2018

Network Data Analysis of Word Graphs With Applications to Authorship Attribution

Timothy Leonard
University of Rhode Island, timothy_leonard@uri.edu

NETWORK DATA ANALYSIS OF WORD GRAPHS WITH APPLICATIONS TO AUTHORSHIP ATTRIBUTION

BY

TIMOTHY LEONARD

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF THE

REQUIREMENTS FOR THE DEGREE OF

MASTER OF SCIENCE

IN

COMPUTER SCIENCE

UNIVERSITY OF RHODE ISLAND

2018

MASTER OF SCIENCE THESIS

OF

TIMOTHY LEONARD

APPROVED:

Thesis Committee:

Major Professor Noah Daniels

Natallia Katenka

Lutz Hamel

Lubos Thoma

Nasser H. Zawia

DEAN OF THE GRADUATE SCHOOL

UNIVERSITY OF RHODE ISLAND

2018

ABSTRACT

Network data analysis is an emerging area of study that applies quantitative analysis to complex data from a variety of application fields. Methods used in network data analysis enable visualization of relational data in the form of graphs and also yield descriptive characteristics and predictive graph models. This thesis shows that a representation of text as a word graph produces the well documented feature sets used in authorship attribution tasks such as the word frequency model and the part-of-speech (POS) bigram model. This thesis applies nominal assortativity of parts of speech, a network data characteristic of word graphs, to the problem of authorship attribution and shows how these features are produced from a word graph model. Specifically, it is shown that the nominal assortative mixture of parts of speech, a statistic that measures the tendency of words of the same POS in a word network to be connected by an edge, produces a feature set that can be used to predict authorship. These results are compared to the POS bigram model, a highly accurate authorship attribution model, and show that the nominal assortativity model is competitive. Analysis of these models along with word graph characteristics provides insights into the English language. Particularly, analysis of the nominal assortative mixture of parts of speech reveals regular structural properties of English grammar.

ACKNOWLEDGMENTS

I would like to thank Dr. Natallia Katenka, Dr. Noah Daniels, Dr. Marco Alvarez, and Dr. Lutz Hamel for their input into the content of this thesis. Dr. Alvarez suggested I attempt the problem of authorship attribution. Dr. Katenka and Dr. Hamel taught me the tools used to build and validate the model. Dr. Katenka and Dr. Daniels urged me to publish the material presented here.

TABLE OF CONTENTS

Page

Page

LIST OF FIGURES

Figure **Page**

LIST OF TABLES

1 Introduction [1]

Attempts to quantitatively evaluate writing date as early as the 19th century with studies on Shakespeare's plays [3]. By the mid 20th century Bayesian statistical analysis was applied to a small set of common words to speculate over the authorship of the Federalist Papers [3] [4]. The problem of authorship attribution falls into the domain of natural language processing (NLP) and includes uncovering plagiarism, determining ghost-writership and pen names, and speculating over the authorship of unsigned supreme court decisions or anonymous blogs [5] [6]. As is the case with most data applications in the 21st century, there is a wealth of written data: large online text resources, blogs, community message boards such as Twitter, Reddit, and Facebook, and traditional print sources such as newspapers and books. Not surprisingly, there are a multitude of models designed to classify authorship. These models are often quite complex both theoretically and computationally and do not produce straightforward descriptions of language. The network analysis techniques outlined in this paper are straightforward and provide interesting descriptions of the English language. The purpose of this article is to replicate a part-of-speech bigram model that has been used with prior success at authorship prediction [6] [7] and show that it is part of a larger word-network model [8]. From this network model, network data analysis allows us to observe structural regularities of English grammar.

The part-of-speech (POS) bigram model is described as frequencies of pairs of consecutive parts of speech in a sentence. For instance the sentence "the dog ate" comprises two POS bigrams: a determiner and noun pair, and a noun and verb pair. There are 36 parts of speech identified by the Penn Treebank. The Cartesian product of these 36 parts of speech is the POS bigram feature set.

One desirable trait of the POS bigram model is its reduced feature space.

In contrast to the 36 parts of speech, there are hundreds of thousands of English words. While modern technology can manage large volumes of data faster than ever before, and machine learning techniques such as random forests, support vector machines (SVMs), and neural networks can handle numerous dimensions efficiently, the high dimensionality of language data requires transformation or reduction of the data to manage dimensionality [6] [7] [9]. The POS bigram model enables speedy analysis of the English language through a reduced yet informative feature space.

2 Literature Review [1]

To compensate for high dimensionality it is common to target a single characteristic of language for inclusion in a feature set, ignoring others. Broadly speaking there are two types of authorship attribution models: those that model the text's content and those that model its structure [6] [7]. Content models encompass topical information such as word frequencies. Stylometric models capture the structure of a language such as the frequency of POS bigrams in a sentence.

Diedrich, *et al.* [6] applied support vector machines to the two categories of models mentioned above in order to evaluate their authorship prediction performance. The first model captured content by recording word frequencies. The second model combined function-word frequencies with frequencies of POS bigrams between tagged nouns, verbs, and adjectives. Diedrich, *et al.* regarded this as a structural model since it modeled grammar [6].

Diedrich, *et al.* found that reducing content words to generic parts of speech had a deleterious effect on accuracy. However, their method was a hybrid POS/word frequency model and represented POS bigrams incompletely [6]. Hirst and Feiguina, on the other hand, found that a POS bigram model could discern accurately between Charlotte and Emily Bronte, sisters whose writing is known to be difficult to distinguish from each other's [7]. We present a model that offers a more complete representation of POS bigrams than those of Diedrich, *et al.* and Hirst, *et al.*, and is also highly predictive of authorship, validating Hirst and Feiguina's conclusions and justifying a deeper analysis of the POS bigram model.

As a vehicle for our analysis, this paper examines the nominal assortative mixture of English parts of speech and applies the results to the problem of authorship attribution. Nominal assortative mixture measures the selective linking, by attribute, of vertices in a network model [10] [11]. In a word-network model,

words are vertices and edges occur between sequential words that appear in a sample of writing. To calculate nominal assortativity of parts of speech and generate a feature set for authorship attribution tasks, each vertex has as an attribute a part of speech that can be compared to other vertices. The nominal assortative mixture coefficients for parts of speech derive directly from the POS bigram model which is part of this more complex word-network model.

Previous research applying network data analysis to authorship attribution has not revealed network statistics to have predictive value [1]. In 2006 Antiqueira *et al.* proposed applying network analysis to the problem of authorship attribution adding the observation that some authors, while less than the majority represented in the sample, cluster by network characteristic when plotted visually [13]. Research on word graphs sought to discover predictive features using network data analysis but did not include the part of speech as an attribute of a vertex. Lahiri and Mihalcea showed that although descriptive characteristics such as transitivity, clustering coefficient, density, etc., are not significant predictors, they may be beneficial when included alongside other features such as word frequencies [14]. Amancio *et al.* also concluded that word network characteristics could be used in conjunction more traditional approaches by examining additional network characteristics such as shortest path, betweeness, and intermittency but did not achieve remarkable accuracy with any feature set created by a combination of 15 network characteristics [15] [16]. Mihalcea and Radev measured degree assortativity which measures the tendency for vertices of same degree to be connected by an edge [17], as opposed to nominal assortativity which compares the attributes of a vertex.

[1] Marinho, Hirst, and Amancio report the results from various authorship attribution techniques all with varying degree of success [12]. The problem with comparing accuracy results across different experiments is the lack of control over experimental design. Different trials include different number of authors, different size text, in addition to different scoring techniques. The bigram model proposed by Hirst and Feiguina achieved very high accuracy, but on only two authors [7] [12].

Foster *et al.* recognized the strongly negative degree assortativity when applying the pearson corellation, however, also kept their discussion limited to degree-degree assortativity [18]. These previous studies did not examine nominal assortativity of parts of speech. We find that while not as powerful as the POS bigram model, the POS assortiative mixture model is competitive at authorship prediction.

More to the purpose of this article, we offer a meaningful description of the English language in a way other discussions on authorship attribution regularly fail to produce. Neural networks are especially criticized for being "black box" models because neuron weights hidden in multiple layers do not naturally correspond to language features. Comparatively, while it may be natural to count word frequencies, the feature set by itself does not offer intuition about language. Zipf discovered that word frequency is inversely proportional to rank [19] [6] [20], an empirical law [21] observable by plotting word frequencies in sorted order. In simplified mathematical terms, the sum of the relative frequencies is the harmonic series [21]. The application of Zipf's law is an approximation, not all corpora follow identical word frequency distributions, however, in general Zipf's approximation holds across languages including English and Chinese [22]. With the goal of continuing statistical insight into language, network data analysis provides an avenue for further exploration of linguistic relationships.

By measuring the tendency for same parts of speech to collocate, the nominal assortative mixture model offers a glimpse into the characteristics of English stylometry. Assortative mixture captures fundamental language characteristics such as what same parts of speech pairs do and do not regularly occur in writing. For instance, we find that the determiner-noun pair is a disassortative bigram that occurs often in the English language. These frequent diassortative pairs contribute to an overall distribution of English parts of speech that is disassortative. However there

are parts of speech that do exhibit assortative qualities such as adjective-adjective pairs. As a feature set the nominal coefficients distinguish stylistic preferences between authors, yet regularities across authors reveal a grammar "signature" for the English language.

The rest of this paper is outlined as follows: We begin with a definition of the authorship classification problem in Section 3.1 and a description of the data in Section 3.2. In Section 3.3 we describe the network model. In Section 3.4 we show how the POS bigram model is constructed from the graph object, and explain how to calculate the nominal assortativity coefficients from POS bigram data. In section 4.1 we visualize the data, describe network characteristics such as graph density, and discuss the relationship between nominal assortativity and degree assortativity for word graphs. In Section 4.2 we report the model testing results for authorship prediction and conclude with a brief discussion in Section 5.

3 Methodology [1]

3.1 Authorship Attribution Problem Formulation

Authorship attribution is a classification problem. The authors of text are the classes and their related works are represented as a feature set. The task is to apply the feature set as a labeling function to accurately discern authorship given labeled training data. While the first step is described in Sections 3.4, the second step is formally presented as follows, given:

1. a universe X of n written works by m authors $A = \{A_1, \cdots, A_m\}$, such that for each author $A_i | i \in \{1, \cdots, m\}$ there exists a vector $A'_i = [a_{i_1}, \cdots, a_{i_k}]$, where k is the number of works written by author A_i and a_{i_j} is the j^{th} work by author A_i,

2. a sample $S | S \subset X$,

3. a target labeling function $f(x) : S \rightarrow A$,

4. a labeled training set D, where D are all the pairs (x, y) such that $x \in S$ and $y = f(x)$

compute a function $\hat{f}(x) : S \rightarrow A$ from D such that $\hat{f}(x) \simeq f(x)$ for all x in X.

In the above definition, the universe of written works X includes n written works from m distinct authors. Each label a_i corresponds to an author labeled 1 to m. Labels are applied to the sample S of X such that each data point is labeled with a single author to produce training dataset D. Since the target labeling function $f(x)$ is not truly known and can be applied in retrospect only, the function $\hat{f}(x)$ is an approximation of the original function $f(x)$ [23].

3.2 Data

The data set analyzed included 5 authors chosen from a subset of the Gutenberg data set made available by Michigan University [24]. The authors were Jerome

Klapka Jerome (1859-1927), Thomas Hardy (1840-1928), Sir Arthur Conan Doyle (1859-1930), Jane Austen (1775-1817), and Nathaniel Hawthorne (1804-1864). For each experiment each author was represented by 30 fixed length excerpts (between 125 and 10000 words) taken from their larger written works. Each sample was manually pre-processed to remove the author‘s names, chapter titles, chapter numbers, subtitles, author‘s notes, editor‘s notes, and extraneous syntax such as brackets and asterisks. The main purpose of cleaning the data was to avoid speech tagger errors and remove extraneous information.

3.3 Word-Network Model

A word-network model is a directed graph *G=(V, E)* with a set V of vertices represented as unique words and a set E of edges, where elements of E are ordered pairs u, v, or bigrams, of distinct words $u, v \in V$ appearing consecutively within sentences in a sample text. The direction of an edge is consistent with the order in which two words occur within each sentence, but edges do not span from a word that ends a sentence to one that begins the next. Each edge represents a unique word bigram, and its weight corresponds to its frequency. The degree d_v of a vertex v, in a word graph G, counts the number of edges (bigrams) in E incident upon v. By computing the out-degree of each vertex in G, one can construct the word frequency model[2] discussed in the introduction. Each vertex in a word graph is attributed with its part of speech. By reducing the word graph G to a POS graph $G_p = (V_p, E_p)$, where a set V_p of vertices represent unique POSs and a set E_p of directed edges represents unique POS bigrams, we can count edge weights to produce the POS bigram frequency model.

The directed graph in Figure 1 represents the following sentences:

[2]it is necessary for calculating vertex degree to connect words that end a sentence to a dummy *end* vertex.

- The quick brown fox jumped over the lazy dog.
- A fox jumped over Sir Walters the lazy dog

The representation of text as a word graph produces the well documented feature sets such as the word frequency model and the part-of-speech bigram model used in authorship attribution tasks (see Table 1 and Table 2). Additionally, this graph representation allows application of various network data analysis methods, such as the reporting of network characteristics including degree distribution, graph density, and nominal assortativity.

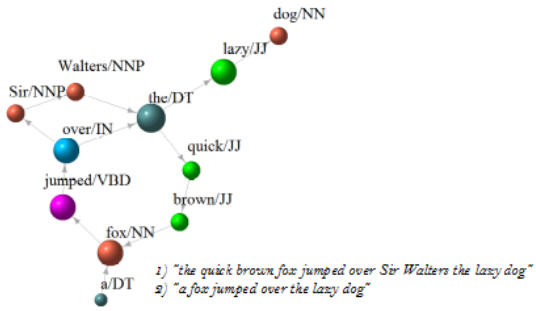

Figure 1. Two sentences represented as a directed word graph. Each vertex is a word attributed with the part of speech in the form `word/POS`. The direction of an edge is consistent with the order of the words written.

3.4 Feature Sets

In this section we take a closer look at word graph analysis and the part-of-speech bigram model as tools for feature set selection for authorship authentication and outlook for the structure of English grammar.

Part of Speech Bigrams

POS bigrams represent adjacency between two consecutive parts of speech as described in Section 3.3. POS bigram frequencies are derived from the word graph representation in Table 2. The feature set includes 34 of 36 Penn Treebank parts

Table 1. Word bigram matrix of the two example sentences.

	the	a	quick	brown	fox	jumped	over	Sir	Walters	lazy	dog
the/DT	0	0	**1**	0	0	0	0	0	0	**1**	0
a/DT	0	0	0	0	**1**	0	0	0	0	0	0
quick/JJ	0	0	0	**1**	0	0	0	0	0	0	0
brown/JJ	0	0	0	0	**1**	0	0	0	0	0	0
fox/NN	0	0	0	0	0	**1**	0	0	0	0	0
jumped/VBD	0	0	0	0	0	0	**1**	0	0	0	0
over/IN	0	0	0	0	0	0	0	**1**	0	0	0
Sir/NNP	0	0	0	0	0	0	0	0	**1**	0	0
Walters/NNP	**1**	0	0	0	0	0	0	0	0	0	0
lazy/JJ	0	0	0	0	0	0	0	0	0	0	**1**
dog/NN	0	0	0	0	0	0	0	0	0	0	0

Table 2. POS bigram matrix of the two example sentences.

	DT	JJ	NN	VBD	IN	NNP
DT	0	**2**	**1**	0	0	0
JJ	0	**1**	**2**	0	0	0
NN	0	0	0	**1**	0	0
VBD	0	0	0	0	**1**	0
IN	**1**	0	0	0	0	**1**
NNP	**1**	0	0	0	0	**1**

of speech.[3] Hence, the resulting feature space is a 1156 (34 x 34) element vector where each element is an ordered pair of sequential parts of speech.

Hirst and Feiguina used Cass, a partial parser to tag parts of speech from short text and construct POS bigrams. The choice to do partial parsing was a compromise between quick computations and complete parsing. While not as accurate as complete parsing it was "accurate enough" [7].

This paper uses the POS tagger from the Stanford Natural Language Processing Group. It achieves high accuracy and is very fast even on large documents [25]. The accuracy of the Stanford POS tagger enables more complete parsing of syntactic labels compared to the partial parser used by Hirst and Feguina. We do not

[3]excluding symbols and list items markers

feel it is necessary to compare Hirst and Feguina's POS bigram model to our own, however, since our results support their conclusion that the POS bigram model distinguishes between authorship on small samples. Instead we expect that more accurate and detailed parsing will improve results.

Tagging parts of speech for each sample, we represent these samples as word graphs and produce the the feature set from the POS bigram frequencies. We apply classification tools including random forest and support vector machines for authorship classification. The results are summarized in Section 4.

Assortative Mixture of Parts of Speech

A word-graph model can be summarized by characteristics including degree distribution, density, and assortativity. Nominal Assortativity is a vector of coefficients ranging between 1 and -1, each of which measures the absolute tendency for graph vertices with the same attribute to share an edge. Positive coefficients indicate positive assortativity, negative coefficients indicate negative assortativity. The assortativity coefficient is analogous to the Pearson correlation coefficient[11][26]. The attribute being measured in this paper is the part of speech. For parts of speech in a word graph, a positive assortativity coefficient suggests words of the same POS occur sequentially, while negative assortativity suggests they do not. The assortativity coefficient is calculated for each of 34 POSs to generate a feature set of 34 elements. For each POS i, a nominal assortativity coefficient r_i can be computed as:

$$r_i = \frac{\sum f_{ii} - \sum f_{ir} f_{ic}}{1 - \sum f_{ir} f_{ic}} \tag{1}$$

where f_{ii} is the fraction of edges in a graph G that join a vertex in the ith category to a vertex in the same (ith) category, f_{ir} and f_{ic} are the marginal row and column sums respectively (see [10] [11] for more details). For the results in

this paper, the assortativity coefficient was calculated using word graph objects with directed edges.

As an illustration, we computed nominal assortativity coefficients for the toy sentences represented as a directed word graph visualized in Figure 1. Specifically, to calculate the assortativity for POS, applying Equation (1) to the values in Table 2 reveals that the parts of speech DT, NN, VBD, and IN all have assortativity coefficients of -1, while JJ and NNP have coefficients -0.202 and 0.333 respectively. Here the parts of speech constitute a feature set of five elements. This example supports our more general finding for larger data samples that parts of speech possess some assortative (disassortative) properties.

3.5 Motifs

Motifs are defined as "small subgraphs occurring far more frequently in a given network than in comparable random graphs" [11]. To Marinho, Hirst, and Amancio, "the topology of a complex network is characterized by the number of motifs found on its structure" [12]. Marinho *et al.* examined word graph motifs between three vertices (see [27] [11] for more information on motifs. Figure 2 shows the 13 possible motifs from three vertices.) and applied motif frequency to the problem of authorship attribution [12]. In their discussion of past applications of motifs to word networks, Marinho *et al.* commented on the work of Milo *et al.* who showed that the languages of English, Spanish, French, and Japanese shared similar motif profiles [12] [2]. Marinho *et al.* suggested as an explanation for similarities between languages that "languages possess an intrinsic structure, which divides words into categories" where "words from one category (e.g. prepositions) tend to be with others from different categories (e.g. nouns or articles) [12]. Marinho's conjecture that word category collocation is dissasortative is supported by the results in this paper.

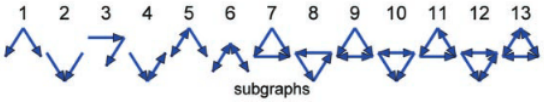

Figure 2. 13 motifs from three vertices. Motifs 1 - 6 are connected triples but not triangles. Motifs 7 - 13 are triangles. Image taken from [2]

Clustering Coefficient

If the three vertex motifs described in [12] [2] are not random , thus constituting motifs, one indication would be that network characteristics generated by random graphs are significantly different than characteristics of the true network. One characteristic that applies to three vertex motifs is the clustering coefficient. The clustering coefficient, also known as transitivity, measures the proportion of connected triples that are triangles to those that are not [28] [11]. The ratio between these two generic motifs (triangles, motifs 7 - 13 figure 2 vs connected non-triangles, motifs 1 - 6 figure 2) describes which motif has a larger presence. A graph exhibiting transitivity suggests there is a high proportion of triangles compared to connected triples (that are not triangles), and vice verse.

To test the assumption that the presence or absence of these motifs is not random, the transitivity for every sample of 1000 POS or more is compared correspondingly to the transitivity of 1000 random graphs produced using the same in-degree and out-degree as each sample (for details see degree.sequence.game random graph generator from iGraph [11]).

The transitivity for word graphs of writing samples greater than or equal to 1000 POS is modeled using linear regression and visualized in figure 8. In equation 2, the dependent variable y is the clustering coefficient for a word graph from a sample of n POS greater than or equal to 1000. The independent variable x_{11} is the number of vertices for a word graph representing a single sample. The discreet categories of sample size (1000, 2000, 3000, 4000, 5000, 6000, 7000, 8000,

9000, 10000) are categorical dummy variables x_1 through x_{10}. The interaction between number of vertices x_{11} and the dummy variables x_1 through x_{10} produces 10 separate regression lines, one for each category of fixed sample size.

$$y = \beta_0 + \beta_1 * x_2 + \beta_2 * x_3 + \ldots + \beta_{10} * x_{11} + \beta_{11} * x_2 * x_{11} + \beta_{12} * x_3 * x_{11} + \ldots + \beta_{19} * x_{10} * x_{11} + \epsilon \quad (2)$$

In equation 2 above, β_0 is the coefficient for the slope. β_1 through β_{10} are the coefficients for the dummy categorical variables for sample size (x_2 through x_{10}). β_{11} is the coefficient for the dependent variable x_{11}. β_{12} through β_{19} are the coefficients for the interaction terms. Lastly, ϵ is the error term. Regression was applied under the assumptions that there is a linear relationship between the dependent and independent variables y and x_{11} and that the distributions of the variables are multivariate normal. Additionally it is assumed that there is no collinearity between variables and that the error values are evenly distributed along the regression line.

4 Results [1]

4.1 Visualizations And Descriptive Analysis

In this section we demonstrate how the written text samples of the selected authors listed in Section 3.2 can be visualized and characterized using word graphs described in Section 3.3. Focusing on nominal assortativity coefficients of parts of speech, we compare different authors and explore if their individual preferences for POS usage elucidate their writing structure. Next, we apply the POS bigram and POS assortative mixture models outlined in Section 3.4 to the problem of authorship attribution using the data described in Section 3.2. We compare these models in terms of predictive accuracy for authorship of various writing sample sizes.

We begin by constructing a word graph for a selected writing sample. Figure 5 shows a visualization of the word graph obtained from a random sample from Hawthorne. Descriptive characteristics of the networks for randomly selected writing samples from each author, including the total number of words, number of vertices and edges, graph density, and degree assortativity are summarized in Table 3. We will discuss these characteristics later in this section.

Turning our attention to nominal assortativity, we computed POS nominal assortativity coefficients for each author and visualized them in Figure 3. Note that each colored bar in Figure 3 corresponds to an author and shows the range of the assortativity coefficient for a POS. Positive assortative values indicate an author's preference for selectively linking the same POS, while negative values indicate preference for dissassortative relationships between twin bigram pairs. While the magnitude of each coefficient does not reliably measure the magnitude of assortativity, the sign of the coefficient does distinguish between assortative and disassortative preferences [26] [29]. Comparing different authors, it appears that individual preferences for POS usage differentiates writing style. While only

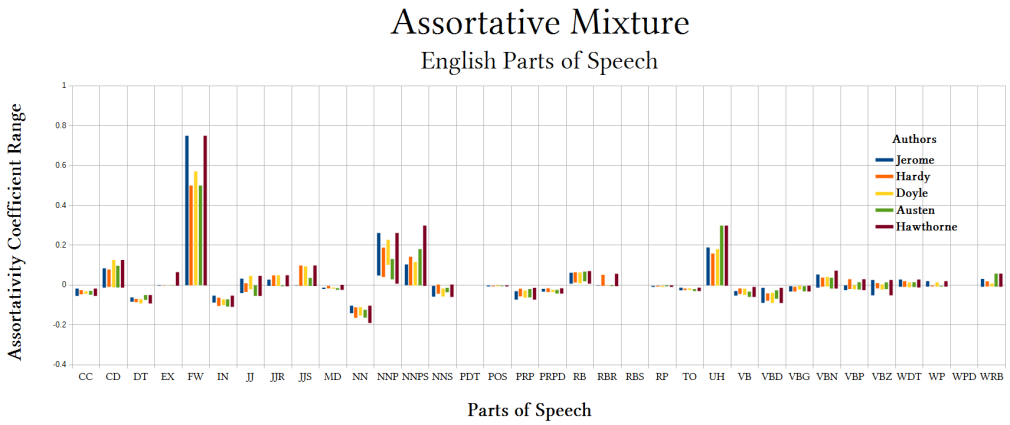

Figure 3. A visual representation of the vector space for each author. Each color corresponds to an author, and each bar shows the range of the assortativity coefficient for a part of speech.

	Tag	Description			
1.	CC	Cordinating conjunction	19.	PRPD	Possessive pronoun
2.	CD	Cardinal Number	20.	RB	Adverb
3.	DT	Determiner	21.	RBR	Adverb, comparative
4.	EX	Existential there	22.	RBS	Adverb, superlative
5.	FW	Foreign word	23.	RP	Particle
6.	IN	Preposition or subordinating conjunction	24.	SYM	Symbol
7.	JJ	Adjective	25.	TO	to
8.	JJR	Adjective, comparative	26.	UH	Interjection
9.	JJS	Adjective, superlative	27.	VB	Verb, base form
10.	LS	List item marker	28.	VBD	Verb, past tense
11.	MD	Modal	29.	VBG	Verb, gerund or present participle
12.	NN	Noun, singular or mass	30.	VBN	Verb, past participle
13.	NNS	Noun, plural	31.	VBP	Verb, non-3rd person singular present
14.	NNP	Proper noun, singular	32.	VBZ	Verd, 3rd person singular present
15.	NNPS	Proper noun, plural	33.	WDT	Wh-determiner
16.	PDT	Predeterminer	34.	WP	Wh-pronoun
17.	POS	Possessive ending	35.	WPD	Possessive wh-pronoun
18.	PRP	Personal pronoun	36.	WRB	Wh- adverb

Figure 4. Penn TreeBank POS tags color coded to match groupings used in Figure 5 and Figure 6

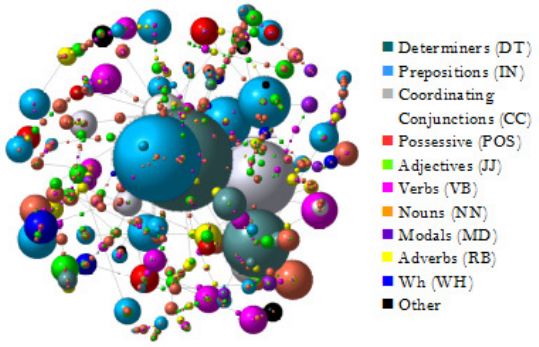

Figure 5. Minimum spanning tree representation of a writing sample from Hawthorne. Each vertex is a word with size proportional to frequency colored by part of speech.

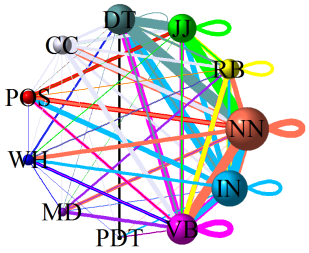

Figure 6. A directed word graph where vertices represents parts of speech. The size of a vertex is proportional to POS (word) frequency. Each edge is weighted by frequency of the out-degree of a vertex and edge colors are the direction of the edge. Self loops indicate positive nominal assortativity, while edges between nodes are disassortative connections.

five authors are tested here, experiments using fewer authors yielded improved results. It would be expected that more authors would see diminished results, but not dramatically for the inclusion of just one additional author. While not significantly as predictive as the POS bigram model, when applied to authorship attribution tasks the POS assortativity model is competitive (see Section 4.2).

In general, across authors, one can observe from Figure 3 that determiners (DT) and coordinating conjunctions (CC) exhibit disassortativity, while proper nouns (NNP) exhibit positive assortativity. The coefficients for parts of speech where the range includes zero indicate parts of speech that do occur sequentially some of the time but not necessarily always, e.g. adjectives (JJ). Foreign words (FW) are generically labeled as such and are not differentiated by the POS tagger, an artifact that produces highly assortative values. These findings seem indicative of the English language structure in general. Consistencies across authors suggests an assortative "signature" for English grammar.

As opposed to nominal assortativity, which produces an assortativity coefficient for each distinct attribute of a vertex (the part of speech in the case of word graphs), degree assortativity is a single coefficient that measures the likelihood for vertices of fixed degree to attach to other vertices of the same degree (see [11]). For word graphs, degree assortativity is negative, reflecting the fact that the most used parts of speech (nouns, verbs prepositions, and determiners) are disassortative by type (see Figure 6). However, degree disassortativity also accounts for the fact that, for instance, that there are many determiner-to-noun transitions in English speech [4] (see Figure 6), with a few hub determiners connected to a diverse set of smaller-degree nouns (see Figure 5).

Table 3 gives descriptive statistics of a single writing sample of size 3000 POS. Recall that a vertex in a word graph represents a unique word, and the

[4] consider the French use of articles *le* and *la*

Table 3. Descriptive statistics of five writing Samples with 3000 words

	Hawthorne	Austen	Jerome	Hardy	Doyle
Vertices	1359	896	1213	1254	1181
Edges	2661	2591	2586	2662	2513
Density	0.001420194	0.00287061	0.001671286	0.001621667	0.001734514
Degree Assortativity	-0.2302739	-0.26577	-0.2638668	-0.2265096	-0.287129
Transitivity	0.01468575	0.03807761	0.01992584	0.01811918	0.02269955
Number of Words	3256	3154	3230	3228	3224

total number of vertices in a word graph represents the total number of unique words used by a particular author within a single writing sample. Since the the number of vertices (unique words) determines the number of possible unique word bigram edges, it follows that the more unique words an author uses within a span of 3000 POS, the less dense the graph. This is a consequence of Zipf's observation that word frequency is inversely proportional to rank usage. In order for a new vertex to contribute to an increase in word graph density it must form unique edges with enough other vertices to exceed the ratio of vertices to edges from the previous graph missing the new vertex. However, by Zipf's law, a new word should contribute fewer edges to the rest of the graph than previous words because, by virtue of having a low frequency rank (a new word occurs once), also as a consequence of occurring later in time, is it used less frequently given a fixed sample size. This is supported by the information in table 3. Austen, with the fewest number of unique words used, has the most dense graph, while Hawthorne has the least dense graph because of his broader vocabulary use. However, the same correlation does not occur with degree assortativity. Austen and Jerome have approximately the same degree assortativity but the number of vertices differ substantially.

4.2 Model Testing

In this section we apply the POS bigram and POS nominal assortativity model to authorship prediction. For each of five authors, we use 30 written text samples of fixed length to derive the two models following the procedures described above. We evaluate and compare these models in terms of accuracy for various text sizes using support vector machines (SVMs) and random forests via 10 fold cross validation.

The results reported in Table 4 suggest the POS bigram model with support vector machines is highly accurate confirming the power of the POS bigram model. Even on small data sets of only 125 words the POS bigram model performed well. The assortative mixture model, on the other hand, performed best using random forests but could not achieve the near perfect accuracy of the POS bigram model on samples of larger text. The 90% confidence intervals were calculated using the bootstrap method on the 10 fold test statistic for the best performing classifier given each model. While significantly not as predictive, the assortative mixture model does perform competitively, especially for larger text sizes.

The assortativity model was also applied using a single layer neural network, however, the results were mediocre compared to support vector machines and random forests. While a more complicated neural network may perform more optimally, the pursuit of such a network is not within the scope of this paper.

With five authors each with 30 samples, the sample space for each experiment was not very large. This introduces the possibility of overfitting. The POS bigram model using support vector machines was highly accurate, however, a large feature space with few instances of test targets (only approximately three target authors per cross validated test sample) may be the reason for exceptionally high accuracy. Using fewer folds for validation or including more samples per author might might remedy the potential for overfitting.

Table 4. 10 fold cross validation and confidence interval for POS bigram and POS assortativity feature sets using random forests and support vector machines.

	Random Forests		Support Vector Machines	
Number of words	Bigram Assortativity	POS Bigram	Bigram Assortativity	POS Bigram
125	54.00%	52.66%	58.00%	75.33%
250	74.08%	65.33%	69.33%	86.00%
500	79.48%	68.00%	72.00%	93.33%
1000	78.74%	72.00%	75.33%	96.67%
2000	87.58%	82.67%	79.33%	100.0%
3000	90.62%	88.00%	89.33%	100.0%
4000	94.02%	90.67%	88.00%	100.0%
5000	91.36%	92.00%	87.33%	100.0%
6000	94.04%	92.00%	87.33%	98.66%
7000	94.50%	93.33%	89.33%	93.33%
8000	93.71%	94.00%	86.00%	99.33%
9000	94.45%	95.33%	90.00%	99.33%
10000	96.45%	96.00%	92.00%	98.67%
	90% Confidence Interval			
	Bigram Assortativity	POS Bigram	Bigram Assortativity	POS Bigram
125	48.67 - 60.67%	46.67 - 59.33%	52.67 - 60.67%	76.00 - 81.33%
250	54.67 - 63.33%	56.67 - 65.33%	64.00 - 71.33%	84.00 - 88.67%
500	62.67 - 71.33%	61.33 - 72.00%	67.33 - 73.33%	91.33 - 94.67%
1000	76.67 - 81.33%	70.67 - 74.67%	71.33 - 76.67%	94.00 - 96.00%
2000	85.33 - 90.00%	74.67 - 85.33%	76.00 - 81.33%	99.33 - 100.0%
3000	89.33 - 92.00%	86.00 - 90.00%	86.67 - 90.67%	100.0 - 100.0%
4000	92.67 - 95.33%	89.33 - 92.00%	86.00 - 90.67%	100.0 - 100.0%
5000	89.33 - 93.33%	90.67 - 93.33%	85.33 - 89.33%	99.33 - 100.0%
6000	92.67 - 95.33%	90.67 - 93.33%	85.33 - 89.33%	99.33 - 99.33%
7000	93.33 - 95.33%	91.33 - 94.00%	88.00 - 91.33%	100.0 - 100.0%
8000	92.00 - 95.33%	92.67 - 94.67%	85.33 - 89.33%	93.33 - 100.0%
9000	92.67 - 96.00%	94.00 - 96.00%	89.33 - 92.00%	100.0 - 100.0%
10000	96.45 - 97.33%	95.33 - 96.67%	91.33 - 94.00%	98.67 - 100.0%

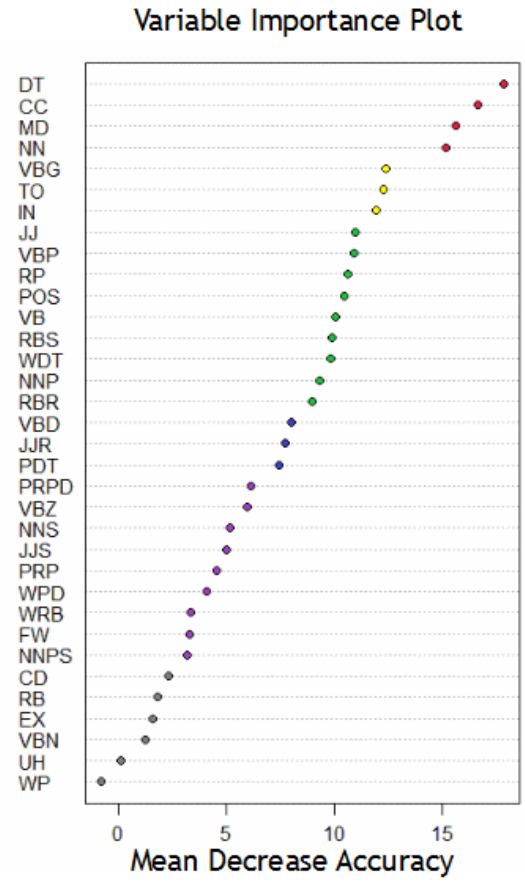

Figure 7. The variable importance plot ranks the most important parts of speech from the nominal assortativity feature set when trained on random forests.

Variable Importance of Features

The variable importance plot in figure 7 returns by rank the most predictive features when using random forests. For word graphs the nominal assortativity of DT, CC, MD, and NN were the most important features. It should be noted that different iterations of the random forest algorithm will produce different rankings, however, in general the most used parts of speech were among the most important variables and the ranking in figure 7 is reflective of that distribution.

4.3 Using Assortativity to Compare Word Networks

Caution is required when using the assortativity coefficient. Hofstad and Litvak showed using a synthetic graph technique, as well as real-world network

data, that for disassortative networks the magnitude of the correlation coefficient (also known as assortative mixture) decreases as the network increases in size. The assortativity calculation does give the correct sign of the coefficient (assortativity vs disassortativity), however, the inconsistent magnitude introduces the problem that the assortativity coefficient is not a good measurement for comparison between graphs of different size [26] [29]. Table 5 shows the results using the assortativity model on varied sample sizes between 9000 and 14000 words. While the model still exceeds 90%, compared to tests on fixed sample size the model suffers considerably (see table 4). The improved results when controlling for sample size reinforces the conclusions of Hofstad and Litvak[5].

Table 5. Accuracy of assortativity model with varying sample size between 9000 - 14000 words

Classifier	10-fold cross	95% conf interval	5% conf interval
SVM	90.67%	95.25%	85.00%
Random Forest	92.00%	97.72%	86.31%

However, with word graphs, it appears that the magnitude of the assortativity coefficient does not decrease in magnitude indefinitely. Instead, once a word graph is large enough, approximately 3000 POS or more, the magnitude of the assortativity coefficient stabilizes. This may explain why the model becomes more than 90% accurate once the word graph includes in excess of 3000 POS. The charts in figure 10 show the average nominal assortativity for different parts of speech for each of the five authors. In general, but not in all cases, the charts show a decrease in the magnitude of the average nominal assortativity coefficient for disassortative parts of speech. This observation supports Hofstad and Litvak. However, for 3000 parts of speech or more, the plot of the average nominal assortativity flattens out.

[5]Controlling for sample size has shown to improve results for other types of models. On smaller sample sizes, Sanderson and Guenter saw as much as a 15% improvement in accuracy when controlling for sample size. [30]

The stabilizing behavior of the assortativity coefficient raises the possibility that, for word networks, nominal assortative mixture is a valid measurement for comparison between networks of different sizes. With 90% accuracy given word networks of varied sizes suggests that for word graphs, nominal assortative mixture is a valid comparator as long as the network sizes are comparable and of sufficient size.

To observe more closely the behavior of the nominal assortativity coefficient, the charts in figures 11 through 13 plot the nominal assortativity coefficient against the size of the word graph for a particular author. The charts in figure 11 plot the nominal assortativity coefficient for determiners (DT) for each author. The charts in figure 12 plot the nominal assortativity coefficient for nouns, singular and mass, (NN) for each author. The charts in figures 13 plot the nominal assortativity coefficient for adjectives (JJ), verbs (VB), prepositions (IN), and coordinating conjuctions (CC) for Nathaniel Hawthorne.

The charts in figures 11 through 13 show the behavior of the assortativity coefficient for word networks of fixed size by cumulative word (POS) frequency that increase step wise (see sections 3.2 and 4 for break down of sample sizes). For smaller word networks the assortativity coefficient appears highly varied. As the size of the word network increases, the assortativity coefficient converges to tight bounds. For DT and NN the assortativity coefficient for all fives authors exhibits this behavior. The same behavior is observed for JJ, VB, IN, and CC, for Hawthorne. It appears that for all parts of speech the assortativity coefficient is more varied for smaller networks and converges as the size of the network increases.

4.4 Motifs

For an example test of the assumption that three vertex motifs are not random table 6 gives the transitivity of the 5 writing samples from table 3, as well as the

transitivity range for 1000 random graphs with the same degree distribution as the comparable sample of writing. In all five cases the true clustering coefficient of the writing sample is below the range for 1000 random graphs. This indicates that the clustering coefficient of word graphs is not random, and is consistently lower than for random graphs. The low clustering coefficient means the absence of triangles is not random for word networks. The low transitivity for words graphs is explained by Milo *et al.* who found that three vertex motifs that were not triangles had a higher significance of occurring than three vertex motifs that formed triangles [2].

Table 6. Clustering coefficient range for 1000 random graphs of same in/out-degree distribution from five writing samples with sample size of 3000 words

Author	Transitivity	Range	
Hawthorne	0.01468575	0.02981018	0.04172314
Austen	0.03807761	0.05655707	0.07976013
Jerome	0.01992584	0.03751943	0.05475122
Hardy	0.01811918	0.03171623	0.04831222
Doyle	0.02269955	0.04085767	0.06236037

The charts in figure 9 plot the clustering coefficient against vertex size for all samples of each of the five authors. For large enough words graphs (about 1000 POS cumulative, or roughly 500 vertices) the clustering coefficient is consistently below the lower bound for the transitivity of random graphs. However, for smaller networks, the same is not true. For smaller networks the coefficient shows high variance. As the network size increases the coefficient falls within bounds. Once the word networks are large enough the clustering coefficient falls within regular bounds for each of the five authors.

The charts in figure 9 show clustering of transitivity value for word graph with similar number of vertices. In figure 8 vertical strips of points constitute samples of a fixed frequency of POS. It appears by the downward slope of the bands that for samples of fixed size an increase in the number of vertices results in

lower transitivity. On a more macro scale, however, for large enough word graphs, the range of the clustering coefficient between bands is consistent. This suggests that while an increase in the word graph size results in lower trasnsitivy for fixed samples sizes, transitivity is range bound for most of an author's writing. Graph size does appear to have a macro effect on transitivity in that the more unique words an author uses, the lower the transitivity. Hawthorne, with the largest word networks, produced the lowest clustering coefficient values, while Austen, with the smallest word networks, produced the largest values.

Since the absence of triangles is significant for word graphs, it is worth asking what triangles, by part-of-speech, do appear often. Given a sample of size 3000 POS, the triangle composed of verticex types DT, IN, and NN occurs singly more often than any other triangle for each of the five authors. Compared to most other POS triangles, with frequencies less than five in most cases, the DT, IN, and NN triangle occurred tens of times more often.

4.5 Future Research

Future research could explore the relationships among natural languages (and languages families) using nominal assortativity similar to the comparative work done on motifs by Milo *et al.* Milo found that three vertex motifs occurred at similar rates for different language such as French and Japanese.

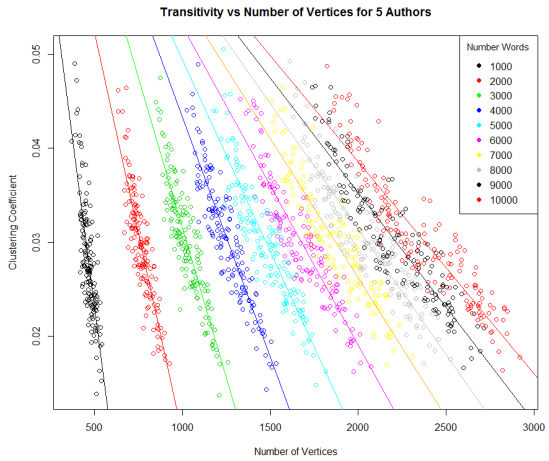

Figure 8. Clustering coefficient verse graph size. The sample size i.e. the number of words is controlled and used as a dummy interaction term. The linear regression lines show a negative correlation between clustering coefficient and the number of vertices. As the number of words increases the slope becomes less steep.

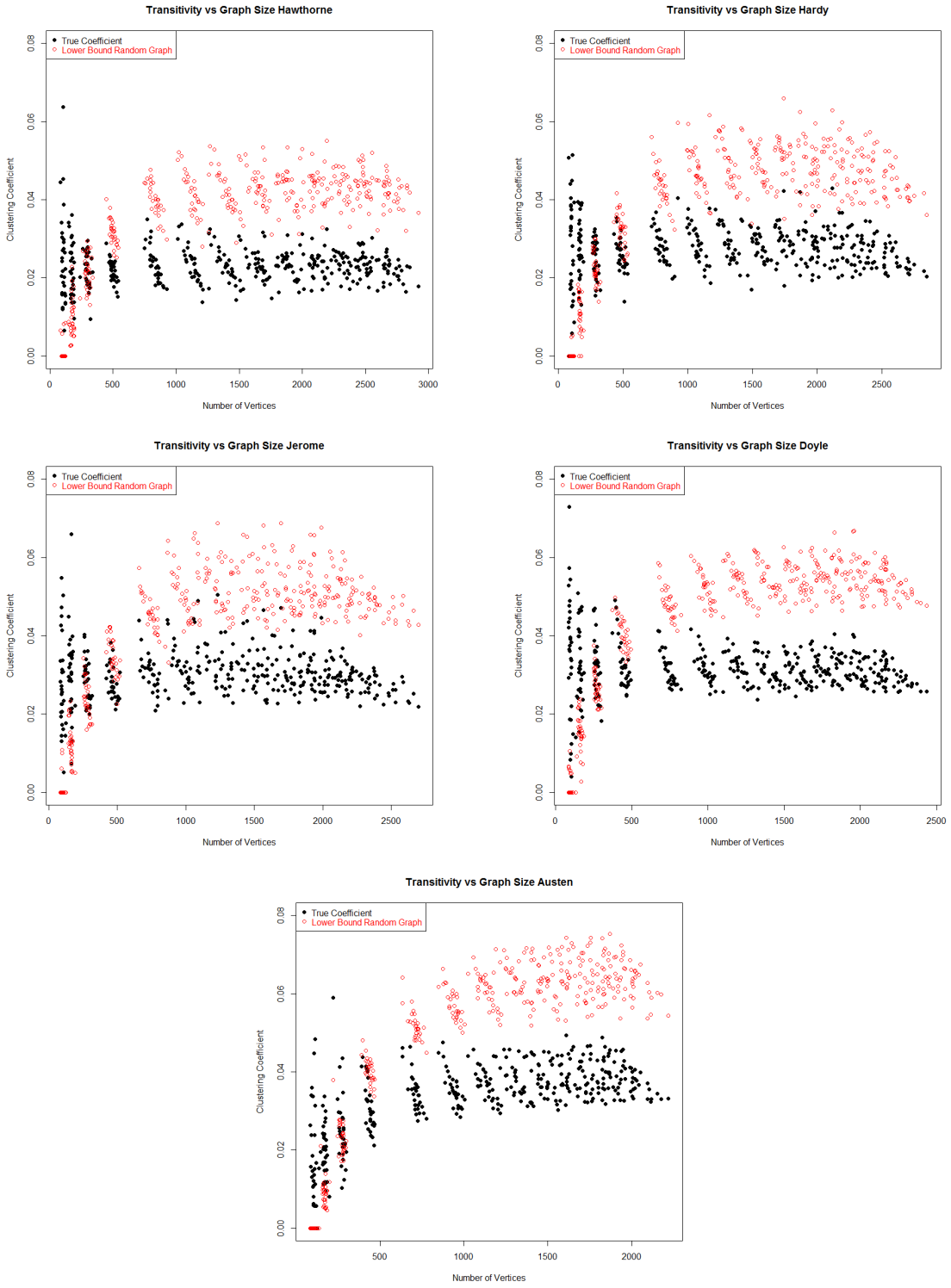

Figure 9. Clustering coefficient for five different authors over all samples. The x axis is the number of vertices, the y axis is the clustering coefficient. Black points are true transitivity and red points are the lower bounds for 1000 random graphs of same in/out-degree distribution. Counting the vertical bands left to right constitutes the number of words in the sample: 125, 250, 500, 1000, 2000, 3000, 4000, 5000, 6000, 7000, 8000, 9000, 10000.

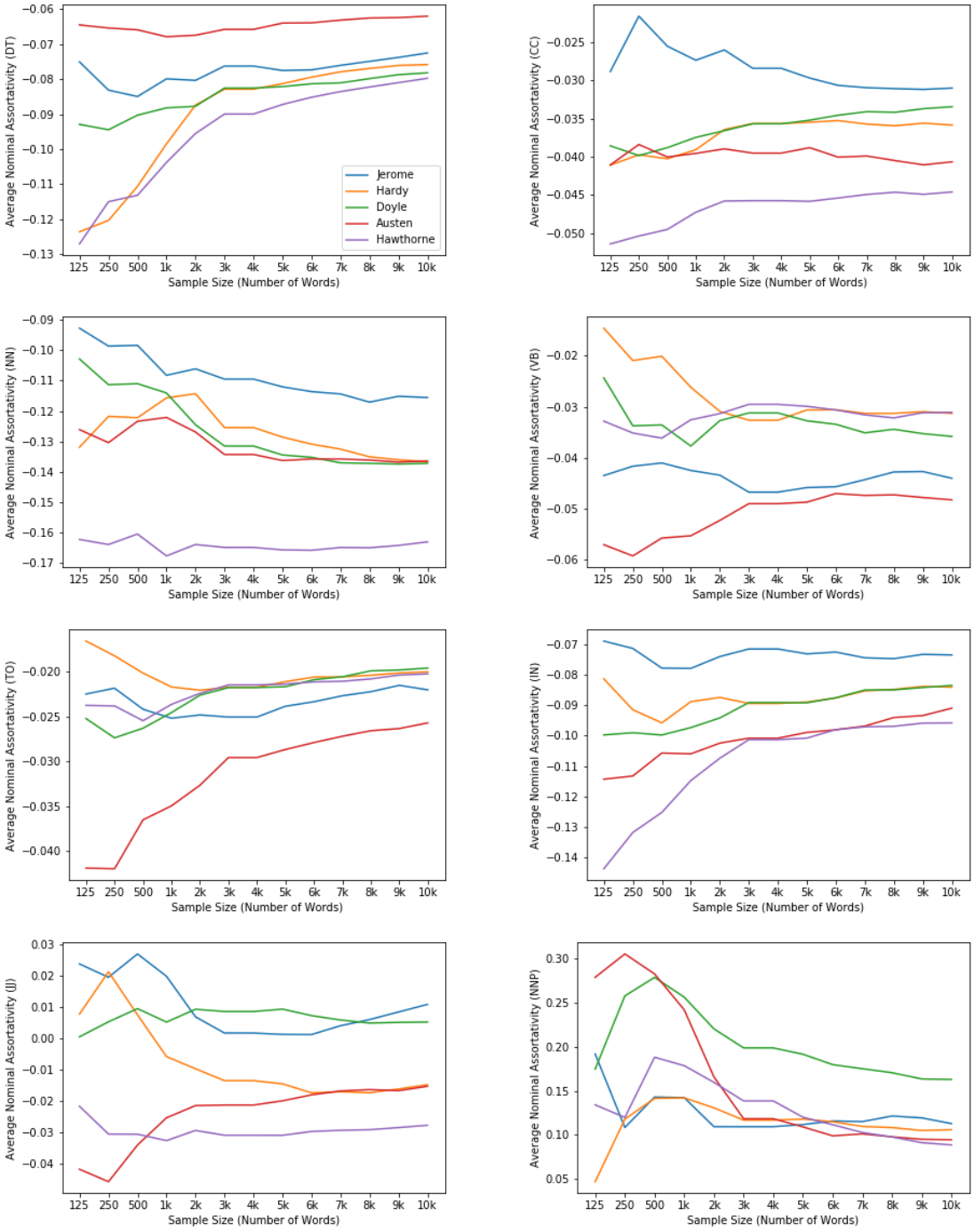

Figure 10. Average nominal assortativity for DT, CC, NN, VB, IN, TO, JJ, and NNP for five different authors.

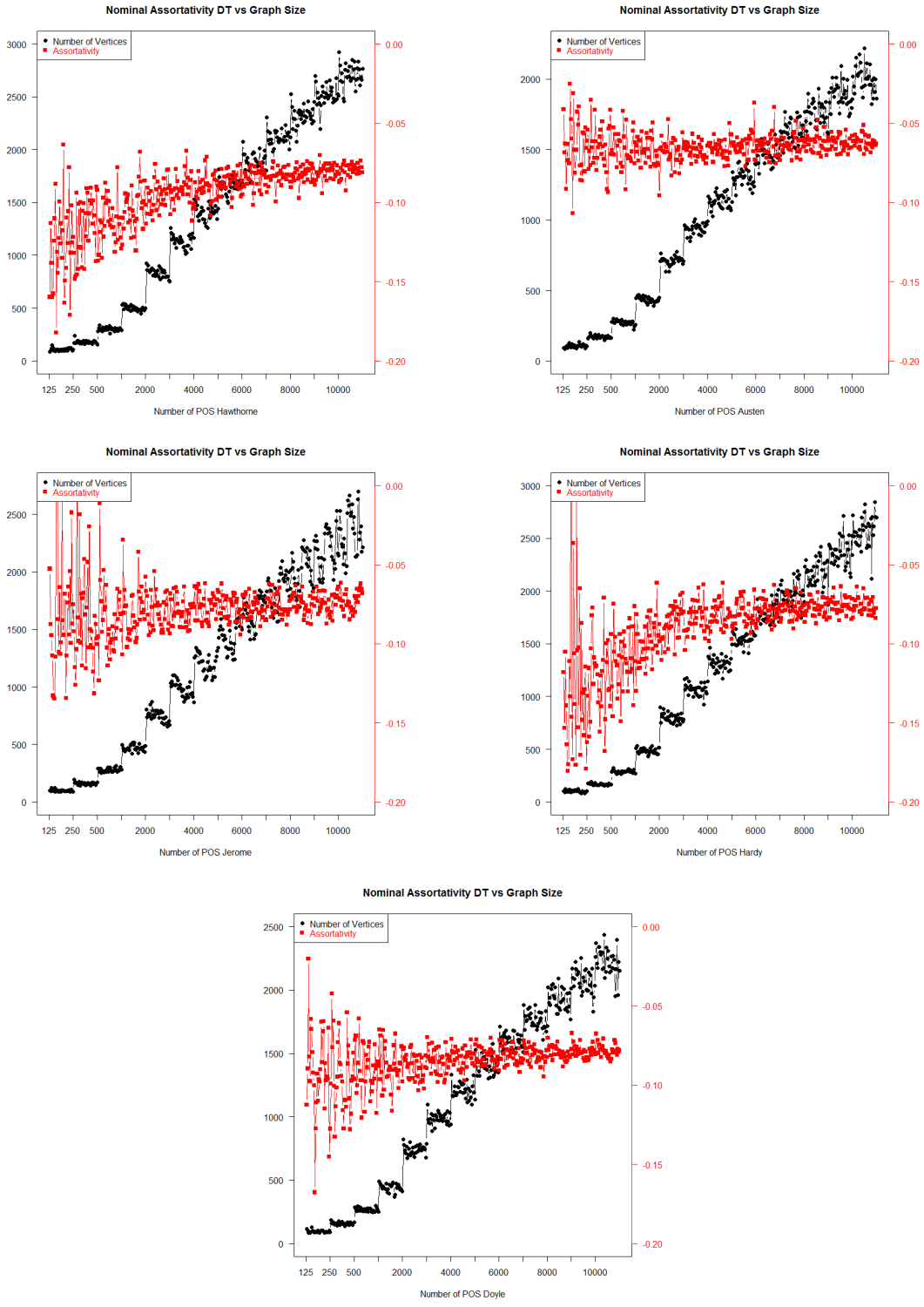

Figure 11. Nominal assortativity DT plotted against graph size for five different authors. The x axis is the number of words in the sample, the left (black) axis is the number of vertices, and the right (red) axis is the assortativity coefficient.

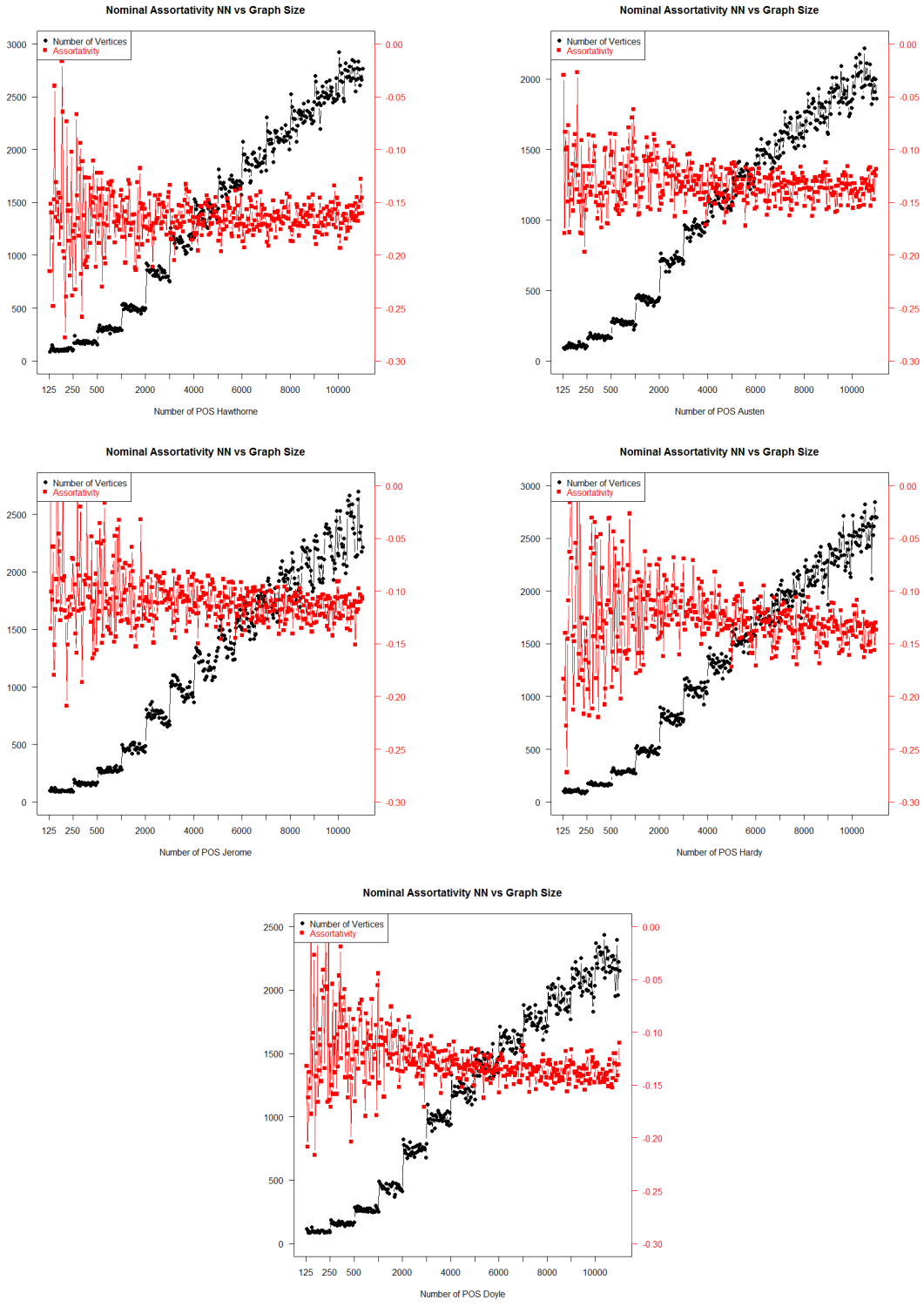

Figure 12. Nominal assortativity NN plotted against graph size for five different authors. The x axis is the number of words in the sample, the left (black) axis is the number of vertices, and the right (red) axis is the assortativity coefficient.

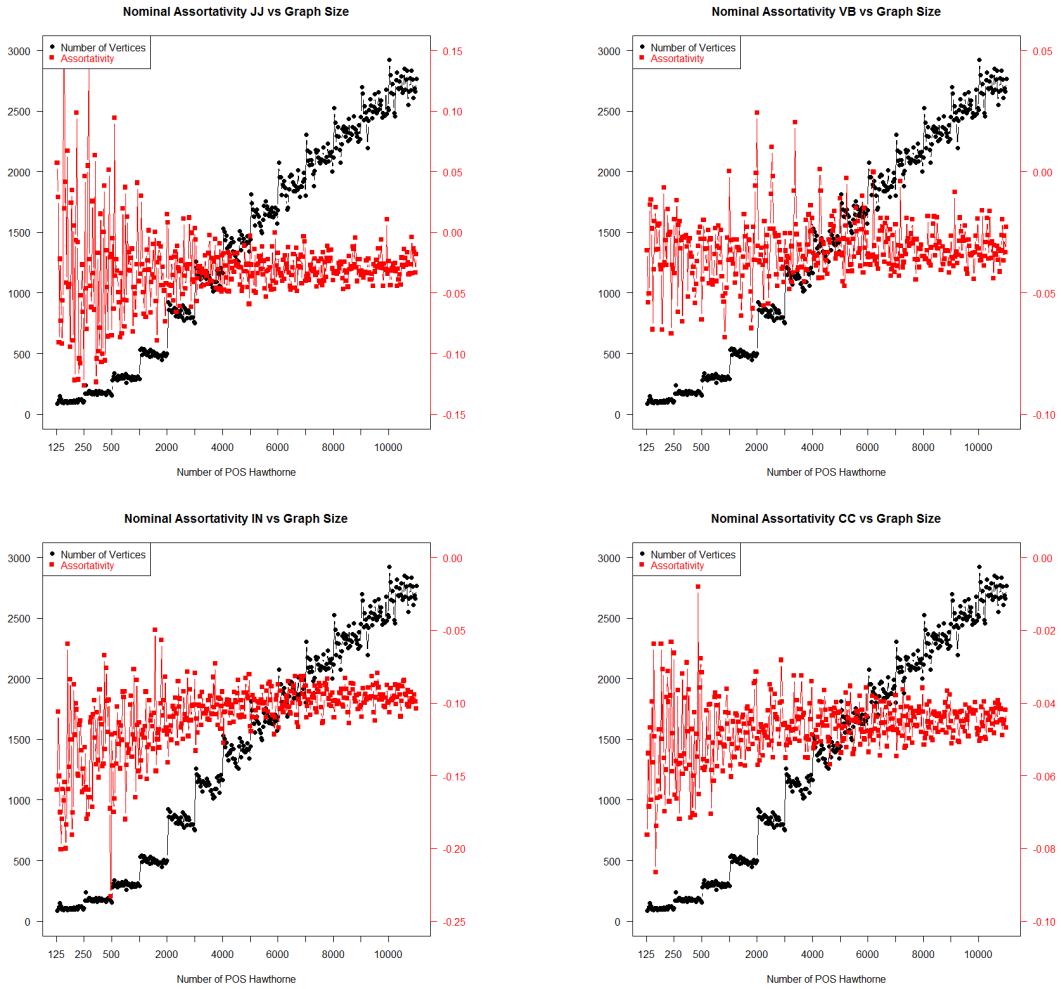

Figure 13. Nominal assortativity JJ, VB, IN, CC plotted against graph size for Hawthorne. The x axis is the number of words in the sample, the left (black) axis is the number of vertices, and the right (red) axis is the assortativity coefficient.

5 Conclusions

Representation of text as relational data provides many advantages. The information contained in a word graph produces several well documented feature sets used in authorship attribution tasks including the part of speech bigram model examined in this article. Since these feature sets have shown success in authorship attribution tasks, it is worthwhile to analyze these models for insights into the English language. Network data analysis provides an avenue to explore language in this way.

When computed for different authors, nominal assortativity by parts of speech appears to distinguish between the individual preferences of authors for part of speech usage, revealing aspects of authors style. Assortative regularities across authors reveals a grammar "signature" for the English language that exhibits mostly disassortative properties but permits some assortative relationships. These disassortative properties span the layered components of language. At the phonetic level different speech sounds are put together to form syllables and make distinct words. The same principle describes using different letters to write words. At the grammar level, words of different parts of speech collocate to create sentences. The combination of differing components enables the structured use of sound and meaning.

LIST OF REFERENCES

[1] T. Leonard, N. Daniels, L. Hamel, and N. Katenka, "Assortative mixture of english parts of speech," in *Proceedings of 6th International Conference on Complex Networks and Their Applications*. Lyon, France: Springer, December 2017, pp. 455–466.

[2] R. Milo, S. Itzkovitz, N. Kashtan, R. Levitt, S. Shen-Orr, I. Ayzenshtat, M. Sheffer, and U. Alon, "Superfamilies of evolved and designed networks," *Science*, vol. 303, no. 5663, pp. 1538–1542, 2004. [Online]. Available: http://science.sciencemag.org/content/303/5663/1538

[3] E. Stamatatos, "A survey of modern authorship attribution methods," *Journal of the American Society for Information Science and Technology*, vol. 60, no. 3, pp. 538–556, 2009. [Online]. Available: http://dx.doi.org/10.1002/asi.21001

[4] F. Mosteller and D. L. Wallace, *Inference and disputed authorship : the Federalist Papers*. Addison-Wesley, 1964.

[5] P. Azar, "Using algorithmic attribution techniques to determine authorship in unsigned judicial opinions," *Stanford Technology Law Review*, vol. Volume 16, Number 3, 2013. [Online]. Available: https://journals.law.stanford.edu/sites/default/files/stanford-technology-law-review-stlr/online/algorithmicattribution.pdf

[6] J. Diederich, J. Kindermann, E. Leopold, and G. Paass, "Authorship attribution with support vector machines," *Applied Intelligence*, vol. 19, no. 1-2, pp. 109–123, May 2003. [Online]. Available: http://dx.doi.org/10.1023/A:1023824908771

[7] G. Hirst and O. Feiguina, "Bigrams of syntactic labels for authorship discrimination of short texts," in *Literary and Linguistic Computing*, 2007.

[8] A. Mehri, A. Darooneh, and A. Shariati, "The complex networks approach for authorship attribution of books," vol. 391, p. 24292437, 04 2012.

[9] Y. Seroussi, I. Zukerman, and F. Bohnert, "Authorship attribution with topic models," *Computational Linguistics*, vol. 40, no. 2, pp. 269–310, 2014.

[10] M. E. J. Newman, "Assortative mixing in networks," *Physical Review Letters*, vol. 89, no. 20, p. 208701, Oct. 2002. [Online]. Available: http://link.aps.org/doi/10.1103/PhysRevLett.89.208701

[11] E. D. Kolaczyk and G. Csardi, *Statistical Analysis of Network Data with R (Use R!)*. Springer Science and Business Media, 2014.

[12] V. Q. Marinho, G. Hirst, and D. R. Amancio, "Authorship attribution via network motifs identification," in *2016 5th Brazilian Conference on Intelligent Systems (BRACIS)*, Oct 2016, pp. 355–360.

[13] L. Antiqueira, T. Pardo, M. Nunes, and O. Oliveira, "Some issues on complex networks for author characterization," in *Inteligencia artificial: Revista Iberoamericana de Inteligencia Artificial, ISSN 1137-3601, N. 36, 2007 (Ejemplar dedicado a: From Natural Language Processing to Information and Human Language Technology), pags. 51-58*, vol. 11, 12 2007.

[14] S. Lahiri and R. Mihalcea, "Authorship attribution using word network features," *CoRR*, vol. abs/1311.2978, 2013. [Online]. Available: http://arxiv.org/abs/1311.2978

[15] D. R. Amancio, E. G. Altmann, O. N. O. Jr, and L. da Fontoura Costa, "Comparing intermittency and network measurements of words and their dependence on authorship," *New Journal of Physics*, vol. 13, no. 12, p. 123024, 2011. [Online]. Available: http://stacks.iop.org/1367-2630/13/i=12/a=123024

[16] D. R. Amancio, "A complex network approach to stylometry," *PLoS ONE*, vol. 10, no. 8, 2015. [Online]. Available: https://doi.org/10.1371/journal.pone.0136076

[17] R. Mihalcea and D. Radev, *Graph-based natural language processing and information retrieval*. United Kingdom: Cambridge University Press, 1 2011.

[18] J. G. Foster, D. V. Foster, P. Grassberger, and M. Paczuski, "Edge direction and the structure of networks," *Proceedings of the National Academy of Sciences*, vol. 107, no. 24, pp. 10 815–10 820, 2010. [Online]. Available: http://www.pnas.org/content/107/24/10815

[19] G. K. Zipf, *Human Behaviour and the Principle of Least Effort*. Addison-Wesley, 1949.

[20] S. T. Piantadosi, "Zipf's word frequency law in natural language: A critical review and future directions," *Psychonomic Bulletin & Review*, vol. 21, no. 5, pp. 1112–1130, Oct 2014. [Online]. Available: https://doi.org/10.3758/s13423-014-0585-6

[21] Wikipedia contributors, "Zipf's law — Wikipedia, the free encyclopedia," 2018, [Online; accessed 11-April-2018]. [Online]. Available: https://en.wikipedia.org/w/index.php?title=Zipf%27s_law&oldid=831932108

[22] S. Shtrikman, "Some comments on zipf's law for the chinese language," *Journal of Information Science*, vol. 20, no. 2, pp. 142–143, 1994.

[23] L. H. Hamel, *Knowledge Discovery with Support Vector Machines (Wiley Series on Methods and Applications in Data Mining)*. Wiley-Interscience, 2011.

[24] S. Lahiri, "Complexity of Word Collocation Networks: A Preliminary Structural Analysis," in *Proceedings of the Student Research Workshop at the 14th Conference of the European Chapter of the Association for Computational Linguistics*. Gothenburg, Sweden: Association for Computational Linguistics, April 2014, pp. 96–105. [Online]. Available: http://www.aclweb.org/anthology/E14-3011

[25] K. Toutanove and C. D. Manning, "Enriching the knowledge sources used in a maximum entropy part-of-speech tagger," in *In Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*, 2000, pp. 63–70. [Online]. Available: https://nlp.stanford.edu/software/tagger.shtml

[26] N. Litvak and R. van der Hofstad, "Degree-degree correlations in random graphs with heavy-tailed degrees," *ArXiv e-prints*, Feb. 2012.

[27] R. Milo, S. Shen-Orr, S. Itzkovitz, N. Kashtan, D. Chklovskii, and U. Alon, "Network motifs: Simple building blocks of complex networks," *Science*, vol. 298, no. 5594, pp. 824–827, 2002. [Online]. Available: http://science.sciencemag.org/content/298/5594/824

[28] M. E. J. Newman, "The structure and function of complex networks," *SIAM REVIEW*, vol. 45, pp. 167–256, 2003.

[29] N. Litvak and R. van der Hofstad, "Uncovering disassortativity in large scale-free networks," *ArXiv e-prints*, vol. 87, no. 2, p. 022801, Feb. 2013.

[30] C. Sanderson and S. Guenter, "Short text authorship attribution via sequence kernels, markov chains and author unmasking: An investigation," in *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, ser. EMNLP '06. Stroudsburg, PA, USA: Association for Computational Linguistics, 2006, pp. 482–491. [Online]. Available: http://dl.acm.org/citation.cfm?id=1610075.1610142

BIBLIOGRAPHY

Amancio, D. R., "A complex network approach to stylometry," *PLoS ONE*, vol. 10, no. 8, 2015. [Online]. Available: https://doi.org/10.1371/journal.pone.0136076

Amancio, D. R., Altmann, E. G., Jr, O. N. O., and da Fontoura Costa, L., "Comparing intermittency and network measurements of words and their dependence on authorship," *New Journal of Physics*, vol. 13, no. 12, p. 123024, 2011. [Online]. Available: http://stacks.iop.org/1367-2630/13/i=12/a=123024

Antiqueira, L., Pardo, T., Nunes, M., and Oliveira, O., "Some issues on complex networks for author characterization," in *Inteligencia artificial: Revista Iberoamericana de Inteligencia Artificial, ISSN 1137-3601, N. 36, 2007 (Ejemplar dedicado a: From Natural Language Processing to Information and Human Language Technology), pags. 51-58*, vol. 11, 12 2007.

Azar, P., "Using algorithmic attribution techniques to determine authorship in unsigned judicial opinions," *Stanford Technology Law Review*, vol. Volume 16, Number 3, 2013. [Online]. Available: https://journals.law.stanford.edu/sites/default/files/stanford-technology-law-review-stlr/online/algorithmicattribution.pdf

Diederich, J., Kindermann, J., Leopold, E., and Paass, G., "Authorship attribution with support vector machines," *Applied Intelligence*, vol. 19, no. 1-2, pp. 109–123, May 2003. [Online]. Available: http://dx.doi.org/10.1023/A:1023824908771

Foster, J. G., Foster, D. V., Grassberger, P., and Paczuski, M., "Edge direction and the structure of networks," *Proceedings of the National Academy of Sciences*, vol. 107, no. 24, pp. 10 815–10 820, 2010. [Online]. Available: http://www.pnas.org/content/107/24/10815

Hamel, L. H., *Knowledge Discovery with Support Vector Machines (Wiley Series on Methods and Applications in Data Mining)*. Wiley-Interscience, 2011.

Hirst, G. and Feiguina, O., "Bigrams of syntactic labels for authorship discrimination of short texts," in *Literary and Linguistic Computing*, 2007.

Kolaczyk, E. D. and Csardi, G., *Statistical Analysis of Network Data with R (Use R!)*. Springer Science and Business Media, 2014.

Lahiri, S., "Complexity of Word Collocation Networks: A Preliminary Structural Analysis," in *Proceedings of the Student Research Workshop at the 14th Conference of the European Chapter of the Association for Computational Linguistics*. Gothenburg, Sweden: Association for Computational Linguistics, April 2014, pp. 96–105. [Online]. Available: http://www.aclweb.org/anthology/E14-3011

Lahiri, S. and Mihalcea, R., "Authorship attribution using word network features," *CoRR*, vol. abs/1311.2978, 2013. [Online]. Available: http://arxiv.org/abs/1311.2978

Leonard, T., Daniels, N., Hamel, L., and Katenka, N., "Assortative mixture of english parts of speech," in *Proceedings of 6th International Conference on Complex Networks and Their Applications*. Lyon, France: Springer, December 2017, pp. 455–466.

Litvak, N. and van der Hofstad, R., "Degree-degree correlations in random graphs with heavy-tailed degrees," *ArXiv e-prints*, Feb. 2012.

Litvak, N. and van der Hofstad, R., "Uncovering disassortativity in large scale-free networks," *ArXiv e-prints*, vol. 87, no. 2, p. 022801, Feb. 2013.

Marinho, V. Q., Hirst, G., and Amancio, D. R., "Authorship attribution via network motifs identification," in *2016 5th Brazilian Conference on Intelligent Systems (BRACIS)*, Oct 2016, pp. 355–360.

Mehri, A., Darooneh, A., and Shariati, A., "The complex networks approach for authorship attribution of books," vol. 391, p. 24292437, 04 2012.

Mihalcea, R. and Radev, D., *Graph-based natural language processing and information retrieval*. United Kingdom: Cambridge University Press, 1 2011.

Milo, R., Shen-Orr, S., Itzkovitz, S., Kashtan, N., Chklovskii, D., and Alon, U., "Network motifs: Simple building blocks of complex networks," *Science*, vol. 298, no. 5594, pp. 824–827, 2002. [Online]. Available: http://science.sciencemag.org/content/298/5594/824

Milo, R., Itzkovitz, S., Kashtan, N., Levitt, R., Shen-Orr, S., Ayzenshtat, I., Sheffer, M., and Alon, U., "Superfamilies of evolved and designed networks," *Science*, vol. 303, no. 5663, pp. 1538–1542, 2004. [Online]. Available: http://science.sciencemag.org/content/303/5663/1538

Moore, C., Yan, X., Zhu, Y., Rouquier, J.-B., and Lane, T., "Active learning for node classification in assortative and disassortative networks," in *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD '11. New York, NY, USA: ACM, 2011, pp. 841–849. [Online]. Available: http://doi.acm.org/10.1145/2020408.2020552

Mosteller, F. and Wallace, D. L., *Inference and disputed authorship : the Federalist Papers*. Addison-Wesley, 1964.

Newman, M. E. J., "Assortative mixing in networks," *Physical Review Letters*, vol. 89, no. 20, p. 208701, Oct. 2002. [Online]. Available: http://link.aps.org/doi/10.1103/PhysRevLett.89.208701

Newman, M. E. J., "The structure and function of complex networks," *SIAM REVIEW*, vol. 45, pp. 167–256, 2003.

Piantadosi, S. T., "Zipf's word frequency law in natural language: A critical review and future directions," *Psychonomic Bulletin & Review*, vol. 21, no. 5, pp. 1112–1130, Oct 2014. [Online]. Available: https://doi.org/10.3758/s13423-014-0585-6

Sanderson, C. and Guenter, S., "Short text authorship attribution via sequence kernels, markov chains and author unmasking: An investigation," in *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, ser. EMNLP '06. Stroudsburg, PA, USA: Association for Computational Linguistics, 2006, pp. 482–491. [Online]. Available: http://dl.acm.org/citation.cfm?id=1610075.1610142

Seroussi, Y., Zukerman, I., and Bohnert, F., "Authorship attribution with topic models," *Computational Linguistics*, vol. 40, no. 2, pp. 269–310, 2014.

Shtrikman, S., "Some comments on zipf's law for the chinese language," *Journal of Information Science*, vol. 20, no. 2, pp. 142–143, 1994.

Stamatatos, E., "A survey of modern authorship attribution methods," *Journal of the American Society for Information Science and Technology*, vol. 60, no. 3, pp. 538–556, 2009. [Online]. Available: http://dx.doi.org/10.1002/asi.21001

Toutanove, K. and Manning, C. D., "Enriching the knowledge sources used in a maximum entropy part-of-speech tagger," in *In Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*, 2000, pp. 63–70. [Online]. Available: https://nlp.stanford.edu/software/tagger.shtml

Wikipedia contributors, "Zipf's law — Wikipedia, the free encyclopedia," 2018, [Online; accessed 11-April-2018]. [Online]. Available: https://en.wikipedia.org/w/index.php?title=Zipf%27s_law&oldid=831932108

Zipf, G. K., *Human Behaviour and the Principle of Least Effort*. Addison-Wesley, 1949.