# Authorship attribution via network motifs identification

Vanessa Queiroz Marinho[1,2], Graeme Hirst[1]
[1]Department of Computer Science
University of Toronto
Toronto, ON, M5S 3G4, Canada
Email: vanessaqm@usp.br, gh@cs.toronto.edu

Diego Raphael Amancio[2]
[2]Department of Computer Science
University of São Paulo
São Carlos, SP, 13566-590, Brazil
Email: diego@icmc.usp.br

*Abstract*—Concepts and methods of complex networks can be used to analyse texts at their different complexity levels. Examples of natural language processing (NLP) tasks studied via topological analysis of networks are keyword identification, automatic extractive summarization and authorship attribution. Even though a myriad of network measurements have been applied to study the authorship attribution problem, the use of motifs for text analysis has been restricted to a few works. The goal of this paper is to apply the concept of motifs, recurrent interconnection patterns, in the authorship attribution task. The absolute frequencies of all thirteen directed motifs with three nodes were extracted from the co-occurrence networks and used as classification features. The effectiveness of these features was verified with four machine learning methods. The results show that motifs are able to distinguish the writing style of different authors. In our best scenario, 57.5% of the books were correctly classified. The chance baseline for this problem is 12.5%. In addition, we have found that function words play an important role in these recurrent patterns. Taken together, our findings suggest that motifs should be further explored in other related linguistic tasks.

*Keywords—Natural language processing, complex networks, authorship attribution, network motifs.*

## I. INTRODUCTION

The representation and characterization of real-world systems with complex networks has been useful to describe a large variety of systems found in the real world [1]. Some examples include the cell, which can be described as a network of substances linked by chemical reactions, and the Internet, a network of routers and computers connected by physical links [2]. Traditionally, the study of networks was mostly limited to the analysis of random graphs. One of the pioneers in graph theory was the mathematician Leonard Euler who, in 1736, solved the famous problem known as The Königsberg Bridges [3]. Since then, graph theory has benefited from major advances [4], mainly after the works of Watts and Strogatz [5] and Barabási and Albert [6] on the ubiquitous properties of real-world networks.

Mathematically, a network is an ordered pair $G = \{V, E\}$ formed by a set $V = \{v_1, v_2, ..., v_n\}$ of vertices (or nodes) and a set $E = \{e_1, e_2, ..., e_m\}$ of edges. The network connectivity can be represented with an adjacency matrix $A$. In this matrix, the possible values for the element $A_{ij}$ are 0 or 1, where $A_{ij} = 1$ iff nodes $i$ and $j$ are connected. Networks can be directed or undirected, a property that depends on the reciprocity of the modelled system. In addition, networks can be formed of weighted or unweighted edges [7].

In the last few years, the finding that many real systems could be characterized by networks with non-trivial patterns [8] allowed rapid development in the area. Some non-trivial patterns include the universal properties known as small-world [5] and scale-free [6]. A rapid increase in data availability and computational capacity allowed the analysis and development of efficient algorithms in several applications, including text analysis via topological characterization of networks. There are several representations of texts as networks, where both nodes and edges may represent distinct textual aspects. In the most common models, nodes represent words and edges are established according to syntactic [9], semantic [10] or empirical [11] relationships. Interestingly, it has been shown that the small-world and scale-free properties arise in such networks [12], in many cases as a consequence of Zipf's Law [13].

A special case of syntactic networks is the word co-occurrence network (or adjacency network) [12]. In this type of network, links are established by connecting adjacent words, since most of the syntactical relations occur among neighbouring words [14]. The representation of texts as word co-occurrence networks has proven useful to tackle different tasks, for example, to create automatic extractive summarizers [9] and to identify the authorship of books [15], [16].

Authorship attribution methods are relevant in practice because they can be applied to classify literary works and solve copyright disputes [17]. The first statistical authorship attribution techniques were devised by Mosteller and Wallace to probe the authorship of the so-called Federalist Papers [18]. Since then, many works tried to define characteristics to quantify the writing style of authors [19].

Recent authorship attribution approaches based on networks [15], [16], [20] quantify authors' writing style by using various measurements that characterize the topology of networks. In general, the topology of a complex network can also be characterized by the number of *motifs* found on its structure. Motifs are small interconnection patterns (subgraphs) forming the topology of networks. Such structures have been able to characterize networks from different fields, as biochemistry, neurobiology, ecology, engineering and social networks [21], [22], [23]. In most of these areas, specific network motifs have been found to play specific roles in the

functional network activity. As a consequence, it has been shown that the prominence of particular subgraphs might be used to automatically identify the function of networks in many fields. Although motifs have been used to detect universal patterns in different languages [23] and to perform textual analysis [24], no comprehensive study has used these structures to characterize and distinguish different writing styles. In particular, we test the hypothesis that authors' writing preferences can be captured in an artlessly manner via identification of network motifs. As such, the main goal of this paper is to probe the relevance of network motifs as features for the authorship attribution problem.

## II. Related Work

In a typical authorship attribution problem, a text whose authorship is unknown is assigned to an author from a set of candidates. The first authorship attribution activities supported by statistical methods are from the XIX century. The goal of these methods is to maximize the probability for a text $x$ to belong to a possible author $a$. Mosteller and Wallace [18] analysed the authorship of a collection of political essays, known as The Federalist Papers. With that study, they made a paramount contribution to the area showing that the frequency of common (function) words (as *and* and *to*) can distinguish different authors.

The research carried out by Mosteller and Wallace [18] was based mostly on simple textual statistics. After that seminal study, researchers have been proposing new attributes in order to characterize writing styles [25]. Some attributes traditionally used in the task include statistical properties of words (e.g., the average length, the frequency, burstiness and the vocabulary richness) [26] and characters (frequencies and long-range correlations) [17]. In addition, syntactic (for example, frequency of specific chunks) and semantic attributes have been used as relevant attributes [26], [27].

In the context of authorship attribution, Uzuner and Katz [28] used a corpus comprising 49 books. They reported accuracies between 34% and 87%. Their best performance was achieved with the frequency of function words. Grieve [29] used 34 features and analysed their performance for different sets of authors. An accuracy of 80% was reached with the frequency of words and punctuation marks using a set of 10 authors. Hirst and Feiguina [27] combined the idea of bigram frequencies with syntactic analysis. Their syntactic label bigrams were found useful to distinguish the works of Anne Brontë and Charlotte Brontë, with an accuracy of 99.5%. Jankowska et al. [30] observed that differences in the usage frequencies of the most common n-grams of characters and words could distinguish different authors. Sapkota et al. [31] demonstrated that character 3-grams that capture information about affixes and punctuation marks are important features for the authorship attribution task.

The dependencies between the topology of word co-occurrence networks and the writing style of authors were observed in some works [15], [16], [20]. The measurements extracted from the networks by Antiqueira et al. [20] were able to distinguish different authors. After the combination of some measurements, it was also possible to cluster different authors, characterizing a common writing style. Amancio et al. [15]

obtained an accuracy rate of 65% when network measurements were combined with the intermittency of the distribution of words along the text. They found that most topological measurements captures syntactic and stylistic characteristics of the language. Lastly Mehri, Darooneh and Shariati [16] modelled 36 Persian books as co-occurrence networks to be used in the authorship attribution task. They did not employ any pre-processing step. The authorship was correctly assigned for 28 books, which represents an accuracy of 77.7%.

In addition to some measurements in complex networks, motifs are structures that can be used to characterize the topology of co-occurrence networks. Motifs are recurrent interconnection patterns found more frequently in real-world networks than in randomized ones [21]. These structures are represented by small subgraphs, generally involving three or four nodes. Milo et al. [21] used a variety of real-world networks (as such transcription networks, food webs and neuronal networks) and extracted the frequency of all possible subgraphs in those networks. The frequency of each subgraph was analysed and compared to its frequency in random graphs. As a consequence, the authors discovered the significant presence of motifs in all networks. As one of their conclusions, they pointed out a structural similarity between the neuronal connectivity and the transcription networks, because these networks presented similar frequencies of their motifs. Moreover, completely different frequencies of motifs were found in electronic circuits with distinct functionalities [21], [22].

Motifs were already used to analyse text [23], [24]. Milo et al. [23] analysed the frequency of motifs in word adjacency networks from texts in English, French, Spanish and Japanese. Despite the differences among these four languages, the authors discovered that they present very similar interconnection patterns. One possible explanation for that is that languages possess an intrinsic structure, which divides words into categories. Therefore, words from one category (e.g. prepositions) tend to be adjacent with others from different categories (e.g. nouns or articles) [32]. Amancio et al. [24] realized that some motifs, as the ones labelled as 12 and 13 in Figure 2 below, rarely happen in real texts.

## III. Methodology

### A. Dataset

The dataset employed in this work comprises 40 novels, written by eight authors. The novels were published between 1835 and 1922. We obtained electronic versions of the books from the Project Gutenberg repository[1]. The full list with the 40 books is summarized in Table I. To avoid the influence of different text lengths in the frequency of motifs and network measurements, each novel was truncated to the length of the shortest book in each scenario used in this paper.

### B. Pre-processing of the text

The task of modelling text as a complex network can be divided into two steps, the pre-processing of the text and the connection of words. One of the first pre-processing activities is to remove punctuation marks. In this paper, contractions (as *he's* and *isn't*) were kept. We believe that the choice of writing

---

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

TABLE I. LIST OF BOOKS EMPLOYED IN THE AUTHORSHIP ATTRIBUTION TASK.

| 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 II. EXAMPLE OF THE TWO PRE-PROCESSING STEPS APPLIED TO SENTENCES FROM THE BOOK *The Adventures of Sherlock Holmes*.

| Original extract | Without stopwords + lemmatization |
|---|---|
| "There are three men waiting for him at the door", said Holmes. "Oh, indeed! You seem to have done the thing very completely. I must compliment you." "And I you", Holmes answered. | three men wait door say holmes oh indeed seem do thing completely must complement holmes answer |

*isn't* instead of its equivalent form *is not* is related to the writing style and, therefore, this should be mapped differently in the network.

One of the pre-processing steps usually applied in text analysis via networked models is the removal of function words or stopwords. These words are mainly prepositions, articles and pronouns that convey little semantic content. The set of stopwords removed in some of our scenarios is presented in Appendix A. However, these words will be included in the creation of the network in some of our experiments. This is based on the fact that these words are very useful in traditional techniques of authorship attribution [18].

The lemmatization of the text is another pre-processing step that can be applied. It is performed with a part-of-speech tagger. In this step, plural words are changed to their singular version, verbs to their infinitive form and names to their masculine form. After this step, words with the same lemma are mapped into the same node. The part-of-speech tagger used to perform the lemmatization process is described in [33].

To demonstrate the two different pre-processing steps that will be used in this paper, Table II illustrates their applications in an extract of the book *The Adventures of Sherlock Holmes*, by Arthur Conan Doyle. Here, these two steps will be combined and evaluated in order to find the best scenario for the authorship attribution task.
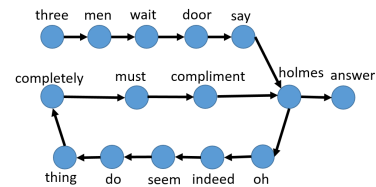


Fig. 1. The graph represents the directed co-occurrence network for the pre-processed sentence "three men wait door say holmes oh indeed seem do thing completely must compliment holmes answer". This network was created connecting each word to its adjacent word and the direction of the edges is defined from the word in the left to the one in the right.
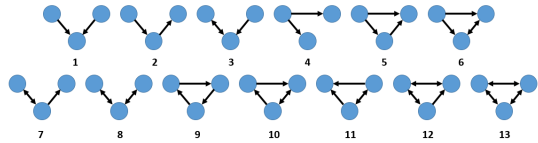


Fig. 2. All thirteen directed motifs involving three nodes.

### C. Co-occurrence networks

After the pre-processing steps, we need to connect each word to create the network. The written language is formed by linear chains of words; therefore, the easiest way to represent it is to connect adjacent words. This type of network, known as a co-occurrence or adjacency network, is widely used in the literature [12], [15], [34]. In a word co-occurrence network, words are mapped into nodes, and a link between two nodes is established if the corresponding words appear adjacent at least once in the (pre-processed) text. The directed network that represents the pre-processed sentences of Table II is illustrated in Figure 1.

### D. Measurements

The topological structure of a complex network can be characterized by several metrics. Firstly, the directed networks derived from our dataset will be characterized by the absolute frequency of all directed motifs involving three nodes. We did not employ motifs involving more nodes because we aimed to use a small set of attributes. The set of directed motifs with three nodes is presented in Figure 2. The frequencies were extracted with the software *mfinder* [35].

In order to compare the classification results obtained with the frequency of motifs and the ones when other networked measurements are used, we extracted five measurements from complex networks. We modified the directed network to obtain its undirected version. This was easily achieved by transforming the adjacency matrix $A$ into its symmetric form. Each one of the measurements is described below:

- Average Degree of Neighbors ($ADN_i$): The degree indicates the amount of distinct neighbors. The $ADN_i$ quantifies the average degree of all neighbors of $i$ [36].

- Average Shortest Path Length ($L_i$): This measurement represents the average distance between node $i$ and all other nodes. In textual networks, $L_i$ quantifies the relevance of each word $i$ according to its distance to the most frequent words [15].

- Betweenness Centrality ($B_i$): The betweenness centrality measures the relevance of a node $i$ according to the number of shortest paths that include this node [4]. In textual networks, frequent words tend to have a high value for this measurement. In addition, some words may act as *bridges* (or articulation points) connecting concepts from different communities and thus they also may take high values of betweenness, regardless of their frequency. Therefore, this measurement may quantify the diversity of contexts in which a word can be used [15].

- Clustering Coefficient ($CC_i$): The clustering coefficient of a node $i$ indicates the probability of two of its neighbors being connected. As it happens in many small-world networks, Cancho and Solé [12] found that the clustering coefficient in textual networks is much higher than the expected value in equivalent random networks.

- Assortativity ($r$): In some networks, nodes of certain degree tend to connect with similar nodes [37]. This measurement quantifies the degree correlation and it can be calculated with the Pearson correlation coefficient, $r$ [38]. In most of the cases, co-occurrence networks are disassortative, $r < 0$.

### E. Extracting properties of books from properties of words

Apart from assortativity and motifs, all measurements are locally defined, i.e. a value is assigned for each word. The goal is to obtain values that can be used as global measurements of each book, not only of each single word. Therefore, we adopted a summarization procedure. For each local measurement $X$, the most natural choice is to calculate the average $\langle X \rangle$, i.e. the average of $X$ over all the $M$ unique words. We also computed the deviation and skewness of the distribution, denoted as $\sigma(X)$ and $\gamma(X)$, respectively. Note that this type of summarization has been performed in related studies [39]. In summary, the three features we use for each measurement $X = \{ADN, L, B, CC\}$ are:

$$\text{Average:} \qquad \langle X \rangle = \frac{1}{M} \sum_{i=1}^{M} X_i,$$

$$\text{Deviation:} \qquad \sigma(X) = \left[ \frac{1}{M-1} \sum_{i=1}^{M} (X_i - \langle X \rangle)^2 \right]^{1/2},$$

$$\text{Skewness:} \qquad \gamma(X) = \left\langle \left( \frac{X - \langle X \rangle}{\sigma(X)} \right)^3 \right\rangle.$$

### F. Machine Learning Methods

In order to quantify the ability of the motifs to distinguish among authors, we employed four machine learning algorithms to induce classifiers from a training set. The techniques are *Support vector machines* (SVM), kNN, *Naive Bayes* and C4.5 [40]. We used the default configuration of these methods available in Weka [41]. It has been shown that the default configuration of parameters in Weka provides near optimal performance for most of the cases [42]. These methods were

TABLE III. PERCENTAGE OF BOOKS CORRECTLY CLASSIFIED WHEN THE ABSOLUTE FREQUENCY OF DIRECTED MOTIFS WAS USED AS THE ONLY CLASSIFICATION FEATURE

|  | C4.5 | kNN | SVM | Naive Bayes |
|---|---|---|---|---|
| (i) Original text | 40.0 | 55.0 | 45.0 | 45.0 |
| (ii) Without stopwords | 27.5 | 32.5 | 0.0 | 30.0 |
| (iii) Lemmatization process | 57.5 | 45.0 | 45.0 | 52.5 |
| (iv) Without stopwords + lemmatization | 22.5 | 27.5 | 2.5 | 30.0 |

applied to a training set independent of the test set using the cross validation technique with 10 folds [42]. In this technique, at each cycle one tenth of the books are used as test while the other nine tenths are used for training the algorithm.

## IV. RESULTS

As we described in Section III-B, different pre-processing steps may be applied before creating the network of each text. For each one of our experiments, we considered four different input scenarios: (i) original text, (ii) without stopwords, (iii) after the lemmatization process and (iv) after the removal of stopwords and the lemmatization process. For each input scenario, we derived one network for each book. In the first experiment, the absolute frequency of all thirteen motifs was extracted. Finding the set of motifs of size 3 is computationally feasible. The average time to extract all motifs in scenarios (ii) and (iv) was around 30 seconds per book. The derived networks in scenarios (i) and (iii) have more nodes and edges and, therefore, the time to extract the motifs increased to an average of 4 minutes per book. We applied the machine learning methods described in Section III-F. The results are presented in Table III. In these results, the function words play an important role during the extraction of the motifs because the overall performance drops to very low values when such very frequent words are disregarded.

As a way to evaluate whether the motifs are actually extracting a real pattern, a random verification step was performed. Only the input scenario (iii) was selected, since it led to the best results. In this process, each instance keeps the same frequencies as the original one; however, the class $y$ is randomly selected from the set of all possible authors. Note that the correct author is also included in the set. This process was performed 10 times and the averages of books correctly classified are 12% (C4.5), 9% (kNN), 13% (SVM) and 10.8% (Naive Bayes). These results are similar to the chance baseline for this problem, 12.5%, since each one of the 8 authors has the same probability of being randomly selected. Therefore, these results and the ones presented in Table III confirm that the frequency of motifs is able to significantly extract patterns related to the authorship of each book.

A visualization of the classification for two different input scenarios is provided in Figure 3. Each one of the 8 authors is represented by a different symbol and each element in the plot is a distinct book. The highest accuracy rate found for scenario (i) was 55%. The ability to distinguish among different authors is more evident in Figure 3 (A), which represents scenario (i). The lowest accuracy rates were found for scenario (ii), which is presented in Figure 3 (B). This poor discriminability of authors can be attributed to the fact that stopwords were disregarded in that scenario.
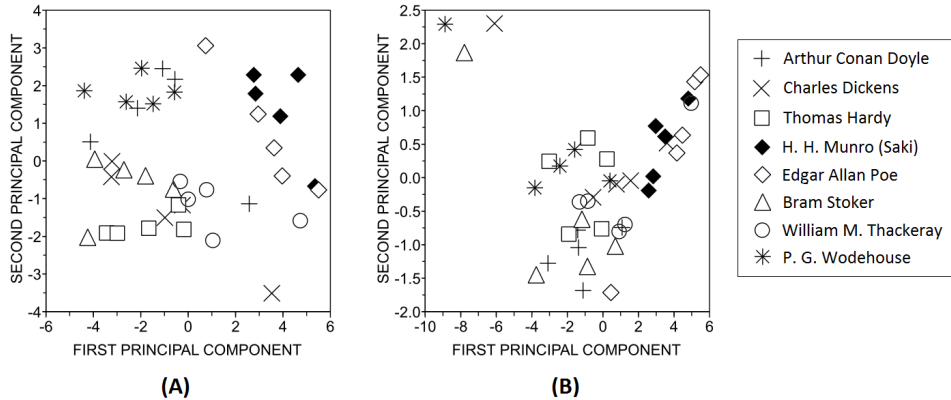
Fig. 3. Principal component analysis performed using network motifs as attributes. The input scenarios were: (A) Scenario (i), original text; and (B) Scenario (ii), text without stopwords. Note that the discriminability increases when the stopwords are considered.

TABLE IV. PERCENTAGE OF BOOKS CORRECTLY CLASSIFIED WHEN SEVERAL FEATURES WERE EXTRACTED FROM THE BOOKS

| | C4.5 | kNN | SVM | Naive Bayes |
|---|---|---|---|---|
| (i) Original text | 50.0 | 42.5 | 42.5 | 55.0 |
| (ii) Without stopwords | 37.5 | 45.0 | 27.5 | 37.5 |
| (iii) Lemmatization process | 47.5 | 50.0 | 37.5 | 45.0 |
| (iv) Without stopwords + lemmatization | 32.5 | 37.5 | 32.5 | 40.0 |
| Frequency of the 20 most frequent words | 55.0 | 67.5 | 72.5 | 55.0 |

In our last experiment, several features were extracted from the books. Firstly, the network measurements explained in Section III-D were extracted from the undirected networks of each book. The results for the four scenarios are listed in Table IV. In addition, the relative frequency of the 20 most frequent words in all 40 books were extracted from the input scenario (i) and used as a classification feature. This result is presented in the last row of Table IV. Comparing the results of Tables III and IV, the motif extraction for the scenarios (i) and (iii) performed almost as well as the results retrieved with traditional network measurements. However, some complex network measurements are correlated with the frequency of words, e.g. betweenness centrality, and this may improve the performance of such measurements. Moreover, the network measurements employed in this paper are not so sensitive to the removal of the function words. The results obtained with the frequency of common words were relatively high. The 20 most frequent words were mainly function words (e.g. *and*, *he*, *his*). These results were already expected, since function words can be used as a strong tool to detect writing styles. However, such features are prone to manipulation, a disadvantage not present in networked approaches.

## V. CONCLUSION

The authorship attribution task has been studied with some success through the representation of word co-occurrence networks. In this paper, we investigated whether network motifs found on these networks are able to discriminate different authors.

The accuracies obtained in this paper are relevant, which confirms that motifs are able to capture aspects of the writing style of different authors. Moreover, the results are considerably higher than the chance baseline expected for this problem, 12.5%. Therefore, we can conclude that there is a dependency between the frequency of the motifs and the writing style of different authors. Another interesting result is the fact that function words played a crucial role to the success of our approach. When the motifs extraction was performed in networks without those words, the accuracies decreased and even reached a level that one of the classifiers could not find any pattern to perform the classification.

The focus of this paper was to evaluate the accuracies when motifs were used as the only classification attributes. To our knowledge, this is the first time that network motifs are used to identify authorship in written texts. Our best result, 57.5% is lower than some traditional approaches found in the literature, that are around 90% [27], [28], [29]. Comparing our results with others that use co-occurrence networks, Amancio et al. [15] obtained an accuracy rate of 65%. From the dataset used by Mehri, Darooneh and Shariati [16], 28 books (77.7%) were correctly classified. The results retrieved in this paper may be seen as complementary measurements and can be combined with traditional techniques usually employed in authorship attribution. Moreover, the significative results obtained here suggest that the characterization based on motifs should be further explored in other related linguistic tasks, such as language identification or complexity in authors' writing style.

## APPENDIX A
## LIST OF STOPWORDS

*i, me, my, myself, we, our, ours, ourselves, you, your, yours, yourself, yourselves, he, him, his, himself, she, her, hers, herself, it, its, itself, they, them, their, theirs, themselves, what, which, who, whom, this, that, these, those, am, is, are, was, were, be, been, being, have, has, had, having, do, does, did, doing, a, an, the, and, but, if, or, because, as, until, while, of, at, by, for, with, about, against, between, into, through, during, before, after, above, below, to, from, up, down, in, out, on, off, over, under, again, further, then, once, here, there, when, where, why, how, all, any, both, each, few, more, most, other, some,*

*such, no, nor, not, only, own, same, so, than, too, very, s, t, can, will, just, don, should, now*

## REFERENCES

[1] R. Albert and A.-L. Barabási, "Statistical mechanics of complex networks," *Reviews of Modern Physics*, 2002.

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

[3] A.-L. Barabási, *Linked: How Everything Is Connected to Everything Else and What It Means for Business, Science, and Everyday Life*. Plume Books, April 2003.

[4] S. Boccaletti, V. Latora, Y. Moreno, M. Chavez, and D.-U. Hwang, "Complex networks: Structure and dynamics," *Physics Reports*, vol. 424, no. 4-5, pp. 175–308, February 2006.

[5] D. J. Watts and S. H. Strogatz, "Collective dynamics of small-world networks," *Nature*, no. 393, pp. 440–442, 1998.

[6] A.-L. Barabási and R. Albert, "Emergence of scaling in random networks," *Science*, vol. 286, no. 5439, pp. 509–512, 1999.

[7] L. F. Costa, F. A. Rodrigues, G. Travieso, and P. R. V. Boas, "Characterization of complex networks: A survey of measurements," *Advances in Physics*, vol. 56, no. 1, pp. 167–242, January 2007.

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

[9] D. R. Amancio, M. G. V. Nunes, O. N. Oliveira, Jr, and L. F. Costa, "Extractive summarization using complex networks and syntactic dependency," *Physica A: Statistical Mechanics and its Applications*, vol. 391, no. 4, pp. 1855 – 1864, 2012.

[10] H. Liu, "Statistical properties of Chinese semantic networks," *Chinese Science Bulletin*, vol. 54, no. 16, pp. 2781–2785, 2009.

[11] G. A. Ludueña, M. D. Behzad, and C. Gros, "Exploration in free word association networks: models and experiment," *Cognitive Processing*, vol. 15, no. 2, pp. 195–200, 2014.

[12] R. Ferrer i Cancho and R. V. Solé, "The small world of human language," *Proceedings of The Royal Society of London. Series B, Biological Sciences*, vol. 268, pp. 2261–2266, 2001.

[13] G. Zipf, "Human behaviour and the principle of least-effort." Cambridge, MA: Addison-Wesley, 1949.

[14] R. Ferrer i Cancho, "Euclidean distance between syntactically linked words," *Physical Review E*, vol. 70, p. 056135, Nov 2004.

[15] D. R. Amancio, E. G. Altmann, O. N. Oliveira, Jr, and L. F. Costa, "Comparing intermittency and network measurements of words and their dependence on authorship," *New Journal of Physics*, vol. 13, no. 12, p. 123024, 2011.

[16] A. Mehri, A. H. Darooneh, and A. Shariati, "The complex networks approach for authorship attribution of books," *Physica A: Statistical Mechanics and its Applications*, vol. 391, no. 7, pp. 2429–2437, 2012.

[17] T. D. Grant, "Quantifying evidence for forensic authorship analysis," *International Journal of Speech, Language and the Law*, 2007.

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

[19] D. I. Holmes, "Authorship attribution," *Computers and the Humanities*, vol. 28, no. 2, pp. 87–106, 1994.

[20] L. Antiqueira, T. A. S. Pardo, M. G. V. Nunes, O. N. Oliveira, Jr, and L. F. Costa, "Some issues on complex networks for author characterization," in *Fourth Workshop in Information and Human Language Technology (TIL'06) in the Proceedings of International Joint Conference IBERAMIA-SBIA-SBRN*. Ribeirão Preto, Brazil: ICMC-USP, October 23-28 2006.

[21] R. Milo, S. Shen-Orr, S. Itzkovitz, N. Kashtan, D. Chklovskii, and U. Alon, "Network motifs: simple building blocks of complex networks." *Science*, vol. 298, no. 5594, pp. 824–827, October 2002.

[22] N. Kashtan, S. Itzkovitz, R. Milo, and U. Alon, "Topological generalizations of network motifs," *Physical Review E*, vol. 70, p. 031909, Sep 2004.

[23] R. Milo, S. Itzkovitz, N. Kashtan, R. Levitt, S. Shen-Orr, I. Ayzenshtat, M. Sheffer, and U. Alon, "Superfamilies of evolved and designed networks." *Science*, vol. 303, no. 5663, pp. 1538–1542, March 2004.

[24] D. R. Amancio, E. G. Altmann, D. Rybski, O. N. Oliveira, Jr, and L. F. Costa, "Probing the statistical properties of unknown texts: application to the Voynich Manuscript," *PLoS ONE*, vol. 8, no. 7, p. e67310, 07 2013.

[25] P. Juola, "Authorship attribution," *Foundations and Trends in Information Retrieval*, vol. 1, no. 3, pp. 233–334, Dec. 2006.

[26] E. Stamatatos, "A survey of modern authorship attribution methods," *Journal of the American Society for Information Science and Technology*, vol. 60, no. 3, pp. 538–556, Mar. 2009.

[27] G. Hirst and O. Feiguina, "Bigrams of syntactic labels for authorship discrimination of short texts," *Literary and Linguistic Computing*, vol. 22, no. 4, pp. 405–417, 2007.

[28] Ö. Uzuner and B. Katz, *Natural Language Processing – IJCNLP 2005: Second International Joint Conference, Jeju Island, Korea, October 11-13, 2005. Proceedings*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2005, ch. A Comparative Study of Language Models for Book and Author Recognition, pp. 969–980.

[29] J. Grieve, "Quantitative authorship attribution: An evaluation of techniques." *Literary and Linguistic Computing*, vol. 22, no. 3, pp. 251–270, 2007.

[30] M. Jankowska, E. Milios, and V. Keselj, "Author verification using common n-gram profiles of text documents," in *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*. Dublin City University and Association for Computational Linguistics, 2014.

[31] U. Sapkota, S. Bethard, M. Montes, and T. Solorio, "Not all character n-grams are created equal: A study in authorship attribution," in *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Denver, Colorado, May–June 2015, pp. 93–102.

[32] J. Allen, *Natural Language Understanding (2nd Edition)*. Redwood City, CA, USA: Benjamin-Cummings Publishing Co., Inc., 1995.

[33] S. Bird, E. Klein, and E. Loper, *Natural Language Processing with Python*, 1st ed. O'Reilly Media, Inc., 2009.

[34] R. M. Roxas and G. Tapang, "Prose and poetry classification and boundary detection using word adjacency network analysis," *International Journal of Modern Physics C*, vol. 21, pp. 503–512, 2010.

[35] N. Kashtan, S. Itzkovitz, R. Milo, and U. Alon, "Efficient sampling algorithm for estimating subgraph concentrations and detecting network motifs," *Bioinformatics*, vol. 20, no. 11, pp. 1746–1758, Jul. 2004.

[36] R. Pastor-Satorras, A. Vázquez, and A. Vespignani, "Dynamical and correlation properties of the Internet," *Physical Review Letters*, vol. 87, no. 25, p. 258701, 2001.

[37] M. E. Newman, "Assortative mixing in networks," *Physical Review Letters*, vol. 89, no. 20, p. 208701, 2002.

[38] L. F. Costa, O. N. Oliveira, Jr, G. Travieso, F. A. Rodrigues, P. R. Villas Boas, L. Antiqueira, M. P. Viana, and L. E. C. Rocha, "Analyzing and modeling real-world phenomena with complex networks: a survey of applications," *Advances in Physics*, vol. 60, no. 3, pp. 329–412, 2011.

[39] D. R. Amancio, "A complex network approach to stylometry," *PloS One*, vol. 10, no. 8, p. e0136076, 2015.

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

[41] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The weka data mining software: An update," *SIGKDD Explorations Newsletter*, vol. 11, no. 1, pp. 10–18, Nov. 2009.

[42] D. R. Amancio, C. H. Comin, D. Casanova, G. Travieso, O. M. Bruno, F. A. Rodrigues, and L. d. F. Costa, "A systematic comparison of supervised classifiers," *PLoS one*, vol. 9, no. 4, 2014.