the images were converted to monochromatic versions (called binary images). A preprocessing step was employed in which the images are dilated and eroded [35] in order to remove small holes. We considered the object size and the higher eigenvalue (λ_1) of the Principal Component Analysis (PCA) [36] to find the dilation

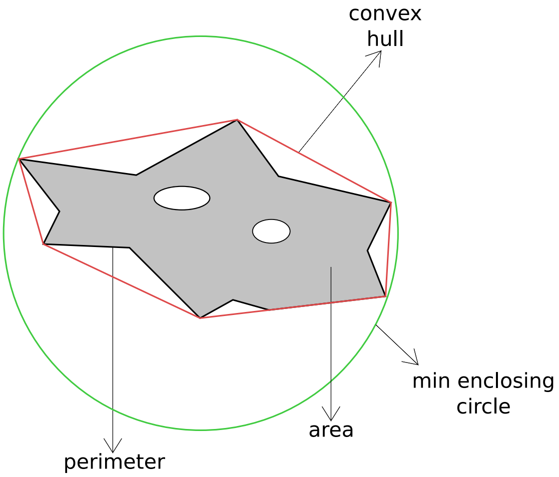


Fig. 3. Example of the following image features: area, perimeter, minimum enclosing circle, and convex hull. Note that the Euler number of this image is $e = -1$ because $N_o = 1$ and $N_h = 2$.

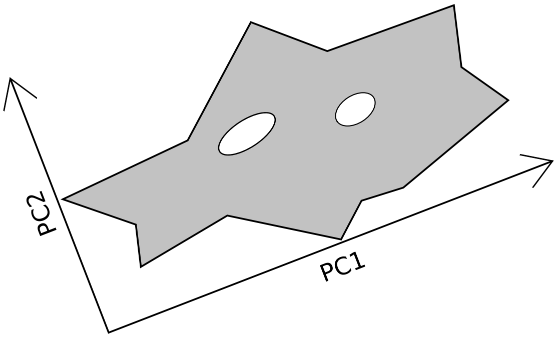


Fig. 4. Example of elongation measurement, where and λ_2 are associated to PC1 and PC2, respectively. In this example $\lambda_1 = 100$ and $\lambda_2 = 50$, so the elongation is 2.

kernel as follows:

$$c_{dim} = c\lambda_1, \tag{2}$$

where c is a constant, set as $c = 3 \cdot 10^{-4}$. In other words, the larger the object is, the larger the kernel size.

The following features were extracted from the preprocessed images:

1. ***Area***: The object area A is calculated as the sum of all pixels of the object in the image (see Fig. 3);
2. ***Perimeter***: The perimeter of an object, P, is the arc length of the external border of the object. An example is shown in Fig. 3;
3. ***Euler number***: This measurement is derived from the number of holes in the image. It is computed as:

$$e = N_o - N_h, \tag{3}$$

   where N_h is the number of holes in the image and N_o is the number of objects. An example is shown in Fig. 3, where $N_o = 1$, $N_h = 2$ and $e = -1$. Note that for all images in this study, $N_o = 1$;
4. ***Minimum enclosing circle***: The minimum enclosing circle M_c is the smallest circle that includes all of the objects' pixels [37]. We measured the radius and the x and y coordinates of the central point of this circle. Fig. 3 illustrates these measurements;
5. ***Convex hull***: similar to the minimum enclosing circle, the convex hull consists of the smallest convex polygon including all the pixels from the object [38] (see Fig. 3). We compute the polygons' area C_a, perimeter C_p, and residual area C_r. The latter is calculated as:

$$C_r = C_a - A. \tag{4}$$

6. ***Elongation***: This measurement quantifies how stretched an object is. It is calculated as the ratio between the variances of the first (λ_1) and second (λ_2) axes of the PCA of the image [39]. An example of such PCA is shown in Fig. 4.

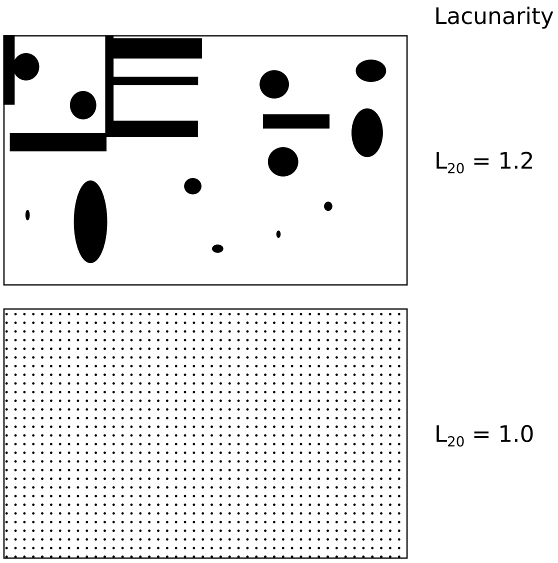


Fig. 5. Two examples of images and their respective lacunarity with $r = 20$. Here, the objects are shown in black, and the holes are white.

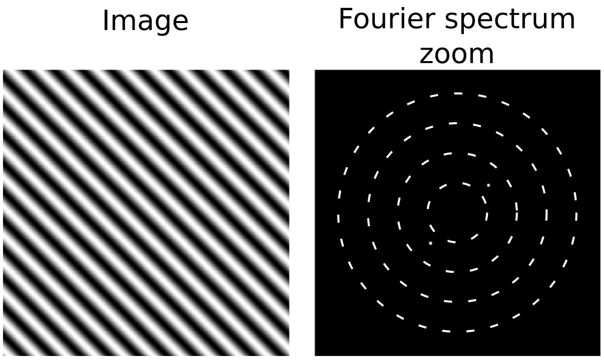


Fig. 6. Example of an image with a 2D plane wave and its respective magnitude Fourier spectrum. Note that there are only two points in the spectra. Observe that the orientation defined by these two points coincides with that of the plane wave. The dashed lines are rings with fixed width used to determine the features.

7. ***Lacunarity***: This feature [40] quantifies the variation of the size of the 'voids' of the objects in an image, as shown in Fig. 5. Note that when the holes are more uniform, the lacunarity (L) value is lower. In this study, we employed the self-referred approach to lacunarity proposed by Rodrigues et al. [41], which involves circular windows of radius r.
8. ***Fourier***: The Fourier transform represents the signal (image) in the frequency domain [39]. In this paper, we consider the magnitude of this transform. Non-overlapping rings with fixed width are defined in the magnitude Fourier space (see Fig. 6). The entropy, average, and standard deviation value of magnitudes inside each ring are obtained and employed as features.

4. Results and discussion

In this section, we will provide a qualitative analysis of the visual properties of the mesoscopic networks studied in this work, complemented by the discussion of the accuracy of the authorship attribution results obtained from experiments considering all authors and experiments considering only pairs of authors. We also provide a principal component analysis in order to visualize the data clusterization. Finally, we discuss how the proposed model can be used to analyze texts written in different languages.

4.1. Image analysis

So as to illustrate some of the visual specificities present in the constructed networks in this paper, we used a visualization method to generate three networks from each of four different authors. These were: Edgar Allan Poe, Saki, Mark Twain and

Table 2
Classification accuracy and number of used features considering different measurements.

	EM		KMeans	
	Accuracy	Features	Accuracy	Features
IF	54%	14	48%	13
NF	54%	15	50%	3
IF+NF	58%	19	54%	34

Henry James. These networks are presented in Fig. 7. In order to better visualize the narrative flow, the color was added to the images. Several of the loops that appear in the images are a consequence of the network visualization. These loops could have been removed manually, but we chose not to interfere in any step of the process.

The obtained images suggest a kind of *visual network* whose edges are narrow strings while the nodes correspond to intersections appearing along these strings. Observe that the string extremities also define nodes. Pieces of the same string initiating and terminating in a node are henceforth called *visual loops*. Nodes with several connections are called *visual hubs*. Two types of these hubs can be identified in Fig. 7: *tight hubs*, as seen in Huckleberry Finn's network; and *loose hubs*, found in The Turn of the Screw.

The images' most noticeable feature is clearly the absence or presence of visual nodes in the networks. Both Saki and Poe have networks without many nodes. It can easily be explained since them both favor collections of unrelated short stories. Nonetheless, there are some small differences between the aforementioned networks; Poe's networks depict some small-scale loops that are absent in Saki's networks. We hypothesize that, although these differences are visually minor, they will have a notable effect on the respective mesoscopic network measurements.

Henry James' The Real Thing and Other Tales also present small-scale loops, but in this case, they are more pronounced. Contrariwise, in the other books from James, especially in The Turn of The Screw, the presence of longer loops is more prevalent. We attribute this effect to the book's plot, where almost all the story consists of the narration of a single character and her interactions with only three other people inside a house, using therefore similar vocabulary throughout the book. We also suspect that by being a suspense/horror story the narrative keeps returning to the same place to build up suspense.

Twain's books also have their peculiarities. His books' networks are much more convoluted than the other authors', having many more long loops. It is indicative that he had a preference for re-using specific patterns throughout his novels. These patterns could be attributed to either characters' speech patterns or situational vocabulary. The former is easily seen in Huckleberry Finn's network, organized around a tight visual hub. In this novel, Twain tried to represent the phonetic differences of a slave's speech pattern. Therefore, whenever Jim (a recently freed slave) interacts with other characters, several distinctive words are used, causing it to have a very typical tf–idf distribution, leading to the tight hub.

It is interesting to notice that both Huckleberry Finn and The Turn of the Screw have many long loops and also present a central hub, but there are clear differences. We suppose that these differences are due to distinct underlying reasons. In Huckleberry Finn, the long loops appear to be caused by a single character distinctive vocabulary. On the other hand, The Turn of the Screw is a book that takes place mostly in the same location, with few different characters, and therefore shares similar situations throughout the book, making these long loops more spread out.

4.2. *Authorship attribution*

In order to evaluate the capability of the proposed method to characterize books according to their respective authors, we calculated all the image features, IF henceforth, described in Section 3.3. These features were selected and ordered according to SVM's (Support Vector Machine) attribute selection [42]. Furthermore, we used expectation maximization (EM) [43] and KMeans [44], which are unsupervised classifier algorithms. The number of classes was set to 10 to reflect the number of authors. In order to find the best number of features, we tested the n first features in the ranking by varying n from 1 to $\mathcal{F}$, where $\mathcal{F}$ is the total number of features (see Table 2). The best classification accuracy, 54%, was achieved using $n = 14$. This result supports the hypothesis that there are visual differences among the visualizations of the mesoscopic networks, which are captured by the image features.

An alternative method to classify these texts is to use features obtained from the complex network measurements, or network features (NF). The selected measurements were presented in Section 3.2. We employed the same machine learning methodology that was applied to the image features and obtained an accuracy of 54% for $n = 15$. We also combined all features (IF + NF) and classified the texts using the same methodology. In this case, a better accuracy rate was achieved: 58%, with $n = 19$.

4.3. *Pairwise authorship attribution*

In Fig. 8 we have the results for pairwise classifications, where only two authors are considered for each experiment. We utilize the classifiers with the best accuracy from the previous subsection. We can see that, as expected, the results that consider both image and network features are better than those that consider only one type of feature.

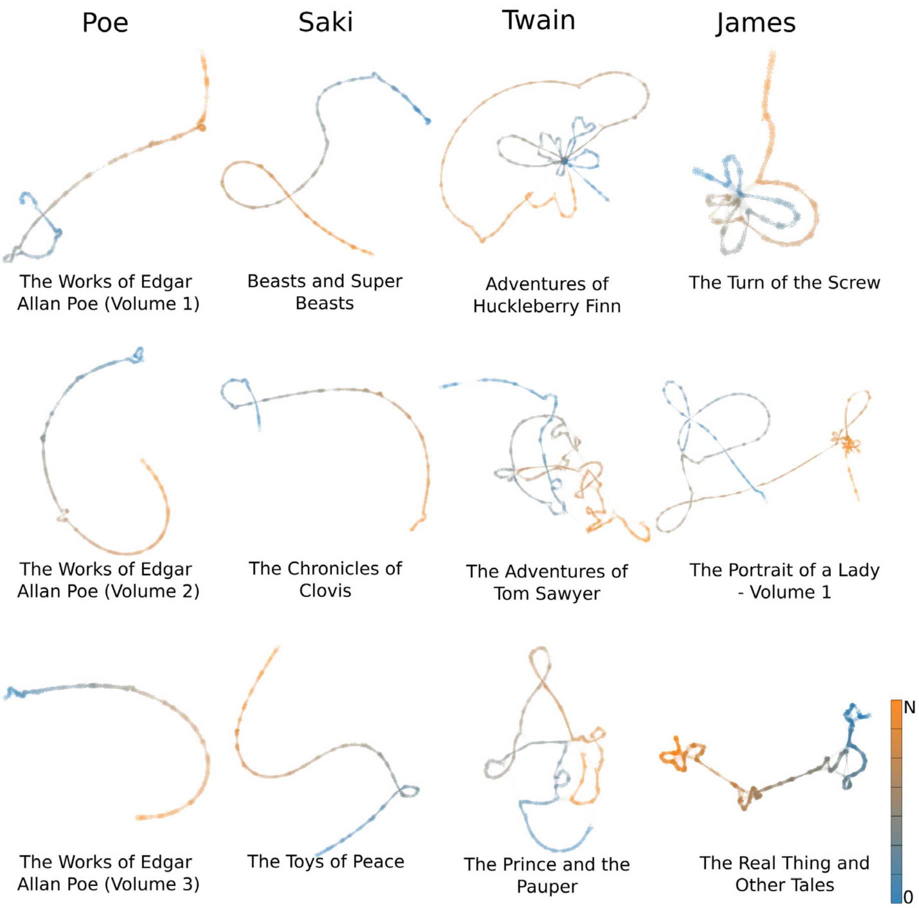
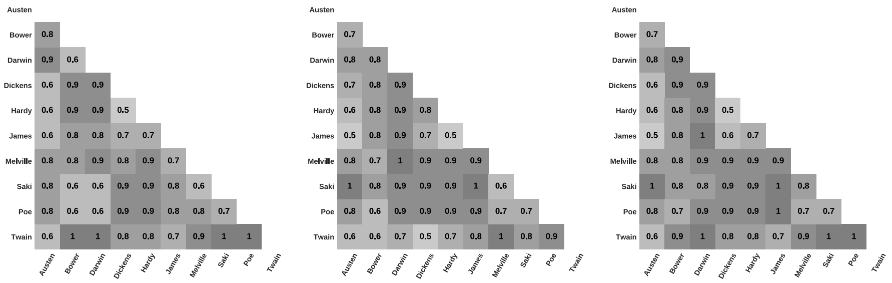

Poe
Saki
Twain
James
The Works of Edgar Allan Poe (Volume 1)
Beasts and Super Beasts
Adventures of Huckleberry Finn
The Turn of the Screw
The Works of Edgar Allan Poe (Volume 2)
The Chronicles of Clovis
The Adventures of Tom Sawyer
The Portrait of a Lady - Volume 1
The Works of Edgar Allan Poe (Volume 3)
The Toys of Peace
The Prince and the Pauper
The Real Thing and Other Tales
N
0

Fig. 7. The discussed mesoscopic networks. The colors represent the node order. The colors represent the node order.



	Austen	Bower	Darwin	Dickens	Hardy	James	Melville	Saki	Poe	Twain
Austen										
Bower	0.8									
Darwin	0.9	0.6								
Dickens	0.6	0.9	0.9							
Hardy	0.6	0.9	0.9	0.5						
James	0.6	0.8	0.8	0.7	0.7					
Melville	0.8	0.8	0.9	0.8	0.9	0.7				
Saki	0.8	0.6	0.6	0.9	0.9	0.8	0.6			
Poe	0.8	0.6	0.6	0.9	0.9	0.8	0.8	0.7		
Twain	0.6	1	1	0.8	0.8	0.7	0.9	1	1	

(a) IF

	Austen	Bower	Darwin	Dickens	Hardy	James	Melville	Saki	Poe	Twain
Austen										
Bower	0.7									
Darwin	0.8	0.8								
Dickens	0.7	0.8	0.9							
Hardy	0.6	0.8	0.9	0.8						
James	0.5	0.8	0.9	0.7	0.5					
Melville	0.8	0.7	1	0.9	0.9	0.9				
Saki	1	0.8	0.9	0.9	0.9	1	0.6			
Poe	0.8	0.6	0.9	0.9	0.9	0.9	0.7	0.7		
Twain	0.6	0.6	0.7	0.5	0.7	0.8	1	0.8	0.9	

(b) NF

	Austen	Bower	Darwin	Dickens	Hardy	James	Melville	Saki	Poe	Twain
Austen										
Bower	0.7									
Darwin	0.8	0.9								
Dickens	0.6	0.9	0.9							
Hardy	0.6	0.8	0.9	0.5						
James	0.5	0.8	1	0.6	0.7					
Melville	0.8	0.8	0.9	0.9	0.9	0.9				
Saki	1	0.8	0.8	0.9	0.9	1	0.8			
Poe	0.8	0.7	0.9	0.9	0.9	1	0.7	0.7		
Twain	0.6	0.9	1	0.8	0.8	0.7	0.9	1	1	

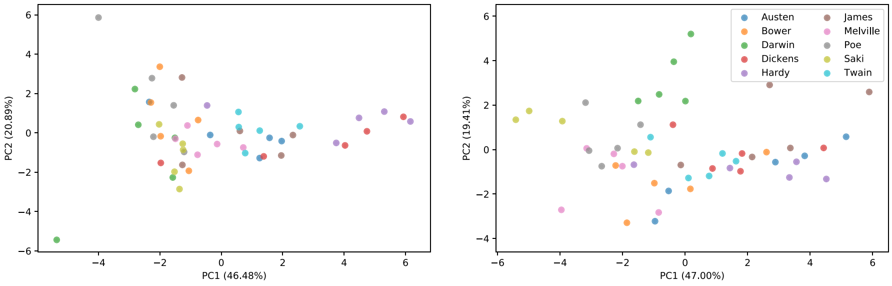
(c) IF + NF

Fig. 8. Accuracy rates (from 0 to 1) in the pairwise classification using only image features (a) - IF, only network features (b) - NF, and the combination of the two strategies (c) - IF + NF. Note that, in general, the accuracies increase when both strategies are combined and more pairs are classified with accuracy of 1.

Note that the results reflect our analysis in subSection 4.1: while the result for the comparison between Poe and Saki, and James and Twain yielded an accuracy of 70%, all other results have a 100% accuracy. It is a clear indication that the features selected reflect our intuitions and that it is even possible to discriminate between similar networks, albeit improvements are desirable.

4.4. PCA

Finally, PCA was used to check if our image and network features properly represent the desired characteristics. Fourteen features were selected according to SVM feature selector. We applied PCA to the image and network features separately



(a)

(b)

Fig. 9. PCA projections of the generated networks using only image (a) or network (b) features.

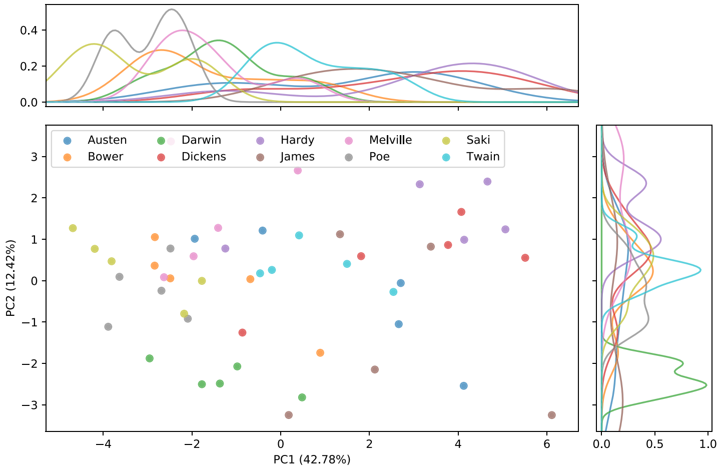
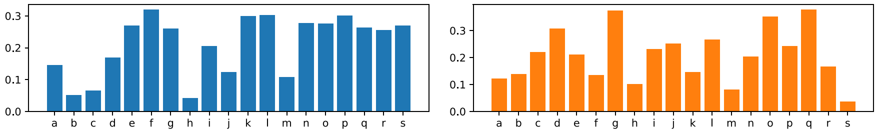


Fig. 10. PCA projections of the generated networks using both image and network features.



(a) PC1

(b) PC2

Fig. 11. Weights (in absolute values) of PCA (shown in Fig. 10) computed from the IF+NF measurements. Item (a) and (b) represents the first and second principal components, respectively. The considered measurements are: *a* - entropy of Fourier spectrum (ring with ratio 30); *b* - skewness of backbone symmetry ($h = 4$) *c* - standard deviation of backbone symmetry ($h = 3$); *d* - average of backbone symmetry ($h = 3$); *e* - skewness of clustering coefficient; *f* - entropy of Fourier spectrum (ratio 120); *g* - standard deviation of Fourier spectrum (ratio 105); *h* - entropy of Fourier spectrum (ring with ratio 90); *i* - average knn; *j* - standard deviation of backbone symmetry ($h = 2$); *k* - lacunarity (ratio 31); *l* - average merged symmetry ($h = 3$); *m* - elongation; *n* - skewness of merged symmetry ($h = 4$); *o* - average of Fourier spectrum (ring with ratio 45); *p* - average merged symmetry ($h = 4$); *q* - average of Fourier spectrum (ring with ratio 105); *r* - skewness of degree, and *s* - standard deviation of accessibility ($h = 2$).

and achieved reasonable separation, especially with network features, as shown in Fig. 9. Network and image features were combined achieving a better separation, as depicted in Fig. 10. According to this figure, it is clear that our features are capable of capturing some stylistic choices of the authors. For example, Poe and Saki's generated networks have similar structures and are fairly close to one another, but the features are distinctive enough to produce a little actual overlap. It is also interesting to note that the PCA weights in Fig. 11 show that both image and network features have a substantial contribution to the separation.

4.5. Comparing different languages

In this section, we show that the construction the proposed network representation is straightforward for languages other than English. In order to compare the topology of networks created from distinct languages, we considered different versions of the Holy Bible dataset [14]. The dataset comprises a fraction of the New Testament written in 16 different languages: Arabic, German, Korean, Russian, Basque, Greek, Latin, Swahili, English, Hebrew, Maori, Vietnamese, Esperanto, Hungarian,

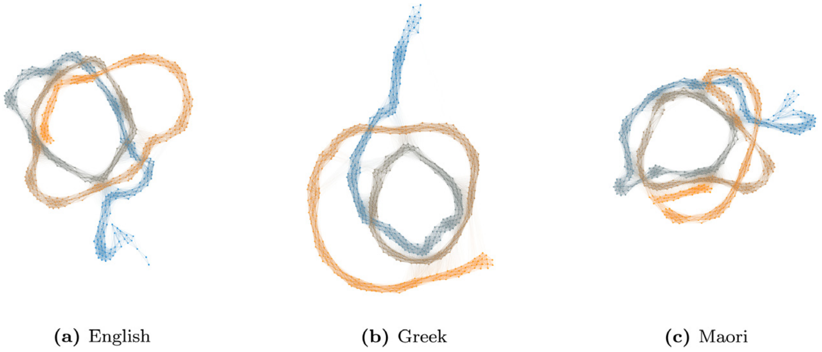


Fig. 12. Examples of network visualizations of books extracted from the Bible New Testament. The colors represent the nodes order. The first and last paragraphs are in blue and orange colors, respectively. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Portuguese, and Xhosa. For all considered languages, we analyzed three books: Matthew, Mark and Luke. As mentioned in the methodology, the text undergoes a preprocessing step before being mapped into a network. We decided not to fully preprocess the bible in order to show that, even if linguistic resources and tools were not available in other languages, it would still be possible to obtain insights from the networks. For the present analysis, only punctuation marks were removed.

All paragraphs were manually separated, and we considered the same network parameters as in the previous section. Fig. 12 shows the visualization of three examples of the networks. The network visualizations obtained from all the considered languages revealed that even if the languages are different, the obtained networks resulted similar, suggesting that the method can be used to identify patterns that go beyond language specificities.

5. Conclusions

Text networks derived from books have been extensively studied, especially regarding authorship recognition. A great deal of the reported approaches are based on frequency of words or adjacency networks. More recently [17], an approach has been reported that attempts at extending the context of text networks to a mesoscopic topological level by considering more context as given by long range words relationships. In the present work, we went a step further and, instead of only visualizing the text networks as in [18], employed a set of geometrical measurements derived from image analysis research in order to obtain additional information about the networks with potential for contributing to improving text networks analyses, especially regarding authorship characterization.

The approach involves the application of a comprehensive set of image analysis techniques ranging from simple measurements such as area and perimeter to more sophisticated methodologies, such as Fourier spectra and lacunarity. Furthermore, in order to know how informative these visualizations features are, we employed a classifier and compared the results with standard topological network features. We found that quantitative visual analysis of these networks yielded results comparable to those obtained using only topological features. The visual features obtained from mesoscopic networks were found to be meaningful for authorship attribution. We also tried the combination of both types of features (topological and image analysis features), which improved further the results. The proposed methodology has some intrinsic advantages concerning the analysis of narratives. In particular, it allows mesoscopic features of the text to become visible, from which the sequence and interrelationships between parts of the text can be easily identified. For instance, recurrences of subjects can be immediately identified from visual loops derived from the narrative. Such a feature cannot be obtained from traditional word co-occurrence models.

The obtained results motivate further research aimed at authorship attribution, as well as other types of networks derived from time series. In particular, it would be interesting to study if specific literary periods and styles have intrinsic visual patterns when represented by visualizations and respective measurements.

Acknowledgments

H.F.A. and T.S.L. thank CAPES for financial support. V.Q.M. and D.R.A. acknowledge financial support from São Paulo Research Foundation (FAPESP), Brazil (grant no. 15/05676-8, 16/19069-9). L.d.F.C. is grateful to CNPq (Brazil) (grant no. 307333/2013-2) and NAP-PRP-USP, Brazil for sponsorship. This work has been supported also by FAPESP, Brazil grants 11/50761-2 and 2015/22308-2.

References

[1] E. Stamatatos, A survey of modern authorship attribution methods, J. Am. Soc. Inf. Sci. Technol. 60 (3) (2009) 538–556.

[2] P. Juola, Authorship attribution, Found. Trends Inf. Retrieval 1 (3) (2006) 233–334.
[3] J. Grieve, Quantitative authorship attribution: An evaluation of techniques, Lit. Linguist. Comput. 22 (3) (2007) 251.
[4] M. Koppel, J. Schler, S. Argamon, Computational methods in authorship attribution, J. Am. Soc. Inf. Sci. Technol. 60 (1) (2009) 9–26.
[5] S. Lahiri, R. Mihalcea, Authorship attribution using word network features, 2013. arXiv:1311.2978.
[6] D.R. Amancio, A complex network approach to stylometry, PLoS One 10 (8) (2015) e0136076.
[7] D.R. Amancio, Probing the topological properties of complex networks modeling short written texts, PLoS One 10 (2) (2015) e0118394.
[8] S. Segarra, M. Eisen, A. Ribeiro, Authorship attribution through function word adjacency networks, IEEE Trans. Signal Process. 63 (20) (2015) 5464–5478.
[9] D.R. Amancio, O.N. Oliveira Jr., L.F. Costa, Structure-semantics interplay in complex networks and its effects on the predictability of similarity in texts, Physica A 391 (18) (2012) 4406–4419.
[10] A. Mehri, A.H. Darooneh, A. Shariati, The complex networks approach for authorship attribution of books, Physica A 391 (7) (2012) 2429–2437.
[11] D.R. Amancio, O. Oliveira Jr., L.F. Costa, Identification of literary movements using complex networks to represent texts, New J. Phys. 14 (4) (2012) 043029.
[12] R. Ferrer i Cancho, R.V. Solé, R. Köhler, Patterns in syntactic dependency networks, Phys. Rev. E 69 (2004) 051915.
[13] H. Liu, Statistical properties of chinese semantic networks, Chin. Sci. Bull. 54 (16) (2009) 2781–2785.
[14] 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.
[15] 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.
[16] A. Utsumi, A complex network approach to distributional semantic models, PLoS One 10 (8) (2015) e0136277.
[17] H.F. de 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.
[18] V.Q. Marinho, H.F. de Arruda, T. Sinelli, L.F. Costa, D.R. Amancio, On the "calligraphy" of books, in: Proceedings of TextGraphs-11: the Workshop on Graph-based Methods for Natural Language Processing, Association for Computational Linguistics, Vancouver, Canada, 2017, pp. 1–10.
[19] F. Mosteller, D.L. Wallace, Inference and Disputed Authorship: The Federalist Papers, Addison-Wesley, Reading, Mass., 1964.
[20] A. Martini, A. Cardillo, P.D.L. Rios, Entropic selection of concepts unveils hidden topics in documents corpora, 2018. arXiv:1705.06510.
[21] C. Carretero-Campos, P. Bernaola-Galvan, A. Coronado, P. Carpena, Improving statistical keyword detection in short texts: Entropic and clustering approaches, Physica A 392 (6) (2013) 1481–1492.
[22] D. Bagnall, Authorship clustering using multi-headed recurrent neural networks—Notebook for PAN at CLEF 2016, 2016.
[23] T. Solorio, P. Rosso, M. Montes-y-Gómez, P. Shrestha, S. Sierra, F.A. González, Convolutional neural networks for authorship attribution of short texts, in: Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2017, Valencia, Spain, April 3–7, 2017, Volume 2: Short Papers, 2017, pp. 669–674.
[24] Y. Sari, A. Vlachos, M. Stevenson, Continuous n-gram representations for authorship attribution, in: M. Lapata, P. Blunsom, A. Koller (Eds.), European Chapter of the Association for Computational Linguistics, EACL 2017, vol. 2, ACL, 2017.
[25] G.A. Miller, WordNet: a lexical database for English, Commun. ACM 38 (11) (1995) 39–41.
[26] C.D. Manning, H. Schütze, Foundations of Statistical Natural Language Processing, MIT Press, Cambridge, MA, USA, 1999.
[27] B. Travencolo, L.F. Costa, Accessibility in complex networks, Phys. Lett. A 373 (1) (2008) 89–95.
[28] L.F. Costa, F.A. Rodrigues, G. Travieso, P.R. Villas Boas, Characterization of complex networks: A survey of measurements, Adv. Phys. 56 (1) (2007) 167–242.
[29] F.N. Silva, C.H. Comin, T.K. Peron, F.A. Rodrigues, C. Ye, R.C. Wilson, E.R. Hancock, L. F. Costa, Concentric network symmetry, Inform. Sci. 333 (2016) 61–80.
[30] 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.
[31] M. Newman, Mixing patterns in networks, Phys. Rev. E 67 (2) (2003) 026126.
[32] R. Pastor-Satorras, A. Vázquez, A. Vespignani, Dynamical and correlation properties of the Internet, Phys. Rev. Lett. 87 (25) (2001) 258701.
[33] T.M.J. Fruchterman, E.M. Reingold, Graph drawing by force-directed placement, Softw. - Pract. Exp. 21 (11) (1991) 1129–1164.
[34] F.N. Silva, D.R. Amancio, M. Bardosova, L.d.F. Costa, O.N. Oliveira Jr., Using network science and text analytics to produce surveys in a scientific topic, J. Inform. 10 (2) (2016) 487–502.
[35] N. Efford, Digital Image Processing: A Practical Introduction using Java, Addison-Wesley, Harlow, England, 2000.
[36] I. Jolliffe, Principal Component Analysis, Wiley Online Library, 2002.
[37] S. Skyum, A simple algorithm for computing the smallest enclosing circle, Inform. Process. Lett. 37 (3) (1991) 121–125.
[38] J. Sklansky, Finding the convex hull of a simple polygon, Pattern Recognit. Lett. 1 (2) (1982) 79–83.
[39] L.d.F. Costa, R.M. Cesar, Shape Analysis and Classification: Theory and Practice, Image Processing Series, Taylor & Francis, 2000.
[40] R.E. Plotnick, R.H. Gardner, R.V. O'Neill, Lacunarity indices as measures of landscape texture, Landsc. Ecol. 8 (3) (1993) 201–211.
[41] E.P. Rodrigues, M.S. Barbosa, L.d.F. Costa, Self-referred approach to lacunarity, Phys. Rev. E 72 (2005) 016707.
[42] I. Guyon, J. Weston, S. Barnhill, V. Vapnik, Gene selection for cancer classification using support vector machines, Mach. Learn. 46 (2002) 389–422.
[43] S. Lloyd, Least squares quantization in PCM, IEEE Trans. Inf. Theory 28 (2) (1982) 129–137.
[44] A.P. Dempster, N.M. Laird, D.B. Rubin, Maximum likelihood from incomplete data via the EM algorithm, J. R. Stat. Soc. Ser. B Stat. Methodol. (1977) 1–38.