



Universidade de São Paulo

Biblioteca Digital da Produção Intelectual - BDPI

Departamento de Física e Ciências Materiais - IFSC/FCM

Artigos e Materiais de Revistas Científicas - IFSC/FCM

2012

Identification of literary movements using complex networks to represent texts

NEW JOURNAL OF PHYSICS, BRISTOL, v. 14, n. 6, supl. 1, Part 2, pp. J217-J222, APR 23, 2012
<http://www.producao.usp.br/handle/BDPI/40906>

Downloaded from: Biblioteca Digital da Produção Intelectual - BDPI, Universidade de São Paulo

Identification of literary movements using complex networks to represent texts

This article has been downloaded from IOPscience. Please scroll down to see the full text article.

2012 New J. Phys. 14 043029

(<http://iopscience.iop.org/1367-2630/14/4/043029>)

View [the table of contents for this issue](#), or go to the [journal homepage](#) for more

Download details:

IP Address: 143.107.154.170

The article was downloaded on 21/03/2013 at 17:12

Please note that [terms and conditions apply](#).

Identification of literary movements using complex networks to represent texts

Diego Raphael Amancio¹, Osvaldo N Oliveira Jr
and Luciano da Fontoura Costa

Institute of Physics of São Carlos University of São Paulo, PO Box 369,
13560-970 São Carlos, São Paulo, Brazil

E-mail: diego.amancio@usp.br

New Journal of Physics **14** (2012) 043029 (15pp)


Received 15 February 2012

Published 23 April 2012

Online at <http://www.njp.org/>

doi:10.1088/1367-2630/14/4/043029

Abstract. The use of statistical methods to analyze large databases of text has been useful in unveiling patterns of human behavior and establishing historical links between cultures and languages. In this study, we identified literary movements by treating books published from 1590 to 1922 as complex networks, whose metrics were analyzed with multivariate techniques to generate six clusters of books. The latter correspond to time periods coinciding with relevant literary movements over the last five centuries. The most important factor contributing to the distinctions between different literary styles was the average shortest path length, in particular the asymmetry of its distribution. Furthermore, over time there has emerged a trend toward larger average shortest path lengths, which is correlated with increased syntactic complexity, and a more uniform use of the words reflected in a smaller power-law coefficient for the distribution of word frequency. Changes in literary style were also found to be driven by opposition to earlier writing styles, as revealed by the analysis performed with geometrical concepts. The approaches adopted here are generic and may be extended to analyze a number of features of languages and cultures.

 Online supplementary data available from stacks.iop.org/NJP/14/043029/mmedia

¹ Author to whom any correspondence should be addressed.

Contents

1. Introduction	2
2. Modeling texts as complex networks	2
2.1. Pre-processing	2
2.2. Complex networks measurements	3
3. The database	6
4. Results and discussion	6
5. Conclusion and further work	13
Acknowledgment	13
Appendix. Mathematical quantification of writing style	13
References	15

1. Introduction

Many findings related to language and culture issues have resulted from the use of statistical methods to treat large quantities of text [1–4]. Recent examples are the analysis of millions of books [1] and the study of twitter messages, where the global variation of mood could be observed through textual analysis of tweets [2]. In several such examples, knowledge is inferred from the analysis of the semantic content in the texts. There are also other methods for analyzing text, including cases where text is represented as a graph (or network) [5]. Of particular relevance was the finding that networks formed from texts are scale-free [6], and their topology can be analyzed, leading to various contributions. For instance, the scale-free structure (which is analogous to the Zipf’s law frequency distribution [7]) of text networks emerged as a consequence of an optimization process for both the reader and the writer, so that the effort at transmitting and obtaining a message was minimized [8]. In addition to allowing for cultural features to be identified and explored, automatic analysis may be useful also for real-world applications, such as automatic text summarization [9], machine translation [10, 11], authorship attribution [12], information retrieval [13] and search engines [14].

In this study, we used the topological metrics of complex networks representing text from 77 books dating from the period 1590–1922 in an attempt to verify changes in writing style. With multivariate statistical analysis of the metrics obtained, we were able to identify periods that correspond to major literary movements. Furthermore, we established the network characteristics responsible for the changes in writing style.

2. Modeling texts as complex networks*2.1. Pre-processing*

The modeling process starts by removing punctuation and words that convey little semantic content (see the supplementary information (SI), section 1, available from stacks.iop.org/NJP/14/043029/mmedia), such as articles and prepositions. Then, the remaining words are transformed into their canonical form, i.e. nouns and verbs are converted into the singular and infinitive forms, respectively. This step is performed using the MXPOST part-of-speech tagger [15], which assists in the resolution of ambiguities. The transformation to the

Table 1. Illustration of the pre-processing (removal of stopwords and punctuation marks) and lemmatization of the extract ‘My father’s family name being Pirrip, and my Christian name Philip, my infant tongue could make of both names nothing longer or more explicit than Pip.’ obtained from the book *Great Expectations* by Charles Dickens.

Original	Without stopwords	After lemmatization
My father’s family name Pirrip, and my, Christian name Philip my infant tongue could make of both names nothing longer or more explicit than Pip	father family name Pirrip Christian name Philip infant tongue could make both names longer more explicit Pip	father family name Pirrip Christian name Philip infant tongue can make both name long more explicit Pip

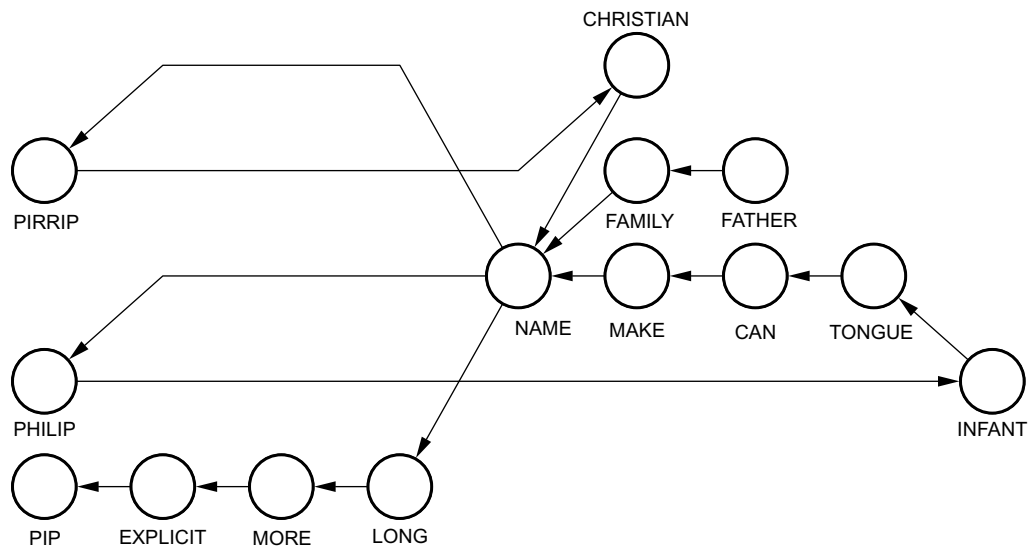


Figure 1. Network obtained from the extract ‘My father’s family name being Pirrip, and my Christian name Philip, my infant tongue could make of both names nothing longer or more explicit than Pip.’ from the book *Great Expectations* by Charles Dickens.

canonical form (lemmatization) is done to cluster words referring to the same concept into a single node of the network despite the differences in flexion. Finally, adjacent words in the written text are connected in the network according to the natural reading order (the left word is the source node and the right word is the target node). The modeling is demonstrated in table 1 for the pre-processing steps, while figure 1 illustrates the network obtained from a small extract from the book *Great Expectations* by Charles Dickens.

2.2. Complex networks measurements

Several metrics extracted from the networks were used to quantify the style of the books. From each local measurement (i.e. which refers to a node) we derived some quantities describing the

distribution of the networks in order to quantify the style of whole books. The measurements and their corresponding distribution descriptors were chosen because they have been useful in quantifying the style of texts in previous studies [12]. The simplest measurement refers to the number N of nodes in the network, which corresponds to the size of the vocabulary used to write the piece of text analyzed. The distribution of word frequency was characterized using the coefficient γ of the frequency distribution p_k :

$$p_k \sim ck^{-\gamma}, \quad (1)$$

where c is a normalization constant (see figure 2(a) for an example of the frequency distribution p_k of a specific book). We did not verify explicitly whether the degree obeys a power-law distribution because k is proportional to the frequency of words. Since the word frequency follows Zipf's law [16, 17], the degree is guaranteed to obey a power-law distribution². To compute γ , we employed a technique based on the accumulated distribution p_k (see figure 2(b)) described in [18]. We also used the frequency of words (or equivalently the degree k of the nodes) to calculate the assortativity Γ [19–21] (or degree–degree correlation) of the network as

$$\Gamma = \frac{\frac{1}{M} \sum_{j>i} k_i k_j a_{ij} - \left[\frac{1}{M} \sum_{j>i} \frac{1}{2} (k_i + k_j) a_{ij} \right]^2}{\frac{1}{M} \sum_{j>i} \frac{1}{2} (k_i^2 + k_j^2) a_{ij} - \left[\frac{1}{M} \sum_{j>i} \frac{1}{2} (k_i + k_j) a_{ij} \right]^2}, \quad (2)$$

where $M = 21\,900^3$ is the number of edges of the network, and $a_{ij} = 1$ if nodes i and j are connected and $a_{ij} = 0$ otherwise. If positive values are obtained for Γ , then highly connected nodes are usually connected to other highly connected nodes, indicating that regions may exist where nodes are highly interconnected [19]. Conversely, if Γ is negative, then highly connected nodes are commonly connected to slightly connected nodes.

In addition to measurements based on the number of nodes of the network and on the degree, the distance between concepts was employed to characterize the structure of the books. This measurement, widely known in the theory of networks as the average shortest path length l [22], is calculated from the distance d_{ij} , which represents the minimum cost (minimum number of edges) required to reach node j , starting from node i . After computing all pairs of values d_{ij} , the average shortest path length l_i of each node i is

$$l_i = \frac{1}{N-1} \sum_{j \neq i} d_{ij}. \quad (3)$$

Since l_i is defined for each node individually, the network is characterized by a distribution of l_i (see the distribution of l_i for a specific book in figure 2(c)). The distribution was characterized quantitatively by computing the average $\langle l \rangle$ and standard deviation Δl . Additionally, we computed the weighted average $(1/\sum k_i) \sum k_i l_i \equiv \langle l_w \rangle$, so that greater importance was given to the most frequent words in the text. The third moment $\zeta(l)$

$$\zeta(l) = \frac{1}{N} \sum_{i=1}^N \left(\frac{l_i - \bar{l}}{\Delta l} \right)^3 = \frac{1}{N(\Delta l)^3} \left(\sum_{i=1}^N l_i^3 - 3\bar{l} \sum_{i=1}^N l_i^2 + 2N\bar{l}^3 \right) \quad (4)$$

was also computed.

² The power-law distribution was verified for all texts of the database.

³ To avoid effects from the size of the books, for obtaining the complex network we used only the first $M + 1$ words of each book.

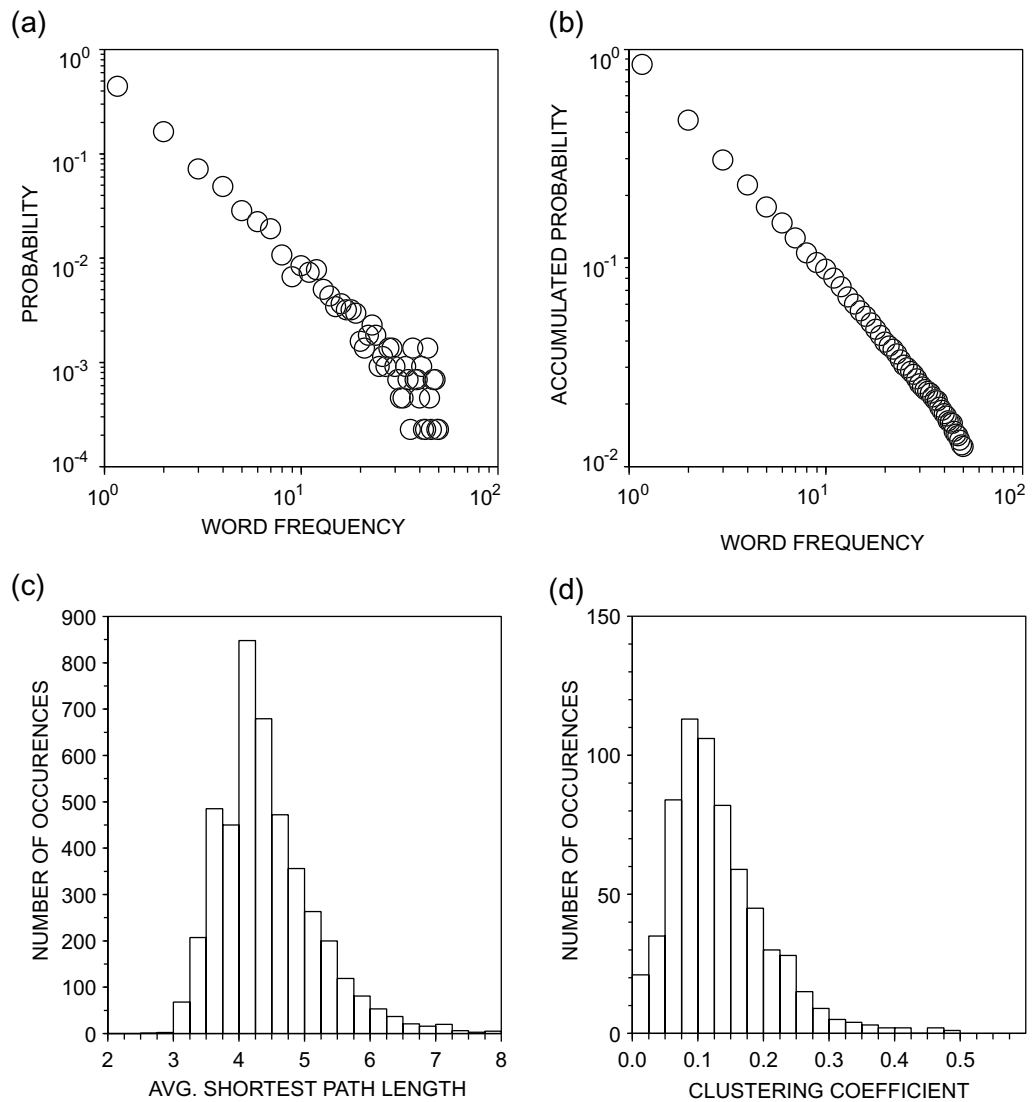


Figure 2. Example of distributions of measurements for the book *Great Expectations* by Charles Dickens. The measurements used were: (a) simple word frequency; (b) accumulated word frequency; (c) average shortest path length; and (d) clustering coefficient. The adjusted R -square found in panel (a) was 0.9348, which confirms that the frequency distribution is very similar to a power-law distribution.

The last metric was the clustering coefficient (C) [22], which quantifies the density of connections between the neighbors of a node i according to

$$C_i = \frac{3 \sum_{k>j>i} a_{ij}a_{ik}a_{jk}}{\sum_{k>j>i} a_{ij}a_{ik} + a_{ji}a_{jk} + a_{ki}a_{kj}}. \quad (5)$$

The clustering coefficient in equation (5) represents the fraction of the number of triangles among all possible connected sets of three nodes and therefore $0 \leq C_i \leq 1$. Similarly to the average shortest path length, it is also necessary to quantitatively characterize the

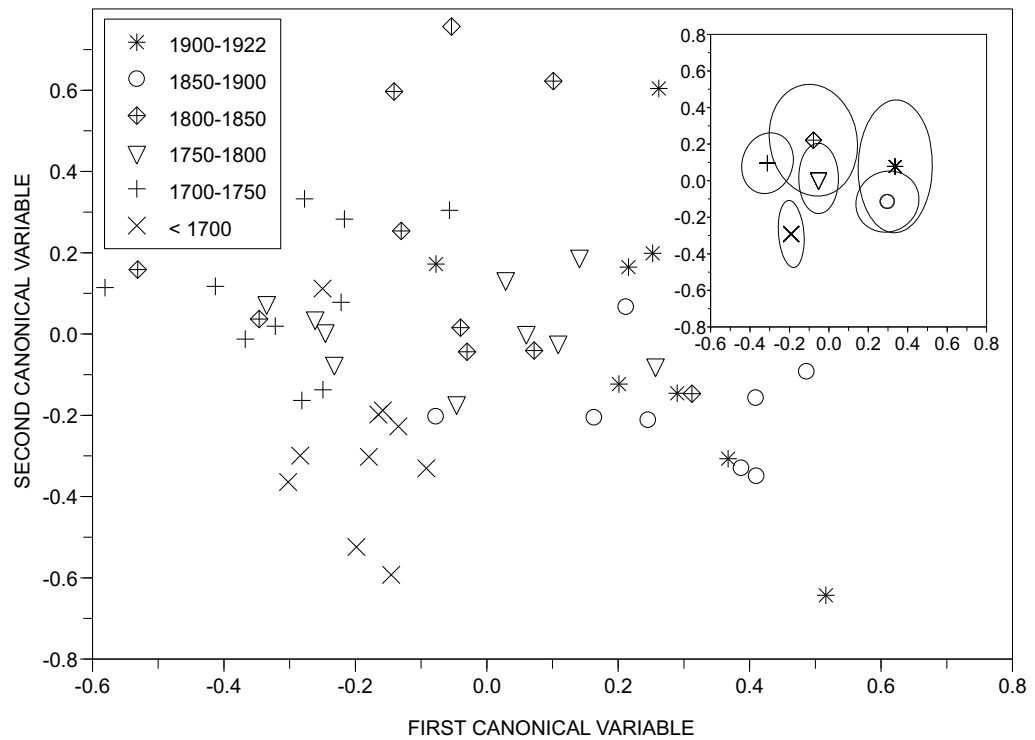


Figure 3. Scatter plot (CVA projection) representing the style of each book using six literary styles. Each style is represented by a set of ten books. The inset displays the dispersion of the literary styles.

distribution of the measurement (for an example of the distribution of C see figure 2(d)). We therefore computed the average $\langle C \rangle$, the standard deviation ΔC , the weighted average $(1/\sum k_i) \sum k_i C_i \equiv \langle C_w \rangle$ and the third moment $\zeta(C)$ to characterize the distribution.

3. The database

The database comprises 77 books available online at the Gutenberg project repository⁴; their publication dates ranged from 1590 to 1922. See tables S1–S3 in SI (section 2, available from stacks.iop.org/NJP/14/043029/mmedia) for the details of the books. The texts were represented with complex networks [8–11, 23–29], in which the edges are defined on the basis of co-occurrence of words (see section 2). The latter procedure has been proven to be suitable for quantifying both the style and structure of texts (see, e.g. [11, 25, 28]). The details of the procedures adopted to model texts as complex networks and a description of the measurements employed to characterize the networks can be found in section 2.

4. Results and discussion

The evolution of literary styles was quantified considering the 11 measurements from complex networks described in section 2.2 for the books from the Project Gutenberg (see footnote 4).

⁴ <http://www.gutenberg.org/>

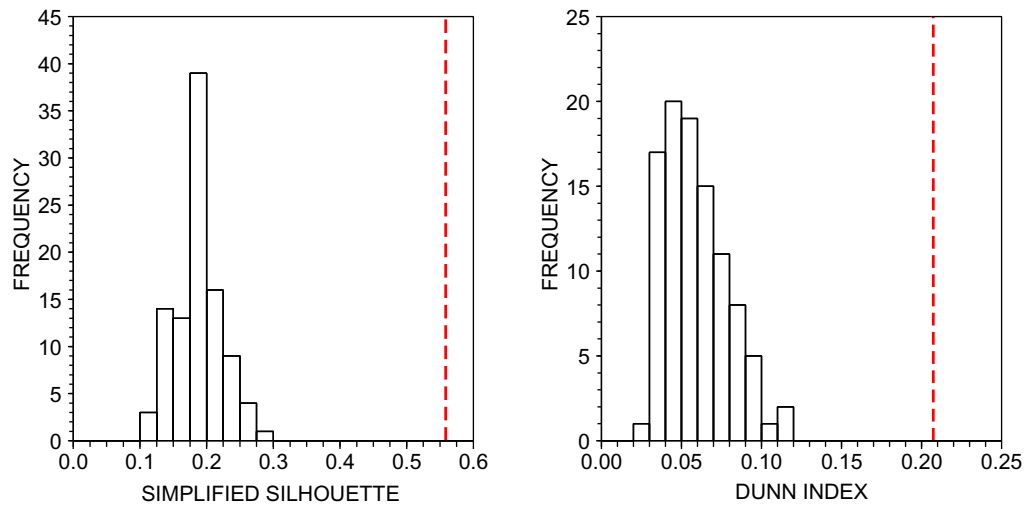


Figure 4. Significance test performed for (a) the simplified silhouette and (b) the Dunn index. The histograms represent the values of the cluster quality indices considering a random distribution of points, while the dotted lines represent the clustering quality indices obtained for the clustering illustrated in figure 5. Because the silhouette for the random case $SWC_{\text{rand}} = 0.187 \pm 0.036$ is smaller than the silhouette $SWC = 0.558$ for the clustering of figure 5, the clustering inferred is significant. The same applies for the Dunn index because $DN_{\text{rand}} = 0.059 < DN = 0.207$.

The main measurements were the shortest path length (l), the clustering coefficient (C), the assortativity (Γ), the power-law coefficient of the degree distribution (γ) and the size of the vocabulary (N). An initial, arbitrary division of the books into six intervals of 50 years, according to their publication date, led to the clusters shown in the canonical variate analysis (CVA; for details see SI, section 3, available from stacks.iop.org/NJP/14/043029/mmedia) plot in figure 3. The distinction was relatively poor, especially considering the standard variation ellipses [30] in the inset of the figure. Good separation was only possible when distant periods in time were compared, as their ellipses did not overlap. This difficulty in distinguishing literary movements should perhaps be expected as there is no reason for sharp transitions to occur only because half-century marks were reached. We also verified the distinguishability of clusters with principal component analysis (PCA; see SI, section 3), but the distinction was also poor.

In order to verify whether books from distinct publication dates could be distinguished at all, we adopted a systematic procedure for the partition of the dataset using an optimization approach. This was performed by assessing the quality of the clustering under the condition that books with consecutive publication dates should either belong to the same cluster or lie in the boundaries of consecutive clusters. More specifically, we varied the delimiters and the number of clusters in the database and quantified the quality of the clustering using two indices, namely the simplified silhouette (SWC) and the Dunn index (DN) (see SI, section 4, available from stacks.iop.org/NJP/14/043029/mmedia). Good distinction between writing styles was obtained for 3, 4, 5, 6 and 7 clusters (see figure S1 in SI), according to the two indices (SWC and DN). The best partition, which was found to be statistically significant (see figure 4), was obtained with SWC and CVA projection, leading to the six clusters in figure 5, where there is almost no

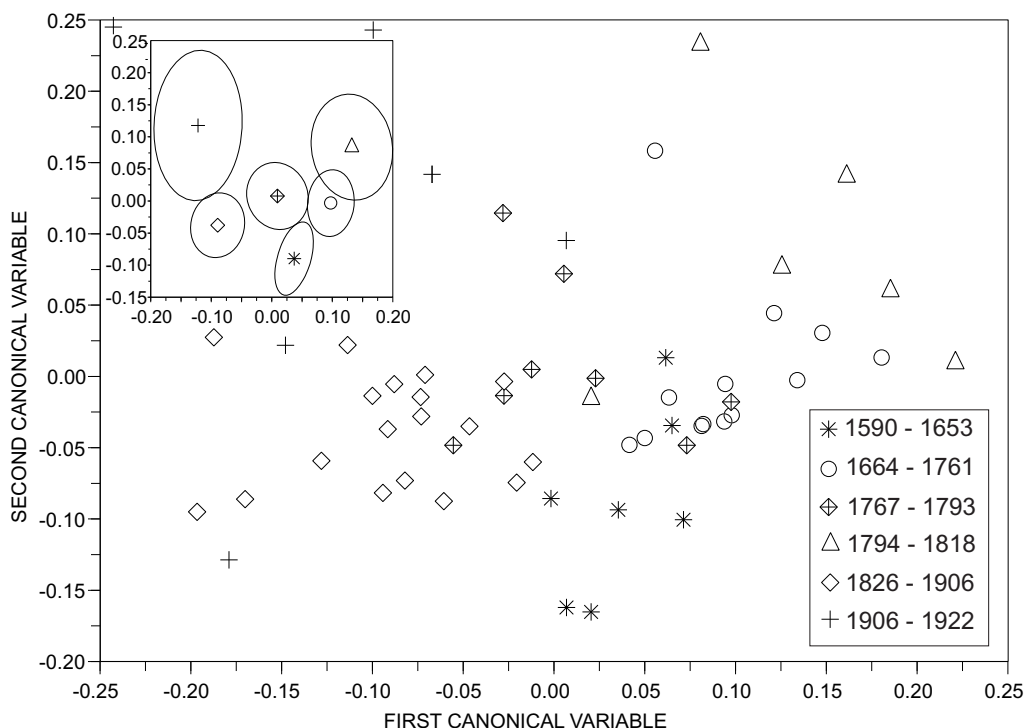


Figure 5. Scatter plot representing the best clustering considering the writing style. Note that besides being a good partitioning scheme, it also keeps a good representation of the original database, since 82% of the variance are kept in the CVA projection.

Table 2. Relationship between the best clustering of writing styles and the traditional classification of literary movements.

Cluster boundary	Literary boundary	Literary movement	Reference
1590–1653	1558–1603	Elizabethan era	[32]
1664–1761	1660–1798	Neoclassicism/Enlightenment	[33–35]
1767–1793	1660–1798	Neoclassicism/Enlightenment	[33, 36]
1794–1818	1764–1820	Gothic fiction	[33, 36]
1826–1906	1830–1900	Realism	[33]
1826–1906	1865–1900	Naturalism	[33, 37]
1906–1922	1890–1940	Modernism	[33, 38]

overlap among clusters, as shown in the inset. Most significantly, the six time periods inferred from this analysis coincide with the well-established literary movements listed in table 2.

Other important features can be inferred from figure 5. Firstly, clusters for subsequent time periods are normally placed next to each other, indicating smooth changes in writing style over time. The same conclusion can be drawn from the analysis of the hierarchical clustering in figure 6 with the Ward [31] distance. The exception to this trend was the major change from the 1794–1818 \rightarrow 1826–1906 period, which may be the consequence of a drastic change in style

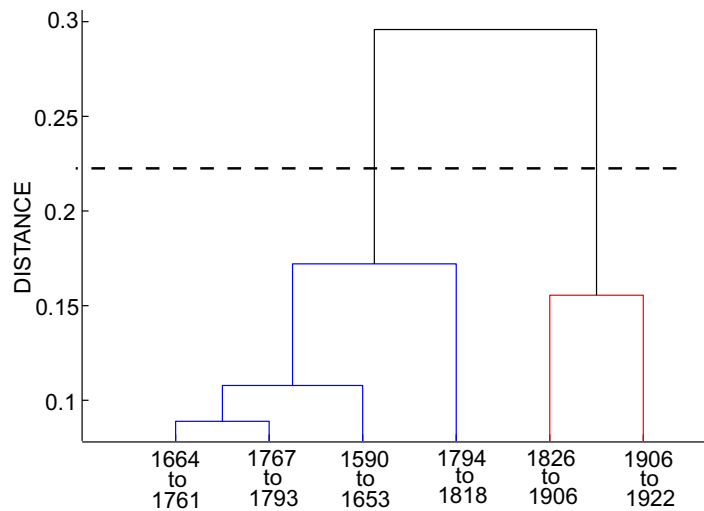


Figure 6. Hierarchical relationship between literary periods using the Ward linkage strategy. The two groups after the division performed with a particular threshold (dotted line) correspond to the oldest and newest books.

triggered by the French Revolution (1789). As to the variance among clusters, the lowest and highest values applied to the 1590–1653 and 1906–1922 periods, respectively. These results are intuitive as few changes in style could be expected in older periods, while in recent periods less uniformity could be the result of the coexistence of many writing styles.

The most important factors contributing to the separation of literary styles were determined in two distinct ways. The first technique considered a feature to be relevant if it was capable of providing significant distinction between groups, regardless of the other features. The list of metrics and the corresponding p -value for the difference of a given measurement between pairs of clusters are given in table 3. The asymmetry in the distribution of the average shortest path length $\zeta(l)$ and the vocabulary size N exhibited the most significant variations. Interestingly, similar results were reported in [12], where these two measurements were also useful in characterizing personal writing styles. In the second evaluation, a feature was considered relevant if it was able to provide good distinction between groups based on the interdependences of features. This evaluation was carried out by computing the importance of each measurement for the axes in the CVA plots. The results of tables 4 and 5 point to the clustering coefficient (C and C_w) as the main factor for the distinction into six clusters. Since there is evidence that the clustering coefficient quantifies whether words are restricted to specific or generic contexts (an explanation of this property is given in [12])⁵, it seems that the extent of the use of generic or specific words varied across history. This change has not been monotonic, as indicated in figure 7(a). In fact, most of the network measurements fluctuated over time, including the size of the vocabulary, whose considerable change was responsible for the most drastic transition, from the 1794–1818→1826–1906 periods. This is clearly seen from figure 7(b). The only metric with

⁵ Context-specific restricted words are those appearing in only a few contexts. For example, the concept ‘teacher’ usually induces concepts related to the learning environment. On the other hand, generic words may appear in a myriad of situations. Examples are ‘red’ (red car, red wall or red skin) and ‘identical’ (identical behaviors, identical grades or identical plates).

Table 3. List of the most significant transitions. Taken individually, the most prominent measurements for discriminating between clusters are the size of the vocabulary N and the third moment of the average shortest path length $\zeta(L)$.

Measurement	Feature	Transition	p -value
Vocabulary	N	1590–1653 \rightarrow 1794–1818	0.048
	N	1664–1761 \rightarrow 1767–1793	0.051
	N	1664–1761 \rightarrow 1826–1906	0.001
	N	1767–1793 \rightarrow 1794–1818	0.011
	N	1794–1818 \rightarrow 1826–1906	$<1.0 \times 10^{-3}$
Assortativity	Γ	1590–1653 \rightarrow 1767–1793	0.008
	Γ	1590–1653 \rightarrow 1826–1906	0.044
	Γ	1664–1761 \rightarrow 1767–1793	0.041
	Γ	1664–1761 \rightarrow 1826–1906	0.006
Shortest path	$\langle l \rangle$	1664–1761 \rightarrow 1826–1906	0.049
	$\langle l_w \rangle$	1664–1761 \rightarrow 1906–1922	0.050
	ΔL	1590–1653 \rightarrow 1906–1922	0.031
	ΔL	1664–1761 \rightarrow 1906–1922	0.022
	ΔL	1767–1793 \rightarrow 1906–1922	0.023
	ΔL	1826–1906 \rightarrow 1906–1922	$<1.0 \times 10^{-3}$
	$\zeta(l)$	1590–1653 \rightarrow 1826–1906	0.028
	$\zeta(l)$	1590–1653 \rightarrow 1906–1922	$<1.0 \times 10^{-3}$
	$\zeta(l)$	1664–1761 \rightarrow 1906–1922	$<1.0 \times 10^{-3}$
	$\zeta(l)$	1767–1793 \rightarrow 1906–1922	0.001
Clustering	$\zeta(l)$	1794–1818 \rightarrow 1906–1922	0.019
	$\zeta(l)$	1826–1906 \rightarrow 1906–1922	$<1.0 \times 10^{-3}$
	$\langle C \rangle$	1664–1761 \rightarrow 1767–1793	0.048
	$\langle C \rangle$	1664–1761 \rightarrow 1826–1906	0.051
	$\langle C_w \rangle$	1664–1761 \rightarrow 1767–1793	0.054
	$\langle C_w \rangle$	1664–1761 \rightarrow 1826–1906	0.055
	ΔC	1664–1761 \rightarrow 1767–1793	0.054
	$\zeta(C)$	1590–1653 \rightarrow 1767–1793	0.045

a well-defined trend over time was the coefficient of the power law for the scale-free networks representing the texts. The decreasing trend in figure 7(c) points to a smoother, and therefore more uniform, frequency distribution, which means that the difference in frequency between low- and high-frequency words decreased with time.

The changes in style between any two consecutive clusters appeared to have been driven by opposition [39] (see appendix), which quantifies the extent to which the current period can be thought of as an opposite movement to the previous literary movements. The coefficient satisfies the inequality $W_{ij} > 0$, with the exception of the 1826–1906 \rightarrow 1909–1922 transition. Furthermore, the opposition movement was more significant than the skewness movement s_{ij} (see appendix), which quantifies how much the change in the current style deviates from the opposition movement. The results are given in table 6. In other words, the innovation of style (\vec{v}_i , see the definition in appendix) was generally driven by contrasting the previous styles (\vec{a}_i , see the definition in appendix). As to the dialectics ρ_{ijk} (see appendix), which quantifies how the

Table 4. Importance of each measurement for the first canonical variable, where the clustering coefficient C and the average shortest path length l were the most prominent.

Measurement (first axis)	Prominence (first axis)
$\langle C_w \rangle$	33.3%
$\langle C \rangle$	31.6%
ΔC	6.6%
$\langle l \rangle$	6.4%
Γ	5.1%

Table 5. Importance of each measurement for the second canonical variable, where the clustering coefficient C and the average shortest path length l were the most prominent.

Measurement (second axis)	Prominence (second axis)
$\langle C \rangle$	34.5%
$\langle C_w \rangle$	33.7%
$\langle l_w \rangle$	9.5%
$\langle l \rangle$	9.4%
ΔC	3.4%

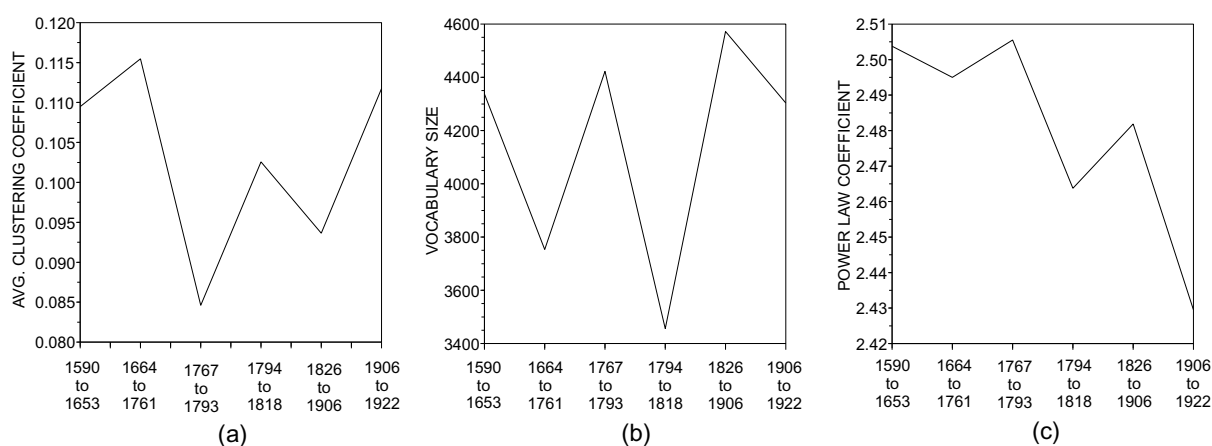


Figure 7. Dynamics of (a) the average clustering coefficient, (b) vocabulary size and (c) coefficient of the power law. While the clustering coefficient and vocabulary size oscillate throughout the periods, the coefficient of the power law tends to decrease, which shows that words were used in a more uniform way in the later periods.

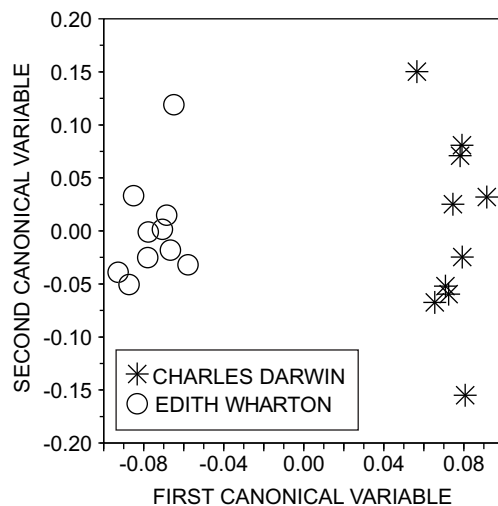
current movement i is an implication of the two previous movements j and k , no clear pattern could be identified in table 7. The lowest ρ_{ijk} (and therefore the highest dialectics) appeared during the 19th century. Thus, realism is a literary style that better approximates as a synthesis of the two previous literary periods.

Table 6. Opposition (W_{ij}) and skewness (s) indices.

Period	W_{ij}	s_{ij}
1590–1653 \rightarrow 1664–1761	1.00	0.00
1664–1761 \rightarrow 1767–1793	0.39	0.08
1767–1793 \rightarrow 1794–1818	0.35	0.18
1794–1818 \rightarrow 1826–1906	1.09	0.07
1826–1906 \rightarrow 1909–1922	-0.01	0.08

Table 7. The counter-dialectics index ρ_{ik} .

Period	ρ_{ik}
1590–1653 \rightarrow 1664–1761 \rightarrow 1767–1793	0.76
1664–1761 \rightarrow 1767–1793 \rightarrow 1794–1818	1.49
1767–1793 \rightarrow 1794–1818 \rightarrow 1826–1906	0.39
1794–1818 \rightarrow 1826–1906 \rightarrow 1909–1922	0.69

**Figure 8.** Comparison of Darwin's and Edith Wharton's styles with CVA projection. A good separation can be observed, indicating that these two authors had quite different styles.

In subsidiary studies, we verified that the complex network metrics used are indeed efficient in distinguishing styles. For this we examined the writing style dynamics of 10 books⁶ of Charles R Darwin (1809–1882) and Edith Wharton (1862–1937), whose styles are known to differ considerably. Indeed, this is confirmed in the CVA plot in figure 8, where again the most significant contributing factor for distinction was the clustering coefficient C , since both $\langle C \rangle$ and $\langle C_w \rangle$ are responsible for 44% of the weights in the first canonical variable axis.

⁶ The list of books is shown in table S3 in SI, section 2 (available from stacks.iop.org/NJP/14/043029/mmedia).

5. Conclusion and further work

We were able to study changes in writing style objectively by analyzing the metrics from complex networks representing texts from books published over several centuries. Significantly, the most appropriate clustering of books matched the traditional literary classification, with the most significant contributing factor for distinguishability being the average shortest path length. We found that it is possible to distinguish between literary movements using only the vocabulary size or the asymmetry of the average shortest path length distribution. Innovation in writing style was found to be driven mainly by opposition, with a growing trend of literary development toward counter-dialectics. Interestingly, these findings represent the generalization of previous results where a dependence was established between network topology and the style of machine translations [10, 11] and the style of authors [12]. We believe that the approach used here may be useful in studying the evolution of any system of interest, since the basic concepts (i.e. characterization through features and use of time series) are completely generic.

In future work, we plan to employ additional complex network measurements in a larger database to verify if the discrimination can be improved further. We shall also examine the relationship between semantics and topology, by generating clusters using the semantics of words to be compared with the clusters obtained from the analysis of network topology. A more challenging endeavor will be to extend the study to other languages, in order to probe whether the patterns revealed in this paper can be generalized.

Acknowledgment

The authors are grateful to FAPESP (2010/00927-9 and 2011/50761-2) for financial support.

Appendix. Mathematical quantification of writing style

In this appendix, we quantify mathematically the variation of writing style. To quantify the change in style over time, we used three concepts, namely *opposition index*, *skewness index* and *counter-dialectics index*, which depend on the measurements computed in each step of the temporal series. For each element i of the temporal series, which represents the value for the measurements described in section 2.2, we defined the 11-dimensional (11D) vector \vec{v}_i :

$$\vec{v}_i = \left[N \ \Gamma \ \gamma \ \langle C \rangle \ \langle C_w \rangle \ \Delta C \ \zeta(C) \ \langle l \rangle \ \langle l_w \rangle \ \Delta l \ \zeta(l) \right]^T. \quad (\text{A.1})$$

The large number of data generated were visualized by projecting \vec{v}_i into a 2D space before computing the indices, and this also helped us to remove undesirable correlations. The projection techniques used are described in SI, section 3 (available from stacks.iop.org/NJP/14/043029/mmedia). Using the projected \vec{v}_i and considering t elements in the time series, \vec{a}_i was defined in the average state at time i , $i \leq t$, as:

$$\vec{a}_i = \frac{1}{i} \sum_{j=1}^i \vec{v}_j. \quad (\text{A.2})$$

Given \vec{a}_i , the *opposite state* of the current state i (see figure A.1(a)) for a geometrical interpretation) is given by

$$\vec{r}_i = \vec{v}_i + 2(\vec{a}_i - \vec{v}_i) = 2\vec{a}_i - \vec{v}_i, \quad (\text{A.3})$$

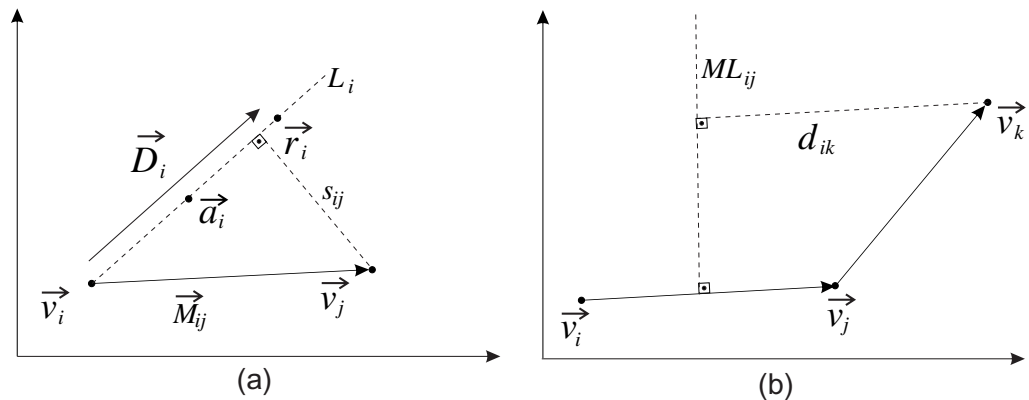


Figure A.1. Illustration of the quantities employed to define the *opposition*, *skewness* and *counter-dialectics* indices.

and given \vec{r}_i and \vec{v}_i , the *opposition vector* \vec{D}_i of state \vec{v}_i (see figure A.1(a)) is given by

$$\vec{D}_i = \vec{r}_i - \vec{v}_i. \quad (\text{A.4})$$

For two consecutive books i and j , the vector representing the style change \vec{M}_{ij} (see figure A.1(a)) is

$$\vec{M}_{ij} = \vec{r}_i - \vec{v}_i. \quad (\text{A.5})$$

The vector \vec{M}_{ij} is important because its norm $\|\vec{M}_{ij}\|$ quantifies the change in style in relation to the previous state \vec{v}_i . With \vec{M}_{ij} , the *opposition index* W_{ij} is the component of \vec{M}_{ij} over \vec{D}_i :

$$W_{ij} = \frac{\vec{M}_{ij} \cdot \vec{D}_i}{\|\vec{D}_i\|^2}. \quad (\text{A.6})$$

If the current style tends to oppose the previous one, then the component of \vec{M}_{ij} over \vec{D}_i will have a high value. This quantifier is useful, for example, in identifying little stylistic innovation: if opposite movements are repeated over and over again, then there is no innovation at all.

The *skewness index* s_{ij} , which is depicted in figure A.1(a), is defined as the distance between \vec{v}_j and the line defined by \vec{D}_i . This index quantifies how far the stylistic movement is from the opposite movement. It is useful in identifying trivial oscillations within the line L_i , for in this case a series of movements with zero *skewness index* would be observed.

The dialectics between three consecutive styles i , $j = i + 1$ and $k = j + 1 = i + 2$ in the temporal series was quantified as follows. If \vec{v}_k is the outcome of a synthesis of the styles represented by \vec{v}_i and \vec{v}_j , then the distance d_{ik} between \vec{v}_k and the middle line ML_{ij} defined by \vec{v}_i and \vec{v}_j (see figure A.1(a)) will be small. The *counter-dialectics index*⁷ ρ_{ik} is

$$\rho_{ik} = \frac{d_{ik}}{\|\vec{M}_{ij}\|}. \quad (\text{A.7})$$

Further details of the definition of the opposition W_{ij} , skewness s_{ij} and counter-dialectics ρ_{ik} can be found in [39].

⁷ Note that we referred to ρ_{ik} as the *counter-dialectics index* instead of the *dialectics index*, because it is defined as a distance. Hence, there is an inverse proportion between ρ_{ik} and the concept of dialectics.

References

- [1] Michel J B *et al* 2011 *Science* **331** 176
- [2] Golder S A and Macy M W 2011 *Science* **333** 1878
- [3] Evans J A and Foster J G 2011 *Science* **331** 721
- [4] Bohannon J 2011 *Science* **330** 1600
- [5] Newman M E J 2003 *SIAM Rev.* **45** 167
- [6] Barabási A L 2009 *Science* **325** 412–3
- [7] Costa L F, Sporns O, Antiqueira L, Nunes M G V and Oliveira O N Jr 2007 *Appl. Phys. Lett.* **91** 054107
- [8] Ferrer i, Cancho R and Solé R V 2003 *Proc. Natl Acad. Sci. USA* **100** 788
- [9] Antiqueira L, Oliveira O N Jr, Costa L F and Nunes M G V 2009 *Inf. Sci.* **179** 584
- [10] 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
- [11] 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
- [12] Amancio D R, Altmann E G, Oliveira O N Jr and Costa L F 2011 *New J. Phys.* **13** 123024
- [13] Boginski V L 2005 Optimization and information retrieval techniques for complex networks *PhD Thesis* University of Florida
- [14] Page L, Brin L, Motwani R and Winograd T 1999 The PageRank citation ranking: bringing order to the Web *Technical Report* Stanford InfoLab
- [15] Ratnaparki A 1996 *Proc. Empirical Methods in Natural Language Processing Conf.* (University of Pennsylvania, Philadelphia, Pa, USA) pp 133–42
- [16] Manning C D and Schütze H 1999 *Foundations of Statistical Natural Language Processing* (Cambridge, MA: MIT)
- [17] Zipf G K 1949 *Human Behavior and the Principle of Least Effort* (Reading, MA: Addison-Wesley)
- [18] Bauke H 2007 *Eur. Phys. J. B* **58** 167
- [19] Newman M E J 2002 *Phys. Rev. Lett.* **89** 208701
- [20] Newman M E J 2003 *Phys. Rev. E* **67** 026126
- [21] Newman M E J 2006 *Phys. Rev. E* **74** 036104
- [22] Newman M E J 2010 *Networks: An Introduction* (Oxford: Oxford University Press)
- [23] Ferrer i, Cancho R and Solé R V 2001 *Proc. R. Soc. B* **268** 2261
- [24] Solé R V, Corominas-Murtra B, Valverde S and Steels L 2010 *Complexity* **15** 20
- [25] Stevanak J T, Larue D M and Lincoln D C 2010 arXiv:1007.3254
- [26] Ferrer i, Cancho R, Solé R V and Köhler R 2004 *Phys. Rev. E* **69** 051915
- [27] Antiqueira L, Nunes M G V, Oliveira O N Jr and Costa L F 2007 *Physica A* **373** 811
- [28] Roxas R M and Tapang G 2010 *Int. J. Mod. Phys. C* **21** 503
- [29] Masucci A P and Rodgers G J 2006 *Phys. Rev. E* **74** 026102
- [30] Lee J and Wong D W S 2000 *Statistical Analysis with ArcView GIS* (New York: Wiley)
- [31] Ward J H 1963 *J. Am. Stat. Assoc.* **58** 236
- [32] http://en.wikipedia.org/wiki/Elizabethan_era
- [33] <http://sparkcharts.sparknotes.com/lit/literaryterms/section5.php>
- [34] <http://en.wikipedia.org/wiki/Neoclassicism>
- [35] http://en.wikipedia.org/wiki/Age_of_Enlightenment
- [36] http://en.wikipedia.org/wiki/Gothic_fiction
- [37] [http://en.wikipedia.org/wiki/Naturalism_\(literature\)](http://en.wikipedia.org/wiki/Naturalism_(literature))
- [38] <http://en.wikipedia.org/wiki/Modernism>
- [39] Fabbri R, Oliveira O N Jr and Costa L F 2010 arXiv:1010.1880