Comparing intermittency and network measurements of words and their dependence on authorship

Diego Raphael Amancio\textsuperscript{1}, Eduardo G Altmann\textsuperscript{2,3}, Osvaldo N Oliveira Jr\textsuperscript{1} and Luciano da Fontoura Costa\textsuperscript{1}

\textsuperscript{1} Institute of Physics of São Carlos, University of São Paulo, PO Box 369, Postal Code 13560-970 São Carlos, São Paulo, Brazil
\textsuperscript{2} Max Planck Institute for the Physics of Complex Systems, Dresden, Germany

E-mail: edugalt@pks.mpg.de

\textit{New Journal of Physics} 13 (2011) 123024 (17pp)

Received 29 June 2011
Published 16 December 2011
Online at \url{http://www.njp.org/}
doi:10.1088/1367-2630/13/12/123024

\textbf{Abstract.} Many features of texts and languages can now be inferred from statistical analyses using concepts from complex networks and dynamical systems. In this paper, we quantify how topological properties of word co-occurrence networks and intermittency (or burstiness) in word distribution depend on the style of authors. Our database contains 40 books by eight authors who lived in the nineteenth and twentieth centuries, for which the following network measurements were obtained: the clustering coefficient, average shortest path lengths and betweenness. We found that the two factors with stronger dependence on authors were skewness in the distribution of word intermittency and the average shortest paths. Other factors such as betweenness and Zipf’s law exponent show only weak dependence on authorship. Also assessed was the contribution from each measurement to authorship recognition using three machine learning methods. The best performance was about 65\% accuracy upon combining complex networks and intermittency features with the nearest-neighbor algorithm of automatic authorship. From a detailed analysis of the interdependence of the various metrics, it is concluded that the methods used here are complementary for providing short- and long-scale perspectives on texts, which are useful for applications such as the identification of topical words and information retrieval.

\textsuperscript{3} Author to whom any correspondence should be addressed.

\textsuperscript{S} Online supplementary data available from \url{stacks.iop/NJP/13/123024/mmedia}
1. Introduction

The application of the ideas from statistical physics to text analysis has a long tradition since Shannon’s usage of entropy as the central concept in information theory [1]. In recent years, physicists have proposed new approaches based on concepts from complex networks [2–15] and dynamical systems [16–20]. In the former, text is represented as complex networks with words (nodes) being connected (links) using procedures depending on their syntactic or semantic relationships [2]. Several of these networks share topological properties such as the scaling in degree [3, 4] and the small world feature [5, 6]. The co-occurrence networks, where adjacent words are linked to each other, are probably the most popular for applications owing to their ability to capture important syntactic and semantic aspects of texts with a straightforward construction procedure. These networks were employed to evaluate writing quality [7] and machine translations [8, 9], to generate and evaluate summaries [10], to construct spell checkers [11], to recognize patterns in poetry [12] and prose [13, 14] and to study general properties of written language [15]. While co-occurrence networks focus mainly on short scales, an increasingly popular approach addresses longer text scales [16–18, 21–24]. The usefulness of the latter approach stems from the finding that topical words are unevenly distributed along the text when compared with a random process or function words. This observation can be quantitatively investigated using different analogies and measures familiar to the communities of statistical physics and dynamical systems, including level statistics [22, 24], burstiness [17–19], entropy [21] and intermittency measures [20]. The author dependence of the features mentioned above has been noted [14, 17], but very little work has so far been devoted to quantifying the extent of this dependence and to testing its usefulness for the automatic detection of authors.

In the field of authorship recognition (or stylometry), one tries to identify the author of documents whose identity is lacking [25]. Some simple quantitative proposals, such as the use of
word length to distinguish between authors, go back to the mid-nineteenth century (see [26] for a historical account). One important recent contribution was by Mosteller and Wallace [27] who showed that the frequency of function words (such as ‘any’, ‘from’, ‘an’, ‘there’ and ‘upon’) can be used to characterize the style of authors. This feature is so strong that even letter pair frequencies can provide good distinguishability between authors [26]. Frequent words are also responsible for the success of the approach—proposed and investigated by physicists—that consists in quantifying the similarity between two books based on the distance between their word-frequency rankings [28–30]. More recently, new features have been proposed: word length, sentence length; frequency of punctuation marks and contractions; and frequency of graphemes, collocations and words. A summary of these recent results is given in [31].

In this paper, we investigate how the metrics of complex networks and intermittency—familiar to physicists—depend on the style of authors. We start quantifying the metrics for each word (section 2). These are used in the definition of global features for each book that are tested according to their efficiency in algorithms of authorship classification (section 3). Finally, we discuss the importance of each feature (section 4). The primary goal of this paper is not to improve state-of-the-art methods of automatic authorship recognition. Instead, we wish to estimate the authorship dependence of the selected metrics. We evaluate our results using authorship classification tests because they provide a statistical rigorous method for quantifying the importance of different features. Nevertheless, our study reveals interesting insights that are potentially useful in real applications and therefore we include a comparison with more traditional statistical natural language methods (that increase the efficiency from 62.5 to 90.0% correct attributions, see section 4.2).

2. Statistical quantification of the role of words in texts

2.1. Database

Our database contains five novels from each of eight authors who lived between 1809 and 1975, which are available in an online repository (http://www.gutenberg.org/). The list of books is given in section 1 of the supplementary information (SI) available from stacks.iop.org/NJP/13/123024/mmedia. To avoid the effects from the length of the texts, each book was limited to their first 18 200 tokens, which correspond to the length of the shortest book. In the remainder of this section, we illustrate our results using the book The Adventures of Sally by P G Wodehouse. The results for all books appear in SI (section 3) and are discussed in section 3 below.

2.1.1. Pre-processing of the text. Before extracting complex networks and intermittency measurements from the texts, some preprocessing steps were applied. Initially, a pre-compiled list of stopwords including articles, prepositions and adverbs was removed from the text (see SI (section 2), available from stacks.iop.org/NJP/13/123024/mmedia). Previous work for recognizing authorship used the frequency of function words, but we decided not to use them in our study because we are interested in the interrelation between words with a pronounced semantic content. This procedure has been employed in many works (see, e.g., [7–9, 12, 15]) and it is crucial to determine how these techniques depend on the writing style of each author. Next, a lemmatization step was applied to the remaining words using the MXPost part-of-speech Tagger based on Ratnaparki’s model [32]. Table 1 exemplifies the application of these
Table 1. Example of the pre-processing steps applied to the texts. An extract (first column) obtained from the book *The Adventures of Sally* by P G Wodehouse is shown after the removal of the stopwords (second column) and after lemmatization (third column).

| Original                                      | Without stopwords | After lemmatization |
|-----------------------------------------------|-------------------|---------------------|
| What’s that? asked Sally.                    | asked Sally       | ask Sally           |
| Pay my bill for last week, pay bill last week| morning Sally got | morning Sally get   |
| due this morning. Sally got up quickly, and flitting down the table, put her arm round her friend’s shoulder and whispered in her ear. | quickly flitting | quickly flit table put arm friend shoulder whispered ear |

pre-processing steps. With this standardization, we grouped together all words referring to the same concept, despite the differences in flexion.

2.2. Network measurements

Complex networks have been used to characterize different properties of languages [5–10, 14, 33]. Here we are interested in author-specific characteristics and therefore we adopt a network description based on word co-occurrence [5, 7–9, 14, 33], where nodes are words and links are established between subsequent words. This procedure is illustrated in figure 1. The network is defined by a set $V = \{v_1, v_2, \ldots, v_n\}$ of vertices, and a set $E$ of edges and is represented as a nonsymmetric weighted matrix $W$. By construction, $W$ is a square matrix of size $n$, where $n$ is the number of distinct words after the pre-processing step. The elements $w_{ij}$ of $W$ indicate the strength that $v_i$ is connected to $v_j$ ($v_i \rightarrow v_j$), i.e. the number of times word $v_j$ appears immediately after word $v_i$. Additionally, we used the nonweighted and undirected network corresponding to $W$, denoted by the matrix $A$ whose elements $a_{ij} = 1$ if the words represented by the vertices $v_i$ and $v_j$ appeared as neighbors at least once in the text. Otherwise $a_{ij} = 0$.

We shall use the statistical properties of complex network measurements in the networks $W$ and $A$. In this section, we discuss the word-specific local measurements, and in section 3.1, we show how to connect them to obtain a global characterization of the network. The number of occurrences $N_i$ of each word $i$ represented by node $v_i$ is

$$N_i = s_i^{\text{in}} = s_i^{\text{out}} = \sum_j w_{ij}, \quad (1)$$

where $s_i^{\text{in}}$ and $s_i^{\text{out}}$ are, respectively, the weight (or strength [34, 35]) of the ingoing and outgoing edges of node $v_i$. Therefore, the degree of each node $v_i$ is the frequency of appearance of word $v_i$ and the degree distribution of $W$ is proportional to the normalized frequency ($f_i = N_i/N_T$, $N_T = \sum_i N_i$) distribution of words, given by Zipf’s law [36, 37]. Below we discuss three typical measurements: the clustering coefficient, average shortest path length and betweenness centrality.

---

The triple equality in equation (1) is not valid for the first and for the last word in the text. The correct expression for the first word is $N_i = k_{\text{out}}(i) = \sum_j w_{ij}$ and that for the last word is $N_i = k_{\text{in}}(i) = \sum_j w_{ij}$.
2.2.1. Clustering coefficient $C$.

The clustering coefficient ($C$) measures the probability that the neighbors of a given vertex $v_i$ are connected. This measurement has been widely employed in complex networks, e.g. to verify the presence of communities [38–40] and to distinguish random networks from other small world networks [5, 41]. Traditionally, the clustering coefficient is defined without considering weights or directions as

$$C_i = \frac{3}{2} \frac{\sum_{k>j>i} a_{ij} a_{ik} a_{jk}}{\sum_{k>j>i} a_{ij} a_{ik} + a_{ij} a_{jk} + a_{ki} a_{kj}},$$

which is equivalent to the fraction of the number of triangles among all possible triads of connected nodes and therefore ranges from 0 to 1. With regard to the interpretation of this measure, Ferrer i Cancho and Solé [5] found that the clustering coefficient of networks representing text was much larger than that expected just by chance (i.e. the value expected for the corresponding random networks).

We singled out the words (and their neighbors) with the highest and lowest values of the clustering coefficient in the book *The Adventures of Sally* by P. G. Wodehouse, which are shown in table 2 for the frequency $N = 5$. From the definition, one expects words with the highest $C$ to have neighbors also connected to each other. This is the case for the words ‘sand’ and ‘excitement’. On the other hand, words whose neighbors are not related to each other at all display low values of $C$ (e.g. there is no link between the neighbors of ‘full’ or between the neighbors of ‘high’). Qualitatively, the clustering coefficient quantifies how words are connected to specific contexts. Indeed, the words ‘sand’ and ‘excitement’ tend to be more restricted to a specific context, while ‘full’ tends to appear in a myriad of contexts. Therefore, it seems that the clustering coefficient can be useful in detecting authorship by quantifying the tendency for using semantic-specific or generic words.

The words with $N_i < 5$ were considered to lack statistics and were disregarded in all the analyses involving the clustering $C$.  

---

**Figure 1.** Example of networks: (a) the subgraph obtained for the sentences shown in table 1; and (b) the global network obtained from the first 1000 associations of the same book.
Table 2. Words of the book *The Adventures of Sally* with the highest and lowest clustering coefficients (the average clustering $\langle C \rangle = 0.085$), for words with $N_i = 5$. The five words with $C = 0$ were randomly selected among the 18 words with $N = 5$ and $C = 0$.

| Word   | Neighbors                                                                 | $N$ | $C$ |
|--------|---------------------------------------------------------------------------|-----|-----|
| Shortly | twelve, see, say, Sally, news, never, heaven, find, enter and Carmyle     | 5   | 0.27|
| Excitement | thing, suppressed, Sally, mince, last, can, come, bristle, brief and apart | 5   | 0.25|
| Sand    | watch, want, sit, shuffling, seat, here, golden, first, dark and Roville   | 5   | 0.18|
| Nose    | voice, tip, sort, smut, smooth, Sally, oh, glance, tell and come          | 5   | 0.18|
| Country | time, still, somewhere, say, place may, happen, great and glorious         | 5   | 0.18|
| Startle | shy, seem, mill, little, gratify, gather, first, everyday, displeased and considerably | 5   | 0.00|
| High    | recess, mouth, motive, lapse, figure, even, disposal, critical, collar and check | 5   | 0.00|
| Gold    | voice, spin, pencil, loan, knob, information, high, heavy, frame and buy  | 5   | 0.00|
| Gift    | tongue, take, sort, potential, mean, few, easily, compensating, blessing and acquire | 5   | 0.00|
| Full    | Tuesday, peal, later, home, happy, gratitude, gleaming, glance, color and battle | 5   | 0.00|

2.2.2. Average shortest path length $L$. A shortest path (or geodesic path) between two nodes is defined as the path whose sum of edge weights is minimum. We start by defining $d_{ij}$ as the length of the shortest path between $v_i$ and $v_j$ (in this case $A$ is employed). Then the average shortest (or geodesic) path length for $v_i$ ($L_i$) is the average shortest path to all other ($n - 1$) nodes of the network:

$$L_i = \frac{1}{n-1} \sum_{j=1}^{n} d_{ij}, \quad (3)$$

which takes low values if $v_i$ is close to the other nodes.

The words with the lowest $L$ include the characters ‘Sally’ ($L = 2.35$, $N = 347$) and ‘Fillmore’ ($L = 2.51$, $N = 138$), in addition to high-frequency words, such as ‘say’ ($L = 2.45$, $N = 349$), ‘good’ ($L = 2.46$, $N = 107$) and ‘man’ ($L = 2.50$, $N = 193$). As for the words with the highest $L$, we found: ‘white-clad’ ($L = 6.33$, $N = 1$), ‘affability’ ($L = 6.31$, $N = 1$), ‘whirl’ ($L = 5.89$, $N = 1$), ‘jazz’ ($L = 5.87$, $N = 1$) and ‘war-aims’ ($L = 5.84$, $N = 1$). Interestingly, all these five words appeared only once in the text, indicating that one of the reasons for a high $L$ could be the low frequency $N$. However, $L$ is not only a consequence of the frequency $N$ of the words, as low-frequency words can also take low values of $L$. This is illustrated in table 3, which compares words with the same $N$ but different $L$. The frequency has a limited influence on $L$, with a Pearson correlation $\text{Corr}(L, N) = -0.36$ calculated over

*New Journal of Physics* 13 (2011) 123024 (http://www.njp.org/)
Table 3. Comparing the average shortest path length $L$ for words with the same frequency $N$ of the book *The Adventures of Sally*. For a given $N$, $L$ may vary widely, which shows the dependence of $L$ on the neighborhood connectivity.

| Word     | $N_i$ | $L_i$ | Word     | $N_i$ | $L_i$ |
|----------|-------|-------|----------|-------|-------|
| Red      | 5     | 3.71  | Earth    | 5     | 2.99  |
| Shudder  | 4     | 3.97  | Lucky    | 4     | 3.00  |
| Maxwell  | 3     | 5.55  | Funny    | 3     | 3.10  |
| Dark     | 2     | 5.15  | Kiss     | 2     | 3.08  |
| Affability | 1   | 6.34  | Repose   | 1     | 3.11  |

all words. Actually, the determining factor is the neighborhood of the word. To understand why this happens, consider the words ‘affability’ and ‘repose’. While the former has as neighbors the words ‘jaunty’ ($N = 1$) and ‘white-clad’ ($N = 1$), the latter has as neighbors the words ‘Sally’ ($N = 347$) and night ($N = 20$). Therefore, one may infer that $L$ actually quantifies the importance of a word according to its distance to the most frequent words. Since we removed stopwords, the shortest path may be thought of as quantifying the distance from a word to the core-content words of the book.

2.2.3. *Betweenness.* Betweenness $B$ is a measurement of centrality, with higher values being assigned to the nodes considered as the most relevant in terms of linking different words. In other words, with $B$ one attempts to quantify the frequency of access of each node, assuming that a given target node in the network is reached from a specific source node via shortest paths. Betweenness is defined as follows. Let $n_{ij}^1$ be the number of distinct shortest paths between the source node $v_s$ and the target node $v_t$ that pass through the node $v_i$. If $g_{st}$ is the total number of shortest paths between $v_s$ and $v_t$, then $B_i$ is given by

$$B_i = \sum_s \sum_t \frac{n_{ij}^1}{g_{st}}. \tag{4}$$

In the context of text analysis, high-frequency words tend to have high $B$. However, some words may play the role of articulation points by linking concepts related to distinct communities. To illustrate this, we show in table 4 that words with similar $N$ may take very different $B$. A comparison between the left and right columns suggests that words with high $B$ connect concepts because of their probable appearance in various contexts. Therefore, similarly to the clustering coefficient $C$, the betweenness centrality $B$ seems to quantify the variety of contexts in which a word can appear. Note, however, that $B$ is based on a global connectivity pattern, in contrast to $C$.

2.3. *Intermittency measurements*

The uneven distribution of words across different documents is an essential feature exploited in statistical natural language processing. For instance, by investigating words appearing overconcentrated in specific documents (when compared with their overall frequency) one can detect keywords, topics and authorship [27, 42]. This is the basic idea of tf–idf (term frequency–inverse document frequency) and related measures that are also at the core of search
Table 4. Comparing the betweenness $B$ for words of the book *The Adventures of Sally* with the same frequency $N$. For a given $N$, the betweenness may vary widely, since low-frequency words may have high betweenness as they may appear in different contexts.

| Word    | $N_i$ | $B_i$ | Word    | $N_i$ | $B_i$ |
|---------|-------|-------|---------|-------|-------|
| Say     | 349   | 745 634 | Sally   | 347   | 1 192 881 |
| Know    | 143   | 243 357 | Fillmore| 138   | 393 955  |
| Tell    | 65    | 53 904 | Gerald  | 62    | 108 528  |
| Allow   | 20    | 15 816 | Roville | 21    | 32 449   |
| Heaven  | 10    | 1 147 | Second  | 10    | 22 004   |
| Rugger  | 5     | 855   | Worthy  | 5     | 10 503   |
| Fish    | 4     | 174   | Spectator| 4     | 14 746   |
| Paper-Knife| 3 | 233   | Group   | 3     | 8 320    |
| Worship | 2     | 44    | Sell    | 2     | 8 346    |
| Thaw    | 1     | 11    | Price   | 1     | 8 295    |

engines [42]. However, there are numerous situations where the comparison with a general database is not available or is not interesting; for instance, when authorship has to be attributed without previous knowledge of texts written by the potential authors. Here, we approach these problems by taking advantage of the finding that words are unevenly distributed not only across documents but also within them [16–18, 21–24].

The quantification of the uneven distribution of words has been proposed based on measures commonly used by physicists [16, 24]. Following [21, 22, 24], we use the statistics of recurrence times, a standard quantification of intermittency or burstiness in time series [18, 19]. In texts, time is counted by the number of words and for each word $i$ the recurrence time $T_j$ is defined as the number of words between two successive occurrences of $i$ (the $j$ and $j + 1$ occurrence) plus one. For instance, the recurrence times for the word ‘the’ in the previous sentence are $T_1 = 9$ and $T_2 = 7$. A word that appears $N_i$ times in a text of size $N_T$ leads to a sequence of $N_T - 1$ inter-occurrence times $\{T_1, T_2, \ldots, T_{N_T-1}\}$. In order to incorporate also the time until the first $T_1$ and after the last $T_{N_T}$ occurrence of the word, we consider $T_N = T_1 + T_{N_T}$. In this case $\bar{T} = N_T/N_i$, where the overline denotes average over the different $T_j$’s. Note that the mean recurrence time $\bar{T}$ gives no additional information other than the frequency $N_i$. The intermittency of the word appears in the variance of $T_j$’s around $\bar{T}$ and can be quantified by

$I = \sigma_T/\bar{T}$, where $\sigma_T = \sqrt{\bar{T}^2 - \bar{T}^2}$. Randomly distributed words in the text have $I = 1$ (in the limit of large $N$ and small $N_i/N_T$), intermittent words have $I > 1$ and words appearing in regular intervals have $I < 1$. We calculate the intermittency measure $I_i = \sigma_T/\bar{T}$ for all words with $N_i \geq 5$ in each of the books (filtered texts) described in section 2.1. The words with $N_i < 5$ were considered to lack statistics and were disregarded.

In table 5, we compare words with highest $I = \sigma_T/\bar{T}$ and words with similar frequency. It is clear that the most intermittent words (largest $\sigma_T/\bar{T}$) are topical words (e.g. names of characters and locations), regardless of their frequency. Indeed, 15 out of the 16 most intermittent words are directly connected to specific characters. Similar behavior is observed in all books of our database. Intermittency is therefore a good characterization of topical words that in turn plays an
Table 5. In the book *The Adventures of Sally* by P G Wodehouse, there are $N_T = 15\,173$ words (tokens), 3657 different word types and 716 words with $N_i \geq 5$. The five words with highest $\sigma_T / \bar{T}$ are shown in the left part of the table. For comparison, in the right we show for each of these words another word with the closest frequency.

| Word   | $N_i$ | $I_i \equiv \sigma_T / \bar{T}$ | Word   | $N_i$ | $I_i \equiv \sigma_T / \bar{T}$ |
|--------|-------|---------------------------------|--------|-------|---------------------------------|
| Jules  | 26    | 4.31                            | Turn   | 26    | 1.55                            |
| Hobson | 31    | 4.09                            | Here   | 31    | 1.35                            |
| Ginger | 115   | 3.86                            | Get    | 117   | 1.24                            |
| Carmyle| 54    | 3.60                            | Feel   | 53    | 0.87                            |
| Bunbury| 20    | 3.59                            | People | 20    | 1.39                            |

Important role in the author-specific characteristic of the texts. The relationship between $\sigma_T / \bar{T}$ and the function of the words has been investigated in detail in [16–18, 21, 22, 24]. In the next section, we explore the fact that these properties are also author specific [17].

3. Evaluating the author dependence

3.1. From properties of words to properties of books

In the previous section, we introduced five quantities characterizing properties of words in the text: frequency ($N$), average shortest path length ($L$), betweenness ($B$), clustering coefficient ($C$) and intermittency ($I = \{ \sigma_T / \bar{T} \}$). The values of these quantities for all words in the books in our database can be found in SI (section 3), available from stacks.iop.org/NJP/13/123024/mmedia. We now analyze the global distribution of these measurements for all the words in a given book by plotting the empirical probability density function $\rho(X)$ for the measurements $X = \{ N, L, B, C, I \}$. Figure 2 shows the results for one book, and similar distributions were obtained for the other books. The shortest path $L$, clustering $C$ and intermittency $I$ have a well-defined peak and width (akin to a Gaussian distribution), but the frequency $N$ and betweenness $B$ have broad tail distributions (as in power-law distributions $\rho(X) \sim X^{-\alpha}$). The tail in $N$ corresponds to the well-known Zipf's law, which also appears in $B$ as expected from the large correlation between $B$ and $N$ ($\text{Corr}(B, N) = 0.95$ in the book of figure 2). With the two different types of behavior, we propose two sets of measurements, one for $X = \{ L, C, I \}$ and another for $X = \{ N, B \}$.

Our goal is to obtain quantities characterizing important features of these distributions to be used as global measurements of the books. The most natural choice is the average value $\langle X \rangle$, where $\langle \cdots \rangle \equiv \frac{1}{M} \sum_{i=1}^{M} \cdots$ corresponds to an average over the $M$ different words. For the network measures $L$, $C$ and $I$ this corresponds to the average values over nodes, a quantity considered as characteristic of the network [7, 12, 14, 43]. For $X = \{ N, B \}$, the high frequency words contribute strongly to $\langle X \rangle$ due to the long tails. To compensate for this effect, we consider also a modified average defined as $\langle X \rangle_2 \equiv \langle \log X \rangle$ for $X = \{ N, B \}$. For $X = \{ L, C, I \}$ the opposite is true, i.e. $\langle X \rangle$ is dominated by the large number of low-frequency words. Accordingly, we introduce a modified average as $\langle X \rangle_2 \propto \sum_i X_i \log N_i$, i.e. a weighted average with weights proportional to the logarithm of the frequency. The quantities $\langle X \rangle$ and $\langle X \rangle_2$ are expected to give
a good account of ‘typical’ values of $X$. However, in section 2 we mentioned that important information is conveyed by words with large $X$, i.e. in the tails of the distributions shown in figure 2. In order to characterize the fat-tail distributions of $X = \{N, B\}$, we used the coefficient $\alpha_X$ of a power-law fit to the tails of $\rho(X)$. An additional motivation for using $\alpha_N$ comes from the suggestion in [45] that it serves as a quantification of the style of texts. The large values of $X = \{L, C, I\}$ were characterized by calculating the skewness of $\rho(X)$, a measure of the asymmetry of the distribution. In summary, the three features we use for each of the five quantities $X = \{N, B, L, C, I\}$ are

Average value: $\langle X \rangle$ for $X = \{N, B, L, C, I\}$. 

Modified average: $\langle X \rangle_2 = \begin{cases} 
\langle \log(X) \rangle & \text{for } X = \{N, B\}, \\
\langle X \log N \rangle / \langle \log N \rangle & \text{for } X = \{L, C, I\}.
\end{cases}$ 

The fitting was performed to the cumulative distribution with logarithmic binning size, as suggested in [44]. A cut-off $X > 3 \times 10^3$ was used for $X = B$ (see figure 2(e)); no cut-off was used for $X = N$.

Figure 2. Probability density function $\rho(X)$ obtained from the different words of the book *The Adventures of Sally* by P G Wodehouse. (a) $X = L$ shortest path, (b) $X = C$ clustering coefficient, (c) $X = I$ intermittency, (d) $X = N$ frequency and (e) $X = B$ betweenness. In (d, e), the cumulative distribution $\rho(x \geq X) \equiv \int_x^\infty \rho(x)dx$ is shown, with the density $\rho(X)$ depicted in the inset. The legends indicate the features defined in equations (5)–(7) obtained for these distributions, and $\text{Cor}(X, N)$ indicates the Pearson correlation coefficient between $X$ and $N$ calculated over all words.
Right tail: \( \gamma(X) = \begin{cases} \alpha \text{ in } X^{-\alpha} & \text{for } X = \{N, B\}, \\ \text{skewness}(X) \equiv \left( \frac{X-\langle X \rangle}{\sigma_X} \right)^3 & \text{for } X = \{L, C, I\}. \end{cases} \) \hspace{1cm} (7)

These features are given in figure 2 for one book (see SI (section 3), available from stacks.iop.org/NJP/13/123024/mmedia, for all 40 books). Obviously, the choice of the quantities above is inevitably arbitrary. Our choice was intended to capture features of the distribution, rather than giving a parametric description of the full distribution. In particular, the power-law fit in equation (7) does not intend to fully describe the distributions, as is apparent in figures 2(d) and (e).

3.2. Machine learning methods and evaluation

In order to quantify the ability of the features described above to distinguish between authors, we employ machine learning algorithms which induce classifiers from a training database. The robustness of our results is tested with three widely used algorithms based on different principles. The first is known as C4.5 [46], and generates decision trees based on the information gained by each feature; the second algorithm is the Naive Bayes [47], which is based on the Bayes theorem; and the third and simplest algorithm is the nearest neighbor [48], which classifies an unknown instance according to the nearest neighbor of that instance in a normalized space involving all features. For more details, see SI (section 4), available from stacks.iop.org/NJP/13/123024/mmedia.

3.3. Efficiency of the classification

We consider the problem of distinguishing between eight authors, using five books to represent each author’s style. More specifically, each book described in section 2.1 was characterized by the set of 15 features discussed in section 3.1 (\( \langle X \rangle, \langle X \rangle^2 \) and \( \gamma(X) \) for \( X = \{N, B, L, C, I\} \)). The authorship assignment was performed using the algorithms in section 3.2 applied to a training dataset independent of the test book using the cross validation methodology (see SI (section 4)). This technique ensures that the training and evaluation sets are different and it is equivalent to assigning the authorship of one book in an experiment where four books of eight authors were used as a training dataset. The final output of the algorithms is the assignment of a specific author to each book tested, and the efficiency is quantified simply as the fraction of successful assignments.

The results are summarized in table 6 and indicate accuracy rates between 42.5 and 50.0% when all 15 features were used. These results were statistically significant by a large amount, confirming that these features successfully capture author-specific characteristics. To further explore the accuracy of different methods, we considered the cases in which only some of the features were included in the algorithms. We tested all \( 2^{15} = 32768 \) combinations of the 15 features and obtained a best result of 65.0% of correct assignments.

3.4. Relative importance of different features

In evaluating the importance of the different features for the final results, it is essential to identify their mutual dependence. We start from the list of all \( 2^{15} = 32768 \) combinations of features ordered by decreasing accuracy (as shown in SI (tables S1–S4)). We wish to quantify
Table 6. Accuracy rate achieved for the three machine learning algorithms using all 15 features and the best combination of these features. The accuracy is estimated based on 40 authorship assignments. The $p$-values correspond to the probability of getting by chance a higher or equal accuracy in one (all features) and in $2^{15} = 32768$ (the best case) trials. The features included in the best cases can be found in SI (tables S1–S4), available from stacks.iop.org/NJP/13/123024/mmedia.

| Algorithms        | Decision tree C4.5 | Nearest-neighbor kNN | Naive Bayes |
|-------------------|--------------------|----------------------|-------------|
| All 15 features   | 50.0% ($p = 1 \times 10^{-8}$) | 47.5% ($p = 6 \times 10^{-8}$) | 42.5% ($p = 2 \times 10^{-6}$) |
| The best case     | 62.5% ($p = 5 \times 10^{-9}$) | 65.0% ($p = 4 \times 10^{-10}$) | 62.5% ($p = 5 \times 10^{-6}$) |

when feature $y$ appears in the top of this list. To this end, we count the fraction of the $2^{14}$ feature combinations that include $y$ with accuracy higher than or equal to a threshold. The final estimate is then given by the area-under-the-curve of the ROC plot obtained by varying the threshold [49]. This procedure is equivalent to the Mann–Whitney $U$ test [50]. The motivation for using this method is that it evaluates the importance of a specific feature by taking into account how it combines with the other features to improve the accuracy of the prediction. The method depends both on the prediction algorithm and on the other features.

The features were ranked based on the method described above. The results for the three prediction algorithms are given in the first three columns of table 7. The three features appearing as the most prominent are $\langle N \rangle$ (average frequency), $\gamma (I)$ (skewness of intermittency) and $\langle N \rangle^2$ (average logarithmic frequency). In order to state the importance of features beyond a specific algorithm, it is important to quantify to what extent the results obtained for the three algorithms (the first three columns) are consistent with each other. To this end, we compute the Spearman’s rank correlation and obtain the values 0.29 ($p$-value = 0.145), 0.49 ($p$-value = 0.032) and 0.67 ($p$-value = 0.003) for the pairs C4.5/kNN, C4.5/Bayes and kNN/Bayes, respectively. The $p$-values are computed under the null hypothesis that the rankings are independent. Altogether, the three $p$-values indicate that the three rankings are consistent with each other. This is a strong indication that our analysis goes beyond algorithm-specific results and indeed captures the influence of the features.

It is interesting to compare the results with evaluations taking into account each feature separately. This can be done either by considering the accuracy of the prediction using only the specific feature or by comparing the information gained by including the feature [51]. This last method has the advantage of being independent of the prediction algorithm. These results are shown in the last four columns of table 7. Note that some features appearing as very important in multiple features analysis are not informative when taken alone (e.g. the skewness of the intermittency $\gamma (I)$). On the other hand, features that are well ranked in the single feature analysis do not always appear among the most important features when multiple features are considered (e.g. the weighted average of the clustering $\langle C \rangle_2$). These observations show the nontrivial mutual dependence of the features. To further explore this, we performed a factorial analysis (see SI (section 5), available from stacks.iop.org/NJP/13/123024/mmedia) using the 12 most important features in table 7, with the most important combinations being summarized in table 8. As expected from table 7, in fact $\gamma (I)$ appears among the best two combinations.
Table 7. Ranking of features based on the accuracy rate of the classifiers, where 1 in the table means best, 2 means second best and so on. The results for each classifier algorithm (C4.5, kNN and Bayes) are reported using different ranking procedures combined with multiple features (Mann–Whitney U test, columns 1–3, information gain (column 4) and accuracy using only one feature (columns 5–7)). The last column reports the Pearson correlation between each feature and the vocabulary size $M$ (the number of different words) calculated over the 40 books in our database. The features in the table are ordered according to the decreasing geometric mean of the ranks obtained in the three multiple features analysis (this ordering is the same achieved by considering for each feature the likelihood of reaching by chance a ranking as good as the one in each of the three ranking schemes). The areas under the curve in the multiple features analysis ranged between 56 and 69%.

| Feature | C4.5 | kNN | Bayes | Info | C4.5 | kNN | Bayes | Correlation with $M$ |
|---------|------|-----|------|------|------|-----|------|---------------------|
| $\langle N \rangle_2$ | 6 | 1 | 1 | 3 | 2 | 5 | 1 | $-0.90$ |
| $\gamma(I)$ | 2 | 2 | 2 | 10 | 12 | 9 | 10 | $-0.08$ |
| $\langle N \rangle$ | 1 | 6 | 3 | 2 | 1 | 2 | 3 | $-0.96$ |
| $\langle L \rangle$ | 7 | 4 | 6 | 9 | 5 | 3 | 8 | $0.85$ |
| $\langle B \rangle$ | 5 | 8 | 5 | 1 | 3 | 1 | 2 | $0.98$ |
| $\langle I \rangle_2$ | 10 | 3 | 10 | 15 | 15 | 12 | 12 | $-0.34$ |
| $\langle L \rangle_2$ | 8 | 7 | 8 | 8 | 5 | 7 | 5 | $0.85$ |
| $\langle C \rangle$ | 12 | 11 | 4 | 6 | 5 | 5 | 5 | $-0.87$ |
| $\gamma(L)$ | 4 | 13 | 11 | 13 | 10 | 9 | 13 | $-0.13$ |
| $\gamma(B)$ | 3 | 14 | 14 | 11 | 8 | 9 | 9 | $-0.07$ |
| $\langle B \rangle_2$ | 9 | 9 | 9 | 7 | 9 | 14 | 5 | $0.88$ |
| $\langle C \rangle_2$ | 11 | 10 | 7 | 5 | 4 | 3 | 4 | $-0.87$ |
| $\langle I \rangle$ | 13 | 5 | 12 | 12 | 13 | 15 | 10 | $-0.29$ |
| $\gamma(N)$ | 15 | 12 | 13 | 4 | 10 | 8 | 13 | $0.81$ |
| $\gamma(C)$ | 14 | 15 | 15 | 14 | 14 | 12 | 15 | $0.07$ |

of features in all three algorithms, which confirms that its effectiveness is correlated with its interdependence on other features.

4. Discussion and conclusions

4.1. Interpretation of the results

We are now in a position to use the word-specific analysis (section 2) and the distribution (section 3.1) of the quantities $X = \{N, B, L, C, I\}$ to assess the importance of the different features $\langle X \rangle$, $\langle X \rangle_2$ and $\gamma(X)$ in table 7:

$N$ Frequency. This was the most efficient quantity for recognizing authorship with $\langle N \rangle$ and $\langle N \rangle_2$ among the three most important features. Noting that $\langle N \rangle$ is proportional to the inverse number of distinct words $M$:

\[
\langle N \rangle = \frac{\text{length of book}}{M},
\]

New Journal of Physics 13 (2011) 123024 (http://www.njp.org/)
one infers that the distinguishing feature between the authors captured by \(\langle N \rangle\) is the different vocabulary sizes. The modified average \(\langle N \rangle_2 = \langle \log N \rangle\) also captures this aspect, including the proportion of frequent and infrequent words. On the other hand, the poor performance of \(\gamma(N) = \alpha\) in \(\rho(N) \sim N^{-\alpha}\) is a clear signature of the universal, author-independent character of Zipf's law (at a fixed book size [37]).

**B Betweenness.** The average betweenness \(\langle B \rangle\) was useful because of its strong correlation with the vocabulary size of the book \(M\) (the last column in table 7). In network terms, this corresponds to a linear relationship between \(\langle B \rangle\) and network size \((M\) is the number of nodes) and can be understood by noting that the number of terms in the sum of definition of \(B_i\) in equation (4) is proportional to \(M^2\), so that \(\langle B \rangle\) is expected to scale linearly with \(M\) for a fixed book size. The fact that \(\langle B \rangle_2\) and \(\gamma(B)\) had a poor performance indicates that the number of words with high betweenness is not a relevant feature to distinguish between authors.

**L Shortest path.** This was the network quantity with best performance. \(\langle L \rangle\) quantifies the typical distance of words to the central hubs of the network (frequent words). The good performance points to a dependence on the style of the authors. The poorer performance of \(\langle L \rangle_2\) and \(\gamma(L)\) indicates that the style dependence in \(L\) is more prominent in the typical values than, respectively, in the frequent and large \(L\) words.

**C Clustering.** The poor performance of all values related to this quantity suggests that authors have very little freedom in choosing the clustering of words co-occurrence networks. The last position in the ranking of \(\gamma(C)\) in table 7 suggests that the fraction of words used in specific contexts (high \(C\)) is author independent. The two averages \(\langle C \rangle_2\) and \(\gamma(C)\) take similar values (as seen in figure 2 and in SI (section 3), available from stacks.iop.org/NJP/13/123024/mmedia; recall the restriction \(N_i \geq 5\) used in section 2.2). They perform well only when used alone, possibly because of their correlation to vocabulary size \(M\).

**I Intermittency.** Apart from the frequency, intermittency was the most important quantity in table 7 with the skewness of the distribution \(\gamma(I)\) playing a prominent role. In view of the results in section 2.3, \(\gamma(I)\) may be interpreted as the fraction of all words that are topical or 'keyword like'. The poor performance of \(\langle I \rangle\) is not surprising since \(I\) is normalized by frequency \((I \equiv \sigma_i/\bar{T})\) and therefore \(\langle I \rangle \approx 1\) is expected. Indeed, from all five quantities the \(I\) features have shown altogether the smallest absolute value of correlation with vocabulary size (table 7), explaining why even \(\gamma(I)\) has a poor relative performance when used alone. Finally, \(\langle I \rangle_2\) performs better than \(\langle I \rangle\), suggesting that frequent words are more relevant.

---

**Table 8.** The two most relevant combinations of features, as revealed by a factorial analysis for the C4.5, kNN and Bayes classifier algorithms. As expected from table 7, \(\gamma(I)\) provides good results when used in conjunction with other features, such as \(\langle N \rangle\), \(\langle N \rangle_2\), \(\gamma(B)\) and \(\langle I \rangle_2\).

| C4.5          | kNN          | Bayes         |
|---------------|--------------|---------------|
| The first combination | \(\gamma(L), \langle C \rangle\) and \(\langle C \rangle_2\) | \(\gamma(I)\) and \(\gamma(B)\) | \(\langle N \rangle, \langle N \rangle_2\) and \(\langle B \rangle\) |
| The second combination | \(\gamma(I)\) and \(\langle N \rangle\) | \(\gamma(I)\) and \(\langle N \rangle_2\) | \(\langle I \rangle_2\) and \(\gamma(I)\) |
4.2. Comparison with other prediction methods

Even if the main goal of this paper is to evaluate the importance of different factors, it is also useful to compare the accuracy of our results with other methods of authorship attribution. Uzuner and Katz [52] used a database of books similar to ours, produced by eight authors. They used five sets of features, including simple statistics and more sophisticated syntactic analysis (table 3 of [52]). Our best results (accuracy of 65%) is comparable to their second best case obtained using ‘syntactic elements of expression’ (62%), being significantly worse only than their best result, achieved using function words (87%). In an extensive review, Grieve [31] reported accuracies obtained with a set of 34 features varying between 33% and 87% for the case of five authors, and between 18 and 80% for ten authors (table 9 of [31]). Our best results are above the median of their results achieved by using different features. Their best results again are based on the relative frequency of function words. These results are in accordance with the long tradition started by Mosteller and Wallace [27] to use the frequency of function words to distinguish between authors.

In order to confirm this in our database, we implemented a series of prediction schemes using the frequency $N_i$ of frequent (mostly function) words. In contrast with the approach described in this paper that used average and scaling properties as features, now the frequencies of specific words are used directly as input features of the prediction algorithms. We used only the kNN algorithm because the other algorithms did not provide good results when too many features were included. When the list of 70 stopwords from table 2.5 of [27] was used, we obtained an accuracy of 62.5%, i.e. comparable to our best results. Following the work [31], we considered two other lists of words: all 1978 words that appear in at least one book of each of the authors, leading to an accuracy of 90%; and all 209 words that appear at least once in every book in our database, leading to an accuracy of 82.5%. We recall that in order to concentrate on analysis that focuses on words with pronounced semantic content instead of function words, we have deliberately excluded a list of stopwords that contributed 80% of the cases listed in [27]. Therefore, our best combination of features compares well to other methods which demand more sophisticated syntactic analysis of the text. Measurements of complex network and intermittency are indeed able to capture many of the author-dependent characteristics.

In order to illustrate how measures analyzed in this paper can be complementary to traditional methods, we have performed a very simple experiment using as features the frequency and intermittency of the set of words composed of the five most frequent words in each book. The accuracy in classifying the books only by frequency was 72.5% and only by intermittency was 37.5%. Although this last accuracy rate is not impressive, it is statistically significant ($p = 2.2 \times 10^{-4}$) and shows that the intermittency values of specific words across distinct authors are different. The accuracy is increased to 80% when both features are included. It remains to be shown in future works how our results can improve state-of-the-art methods of authorship attribution.

4.3. Summary of conclusions

We have shown that the styles of different authors leave fingerprints in very general statistical measures of texts based on the network of co-occurrence of words and on intermittency or burstiness of words. The statistically significant scores obtained in authorship attribution unequivocally show that the style dependence of these features can be used in practice. Regarding the prominence of the different features, we note that both the results and ranking of
features may depend on the database, selected features and attribution algorithms. Accordingly, as emphasized in [31], different algorithms and features have to be tested in a given corpus before any real application of authorship attribution. However, the robustness of our results using three radically different attribution algorithms strongly suggests that the different features have importance that go beyond specific algorithms. Two features should be highlighted: (i) the skewness of the distribution of intermittent words $\gamma(I)$, which is based on the long-scale distribution of words and detects the extent to which topical words (keywords, large $I \equiv \sigma_I/T$) were used in the book; and (ii) the mean shortest path of the word co-occurrence network $\langle L \rangle$, which is based on the short-range connectivity of words and detects the typical distance of words to all other words. The different nature of these two quantities suggests a complementary role in capturing both short- and long-scale properties of the text, as well as typical and exceptional words.

Our focus in this paper was on the evaluation of the different features, rather than on maximizing the efficiency of the authorship attribution algorithms. This is apparent when comparing the best accuracy rates we achieved using our approach (62.5%) and using previous proposals (90.0%), as discussed in section 4.2. A further limitation of approaches based on intermittency and networks is that they require large pieces of text. While the root of the success of previous methods is the observation that function words are a powerful tool for detecting the style of authors [27, 31, 52], in the complex network and intermittency approaches used in this paper the focus is on content words. In this sense the results we achieve can be thought of as being complementary to the analysis using function words. More specifically, our results suggest that using $\gamma(I)$ and $\langle L \rangle$ can improve authorship recognition techniques when used in combination with the many different features currently employed [31]. Finally, the successful application of these measurements to characterize the style of authors suggests that the quantities discussed here can be further explored in other linguistic tasks, an approach that has been limited to a few works (see, e.g., [14, 53]).

Acknowledgments

We thank Stefan Siegert for valuable suggestions and assistance with the statistical methods and for a careful reading of the manuscript. This work was supported by FAPESP and CNPq (Brazil). DRA acknowledges support from the MPIPKS during his one-month visit to Dresden (Germany).

References

[1] Shannon C E 1948 Bell Syst. Tech. J. 27 379
[2] Solé R V 2010 Complexity 15 20
[3] Ferrer i Cancho R and Solé R V 2003 Proc. Natl Acad. Sci. USA 100 788
[4] Ferrer i Cancho R 2005 Physica A 345 275
[5] Ferrer i Cancho R and Solé R V 2001 Proc. R. Soc. B 268 2261
[6] Ferrer i Cancho R, Solé R V and Köhler R 2004 Phys. Rev. E 69 051915
[7] Antiqueira L, Nunes M G V, Oliveira O N Jr and Costa L F 2007 Physica A 373 811
[8] Amancio D R, Nunes M G V, Oliveira O N Jr, Pardo T A S, Antiqueira L and Costa L F 2011 Physica A 390 131
[9] Amancio D R, Antiqueira L, Pardo T A S, Costa L F, Oliveira O N Jr and Nunes M G V 2008 Int. J. Mod. Phys. C 19 583–98
[10] Antiqueira L, Oliveira O N Jr, Costa L F and Nunes M G V 2009 Inf. Sci. 179 584

New Journal of Physics 13 (2011) 123024 (http://www.njp.org/)
Choudhury M, Thomas M, Mukherjee A, Basu A and Ganguly N 2007 Proc. 2nd Workshop on TextGraphs: Graph-Based Algorithms for Natural Language Processing p 81688

Roxas R M and Tapang G 2010 Int. J. Mod. Phys. C 21 503

Stevanak J T, Larue D M and Lincoln D C 2010 arXiv:1007.3254

Antiqueira L, Pardo T A S, Nunes M G V, Oliveira O N Jr and Costa L F 2006 Proc. Workshop on Information and Human Language Technology

Masucci A P and Rodgers G J 2006 Phys. Rev. E 74 026102

Montemurro M and Zanette D 2002 Adv. Complex Syst. 5 7

Berryman M J, Allison A and Abbott D 2003 Fluctuation Noise Lett. 3 L1

Altmann E G, Pierrehumbert J B and Motter A E 2009 PLoS One 4 e7678

Goh K I and Barabasi A L 2008 Europhys. Lett. 81 48002

Allegreni P, Grigolini P and Palatella L 2004 Chaos Solitons Fractals 20 95

Herrera J P and Pury P A 2008 Eur. Phys. J. B 63 135

Carpena P, Berneola-Galván P, Hackenberg M, Coronado A V and Oliver J L 2009 Phys. Rev. E 79 3

Katz S M 1966 Nat. Lang. Eng. 2 15

Ortuño M, Carpena P, Berneola-Galván P, Muñoz E and Somoza A M 2002 Europhys. Lett. 57 759

Oakes M 2004 Proc. 5th Int. Conf. on Recent Advances in Soft Computing

Tankard W J Jr 2001 Applications of Computer Content Analysis (Westport, CT: Ablex Publishing) chapter 4

Mosteller F and Wallace D L 1963 J. Am. Stat. Assoc. 58 302

Havlin S 1995 Physica A 216 148–50

Vilenksy B 1996 Physica A 231 705–11

Yang A C C, Peng C K, Yien W K and Goldberger A L 2003 Physica A 329 473–83

Grieve J 2007 Lit. Linguist. Comput. 22 3

Ratnaparki A 1996 Proc. Empirical Methods in Natural Language Processing Conf.

Dorogovtsev S N and Mendes J F J 2001 Proc. R. Soc. B 268 2603

Barthelemy M, Barrat A, Pastor-Satorras R and Vespignani A 2005 Physica A 346 34

Costa L F, Sporns O, Antiqueira L, Nunes M G V and Oliveira O N Jr 2007 Appl. Phys. Lett. 91 054107

Zipf G K 1949 Human Behavior and the Principle of Least Effort (Reading, MA: Addison-Wesley)

Bernhardsson B 2007 Eur. Phys. J. B 58 167

Spackman K A 1989 Proc. 6th Int. Workshop on Machine Learning p 160

Mann H B and Whitney D R 1947 Ann. Math. Stat. 18 50660

Witten I H, Frank E and Hall M A 2011 Data Mining: Practical Machine Learning Tools and Techniques 3rd edn (Burlington, MA: Morgan Kaufmann)

Uzuner O and Katz B 2005 SIGIR Workshop on Stylistic Analysis of Text for Information

Paranyuhkin D 2010 Text Network Analysis http://issuu.com/deemeetree/docs/text-network-analysis (accessed June 2011)