Authorship Attribution through Function Word Adjacency Networks

Santiago Segarra, Mark Eisen, and Alejandro Ribeiro

***Abstract*—A method for authorship attribution based on function word adjacency networks (WANs) is introduced. Function words are parts of speech that express grammatical relationships between other words but do not carry lexical meaning on their own. In the WANs in this paper, nodes are function words and directed edges from a source function word to a target function word stand in for the likelihood of finding the latter in the ordered vicinity of the former. WANs of different authors can be interpreted as transition probabilities of a Markov chain and are therefore compared in terms of their relative entropies. Optimal selection of WAN parameters is studied and attribution accuracy is benchmarked across a diverse pool of authors and varying text lengths. This analysis shows that, since function words are independent of content, their use tends to be specific to an author and that the relational data captured by function WANs is a good summary of stylometric fingerprints. Attribution accuracy is observed to exceed the one achieved by methods that rely on word frequencies alone. Further combining WANs with methods that rely on word frequencies, results in larger attribution accuracy, indicating that both sources of information encode different aspects of authorial styles.**

I. Introduction

The discipline of authorship attribution is concerned with matching a text of unknown or disputed authorship to one of a group of potential candidates. More generally, it can be seen as a way of quantifying literary style or uncovering a stylometric fingerprint. The most traditional application of authorship attribution is literary research, but it has also been applied in forensics [2], defense intelligence [3] and plagiarism [4]. Both, the availability of electronic texts and advances in computational power and information processing, have boosted accuracy and interest in computer based authorship attribution methods [5]–[7].

Authorship attribution dates at least to more than a century ago with a work that proposed distinguishing authors by looking at word lengths [8]. This was later improved by [9] where the average length of sentences was considered as a determinant. A seminal development was the introduction of the analysis of function words to characterize authors' styles [10] which inspired the development of several methods. Function words are words like prepositions, conjunctions, and pronouns which on their own carry little meaning but dictate the grammatical relationships between words. The advantage of function words is that they are content independent and, thus, can carry information about the author that is not biased by the topic of the text being analyzed. Since [10], function words appeared in a number of papers where the analysis of the frequency with which different words appear in a text plays a central role one way or another; see e.g., [11]–[16]. Other attribution methods include the stylometric techniques in [17], the use of vocabulary richness as a stylometric marker [18]–[20] – see also [21] for a critique –, the use of stable words defined as those that can be replaced by an equivalent [22], and syntactical markers such as taggers of parts of speech [23]–[25]. Other recent methods have begun to use topic models to distinguish authors [26]–[28].

In this paper, we use function words to build stylometric fingerprints but, instead of focusing on their frequency of usage, we consider their relational structure. We encode these structures as word adjacency networks (WANs) which are asymmetric networks that store information of co-appearance of two function words in the same sentence (Section III). With proper normalization, edges of these networks describe the likelihood that a particular function word is encountered in the text given that we encountered another one. In turn, this implies that WANs can be reinterpreted as Markov chains describing transition probabilities between function words. Given this interpretation it is natural to measure the dissimilarity between different texts in terms of the relative entropy between the associated Markov chains (Section III-A). Markov chains have also been used as a tool for authorship attribution in [29]–[31]. However, the chains in these works represent transitions between letters, not words. Although there is little intuitive reasoning behind the notion that an author's style can be modeled by his usage of individual letters, these approaches generate somewhat positive results.

The classification accuracy of WANs depends on various parameters regarding the generation of the WANs as well as the selection of words chosen as network nodes. We consider the optimal selection of these parameters and develop an adaptive strategy to pick the best network node set given the texts to attribute (Section IV). Using a corpus composed of texts by 21 authors from the 19th century, we illustrate the implementation of our method and analyze the changes in accuracy when modifying the number of candidate authors as well as the length of the text of known (Section V-A) and unknown (Section V-B) authorship. Further, we analyze how the similarity of styles between two authors influences the accuracy when distinguishing their texts (Section V-C). We then incorporate authors from the early 17th century to the corpus and analyze how differences in time period, genre, and gender influence the classification rate of WANs (Sections VI-A to VI-C). We also show that WANs can be

Supported by NSF CAREER CCF-0952867 and NSF CCF-1217963. The authors are with the Department of Electrical and Systems Engineering, University of Pennsylvania, 200 South 33rd Street, Philadelphia, PA 19104. Email: {ssegarra, maeisen, aribeiro}@seas.upenn.edu. Part of the results in this paper appeared in [1].

used to detect collaboration between several authors (Section VI-D). We further demonstrate that our classifier performs better than techniques based on function word frequencies alone (Section VII). Perhaps more important, we show that the stylometric information captured by WANs is not the same as the information captured by word frequencies. Consequently, their combination results in a further increase in classification accuracy.

II. 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 III, 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 $r_U(u) \in A$. Notice that we assume that the real author of every unknown text is contained in the pool of candidate authors. 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. We define the classification accuracy of said estimator $\hat{r}_U$ as the fraction of unknown texts that are correctly attributed. With $\mathbb{I}$ denoting the indicator function we can write the classification accuracy ρ as

$$\rho(\hat{r}_U) = \frac{1}{k} \sum_{u \in U} \mathbb{I}\left\{\hat{r}_U(u) = r_U(u)\right\}. \tag{2}$$

We use $\rho(\hat{r}_U)$ to gauge performance of the proposed classifier in Sections IV to VII.

III. Function Words Adjacency Networks

In order to solve the proposed problem, we construct word adjacency networks (WANs) for the known texts $t \in T$ and unknown texts $u \in U$ and then build an estimator $\hat{r}_U$ based on the comparison of WANS.

WANs are weighted and directed networks that 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. 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 $0 < i(\omega_2) - i(\omega_1) \leq D$ is defined as

Common Function Words									
the	and	a	of	to	in	that	with	as	it
for	but	at	on	this	all	by	which	they	so
from	no	or	one	what	if	an	would	when	will

TABLE I: Most common function words in analyzed texts.

$$d(\omega_1, \omega_2) := \alpha^{i(\omega_2) - i(\omega_1) - 1}. \tag{3}$$

The directed proximity in (3) can be interpreted as the value of a *gappy* bigram [32]–[34] consisting of words ω_1 and ω_2 where α is the decaying factor that quantifies the magnitude of the gap between the pair of words.

In every sentence there are two kind of words: function and non-function words [35]. While in (3) the words w_1 and w_2 need not be function words, in this paper we are interested only in the case in which both w_1 and w_2 are function words. Function words are words that express primarily a grammatical relationship. These words include conjunctions (e.g., *and, or*), prepositions (e.g., *in, at*), quantifiers (e.g., *some, all*), modals (e.g., *may, could*), and determiners (e.g., *the, that*). We exclude gender specific pronouns (*he, she*) as well as pronouns that depend on narration type (*I, you*) from the set of function words to avoid biased similarity between texts written using the same grammatical person. The 30 function words that appear most often in our experiments are listed in Table I. For a full list of the function words considered, see [36]. 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 $\alpha^0 = 1$ and the directed proximity from the first *a* to *in* is $\alpha^1 = 0.8$. The directed proximity to *worth* or *load* is 0 because the indices of these words differ in more than $D = 4$.

To formally define a WAN, from a given text t we construct the network $W_t = (F, Q_t)$ where $F = \{f_1, f_2, ..., f_n\}$ is the set of nodes composed by a collection of function words common to all WANs being compared and $Q_t : F \times F \to \mathbb{R}_+$ is a similarity measure between pairs of nodes. Methods to select the elements of the node set F are discussed in Section IV.

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

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

for all $f_i, f_j \in F$ not necessarily distinct, 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 (3) for all appearances of f_i when the words are found at most D positions apart in the same sentence. Since in general $Q_t(f_i, f_j) \neq Q_t(f_j, f_i)$, the WANs generated are directed. Notice that the function in (4) combines into one similarity number the frequency of co-appearance of two words and the distance between these two words in each appearance, making both effects indistinguishable.

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

$$Q_t = \begin{array}{c} \\ \text{a} \\ \text{in} \\ \text{of} \\ \text{but} \end{array} \begin{array}{c} \begin{array}{cccc} \text{a} & \text{in} & \text{of} & \text{but} \end{array} \\ \begin{pmatrix} 0 & 3\times 0.8^1 & 0.8^1 & 0 \\ 2\times 0.8^3 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 1 & 0.8^2 & 0 & 0 \end{pmatrix} \end{array}. \tag{5}$$

The total similarity value from *a* to *in* is obtained by summing up the three 0.8^1 proximity values that appear in each sentence. Although the word *a* appears twice in every sentence, $Q(\text{a}, \text{a}) = 0$ 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

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

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 that there is at least one positively weighted edge out of every node f_i so that we are not dividing by zero. If this is not the case for some function word f_i, we fix $\hat{Q}_c(f_i, f_j) = 1/|F|$ for all f_j.

Example 3 By applying normalization (7) to the similarity function in Example 2, we obtain the following normalized similarity matrix

$$\hat{Q}_t = \begin{array}{c} \\ \text{a} \\ \text{in} \\ \text{of} \\ \text{but} \end{array} \begin{array}{c} \begin{array}{cccc} \text{a} & \text{in} & \text{of} & \text{but} \end{array} \\ \begin{pmatrix} 0 & 0.75 & 0.25 & 0 \\ 1 & 0 & 0 & 0 \\ 0.25 & 0.25 & 0.25 & 0.25 \\ 0.61 & 0.39 & 0 & 0 \end{pmatrix} \end{array}. \tag{8}$$

Similarity $\hat{Q}_t$ no longer depends on the length of the text t but on the relative frequency of the co-appearances of function words in the text.

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. $\hat{P}_c = (F, \hat{Q}_c)$ estimates P_c with the available known texts written by author a_c.

A. 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

$$H(P_1, P_2) = \sum_{i,j} \pi(f_i) P_1(f_i, f_j) \log \frac{P_1(f_i, f_j)}{P_2(f_i, f_j)}, \tag{9}$$

where π is the limiting distribution on P_1 and we consider $0 \log 0$ to be equal to 0. 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 of being observed in both MCs [37]. Relative entropy (9), also called Kullback-Leibler divergence rate [38], is a common dissimilarity measure among Markov chains and is used in a variety of applications such as face recognition [39] and gene analysis [40]. Notice that the limit distribution π in (9) retains some information about the frequency of appearance of the function words. E.g., for the MC in Example 3, the highest limit probability $\pi(a) = 0.44$ is obtained for the most frequent word *a* while the lowest limit probability $\pi(but) = 0.04$ is achieved by one of the two words that appears only once in the text fragment in Example 1. We point out that relative entropy measures have also been used to compare vectors with function word frequencies [41]. This is unrelated to their use here as measures of the relational information captured in function WANs. Attribution in [42] is also based on the comparison of graphs via information theoretic measures. However, both the graphs constructed and the measure used differ from those developed in this paper.

Using (9), we generate the attribution function $\hat{r}_U(u)$ by assigning the text u to the author with the most similar relational structure

$$\hat{r}_U(u) = a_p, \text{ where } p = \underset{c}{\operatorname{argmin}}\, H(P_u, \hat{P}_c). \tag{10}$$

Whenever a transition between words appears in an unknown text but not in a profile, the relative entropy in (10) takes an infinite value for the corresponding author. In practice we compute the relative entropy in (9) by summing over the non-zero transitions in the profiles,

$$H(P_1, P_2) = \sum_{i,j|P_2(f_i,f_j)\neq 0} \pi(f_i) P_1(f_i, f_j) \log \frac{P_1(f_i, f_j)}{P_2(f_i, f_j)}. \tag{11}$$

Observe that if there is a transition between words that appears often in the text P_1 but never in the profile P_2, the expression in (11) skips the relative entropy summand. This is undesirable because the often appearance of this transition in the text network P_1 is a strong indication that this text was not written by the author whose profile network is P_2. The expression in (9) would capture this difference by producing an infinite value for the relative entropy. However, this infinite value is still produced if a transition between words does not appear in the author profile P_2 and appears just once in the text P_1. In this case, the null contribution to the relative entropy in (11) is more reasonable than the infinity contribution in (9) because the rarity of the transition in both texts is an indication that the text and the profile belong to the same author. Our experiments show that the latter situation is more common than the former. Transitions rare enough so as not to appear in a profile are, for the most part, also infrequent in all texts. This is reasonable because rare combinations of function words are properties of the language more than of individual authors. We have also explored the use of Laplace smoothing to avoid infinite entropies – see e.g., [43, Chapter 13], but (11) still achieves better results in practice.

Most of the computational burden of the method proposed resides in the construction of the WANs, that is, going from the written texts to the dissimilarity function Q_t in (4). Nevertheless, this is a one-time effort given that once the WAN is built, it can be utilized for various attribution problems. The attribution time is based on the computation of relative entropies [cf. (9) and (11)] which, for the network sizes considered in practice, take in the order of 3 ms. In this way, the attribution of a text among, e.g., 10 authors takes approximately 30 ms.

We proceed to compare our network similarity approach with the more conventional maximum likelihood test for Markov chains after the following remark.

Remark 1 For the relative entropies in (10) to be well defined, the MCs P_u associated with the unknown texts have to be ergodic to ensure that the limiting distributions π in (9) and (11) are unique. In practice, this is usually true if the texts that generated P_u are sufficiently long. If this is not true for a particular network, then the limiting distribution π is not well defined since it depends on the state in which the MC is initialized. Hence, for these cases, we replace $\pi(f_i)$ in (9) and (11) by the expected fraction of time $\bar{\pi}(f_i)$ that a randomly initialized walk spends in state f_i. The random initial state – function word – is drawn from a distribution given by the relative function word frequencies in the text. Formally

$$\bar{\pi}^T = \lim_{t\to\infty} p^T {P_1}^t, \tag{12}$$

where p is a vector of length $|F|$ and contains in the i-th position the relative frequency of function word f_i in the unknown text u. Notice that for ergodic P_1, π coincides with $\bar{\pi}$ independently of the probability distribution p.

B. Network Similarity vs. Maximum Likelihood Estimation

A more conventional approach towards the attribution of an unknown text among a group of authors associated to Markov chains would be as follows: we first relate the unknown text to a particular realization of a Markov chain and then assign such text to the author whose MC has the largest probability of outputting such a realization. Essentially, we would be computing the maximum likelihood test [44] for a given unknown text. As it turns out, whenever a Markov chain realization *can* be associated with a text, both the relative entropy and the maximum likelihood approaches are equivalent.

To be more specific, let us define the following problem: we are given two Markov chains P and R defined on the same state space of function words $F = \{f_1, f_2, \ldots, f_n\}$. We are also given a walk W on the state space F, i.e., an ordered sequence of words $W = \{w_i\}$ for $i = 1, \ldots, m$ such that $w_i \in F$ for all i. We want to compare this walk W with the chains P and R. In particular, we want to compare the maximum likelihood approach with the network similarity one proposed in Section III-A. For the former, we find the likelihood of W being generated by P and R and *assign* the walk to the chain with the largest likelihood. For the latter, we empirically generate a third chain Q based on the walk W and compare Q with P and R via relative entropies (9). We then *assign* W to the chain closer to Q.

Indeed, both approaches are equivalent. To see this, observe that the log-likelihood that W is generated by one of the chains, say P is given by

$$\mathcal{L}(W, P) = \sum_{k=1}^{m-1} \log(P(w_k, w_{k+1})). \tag{13}$$

Notice that for two particular function words $f_i, f_j \in F$, in the above summation the term $\log(P(f_i, f_j))$ appears $F_i Q(f_i, f_j)$ times where F_i is the number of times that state f_i appears in walk W without considering its last state. This implies that

$$\mathcal{L}(W, P) = \sum_{i,j} F_i Q(f_i, f_j) \log(P(f_i, f_j)). \tag{14}$$

If we apply any strictly decreasing function $g : \mathbb{R} \to \mathbb{R}$ to $\mathcal{L}$, then the lower the value of $h := g \circ \mathcal{L}$ the more similar W is to the chain P. In particular, pick the function

$$g(x) = -\frac{x}{m-1} + \sum_{i,j} \frac{F_i}{m-1} Q(f_i, f_j) \log(Q(f_i, f_j)). \tag{15}$$

Notice that g does not depend on P. By composing g with $\mathcal{L}$ we obtain

$$h(W,P)=g\circ\mathcal{L}(W,P)=\sum_{i,j}\frac{F_i}{m-1}Q(f_i,f_j)\log\frac{Q(f_i,f_j)}{P(f_i,f_j)}. \tag{16}$$

Finally, for an ergodic MC, the fraction $F_i/(m-1)$ is very close to the limit probability $\pi(f_i)$ except for a minor border effect in the first and last observation. Thus, we obtain that

$$h(W,P)\approx\sum_{i,j}\pi(f_i)Q(f_i,f_j)\log\frac{Q(f_i,f_j)}{P(f_i,f_j)}=H(Q,P), \tag{17}$$

where H stands for relative entropy (9). Hence, the larger the likelihood between W and P, the smaller the entropy between Q and P, making both approaches equivalent.

Even though both approaches are identical when a chain realization can be defined for the text to attribute, for the purposes of our paper the network similarity approach is preferable since each text does not clearly define a walk W. For each appearance of a function word in a text, we do not only consider the transition to the next function word but, in turn, we consider a distribution of possible transitions over the next words with a dampening factor. This distribution is naturally represented by a MC itself. In particular, if in a text two function words appear always as a pair but with a third function word in between, we want to detect this. The associated walk W would not record any transition between these two words whereas the MC built does capture this interaction. For this reason, we choose the relative entropy operator H to compare Markov Chains.

In the next section, we proceed to specify the selection of function words in F for the construction of WANs as well as the choice of the parameters α and D.

IV. Selection of Function Words and WAN Parameters

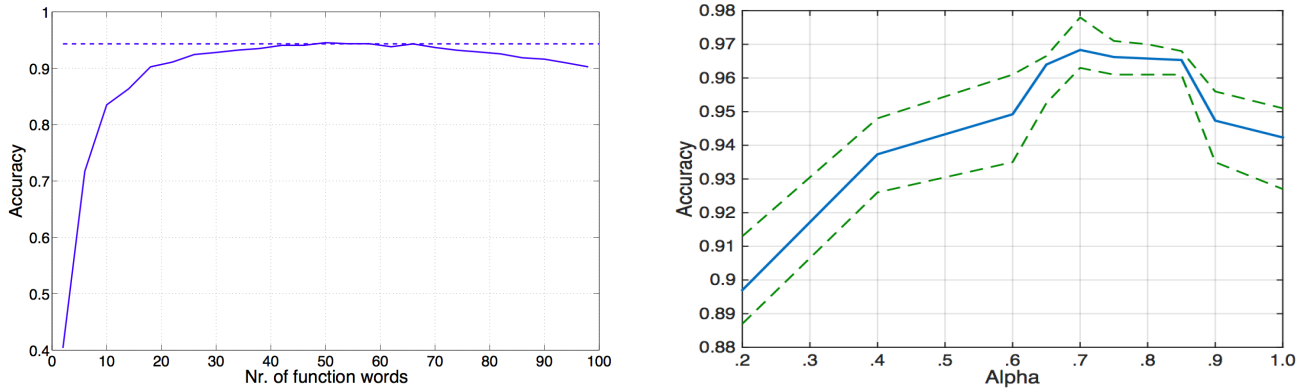
The classification accuracy of the function WANs introduced in Section III depends on the choice of several variables and parameters: the set of sentence delimiters or stopper symbols, the window length D, the discount factor α, and the set of function words F defining the nodes of the adjacency networks. In this section, we study the selection of these parameters to maximize classification accuracy.

The selections of stopper symbols and window lengths are not critical. As stoppers we include the grammatical sentence delimiters '.', '?' and '!', as well as semicolons ';' to form the stopper set $\{.\ ?\ !\ ;\}$. We include semicolons since they are used primarily to connect two independent clauses [35]. In any event, the inclusion or not of the semicolon as a stopper symbol entails a minor change in the generation of WANs due to its infrequent use. As window length we pick $D=10$, i.e., we consider that two words are not related if they appear more than 10 positions apart from each other. Larger values of D lead to higher computational complexity without increase in accuracy since grammatical relations of words more than 10 positions apart are rare.

In order to choose which function words to include when generating the WANs we present two different approaches: a static methodology and an adaptive strategy. The static approach consists in picking the function words – among all the functions words considered; see [36] – most frequently used in the union of *all* the texts being considered in the attribution, i.e, all those that we use to build the profile and those being attributed. By using the most frequent function words we base the attribution on repeated grammatical structures and limit the influence of noise introduced by unusual sequences of words which are not consistent stylometric markers. In our experiments, we see that selecting a number of functions words between 40 and 70 yields optimal accuracy. For way of illustration, we consider in Fig. 1a the attribution of 1,000 texts of length 10,000 words among 7 authors chosen at random from our pool of 19th century authors [36] for a fixed value of $\alpha=0.75$ and profiles of 100,000 words – see also Section V for a description of the corpus. The solid line in this figure represents the accuracy achieved when using a network composed of the n most common function words in the texts analyzed for n going from 2 to 100. Accuracy is maximal when we use exactly 50 function words, but the differences are minimal and likely due to random variations for values of n between $n=42$ and $n=66$. The flatness of the accuracy curve is convenient because it shows that the selection of n is not that critical. In this particular example we can choose any value between, say $n=45$ and $n=60$, without affecting reliability. In a larger test where we also vary the length of the profiles, the length of the texts attributed, and the number of candidate authors, we find that including 60 function words is empirically optimal.

The adaptive approach still uses the most common function words but adapts the number of function words used to the specific attribution problem. In order to choose the number of function words, we implement repeated leave-one-out cross validation as follows. For every candidate author $a_i\in A$, we concatenate all the known texts $T^{(i)}$ written by a_i and then break up this collection into N pieces of equal length. We build a profile for each author by randomly picking $N-1$ pieces for each of them. We then attribute the unused pieces between the authors utilizing WANs containing the n most common function words for n varying in a given interval $[n_{\min},n_{\max}]$. We perform M of these cross validation rounds in which we change the random selection of the $N-1$ texts that build the profiles. The value of n that maximizes accuracy across these M trials is selected as the number of nodes for the WANs. We perform attributions using the corresponding n word WANs for the profiles as well as for the texts to be attributed. In our numerical experiments we have found that using $N=10$, $n_{\min}=20$, $n_{\max}=80$, and M varying between 10 and 100 depending on the available computation time are sufficient to find values of n that yield good performance.

The dashed line in Fig. 1a represents the accuracy obtained by implementing the adaptive strategy with $N=10$, $n_{\min}=20$, $n_{\max}=80$, and $M=100$ for the same attribution problem considered in the static method – i.e., attribution of 1,000 texts of length 10,000 words among 7 authors for $\alpha=0.75$ and



(a) Attribution accuracy as a function of the network size.

(b) Attribution accuracy as a function of the discount factor α.

Fig. 1: Both figures present the accuracy for the attribution of 1,000 texts of length 10,000 words among 7 authors chosen at random with 100,000 words profiles. (a) The solid line represents the accuracy achieved for static networks of increasing size. The dashed line is the accuracy obtained by the adaptive method. (b) Mean accuracy (blue) is maximized for values of the discount factor α in the range between 0.65 and 0.85. Percentiles - 25 and 75 - are depicted in dashed green.

profiles of 100,000 words. The accuracy is very similar to the best correct classification rate achieved by the static method. This is not just true of this particular example but also true in general. The static approach is faster because it requires no online training to select the number of words n to use in the WANs. The adaptive strategy is suitable for a wider range of problems because it contains less assumptions than the static method about the best structure to differentiate between the candidate authors. E.g., when shorter texts are analyzed, experiments show that the optimal static method uses slightly less than 60 words. Likewise, the optimal choice of the number of words in the WANs changes slightly with the time period of the authors, the specific authors considered, and the choice of parameter α. These changes are captured by the adaptive approach. We advocate adaptation in general and reserve the static method for rapid attribution of texts or cases when the number of texts available to build profiles is too small for effective cross-validation.

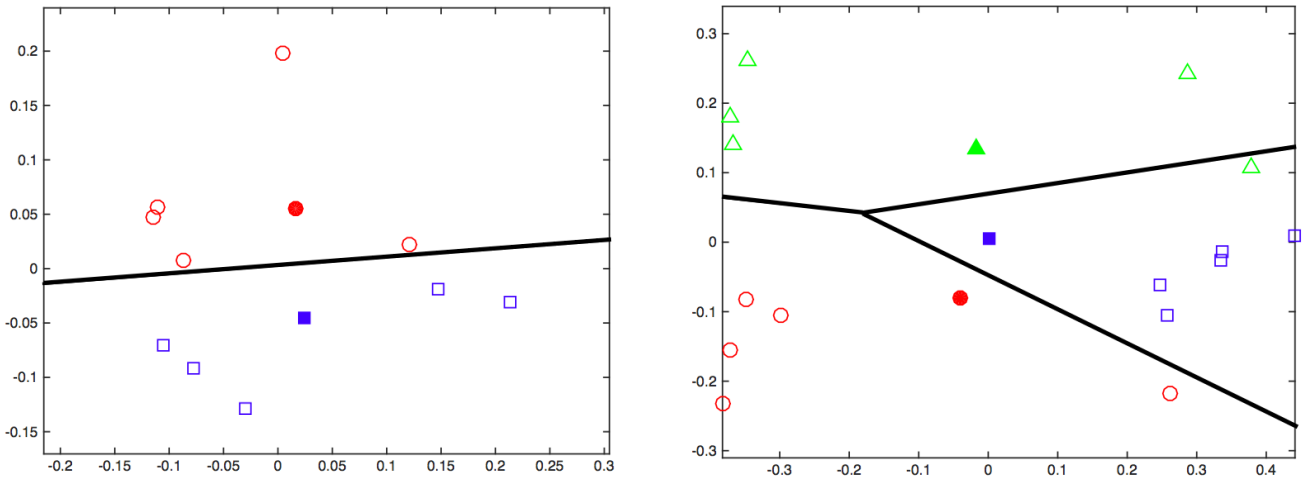
To select the decay parameter we use the adaptive leave-one-out cross validation method for different values of α and study the variation of the correct classification rate as α varies. In Fig. 1b we show the variation of the correct classification rate with α when attributing 1,000 texts of length 10,000 words between 7 authors of the 19th century picked at random from our text corpus [36] using profiles with 100,000 words – see also Section V for a description of the corpus. As in the case of the number of words used in the WANs there is a wide range of values for which variations are minimal and likely due to randomness. This range lies approximately between $\alpha = 0.65$ and $\alpha = 0.85$. Notice that for the particular case of $\alpha = 1$, the WANs store the frequencies of appearances for pairs of function words within the window length D. However, Fig. 1b reveals that the discounted approach where $\alpha < 1$ achieves better results when α is optimized. In a larger test where we also vary text and profile lengths as well as the number of candidate authors we find that $\alpha = 0.75$ is optimal. We found no gains in an adaptive method to choose α.

V. Attribution Accuracy

Henceforth, we fix the WAN generation parameters to the optimal values found in Section IV, i.e., the set of sentence delimiters is $\{ \ . \ ? \ ! \ ; \ \}$, the discount factor is $\alpha = 0.75$, and the window length is $D = 10$. The set of function words F is picked adaptively for every attribution problem by performing $M = 10$ cross validation rounds.

The text corpus used for the simulations consists of authors from two different periods [36]. The first group corresponds to 21 authors spanning the 19th century, both American – such as Nathaniel Hawthorne and Herman Melville – and British – such as Jane Austen and Charles Dickens. For these 21 authors, we have an average of 6.5 books per author with a minimum of 4 books for Charlotte Bronte and a maximum of 10 books for Herman Melville and Mark Twain. In terms of words, this translates into an average of 560,000 words available per author with a minimum of 284,000 words for Louisa May Alcott and a maximum of 1,096,000 for Mark Twain. The second group of authors corresponds to 7 Early Modern English playwrights spanning the late 16th century and the early 17th century, namely William Shakespeare, George Chapman, John Fletcher, Ben Jonson, Christopher Marlowe, Thomas Middleton, and George Peele. For these authors we have an average of 22 plays per author with a minimum of 4 plays for Peele and a maximum of 47 plays written either completely or partially by Fletcher. In terms of word length, we count with an average length of 400,000 words per author with a minimum of 50,000 for Peele and a maximum of 900,000 for Fletcher.

To illustrate authorship attribution with function WANs, we solve an authorship attribution problem with two candidate authors: Mark Twain and Herman Melville. For each candidate author we are given five known texts and are asked to attribute ten unknown texts, five of which were written by Twain while the other five belong to Melville [36]. Every text in this attribution belongs to a different book and corresponds to a 10,000 word extract, i.e. around 25 pages of a paper back



(a) MDS representation for two authors.

(b) MDS representation for three authors.

Fig. 2: (a) Perfect accuracy is attained for two candidate authors. Every empty marker falls in the half plane corresponding to the associated filled marker of their color. (b) One mistake is made for three authors. One green empty triangle falls in the region attributable to the blue square author.

midsize edition. The five known texts from each author are used to generate corresponding profiles as described in Section III. Relative entropies in (11) from each of the ten unknown texts to each of the two resulting profiles are then computed.

Since relative entropies are not metrics, we use multidimensional scaling (MDS) [45] to embed the two profiles and the ten unknown texts in 2-dimensional Euclidean metric space with minimum distortion. The result is illustrated in Fig. 2a. Twain's profile is depicted as a filled red circle whereas Melville's profile is depicted as a filled blue square. Unknown texts are depicted as empty circles and squares, where the color and the shape indicates the real author, i.e. red circles for Twain and blue squares for Melville. A solid black line composed of points equidistant to both profiles is also plotted. This line delimits the two half planes that result in attribution to one author or the other. From Fig. 2a, we see that the attribution is perfect for these two authors. All red (Twain) empty circles fall in the half plane closer to the filled red circle and all blue (Melville) empty squares fall in the half plane closer to the filled blue square. We emphasize that the WAN attributions are not based on these Euclidean distances but on the non-metric dissimilarities given by the relative entropies. Since the number of points is small, the MDS distortion is minor and the distances in Fig. 2a are close to the relative entropies. The latter separate the points better, i.e., relative entropies are smaller for texts of the same author and larger for texts of different authors.

We also illustrate an attribution between three authors by creating a profile for Jane Austen using five 10,000 word excerpts and adding five 10,000 word excerpts of texts written by Jane Austen to the ten excerpts to attribute from Twain and Melville's books. We then perform an attribution of the 15 texts to the three profiles constructed. An MDS approximate representation of the relative entropies between texts and profiles is shown in Fig. 2b where the filled green triangle represents Austen's profile and the empty green triangles represent her texts to attribute. Circles and squares still represent Twain's and Melville's works, respectively. We also plot the Voronoi tessellation induced by the three profiles, which specify the regions of the plane that are attributable to each author. Different from the case in Fig. 2a, attribution is not perfect since one of Austen's texts is mistakenly attributed to Melville. This is represented in Fig. 2b by the green empty triangle that appears in the section of the Voronoi tessellation that corresponds to the blue square profile. In general, for larger number of candidate authors, the distortion introduced by the MDS embedding is higher, compromising the reliability of any classifier based on the low-dimensional metric representation. Notice that this does not affect the WAN attribution, which is based on the non-metric dissimilarities given by the relative entropies.

Besides the number of authors, the other principal determinants of classification accuracy are the length of the profiles (training set), the length of the texts of unknown authorship (testing set), and the similarity of writing styles as captured by the relative entropy dissimilarities between profiles. We study these effects in sections V-A,V-B, and V-C, respectively.

A. *Varying the Training Set: Length of Profiles*

The profile (training set) length is defined as the total number of words, function or otherwise, used to construct the profile. To study the effect of varying the training set size, we fix $\alpha = 0.75$, $D = 10$, and vary the length of author profiles from 10,000 to 100,000 words in increments of 10,000 words. For each profile length, we attribute texts containing 25,000, 5,000 and 1,000 words, i.e., we consider three different testing set sizes. Moreover, for each given combination of profile and text length, we consider problems ranging from binary attribution to attribution between ten authors. To build profiles, we use ten texts of the same length randomly chosen among all the texts written by a given author. The length of each excerpt is such that the ten pieces add up to the desired

Nr. of authors	Number of words in profile (thousands)										Rand.
	10	**20**	**30**	**40**	**50**	**60**	**70**	**80**	**90**	**100**	
2	0.927	0.964	0.984	0.985	0.981	0.979	0.981	0.986	0.992	0.988	0.500
3	0.871	0.934	0.949	0.962	0.968	0.975	0.982	0.978	0.974	0.978	0.333
4	0.833	0.905	0.931	0.949	0.948	0.964	0.963	0.968	0.969	0.977	0.250
5	0.800	0.887	0.923	0.950	0.945	0.951	0.953	0.961	0.961	0.969	0.200
6	0.760	0.880	0.929	0.932	0.937	0.941	0.948	0.952	0.950	0.973	0.167
7	0.755	0.851	0.909	0.924	0.937	0.943	0.937	0.957	0.960	0.957	0.143
8	0.722	0.841	0.898	0.911	0.932	0.941	0.938	0.947	0.952	0.955	0.125
9	0.711	0.855	0.882	0.905	0.915	0.931	0.932	0.944	0.952	0.955	0.111
10	0.701	0.827	0.882	0.910	0.923	0.923	0.934	0.935	0.943	0.935	0.100

TABLE II: Profile length vs. accuracy for different number of authors (text length = 25,000)

Nr. of authors	Number of words in profile (thousands)										Rand.
	10	**20**	**30**	**40**	**50**	**60**	**70**	**80**	**90**	**100**	
2	0.863	0.930	0.932	0.945	0.928	0.952	0.942	0.907	0.942	0.967	0.500
3	0.821	0.884	0.886	0.890	0.910	0.901	0.943	0.912	0.911	0.914	0.333
4	0.728	0.833	0.849	0.862	0.892	0.867	0.888	0.905	0.882	0.885	0.250
5	0.698	0.819	0.825	0.839	0.862	0.884	0.859	0.865	0.882	0.893	0.200
6	0.673	0.754	0.789	0.798	0.832	0.837	0.863	0.870	0.896	0.878	0.167
7	0.616	0.754	0.806	0.838	0.812	0.848	0.859	0.854	0.873	0.868	0.143
8	0.600	0.720	0.748	0.820	0.805	0.831	0.831	0.854	0.857	0.850	0.125
9	0.587	0.718	0.767	0.781	0.796	0.809	0.833	0.849	0.843	0.847	0.111
10	0.556	0.693	0.737	0.753	0.805	0.827	0.829	0.824	0.843	0.846	0.100

TABLE III: Profile length vs. accuracy for different number of authors (text length = 5,000)

Nr. of authors	Number of words in profile (thousands)										Rand.
	10	**20**	**30**	**40**	**50**	**60**	**70**	**80**	**90**	**100**	
2	0.738	0.788	0.747	0.823	0.803	0.803	0.802	0.800	0.812	0.793	0.500
3	0.599	0.698	0.690	0.737	0.713	0.744	0.724	0.726	0.757	0.701	0.333
4	0.528	0.638	0.640	0.672	0.658	0.663	0.656	0.663	0.651	0.707	0.250
5	0.491	0.561	0.598	0.627	0.686	0.621	0.633	0.661	0.674	0.632	0.200
6	0.469	0.549	0.578	0.593	0.626	0.594	0.598	0.617	0.606	0.582	0.167
7	0.420	0.469	0.539	0.551	0.583	0.564	0.603	0.593	0.583	0.598	0.143
8	0.392	0.454	0.544	0.540	0.572	0.551	0.583	0.589	0.563	0.599	0.125
9	0.385	0.449	0.489	0.528	0.519	0.556	0.551	0.580	0.560	0.576	0.111
10	0.353	0.410	0.466	0.480	0.506	0.536	0.529	0.542	0.556	0.553	0.100

TABLE IV: Profile length vs. accuracy for different number of authors (text length = 1,000)

profile length. E.g., to build a profile of length 50,000 words for Melville, we randomly pick ten excerpts of 5,000 words each among all the texts written by him. For the texts to be attributed, however, we always select contiguous extracts of the desired length. E.g., for texts of length 25,000 words, we randomly pick excerpts of this length written by some author – as opposed to the selection of ten pieces of different origin we do for the profiles. This resembles the usual situation where the profiles are built from several sources but the texts to attribute correspond to a single literary creation. For a given profile size and number of authors, several attribution experiments were run by randomly choosing the set of authors among those from the 19th century [36] and randomly choosing the texts forming the profiles. The amount of attribution experiments was chosen large enough to ensure that every accuracy value in tables II - IV is based on the attribution of at least 600 texts.

The accuracy results of attributing a text of 25,000 words are stated in Table II. This word length is equivalent to around 60 pages of a midsize paperback novel – i.e., a novella, or a few book chapters – or the typical length of a Shakespeare play. In the last column of the table we inform the expected accuracy of random attribution between the candidate authors. The purpose of this column is *not* to provide a performance benchmark. However, the difference between the accuracies of this column and the rest of the table indicates that WANs *do* carry stylometric information useful for authorship attribution. For a comparison of the performance of WAN attribution with state of the art classifiers see Section VII. Overall, attribution of texts with 25,000 words can be done with high accuracy even when attributing among a large number of authors if reasonably large corpora are available to build author profiles with 60,000 to 100,000 words. E.g., for a profile containing 40,000 words, our method achieves an accuracy of 0.985 for binary attributions whereas the corresponding random accuracy is 0.5. As expected, accuracy decreases when the number of candidate authors increases. E.g., for profiles of 80,000 words, an accuracy of 0.986 is obtained for binary attributions whereas an accuracy of 0.935 is obtained when the pool of candidates contains ten authors.

Accuracy increases with longer profiles. E.g., when performing attributions of 25,000 word texts among 6 authors, the accuracy obtained for profiles of length 10,000 is 0.760 whereas the accuracy obtained for profiles of length 60,000 is 0.941. There is a saturation effect concerning the length of the profile that depends on the number of authors being considered. For binary attributions there is no major increase in accuracy beyond profiles of length 30,000. However, when the number of candidate authors is 7, accuracy stabilizes for profiles of length in the order of 80,000 words. There seems to be little benefit in using profiles containing more than 100,000

words, which corresponds to a short novel of about 250 pages.

Correct attribution rates of shorter excerpts containing 5,000 words are shown in Table III for the same profile lengths and number of candidate authors considered in Table II. A text of this length corresponds to about 13 pages of a novel – something in the order of the chapter of a book – or an act in a Shakespeare play. When considering these shorter texts, acceptable classification accuracy is achieved except for very short profiles and large number of authors, while reliable attribution requires a small number of candidate authors or a large profile. E.g., attribution between three authors with profiles of 70,000 words has an average accuracy of 0.943. While smaller than the corresponding correct attribution rate of 0.982 for texts of length 25,000 words, this is still a respectable number. To achieve an accuracy in excess of 0.9 for the case of three authors we need a profile of at least 50,000 words.

For very short texts of 1,000 words, which is about the length of an opinion piece in a newspaper, a couple pages in a novel, or a scene in a Shakespeare play, we can provide indications of authorship but cannot make definitive claims. As shown in Table IV, the best accuracies are for binary attributions that hover at around 0.8 when we use profiles longer than 40,000 words. For attributions between more than 2 authors, maximum correct attribution rates are achieved for profiles containing 90,000 or 100,000 words and range from 0.757 for the case of three authors to 0.556 when considering ten authors. These rates are markedly better than random attribution but not sufficient for definitive statements. The results can be of use as circumstantial evidence in support of attribution claims substantiated by further proof.

B. Varying the Testing Set: Length of Texts to Attribute

In this section we analyze the effect of varying the length of the texts to attribute (testing set) in attribution accuracy for different profile lengths (training set) and number of candidate authors. Using $\alpha = 0.75$ and $D = 10$, we consider profiles of length 100,000, 20,000 and 5,000 words and vary the number of candidate authors from two to ten. The lengths of the texts to attribute considered are 1,000 words to 6,000 words in 1,000 word increments, 8,000 words, and 10,000 to 30,000 words in 5,000 word increments. We use the finer resolution of 1,000 word increments for short texts, since the attribution accuracy is very sensitive to the text length in this regime. As in Section V-A, for every combination of number of authors and text length, enough independent attribution experiments were performed to ensure that every accuracy value in tables V - VII is based on at least 600 attributions.

For profiles of length 100,000 words, the results are reported in Table V. As done in tables II-IV, we state the expected accuracy of random attribution in the last column of the table. Accuracies reported towards the right end of the table, i.e. 20,000-30,000 words, correspond to the attribution of a dramatic play or around 60 pages of a novel, which we will refer to as long texts. Accuracies for columns in the middle of the table, i.e. 5,000-8,000 words, correspond to an act in a dramatic play or between 12 and 20 pages of a novel, which we will refer to as medium texts. The left columns of this table, i.e. 1,000-3,000 words, correspond to a scene in a play, 2 to 7 pages in a novel or an article in a newspaper, which we will refer to as short texts. For the attribution of long texts, we achieve a mean accuracy of 0.988 for binary attributions which decreases to an average accuracy of 0.945 when the number of candidate authors is increased to ten. For medium texts, the decrease in accuracy is not very significant for binary attributions, with a mean accuracy of 0.955, but the accuracy is reduced to 0.856 for attributions among ten authors. The accuracy is decreased further when attributing short texts, with a mean accuracy of 0.894 for binary attributions and 0.700 for the case with ten candidates. This indicates that when profiles of around 100,000 are available, WANS achieve accuracies over 0.95 for medium to long texts. For short texts, acceptable classification rates are achieved if the number of candidate authors is between two and four.

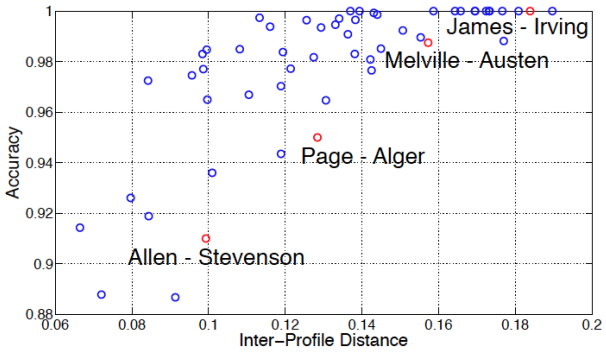


Fig. 3: Binary attribution accuracy as a function of the inter-profile dissimilarity. Higher accuracy is attained for attribution between authors which are more dissimilar.

If we reduce the length of the profiles to 20,000 words, reasonable accuracies are attained for small pools of candidate authors; see Table VI. E.g, for binary attributions, the range of correct classification varies between 0.812 for texts of 1,000 words to 0.969 for texts with 30,000 words. The first of these numbers means that we can correctly attribute a newspaper opinion piece with accuracy 0.812 if we are given corpora of 20 opinion pieces by the candidate authors. The second of these numbers means that we can correctly attribute a play between two authors with accuracy 0.969 if we are given corpora of 20,000 words by the candidate authors. Further reducing the profile length to 5,000 words results in classification accuracies that are acceptable only when we consider binary attributions and texts of at least 10,000 words; see Table VII. For shorter texts or larger number of candidate authors, WANs can provide supporting evidence but not definitive proof.

In Sections V-A and V-B, the profiles across all candidates authors are balanced, i.e. they contain the same number of words. Attribution can be performed in scenarios with unbalanced profiles where the shortest profile contains n_{short} words and the longest one contains n_{long} words. In this case, the accuracy obtained is lower than the one corresponding to a balanced scenario with the same number of candidate authors and every profile of length n_{long} and larger than that of a balanced scenario with profiles of length n_{short} words.

Nr. of authors	Number of words in texts (thousands)												Rand.
	1	**2**	**3**	**4**	**5**	**6**	**8**	**10**	**15**	**20**	**25**	**30**	
2	0.840	0.917	0.925	0.938	0.940	0.967	0.958	0.977	0.967	0.989	0.988	0.986	0.500
3	0.789	0.873	0.890	0.919	0.913	0.932	0.936	0.956	0.952	0.979	0.979	0.975	0.333
4	0.736	0.842	0.870	0.902	0.906	0.933	0.937	0.952	0.965	0.970	0.973	0.974	0.250
5	0.711	0.797	0.858	0.874	0.891	0.906	0.924	0.925	0.955	0.971	0.980	0.964	0.200
6	0.690	0.796	0.828	0.886	0.884	0.911	0.919	0.922	0.944	0.957	0.969	0.961	0.167
7	0.633	0.730	0.814	0.855	0.874	0.890	0.910	0.911	0.928	0.947	0.956	0.951	0.143
8	0.602	0.740	0.811	0.846	0.882	0.887	0.915	0.910	0.930	0.944	0.957	0.963	0.125
9	0.607	0.721	0.774	0.826	0.845	0.870	0.889	0.890	0.918	0.948	0.951	0.953	0.111
10	0.578	0.731	0.792	0.816	0.842	0.855	0.872	0.893	0.921	0.933	0.942	0.961	0.100

TABLE V: Text length vs. accuracy for different number of authors (profile length = 100,000)

Nr. of authors	Number of words in texts (thousands)												Rand.
	1	**2**	**3**	**4**	**5**	**6**	**8**	**10**	**15**	**20**	**25**	**30**	
2	0.812	0.850	0.903	0.912	0.913	0.912	0.938	0.945	0.918	0.964	0.964	0.969	0.500
3	0.760	0.797	0.858	0.899	0.887	0.918	0.920	0.918	0.919	0.938	0.929	0.928	0.333
4	0.670	0.747	0.813	0.852	0.868	0.887	0.889	0.906	0.918	0.915	0.900	0.913	0.250
5	0.621	0.721	0.749	0.813	0.823	0.819	0.859	0.878	0.876	0.887	0.889	0.893	0.200
6	0.557	0.681	0.754	0.782	0.799	0.831	0.852	0.866	0.871	0.879	0.881	0.872	0.167
7	0.493	0.610	0.674	0.706	0.731	0.770	0.798	0.807	0.828	0.862	0.867	0.858	0.143
8	0.467	0.623	0.675	0.721	0.741	0.769	0.790	0.826	0.822	0.857	0.841	0.857	0.125
9	0.474	0.574	0.656	0.672	0.710	0.734	0.781	0.783	0.813	0.845	0.837	0.841	0.111
10	0.433	0.535	0.612	0.663	0.684	0.706	0.752	0.772	0.836	0.840	0.848	0.848	0.100

TABLE VI: Text length vs. accuracy for different number of authors (profile length = 20,000)

Nr. of authors	Number of words in texts (thousands)												Rand.
	1	**2**	**3**	**4**	**5**	**6**	**8**	**10**	**15**	**20**	**25**	**30**	
2	0.672	0.740	0.747	0.707	0.803	0.823	0.788	0.848	0.820	0.802	0.827	0.832	0.500
3	0.547	0.623	0.626	0.653	0.744	0.710	0.712	0.757	0.736	0.764	0.734	0.733	0.333
4	0.452	0.487	0.528	0.597	0.652	0.623	0.623	0.662	0.682	0.661	0.632	0.694	0.250
5	0.403	0.493	0.535	0.538	0.505	0.573	0.618	0.592	0.681	0.606	0.638	0.570	0.200
6	0.372	0.457	0.480	0.485	0.529	0.518	0.545	0.577	0.605	0.631	0.599	0.601	0.167
7	0.349	0.382	0.460	0.469	0.475	0.504	0.522	0.539	0.528	0.568	0.588	0.562	0.143
8	0.302	0.390	0.453	0.440	0.473	0.510	0.505	0.517	0.541	0.530	0.534	0.549	0.125
9	0.296	0.347	0.370	0.427	0.477	0.439	0.485	0.492	0.506	0.530	0.557	0.532	0.111
10	0.254	0.337	0.373	0.405	0.413	0.427	0.455	0.487	0.480	0.460	0.443	0.463	0.100

TABLE VII: Text length vs. accuracy for different number of authors (profile length = 5,000)

C. Inter-Profile Dissimilarities

Besides the number of candidate authors and the length of the texts and profiles, the correct attribution of a text is also dependent on the similarity of the writing styles of the authors themselves. Indeed, repeated binary attributions between Henry James and Washington Irving with random generation of 100,000 word profiles yield a perfect accuracy of 1.0 on the classification of 400 texts of 10,000 words each. The same exercise when attributing between Grant Allen and Robert Louis Stevenson yields a classification rate of 0.91. This occurs because the stylometric fingerprints of Allen and Stevenson are harder to distinguish than those of James and Irving.

Dissimilarity of writing styles can be quantified by computing the relative entropies between the profiles [cf. (11)]. Since relative entropies are asymmetric, i.e., $H(P_1, P_2) \neq H(P_2, P_1)$ in general, we consider the average of the two relative entropies between two profiles as a measure of their dissimilarity. For each pair of authors, the relative entropy is computed based on the set of function words chosen adaptively to maximize the cross validation accuracy. For the 100,000 word profiles of James and Irving, the inter-profile dissimilarity resulting from the average of relative entropies is 0.184. The inter-profile dissimilarity between Allen and Stevenson is 0.099. This provides a formal measure of similarity of writing styles which explains the higher accuracy of attributions between James and Irving with respect to attributions between Allen and Stevenson.

The correlation between inter-profile dissimilarities and attribution accuracy is corroborated by Fig. 3. Each point in this plot corresponds to the selection of two authors at random from our pool of 21 authors from the 19th century. For each pair we select ten texts of 10,000 words each to generate profiles of length 100,000 words. We then attribute ten of the remaining excerpts of length 10,000 words of each of these two authors among the two profiles and record the correct attribution rate as well as the dissimilarity between the random profiles generated. The process is repeated twenty times for these two authors to produce the average dissimilarity and accuracy that yield the corresponding point in Fig. 3. E.g., consider two randomly chosen authors for which we have 50 excerpts of 10,000 word available. We select ten random texts to form a profile and attribute 20 out of the remaining 80 excerpts – 10 for each author. After repeating this procedure twenty times we get the average accuracy of attributing 400 texts of length 10,000 words between the two authors.

Besides the positive correlation between inter-profile dissimilarities and attribution accuracies, Fig. 3 shows that classification is perfect for 11 out of 12 instances where the inter-profile dissimilarity exceeds 0.16. Errors are rare for

	Stevenson	Alger	Melville	Allen	James	Alcott	Abbott	Austen	Garland	Hawthorne
Stevenson	3.3 / 93.8	10.0 / 0.0	**6.1 / 2.0**	6.7 / 0.5	10.9 / 0.0	10.7 / 0.2	10.0 / 0.5	12.7 / 0.0	**6.4 / 2.0**	7.9 / 1.0
Alger	10.0 / 0.80	4.2 / 96.6	10.9 / 0.5	11.6 / 0.0	10.4 / 0.5	11.8 / 0.0	13.4 / 0.0	11.6 / 0.2	**10.0 / 1.2**	11.0 / 0.2
Melville	**6.1 / 5.8**	10.9 / 0.5	3.4 / 76.0	**6.2 / 7.0**	10.8 / 1.0	12.7 / 1.4	8.6 / 2.5	11.7 / 0.8	7.1 / 1.0	7.2 / 4.0
Allen	**6.7 / 9.5**	11.6 / 0.5	**6.2 / 6.2**	3.8 / 65.0	11.4 / 1.2	13.0 / 0.2	**7.7 / 6.0**	11.1 / 0.8	**7.2 / 8.0**	7.9 / 2.4
James	10.9 / 0.0	**10.4 / 1.2**	10.8 / 0.0	11.4 / 0.5	3.8 / 96.2	14.0 / 0.0	13.8 / 0.0	9.7 / 0.0	11.8 / 0.8	**8.7 / 1.3**
Alcott	10.7 / 0.0	**11.8 / 0.9**	12.7 / 0.0	13.0 / 0.0	14.0 / 0.0	3.5 / 98.8	16.1 / 0.0	15.0 / 0.0	9.9 / 0.3	12.9 / 0.0
Abbott	10.0 / 1.8	13.4 / 1.8	8.6 / 0.8	7.7 / 1.5	13.8 / 0.5	16.1 / 5.0	2.7 / 88.2	11.6 / 0.0	9.7 / 0.2	9.1 / 0.2
Austen	12.7 / 0.0	11.6 / 0.0	11.7 / 0.0	11.1 / 0.0	9.7 / 0.0	15.0 / 0.0	**11.6 / 0.5**	3.5 / 99.5	12.7 / 0.0	9.6 / 0.0
Garland	6.4 / 1.2	10.0 / 0.0	**7.1 / 3.2**	**7.2 / 2.8**	11.8 / 0.0	9.9 / 0.5	9.7 / 0.0	12.8 / 0.0	3.6 / 91.8	8.9 / 0.5
Hawthorne	7.9 / 0.0	11.0 / 0.0	**7.2 / 1.2**	7.9 / 0.2	8.7 / 0.0	12.9 / 0.0	9.1 / 0.2	9.6 / 0.0	8.9 / 0.0	2.9 / 98.4

TABLE VIII: Confusion matrix for ten 19th century authors. The first value in each cell is the profile dissimilarity in cn between authors. The second value is the percentage of texts from the author in the row which are attributed to the author in the column. As marked in bold, high confusion rates are related to low profile dissimilarities.

profile dissimilarities between 0.10 and 0.16 since correct classifications average 0.984 and account for at least 0.96 of the attribution results in all but three outliers. For pairs of authors with dissimilarities smaller than 0.1 the average accuracy is 0.942.

To further emphasize the effect of inter-profile dissimilarities in the attribution accuracy, we present the confusion matrix for a 10-class classification among ten of the authors analyzed; see Table VIII. Inter-profile dissimilarities were computed based on the WANs that maximize the 10-class cross validation accuracy. Intra-profile dissimilarities, i.e. dissimilarities between an author and himself, were computed as the average relative entropy among various random partitions of his work into two pieces. Observe that intra-profile dissimilarities are markedly smaller than inter-profile dissimilarities, as expected. E.g., the intra-profile dissimilarity for Melville is $3.4cn$ whereas his average inter-profile dissimilarity with the remaining nine authors is $9.0cn$. In Table VIII we also inform the confusion rate in attribution, i.e., the percentage of an author's texts that are attributed to another author. Notice that, in general, higher confusion rates are associated with lower inter-profile dissimilarities. E.g., the most common mistake when attributing Hawthorne's texts is to assign them to Melville – 1.2% error – which coincides with being the author closest to him with an inter-profile dissimilarity of $7.2cn$.

VI. Meta attribution studies

WANs can also be used to study problems other than attribution between authors. In this section we demonstrate that WANs carry information about time periods, the genre of the composition, and the gender of the authors. We also illustrate the use of WANs in detecting collaborations.

A. Time

WANs carry information about the point in time in which a text was written. If we build random profiles of 200,000 words for Shakespeare, Chapman, and Melville and compute the inter-profile dissimilarity as in Section V-C, we obtain a dissimilarity of 0.04 between Shakespeare and Chapman and of 0.17 between Shakespeare and Melville. Since inter-profile dissimilarity is a measure of difference in style, these values are reasonable given that Shakespeare and Chapman were contemporaries but Melville lived more than two centuries after them.

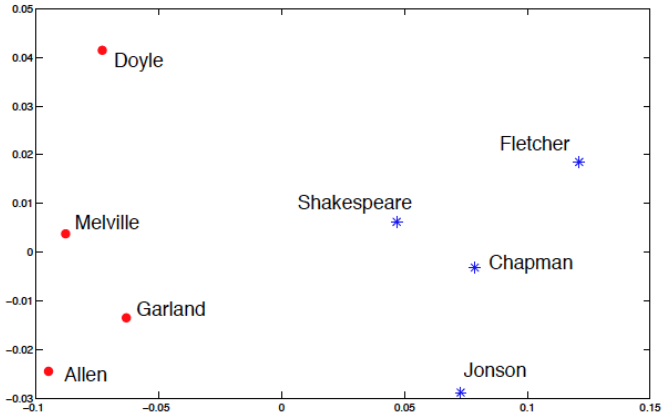
	Marlowe	Chapman
Shakespeare (Com.)	11.6	7.7
Shakespeare (His.)	7.6	9.3

TABLE IX: Inter-profile dissimilarities (x100) between authors of different genres.

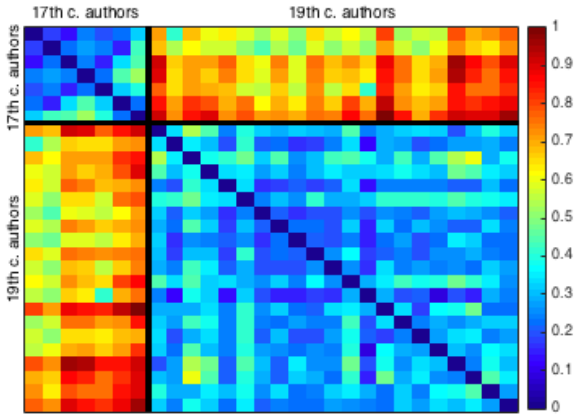
To further illustrate this point, in Fig. 4a we plot a two dimensional MDS representation of the dissimilarity between eight authors whose profiles were built with all their available texts in our corpus [36]. Four of the profiles correspond to early 17th century authors – Shakespeare, Chapman, Jonson, and Fletcher – and are represented by blue stars while the other four – Doyle, Melville, Garland, and Allen – correspond to 19th century authors and are represented by red dots. Notice that authors tend to have a smaller distance with their contemporaries and a larger distance with authors from other periods. This fact is also illustrated by the heat map of inter-profile dissimilarities in Fig. 4b where bluish colors represent smaller entropies. The first 7 rows and columns correspond to authors of the 17th century whereas the remaining 21 correspond to authors of the 19th century, where profiles were built with all the available texts. Notice that the blocks of blue color along the diagonal are in perfect correspondence with the time period of the authors, verifying that WANs can be used to determine the time in which a text was written. The average entropies among authors of the 17th century and among those of the 19th century are 0.096 and 0.098 respectively, whereas the average entropies between authors of different epochs is 0.273. I.e., the relative entropy between authors of different epochs almost triples that of authors belonging to the same time period.

B. Genre

Even though function words by themselves do not carry content, WANs constructed from a text contain, rather surprisingly, information about its genre. We illustrate this fact in Fig. 5, where we present the relative entropy between 20 pieces of texts written by Shakespeare of length 20,000 words, where 10 of them are history plays – e.g., *Richard II, King John, Henry VIII* – and 10 of them are comedies – e.g., *The Tempest, Measure for Measure, The Merchant of Venice*. As in Fig. 4b, bluish colors in Fig. 5 represent smaller relative entropies. Two blocks along the diagonal can be distinguished that coincide with the plays of the two different genres. Indeed,



(a) MDS plot for authors of different time periods.



(b) Heat map of inter-profile dissimilarities.

Fig. 4: (a) Authors from the early 17th century are depicted as blue stars while authors from the 19th century are depicted as red dots. Inter-profile dissimilarities are small within the groups and large between them. (b) High inter-profile dissimilarities are illustrated with warmer colors. Two groups of authors with small inter-profile dissimilarities are apparent: the first seven correspond to 17th century authors and the rest to 19th century authors.

Sh.	Jon.	Fle.	Mid.	Cha.	Marl.
19.1	20.0	**18.2**	20.2	19.5	20.9

TABLE X: Relative entropies from *Two Noble Kinsmen* to different profiles (x100).

if we sequentially extract one text from the group and attribute it to a genre by computing the average relative entropies to the remaining histories and comedies, the 20 pieces are correctly attributed to their genre.

More generally, inter-profile dissimilarities between authors that write in the same genre tend to be smaller than between authors that write in different genres. As an example, in Table IX we compute the dissimilarity between two Shakespeare profiles – one built with comedies and the other with histories – and two contemporary authors: Marlowe and Chapman. All profiles contain 100,000 words formed by randomly picking 10 extracts of 10,000 words. Marlowe never wrote a comedy and mainly focused on histories – *Edward II, The Massacre at Paris* – and tragedies – *The Jew of Malta, Dido* –, while the majority of Chapman's plays are comedies – *All Fools, May Day*. Genre choice impacts the inter-profile dissimilarity since the comedy profile of Shakespeare is closer to Chapman than to Marlowe and vice versa for the history profile of Shakespeare. The inter-profile dissimilarity between Shakespeare profiles is 6.2, which is still smaller than any dissimilarity in Table IX. This points towards the conclusion that the identity of the author is the main determinant of the writing style but that the genre of the text being written also contributes to the word choice. In general, two texts of the same author but different genres are more similar than two texts of the same genre but different authors which, in turn, are more similar than two texts of different authors and genres.

C. Gender

Word usage can be used for author profiling [46] and, in particular, to infer the gender of an author from the written

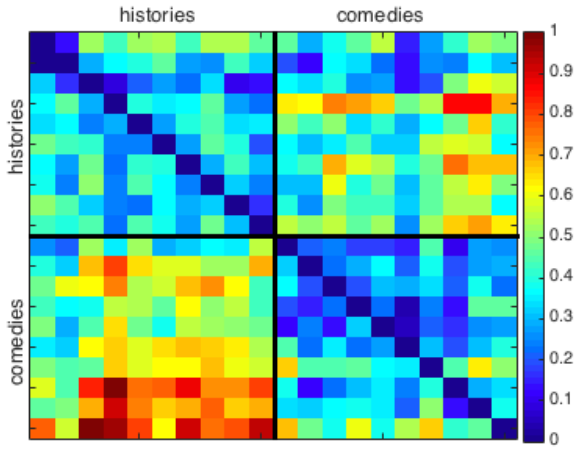


Fig. 5: Heat map of relative entropies between 20 Shakespeare extracts. The first 10 texts correspond to history plays while the last 10 correspond to comedy plays. Relative entropies within texts of the same genre are smaller than across genres.

	Sh.	Jon.	Fle.	Mid.	Cha.	Marl.
Sh.	19.1	19.2	**17.9**	19.0	19.1	19.3
Jon.	19.2	20.0	18.4	19.5	19.3	19.3
Fle.	**17.9**	18.4	18.2	18.4	18.2	18.1
Mid.	19.0	19.5	18.4	20.2	19.4	18.9
Cha.	19.1	19.3	18.2	19.4	19.5	19.4
Marl.	19.3	19.3	18.1	18.9	19.4	20.9

TABLE XI: Relative entropies from *Two Noble Kinsmen* to hybrid profiles composed of two authors (x100).

text. To illustrate this, we divide the 21 authors from the 19th century [36] into females – five of them – and males. We pick a gender at random and pick an excerpt of 10,000 words from any author of the selected gender. We then build two 100,000 words profiles, one containing pieces of texts written by male

Nr. of authors	N. Bayes	1-NN	3-NN	DT-gdi	DT-ce	SVM	WAN	Voting
2	2.6	3.5	5.2	12.2	12.2	2.7	**1.6**	0.9
4	6.0	9.2	12.4	25.3	25.5	6.8	**4.6**	3.3
6	8.1	11.7	15.2	31.9	32.2	7.9	**5.3**	3.8
8	9.6	15.4	19.2	36.4	37.2	11.1	**6.7**	5.2
10	10.8	16.7	21.4	42.1	42.1	11.5	**8.3**	6.0

TABLE XII: Error rates in % achieved by different methods for profiles of 100,000 words and texts of 10,000 words. The WANs achieve the smallest error rate among the methods considered separately. Voting decreases the error even further by combining the relational data of the WANs with the frequency data of other methods.

authors and the other by female authors. In order to avoid bias, we do not include any text of the author from which the text to attribute was chosen in the gender profiles. We then attribute the chosen text between the two gender profiles. After repeating this procedure 5,000 times, we obtain a mean accuracy of 0.63. Although this accuracy is lower than state-of-the-art gender profiling methods [47], the difference with random attribution, i.e. accuracy of 0.5, validates the fact that WANs carry gender information about the authors.

D. Collaborations

WANs can also be used for the attribution of texts written collaboratively between two or more authors. Since collaboration was a common practice for playwrights in the early 17th century, we consider the attribution of Early Modern English plays [36]. For a given play, we compute its relative entropy to six contemporary authors – Shakespeare, Jonson, Fletcher, Middleton, Chapman, and Marlowe – by generating 50 random profiles for each author of length 80,000 words and averaging the 50 entropies to obtain one representative number. We do not consider Peele in the analysis due to the short total length of available texts.

When two authors collaborate to write a play, the resulting word adjacency network is close to the profiles of both authors, even though these profiles are built with plays of their sole authorship. As an example, consider the play *Two Noble Kinsmen* which is an accepted collaboration between Fletcher and Shakespeare [48]. In Table X, we present the relative entropies between the play and the six analyzed authors. Notice that the two minimum entropies correspond to those who collaborated in writing it.

Collaboration can be further confirmed by the construction of hybrid profiles, i.e. profiles built containing 40,000 words of two different authors. Each entry in Table XI corresponds to the relative entropy from *Two Noble Kinsmen* to a hybrid profile composed by the authors in the row and column of that entry. Notice that the diagonal of Table XI corresponds to profiles of sole authors and, thus, coincides with Table X. The smallest relative entropy in Table XI is achieved by the hybrid profile composed by Fletcher and Shakespeare, which is consistent with the accepted attribution of the play.

VII. Comparison and Combination with Frequency Based Methods

Machine learning tools have been used to solve attribution problems by relying on the frequency of appearance of function words [49]. These methods consider the number of times an author uses different function words but, unlike WANs, do not contemplate the order in which the function words appear. The most common techniques include naive Bayes [50, Chapter 8], nearest neighbors (NN) [50, Chapter 2], decision trees (DT) [50, Chapter 14], and support vector machines (SVM) [50, Chapter 7].

In Table XII we inform the percentage of errors obtained by different methods when attributing texts of 10,000 words among profiles of 100,000 words for a number of authors ranging from two to ten. For a given number of candidate authors, we randomly pick them from the pool of 19th century authors [36] and attribute ten excerpts of each of them using the different methods. We then repeat the random choice of authors 100 times and average the error rate. For each of the methods based on function word frequencies, we pick the set of parameters and preprocessing that minimize the attribution error rate. E.g., for SVM the error is minimized when considering a polynomial kernel of degree 3 and normalizing the frequencies by text length. For the nearest neighbors method we consider two strategies based on one (1-NN) and three (3-NN) nearest neighbors as given by the l_2 metric in Euclidean space. Also, for decision trees we consider two types of split criteria: the Gini Diversity Index (DT-gdi) and the cross-entropy (DT-ce) [51].

The WANs achieve a lower attribution error than frequency based methods; see Table XII. For binary attributions, naive Bayes and SVM achieve error rates of 2.6% and 2.7% respectively and, thus, outperform nearest neighbors and decision trees. However, WANs outperform the aforementioned methods by obtaining an error rate of 1.6%. This implies a reduction of 38% in the error rate. For 6 authors, WANs achieve an error rate of 5.3% that outperform SVMs achieving 7.9% entailing a 33% reduction. This trend is consistent across different number of candidate authors, with WANs achieving an average error reduction of 29% compared with the best traditional machine learning method.

More important than the fact that WANs tend to outperform methods based on word frequencies, is the fact that they carry different stylometric information. Thus, we can combine both methodologies to further increase attribution accuracy. We do this via a voting method, where we perform majority voting between WANs and the two best performing frequency based methods, namely, naive Bayes and SVMs. More specifically, each of the three methods gives one vote to its preferred candidate author and then the voting method chooses the author with more votes. In case of a three-way tie, the candidate author

voted by the WANs is chosen. In the last column of Table XII we inform the error rate of majority voting. These error rates are consistently smaller than those achieved by WANs and, hence, by the other frequency based methods as well. E.g., for attributions among four authors, voting achieves an error of 3.3% compared to an error of 4.6% of WANs. This corresponds to a 28% reduction in error. Averaging among attributions for different number of candidate authors, majority voting entails a reduction of 30% compared with WANs. The combination of WANs and function word frequencies halves the attribution error rate with respect to the traditional machine learning methods.

Although the error rates presented in Table XII correspond to profiles of balanced length, the results also hold for scenarios where different profiles contain different number of words. This means that, for unbalanced scenarios, the WANs still outperform traditional classifiers and the voting method also achieves the lowest error rates.

VIII. Conclusions and Future Work

Relational data between function words was used as stylometric information to solve authorship attribution problems. Normalized word adjacency networks (WANs) were used as relational structures. We interpreted these networks as Markov chains in order to facilitate their comparison using relative entropies. The accuracy of WANs was analyzed for varying number of candidate authors, text lengths, profile lengths and different levels of heterogeneity among the candidate authors, regarding genre, gender, and time period. The method works best when the corpora of known texts is of substantial length, when the texts being attributed are long, or when the number of candidate authors is small. If long profiles are available – more than 60,000 words, corresponding to 150 pages of a midsize paperback book –, we demonstrated very high attribution accuracy for texts longer than a few typical novel chapters even when attributing between a large number of authors, high accuracy for texts as long as a play act or a novel chapter, and reasonable rates for short texts such as newspaper opinion pieces if the number of candidate authors is small. WANs were also shown to classify accurately the time period when a text was written, to acceptably estimate the genre of a piece, and to have some predictive power on the gender of the author. The applicability of WANs to identify multiple authors in collaborative works was also demonstrated. With regards to existing methods based on the frequency with which different function words appear in the text, we observed that WANs exceed their classification accuracy. More importantly, we showed that WANs and frequencies captured different stylometric aspects so that their combination is possible and ends up halving the error rate of existing methods.

As directions for future research, we plan to investigate more sophisticated ways to combine the attribution power of WANs with existing methods in order to improve the attribution accuracy in general and to gain discriminating power for short texts. Moreover, we want to extend our method towards the attribution of other types of creative work such as the identification of a composer based on a musical piece.

References

[1] S. Segarra, M. Eisen, and A. Ribeiro, "Authorship attribution using function words adjacency networks," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, 2013, pp. 5563–5567.

[2] T. Grant, "Quantifying evidence in forensic authorship analysis," *International Journal of Speech Language and the Law*, vol. 14, no. 1, 2007.

[3] A. Abbasi and H. Chen, "Applying authorship analysis to extremist-group web forum messages," *Intelligent Systems, IEEE*, vol. 20, no. 5, pp. 67–75, Sept 2005.

[4] S. Meyer zu Eissen, B. Stein, and M. Kulig, "Plagiarism detection without reference collections," in *Advances in Data Analysis*, ser. Studies in Classification, Data Analysis, and Knowledge Organization, R. Decker and H.-J. Lenz, Eds. Springer Berlin Heidelberg, 2007, pp. 359–366.

[5] D. I. Holmes, "Authorship attribution," *Computers and the Humanities*, vol. 28, no. 2, pp. 87–106, 1994.

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

[7] 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.

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

[9] 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.

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

[11] 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.

[12] 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.

[13] D. L. Hoover, "Delta prime?" *Literary and Linguistic Computing*, vol. 19, no. 4, pp. 477–495, 2004.

[14] H. van Halteren, R. H. Baayen, F. Tweedie, M. Haverkort, and A. Neijt, "New machine learning methods demonstrate the existence of a human stylome," *Journal of Quantitative Linguistics*, vol. 12, no. 1, pp. 65–77, 2005.

[15] B. Yu, "Function words for chinese authorship attribution," in *Proceedings of the NAACL-HLT 2012 Workshop on Computational Linguistics for Literature*. Association for Computational Linguistics, 2012, pp. 45–53.

[16] M. Ebrahimpour, T. J. Putniņš, M. J. Berryman, A. Allison, B. W.-H. Ng, and D. Abbott, "Automated authorship attribution using advanced signal classification techniques," *PloS one*, vol. 8, no. 2, p. e54998, 2013.

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

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

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

[20] 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.

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

[22] 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.

[23] 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.

[24] T. Solorio, S. Pillay, S. Raghavan, and M. Montes-y Gómez, "Modality specific meta features for authorship attribution in web forum posts." in *IJCNLP*, 2011, pp. 156–164.

[25] G. Sidorov, F. Velasquez, E. Stamatatos, A. Gelbukh, and L. Chanona-Hernández, "Syntactic n-grams as machine learning features for natural language processing," *Expert Systems with Applications*, vol. 41, no. 3, pp. 853–860, 2014.

[26] Y. Seroussi, I. Zukerman, and F. Bohnert, "Authorship attribution with latent dirichlet allocation," in *Proceedings of the fifteenth conference on computational natural language learning*. Association for Computational Linguistics, 2011, pp. 181–189.

[27] L. Pearl and M. Steyvers, "Detecting authorship deception: a supervised machine learning approach using author writeprints," *Literary and linguistic computing*, p. fqs003, 2012.

[28] J. Savoy, "Authorship attribution based on a probabilistic topic model," *Information Processing & Management*, vol. 49, no. 1, pp. 341–354, 2013.

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

[30] 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.

[31] 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*. Association for Computational Linguistics, 2006, pp. 482–491.

[32] M. J. Collins, "A new statistical parser based on bigram lexical dependencies," in *Proceedings of the 34th annual meeting on Association for Computational Linguistics*. Association for Computational Linguistics, 1996, pp. 184–191.

[33] H. Lodhi, C. Saunders, J. Shawe-Taylor, N. Cristianini, and C. Watkins, "Text classification using string kernels," *The Journal of Machine Learning Research*, vol. 2, pp. 419–444, 2002.

[34] C. Cortes, P. Haffner, and M. Mohri, "A machine learning framework for spoken-dialog classification," in *Springer Handbook of Speech Processing*. Springer, 2008, pp. 585–596.

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

[36] S. Segarra, M. Eisen, and A. Ribeiro, "Compilation of texts used for the numerical experiments (journal materials)," *https://fling.seas.upenn.edu/~maeisen/wiki/index.php?n=Main.TextAttribution2*, 2014.

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

[38] Z. Rached, F. Alajaji, and L. Campbell, "The kullback-leibler divergence rate between markov sources," *Information Theory, IEEE Transactions on*, vol. 50, no. 5, pp. 917–921, May 2004.

[39] H. Tang, M. Hasegawa-Johnson, and T. Huang, "Non-frontal view facial expression recognition based on ergodic hidden markov model supervectors," in *Multimedia and Expo (ICME), 2010 IEEE International Conference on*, July 2010, pp. 1202–1207.

[40] M. Vidyasagar, S. Mande, C. Reddy, and V. Rao, "The 4m (mixed memory markov model) algorithm for finding genes in prokaryotic genomes," *Automatic Control, IEEE Transactions on*, vol. 53, no. Special Issue, pp. 26–37, Jan 2008.

[41] Y. Zhao, J. Zobel, and P. Vines, "Using relative entropy for authorship attribution," in *Information Retrieval Technology*, ser. Lecture Notes in Computer Science, H. Ng, M.-K. Leong, M.-Y. Kan, and D. Ji, Eds. Springer Berlin Heidelberg, 2006, vol. 4182, pp. 92–105.

[42] R. Arun, V. Suresh, and C. V. Madhavan, "Stopword graphs and authorship attribution in text corpora," in *Semantic Computing, 2009. ICSC'09. IEEE International Conference on*. IEEE, 2009, pp. 192–196.

[43] C. D. Manning, P. Raghavan, and H. Schütze, *Introduction to Information Retrieval*. New York, NY, USA: Cambridge University Press, 2008.

[44] T. W. Anderson and L. A. Goodman, "Statistical inference about markov chains," *The Annals of Mathematical Statistics*, pp. 89–110, 1957.

[45] M. A. A. Cox and T. F. Cox, "Multidimensional scaling," in *Handbook of Data Visualization*, ser. Springer Handbooks Comp.Statistics. Springer Berlin Heidelberg, 2008, pp. 315–347.

[46] S. Argamon, M. Koppel, J. W. Pennebaker, and J. Schler, "Automatically profiling the author of an anonymous text," *Commun. ACM*, vol. 52, no. 2, pp. 119–123, Feb. 2009.

[47] M. Koppel, S. Argamon, and A. R. Shimoni, "Automatically categorizing written texts by author gender," *Literary and Linguistic Computing*, vol. 17, no. 4, pp. 401–412, 2002.

[48] E. A. B. Farmer and Z. Lesser, "Deep: Database of Early English Playbooks," *http://deep.sas.upenn.edu/*, 2007. [Online]. Available: http://deep.sas.upenn.edu/

[49] Y. Zhao and J. Zobel, "Effective and scalable authorship attribution using function words," in *Information Retrieval Technology*, ser. Lecture Notes in Computer Science, G. Lee, A. Yamada, H. Meng, and S. Myaeng, Eds. Springer Berlin Heidelberg, 2005, vol. 3689, pp. 174–189.

[50] C. M. Bishop, *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Secaucus, NJ, USA: Springer-Verlag New York, Inc., 2006.

[51] L. Breiman, J. Friedman, C. J. Stone, and R. A. Olshen, *Classification and regression trees*. CRC press, 1984.