Authorship Attribution with Assortative Mixture of English Parts of Speech

Timothy Leonard (Advisor Dr. Natallia Katenka)
University of Rhode Island

INTRODUCTION

Authorship attribution is a **classification problem** with two main objectives: *1)* to accurately predict some characteristic of a piece of text (e.g. authorship), and *2)* to provide a descriptive model of writing that contributes to our knowledge of language. This article presents an assortative mixture model of English parts of speech that accurately predicts authorship in a supervised learning environment. By measuring the tendency for same parts of speech to collocate, the model offers a detailed and unbiased glimpse into the stylometric features of grammar. Assortative mixture is a single coefficient that can be applied to each part of speech in a word graph to generate a small but inclusive feature set. Comprised of a homogenous estimator, the assortative mixture model is simple yet captures many fundamental language characteristics including what grammar types exhibit selective linking. As a **network graph model**, words are vertices and edges represent sequential words (i.e. word bigrams or word adjacencies) that appear in a sample of writing. To calculate **nominal assortativity** and generate a feature set, vertices have as an attribute a part of speech that can be compared to other vertices. Such graphs are not new to the literature, however, previous models ignore grammar or fail to represent all grammar types due to computational limitations or deliberate choice of the model. Research on word graphs sought to discover predictive features using network analysis but did not include the part of speech as an attribute of a vertex. These studies showed that other descriptive characteristics such as transitivity, density, degree assortativity, etc., do not stand alone as significant predictors in a feature set. By comparison, nominal grammar assortativity alone is highly predictive of authorship. The statistical analysis of graphs aided with an accurate speech tagger empowers a more **mathematically descriptive** examination of grammar now that entire collections of writing can be tagged efficiently.

AUTHORSHIP ATTRIBUTION FORMAL DEFINITION

Authorship attribution is a classification problem. Given samples of writing apply some **labeling function** to accurately discern authorship given labeled training data. More formally[1], **given**:

1) a universe $\boldsymbol{X}$ of written works by authors $\boldsymbol{A}_i = [a_1, a_2, \ldots, a_k]$,
2) a sample $\boldsymbol{S}$ where $\boldsymbol{S}$ is a subset of $\boldsymbol{X}$,
3) a target labeling function $f(\boldsymbol{x}) : \boldsymbol{X} \rightarrow \{a_1, a_2, \ldots, a_k\}$,
4) a labeled training set $\boldsymbol{D}$, where $\boldsymbol{D}$ are all the pairs $(\boldsymbol{x}, \boldsymbol{y})$ such that $\boldsymbol{x}$ is in $\boldsymbol{S}$ and $\boldsymbol{y} = f(\boldsymbol{x})$,

compute a function $\bar{f}(\boldsymbol{x}) : \boldsymbol{X} \rightarrow \{a_1, a_2, \ldots, a_k\}$ using D such that:

$\bar{f}(x) \approx f(x)$ **for all x in X.**

DATA

The data set includes **5 authors** chosen from a subset of the Gutenberg data set made available by Michigan University. Each author is represented by **30 writing samples (150 total)** of **between 9000 and 14000 words**. The authors are Jerome Klapka Jerome (1859-1827), Thomas Hardy (1840-1928), Sir Arthur Conan Doyle (1859-1930), Jane Austen (1775-1817), and Nathaniel Hawthorne (1804-1864). Each sample is manually pre-processed to remove symbols and fix punctuation that might hinder the speech tagger. Stanford CoreNLP tags each sample with parts of speech. Stanford CoreNLP uses the Penn Treebank parts of speech categories.

ASSORTATIVE MIXTURE AS FEATURE SET

Assortativity is a coefficient within a range between 1 and -1 that measures the tendency for like vertices in a graph to share an edge. For parts of speech in a word graph, a **positive assortativity** coefficient suggests words of the same part of speech occur sequentially, while **negative assortativity** suggests they do not. The assortativity coefficient is calculated for each part of speech to generate a feature set used in supervised learning algorithms such as SVMs and random forests.

$$r_i = \frac{\sum f_{ii} - \sum f_{ir} f_{ic}}{1 - \sum f_{ir} f_{ic}} \quad (1)$$

In equation (1) r_i is the assortativity coefficient, f_{ij} is the fraction of edges in a graph G that join a vertex in the i_{th} category to a vertex in the j_{th} category, f_{ir} and f_{ic} are the marginal row and column sums respectively.[2]

CITATIONS

[1]Lutz Hamel. *Knowledge Discovery with Support Vector Machines.* John Wiley & Sons, Inc, 2009.

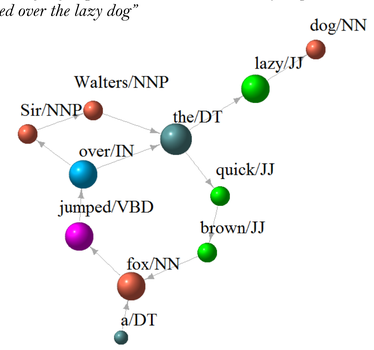
[2]Eric D. Kolaczyk, Gabor Csárdi. *Statistical Analysis of Network Data with R.* Springer Science & Business Media, 2014.

NETWORK MODEL

A word network is a **directed** graph model. Words are **vertices** and **edges** occur between sequential words in sentences that are present in sample text. Each vertex has for an **attribute** its part of speech. **Edge weights** represent word frequency.

Example Sentences:

1) "the quick brown fox jumped over Sir Walters the lazy dog"
2) "a fox jumped over the lazy dog"



Graph 1. The directed graph above represents the two example sentences. Each vertex is a word with the part of speech as an attribute in the form ***word/POS***. Edges are between sequential words.

	the	a	quick	brown	fox	jumped	over	Sir	Walters	lazy	dog
the	0	0	1	0	0	0	0	0	0	1	0
a	0	0	0	0	1	0	0	0	0	0	0
quick	0	0	0	1	0	0	0	0	0	0	0
brown	0	0	0	0	1	0	0	0	0	0	0
fox	0	0	0	0	0	1	0	0	0	0	0
jumped	0	0	0	0	0	0	1	0	0	0	0
over	1	0	0	0	0	0	0	1	0	0	0
Sir	0	0	0	0	0	0	0	0	1	0	0
Walters	1	0	0	0	0	0	0	0	0	0	0
lazy	0	0	0	0	0	0	0	0	0	0	1
dog	0	0	0	0	0	0	0	0	0	0	0

Matrix 1. A word adjacency 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

Matrix 2. A POS adjacency matrix of the two sentences.

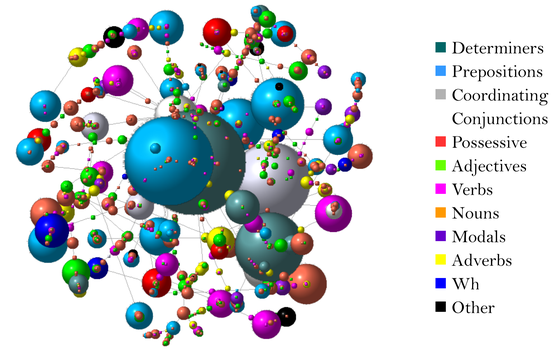
DT	-1
JJ	-.202
NN	-1
VBD	-1
IN	-1
NNP	.333

Table 1. Assortativity for each part of speech in Matrix 2.

	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

Table 2. Penn Treebank parts of speech grouped generally by color.

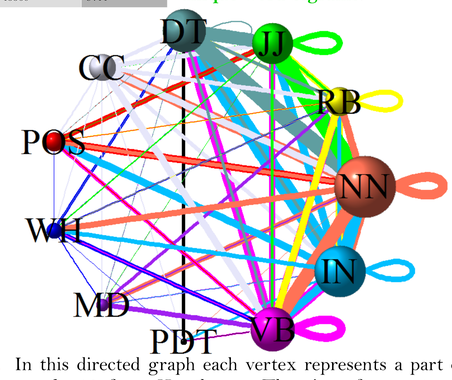
VISUALIZATIONS AND DESCRIPTIVE ANALYSIS



Graph 2. Minimum spanning tree representation of sample 12 from Hawthorne. Each vertex is a word with size proportional to frequency. In a directed word graph word frequency is **edge strength**.

	Hawthorne Sample 12	Austen Sample 4
Vertices	3500	2003
Edges	10366	7171
Density	0.0008464459	0.001788277
Degree Assortativity	-0.2461854	-0.2942947
Total words	13360	9714

Table 4. Descriptive analysis of two samples from two different authors. The number of vertices is the number of **unique words**. The number of edges is the number of **unique word bigrams**.



Graph 3. In this directed graph each vertex represents a part of speech in sample 12 from Hawthorne. The size of a vertex is proportional to POS frequency. Edges are 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. Notice nouns and verbs are among the largest categories in graph 3 but individually have low frequency in graph 2. Determiners and prepositions individually have highest frequency in graph 2 but comparatively are smaller overall categories.

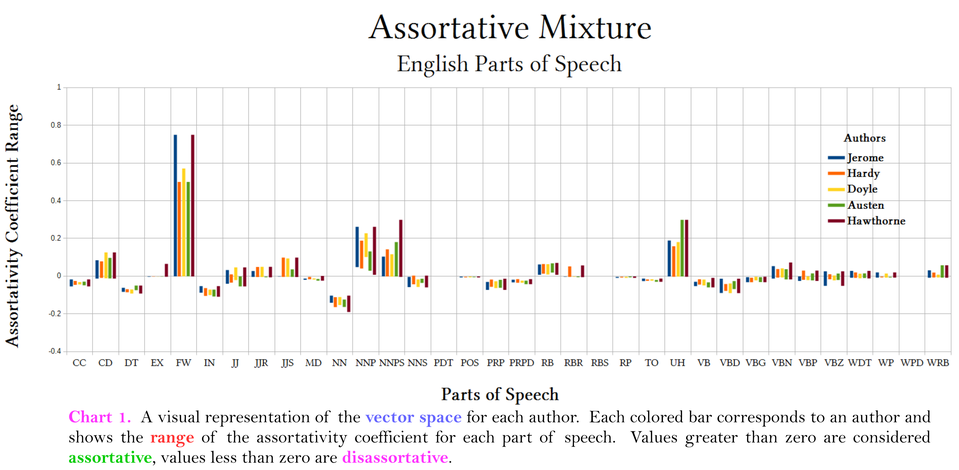
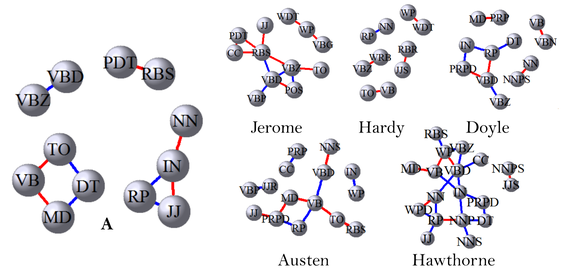


Chart 1. A visual representation of the **vector space** for each author. Each colored bar corresponds to an author and shows the **range** of the assortativity coefficient for each part of speech. Values greater than zero are considered **assortative**, values less than zero are **disassortative**.

CORRELATION MATRIX



Graph 4. Graph A on the left represents the correlation matrix of the nominal assortativity for each part of speech over all authors. The smaller graphs on the right represent matrices for each author individually. Edges are present when the absolute value of the correlation is greater than 0.5. **Red edges** characterize positive correlation, while **blue edges** signify negative correlation.

MODEL TESTING

The model was tested using **support vector machines** (SVMs) and **random forests**. Both algorithms were tested using the leave one out, 10 fold cross validation, and bootstrap method. The accuracy results using the leave one out method and 10 fold cross validation, as well as the bootstrap mean and bootstrap 90% mean confidence interval, are reported in table 3 alongside the resubstitution (training) error. The results using random forests are slightly better that using SVMs.

RESULTS

Classifier	Training error	Leave one out	10-fold cross	Bootstrap mean	95% conf. interval	5% conf. interval
SVM *linear kernel*	0.00%	90.67%	90.67%	90.30%	95.23%	85.00%
Random Forest	0.00%	91.33%	92.00%	91.38%	97.72%	86.31%

Table 5. Training error and model accuracy for each test method.

CONCLUSIONS

The assortative mixture model of parts of speech is highly predictive of authorship. Analysis of word graphs reveals what parts of speech are assortative and what are disassortative. Overall **word graphs exhibit disassortative properties**, but there are lots of exceptions. Determiners, coordinating conjunctions, prepositions, nouns, and some verbs are strongly disassortative. Words from these parts of speech categories are the most frequently used and therefore drive negative assortativity for parts of speech as a whole, explaining why word graphs are dissassortative by degree. On the other hand, adjectives, adverbs, and proper nouns do show assortative qualities.

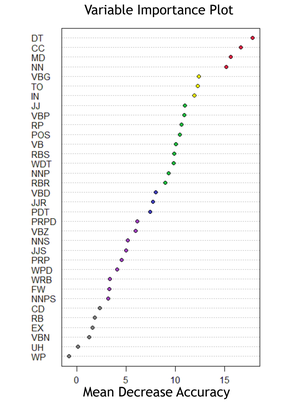


Chart 2. The variable importance plot shows which parts of speech are **most significant** to the model using random forests.