Journal of Complex Networks (2018) **6**, 620–638
doi: 10.1093/comnet/cnx047
Advance Access Publication on 10 October 2017

Labelled network subgraphs reveal stylistic subtleties in written texts

VANESSA QUEIROZ MARINHO
Institute of Mathematics and Computer Science, University of São Paulo, Avenida Trabalhador Sancarlense, 400 - Centro, São Carlos - SP, 13566-590, Brazil

GRAEME HIRST
Department of Computer Science, University of Toronto, Toronto, Ontario, Canada M5S3G4

AND

DIEGO RAPHAEL AMANCIO[†]
Institute of Mathematics and Computer Science, University of São Paulo, Avenida Trabalhador Sancarlense, 400 - Centro, São Carlos - SP, 13566-590, Brazil
[†]Corresponding author. Email: diegoraphael@gmail.com

Edited by: Edwin Hancock

[Received on 6 May 2017; editorial decision on 5 September 2017; accepted on 7 September 2017]

The vast amount of data and increase of computational capacity have allowed the analysis of texts from several perspectives, including the representation of texts as complex networks. Nodes of the network represent the words, and edges represent some relationship, usually word co-occurrence. Even though networked representations have been applied to study some tasks, such approaches are not usually combined with traditional models relying upon statistical paradigms. Because networked models are able to grasp textual patterns, we devised a hybrid classifier, called *labelled subgraphs*, that combines the frequency of common words with small structures found in the topology of the network. Our approach is illustrated in two contexts, authorship attribution and translationese identification. In the former, a set of novels written by different authors is analysed. To identify translationese, texts from the Canadian Hansard and the European Parliament were classified as to original and translated instances. Our results suggest that labelled subgraphs are able to represent texts and it should be further explored in other tasks, such as the analysis of text complexity, language proficiency and machine translation.

Keywords: complex networks; motifs; natural language processing; authorship attribution; translationese; labelled subgraphs.

1. Introduction

The advent of Internet has allowed immediate access to an enormous amount of texts. The need to process and analyse these texts, in the form of emails, blog posts, tweets and news, has fostered the development of methods in a variety of natural language processing (NLP) tasks, such as automatic summarization, authorship attribution, machine translation, sentiment analysis and others. Commonly used in some methods, word frequency is a simple, yet useful attribute employed to address some of these tasks [1, 2]. In many contexts, however, the use of this attribute alone has not led to optimized results. Even when frequency attributes yield good performance, the robustness of classification systems might be undermined [3]. In the case of the authorship attribution task, for example, several works have

reported excellent results when word frequency and other simple features are taken into account [2, 4, 5]. However, recent works have shown that such features are prone to manipulation, as simple word statistics patterns can be easily mimicked by authors trying to conceal their identities [3]. This drawback to the use of simple frequency counts in some NLP applications paves the way for the exploration of novel informative textual features, so as to provide both performance and robustness to the problems addressed. In this scenario, some network approaches have been proposed to analyse texts using a topological point of view [6, 7].

In recent years, network theory has drawn the attention of a myriad of scientists from distinct research areas [8–11]. Of particular interest to the aims of this article, networks have also been applied as a complementary tool in text analysis [7, 12, 13]. A well-known model, the co-occurrence (or word adjacency) representation has been extensively used in the study of text complexity, machine translations, stylometry and disease diagnosis [7, 13, 14]. In this model, words are modelled as nodes in the network while the edges may represent syntactic [15], semantic [12], or empirical [16] relationships. The complementary role played by co-occurrence networks in text analysis stems from their ability in considering both meso- and large-scale structure of texts, a feature markedly overlooked by bag-of-word models [7]. The structure of a text is typically analysed in terms of topological measurements [9, 17], with reinterpretations in the context of text analysis [18].

While much study has been devoted to create text analysis techniques based either on statistical or networked representation and characterization, only a few works have probed the benefits of combining such distinct paradigms. For this reason, the main goal of this article is to combine networked representations with the frequency of words. In order to do so, we explore the concept of *network motif* to complement the information provided by frequency statistics in text analysis. In the current study, the combination of frequency and local structure as attributes for words is accomplished by considering node labels in each distinct subgraph. To illustrate the effectiveness of the proposed method in text analysis, we tackle the problems of identifying the authorship of texts, known as authorship attribution, and the identification of translationese. In the latter, the goal is to distinguish content originally produced in a language from content translated into that language. As we shall show, our approach is able to represent texts in a more adequate and accurate manner. This is particularly clear when we compare the performance of the traditional approaches (based either on network or textual features) with the performance of the proposed technique, which is based on the combination of both textual and network features.

This article is organized as follows: Section 2 briefly describes related work in the field of complex networks. Next, we explain our methodology in Section 3. A case study and the results of our hybrid classifier in the contexts of authorship attribution and translationese identification are presented in Section 4. Finally, Section 5 presents a summary of our article and the perspectives for further studies.

2. Related work

Methods and concepts from complex networks have been successfully applied to analyse written texts. In several studies, texts are modelled as co-occurrence (or word adjacency) networks, where nodes represent distinct words and edges connect adjacent words. Co-occurrence networks have already been used to identify the authorship of texts [19–22, 24], to distinguish prose from poetry [25] and to discriminate informative and imaginative documents [26]. Moreover, the structure of these networks has also proven useful to discriminate word senses [27]. After modelling texts as co-occurrence networks, most of the approaches extract several network measurements in order to characterize the topology of the networks [17]. While most of these measurements are able to provide significant performance of the studied task, in most of the studies the textual context is disregarded after the network is obtained—i.e.

semantic elements are not fully considered in the analysis. Because node labels (i.e. concepts associated with each node) may also play a complementary role in the analysis, the study of strategies for combining structure and semantics becomes relevant. While the combination of distinct sources of information in classification problems has been greatly investigated by the pattern recognition community for several years, such methods do not consider the particularities of each complex system under analysis. Here, we take the view that textual structure and semantics can be easily combined via extraction of small subgraphs.

In network theory, recurrent subgraphs (or *motifs*) are used in a large number of applications [22, 23, 28–34]. Usually, motifs are those whose frequency is larger than the expected (in a null model). These motifs are responsible for particular functions in biological and social networks [34–36]. In textual networks, motifs have also been employed to extract relevant information. Milo *et al.* [34] analysed texts written in four different languages, namely English, French, Japanese and Spanish. They observed that their respective word adjacency networks have similar motif sets. In a similar approach, Cabatbat *et al.* [31] compared co-occurrence networks based on translations of the Bible and the Universal Declaration of Human Rights. They found that the frequency distribution of motifs is preserved across translations. El-Fiqi *et al.* [32] used motifs to detect and identify the translator for the meanings of the Holy Qur'an. Their proposed method identified the corresponding translators of the texts with an accuracy of 70%. Biemann *et al.* [30] extracted the frequency of directed and undirected motifs from texts in six languages to successfully distinguish human-generated texts from those generated with n-gram models. Amancio *et al.* [28] analysed the connectivity patterns in textual networks and found that the frequency distribution of motifs in real texts is uneven. According to their results, some motifs rarely occur in natural language texts. Marinho *et al.* [22] extracted the frequency of all directed subgraphs comprising three nodes to reveal the authorship of several books. In their experiments, the authorship was correctly assigned for almost 60% of the books using only a small set of attributes.

While the characterization by network motifs has already been used in the context of text analysis, there is no systematic evaluation of the benefits in considering node labels in such structures. For this reason, our study focuses on devising strategies to combine structure and semantics in an effective way.

3. Methodology

In this section, we describe the creation of networks from raw texts. We also detail the proposed approach to characterize texts in terms of their semantics and structure.

3.1 *From texts to networks*

There are some pre-processing steps that can be performed prior to the creation of the co-occurrence networks, such as the removal of punctuation marks, the lemmatization of words and the removal of function words. In this article, we lower-cased the words and removed numbers and punctuation marks from the texts. Table 1 presents an extract before and after the pre-processing steps.

A word co-occurrence network can be represented by a directed graph $G = (V, A)$, where V and A are the set of nodes and edges, respectively. Each node $v_i \in V$ denotes a word from the vocabulary of the pre-processed text. Two vertices are connected by an arc $a \in A$ if the corresponding words are adjacent in at least one sentence. The direction of an arc is from the first to the following word. Here, we disregarded sentence and paragraph boundaries while determining the adjacent words. Therefore, the last word of a sentence or paragraph is connected to the first word of the next sentence or paragraph. Figure 1 presents the co-occurrence network obtained from the sentences in Table 1.

TABLE 1 *Example of an extract after the pre-processing steps. The sentences are from the book* Hard Times *written by Charles Dickens*

Original extract	NOW, what I want is, Facts. Teach these boys and girls nothing but Facts.
Pre-processed extract	now what i want is facts teach these boys and girls nothing but facts

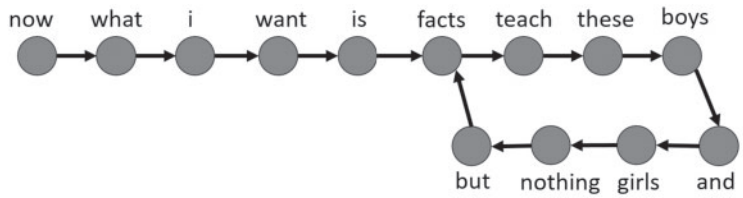

FIG. 1. Co-occurrence network from the pre-processed extract presented in Table 1.

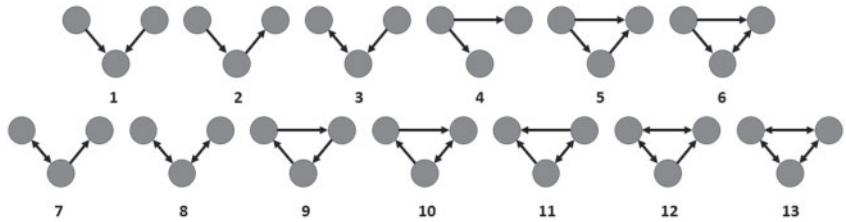

FIG. 2. All 13 possible directed motifs involving three nodes.

3.2 *Characterization via labelled subgraphs*

The topology of a complex network can be characterized by several metrics. One of these metrics is the frequency of directed motifs involving a few nodes. The set of directed motifs with three nodes is shown in Fig. 2. This type of representation has been used in several characterizations of complex systems [37, 38]. Because we are interested in analysing texts (i.e. networks with relevant information stored in each node), we introduce the concept of *labelled subgraphs* to take into account the information of the node labels in the subgraphs considered. In a typical motif analysis, the occurrences of the motifs are compared with the expected values in random networks, in order to evaluate the significance of the appearance of a given motif in the network. In this article, we did not count the occurrences of those structures in random networks, instead we decided to only calculate the frequency of all the subgraphs involving three nodes, regardless of their significance.

Labelled subgraphs are used in the strategy adopted to characterize texts by considering their frequency of appearance in the respective text networks. Instead of considering the frequency of subgraphs, we extracted the relative frequency of a given word w in each one of the 13 directed subgraphs displayed in Fig. 2. More specifically, the frequency ($n_{w,m}$) of a labelled subgraph that combines word w and subgraph

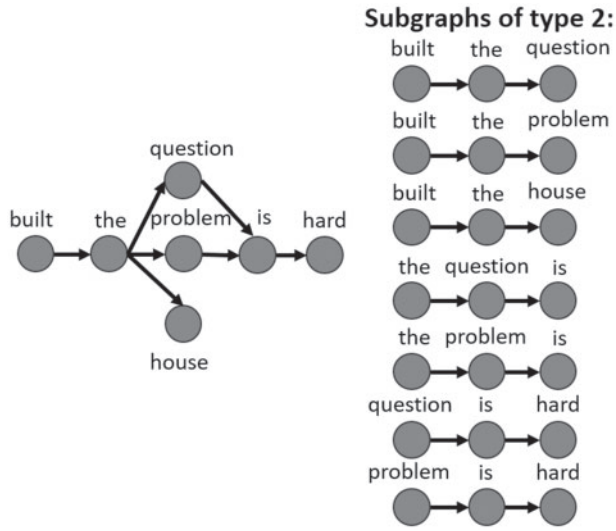

FIG. 3. An example of a co-occurrence network is presented on the left. On the right, we show all subgraphs of type 2 extracted from this network, with a total of 7, that is $n_2 = 7$.

m is given by:

$$n_{w,m} = \frac{\tilde{n}_{w,m}}{n_m}, \tag{3.1}$$

where $\tilde{n}_{w,m}$ is the total number of occurrences of word w in subgraph m and n_m is the total number of occurrences of subgraph m, irrespective of any node labels. Here, we considered w as being a word from the set of the most frequent words W from the training data set. This is the first version (V1) of the proposed method. We selected the most frequent words because they are usually useful to characterize writing styles [28, 39].

In the version V1, a word may appear in any of the three nodes forming a subgraph. The number of features in version V1 is equal to $|W| * 13$. To take into consideration the possibility of a word appearing in different nodes of the same subgraph, we also considered the word position inside the subgraph according to the different configurations of nodes—this is referred to as second version (V2). Note that, in this version, some subgraphs may have equivalent nodes: this is the case of border nodes in subgraph type 1 and all nodes in subgraph type 9). In these symmetrical cases, we considered only one configuration, in order to avoid duplicated features. As a consequence, the total number of features in version V2 is equal to $|W| * 30$. Examples of possible features for the toy network depicted in Fig. 3 are described below:

- V1: The frequency of word *the* in Subgraph 2. For example, if we extract labelled subgraphs from the network presented in Fig. 3, the word *the* is one of the nodes in 5 occurrences of Subgraph type 2, that is $\tilde{n}_{\text{'the',2}} = 5$. Subgraph 2 occurs 7 times ($n_2 = 7$), therefore, the frequency of word *the* in Subgraph 2 is $n_{\text{'the',2}} = 5/7$.

- V2: The frequency of word *the* as the central node in Subgraph 2. In Fig. 3, the word *the* appears three times as the central node in Subgraph type 2, that is $\tilde{n}_{\text{'the',central,2}} = 3$. Therefore, the frequency of word *the* in this node configuration in Subgraph 2 is $n_{\text{'the',central,2}} = 3/7$.

The analysis of the computational complexity of the proposed method is presented in Appendix B. We note here that even though the complexity of the proposed method is worse than traditional ones based on frequency counts, in practical occasions the input size (i.e. the text length) does not hamper the time efficiency of our approach because usually short extracts of texts are available for a typical authorship attribution scenario.

4. Results

This section describes a qualitative and quantitative analysis of our method. Initially, we present some details about the conducted experiments, which include information about the data sets and the machine learning methods. Then, in the context of the authorship attribution task, we present a case study in which *labelled subgraphs* are used to characterize novels written by the Brontë sisters, known to have very similar writing styles. We also use our method to extract features in a typical authorship attribution task. In order to identify the authorship of several novels, we used two different data sets with several books from different authors. We also employed *labelled subgraphs* in the identification of translationese. The goal of this experiment was to evaluate whether a text was originally produced in its language or it was translated into that language. In the tables presented in this section, we use the following terminology: *LSV1* and *LSV2* denote the results obtained with the versions V1 and V2 of our proposed technique, respectively. We also compare our results with the ones obtained with the frequency of the most frequent words. This is denoted as *MFW*. The number of words used in each experiment is represented by $|W|$. In addition, we present the accuracies when the relative frequencies of the 13 directed subgraphs involving three nodes were extracted and used as the only classification features, which is denoted as *SUBGRAPHS*.

4.1 *Experimental setup*

This section presents some details of the data sets used in the authorship attribution and translationese identification tasks. In addition, we present the machine learning methods and parameters used to evaluate the proposed technique.

4.1.1 *Data sets* In our experiments, two types of data sets were used. The first type includes the data sets of books for the authorship attribution experiments. For those experiments, two data sets were selected. The first one, henceforth referred to as Data set 1, is a diverse data set comprising 40 books written by 8 authors, which is described in Table A1. The second dataset, henceforth referred to as Data set 2, has only 19 books written by 9 different authors. This data set is presented in Table A2. A small subset of Data set 2—five books written by the Brontë sisters—was selected to illustrate our method in an initial qualitative analysis. The second type of data sets comprises two parallel corpora used for translationese identification, the Canadian Hansard and the European Parliament (EP). We chose them for two main reasons. First, pieces of text from the Canadian Hansard and the EP debates are tagged according to the original language. Second, these translations are generally produced following good translation standards which are reflected in their quality. This makes the task of translationese identification more challenging, providing thus another ideal scenario to probe the capabilities of the proposed methodology.

Each data set was used differently in the experiments. The books from Data set 1 were truncated to the size of the shortest novel and used individually. Then, one co-occurrence network is created for each book and the features were extracted from those networks. In this case, the classification task used only 40 instances, 5 instances (books) per author. Because Data set 2 has only 19 books and fewer books per author—for instance, Emily Brontë and Nathaniel Hawthorne have only one book each—, we decided to split each book in shorter pieces with 8,000 words each. Then, one co-occurrence network is created for each partition and the features were extracted from those networks. To avoid issues with imbalanced data, we selected approximately the same number of partitions per author. In this case, the classification task used 138 instances, an average of 15.3 instances (partitions) per author. The small subset of books used in the qualitative analysis was also split in several 8,000 words partitions.

For the Canadian Hansard, each one of the 463 collected sessions was divided into two files, one with all sentences originally produced in a target language L and the other with the sentences translated into L. One co-occurrence network was created for each file and the features were extracted from those networks. The classification task used 926 instances, equally divided into the two classes. As in Data set 2, the debates from the EP were divided in shorter pieces with 8,000 words each. Then, one co-occurrence network is created for each partition and the features were extracted from those networks. We selected approximately $5n$ partitions originally produced in the target language L and n partitions from each one of the other five source languages. Other details about these procedures are presented in the remaining of this section.

4.1.2 *Machine learning methods* In order to evaluate the performance of the proposed technique to identify stylistic subtleties in texts, we employed four machine learning algorithms to induce classifiers from a training data set. The techniques are Decision Tree, kNN, Support Vector Machines and Naive Bayes [40]. We did not change the default parameters of these methods implemented in Weka [41][1], because comparative studies have shown that the default parameters yield, in most cases, near-optimal results. We used a cross-validation technique with 10 folds, in which one-tenth of the texts are used as a test set whereas the other nine-tenths are used in the training process.

4.2 *Authorship attribution task*

Methods of authorship attribution identify the most likely author of a text whose authorship is unknown or disputed [2]. These texts could be email messages, blog posts, or literary works, such as books and poetry. The authorship attribution problem was the first NLP task to which we applied our method. Typically, the frequency of function words is informative as features for the problem, as noted in several works [2, 4, 5, 39]. However, in specific cases, these features might not perform well to distinguish very similar writing styles. This disadvantage can be overcome with our proposed technique, as we illustrate below.

In order to illustrate the ability of *labelled subgraphs* in discriminating texts with subtle differences in style, we selected a small subset of Data set 2 comprising the following books: *Agnes Grey* and *The Tenant of Wildfell Hall*, written by Anne Brontë, *Jane Eyre* and *The Professor* from Charlotte Brontë and *Wuthering Heights* from Emily Brontë. According to [1], all three sisters are very hard to distinguish. In

[1] The selected algorithms and default parameters are: J48 (with confidence factor equals to 0.25 and minimum number of instances per leaf equals to 2), IBk (with number of neighbours equals to 1 and window size equals to 0), SMO (with a polynomial kernel and the complexity parameter C equals to 1) and NaiveBayes.

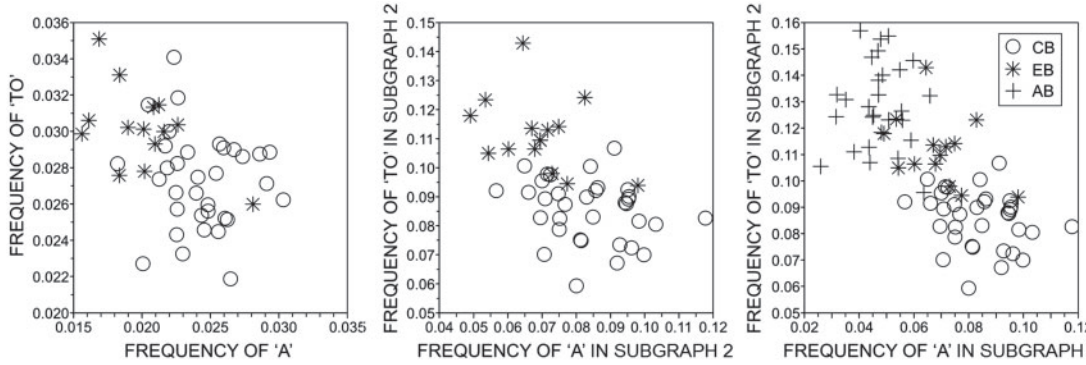

FIG. 4. Two different feature sets are extracted from 76 partitions from books of Anne Brontë (AB), Charlotte Brontë (CB) and Emily Brontë (EB). In (a), the partitions from books of Charlotte and Emily Brontë are characterized by the frequency of two words, *a* and *to*. In (b), the same data is visualized according to the frequency of word *a* and *to* in Subgraph 2, that is $n_{\text{'a',2}}$ and $n_{\text{'to',2}}$. Finally, the partitions from Anne Brontë are added in (c).

addition, by considering books written by the three sisters, we guarantee that all authors share the same period of time, gender and similar education when we compare them. Therefore, the differences among the Brontë sisters arise from their individual writing styles [42, 43].

In our analysis, each one of the considered books was split in non-overlapping partitions comprising 8,000 words each, with a total of 76 instances. For this application, we only removed punctuation marks; we did not employ any other pre-processing step. One co-occurrence network was created for each one of the 76 partitions. For simplicity's sake, we illustrate the potential of the proposed technique by considering just two words in this example. The frequency of the words *a* and *to* was extracted from each partition and these values are presented in Fig. 4(a). The results show that there is a large overlapping region between Emily Brontë (represented by stars) and Charlotte Brontë (represented by circles) for the considered features. However, if one considers also the frequency of both words in Subgraph 2 (as described in Section 3.2), a much better discrimination can be obtained, as shown in Fig. 4(b). This result confirms the suitability of the labelled subgraphs in discriminating texts, as such a good discrimination could not be obtained if only the frequency of two words were considered. The use of such features also allows a clear distinction between Anne and Charlotte Brontë, though the use of these two attributes is not enough to discriminate Anne from Emily Brontë (see Fig. 4(c)).

In a typical authorship attribution task, the objective is to identify the author of an unknown document. For this aim, a set of documents is used to train supervised classifiers. To address this task, we firstly considered Data set 1, which is described in Table A1. For the analysis, each book was truncated to the size of the shortest novel. In other words, the texts were truncated by keeping the first 46,025 words. Note that the truncation procedure is necessary because there might exist non-trivial dependencies on the text size. The truncation has also been considered in similar analyses [18, 27]. We created one co-occurrence network for each truncated book. We considered the set W of the most frequent words. Therefore, the set of features consists of all combinations of words in W appearing in one node of the subgraphs considered. For comparison purposes, we also calculated the classification accuracies when the frequencies of the W most frequent words were used as features. This frequency is calculated as the number of occurrences of each word divided by the number of tokens. Thus, it becomes possible to quantify the information gain provided by the inclusion of subgraphs in the traditional analysis based solely on frequency. Moreover,

TABLE 2 *Accuracy rate (%) and standard error of the mean in discriminating the authorship of books in Data set* 1. *The best result obtained with the proposed technique surpasses by* 15 *percentage points the best performance obtained with traditional features based on the frequency of function words*

Methods	$\|W\|$	**No . of features**	**J48**	**kNN**	**SVM**	**Bayes**
LSV1	5	65	45.00 ± 4.74	65.00 ± 7.25	62.50 ± 3.95	30.00 ± 6.89
LSV1	10	130	37.50 ± 5.30	60.00 ± 8.06	67.50 ± 6.17	27.50 ± 5.53
LSV1	20	260	60.00 ± 6.32	65.00 ± 6.32	$\mathbf{75.00} \pm 7.91$	25.00 ± 7.07
LSV2	5	150	55.00 ± 5.92	50.00 ± 8.66	62.50 ± 6.37	22.50 ± 6.57
LSV2	10	300	47.50 ± 7.46	65.00 ± 8.80	77.50 ± 5.53	15.00 ± 5.24
LSV2	20	600	45.00 ± 7.75	60.00 ± 6.32	$\mathbf{80.00} \pm 7.75$	25.00 ± 6.12
MFW	5	5	30.00 ± 5.92	57.50 ± 7.12	22.50 ± 6.57	50.00 ± 7.91
MFW	10	10	45.00 ± 6.89	52.50 ± 8.25	27.50 ± 6.57	42.50 ± 8.70
MFW	20	20	52.50 ± 5.53	62.50 ± 6.37	$\mathbf{65.00} \pm 6.32$	45.00 ± 7.75
SUBGRAPHS	—	13	47.50 ± 5.53	50.00 ± 9.35	40.00 ± 7.25	$\mathbf{55.00} \pm 6.89$

The highest accuracy rate for each method is presented in bold.

we also compared our results with those obtained when the relative frequencies of the 13 subgraphs involving three nodes were used as the only classification features.

The average classification accuracies (considering the 10 folds) for $|W| = \{5, 10, 20\}$ in Data set 1 are presented in Table 2. In addition, we present the standard error of the mean, which indicates the reliability of the mean and it is given by the standard deviation times the inverse of the square root of the number of samples [44]. The best results were obtained with the SVM, in general. For this reason, the discussion here is focused on the results obtained by this classifier. We note that, when comparing both versions of the proposed technique for the same $|W|$, the second version yielded best results, which reinforces the importance of function words in specific nodes. The relevance of using the local structure becomes evident when we analyse the results obtained with frequency features. While the best performance of the proposed technique reached 80% of accuracy, the best performance obtained with word frequency features was only 65%. The best result obtained with only the frequency of subgraphs was even lower, 55%.

An interesting pattern arising from the results in Table 2 is the apparent steady improvement in accuracy (for the SVM at least) as the number of features ($|W|$) increases. Therefore, we may expect that larger values of $|W|$ could yield better classification accuracies, with a corresponding loss in temporal efficiency. To further probe the correlation between accuracy and the value of the parameter $|W|$, we evaluated the performance of the same authorship attribution task for $1 \leq |W| \leq 40$. The percentage of books correctly classified for each value of $|W|$ is presented in Fig. 5(a) (LSV1) and Fig. 5(b) (LSV2). Considering the best scenario for each classifier, the SVM still outperforms all other methods. However, we did not observe an improvement in the discrimination, as the SVM does not benefit much from the addition of features. Conversely, the kNN is much benefited from the inclusion of new features. This behaviour is markedly visible in the LSV2 variation, with optimized results leading to a performance similar to that obtained with the SVM. Based on these results and considering the efficiency loss associated with the inclusion of features, we used at most $|W| = 20$ in most of the remaining experiments.

The authorship attribution task was also evaluated using Data set 2, which is presented in Table A2. The goal of this second experiment was to evaluate the performance of *Labelled subgraphs* in characterizing

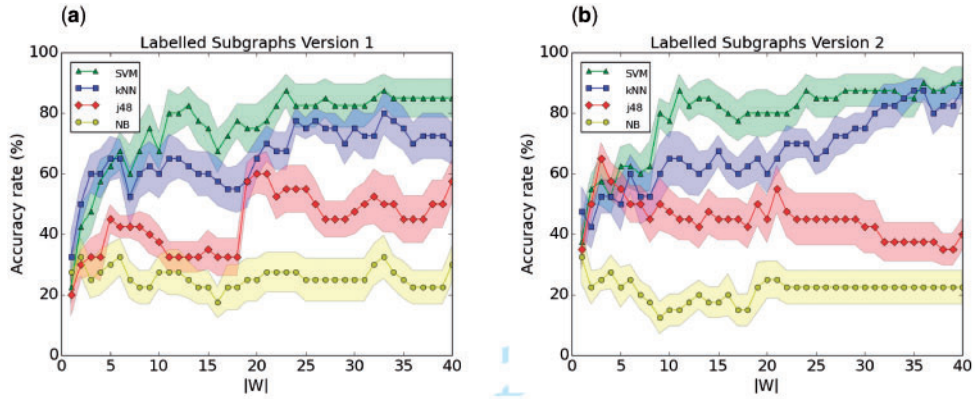

FIG. 5. Accuracy rate (%) and standard error of the mean in discriminating the authorship of books in Data set 1 for different classifiers when several values of $|W|$ are used. While some classifiers benefit from the addition of new features, the best classifier—the SVM—does not require more than 20 words in W to provide an excellent accuracy rate. (a) Labelled Subgraphs Version 1. (b) Labelled Subgraphs Version 2.

TABLE 3 *Accuracy rate (%) and standard error of the mean in discriminating the authorship of books in Data set 2. The best performance was obtained when the proposed technique* LSV2 *was used as attribute to train the SVM classifier*

Methods	$\|W\|$	**No. of features**	**J48**	**kNN**	**SVM**	**Bayes**
LSV1	5	65	58.75 ± 1.38	65.16 ± 1.61	74.30 ± 2.60	69.29 ± 1.63
LSV1	10	130	61.76 ± 1.93	83.78 ± 1.04	91.61 ± 0.93	81.31 ± 1.38
LSV1	20	260	66.84 ± 1.21	88.32 ± 1.01	**95.48** ± 0.74	78.76 ± 2.70
LSV2	5	150	62.16 ± 1.74	67.43 ± 2.03	82.07 ± 1.83	69.10 ± 2.22
LSV2	10	300	65.58 ± 1.64	80.91 ± 0.82	91.60 ± 0.87	75.30 ± 1.55
LSV2	20	600	68.16 ± 1.82	88.06 ± 0.84	**96.09** ± 0.89	77.53 ± 2.81
MFW	5	5	58.36 ± 1.01	67.80 ± 1.76	57.84 ± 2.31	73.54 ± 2.05
MFW	10	10	65.89 ± 1.11	83.66 ± 1.61	85.39 ± 1.21	83.82 ± 1.31
MFW	20	20	70.39 ± 2.04	91.12 ± 0.99	**94.47** ± 1.06	91.56 ± 1.09
SUBGRAPHS	—	13	57.60 ± 2.10	58.30 ± 1.64	**64.19** ± 2.06	63.68 ± 2.55

The highest accuracy rate for each method is presented in bold.

shorter pieces of text. In this data set, each book was split in several non-overlapping partitions with 8,000 words each. We selected the same number of partitions per author. We created one co-occurrence network for each partition and we extracted the features from those networks. The classification accuracies are presented in Table 3. The results show, for this data set, the best classifier for all studied features is the SVM. The best accuracies occurred for $|W| = 20$. Different from the results obtained in Data set 1, the gain in performance provided by the proposed technique is around 1.6 percentage points, when we compare our results with the best obtained with word frequency features. This result suggests that, in an easier authorship identification scenario, structural information plays a minor role in characterizing authors'

styles. We considered this last scenario easier for two main reasons. First and foremost, assuming that an author usually keeps his/her writing style throughout the book [45], we had several similar partitions extracted from each book. Therefore, the classifier was probably tested with instances very similar to others it has seen during the training phase. Second, we had fewer books per author (an average of 2.1); therefore there was less variance of writing styles per author.

The results obtained by the proposed technique in both authorship attribution experiments outperformed others that used only networked representations. Amancio *et al.* [18] assigned the authorship of books from a similar data set with an accuracy rate of 65%. When the frequency of all directed subgraphs with three nodes was used as attributes, Marinho *et al.* [22] achieved an accuracy of 57.5%. In a similar study, Mehri *et al.* [24] identified the authorship of several Persian books written by five authors. The authorship was correctly assigned in 77.7% of the books. Here, our results highlight the importance of function words to characterize writing styles because most (when not all) of the words from W are function words.

4.3 *Translationese*

The term *translationese* was first proposed by [46], who analysed texts originally written in Swedish and texts translated into the same language, and concluded that the main differences between them are not related to poor translation. These differences were rather an influence of the source language on the target one. Several works have been dedicated to the task of translationese identification, which consists of automatically detecting original and human-translated texts [47–54]. These methods are usually applied in a range of parallel resources, such as literary works, news articles and transcripts from parliamentary proceedings in several languages.

We start our analysis with data from the Canadian Hansard, which provides transcripts of debates from the Canadian parliament in the two official languages of the country, English and French. All debates are available online[2] in an XML format. During the debates, the members are allowed to speak either in English or French. Therefore, this is a rich parallel resource in which the original language of the sentences is indicated. Kurokawa *et al.* [51] identified translationese using the 35th to 39th Parliaments from the Canadian Hansard. They analysed the data in two granularity levels, the sentence and the block. Their blocks presented very different sizes, ranging from 3 to thousands of words. They achieved accuracies of up to 90% with word bigram frequencies.

In our experiment, we used 463 sessions from the 39th to 41st Parliaments, spanning the years 2006–2013. For the English side of the experiment, we divided each one of the 463 sessions into two files, one with all sentences originally produced in English (class *Original*) and the other with the sentences translated into English (class *Translated*). Apart from removing punctuation marks, no pre-processing step was performed in these files. We created one co-occurrence network for each file and we extracted the *labelled subgraphs* as features for the classification. We also extracted the frequency of some of the most frequent words to compare with our results. We proceeded in a similar way for the French side.

The results obtained with the Canadian Hansard are presented in Table 4. The results suggest that *labelled subgraphs* are extracting information about French to English (and vice-versa) translation and these features lead to accuracies higher than the ones obtained with only the frequency of the most frequent words and the frequencies of directed subgraphs.

[2] http://www.parl.gc.ca/

TABLE 4 *Accuracy rate (%) and standard error of the mean in discriminating the debates from the Canadian Hansard into two classes (original and translated). The highest accuracies were obtained with the strategy based on labelled subgraphs*

Language	Methods	$\|W\|$	No. of features	J48	kNN	SVM	Bayes
English	LSV1	20	260	90.61 ± 0.69	89.63 ± 0.94	97.19 ± 0.44	90.71 ± 1.09
	LSV2	20	600	90.71 ± 0.59	94.06 ± 0.82	**98.38** ± 0.17	88.45 ± 1.00
	MFW	5	5	72.24 ± 1.86	75.39 ± 1.26	57.85 ± 4.16	53.19 ± 2.65
	MFW	10	10	74.84 ± 1.84	76.05 ± 1.44	60.67 ± 4.96	53.73 ± 3.04
	MFW	20	20	78.52 ± 1.34	**80.25** ± 1.56	64.47 ± 4.64	54.36 ± 3.67
	SUBGRAPHS	—	13	88.66 ± 1.19	87.58 ± 0.53	**91.04** ± 0.90	88.55 ± 1.45
French	LSV1	20	260	94.60 ± 0.48	87.05 ± 1.02	98.38 ± 0.56	89.84 ± 0.86
	LSV2	20	600	95.57 ± 0.75	89.42 ± 0.95	**98.70** ± 0.43	89.30 ± 1.20
	MFW	5	5	70.76 ± 2.13	70.08 ± 1.59	59.30 ± 3.94	53.51 ± 2.48
	MFW	10	10	72.04 ± 2.08	72.57 ± 1.41	56.47 ± 4.10	53.83 ± 2.71
	MFW	20	20	**87.39** ± 1.42	87.26 ± 1.19	63.02 ± 5.19	55.32 ± 3.94
	SUBGRAPHS	—	13	95.14 ± 0.41	94.93 ± 0.82	**96.55** ± 0.66	91.46 ± 1.22

The highest accuracy rate for each method is presented in bold.

The ability of *labelled subgraphs* in identify original vs. translated texts was also investigated in the Europarl parallel corpus [55], which was extracted from the Proceedings of the EP. This parallel data set includes versions in more than 20 European languages. As in the Canadian Hansard, blocks of text are annotated with their original language. However, there were a few sentences with inconsistent source language tags, in which more than one language was claimed to be the source language. Those sentences were discarded in our analysis. For our purposes, we investigated translationese using four target languages (English, French, Italian and Spanish) and six source languages (English, French, Spanish, Italian, Finnish and German) from the fifth version of the corpus. Apart from removing punctuation marks, we did not employ any pre-processing step. For the English side of the experiment, we combined all sentences originally produced in English in one file (class *Original*). Then, all sentences translated into English from the other five source languages (French, German, Italian, Spanish and Finnish) were combined in one file per language (class *Translated*). These six long files were split in non-overlapping partitions with 8,000 words each. We then selected approximately n partitions from each one of the five source languages and $5n$ partitions from English, with $n = 180$. We did this because we wanted to avoid issues with imbalanced classes. We proceeded in a similar way for the other three target languages. For French, Spanish and Italian, we used $n = 128$, $n = 69$ and $n = 55$ partitions, respectively. Here, one co-occurrence network was created for each partition. The other steps are similar to the ones applied to the Canadian Hansard data set.

The results obtained with the EP are shown in Table 5. We presented the accuracies obtained with the four classifiers. Our results confirm the suitability of frequent words as relevant features for translationese identification, as described by Koppel and Ordan [50]. In a similar approach, Koppel and Ordan [50] identified translationese in 2,000 English chunks from the Europarl corpus. They achieved an accuracy of 96.7% using the frequency of 300 function words. However, they did not detect translationese with target languages other than English. Once again, we have found that the characterization by *labelled subgraphs* is extracting information about translationese. The gain in performance depends on the language being

TABLE 5 *Accuracy rate (%) and standard error of the mean in discriminating the debates from the EP into two classes (original and translated)*

Language	Methods	$\|W\|$	No. of features	J48	kNN	SVM	Bayes
English	LSV1	20	260	73.89 ± 1.07	71.17 ± 1.17	90.61 ± 0.33	81.17 ± 0.97
	LSV2	20	600	75.89 ± 1.28	71.28 ± 1.24	**92.89** ± 0.63	82.28 ± 1.10
	MFW	5	5	66.39 ± 0.59	60.89 ± 1.03	68.67 ± 1.02	67.17 ± 0.92
	MFW	10	10	69.17 ± 0.79	64.72 ± 1.05	68.83 ± 0.85	67.06 ± 0.99
	MFW	20	20	71.67 ± 0.95	73.39 ± 0.95	**78.89** ± 1.11	76.78 ± 0.73
	SUBGRAPHS	—	13	62.00 ± 1.31	59.94 ± 1.29	63.94 ± 1.03	**64.28** ± 1.02
French	LSV1	20	260	75.23 ± 1.55	69.06 ± 1.12	87.34 ± 0.71	79.22 ± 0.99
	LSV2	20	600	75.63 ± 0.93	68.05 ± 0.93	**88.13** ± 0.71	76.48 ± 1.16
	MFW	5	5	64.92 ± 1.12	59.38 ± 0.92	62.27 ± 1.92	63.83 ± 1.22
	MFW	10	10	73.28 ± 0.98	73.44 ± 0.94	78.13 ± 1.08	76.56 ± 0.86
	MFW	20	20	80.63 ± 0.95	79.45 ± 0.81	82.73 ± 0.77	**83.83** ± 0.86
	SUBGRAPHS	—	13	69.30 ± 1.61	61.48 ± 1.45	**71.02** ± 1.38	69.45 ± 1.33
Italian	LSV1	20	260	81.47 ± 1.61	78.72 ± 1.22	93.59 ± 1.15	90.92 ± 1.08
	LSV2	20	600	81.28 ± 1.68	76.16 ± 1.81	**95.99** ± 0.77	86.38 ± 1.63
	MFW	5	5	79.05 ± 1.90	78.53 ± 1.30	85.63 ± 1.26	84.38 ± 1.25
	MFW	10	10	80.91 ± 1.70	83.82 ± 1.13	90.72 ± 0.65	89.79 ± 0.86
	MFW	20	20	82.74 ± 1.99	86.49 ± 1.80	**92.71** ± 1.15	91.26 ± 1.09
	SUBGRAPHS	—	13	68.91 ± 2.00	62.36 ± 2.18	**74.91** ± 1.58	71.09 ± 2.06
Spanish	LSV1	20	260	79.65 ± 1.96	74.17 ± 1.35	91.03 ± 1.05	88.38 ± 0.78
	LSV2	20	600	78.35 ± 2.39	73.30 ± 1.26	**93.48** ± 0.93	86.19 ± 1.47
	MFW	5	5	85.32 ± 1.37	85.18 ± 1.25	88.40 ± 0.66	84.18 ± 1.46
	MFW	10	10	85.63 ± 1.50	87.67 ± 0.90	90.85 ± 0.87	89.85 ± 1.21
	MFW	20	20	86.50 ± 1.81	85.20 ± 1.23	**92.01** ± 0.92	89.42 ± 1.03
	SUBGRAPHS	—	13	66.96 ± 2.83	65.36 ± 2.01	**71.45** ± 1.81	71.01 ± 1.18

The highest accuracy rate for each method is presented in bold.

analysed: for the English, our method surpasses the best of the word frequency results by a margin of 14 percentage points. However, for the Spanish language, the gain is only 1.47 percentage points, when we consider the best of the word frequency results. In the latter case, however, note that an excellent discrimination can already be obtained with the frequency of the five most frequent words.

5. Conclusion

The enormous amount of texts available on the Web has increased the need for methods that automatically process and analyse this content. Therefore, several NLP tasks, such as authorship attribution and machine translation, have received great attention in recent years. In traditional approaches, texts are usually characterized by attributes derived from statistical properties of words (e.g. frequency, part-of-speech

tags and vocabulary size) [2] and characters (e.g. frequency of characters and punctuation marks) [56]. In addition, syntactic and semantic features have been used as relevant attributes [2]. More recently, interdisciplinary methodologies have also been proposed to study several aspects of texts. A well-known approach is the use of complex networks to analyse many levels of complexity of written documents. In this study, we advocated that the use of complex networks in combination with traditional features can improve the characterization of texts.

Even though the idea of adding labels to network nodes has already been used in biological networks [57], the relevance of such approach was not probed in many contexts, such as the characterization of written texts. In order to combine networked methods with traditional techniques usually employed in many NLP tasks, we proposed a hybrid method that combines the frequency of the most frequent words (mostly function words) with the occurrence of small subgraphs, called *labelled subgraphs*. By doing so, in the context of authorship and translationese identification, we could reveal stylistic subtleties in written texts that were not extracted with only the frequency of the words and only the frequency of subgraphs. In future works, we intend to analyse the robustness of the proposed method with regard to text length, as it is known that network properties may considerably vary when small networks are considered [58, 59]. We also intend to extend our method by considering network subgraphs comprising more than three nodes. Another possibility is to consider other structures particularly present in some textual networks, as paths and stars in knitted and word association networks [60], respectively.

The results obtained in this article suggest that the proposed approach could be applied in related tasks, such as the analysis of text complexity or the evaluation of proficiency in language learning. We believe these two tasks could be approached with our method because higher or lower complexities and proficiency levels may result in different word connections. Moreover, labelled subgraphs may also be used to detect the translation direction, that is given two parallel texts in different languages, which one is the original and which one was translated from the original. This information has a significant impact on statistical machine translation (SMT) systems for two main reasons. First, it has been proved that translation models trained on texts produced in the same direction of the SMT task usually perform better than the ones trained on the opposite direction [51]. Second, translated sentences are better represented by language models compiled from translated texts [61]. Therefore, it is of paramount importance to automatically find the translation direction.

Acknowledgements

V.Q.M. and D.R.A. acknowledge São Paulo Research Foundation for financial support. V.Q.M. and G.H. acknowledge The Natural Sciences and Engineering Research Council of Canada.

Funding

São Paulo Research Foundation (2014/20830-0, 2015/05676-8, 2015/23803-7 and 2016/19069-9) to V.Q.M. and D.R.A.

References

1. Koppel, M., Schler, J. & Mughaz, D. (2004) Text categorization for authorship verification. *Eighth International Symposium on Artificial Intelligence and Mathematics.*
2. Stamatatos, E. (2009) A survey of modern authorship attribution methods. *J. Amer. Soc. Inform. Sci. Technol.*, **60**, 538–556.

3. Brennan, M. R. & Greenstadt, R. (2009) Practical attacks against authorship recognition techniques. *Proceedings of the 21st Innovative Applications of Artificial Intelligence Conference*. Pasadena, California, USA: AAAI Digital Library.

4. Grieve, J. (2007) Quantitative authorship attribution: an evaluation of techniques. *Literary Linguist. Comput.*, **22**, 251.

5. Koppel, M., Schler, J. & Argamon, S. (2009) Computational methods in authorship attribution. *J. Amer. Soc. Inform. Sci. Technol.*, **60**, 9–26.

6. Amancio, D. R., Nunes, M. G. V., Oliveira Jr., O. N. & da F. Costa, L. (2012). Using complex networks concepts to assess approaches for citations in scientific papers. *Scientometrics*, **91**, 827–842.

7. Cong, J. & Liu, H. (2014). Approaching human language with complex networks. *Phys. Life. Rev.*, **11**, 598–618.

8. Ferretti, S. (2017). On the modeling of musical solos as complex networks. *Inform. Sci.*, **375**, 271–295.

9. Newman, M. (2010). *Networks: An Introduction*. New York: Oxford University Press.

10. Serrà, J., Corral, A., Boguñá, M., Haro, M. & Arcos, J. L. (2012). Measuring the evolution of contemporary western popular music. *Sci. Rep.*, **2**, 521 pages.

11. Xin, C., Zhang, H. & Huang, J. (2016). Complex network approach to classifying classical piano compositions. *Eurphys. Lett.*, **116**, 18008.

12. Liu, H. (2009). Statistical properties of Chinese semantic networks. *Chin. Sci. Bull.*, **54**, 2781–2785.

13. Mihalcea, R. & Radev, D. (2011). *Graph-Based Natural Language Processing and Information Retrieval*. Cambridge: Cambridge University Press.

14. Ferrer i Cancho, R. & Solé, R. V. (2001). The small world of human language. *Proc. Soc. London Ser. B Biol. Sci.*, **268**, 2261–2266.

15. Čech, R., Mačutek, J. & Žabokrtský, Z. (2011). The role of syntax in complex networks: local and global importance of verbs in a syntactic dependency network. *Phys. A.*, **390**, 3614–3623.

16. Ludueña, G. A., Behzad, M. D. & Gros, C. (2014). Exploration in free word association networks: models and experiment. *Cogn. Process.*, **15**, 195–200.

17. Costa, L. F., Rodrigues, F. A., Travieso, G. & Boas, P. R. V. (2007). Characterization of complex networks: a survey of measurements. *Adv. Phys.*, **56**, 167–242.

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

19. Amancio, D. R. (2015a). Authorship recognition via fluctuation analysis of network topology and word intermittency. *J. Stat. Mech. Theory Exp.*, **2015**, P03005.

20. Amancio, D. R. (2015). A complex network approach to stylometry. *PLoS One*, **10**, 1–21.

21. Lahiri, S. & Mihalcea, R. (2013). Authorship attribution using word network features. *arXiv preprint arXiv:1311.2978*.

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

23. Mesgar, M. & Strube, M. (2015) Graph-based coherence modeling for assessing readability. *Proceedings of the Fourth Joint Conference on Lexical and Computational Semantics*. Denver, Colorado: Association for Computational Linguistics, pp. 309–318.

24. Mehri, A., Darooneh, A. H. & Shariati, A. (2012). The complex networks approach for authorship attribution of books. *Phys. A.*, **391**, 2429–2437.

25. Roxas, R. M. & Tapang, G. (2010). Prose and poetry classification and boundary detection using word adjacency network analysis. *Int. J. Modern Phys. C*, **21**, 503–512.

26. de Arruda, H. F., da F. Costa, L. & Amancio, D. R. (2016). Using complex networks for text classification: discriminating informative and imaginative documents. *Europhys. Lett.*, **113**, 28007.

27. Silva, T. C. & Amancio, D. R. (2012). Word sense disambiguation via high order of learning in complex networks. *Europhys. Lett.*, **98**, 58001.

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

29. Amancio, D. R., Oliveira Jr, O. N. & Costa, L. da F. (2012) Structure-semantics interplay in complex networks and its effects on the predictability of similarity in texts. *Phys. A.*, **391**, 4406–4419.

30. Biemann, C., Roos, S. & Weihe, K. (2012). Quantifying semantics using complex network analysis. *Proceedings of the 24th International Conference on Computational Linguistics*. Mumbai: The COLING 2012 Organizing Committee, pp. 263–278.

31. Cabatbat, J. J. T., Monsanto, J. P. & Tapang, G. A. (2014). Preserved network metrics across translated texts. *Int. J. Mod. Phys. C*, **25**, 1350092.

32. El-Fiqi, H., Petraki, E. & Abbass, H. A. (2011). A computational linguistic approach for the identification of translator stylometry using Arabic-English text. *IEEE International Conference on Fuzzy Systems.* Taipei, Taiwan: IEEE, pp. 2039–2045.

33. Krumov, L., Fretter, C., Müller-Hannemann, M., Weihe, K. & Hütt, M. (2011). Motifs in co-authorship networks and their relation to the impact of scientific publications. *Eur. Phys. J. B*, **84**, 535–540.

34. Milo, R., Itzkovitz, S., Kashtan, N., Levitt, R., Shen-Orr, S., Ayzenshtat, I., Sheffer, M. & Alon, U. (2004). Superfamilies of evolved and designed networks. *Science*, **303**, 1538–1542.

35. Kashtan, N., Itzkovitz, S., Milo, R. & Alon, U. (2004). Topological generalizations of network motifs. *Phys. Rev. E*, **70**, 031909.

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

37. Campbell, C., Shea, K., Yang, S. & Albert, R. (2016). Motif profile dynamics and transient species in a boolean model of mutualistic ecological communities. *J. Complex Netw.*, **4**, 127.

38. Zenil, H., Kiani, N. A. & Tegnér, J. (2016). Quantifying loss of information in network-based dimensionality reduction techniques. *J. Complex Netw.*, **4**, 342.

39. García, A. M. & Martín, J. C. (2007). Function words in authorship attribution studies. *Literary Linguist. Comput.,* **22**, 49.

40. Duda, R. O., Hart, P. E. & Stork, D. G. (2000). *Pattern Classification*, 2nd edn. New York, NY: Wiley.

41. Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P. & Witten, I. H. (2009). The weka data mining software: an update. *SIGKDD Explor. Newsl.*, **11**, 10–18.

42. Gamon, M. (2004). Linguistic correlates of style: authorship classification with deep linguistic analysis features. *Proceedings of the 20th International Conference on Computational Linguistics*. *COLING '04*. Geneva, Switzerland: Association for Computational Linguistics.

43. Hirst, G. & Feiguina, O. (2007). Bigrams of syntactic labels for authorship discrimination of short texts. *Literary Linguist. Comput.*, **22**, 405–417.

44. Cramer, H. (1946). *Mathematical Methods of Statistics*. Uppsala, Sweden: Princeton University Press.

45. Juola, P. (2006). Authorship attribution. *Found. Trends Inf. Retr.*, **1**, 233–334.

46. Gellerstam, M. (1986). Translationese in Swedish novels translated from English. *Translation Studies in Scandinavia* L. Wollin & H. Lindquist (eds). Lund: CWK Gleerup, pp. 88–95.

47. Avner, E. A., Ordan, N. & Wintner, S. (2016). Identifying translationese at the word and sub-word level. *Digit. Scholarship Humanities*, **31**, 30–54.

48. Baroni, M. & Bernardini, S. (2006). A new approach to the study of translationese: machine-learning the difference between original and translated text. *Literary Linguist. Comput.*, **21**, 259–274.

49. Ilisei, I., Inkpen, D., Pastor, G. C. & Mitkov, R. (2010). Identification of translationese: a machine learning approach. *11th International Conference on Computational Linguistics and Intelligent Text Processing (CICLing)*, vol. 6008. Springer, pp. 503–511.

50. Koppel, M. & Ordan, N. (2011). Translationese and its dialects. *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, vol. 1. pp. 1318–1326.

51. Kurokawa, D., Goutte, C. & Isabelle, P. (2009). Automatic detection of translated text and its impact on machine translation. *Proceedings of MT Summit XII*. pp. 81–88.

52. Popescu, M. (2011). Studying translationese at the character level. *Recent Advances in Natural Language Processing.* pp. 634–639.

53. Rabinovich, E. & Wintner, S. (2015). Unsupervised identification of translationese. *Trans. Assoc. Comput. Linguist.*, **3**, 419–432.
54. van Halteren, H. (2008). Source language markers in europarl translations. *Proceedings of the 22nd International Conference on Computational Linguistics*. *COLING '08*, vol. 1. Manchester, UK: Association for Computational Linguistics, pp. 937–944.
55. Koehn, P. (2005). Europarl: a parallel corpus for statistical machine translation. *Conference Proceedings: The Tenth Machine Translation Summit*. pp. 79–86.
56. Grant, T. D. (2007). Quantifying evidence for forensic authorship analysis. *Int. J. Speech, Lang. Law*, **14**, 1–25.
57. Chen, J., Hsu, W., Lee, M. L. & Ng, S.-K. (2007). Labeling network motifs in protein interactomes for protein function prediction. *23rd International Conference on Data Engineering*. Istanbul, Turkey, pp. 546–555.
58. Amancio, D. R. (2015c) Probing the topological properties of complex networks modeling short written texts. *PLoS One*, (2):e0118394.
59. Boas, P. R. V., Rodrigues, F. A., Travieso, G. & Costa, L. da F. (2010). Sensitivity of complex networks measurements. *J. Stat. Mech. Theory Exp.*, (03):P03009.
60. Palla, G., Derényi, I., Farkas, I. & Vicsek, T. (2005). Uncovering the overlapping community structure of complex networks in nature and society. *Nature*, **435**, 814–818.
61. Lembersky, G., Ordan, N. & Wintner, S. (2012). Language models for machine translation: original vs. translated texts. *Comput. Linguist.*, **38**, 799–825.
62. Heaps, H. S. (1978). *Information Retrieval: Computational and Theoretical Aspects*. Orlando, FL: Academic Press, Inc.
63. Manning, C. D., Raghavan, P. & Schütze, H. (2008) *Introduction to Information Retrieval*. New York, NY: Cambridge University Press.

Appendix A. Data sets for authorship attribution

Table A1 *Data set 1—List of 40 books written by 8 different authors*

Author	Books
Arthur Conan Doyle	*The Adventures of Sherlock Holmes*(1892), *The Tragedy of the Korosko*(1897), *The Valley of Fear*(1914), *Through the Magic Door*(1907), *Uncle Bernac - A Memory of the Empire*(1896).
Bram Stoker	*Dracula's Guest*(1914), *Lair of the White Worm*(1911), *The Jewel Of Seven Stars*(1903), *The Man*(1905), *The Mystery of the sea*(1902).
Charles Dickens	*A Tale of Two Cities*(1859), *American Notes*(1842), *Barnaby Rudge: A Tale of the Riots of Eighty*(1841), *Great Expectations*(1861), *Hard Times*(1854).
Edgar Allan Poe	*The Works of Edgar Allan Poe, Volume 1–5,*(1835).
H. H. Munro (Saki)	*Beasts and Super Beasts*(1914), *The Chronicles of Clovis*(1912), *The Toys of Peace*(1919), *When William Came*(1913), *The Unbearable Bassington*(1912).
P. G. Wodehouse	*Girl on the Boat*(1920), *My Man Jeeves*(1919), *Something New*(1915), *The Adventures of Sally*(1922), *The Clicking of Cuthbert*(1922).
Thomas Hardy	*A Pair of Blue Eyes*(1873), *Far from the Madding Crowd*(1874), *Jude the Obscure*(1895), *Mayor Casterbridge*(1886), *The Hand of Ethelberta*(1875).
William M. Thackeray	*Barry Lyndon*(1844), *The Book of Snobs*(1848), *The History of Pendennis*(1848), *The Virginians*(1859), *Vanity Fair*(1848).

TABLE A2 *Data set 2—List of 19 books written by 9 different authors*

Author	Books
Anne Brontë	*Agnes Grey*(1847),*The Tenant of Wildfell Hall*(1848)
Jane Austen	*Emma*(1815),*Mansfield Park*(1814),*Sense and Sensibility*(1811)
Charlotte Brontë	*Jane Eyre*(1847),*The Professor*(1857)
James Fenimore Cooper	*The Last of the Mohicans*(1826),*The Spy*(1821),*The Water Witch*(1831)
Charles Dickens	*Bleak House*(1853),*Dombey and Son*(1848),*Great Expectations*(1861)
Ralph Waldo Emerson	*The Conduct of Life*(1860),*English Traits*(1853)
Emily Brontë	*Wuthering Heights*(1847)
Nathaniel Hawthorne	*The House of the Seven Gables*(1851)
Herman Melville	*Moby Dick*(1851),*Redburn*(1849)

Appendix B. Computational complexity analysis

The computational complexity of the proposed method is, in average $\mathcal{O}(n\,\langle k\rangle^2)$, where n is the number of nodes and $\langle k\rangle$ is the average degree of the network. Such a complexity value can be obtained as follows. For each node in the network, the algorithm visits all its neighbours (the number of neighbours of a node is, in average, the average degree of the network $\langle k\rangle$) and, then, the neighbours of its neighbours. The computational complexity can also be computed in terms of the number of edges e:

$$n\,\langle k\rangle^2 = n\left(\frac{2e}{n}\right)^2 = \frac{4e^2}{n}. \tag{B.1}$$

Therefore, $\mathcal{O}(n\,\langle k\rangle^2) = \mathcal{O}(e^2/n)$.

The computational complexity can also be defined in terms of the number of words (tokens) M of the text. The number of nodes n, which is the number of unique words in the text (i.e. the vocabulary size), can be written as a function of the number of tokens M, so that $n = cM^b$ (from Heaps' law [62]), where both c and b are constant values. In addition, the number of words M can be used as an upper bound for the number of edges e. Therefore, equation B.1 can be rewritten as:

$$\mathcal{O}\left(\frac{4e^2}{n}\right) = \mathcal{O}\left(\frac{4M^2}{cM^b}\right) = \mathcal{O}(M^{2-b}). \tag{B.2}$$

Note that c and b are parameters usually defined as $30 \leq c \leq 100$ and $b \approx 0.5$ [63]; therefore the complexity of the proposed method can be approximated to $\mathcal{O}(M\sqrt{M})$.

To measure the execution time, we randomly selected 5% of the partitions from each target language in the European Parliament (EP), and 50% of the books and partitions from Data set 1 and 2, respectively. In this analysis, we did not include the files from the Canadian Hansard because those texts have varied lengths, which would reflect in very different execution times. We present the average times to extract three kinds of features, considering that the set of the 20 most frequent words W is known beforehand. Therefore, the times presented in Table B.1 represent the average time to extract the frequency of the words from W (MFW), the average time to extract the frequency of all 13 subgraphs with three nodes

TABLE B.1 *Average times to extract three kinds of features*

Data set	No.# of samples	MFW	Subgraphs	LSV1 and LSV2
Data set 1	20	0.015 ms	2,264.74 s	2,275.03 s
Data set 2	69	0.012 ms	68.59 s	69.32 s
EP—English	90	0.012 ms	54.24 s	54.88 s
EP—French	64	0.014 ms	81.60 s	82.17 s
EP—Italian	27	0.011 ms	96.92 s	97.51 s
EP—Spanish	34	0.011 ms	74.00 s	74.64 s

(Subgraphs), and the average time to extract the labelled subgraphs for the words in W, in both versions (LSV1 and LSV2). All average times reported were obtained using a computer with i7-3770 CPU and 32 GB of RAM. Apart from the results of Data set 1, which were calculated in texts with 46,025 words, all the others were obtained with partitions of 8,000 words each, which is reflected in faster execution times (less than 100 s compared to an average of 38 min). An important procedure is to save the labelled subgraphs that were already extracted from each file, so that some labelled subgraphs frequencies are already available (and there is no need to be extracted again) in the several cross validation executions. As future work, some improvements could be done in the code that extracts labelled subgraphs. A straightforward improvement is to use parallel threads to extract the subgraphs from different parts of the network.