AUTHORSHIP ATTRIBUTION USING FUNCTION WORDS ADJACENCY NETWORKS

Santiago Segarra, Mark Eisen, and Alejandro Ribeiro

Department of Electrical and Systems Engineering, University of Pennsylvania

ABSTRACT

We present an authorship attribution method based on relational data between function words. These are content independent words that help define grammatical relationships. As relational structures we use normalized word adjacency networks. We interpret these networks as Markov chains and compare them using entropy measures. We illustrate the accuracy of the method developed through a series of numerical experiments including comparisons with frequency based methods. We show that accuracy increases when combining relational and frequency based data, indicating that both sources of information encode different aspects of authorial styles.

Index Terms— Authorship attribution, word adjacency network, Markov chain, relative entropy

1. INTRODUCTION

The goal of authorship attribution is to match a text of unknown or disputed authorship to one of a group of potential candidates. More generally, it can be seen as the search for a compact representation of an author's writing style, or stylometric fingerprint. Applications of this study range from forensics to questions of plagiarism in the works of both published authors as well as students. With recent developments in computational efficiency and information processing, authorship attribution studies are of both increasing interest and accuracy [1, 2]. The study of authorship attribution, sometimes called stylometry, has its beginnings in works published over a century ago [3] which proposed distinguishing authors by looking at word lengths. This was later improved upon by [4] to consider average sentence length as a determinant.

These two rudimentary ideas have improved since. A significant development came with the introduction of the influential idea of analyzing function words as a way to characterize authors' styles [5]. Function words are words like prepositions, conjunctions, and pronouns which on their own carry little meaning but instead help define grammatical relationships between words. The study of function words is beneficial as they primarily inform about syntax rather than content. Since [5], function words appeared in a number of papers such as [6] where principal component analysis was performed regarding the frequencies of the most common words in a text. A similar look at commonly appearing words was done in [7]. Attention has also been given to analyzing features other than appearances of high-frequency words. Examples of these are the stylometric techniques in [8] and the use of vocabulary richness as a stylometric marker [9–11] – see also [12] for a critique. Furher examples are word stability, the extent to which a word can be replaced by an equivalent [13], or syntactical markers like part-of-speech taggers [14].

Frequency based feature analysis have been expanded into the application of Markov chains in stylometry. Studies done in [15] and [16] use letter based Markov chains to model texts. Although this approach generates positive results, there is little intuitive reasoning behind the notion that an author's style can be better modeled by his usage of individual letters rather than words. Research in [17] has looked at using word based Markov chains but the author did not focus on function words and had to introduce smoothing techniques to account for transitions not encountered in the training sets, deteriorating the accuracy.

In this paper, we focus on function words but instead of using their frequency distribution as an author signature [5] we propose the use of the relational structure of function words. In order to classify the authorship of a text we compute an asymmetric network of function word adjacencies capturing how likely it is to find a particular function word within the next few words conditional on the occurrence of another given word (Section 3). The resulting matrices can be interpreted as transition probabilities of a Markov chain. The similarity of different texts is estimated by the relative entropy of these transition probabilities (Section 3.1). We test the proposed methodology in authorship attribution problems including texts from up to 18 different authors using training sets consisting of between 1 and 6 known texts per author. Estimation accuracy in the order of at least 90% is observed in most cases (Section 4). We further demonstrate that our classifier performs better that classifiers based in word frequencies (Section 4.1). Perhaps more important, numerical experiments show that classifiers based on word frequencies encode different stylometric fingerprints than the classifiers proposed here and can then be combined for increased attribution correctness.

2. PROBLEM FORMULATION

We are given a set of n authors $A = \{a_1, a_2, ..., a_n\}$, a set of m known texts $T = \{t_1, t_2, ..., t_m\}$ and a set of k unknown texts $U = \{u_1, u_2, ..., u_k\}$. We are also given an authorship attribution function $r_T : T \to A$ mapping every known text in T to its corresponding author in A, i.e. $r_T(t) \in A$ is the author of text t for all $t \in T$. We further assume r_T to be surjective, this implies that for every author $a_i \in A$ there is at least one text $t_j \in T$ with $r_T(t_j) = a_i$. Denote as $T^{(i)} \subset T$ the subset of known texts written by author a_i, i.e.

$$T^{(i)} = \{t \mid t \in T, r_T(t) = a_i\}. \tag{1}$$

According to the above discussion, it must be that $|T^{(i)}| > 0$ for all i and $\{T^{(i)}\}_{i=1}^n$ must be a partition of T. In Section 3, we use the texts contained in $T^{(i)}$ to generate a relational profile for author a_i. There exists an unknown attribution function $r_U : U \to A$ which assigns each text $u \in U$ to its actual author $r_U(u) \in A$. Our objective is to approximate this unknown function with an estimator $\hat{r}_U$ built with the information provided by the attribution function r_T. In particular, we construct word adjacency networks (WAN) for the known texts $t \in T$ and unknown texts $u \in U$. We attribute texts by comparing the WANs of the unknown texts $u \in U$ to the WANs of the known texts $t \in T$.

In constructing WANs the concepts of sentence, proximity, and function words are important. Every text consists of a sequence of sentences, where a sentence is defined as an indexed sequence of words between two stopper symbols. We think of these symbols as grammatical sentence delimiters, but this is not required. For a given sentence we define a directed proximity between two words parametric on a discount factor $\alpha \in (0, 1)$ and a window length D. If we denote as $i(\omega)$ the position of word ω within its sentence the directed proximity $d(\omega_1, \omega_2)$ from word ω_1 to word ω_2 when $0 < i(\omega_2) - i(\omega_1) \le D$ is defined as

$$d(\omega_1, \omega_2) := \alpha^{i(\omega_2)-i(\omega_1)}. \tag{2}$$

Supported by NSF CAREER CCF-0952867 and NSF CCF-1217963.

Common Function Words				
the	and	a	of	to
in	that	with	but	it

Table 1. 10 most common function words found in the texts

In every sentence there are two kind of words: function and non-function words [18]. Function words are words that express primarily a grammatical relationship. Examples of function words include articles, prepositions, and pronouns. The 10 most common function words are listed in Table 1. We exclude gender specific pronouns ("he" and "she") as well as pronouns that depend on narration type ("I" and "you") from the set of function words to avoid biased similarity between texts written using the same grammatical person – see Section 3 for details. The concepts of sentence, proximity, and function words are illustrated in the following example.

Example 1 Define the set of stopper symbols as $\{.\ ;\}$, let the parameter $\alpha = 0.8$, the window $D = 4$, and consider the text

> "A swarm in May is worth a load of hay; a swarm in June is worth a silver spoon; but a swarm in July is not worth a fly."

The text is composed of three sentences separated by the delimiter $\{\ ;\}$. We then divide the text into its three constituent sentences and highlight the function words

> **a** swarm **in** May is worth **a** load **of** hay
>
> **a** swarm **in** June is worth **a** silver spoon
>
> **but a** swarm **in** July is not worth **a** fly

The directed proximity from the first "a" to "swarm" in the first sentence is $\alpha^1 = 0.8$ and the directed proximity from the first "a" to "in" is $\alpha^2 = 0.64$. The directed proximity to "worth" or "load" is 0 because the indices of these words differ in more than $D = 4$.

Define the relative accuracy as the fraction of unknown texts that are correctly attributed. With $\mathbb{I}$ denoting the indicator function we can write the estimation accuracy ρ as

$$\rho(\hat{r}_U) = \frac{1}{k} \sum_{u \in U} \mathbb{I}\left\{\hat{r}_U(u) = r_U(u)\right\}, \tag{3}$$

We use $\rho(\hat{r}_U)$ to gauge performance of the classifier in Section 4.

3. WORD ADJACENCY NETWORK

As relational structures we construct WANs for each text. These weighted and directed networks contain function words as nodes. The weight of a given edge represents the likelihood of finding the words connected by this edge close to each other in the text. Formally, from a given text t we construct the network $W_t = (F, Q_t)$ where $F = \{f_1, f_2, ..., f_f\}$ is the set of nodes composed by a collection of function words common to all WANs and $Q_t : F \times F \to \mathbb{R}_+$ is a similarity measure between pairs of nodes. We choose F as the set of the f most common function words among the texts analyzed. The choice of the number $|F|$ of function words is discussed in Section 4.1.

In order to calculate the similarity function Q_t we first divide the text t into sentences s_t^h where h ranges from 1 to the total number of sentences. We denote by $s_t^h(e)$ the word in the e-th position within sentence h of text t. In this way, we define

$$Q_t(f_i, f_j) = \sum_{h,e} \mathbb{I}\{s_t^h(e) = f_i\} \sum_{d=1}^{D} \alpha^d\ \mathbb{I}\{s_t^h(e+d) = f_j\}, \tag{4}$$

for all $f_i, f_j \in F$, where $\alpha \in (0, 1)$ is the discount factor that decreases the assigned weight as the words are found further apart from each other and D is the window limit to consider that two words are related. The similarity measure in (4) is the sum of the directed proximities from f_i to f_j defined in (2) for all appearances of f_i when the words are found at most D positions apart. Since in general $Q_t(f_i, f_j) \neq Q_t(f_j, f_i)$ the WANs generated are directed.

Example 2 Consider the same text and parameters of Example 1. There are four function words yielding the set $F = \{\text{a, in, of, but}\}$. The matrix representation of the similarity function Q_t is

$$Q = \begin{array}{c} \\ \text{a} \\ \text{in} \\ \text{of} \\ \text{but} \end{array} \begin{array}{c} \begin{array}{cccc} \text{a} & \text{in} & \text{of} & \text{but} \end{array} \\ \begin{pmatrix} 0 & 3 \times 0.8^2 & 0.8^2 & 0 \\ 2 \times 0.8^4 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0.8 & 0.8^3 & 0 & 0 \end{pmatrix} \end{array}. \tag{5}$$

The total similarity score from "a" to "in" is obtained by summing up the three 0.8^2 proximity scores that appear in each sentence. Although the word "a" appears twice in every sentence, $Q(\text{a, a}) = 0$ because its appearances are more than $D = 4$ words apart.

Using text WANs, we generate a network W_c for every author $a_c \in A$ as $W_c = (F, Q_c)$ where

$$Q_c = \sum_{t \in T^{(c)}} Q_t. \tag{6}$$

Similarities in Q_c depend on the amount and length of the texts written by author a_c. This is undesirable since we want to be able to compare relational structures among different authors. Hence, we normalize the similarity measures as

$$\hat{Q}_c(f_i, f_j) = \frac{Q_c(f_i, f_j)}{\sum_j Q_c(f_i, f_j)}, \tag{7}$$

for all $f_i, f_j \in F$. In this way we achieve normalized networks $\hat{P}_c = (F, \hat{Q}_c)$ for each author a_c. In (7) we assume $|F|$ small enough or texts long enough to guarantee a non zero denominator.

Our claim is that every author a_c has an inherent relational structure P_c that serves as an authorial fingerprint and can be used towards the solution of authorship attribution problems. $\hat{P}_c$ estimates P_c with the available known texts written by author a_c.

3.1. Network Similarity

The normalized networks $\hat{P}_c$ can be interpreted as discrete time Markov chains (MC) since the similarities out of every node sum up to 1. Thus, the normalized similarity between words f_i and f_j is a measure of the probability of finding f_j in the words following an encounter of f_i. In a similar manner, we can build a MC P_u for each unknown text $u \in U$.

Since every MC has the same state space F we use the relative entropy $H(P_1, P_2)$ as a dissimilarity measure between the chains P_1 and P_2. The relative entropy is given by

$$H(P_1, P_2) = \sum_{i,j} \pi(f_i) P_1(f_i, f_j) \log \frac{P_1(f_i, f_j)}{P_2(f_i, f_j)}, \tag{8}$$

where π is the limiting distribution on P_1. The choice of H as a measure of dissimilarity is not arbitrary. In fact, if we denote as w_1 a realization of the MC P_1, $H(P_1, P_2)$ is proportional to the logarithm of the ratio between the probability that w_1 is a realization of P_1 and the probability that w_1 is a realization of P_2. In particular, when $H(P_1, P_2)$ is null, the ratio is 1 meaning that a given realization of P_1 has the same probability

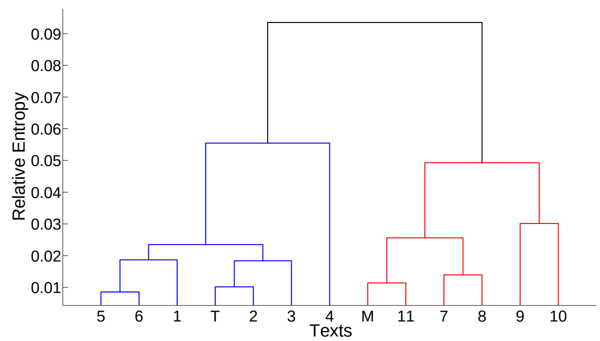

Fig. 1. Clustering of Twain's (T) and Melville's (M) relational profiles and the 11 unknown texts. Two clusters, blue and red, emerge corresponding to both authors with perfect accuracy.

Author	Texts	Author	Texts
Shakespeare, W.	10	Twain, M.	9
Austen, J.	7	Allen, G.	7
Cooper, J.F.	6	Dickens, C.	6
Marlowe, C.	8	Bacon, F.	6
Beaumont, F. & Fletcher, J.	7	Hawthorne, N.	6
Abbott, J.	7	James, H.	8
Alger, H.	7	Jonson, B.	6
Alcott, L.M.	7	Aldrich, T.B.	7
Garland, H	8	Melville, H.	8

Table 2. Authors and total number of texts per author considered for the second numerical experiment, Table 3. See details in [20].

of being observed in both MCs [19]. Using (8) we generate the attribution function $\hat{r}_U(u)$ by assigning the text u to the author with the most similar relational structure

$$\hat{r}_U(u) = a_p, \quad \text{where} \quad p = \operatorname*{argmin}_c H(P_u, \hat{P}_c). \tag{9}$$

We evaluate this classifier in the next section after the following remark.

Remark 1 In (9) we assume that the unknown texts are long enough for the corresponding MC to be ergodic. This ensures that the limiting distribution π is well defined. If this is not achieved, we replace $\pi(f_i)$ with the expected fraction of time a randomly initialized walk spends in state f_i. The random initial function word is drawn from a distribution proportional to the word frequencies in the text.

4. NUMERICAL RESULTS

In this section we fix $\alpha = 0.8$, $D = 10$ and the set of sentence delimiters to be { . ( ) ? ! ; : }. Moreover, we consider state spaces of 10 function words except in Section 4.1 where we vary the number of function words considered.

To illustrate the method developed, we begin by solving an authorship attribution problem with two candidate authors: Mark Twain and Herman Melville. For each author we have 3 known texts. We are given 11 unknown texts where the first 6 belong to Twain and the other 5 were written by Melville [20]. Every text in this simulation belongs to a different book and corresponds to a 10,000 words extract, i.e. around 25 pages of a paper back mid size edition. With the method here developed, the 11 unknown texts are attributed with perfect accuracy. An intuitive reason of why this works is depicted in Fig. 1. In this figure, we plot

		Known texts per author						Rnd.
		1	2	3	4	5	6	Attr.
Number of Authors	2	1.00	1.00	1.00	1.00	1.00	1.00	.50
	3	.87	1.00	1.00	1.00	1.00	1.00	.33
	4	.76	.92	1.00	1.00	1.00	1.00	.25
	5	.74	.90	1.00	1.00	1.00	1.00	.20
	6	.74	.82	.93	.90	.93	1.00	.17
	7	.63	.85	.94	.92	.94	1.00	.14
	8	.67	.86	.94	.93	.95	1.00	.13
	9	.60	.83	.92	.90	.95	.92	.11
	10	.55	.77	.90	.91	.95	.92	.10
	11	.53	.74	.89	.86	.88	.85	.09
	12	.57	.76	.90	.87	.89	.87	.08
	13	.60	.78	.91	.88	.90	.88	.08
	14	.59	.78	.90	.86	.87	.88	.07
	15	.61	.77	.89	.87	.88	.88	.07
	16	.57	.73	.85	.84	.88	.89	.06
	17	.58	.74	.85	.85	.89	.90	.06
	18	.54	.69	.79	.83	.88	.86	.06

Table 3. Accuracy for different number of candidate authors and number of known texts per author. Expected accuracy of random attribution is also informed. Accuracy decreases with increasing number of authors and decreasing number of training texts per author.

the average linkage hierarchical clustering dendrogram [21] of the author profiles (T and M) and the eleven unknown texts. Relative entropy (8) is used as a dissimilarity measure. Two different clusters arise, corresponding to the two authors. This means that in average two texts by the same author are not further apart than 0.06 but two texts from different authors are at a distance greater than 0.09.

The second numerical experiment varies the number of authors, see Table 2, as well as the number of known texts per author. The corpus of texts analyzed can be found in [20]. The text lengths vary from just over 4,000 to 100,000 words each. Texts longer than this were truncated to this maximum word count. The accuracy obtained can be observed in Table 3. E.g. focus on the 92% accuracy of the attribution with 4 authors and 2 known texts per author. To understand the source of this accuracy value, consider the first four authors in Table 2, these are Shakespeare, Twain, Austen, and Allen. Take 2 of their texts as known. In this way, there are $8+7+5+5 = 25$ unknown texts to attribute among these four authors. The accuracy of 92% indicates that 23 out of the 25 texts were correctly attributed by our method. The expected accuracy of random attribution is also informed in the last column of the table. The consistent difference between the accuracy of the proposed method and the one corresponding to random attribution is an indicator that the relational data in the MCs capture stylistic features of the authors.

The attribution between two authors in the first row of Table 3 is done between Shakespeare and Twain, who lived more than two centuries apart. Perfect accuracy is achieved with one known text from each author. This hints that little information is needed to distinguish between authors with marked differences in writing styles. Moreover, based on 2 known texts per author, the method can distinguish with maximum accuracy between three authors, these are Shakespeare, Twain, and Austen. For 3 known texts, the perfect attribution holds for 5 authors and for 6 known texts the method can correctly attribute the texts among 8 authors. The accuracy is deteriorated by increasing the number of candidate authors. For example, if we fix the known texts per author to be 3, then by increasing the number of candidates authors from 4 to 16 the accuracy is reduced from 100% to 85%. Furthermore, the accuracy increases when the number of known texts per author is increased. E.g., fixing the number of authors as 8, if we go from 1 training text to 6, we increase the accuracy from 67% to 100%. In columns with higher number of known texts, the accuracy deteriorates with the incorporation of more authors

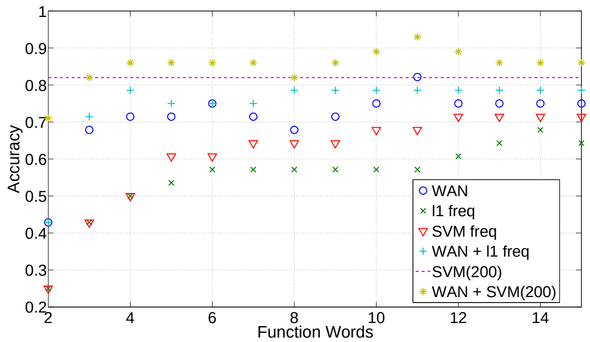

Fig. 2. Accuracy as a function of function words for a problem with 8 authors and 4 known texts per author [20]. The accuracy of the WAN method developed is higher than that of a frequency based SVM and a frequency based $l1$ norm for this range of function words. The accuracy of the combined methods is higher than that of their constituent methods indicating that relational and frequency based information encode different stylistic features.

at a lower rate. This indicates that when considering more known texts, a more reliable relational profile is built for each author, improving the accuracy of the attributions.

4.1. Comparison and Combination with Existing Methods

The method developed correctly attributes both anonymous texts in the Authorship Simulations proposed in [22] where 20 books are considered from 8 different authors. Furthermore, we present Fig. 2 where we depict the accuracy of a number of methods as a function of the amount of function words considered. This experiment is done for a pool of 8 authors with 4 known texts per author [20]. The $l1$ method consists in generating normalized frequency patterns of function words for each author from the known texts. A given unknown text is attributed to the author with minimal $l1$ distance to the frequency vector of such text. The support vector machine (SVM) method is a refinement of the $l1$ method. In the former, we start by applying a lineal, one-against-all SVM filter, i.e. we undertake a binary attribution between an author and every other author considered together. If the text is attributed to the single author, this author passes the filter. The $l1$ method is used to decide in case multiple candidates pass the filter. The method developed in this paper outperforms the $l1$ and the SVM based methods for this range of function words; see Fig 2. In fact, the SVM method based on 200 function words achieves the same accuracy obtained by the WAN method here described with 11 words. For low number of function words, almost 70% of the texts are correctly classified by the WAN method when focusing on the relation between the words *the, and,* and *a*.

When normalizing the word adjacency networks (7), purely relational information is retained. It has been critiqued that in methods that combine frequency with relational data [22], the source of the attribution accuracy is unclear [23]. By relying exclusively on relational data we settle this dispute. However, frequency data *is* useful and a more accurate method would be one that combines both sources of information. We present two ways of combining them.

The first one is by measuring dissimilarities between texts as the sum of the relative entropy measure (8) and a frequency based measure, e.g. the $l1$ norm of the frequency vectors. Fig. 2 shows that the combination achieves a higher accuracy than both methods considered separately. For example, for 8 function words the WAN and the $l1$ methods have accuracies of 68% and 57% respectively while their combination has an accuracy of 79%.

Another possible combination is to apply a frequency SVM filter as the one described for the pure SVM method based on a high number of function words, e.g. 200, and then choose among the candidates that pass the filter using a relational measure such as relative entropy (8). Figure 2 shows that this combined method achieves the highest accuracy among the methods considered with a maximum of 93% for 11 function words. Both combined methods show that the mixture of relational and frequency data yields higher accuracy than both information sources considered separately. This indicates that frequency and relational data encode different stylistic features.

The accuracy decreases when shorter texts are considered. We repeat the experiment in Fig. 2 with 500 word extracts of the previously utilized texts. The best WAN accuracy is 36% and is achieved for a network with 6 function words. This is approximately three times the expected accuracy of random attribution among 8 authors. The best SVM accuracy is also 36%. However, the accuracy when 200 function words are considered is 25%. Nevertheless, when considering the WAN+SVM method proposed, the combined accuracies of 25% and 36% yield a total accuracy of 43%. This reinforces the idea that both methods rely on complementary information.

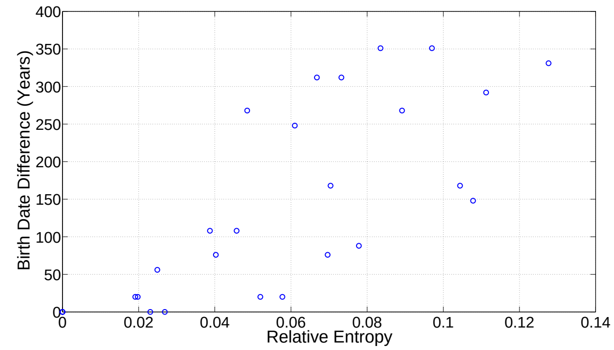

Fig. 3. Relative entropy between author profiles as a function of the year difference between birth dates for 9 authors born between 1564 and 1915 [20]. The increasing tendency of the scattered points indicates that relational structures encode literature style features common to contemporary authors.

4.2. Temporal Profiling

In Fig. 3 we depict the relative entropy, i.e. the dissimilarity between authorial styles given by our method, as a function of the year difference between the birth dates of the authors. The positive correlation observed hints that the historical period has a direct influence on the authorial style, even when considering content independent data as the use of function words. Therefore, we can use this authorship attribution method to estimate the period of time when the author of a given text lived. Profiling studies can be expanded to consider other characteristics such as gender and nationality.

5. CONCLUSION

An authorship attribution method based on relational data between function words was developed. Normalized word adjacency networks were used as relational structures. These networks were interpreted as Markov chains in order to facilitate their comparison using entropy measures. The accuracy of the method developed for long texts was presented using an ad-hoc corpus, comparisons with existing methods and an application in temporal profiling. Further, it was shown that an increase in accuracy can be achieved through combination with frequency based methods. Thus, unveiling the fact that relational and frequency based methods capture different aspects of stylometric information.

6. REFERENCES

[1] P. Juola, "Authorship attribution," *Foundations and Trends in Information Retrieval*, vol. 1, pp. 233–334, 2006.

[2] E. Stamatatos, "A survey of modern authorship attribution methods," *Journal of the American Society for Information Science and Technology*, vol. 60, pp. 538–556, March 2009.

[3] T. C. Mendenhall, "The characteristic curves of composition," *Science*, vol. 9, pp. 237–246, 1887.

[4] G. U. Yule, "On sentence-length as a statistical characteristic of style in prose: With application to two cases of disputed authorship," *Biometrika*, vol. 30, pp. 363–390, 1939.

[5] F. Mosteller and D. Wallace, "Inference and disputed authorship: The federalist," *Addison-Wesley*, 1964.

[6] J. F. Burrows, "an ocean where each kind...: Statistical analysis and some major determinants of literary style," *Computers and the Humanities*, vol. 23, pp. 309–321, 1989.

[7] D. I. Holmes and R. S. Forsyth, "The federalist revisited: New directions in authorship attribution," *Literary and Linguistic Computing*, vol. 10, pp. 111–127, 1995.

[8] R. S. Forsyth and D. I. Holmes, "Feature-finding for test classification," *Literary and Linguistic Computing*, vol. 11, pp. 163–174, 1996.

[9] G. U. Yule, "The statistical study of literary vocabulary," *CUP Archive*, 1944.

[10] D. I. Holmes, "Vocabulary richness and the prophetic voice," *Literary and Linguistic Computing*, vol. 6, pp. 259–268, 1991.

[11] F. J. Tweedie and R. H. Baayen., "How variable may a constant be? measures of lexical richness in perspective," *Computers and the Humanities*, vol. 32, pp. 323–352, 1998.

[12] D.L. Hoover, "Another perspective on vocabulary richness," *Computers and the Humanities*, vol. 37, pp. 151–178, 2003.

[13] M. Koppel, N. Akiva, and I. Dagan, "Feature instability as a criterion for selecting potential style markers," *Journal of the American Society for Information Science and Technology*, vol. 57, pp. 1519–1525, September 2006.

[14] D. Cutting, J. Kupiec, J. Pedersen, and P. Sibun, "A practical part-of-speech tagger," *Proceedings of the third conference on Applied Natural Language Processing*, pp. 133–140, 1992.

[15] D. V. Khemelev and F.J. Tweedie, "Using markov chains for identification of writers," *Literary and linguistic computing*, vol. 16, pp. 299–307, 2001.

[16] O. V. Kukushkina, A. A. Polikarpov, and D. V. Khmelev, "Using literal and grammatical statistics for authorship attribution," *Problems of Information Transmission*, vol. 37, pp. 172–184, 2001.

[17] C. Sanderson and S. Guenter, "Short text authorship attribution via sequence kernels, markov chains and author unmasking: An investigation," *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, vol. 37, pp. 482–491, 2006.

[18] R. Quirk, S. Greenbaum, G. Leech, and J. Svartvik, "A comprehensive grammar of the english language," *Longman*, 1985.

[19] G. Kesidis and J. Walrand, "Relative entropy between markov transition rate matrices," *IEEE Trans. Information Theory*, vol. 39, pp. 1056–1057, May 1993.

[20] S. Segarra, M. Eisen, and A. Ribeiro, "Compilation of texts used for the numerical experiments," *https://fling.seas.upenn.edu/ maeisen/wiki/index.php?n=Main.TextAttribution*.

[21] D. Müllner, "Modern hierarchical, agglomerative clustering algorithms," *arXiv:1109.2378*, September 2011.

[22] D.L. Hoover, "Frequent collocations and authorial style," *Literary and Linguistic Computing*, vol. 18, 2003.

[23] S. Argamon and S. Levitan, "Measuring the usefulness of function words for authorship attribution," *Proceedings of the 2005 ACH/ALLC Conference*, 2005.