On the “Calligraphy” of Books

Vanessa Queiroz Marinho[*] **Henrique Ferraz de Arruda**[*] **Thales Sinelli Lima**[*]
Luciano da Fontoura Costa[†] **Diego Raphael Amancio**[*]
[*]Institute of Mathematics and Computer Science, University of São Paulo, Brazil
[†]São Carlos Institute of Physics, University of São Paulo, Brazil
Corresponding authors: {vanessa.qm.1,diegoraphael}@gmail.com

Abstract

Authorship attribution is a natural language processing task that has been widely studied, often by considering small order statistics. In this paper, we explore a complex network approach to assign the authorship of texts based on their mesoscopic representation, in an attempt to capture the flow of the narrative. Indeed, as reported in this work, such an approach allowed the identification of the dominant narrative structure of the studied authors. This has been achieved due to the ability of the mesoscopic approach to take into account relationships between different, not necessarily adjacent, parts of the text, which is able to capture the story flow. The potential of the proposed approach has been illustrated through principal component analysis, a comparison with the chance baseline method, and network visualization. Such visualizations reveal individual characteristics of the authors, which can be understood as a kind of calligraphy.

1 Introduction

The ever increasing availability of public content on the Internet – including books, tweets, and blog posts – has implied in many new developments in several natural language processing (NLP) areas such as machine translation, sentiment analysis, and authorship attribution. Recently, advancements in the latter task have been achieved by using complex networks (Antiqueira et al., 2006; Amancio et al., 2011; Lahiri and Mihalcea, 2013; Marinho et al., 2016; Akimushkin et al., 2017).

The network models used in many of these works are based on word co-occurrence. In this approach, each distinct word is represented by a node, and edges connect adjacent words. Although this networked representation has proven successful in many tasks, it is not without its share of problems. Co-occurrence networks do not portray the topical structure found in many texts and are usually devoid of community structure (de Arruda et al., 2016). In order to overcome this disadvantage, some techniques have been devoted to the *mesoscopic* representation of texts (de Arruda et al., 2016, 2017). de Arruda et al. (2017) proposed a novel networked model, in which each node represents a respective set of consecutive paragraphs, while weighted edges express the similarity between nodes. Their proposed network is able to extract the organization and flow of text by effectively capturing the similarity between the blocks of text. In addition, their method was employed to distinguish between real and shuffled texts. However, mesoscopic networks have not been applied to tackle other NLP tasks.

Most researchers in the field of authorship attribution assume that each author has a signature (known as authorial fingerprint) that distinguishes his/her writing from the others (Juola, 2006). So inspired, we decided to test the hypothesis that these authorial fingerprints are also visible at a mesoscopic scale. At this scale, distinctive graphical patterns of the course of the text emerge, akin to a “discourse calligraphy” of the author. Thus, in order to classify texts according to their authorship, we created mesoscopic networks from texts and employed a set of topological measurements. In particular, the main goal of this paper is to probe whether the authors’ writing styles correlate with the story flow of their books.

This paper is structured as follows: Section 2 briefly describes the problem and some complex network approaches for authorship attribu-

Proceedings of TextGraphs-11: the Workshop on Graph-based Methods for Natural Language Processing, ACL 2017, pages 1–10,
Vancouver, Canada, August 3, 2017. ©2017 Association for Computational Linguistics

tion. The process to create mesoscopic networks is explained in Section 3. In addition, we also describe the dataset, the selected measurements and the machine learning algorithms in Section 3. The obtained results are reported in Section 4. Finally, Section 5 outlines our conclusions and prospects for future work.

2 Related Work

Authorship attribution methods attempt to find the most likely author of a document (Stamatatos, 2009). Since the seminal work conducted by Mosteller and Wallace (1964), authorship attribution has been a widely studied problem and several different approaches have been proposed. One of the first approaches consisted in analyzing the frequency of common words, such as *to* or *the*, in order to classify political essays according to their authorship (Mosteller and Wallace, 1964).

Since then, Mosteller and Wallace (1964)'s method has been enhanced to incorporate different attributes capable of qualifying writing styles. These include lexical, character, syntactic, and semantic features (Stamatatos, 2009). Simple lexical and character features (e.g. frequency and burstiness of words and characters, average lengths of texts, and others) have been used in several works, as reported by Grieve (2007), Koppel et al. (2009), and Stamatatos (2009). Most of these works have achieved good results by using, for example, the frequency of stopwords. Examples of syntactic information include the frequencies of POS tags and constituency-based parsing tree rules (Baayen et al., 1996; Gamon, 2004; Hirst and Feiguina, 2007). Finally, semantic features can be extracted from semantic dependency graphs and from the semantic roles associated with some words (Gamon, 2004; Argamon et al., 2007).

The usage of network analysis in authorship attribution has already been studied from different perspectives. Antiqueira et al. (2006), one of the first works in the area, extracted some measurements from co-occurrence networks and discovered that these could be used to characterize the writing style of authors. Amancio et al. (2011) combined network measurements with the distribution of words to characterize the authorship of several books. Lahiri and Mihalcea (2013) carried out an in-depth authorship attribution study using more than 100 features extracted from co-occurrence networks. They found that local features (those extracted from individual nodes) outperform global features in the authorship attribution problem.

Apart from using traditional network measurements, the frequency of network motifs involving three nodes (Milo et al., 2002) was found useful to characterize the writing style (Marinho et al., 2016). Instead of considering the text as a static structure, Akimushkin et al. (2017) studied the topology evolution of co-occurrence networks extracted from different sections of the text. Unlike most of the previous mentioned works, in which stopwords are usually removed, Segarra et al. (2013) proposed an authorship attribution method based on networks formed only by stopwords.

3 Methods

In this section, we describe the process to create mesoscopic networks from raw texts. We also detail the network measurements and machine learning methods.

3.1 Mesoscopic Approach

There are several ways to represent texts as complex networks, such as co-occurrence, syntactic, semantic or similarity networks (Mihalcea and Radev, 2011; Cong and Liu, 2014). In this study, we adopt the mesoscopic network approach proposed by de Arruda et al. (2017). Such networks are able to represent the text unfolding along time, which is normally overlooked by traditional approaches. Moreover, these networks were used to classify documents between real and shuffled texts, using only simple statistics. The high accuracy rate obtained in that classification task led us to infer that mesoscopic networks are able to represent structural aspects of real texts, such as the organization and development of the author's idea.

In order to create the network from a given text (T), some preprocessing steps can be applied. In our study, we removed the stopwords, and the remaining words were lemmatized. Figure 1 illustrates the methodology used to create mesoscopic networks. In the first step, shown in Figure 1(a), the text is partitioned into a set of paragraphs, $T = (p_0, p_1, p_2, \cdots)$, where p_i is a sequence of the preprocessed words belonging to the same paragraph i. Different from the co-occurrence networks, where nodes represent words, in mesoscopic networks nodes encompass sequences of Δ consecutive paragraphs. More

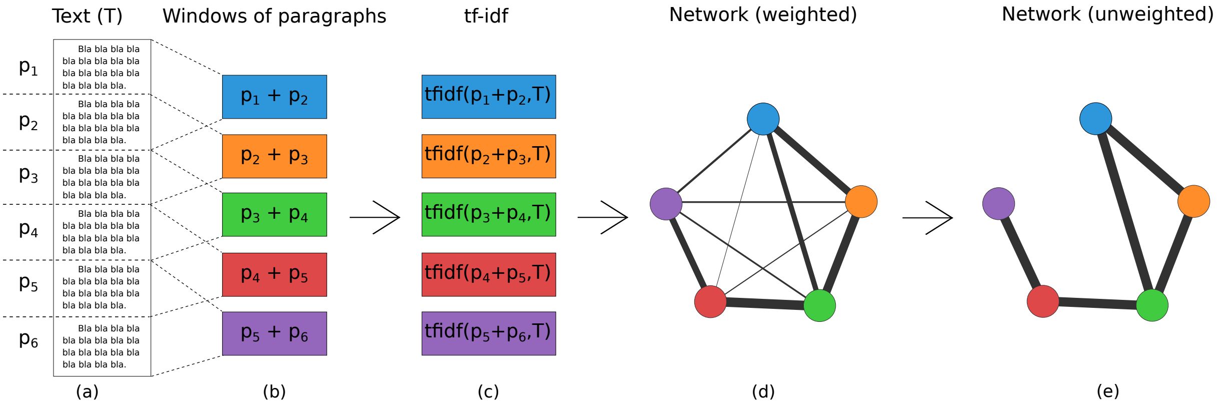


Figure 1: Illustration of the mesoscopic approach proposed by de Arruda et al. (2017). First, the text T is divided into subsequent paragraphs (a). Overlapping windows with $\Delta = 2$ paragraphs are shown in (b). Then, the tf-idf map is computed for all windows (c). Each pair of nodes (windows) i and j is now connected by an edge, weighted by the cosine similarity between their respective tf-idf maps (d). Next, in the network pruning phase, the edges with the lowest weights are removed until the network reaches a given average degree $\langle k \rangle$. The network in (e) illustrates the obtained unweighted mesoscopic network with $\langle k \rangle = 2$.

specifically, each possible subsequent set with Δ paragraphs, $W_i^\Delta = (p_i, p_{i+1}, \cdots, p_{i+\Delta-1})$, represents a network node, as shown in Figure 1(b).

So as to account for the importance of the words in a given paragraph, we applied the *tf-idf* (Manning and Schütze, 1999) statistics, which was originally proposed to quantify the importance of a given word w in a document d given a corpus D. A tf-idf(w, d, D) map is computed as

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

where $f_{w,d}$ is the frequency of word w in the document d, n is the total number of words in the document d, $|D|$ represents the total number of documents and d_w is the number of documents in which w occurs at least once. In order to apply the tf-idf measurement, we considered all the possible windows of subsequent paragraphs, W_i^Δ, as the set of documents D (see Figure 1(c)). Finally, for each pair of nodes i and j, a respective edge is created and its weight is calculated according to the cosine similarity between tf-idf(W_i^Δ, T) and tf-idf(W_j^Δ, T), where tf-idf(W_i^Δ, T) is a tf-idf vector of all words, computed from a given set of paragraphs W_i^Δ. This step is illustrated in Figure 1(d).

In order to convert the network from weighted to unweighted, the edges with the lowest weights can be removed, as described in Section 3.2. It should be noted that edges originating from adjacent paragraphs tend to have higher weights because of the implied overlap. Figure 1(e) shows an example of unweighted network. In our experiment, we set $\Delta = 20$, as empirically determined elsewhere (de Arruda et al., 2017).

3.2 Network Pruning

Mesoscopic networks are complete weighted graphs, i.e. every node is connected to every other node (Newman, 2010). In this paper, we repeatedly removed the edges with the lowest weights until each network reached a fixed network average degree $\langle k \rangle$. The average degree of a network g, with E edges and N nodes, is defined as

$$\langle k \rangle = \frac{2 * E}{N}. \quad (2)$$

We used several values of $\langle k \rangle$, ranging from 5 to 50, by steps of 5.

3.3 Network Measurements

The following network measurements were extracted from the networks[1]. Most of these measurements (apart from assortativity) apply to a single node. So, in order to obtain more global characterization, we calculated the average, standard deviation and skewness (third moment) of each distribution. The obtained statistics from these distributions were then used as features in the machine learning methods.

[1] For most of these measurements, we used the Igraph software package (Csardi and Nepusz, 2006)

Degree: The degree quantifies the number of connections of a node (Costa et al., 2007). Even though the average degree of all networks is the same as a consequence of network pruning, the degree of each node may still vary inside the network. Therefore, we used the standard deviation and skewness of this measurement, disregarding the average.

Average Degree of Neighbors: The average degree of neighbors (Pastor-Satorras et al., 2001) quantifies how well connected are the neighbors of a node.

Assortativity: As described by Newman (2003), the assortativity quantifies how likely it is for a given node to connect to other nodes with similar degree. Lower than zero values of assortativity are obtained when a node tends to connect to others with very different degrees. When a node connects only to others with the same degree, the assortativity becomes one. Null assortativity indicates that there is no correlation.

Clustering Coefficient: This measurement reflects how well interconnected are the neighbors of a given node (Watts and Strogatz, 1998).

Accessibility *($h = \{2, 3\}$)*: The accessibility of a node i is based on Shannon's entropy (Shannon and Weaver, 1963) of the probability of accessing nodes at the h^{th} concentric level, centered at i, by a given dynamics starting at that node (Travençolo and Costa, 2008). Here, we adopted the self-avoiding random walk as the reference dynamics.

Symmetry *($h = \{2, 3, 4\}$)*: This measurement (Silva et al., 2016b), obtained for each node i, quantifies the symmetry of the topology around i. It can be understood as a normalization of the accessibility, and includes two components: *backbone*, where edges between nodes from the same concentric level are discarded, and *merged*, where nodes that share edges in the same level are merged.

Network visualization can provide means to better understand the structure of a given book's story by organizing, into an embedding space, the topology of the obtained network. We applied a visualization methodology based on force-directed graph drawing (Silva et al., 2016a). Specifically, this method is based on the Fruchterman and Reingold (1991) (FR) algorithm, which simulates a system of particles, which attract and repel one another. The attractive force, f_a, reflects the node connectivity, while the repulsive force, f_r, acts between all pair of nodes. A gravitational force, f_g, can also be added. We adopted $f_a = 0.0002$, $f_r = 1.25$, and $f_g = 0.001$.

3.4 Machine Learning Methods

Several classifiers — Decision Trees, Random Forest, kNN, Logistic Regressors, SVM, Naive Bayes (Duda et al., 2000) — were tested in order to choose the most adequate. Support Vector Machines (SVM) and Random Forest were selected. We used the Linear SVM implementation (with default parameters), and Random Forest with 50 trees, both available at *Scikit-learn* (Pedregosa et al., 2011). We employed the *leave-one-out* cross-validation technique, in which only one dataset instance is used as test while all the others are taken for training the classifier. Feature selection was attempted, but no particular subset of features stood out. Therefore, all measurements were considered.

4 Results and Discussion

In this section, we describe the selected dataset and present the obtained results organized in two parts: (i) the complete set of authors; and (ii) four authors representing major types of works.

4.1 Dataset

In order to investigate whether authors can be distinguished by the story flow in their works, we created mesoscopic networks from several texts. Our dataset is composed of 100 English texts written by 20 distinct authors (five texts per author) extracted from Machicao et al. (2016). The selected 20 authors are: Andrew Lang, Arthur Conan Doyle, B. M. Bower, Bram Stoker, Charles Darwin, Charles Dickens, Edgar Allan Poe, H. G. Wells, Hector H. Munro (Saki), Henry James, Herman Melville, Horatio Alger, Jane Austen, Mark Twain, Nathaniel Hawthorne, P. G. Wodehouse, Richard Harding Davis, Thomas Hardy, Washington Irving, and Zane Grey. The whole dataset was obtained from the Project Gutenberg repository[2]. The complete list of used texts is available at this link [3].

4.2 Complete Set of Authors

In the first experiment, we used all the books by all 20 authors, yielding the results presented in Ta-

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

[3] `https://goo.gl/2pJJHG`

ble 1. Remarkably, though the chance baseline for this experiment is only 5% (each author has the same probability of being randomly selected), our best result was as high as 35%. Moreover, 17 (48.5%) out of the 35 books correctly classified by our method were written by only 4 authors: namely Andrew Lang, B. M. Bower, Hector H. Munro (Saki), and Henry James

Table 1: Accuracy rate in discriminating the authorship of texts.

Average Degree	**Random Forest**	**SVM**
$\langle k \rangle = 5$	10%	12%
$\langle k \rangle = 10$	18%	14%
$\langle k \rangle = 15$	22%	25%
$\langle k \rangle = 20$	25%	24%
$\langle k \rangle = 25$	21%	17%
$\langle k \rangle = 30$	21%	23%
$\langle k \rangle = 35$	16%	17%
$\langle k \rangle = 40$	16%	23%
$\langle k \rangle = 45$	18%	25%
$\langle k \rangle = 50$	16%	20%
All combined	26%	**35%**

We also performed a pairwise classification. The obtained results were compared with a traditional approach usually employed in the literature, the analysis of the most frequent words. For this experiment, we used the original texts of each book, extracted the frequency of the 20 most frequent words, and then used a SVM classifier. Figure 2 shows the accuracies for the traditional features, and Figure 3 illustrates the pairwise classification accuracies when mesoscopic networks were used to model each text, we did not select a single average degree $\langle k \rangle$, but rather we combined all the degrees listed in Table 1. The accuracies were obtained with the SVM classifier.

A careful examination of Figure 2 and 3 reveals that for some cases, except the squares with lighter colors, our results are on par with those obtained with the frequency of the 20 most frequent words (mainly stopwords). Moreover, our method even achieved higher accuracies in some combinations. See, for example, authors Grey and Munro, for which 7 and 6, respectively, of our results were better than the traditional approach. One thing that we should note, and which will be revisited in the following subsection, is the fact that it is hard for mesoscopic networks to distinguish Edgar Allan Poe from Charles Darwin. In this case, we obtained an accuracy rate of 50%, contrasted to 80% achieved by the other approach.

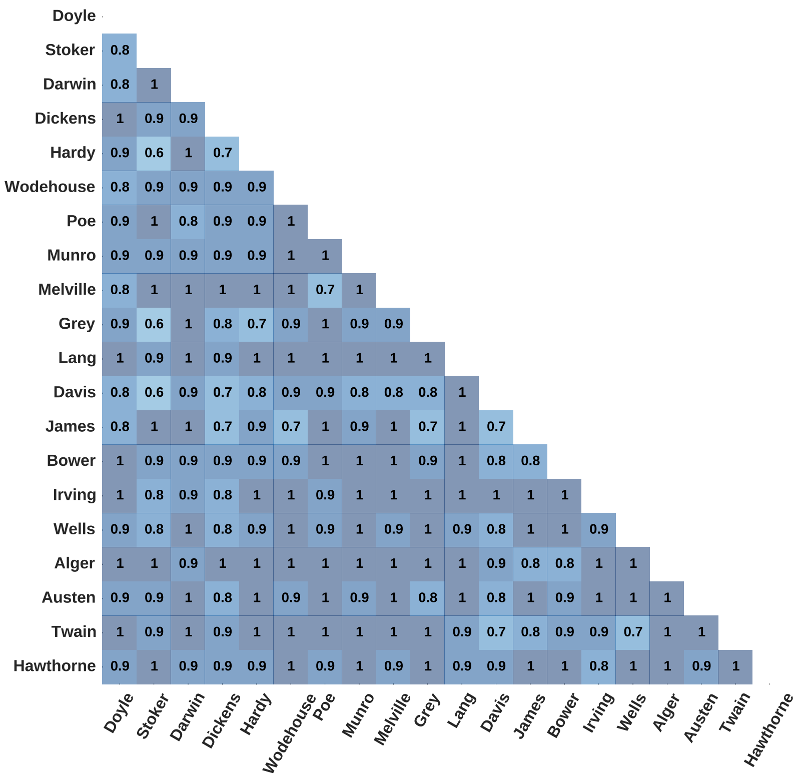
4.3 Small Set of Authors

Out of the 20 authors considered in the previous subsection, we selected four authors, namely Charles Darwin, Thomas Hardy, Edgar Allan Poe, and Mark Twain. They were chosen because two of them have several *novels* (Thomas Hardy and Mark Twain), Edgar Allan Poe is best known for writing *short stories* and Charles Darwin wrote about his *scientific theories* and observations. The now obtained accuracy rate in classifying them was enhanced to 65% (Random Forests) and 50% (SVM) by using the mesoscopic representation, contrasted to the chance baseline of 25% obtained for four authors. The Principal Component Analysis (PCA) (Jolliffe, 2002) considering these four authors is presented in Figure 4.

The PCA results indicate a clear partitioning between the groups of books associated to each author. Remarkably, one of Thomas Hardy's book (*A Changed Man and Other Tales*) resulted between those of Edgar Allan Poe and Charles Darwin. Such a good partitioning is a consequence of the quite different mesoscopic networks obtained for these authors, as depicted in Figure 5.

The mesoscopic networks presented in Figure 5 unveil interesting aspects, including an unexpected similarity to intricate calligraphic shapes. Note that the books which contain tales or short stories, such as those by Edgar Allan Poe, as well as the book *A Changed Man and Other Tales*, present a similar chain-like topology with a few cycles. Moreover, most of these cycles appear at a relatively small scale. Interestingly, the scientific books of Charles Darwin also present this chain-like structure, which is probably related to the nature of his writings, describing his theories, observations, and findings.

It is clear, visually, that the other books present more complex stories, where paragraphs (nodes) from different parts of the book sharing similar content resulted in intersections. For example, the book *Adventures of Huckleberry Finn* tells the story of Huckleberry Finn traveling down the Mississippi river. During most of the book, he goes through different small adventures along the river. Another interesting point is that this book ends in a similar setting as it begins, when Huckleberry Finn returns to his city, which is reflected in the



	Doyle	Stoker	Darwin	Dickens	Hardy	Wodehouse	Poe	Munro	Melville	Grey	Lang	Davis	James	Bower	Irving	Wells	Alger	Austen	Twain	Hawthorne
Doyle																				
Stoker	0.8																			
Darwin	0.8	1																		
Dickens	1	0.9	0.9																	
Hardy	0.9	0.6	1	0.7																
Wodehouse	0.8	0.9	0.9	0.9	0.9															
Poe	0.9	1	0.8	0.9	0.9	1														
Munro	0.9	0.9	0.9	0.9	0.9	1	1													
Melville	0.8	1	1	1	1	1	0.7	1												
Grey	0.9	0.6	1	0.8	0.7	0.9	1	0.9	0.9											
Lang	1	0.9	1	0.9	1	1	1	1	1	1										
Davis	0.8	0.6	0.9	0.7	0.8	0.9	0.9	0.8	0.8	0.8	1									
James	0.8	1	1	0.7	0.9	0.7	1	0.9	1	0.7	1	0.7								
Bower	1	0.9	0.9	0.9	0.9	0.9	1	1	1	0.9	1	0.8	0.8							
Irving	1	0.8	0.9	0.8	1	1	0.9	1	1	1	1	1	1	1						
Wells	0.9	0.8	1	0.8	0.9	1	0.9	1	0.9	1	0.9	0.8	1	1	0.9					
Alger	1	1	0.9	1	1	1	1	1	1	1	1	0.9	0.8	0.8	1	1				
Austen	0.9	0.9	1	0.8	1	0.9	1	0.9	1	0.8	1	0.8	1	0.9	1	1	1			
Twain	1	0.9	1	0.9	1	1	1	1	1	1	0.9	0.7	0.8	0.9	0.9	0.7	1	1		
Hawthorne	0.9	1	0.9	0.9	0.9	1	0.9	1	0.9	1	0.9	0.9	1	1	0.8	1	1	0.9	1	

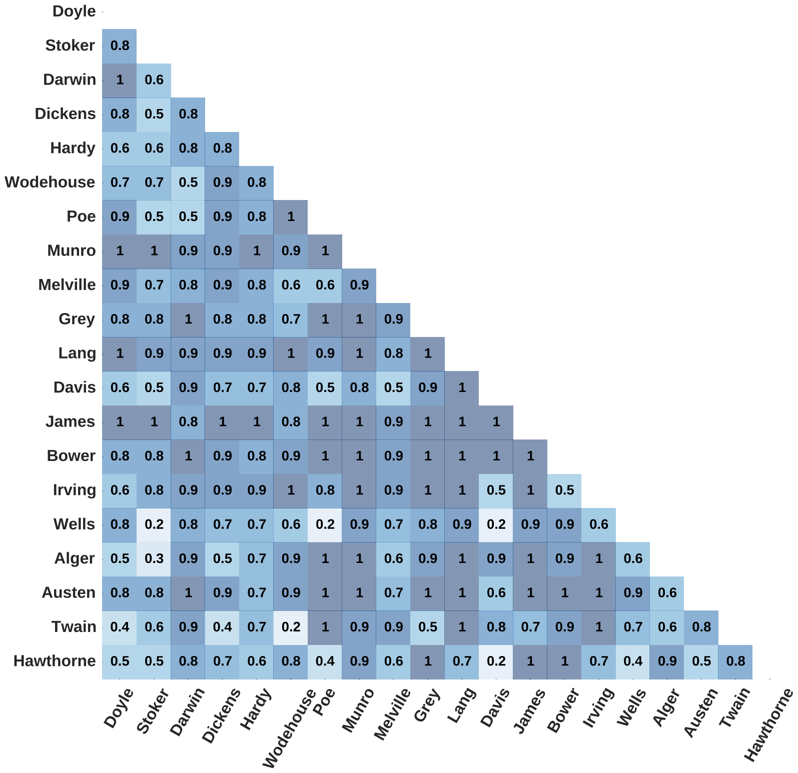
Figure 2: Accuracy rate (from 0 to 1) in the pairwise classification using the frequency of the 20 most frequent words.

respective return of the unfolding trajectory to its beginning. It is important to highlight that a full visual analysis with all the 20 authors was beyond the scope of this experiment. Our primary goal was to perform a preliminary investigation of the books through geometrical approaches.

5 Conclusion

Complex network methods have been applied with growing success to several natural language processing tasks. In some of these approaches, a chunk of text is represented as a co-occurrence network, which reflects the syntactic relationship between words (Cancho and Solé, 2001). Although this is a well-known representation, it is not without its share of problems. Those networks, for example, are unable to represent the topical structure found in many texts. So as to overcome such a limitation, a mesoscopic representation has been recently proposed (de Arruda et al., 2017). The main goal of that approach was to take into account the semantical relationship between chunks of text. More specifically, the network nodes correspond to texts from consecutive paragraphs, while the edges are weighted by the similarity between the respective texts. Statistics of some local topological measurements were used to characterize books' mesoscopic networks. We tested the hypothesis that such a representation is useful at assigning the authorship to documents. In particular, we advocated that fingerprints left by each author are visible at a mesoscopic scale.

The obtained accuracy rates, which in one case surpassed by 40 percentage points the chance baseline, suggest that the proposed approach is capable of revealing writing styles characteristics. In addition, we performed an alternative classification, in which all pairs of distinct authors were considered. In some cases our method provided better results than those obtained with traditional features. Such a result indicates that features obtained from mesoscopic networks can be used as a complement to more traditional features of texts. In order to better understand the unfolding of texts, we selected authors whose works include short stories, novels, and scientific writing. A set of topological features was estimated and PCA projected. Interestingly, in this projected space, a



	Doyle	Stoker	Darwin	Dickens	Hardy	Wodehouse	Poe	Munro	Melville	Grey	Lang	Davis	James	Bower	Irving	Wells	Alger	Austen	Twain	Hawthorne
Doyle																				
Stoker	0.8																			
Darwin	1	0.6																		
Dickens	0.8	0.5	0.8																	
Hardy	0.6	0.6	0.8	0.8																
Wodehouse	0.7	0.7	0.5	0.9	0.8															
Poe	0.9	0.5	0.5	0.9	0.8	1														
Munro	1	1	0.9	0.9	1	0.9	1													
Melville	0.9	0.7	0.8	0.9	0.8	0.6	0.6	0.9												
Grey	0.8	0.8	1	0.8	0.8	0.7	1	1	0.9											
Lang	1	0.9	0.9	0.9	0.9	1	0.9	1	0.8	1										
Davis	0.6	0.5	0.9	0.7	0.7	0.8	0.5	0.8	0.5	0.9	1									
James	1	1	0.8	1	1	0.8	1	1	0.9	1	1	1								
Bower	0.8	0.8	1	0.9	0.8	0.9	1	1	0.9	1	1	1	1							
Irving	0.6	0.8	0.9	0.9	0.9	1	0.8	1	0.9	1	1	0.5	1	0.5						
Wells	0.8	0.2	0.8	0.7	0.7	0.6	0.2	0.9	0.7	0.8	0.9	0.2	0.9	0.9	0.6					
Alger	0.5	0.3	0.9	0.5	0.7	0.9	1	1	0.6	0.9	1	0.9	1	0.9	1	0.6				
Austen	0.8	0.8	1	0.9	0.7	0.9	1	1	0.7	1	1	0.6	1	1	1	0.9	0.6			
Twain	0.4	0.6	0.9	0.4	0.7	0.2	1	0.9	0.9	0.5	1	0.8	0.7	0.9	1	0.7	0.6	0.8		
Hawthorne	0.5	0.5	0.8	0.7	0.6	0.8	0.4	0.9	0.6	1	0.7	0.2	1	1	0.7	0.4	0.9	0.5	0.8	

Figure 3: Accuracy rate (from 0 to 1) in the pairwise classification using network features extracted from mesoscopic networks.

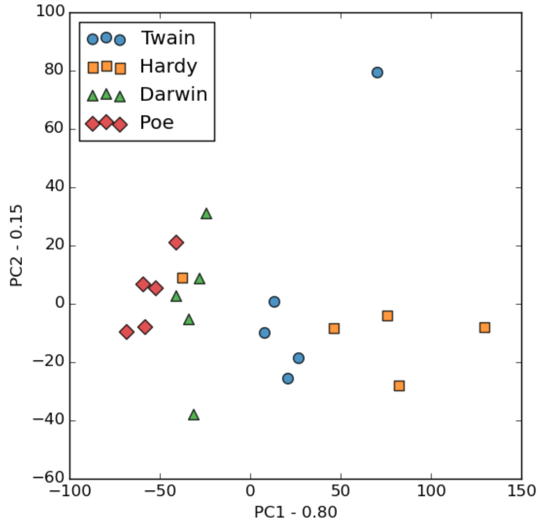


Figure 4: PCA of the books written by Charles Darwin, Thomas Hardy, Edgar Allan Poe, and Mark Twain.

book of tales written by *Thomas Hardy* resulted closer to *Edgar Allan Poe*'s books, which are also composed of short stories. Even more surprising, the patterns obtained by the visualization resulted quite representative of the different types of works, suggesting a "calligraphy". Such visualizations reveal intricate discourse patterns in the books.

The goal of this paper was not to provide state-of-the-art results for authorship attribution, given that most traditional approaches in the literature have achieved results as high as 90% (Grieve, 2007; Koppel et al., 2009). Instead, we report an approach that can be used to obtain novel stylometric features, as well as to complement traditional methods.

Future works could apply a similar approach to other related tasks — such as authorship verification, plagiarism detection, and topic segmentation — and also extend the mesoscopic representation to include different granularity levels, such as sentences or chapters. Another possibility is to investigate the relationship between the emotional content of a text and its topology.

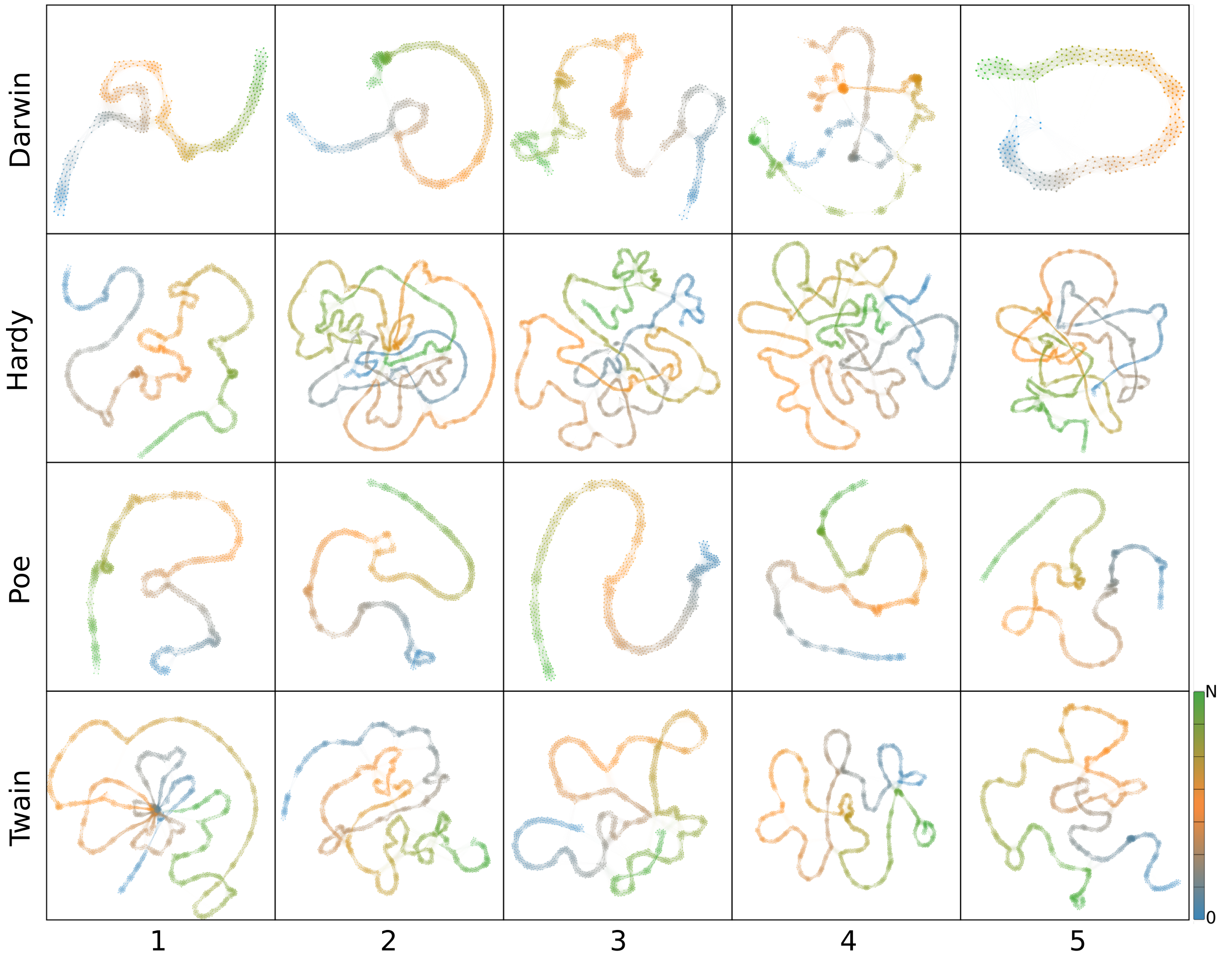


Figure 5: Mesoscopic networks for 20 books of four different authors. **Charles Darwin:** (1) *Coral Reefs*, (2) *The Expression of the Emotions in Man and Animals*, (3) *Geological Observations on South America*, (4) *The Different Forms of Flowers on Plants of the Same Species*, and (5) *Volcanic Islands*. **Thomas Hardy:** (1) *A Changed Man; and Other Tales*, (2) *A Pair of Blue Eyes*, (3) *Far from the Madding Crowd*, (4) *Jude the Obscure*, and (5) *The Hand of Ethelberta*. **Edgar Allan Poe:** *The Works of Edgar Allan Poe - Volume (1) to (5)*. **Mark Twain:** (1) *Adventures of Huckleberry Finn*, (2) *The Adventures of Tom Sawyer*, (3) *The Prince and the Pauper*, (4) *A Connecticut Yankee in King Arthur's Court*, and (5) *Roughing It*. The bluish nodes represent the windows formed by paragraphs from the beginning of the book and the greenish ones represent the windows formed by paragraphs from the end of the book. The order of the windows can be seen in the legend, where N represents the last window.

Acknowledgments

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

References

C. Akimushkin, D. R. Amancio, and O. N. Oliveira Jr. 2017. Text authorship identified using the dynamics of word co-occurrence networks. *PLoS ONE* .

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

L. Antiqueira, T. A. S. Pardo, M. G. V. Nunes, O. N. Oliveira Jr, and L. F. Costa. 2006. Some issues on complex networks for author characterization. In *Fourth Workshop in Information and Human Language Technology in the Proceedings of International Joint Conference IBERAMIA-SBIA-SBRN*. ICMC-USP, Ribeirão Preto, Brazil.

S. Argamon, C. Whitelaw, P. Chase, S. R. Hota, N. Garg, and S. Levitan. 2007. Stylistic text classification using functional lexical features. *Journal of the American Society for Information Science and Technology* 58(6):802–822.

H. Baayen, H. van Halteren, and F. Tweedie. 1996. Outside the cave of shadows: using syntactic annotation to enhance authorship attribution. *Literary and Linguistic Computing* 11(3):121.

R. F. i Cancho and R. V. Solé. 2001. The small world of human language. *Proceedings of The Royal Society of London. Series B, Biological Sciences* 268:2261–2266.

J. Cong and H. Liu. 2014. Approaching human language with complex networks. *Physics of life reviews* 11(4):598–618.

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

G. Csardi and T. Nepusz. 2006. The igraph software package for complex network research. *InterJournal* Complex Systems:1695.

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

H. F. de Arruda, F. N. Silva, V. Q. Marinho, D. R. Amancio, and L. da F. Costa. 2017. Representation of texts as complex networks: a mesoscopic approach. *arXiv preprint arXiv:1606.09636v2* .

R. O. Duda, P. E. Hart, and D. G. Stork. 2000. *Pattern Classification (2nd Edition).* Wiley-Interscience.

T. M. J. Fruchterman and E. M. Reingold. 1991. Graph drawing by force-directed placement. *Software: Practice and experience* 21(11):1129–1164.

M. Gamon. 2004. Linguistic correlates of style: authorship classification with deep linguistic analysis features. In *Proceedings of International Conference on Computational Linguistics (COLING).* Geneva, Switzerland, pages 611–617.

J. Grieve. 2007. Quantitative authorship attribution: An evaluation of techniques. *Literary and Linguistic Computing* 22(3):251.

G. Hirst and O. Feiguina. 2007. Bigrams of syntactic labels for authorship discrimination of short texts. *Literary and Linguistic Computing* 22(4):405–417.

I. Jolliffe. 2002. *Principal component analysis*. Wiley Online Library.

P. Juola. 2006. Authorship attribution. *Foundations and Trends in Information Retrieval* 1(3):233–334.

M. Koppel, J. Schler, and S. Argamon. 2009. Computational methods in authorship attribution. *Journal of the American Society for Information Science and Technology.* 60(1):9–26.

S. Lahiri and R. Mihalcea. 2013. Authorship attribution using word network features. *arXiv preprint arXiv:1311.2978* .

J. Machicao, E. A. Correa Jr., G. H. B. Miranda, D. R. Amancio, and O. M. Bruno. 2016. Authorship attribution based on life-like network automata. *arXiv preprint arXiv:1610.06498* .

C. D. Manning and H. Schütze. 1999. *Foundations of Statistical Natural Language Processing*. MIT Press, Cambridge, MA, USA.

V. Q. Marinho, G. Hirst, and D. R. Amancio. 2016. Authorship attribution via network motifs identification. In *Proceedings of the 5th Brazilian Conference on Intelligent Systems (BRACIS).* Recife, Brazil.

R. Mihalcea and D. Radev. 2011. *Graph-based natural language processing and information retrieval.* Cambridge University Press, Cambridge; New York.

R. Milo, S. Shen-Orr, S. Itzkovitz, N. Kashtan, D. Chklovskii, and U. Alon. 2002. Network motifs: simple building blocks of complex networks. *Science* 298(5594):824–827.

F. Mosteller and D. L. Wallace. 1964. *Inference and Disputed Authorship: The Federalist Papers.* Addison-Wesley, Reading, Mass.

M. Newman. 2003. Mixing patterns in networks. *Physical Review E* 67(2):026126.

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

R. Pastor-Satorras, A. Vázquez, and A. Vespignani. 2001. Dynamical and correlation properties of the Internet. *Physical Review Letters* 87(25):258701.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, et al. 2011. Scikit-learn: Machine learning in python. *Journal of Machine Learning Research.* 12(Oct):2825–2830.

S. Segarra, M. Eisen, and A. Ribeiro. 2013. Authorship attribution using function words adjacency networks. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, pages 5563–5567.

C. E. Shannon and W. Weaver. 1963. *A Mathematical Theory of Communication*. University of Illinois Press, Champaign, IL, USA.

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

F. N. Silva, C. H. Comin, T. K. DM. Peron, F. A. Rodrigues, C. Ye, R. C. Wilson, E. R. Hancock, and L. da F. Costa. 2016b. Concentric network symmetry. *Information Sciences* 333:61–80.

E. Stamatatos. 2009. A survey of modern authorship attribution methods. *Journal of the American Society for Information Science and Technology.* 60(3):538–556.

B. A. N. Travençolo and L. da F. Costa. 2008. Accessibility in complex networks. *Physics Letters A* 373(1):89–95.

D. J. Watts and S. H. Strogatz. 1998. Collective dynamics of small-worldnetworks. *Nature* 393(6684):440–442.