Martins *et al. Appl Netw Sci* (2021) 6:50
https://doi.org/10.1007/s41109-021-00390-7

Applied Network Science

RESEARCH Open Access

# *Historia Augusta* authorship: an approach based on Measurements of Complex Networks

Armando Martins[1,7], Clara Grácio[2,3*], Cláudia Teixeira[1,8], Irene Pimenta Rodrigues[4,5], Juan Luís Garcia Zapata[6] and Lígia Ferreira[4,5]

*Correspondence: mgracio@uevora.pt
[3] CIMA, Évora, Portugal
Full list of author information is available at the end of the article

**Abstract**

In this work, we analyze in detail the topology of the written language network using co-occurrence of words to recognize authorship. The Latin texts object of this study are excerpts from *Historia Augusta*, a collection of biographies of Roman emperors extending from Hadrian, who started to reign in 117 CE, to Carus and his sons Numerian and Carinus, that is, to the years up 284–285 CE. According to the manuscript tradition, the biographies are attributed to six different authors. Scholarship since the late 19th century has been arguing for a single authorship instead. The aim of this paper is to verify this hypothesis.

**Keywords:** Latin text Authorship, Complex networks, Topological parameters, Co-ocurrence networks, Spectral clusterings

## Introduction

The recognition of the authorship and the literary style of a writer have been frequent subjects of investigation. Today, a new approach to representing and modeling complex systems has gained strength and proven powerful: complex networks. They have already modeled many real systems from the internet to the human body. Words are a good example of simple elements that combine to form complex structures such as novels, poems, dictionaries and manuals that are designed to transport or convey information. The written human language is one of the most important examples of a self-organizing system in nature. Historically, the beginning of quantitative analysis of natural language is usually associated with the name of G.K. Zipf, 1949, who was the first to carry out an extensive study of word frequencies in written texts in a few different languages.

Several topological metrics and parameters of the complex networks, extracted from the determined adjacency matrix, are calculated using spectral clustering tools we have developed based on the second eigenvector of the Laplacian graph matrix (Rocha et al. 2015). Some parameters are Degree Density, Betweenness Centrality, Graph Diameter, Graph Radius, Degree Centrality, Cliques, Graph Assortativity, Clustering Coefficient, and Average Shortest Path. This technique avoids the high cost of combinatorial algorithms, using instead linear algebra numerical methods, well established in scientific computing. In this case, the detection of communities identifies marks of style, which

allows us to consider that the attribution of the various texts to each of the authors will not be correct. We use a different computational method to verify the authorship of texts, General Imposters (Koppel and Winter 2014) in Stylo R package (Eder et al. 2016). In order to evaluate our method, we used as control texts the *Res Gestae* by *Ammianus Marcellinus*, and Portuguese texts by José Saramago, Mia Couto and Lobo Antunes.

*Historia Augusta* is a late Roman collection of biographies of Roman Emperors, Caesars and usurpers, covering the period from Hadrian to Carus, Carinus and Numerianus, with a gap spanning the years 244 to 253 CE. Traditionally, the work is attributed to six different authors (collectively known as the *Scriptores Historiae Augustae*). However, the true authorship of the work, as well as its actual date, its reliability, and its purpose have long been matters for controversy among historians and scholars (White 1967). Regarding authorship, Dessau (Dessau 1889) has rejected the traditional attribution to six different authors, i.e., Aelius Spartianus, Julius Capitolinus, Vulcacius Gallicanus, Aelius Lampridius, Trebellius Pollio, and Flavius Vopiscus, proposing instead a single authorship. Internal attribution itself rests on fragile foundations. Stover (Stover 2020) has proved that the textual transmission of HA is also intertwined with problems of codicology to a point that it affects not only readings, but also the literary authorship of each *Vita*. Indeed, it turns out that codicology explains how some of the *Vitae* have been attributed to its fictitious author, e.g. the authorship of *Valeriani Duo* is ascribed to Julius Capitolinus according to the main manuscript, but Stover shows that its *incipit* belongs to another branch of the textual tradition, and ascribes it to Trebellius Pollio.

Dessau's arguments were widely accepted among scholars and historians (White 1967), such as (Syme 1971; Adams 1972; Paschoud 1991; Burgersdijk 2010; Rohrbacher 2016; Cameron 2011; Savino 2017; Stover 2020). However, dissent from such trend emerged first on a paper by Momigliano (1954), who argues for a multiple authorship—a viewpoint, largely dismissed today [a recent attempt to uphold it was made by Baker (2014)]. An intermediate position is that of Hengst et al. (2010), who dilutes the concept of a single author into that of one editor who writes into his own materials from previous biographies.

This subject also attracted the interest of scholars working with computational methods. Marriott (1979) published a paper consisting of two studies: one analyzing the distribution of sentences lengths, and the second taking as base the grammatical types of words that appear in the beginning and at the end of the sentences. Both studies supported single authorship of the HA, but his methods, specifically the use of sentence length, were criticized, namely by Tse et al. (1998). These same authors advanced an approach based on different statistical methodology applied to the occurrence of function words, concluding for multiple authorship. More recently, Stover and Kestemont (2016) shifted the focus from authorship attribution to authorship verification, resorting to the General Imposters framework (GI) and over the results obtained was applied a Principal Components Analysis. This methodology allowed the authors to conclude that the results obtained by GI verification did not support multiple authorship; moreover, besides the existence of stylistic features common to the entire collection, their research displays evidence of two distinct authorial layers "which correspond, more or less, to the categories of the Hauptviten and later lives" (Stover and Kestemont 2016), showing also a stylistic discontinuity after the lacuna.

Complex networks have been the target of intense research. Their vast thematic scope, touching all fields of our world, make this field of science a fascinating one, resulting in an intense and fruitful scientific production. And language could not be excluded from this huge range of issues. Since the early 1990s, studies for the modeling of information, not only of texts but also of multimedia data, have gained more and more attention from researchers, namely, the study through complex networks. These networks, which represent the structure of a text, can be constructed in different ways. In our work, we consider the co-occurrence networks in which each word is a vertex and each edge represents an adjacency relationship between two vertices. This construction was used for the first time in 2001 (i Cancho and Solé 2001; Dorogovtsev and Mendes 2001) and has continued to be used and developed (Amancio et al. 2008, 2011; Mehri et al. 2012; Segarra et al. 2015; Kulig et al. 2015). In particular, the question studied in this work analyzes the authorship of HA texts using the style marks extracted from the complex network.

In this work, an authorship study of a Latin dataset using co-occurrence graphs is presented. The text representation in co-occurrence text graphs is different from the approaches followed in related works. Text is divided in pieces of 100 words each, and each one is represented as a co-occurrence graph. The number of text samples in the dataset increases but it still keeps the author style marks captured in complex measures extracted from the graph. The texts dataset has different characteristics from the datasets used in other works. Historia Augusta is attributed to six authors. It is a set of few texts by each author, all of them with a small number of words. From the co-occurrence graphs, 11 measures are extracted; these measures are used in most works except for Fiedler. The impact of each measure in each text classification is evaluated, since the number of subsets is relatively small ($2^{11}$. This evaluation concluded that it is better not to exclude any of the parameters, since the best parameters subset could vary a lot and the classification results could be similar for different sets. We can conclude that our approach to represent texts in co-occurrence networks can be used with state of the art results in an authorship attribution task even when the text dataset has less than 3 texts by author and texts have less than 2000 words. In Akimushkin et al. (2017) each author text is represented by 4 moments of each time series complex measure, 12 are used, a time series of a measure is obtained in the sequence of 268 co-occurrence weighted graphs built with pieces of 200 tokens/words, resulting in a dataset with 48 parameters that the authors try to reduce using Principal Component Analysis and Isomap. The text dataset used is composed by 8 English and 8 Polish authors with 6 books by each author, and all the books in the dataset have more than 30000 words. The authors report 90%. accuracy on the task.

In Marinho et al. (2016) a text dataset composed by 48 texts, 8 English authors with 6 texts each, is used. A co-occurrence graph is built from which the complex measure frequency of direct motif involving 3 nodes is extracted. The graphs are trunked to get the same dimension for all the texts. This work reports an accuracy of 57.5% in the authorship recognition task that increases up to 65% when network measurements are combined with the intermittency of the distribution of words along the text. In Segarra et al. (2013) the text is represented as a Normalized word adjacency networks combined with word frequency with function words. In their study, they conclude that accuracy

increases by combining relation and frequency of data. The text dataset is composed by 18 English authors with 6 to 10 books each, all books have more than 30000 words. This work reports an accuracy between 54 and 100%, depending on the number of authors and books evaluated. In Quispe et al. (2021) texts are represented as co-occurrence networks, augmented with virtual nodes that are obtained by word similarity calculation and their link. The word similarity is calculated by using word embeddings. This important work shows how language semantics can be incorporated in the co-occurrence networks. They use the dataset in Segarra et al. (2013). For languages with as many resources as English, the calculus of similarity using word embeddings is possible, but for a language with poor resources such as Latin, it will be difficult, even when using the multilingual language models. Another text representation using co-occurrence graphs is presented in de Arruda et al. (2019) where the authors propose a paragraph-based representation of texts.

## Co-occurrence networks

The method we present for assigning an authorship is based on the evolution of the topological structure of the networks. Among the different types of existing word networks, there are the so-called co-occurrence networks, characterized by relating the vertices (words) from their proximity in a text. In this model, a text is represented by a complex network, in which each word is a vertex and each edge represents an adjacency relationship between two vertices. Thus, for each pair of consecutive words, there is a corresponding directed edge on the network.

Therefore, unlike previous approaches, we do not construct one single network from the whole text. Instead, a text is divided into shorter pieces of text comprising the same number of words. Note that we use the same number of words in each partition because some network measurements are sensitive to network size. We first remove all punctuation marks and numbers from the original text and then divide the text into blocks of 100, 200 or 500 words. By making this partition we obtain a sequence of n subsets of the text with an equal number of words. Then we do two series of different calculations: in one series we remove the stop words and in the other series we keep the stop words.

Although there are works analyzing corpus in modern languages which use only stop words (see Segarra et al. 2013), we have not considered this option due to the syntactic characteristics of Latin language.

With each of the subsets, we build a sequence of networks, networks of co-occurrence (word collocation networks), where the nodes are the words/tokens and an edge between two words/tokens indicate that the words/tokens are neighbors (Bollobás and Riordan 2005). After obtaining a sequence of co-occurrence independent networks, we calculated all measures for each of these networks. Our construction is outlined in Fig. 1.

An example of a part of the text and the respective network of co-occurrences, can be seen in Fig. 2.

*...Redeunti sane Romam post bellum civile Nigri aliud bellum civile Clodi Albini nuntiatum est, qui rebellavit in Gallia. Quare postea occisi sunt filii eius cum matre. Albinum igitur statim hostem iudicavit et eos, qui ad illum mollius vel scripserunt vel rescripserunt. Et cum iret contra Albinum, in itinere apud Viminacium filium suum maiorem Bassianum adposito Aurelii Antonini nomine Caesarem appellavit, ut fratrem suum*

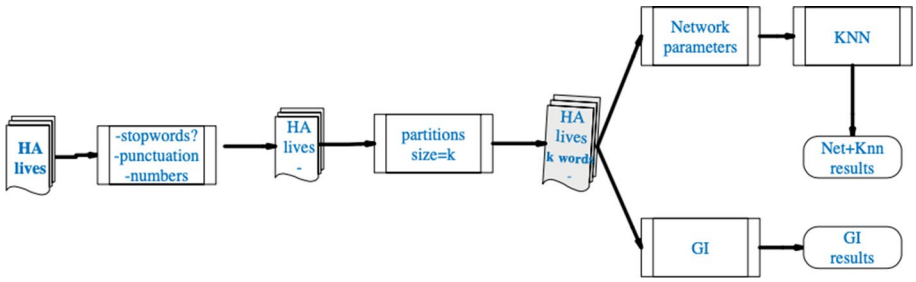


**Fig. 1** Authorship verification systems information flow

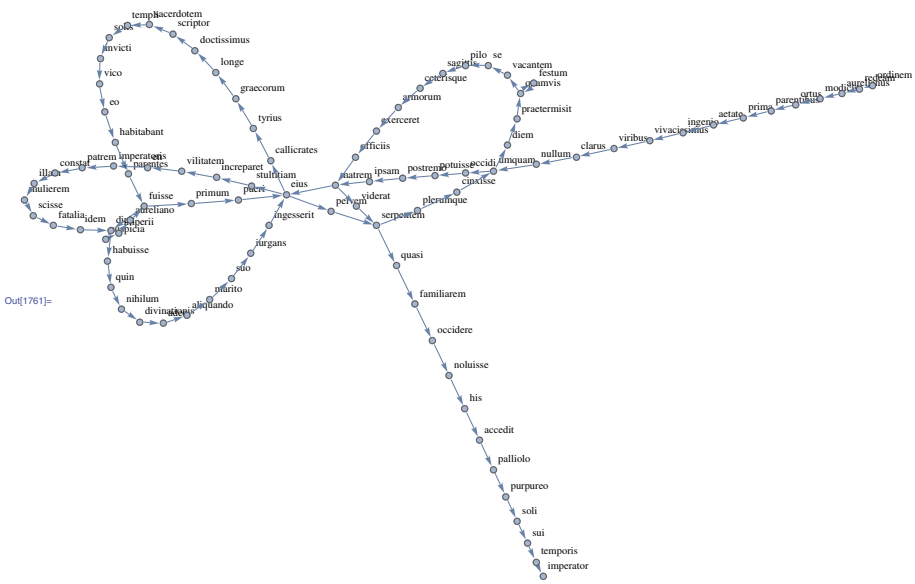


**Fig. 2** Co-occurrence netork corresponding to the extract from the HA text that contains what was previously presented

*Getam ab spe imperii, quam ille conceperat, summoveret. Et nomen quidem Antonini idcirco filio adposuit, quod somniaverat Antoninum sibi successurum. Unde Getam etiam quidam Antoninum putant dictum, ut et ipse succederet in imperio. Aliqui putant idcirco illum Antoninum appellatum, quod Severus ipse in Marci familiam transire voluerit. Et primo quidem ab Albinianis Severi duces victi...*

### Topological network measurements

For each of the text blocks, a co-occurrence network is constructed, which generates a sequence of independent networks for each text of the corpus. In order to represent the author text, we calculated, for each of these networks, all the graph parameters that we consider relevant for this characterization. The main hyperparameter of the model above described is the number of words $w$ in each text block. We tested several options to analyze the best $w$, which is, the value that provides graphs big enough to show a distinctive structure but also allows a large number of graphs for each corpus item (Rodrigues et al. 2020).

Each graph partition is described by the following topological network measurements:

**Total number of nodes** A first measure is the total number of vertices, that is, the measure of the vocabulary of each partition, which is called the order of the graph. The number of words $w$ of each text block is fixed, and the building of the co-occurrence network identifies equal words/tokens in one vertex. Hence the order of the resultant graph, $n$, is the number of different words/tokens among the $w$ in the text block that capture what can be considered a style mark.

**Total number of edges** If one has a network with $n$ nodes, there are $n-1$ directed edges that can lead from it (going to every other node). Therefore, the maximum number of edges is $n(n-1)$. The number of co-occurrences in a linear text block of $w$ words/tokens is $w-1$, a co-occurrence between each two consecutive words. This measure captures the use of n-grams by the authors. When the number of edges is lower than $w-1$, there are repeated occurrences of some n-grams, sequences of words/tokens in the text; if the sequence 'poor boy' appears more than once in the text, it appears only once in the graph.

**Degree of a node** For an undirected graph, the degree of a node $v_i$, is the number of edges incident to it and is represented by $k_i$, that is, $k_i = \sum_{j=1}^{j=N} a_{ij}$. But for a directed graph more information is needed. We define the *in*-degree $k_{in,i}$ of the vertex $i$ as the number of the edges arriving at $i$, $k_{in,i} = \sum_j a_{ji}$, and the *out*-degree $k_{out,i}$ of the vertex $i$ as the number of the edges departing from $i$, $k_{out,i} = \sum_j a_{ij}$. Then, the degree of a vertex $i$ , $k_i$, is defined by the sum of the in-degree and the out-degree, $k_i = k_{in,i} + k_{out,i}$. This measure reflects the author use of vocabulary in different contexts. For instance, if an adjective is used always before the same noun, the *out*-degree of the adjective will be just one.

**Clustering coefficient** The clustering coefficient measures the number of mutual neighbors of adjacent nodes, such that, the average probability for two neighbors of some vertex can be directly connected. As an illustration, we can say that the clustering coefficient of a given vertex $v$ (in this case a word in the text) indicates the probability that any other vertices adjacent to $v$ (in this case two words located next to $v$) will also connect (these two words can be found next to each other).

For instance, a text like 'Oliver Twist' by Charles Dickens has the following words sequence:

"... poor desolate creature ..." and "... poor creature ..."

The words/tokens 'desolate' and 'creature' are neighbors and share a neighbor, 'poor'.

The clustering coefficient $c_i$ of a vertex $v_i$ is given by the proportion of links between the vertices within its neighborhood divided by the number of links that could possibly exist between them. The clustering coeficient can be useful to detect authorship by quantifying the tendency of using semantic-specific or generic words (Amancio et al. 2011).

To characterize the global clustering coefficient of the network we consider (Bollobás and Riordan 2005) the average over all vertices with degree larger than one.

**Path length** Another measure for the structure of a graph is the average short path length. A path in a directed graph $G$, is a sequence of vertices and edges that begins

with a vertex, ends with a vertex, and such that for every edge $(v_i \rightarrow v_j)$ in the path; vertex $v_i$ is the element just before the edge, and vertex $v_j$ is the next element after the edge. We only consider the paths in which all the *vs* are different. A path between $v_0$ and $v_k$ is a sequence of the form $v_0 \rightarrow v_1 \rightarrow ... \rightarrow v_k$. The length of the path is the number of edges, $k$. The shortest path length is the smallest number of edges between two nodes (called distance between two nodes) and the average shortest path length is the average length of all shortest paths between vertices of G. We can say that the average shortest path length is the typical distance between any two nodes in the network. In texts, the average shortest path length quantifies the relevance of words. This parameter tells us that the most important words are those that are closest to the central words (hubs).

**Network diameter and Network radius** Diameter and radius are indices that measure the topological length of a network by counting the number of edges in the shortest path between the most distant vertices. The maximum distance between a vertex and all other vertices is considered as the eccentricity of vertex. The maximum eccentricity from all the vertices is considered as the diameter of the network, and the radius is the value of the smallest eccentricity. In a network of words, it is helpful to detect the size of sentences in the text.

**Betweenness centrality** Let us denote the total number of shortest paths between vertices $s$ and $t$ by $\lambda_{st}$, and the number passing through vertex $v$ by $\lambda_{st(v)}$. Let $\delta_{st(v)}$ denote the fraction of shortest paths between $s$ and $t$ that pass through a particular vertex $v$ i.e., $\delta_{st(v)} = \frac{\lambda_{st(v)}}{\lambda_{st}}$. Betweenness centrality of a vertex $v$ is defined as $Bc(v) = \sum_{s \neq v \neq t} \delta_{st(v)}$

In a simple way, this measure helps to identify words that play a bridge in a text. The betweenness is able to identify the generality of contexts in which a word appears. More generic words tend to have higher betweenness values while more specific words (for a topic) tend to have lower betweenness values (Amancio et al. 2011).

**Assortativity** In several networks obtained from real contexts, it is common to have a tendency for connections between vertices of similar characteristics or, on the contrary, a tendency for connections between vertices with opposite characteristics. This type of analysis can be performed with respect to different properties, although it is more frequent to use the vertex degrees for this study. In a network, degree assortativity is called the phenomenon of tendency of connections between vertices with similar degrees. The phenomenon of a tendency to form connections of vertices of low degrees and vertices of high degrees is called disassortativity. We can estimate (Newman 2006; Murakami et al. 2017) the assortativity of a network, through the correlation coefficient.

In assortativity networks (where $r > 0$), vertices with the same degree connect to each other, whereas in networks disassortativity (where $r < 0$), vertices with low degree establish connections with vertices that have high grade. In word adjacency networks, the assortativity quantifies how words with distinct frequency appear as neighbors.

**Number of clique** A graph is complete if every pair of vertices are adjacent. A clique is a subgraph (a subset of a graph's edges, and associated vertices, that constitutes a

graph) which is complete. The choices that the authors make in a given set of words that are often used together to describe a certain topic, result in cliques in the co-occurrence network.

**Spectral method** The spectral methods for complex networks are based on the eigenvectors of certain matrix associated with the network. It is the Laplacian matrix: $L = D - A_s$, being $A_s$ the symmetrized adjacency matrix and $D$ the matrix that has the total degree of each vertex in the diagonal. It is known that $L$ has real eigenvalues $\lambda_1 \leq \lambda_2 \leq \ldots \lambda_n$, the smallest of them $\lambda_1 = 0$, and that the multiplicity of $\lambda_1$ is the number of connected components (Gera et al. 2018).

There are deeper results about the next eigenvalue $\lambda_2$ and its corresponding eigenvector, called Fiedler vector (Cvetković et al. 1980). The Fiedler value $\lambda_2$ is related to the conductance, defined in the following manner. For a partition of the set of vertices in two parts, its cost is the number of edges that join vertices in different parts. The conductance is the minimum cost that can be obtained by a partition. Besides, the Fiedler vector gives us such partition (Spielman and Teng 1996).

A very important problem in characterizing the behavior of complex networks is their ability to correctly detect communities. Communities are groups of vertices that are more connected to each other than to the other vertices in the network. This is important if a text can break into small, densely connected groups and conductance is a measure of the quality of a community. Conductance is closely related to the value of Fiedler.

An intuitive interpretation of the value of Fiedler or of conductance is that it indicates the connectivity of the graph: in co-occurrence networks, a low value indicates that there are groups of words strongly connected within the group, but weakly outside it. Such groups can be seen as vocabularies of a similar theme or style. A high value indicates that such groups are not clearly defined.

With respect to the computational complexity, the above measurements can be computed in linear time $O(n)$ or quadratic time $O(n^2)$ in the order $n$, the number of the graph vertices. Even for the Fiedler eigenvector the time complexity is quadratic. To calculate the number of cliques, we use the function FindClique from WOLFRAM MATHEMATICA that has time complexity NP-hard.

The time complexity to obtain the texts dataset is not an issue since it can be done in a few hours for the Latin or Portuguese corpus.

Figure 3 represents the values of the parameters calculated in the sequence of graphs obtained from the Life of the Divus Aurelian. The original text, after being processed, originated a sequence of 50 networks. With different colors, the parameter values are drawn in each of the 50 networks.

Vertex indicates the number of vertices for each graph, it is represented by a red line and varies between 62 and 97. Edge indicates the number of edges for each graph, it is represented by a cyan line and varies between 65 and 99.

Cliques indicate the number of cliques for each graph (In this case also includes cliques formed with only one vertex), it is represented by a blue line and varies between 65 and 99.

MeanClusteringCoef of a graph is the mean over all local clustering coefficients of vertices of the graph, it is represented by a brown line and varies between 0 and 0.11.

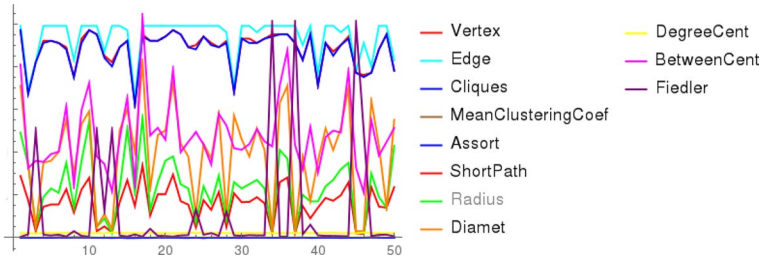


**Fig. 3** Topological network measurements calculated for the Life of Divus Aurelian. Each measure has a different scale. The range of each is commented on in the text

Assort gives the assortativity coefficient of each graph using vertex degrees, it is represented by a blue line and varies between -0.17 and 0.2.

ShortPath is the average length of all shortest paths between vertices of each graph, it is represented by a red line and varies between 1,75 and 33,83.

Radius gives the minimum among all the maximum distances between a vertex to all other vertices in each graph, it is represented by a green line and varies between 2 and 56.

Diameter gives the greatest distance between any pair of vertices in each graph, it is represented by an orange line and varies between 3 and 82. For a directed graph, the in-degree is the number of incoming edges and the out-degree is the number of outgoing edges.

DegreeCent gives the average of the degrees of the vertices of each graph, it is represented by a yellow line and varies between 2 and 2.4.

BetweenCent will give high centralities to vertices that are on many shortest paths of other vertex pairs, varies between 593 and 2892, and its value scaled by 1/28 is represented by a pink line.

Fiedler gives us an algebraic measure of connectivity, and varies between 0.006 and 2. Its value scaled by 50 is plotted in black.

## Authorship Verification

Different authorship verification (attribution) methods (Stamatatos 2009) differ on the representation chosen for the texts, as well as on the similarity measures used to define closeness between texts representations. One common approach to text representation is the replacement of the text words by numbers using vectors. These frequency vectors, more precisely frequency vectors of terms, can be simply words, or lemmas in n-grams (groups of words).

In this paper, the General Imposters method (GI) (Koppel and Winter 2014; Kestemont et al. 2016) in the R package stylo (Eder et al. 2016) is used to compare the results obtained using Topological and Spectral Measurements of Complex Networks to represent the analyzed texts. The use of imposters enables us to compare our results on *Historia Augusta* with those reported in Stover and Kestemont (2016) that use GI.

The use of Complex Networks Measurements to represent the texts for authorship attribution has been done by other authors (Akimushkin et al. 2017; Stanisz et al. 2019) with promising results for text in English and Polish. In this paper, we applied

it to Portuguese and Latin texts (Teixeira and Rodrigues 2018) and the Parameters Measurements used include other, such as Fiedler. The analysis of *Historia Augusta* constitutes a greater challenge due to the size of the texts of each Life, because some of them are very small.

As a classifier, we use K-nearest neighbor (KNN) with Euclidean distance and prior to classification, we standardize the feature values of the complex network measures that represent the text to increase the KNN classifier performance. Feature values standardization eliminates scale effects due to features with different measurement scales (e.g., Total number of edges and Fiedler have different scales).

For each experiment done with complex networks, an equivalent experiment was done with GI (see Fig. 1).

To run GI we need a corpus with texts classified by author, then we can choose the set of texts we want to test the classification (author attribution). For each text tested, GI returns a value between 0 and 1, for each of the candidate classes (authors), which represents the probability that the text belongs to the corresponding author. Accuracy is measured as follows: if the probability obtained for the tested file class is greater than 0.5 and all values for the remaining classes are less than 0.5, we will consider that the method hits the file category (author).

To test the authorship method using Topological and Spectral Measurements of Complex Networks, a dataset is built with the texts network parameters and their classification, the text identification and the author's name (Rodrigues et al. 2020). Then we use R studio class package to standardize the dataset parameters' values and we run K-nearest neighbor (KNN), with K=1 and Euclidean distance function. Since KNN is sensitive to the number of features used to represent each text, i.e. for a dataset, a model built with n parameters can have lower accuracy then a model using n-1 parameters. So, we have a first task to find the best set of features for each dataset.

To find the best set of features to represent the Portuguese authors' texts and the authors of the Latin texts, we built a model for each subset of the text features (if $n$ measures of the complex networks are used to represent a text, we try $2^n$ features subsets), we evaluate it by calculating the accuracy and we choose the set of features that gives rise to the best accuracy.

In our experiments with Portuguese authors, we evaluate the adequacy of representing texts or parts of texts with co-occurrence networks, extract Measurements of Complex Networks from those networks and use a classifier build with KNN: to this method, we call it Net+Knn.

#### Authorship Verification of Portuguese Books

Our Portuguese corpus was built with three books of three Portuguese contemporary authors: nine books total. In this section we present the experiments to evaluate the accuracy of the Portuguese authors books with different approaches: text with and without stop words and representing parts of each text (100, 200 and 500 words) or all the text. In the experiments we use two methods, General Imposters (GI) and Topological and Spectral Measurements of Complex Networks with the classifier KNN (Net+Knn).

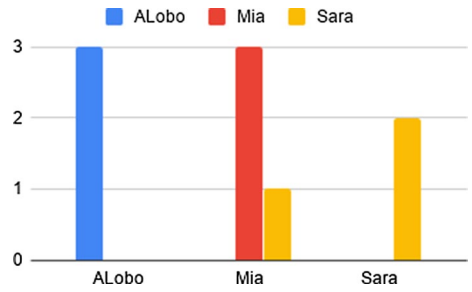


**Fig. 4** Number of works correctly assigned using one Graph by book

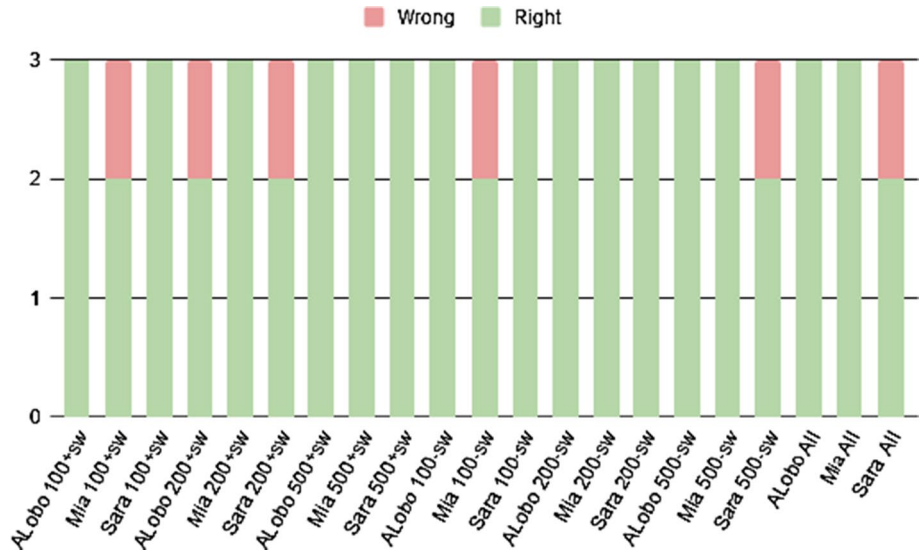


**Fig. 5** Portuguese authors books correctly assigned with network series using Net+Knn

***Accuracy of the Portuguese authors books representing all the text.***

First, for each book, a word co-occurrence network is built, and the Measurements of Complex Networks are calculated in order to obtain a feature value representation that includes the information on the book author. This dataset, with 9 lines, is the input of KNN. To run KNN, we divide the dataset into train and test. Since there are only 3 books for each author, the train will have 8 books and the test 1 book. When KNN runs, it outputs a confusion matrix. KNN was run 9 times, one for each book, and the confusion matrices are added to obtain the final one where the accuracy is calculated. Since KNN is sensitive to the number of features, we calculate the feature subset that gives the best accuracy, which for this case is: "Average shortest path length", "Betweenness centrality" and "Fiedler". The calculus of the best feature subset is NP-hard, but since we have a relatively small n, 11 features, we run KNN with K=1, 2047 times for each book.

The final confusion matrix is represented in Fig. 5. Note that the results with and without stop words are similar. We can see that 8 out of 9 books (88%) have their authorship confirmed. Only 'Sara' (José Saramago) has a book that is attributed to Mia (Mia Couto). The confidence interval can be calculated with the expression

$\pm C\sqrt{accuracy(1-accuracy))/n)}$, with $C = 1.96$ for a 95%, and $n = 9$ the number of observations, $\pm 21.2\%$[1].

To obtain the accuracy of GI in this corpus, we created 3 classes named *Sara*, *Mia* and *ALobo* with three files each. Then GI was run 9 times, one for each text as test, and the text accuracy was calculated. The accuracy of GI for this set of books and authors is 100%, which is higher than the proposed method. This same result is obtained when texts are represented with or without stop words (Fig. 4).

This experiment allows us to conclude that the use of Measurements of Complex Networks to represent authored texts combined with a classifier such as KNN is an accurate method for authorship verification (Mehri et al. 2012), even if other methods can have a better performance. Both methods work well for Portuguese texts.

#### *Accuracy of the Portuguese authors books representing parts of the text.*

In these experiments, we divided each text in 20 parts. The idea was to evaluate the performance of the authorship verification algorithms when texts are smaller, and to determine how smaller a text can be to still convey some stylistic issues that enable authorship recognition.

When the text is divided in 20 pieces, we will get more samples for representing each book, so our dataset for classification with Net+Knn will have 180 samples. The classification will be done by using all the 20 samples of one book as test, and the others as training. The test is repeated for the 9 books and the confusion matrices are added. In this experiment, what is classified is a part of the text, but we can classify the book considering that if more than 50% of the parts are well attributed, then the book authorship is recognized.

We made experiments with pieces of 100, 200 and 500 words, with and without stop words. The results are presented in Fig. 5.

These results were obtained using the features subset that better discriminates the Portuguese authors, for which KNN model has better accuracy. For instance, the features chosen for 100 words pieces are: "Assortativity" and "Fiedler".

In Fig. 5, each column has a label ('Author', 'number of words of each piece', '+/-sw'): 'Author' is the author abbreviation, 'number of words of each piece' is the number of words used to build each of the 20 graphs, '+sw' means that stop words are present and '-sw' means that stop words were removed. The first column, *ALobo 100+sw*, represents the accuracy for the recognition of the authorship of the books of ALobo, a Portuguese author, when the books are represented by 20 pieces of 100 words including stop words.

We have made the same experiments using GI, with 20 pieces of each text varying the size and removing the stop words or not. These results are presented in Fig. 6.

As we can see in Fig. 5, the results of Net+Knn:

- When each work is represented by 20 networks with 100 words each, with and without stop words. The results for 100+sw and 100-sw are the same for the three authors 88%, with the confidence interval 4.7% for 95% probability.

[1] Due to the low number of observations, less than 30, the confidence interval calculus can be considered abusive.

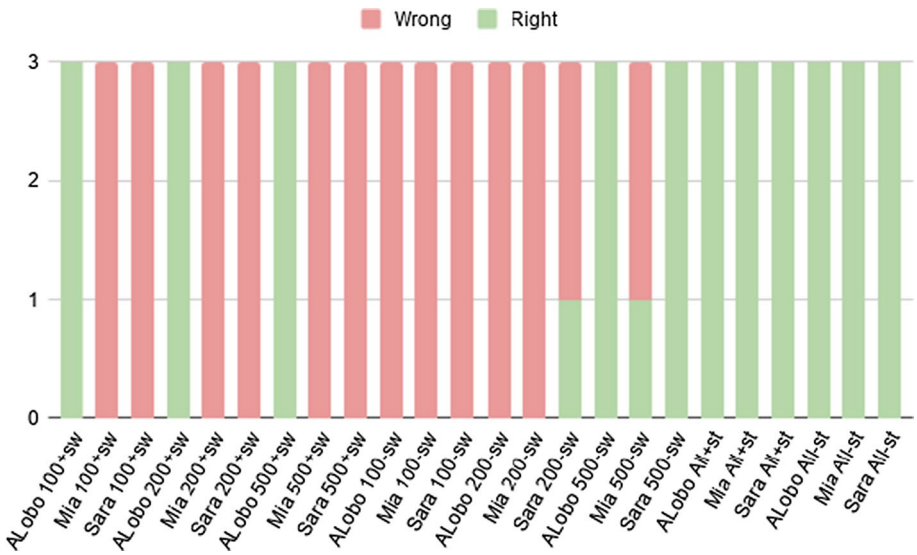


**Fig. 6** Number of Portuguese authors works correctly assigned varying the number of words using GI

- When works are represented by 20 networks with 200 words each, with and without stop words. The results are 100% of accuracy, for 200+sw and 200-sw are lower 77%, the confidence interval 6.1% for 95% probability.
- When works are represented by 20 networks with 500 words each, with and without stop words. Sara 500-sw (José Saramago excluding stop words) has a book that is not recognized as authored by him, with the confidence interval 8.2% for 95% probability (considering n= 60 observations). But when the networks are built without excluding the stop words, Sara 500+sw, the three books have their authorship verified with the confidence interval 0%. The works by the other two authors have their authorship recognized.

These results show that the confidence interval improves; it is smaller in cases where the accuracy is the same when we consider 20 pieces of text, because we increase the number of observations.

The results lead us, also, to conclude that Net+Knn is a good tool to use with small texts like the ones in Historia Augusta, and behaves well when comparing texts of different length. There are other works (Akimushkin et al. 2017) that report similar results for different languages and different tasks.

The results of GI are presented in Fig. 6:

- When each work is represented by 20 pieces of 100 words each, with and without stop words. The results for 100+sw are 33%: only one author, ALobo, has its works recognized, the other authors do not have any of their works well attributed. When the stop words are removed from the text the results get worse, none of the authors has any work recognized.
- When works are represented by 20 pieces with 200 words each, with and without stop words. With the stop words 200+sw, 3 works in 9 are well assigned and without stop words, 200-sw, only 1 work in 9 has the authorship well recognized.

**Table 1** Chapters and Authors chosen

| Author | Chapters |
|---|---|
| Aelius Lampridius, AL | Elagabalus (6067 words), Severus Alexander (11182 words), Diadumenus (1745 words) and Commodus (3693 words) |
| Aelius Spartianus, AS | Septimius Severus (4449 words), Pescennius Niger (2372 words), Hadrianus (5438 words) and Caracalla (2106 words) |
| Flavius Vopiscus, FV | Tacitus (3373 words), Probus (4521 words), Divus Aurelianus (8139 words) and Carus Carinus Numerian (3118 words) |
| Julius Capitolinus, JC | Marcus Aurelius (5829 words), Gordiani Tres (5804 words ), Pertinax (2754 words) and Opellius Macrinus (2599 words) |
| Trebellius Pollio, TP | Gallieni Duo (3838 words), Tyranni Triginta (6919 words), Divus Claudius (3131 words) and Valeriani Duo (1052 words) |
| Marcellinus, AM | "Hist_lib 14" (8368 words), "Hist_lib 18" (5142 words), "Hist_lib 22" (9110 words), "Hist_lib 28" (6248 words) |

- When works are represented by 20 networks with 500 words each, with and without stop words. The performance of Imposters improves for 500-sw, it recognizes the author of 7 works in 9, but for 500+sw it recognizes 3 in 9 only.

GI can have a very good behavior when works or pieces of work are above a size of 2000 words at least.

Our experiments lead us to conclude that the use of Measurements of Complex Networks to obtain a text representation to be classified with a classifier such as KNN, Net+Knn, has state of the art results in authorship verification task. This method, Net+Knn, has good accuracy results when only some pieces of the books are used, and has smaller confidence intervals because the number of observations increases. The fact that we can use series of small pieces of text (100 words) is important when we try to verify the *Historia Augusta* authorship since the amount of text in some Lives is reduced.

## Authorship of *Historia Augusta*

In order to study the authorship of *Historia Augusta* with Net+Knn and GI, we started by selecting four Lives of each author, excluding the only Life authored by Vulcatius Gallicanus, i.e. the Life of Avidius Cassius which has only 1000 words.

We included works of a late-antique author, *Ammianus Marcellinus*, whose *Res Gestae* was used as control for the Net+knn and Imposters methods.

The four Lives[2] of each author were chosen in order to balance the classes for the classification process that is performed using KNN algorithm on the set of networks that represents each text. Table 1 presents the HA texts and the Ammianus Marcellinus texts used.

Table 1 presents the author, the name and the size of the work. Note that the size of the works is indicated through the number of words, and ranges from 1052 words, as in

[2] *Historia Augusta* is divided in Lives (a biography of a Roman Emperor).

**Table 2** Chosen Chapters and Authors

| #words | #Graphs | Exceptions |
|---|---|---|
| 100 | 13 | AL-Diadumenus (12 Graphs)<br>TP-Valeriani Duo (6 Graphs) |
| 200 | 6 | AL-Diadumenus (5 Graphs)<br>TP-Valeriani Duo (2 Graphs) |
| 500 | 5 | TP Divus Claudius, JC Pertinax,<br>FV Carus Carinus Numerian (4 Graphs)<br>AS Pescennius Niger, AS Caracalla,<br>JC Opellius Macrinus (3 Graphs)<br>AL Diadumenus (2 Graphs),<br>TP Valeriani Duo (1 Graph) |

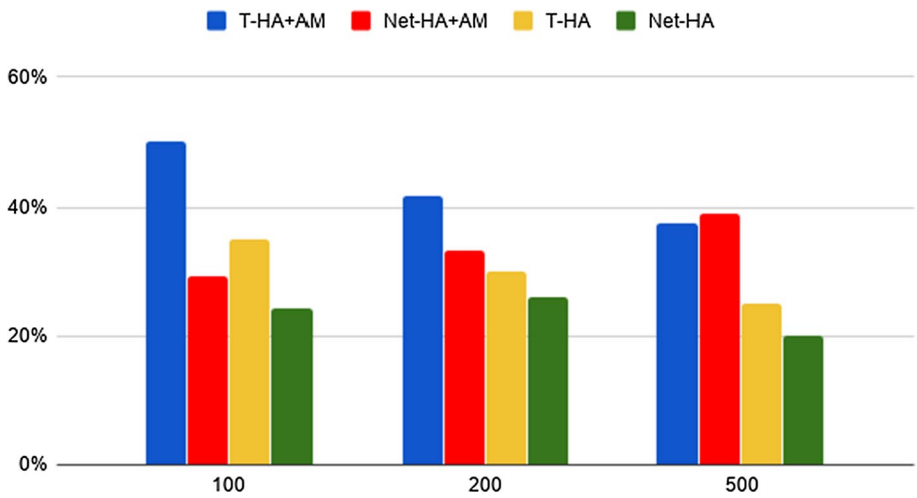


**Fig. 7** Accuracy of HA Authors and Ammianus Marcellinus with Net+Knn

the case of Valeriani Duo by Trebellius Pollio, to 11182 words, as in the case of Severus Alexander by Aelius Lampridius.

#### Authorship of *Historia Augusta* and *Res Gestae* of Ammianus Marcellinus

Our first experiences were done to decide the best number of words in each co-occurrence network.

We tried with 100, 200 and 500 words. To balance our classes, we took samples from the documents to obtain a similar number of graphs to represent each document (see Table 2).

Three datasets were built by computing the Measurements of Complex Networks from the graphs that represent our Latin texts described above. One dataset for each graph (co-occurrence words) size: 100, 200 and 500 words.

Note that, due to the size of some texts, the texts representation is not well balanced. As it can be seen in the above table, there are some texts that have fewer graphs than others. This can influence the KNN performance.

To determine the best number of words in the graphs, we calculated: the accuracy of the classification of the texts and the accuracy of the graphs.

We ran two experiments: one with both HA texts and AM texts (AM is the control author), and the other with just the HA authors, in a total of five authors, since one has only one text in HA.

Figure 7 presents the results of these experiments.

In this figure the results for 100, 200 and 500 words are displayed with four values each:

- T+HA+AM—accuracy of texts by Historia Augusta authors and A. Marcellinus: when the graphs were built with 100 words the accuracy and confidence interval is $50 \pm 20\%$, with 200 words is $42 \pm 19.7\%$ and with 500 words is $38 \pm 19.4\%$.
- Net+HA+AM—accuracy of the graphs of texts by Historia Augusta authors and A. Marcellinus: when the graphs were built with 100 words the accuracy is $29 \pm 5.1\%$, with 200 words is $33 \pm 7.8\%$ and with 500 words is $39 \pm 9.5\%$.
- T+HA—accuracy of texts by Historia Augusta authors: when the graphs were built with 100 words the accuracy is $35 \pm 20.9\%$, with 200 words is 30% ±20% and with 500 words is 25% ±18.9%.
- Net+HA—accuracy of the graphs of texts by Historia Augusta authors: when the graphs were built with 100 words the accuracy is $24 \pm 5.2\%$, with 200 words is $26 \pm 8\%$ and with 500 words is $20 \pm 8.7\%$.

The data presented in Fig. 7 was obtained by using Net+Knn. Repeating the following procedure for each text: the train set is built with the representation of the other texts, KNN classifies the text tested and the resulting confusion matrix is added to the previous ones. At the end the accuracy is calculated with the information on the final confusion matrix.

When we calculate the accuracy of graphs, Net+HA+AM or Net+HA, we use the confusion matrix as KNN returns it. When we calculate the accuracy of the texts, T+HA+AM or T+HA, the confusion matrix returned by KNN is transformed, inserting the value 1 in the position of the author of the tested text, if the number of graphs correctly assigned to the author is greater than 50% of the total number of graphs that represent the text; otherwise, the value inserted is 0.

Figure 7 shows that the texts classification, T-HA+AM, is better (50%) when each text network has 100 words. The accuracy of the graphs classification, Net+HA+AM, has the best result (40%) when the networks are built with 500 words.

When using HA authors texts only, the number of well classified texts, T+HA, achieves its best result (38%) with graphs of 100 words. But each graph, Net+HA, is better classified with 200 words graphs.

These results are not very encouraging. The best result is 50% when the classification is done with 100 words graphs, including the four texts of AM, the author used for control.

Looking at Fig. 8, which shows the text graphs classification resulting from the confusion matrix, we can see that most graphs of the texts by Ammianus Marcellinus (AM) are well classified (29 in 52), unlike the HA authors which have all less than 50% of well classified graphs.

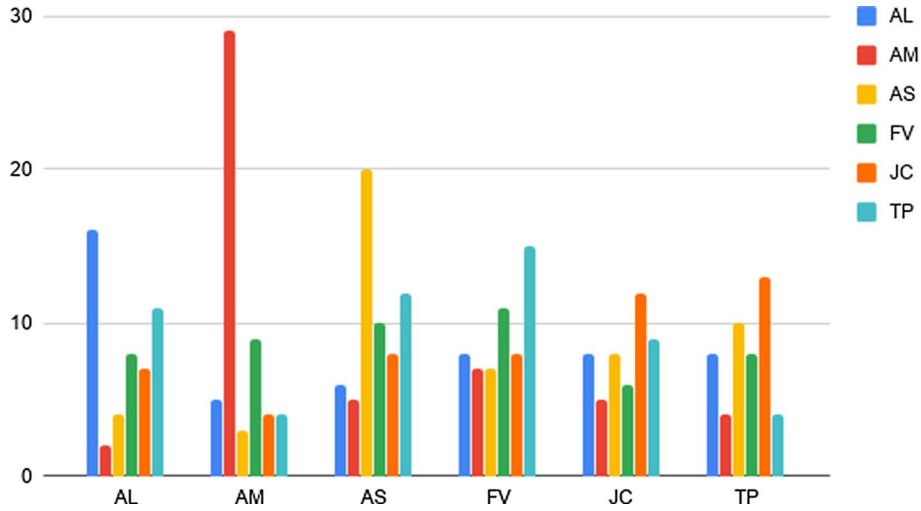


**Fig. 8** Classification of text graphs by HA Authors and AM using Net+Knn

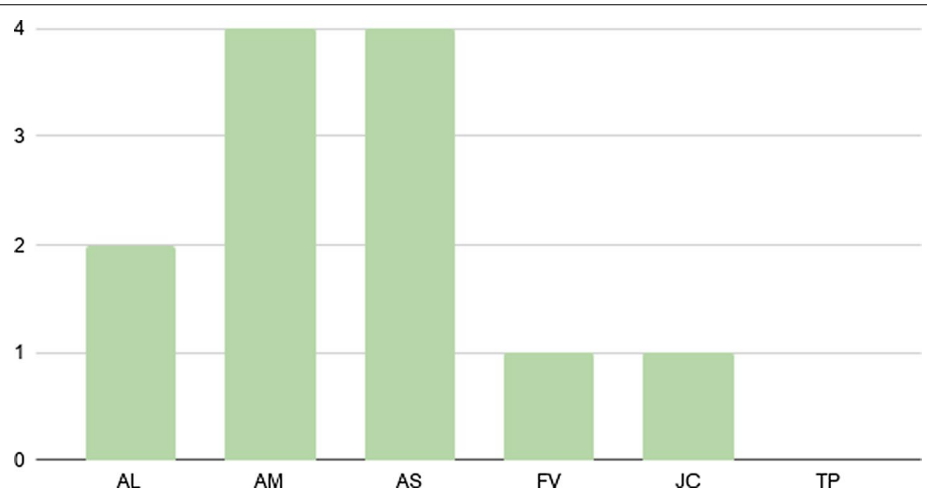


**Fig. 9** Classification of HA Authors and AM texts using Net+Knn

Figure 9 presents the transformed confusion matrix for texts classification. An author is associated to the number of texts well classified.

The four texts by Ammianus Marcellinus are well classified, and some of the Lives of HA are correctly attributed to their traditional authors. Nevertheless, only AS's texts are all correctly attributed (four in four).

From these experiments we conclude that the best results are obtained with 100 words graphs. The Net+Knn method is able to identify Ammianus Marcellinus' texts, though it reveals some problems when applied to HA Lives.

Figure 10 illustrates the results of the same test but using GI.

This figure shows that the classification results are similar: GI identifies Ammianus Marcellinus' authorship but fails on most of the HA Lives.

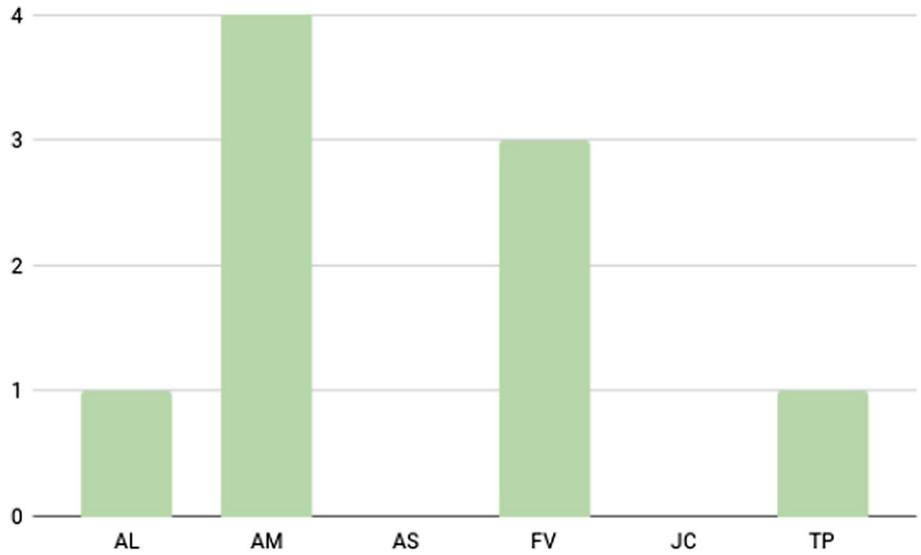


**Fig. 10** Classification of texts by HA Authors and AM using Imposters

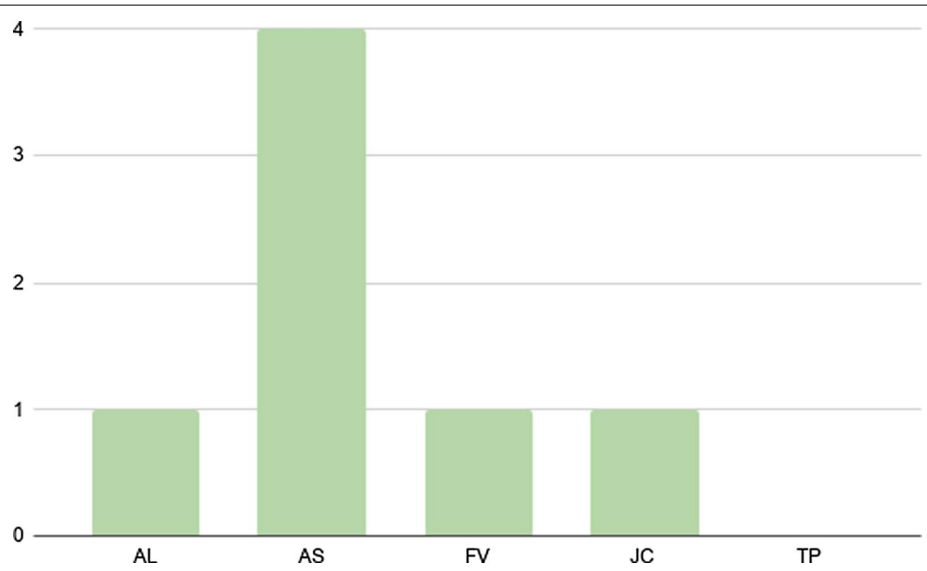


**Fig. 11** Net+Knn text classification of HA authors

We can conclude that the use of 100 words pieces of text still can give a similar result to state-of-the-art techniques such as GI for the Latin language. But we still cannot conclude much about the authorship of HA Lives.

In the next subsection we focus on the authorship of HA Lives.

#### Authorship of *Historia Augusta* Lives

In order to study HA Lives authorship, the same experiences were repeated, this time excluding Ammianus Marcellinus' texts. The number of well classified Lives (texts) using the two authorship methods is represented in Figs. 11 and 12.

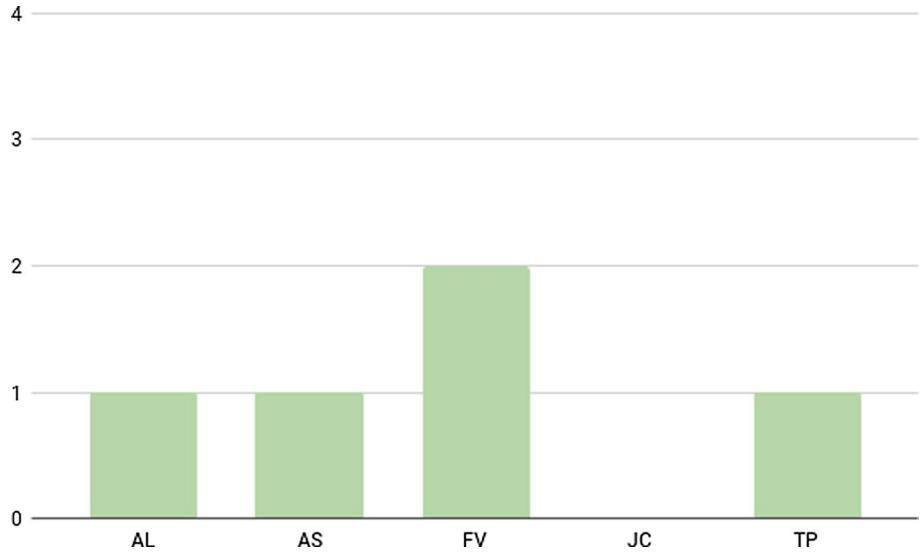


**Fig. 12** GI texts classification of HA authors

The Net+Knn method using 100 words graphs is able to identify four of the Lives by AS, none of the Lives by TP and one of each of the other authors.

Results obtained by our method point clearly to a single authorship, in spite of being surprisingly different in the case of AS. Indeed, the four Lives of this author were consistently attributed to him by Net+Knn.

Confronted with this evidence, we judged necessary to reassess the singularities of AS's Lives against other Lives. So, we proceeded then to evaluate authors in pairs: AS-JC, AS-TP and JC-TP (this pair was intended to function as the control pair). Figure 13 presents the classification using just these pairs of authors, calculated with GI and with Net+Knn.

As Fig. 13 shows, the accuracy when using GI and Net+Knn is substantially different. The results for the pair AS-JC are distinct: using GI, it appears that the authors exhibit the same profile (only the life of Septimius Severus is attributed with a percentage of 88%), whereas Net+Knn points to style differences between both authors. The used method distinguishes between the Lives by AS, on the one hand, and the four tested Lives by JC, on the other: 75% of accuracy for the latter and between 75 and 100% for the former. In addition to that, there are misattributed texts:

- for the pair AS/TP, *Tyranni triginta* by TP;
- for the pair AS/JC the *Septimius Severus* by AS and *Opellius Macrinus* by JC;
- for the pair JC/TP, *Tyranni triginta* and *Gallieni duo* by TP and *Gordiani tres* by JC.

Notwithstanding this result, i.e. the specificity of AS, it was necessary to assess whether it would stand in other contexts. We thought it was worth to conduct an experiment along the lines of Stover and Kestemont (2016), who pointed out stylistic differences between the authors of the so called *Hauptviten* and *Nebenviten*. The goal was to verify whether the specificity of AS would hold in the broader set of the division between the block of *Hauptviten* and that of *Nebenviten*—in fact, according to the traditional

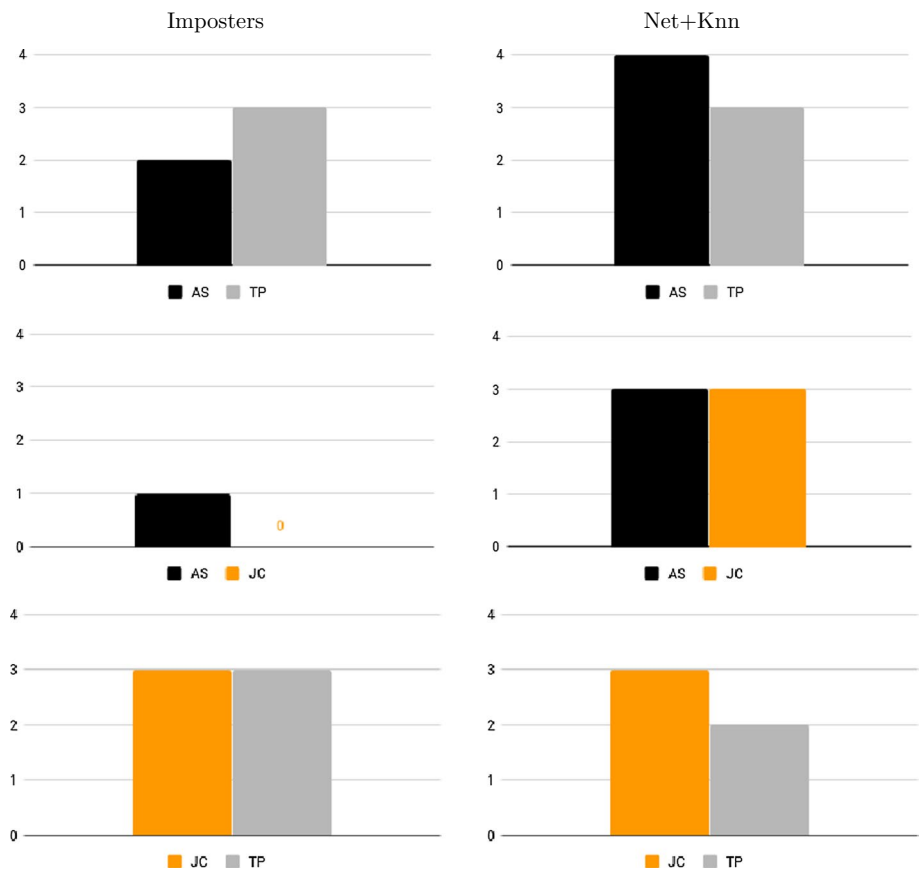


**Fig. 13** Pairs of HA Lives with GI and Net+Knn

authorship attribution, AS belongs exclusively to the *Hauptviten*. Net+Knn was run for the *Hauptviten* and the *Nebenviten*, and the test yielded the following results: out of the 9 *Hauptviten*, 9 Lives were correctly attributed to the group (100%); in the case of the *Nebenviten* 2 Lives (Aelius and Clodius Albinus) were correctly assigned to the group, whereas the remaining (Geta, Pescennius Niger and Avidius Cassius), were left out of the group. This means that within the block of *Hauptviten* the specificity of AS does not reveal itself.

## Conclusions

Net+Knn method yielded results close to 100% accuracy, both in the attribution of 20th century Portuguese authors and in the attribution of segments of Ammianus Marcellinus. Since these texts have no issues as to authorship attribution, the accuracy of the results gave us a sign of robustness of the Net+Knn method, which allowed us to apply it on a text with disputed authorship such as HA.

The partition of text into blocks of 100, 200 or 500 words was proved to be a good technique to improve authorship verification accuracy and the confidence interval. Taking text partitions increases the number of independent observations for each text, providing more evidence to model the text authorship classification. In a method

such as GI that models texts with n-grams frequency, taking text partitions degrades the classification performance.

The style marks encoded in co-occurrence graphs are different of those encoded in a n-gram frequency model. So, a system like Net+Knn can be used together with a system like GI in the study of texts authorship since they highlight different style marks.

The best number of words to split texts depends on the corpus we are studying. As we show, for Portuguese authors 200 or 500 words partition worked very well. Removing stop words or not, does not have a relevant impact on the results for Portuguese or Latin with Net+Knn, but in GI results without removing stop words degrade.

Concerning HA Lives, the following results were obtained (by means of the stated experiments):

(a) based on a set of samples from HA, Net+Knn fails the attribution defined by manuscript tradition. Still, it is able to hit 4 in 4 Lives of AS (Fig. 11);
(b) based on the same set of samples, Imposter's analysis fails the attribution defined by manuscript tradition, although it is more robust in the case of FV, whom it attributes 2 Lives (Fig. 12);
(c) As to the assessment with pairs of Lives (Fig. 13), Net+Knn concludes that AS is an author different from the others. The Imposters, in turn, identifies differences between TP and JC, but does not distinguish TP from JC.

These results do not validate the claim that HA was written by the six different authors designated by the manuscript tradition, since the experiments did not show robust delimitations between four of the traditional authors (JC, FV, TP, AL; VG was left out due to the scarce amount of text attributed to him). This fact is in line with the hypothesis of a single authorship (the most widely accepted perspective on the topic (Syme 1971; Adams 1972; Paschoud 1991; Burgersdijk 2010; Rohrbacher 2016; Cameron 2011; Savino 2017; Stover 2020). The singularity of AS, which emerges as a robust outcome, despite losing its relevance in the overall context—i.e., despite being the only author whom Knn+net attributes Lives matching the manuscript tradition; whereas for the remaining authors, JC, FV, TP, AL Knn+net does not attribute texts in a comparable robust manner—, turns up as an anomaly which requires analysis, and possible confirmation, in the frame of philological research.

**Abbreviations**

AL: Aelius Lampridius; ALobo: António Lobo Antunes; AM: Ammianus Marcellinus; AS: Aelius Spartianus; FV: Flavius Vopiscus; GI: General Imposters framework; JC: Julius Capitolinus; KNN: K-nearest neighbor; HA: Historia Augusta; Mia: Mia Couto; Net+HA: Accuracy of the graphs of texts by Historia Augusta authors; Net+HA+AM: Accuracy of the graphs of texts by Historia Augusta authors and A. Marcellinus; Net+Knn: Measurements of Complex Networks using a classifier build with KNN; Sara: José Saramago; +sw: With stop words; −sw: Without stop words; T+HA: Accuracy of texts by Historia Augusta authors; T+HA+AM: Accuracy of texts by Historia Augusta authors and A. Marcellinus; TP: Trebellius Pollio.

**Acknowledgements**

The present work is financed by National Funds through the Portuguese funding agency, FCT (Fundação para a Ciência e a Tecnologia) within NOVA LINCS center, Project ID UIDB/04516/2020. The authors also acknowledge: Center for Classical and Humanistic Studies, CECH-UC, through the project BioRom: PTDC/LLT-OUT/28431/2017, Centro de Investigação em Matemática e Aplicações (CIMA) through the grant UIDB/04674/2020 , research centers are supported by FCT (Fundação para a Ciência e a Tecnologia, Portugal) and, also, Departamento de Matemáticas, y Escuela Politécnica de Cáceres, de la Universidad de Extremadura, Spain.

**Authors' contributions**
Conceived and projected the calculations: all authors. Performed the calculations: Clara Grácio, Irene Pimenta Rodrigues, Juan Zapata and Lígia Ferreira. Analyzed the data and results and wrote the article: all authors. All authors read and approved the final manuscript.

**Availability of data and materials**
The datasets generated and/or analyzed during the current study are available in the Mendeley Data, repository: http://dx.doi.org/10.17632/fbdg5tjfb6.1, http://dx.doi.org/10.17632/3zx7ykdrh4.1, http://dx.doi.org/10.17632/z9hsp9cj2s.1. Martins, Armando; Grácio, Clara; Teixeira, Cláudia; Pimenta Rodrigues, Irene; Garcia-Zapata, Juan; Ferreira, Lígia (2020), "Latin-texts Authorship evaluation", Mendeley Data, v1http://dx.doi.org/10.17632/fbdg5tjfb6.1. Martins, Armando; Grácio, Clara; Teixeira, Cláudia ; Pimenta Rodrigues, Irene; Garcia-Zapata, Juan; Ferreira, Lígia (2020), "pt-texts authorship evaluation", Mendeley Data, v1http://dx.doi.org/10.17632/3zx7ykdrh4.1. Pimenta Rodrigues, Irene; Grácio, Clara; Ferreira, Lígia; Garcia-Zapata, Juan Luiz (2020), "Co-occurrence network - Augustan History", Mendeley Data, v1http://dx.doi.org/10.17632/z9hsp9cj2s.1

## Declarations

**Competing interests**
The authors declare that they have no competing interests.

**Author details**
[1]Department of Languages and Literature, University of Évora, Évora, Portugal. [2]Department of Mathematics, University of Évora, Rua Romão Ramalho, Évora, Portugal. [3]CIMA, Évora, Portugal. [4]Department of Informatics, University of Évora, Rua Romão Ramalho, Évora, Portugal. [5]NOVAlinks, Évora, Portugal. [6]Department of Mathematics, University of Extremadura, Badajoz, Extremadura, Spain. [7]Centre for Classical Studies, University of Lisbon, Lisbon, Portugal. [8]Center for Classical and Humanistic Studies (CECH), University of Coimbra, Coimbra, Portugal.

Received: 25 February 2021 Accepted: 16 May 2021
Published online: 06 July 2021

**References**

Adams J (1972) On the authorship of the Historia Augusta. Class Q 22(1):186–194
Akimushkin C, Amancio DR, Oliveira ON Jr (2017) Text authorship identified using the dynamics of word co-occurrence networks. PLoS ONE 12(1):1–15. https://doi.org/10.1371/journal.pone.0170527
Amancio DR, Antiqueira L, Pardo TA, da F Costa L, Oliveira Jr ON, Nunes MG (2008) Complex networks analysis of manual and machine translations. Int J Modern Phys C 19(04): 583–598
Amancio DR, Altmann EG, Oliveira ON Jr, da Fontoura Costa L (2011) Comparing intermittency and network measurements of words and their dependence on authorship. New J Phys 13(12):123024
Baker R (2014) A study of a late antique corpus of biographies [historia augusta]. Ph.D. thesis, Oxford University, UK
Bollobás B, Riordan OM (2005) Mathematical results on scale-free random graphs, chap. 1, pp 1–34. Wiley, New York
Burgersdijk DWP, et al (2010) Style and structure of the historia augusta. Ph.D. thesis, Universiteit van Amsterdam [Host]
Cameron A (2011) The last pagans of Rome. Oxford University Press, New York
Cvetković DM, Doob M, Sachs H (1980) Spectra of graphs: theory and application. Academic Press, New York
de Arruda H F, Marinho VQ, da F Costa L, Amancio D R (2019) Paragraph-based representation of texts: a complex networks approach. Inf Process Manag 56(3):479–494 . https://doi.org/10.1016/j.ipm.2018.12.008
Dessau H (1889) Über zeit und persönlichkeit der scriptores historiae augustae. Hermes, pp 337–392
Dorogovtsev SN, Mendes JFF (2001) Language as an evolving word web. Proc R Soc Lond Ser B Biol Sci 268:2603–2606
Eder M, Rybicki J, Kestemont M (2016) Stylometry with r: a package for computational text analysis. R J 8(1):107–121
Gera R, Alonso L, Crawford B, House J, Mendez-Bermudez J, Knuth T, Miller R (2018) Identifying network structure similarity using spectral graph theory. Appl Network Sci 3(1):2
Hengst Dd, Burgersdijk DWP, Waarden JAv (2010) Emperors and historiography: collected essays on the literature of the Roman Empire. Mnemosyne, bibliotheca classica Batava. Supplementum; v. 319. Brill, Leiden
i Cancho RF, Solé RV (2001) The small world of human language. Proc R Soc Lond Ser B Biol Sci 268:2261–2265
Kestemont M, Stover J, Koppel M, Karsdorp F, Daelemans W (2016) Authenticating the writings of julius caesar. Expert Syst Appl 63:86–96
Koppel M, Winter Y (2014) Determining if two documents are written by the same author. J Assoc Inf Sci Technol 65(1):178–187
Kulig A, Drożdż S, Kwapień J, Oświecimka P (2015) Modeling the average shortest-path length in growth of word-adjacency networks. Phys Rev E 91(3):032810
Marinho VQ, Hirst G, Amancio DR (2016) Authorship attribution via network motifs identification. In: 2016 5th Brazilian conference on intelligent systems (BRACIS), pp 355–360 . https://doi.org/10.1109/BRACIS.2016.071
Marriott FH (1979) Barnard's monte carlo tests: how many simulations? J Roy Stat Soc Ser C (Appl Stat) 28(1):75–77
Mehri A, Darooneh AH, Shariati A (2012) The complex networks approach for authorship attribution of books. Physica A 391(7):2429–2437
Momigliano A (1954) An unsolved problem of historical forgery: the scriptores historiae augustae. J Warburg Courtauld Inst 17(1/2):22–46
Murakami M, Ishikura S, Kominami D, Shimokawa T, Murata M (2017) Robustness and efficiency in interconnected networks with changes in network assortativity. Appl Network Sci 2(1):6

Newman MEJ (2006) The structure and dynamics of networks. In: Princeton studies in complexity
Paschoud F (1991) L'Histoire Auguste et Dexippe. Università degli Studi di Macerata
Quispe LV, Tohalino JA, Amancio DR (2021) Using virtual edges to improve the discriminability of co-occurrence text networks. Physica A 562:125344
Rocha JL, Fernandes S, Grácio C, Caneco A (2015) Spectral and dynamical invariants in a complete clustered network. Appl Math Inf Sci 9(5):2367
Rodrigues IP, Ferreira CGL, Garcia-Zapata JL (2020) Co-occurrence network-augustan history. https://doi.org/10.17632/z9hsp9cj2s.2
Rohrbacher D (2016) The play of Allusion in the Historia Augusta. University of Wisconsin Press, Madison
Savino E (2017) Ricerche sull'Historia Augusta. Naus, Naples
Segarra S, Eisen M, Ribeiro A (2015) Authorship attribution through function word adjacency networks. IEEE Trans Signal Process 63(20):5464–5478
Segarra S, Eisen M, Ribeiro A (2013) Authorship attribution using function words adjacency networks. In: 2013 IEEE international conference on acoustics, speech and signal processing, pp 5563–5567 (2013). https://doi.org/10.1109/ICASSP.2013.6638728
Spielman DA, Teng SH (1996) Spectral partitioning works: planar graphs and finite element meshes. In: Proceedings of 37th conference on foundations of computer science, pp 96–105. IEEE
Stamatatos E (2009) A survey of modern authorship attribution methods. J Am Soc Inf Sci Technol 60(3):538–556
Stanisz T, Kwapień J, Drożdż S (2019) Linguistic data mining with complex networks, a stylometric-oriented approach. Inf Sci 482:301–320
Stover JA (2020) New light on the historia augusta. J Roman Stud 110:167–198
Stover JA, Kestemont M (2016) The authorship of the historia augusta: two new computational studies. Bull Inst Class Stud 59(2):140–157
Syme R (1971) Emperors and biography: studies in the "Historia Augusta"/by Sir Ronald Syme. Clarendon Press, Oxford
Teixeira C, Rodrigues I (2018) Deciphering Latin sentences using traditional linguistic resources. Digital Scholarship Humanit 34(4):791–805
Tse EK, Tweedie FJ, Frischer BD (1998) Unravelling the purple thread: function word variability and the scriptores historiae augustae. Literary Linguist Comput 13(3):141–149
White L (1967) The historical roots of our ecologic crisis. Science 155(3767):1203–1207

## Publisher's Note