WORD ADJACENCY NETWORKS FOR AUTHORSHIP ATTRIBUTION: SOLVING SHAKESPEAREAN CONTROVERSIES

Santiago Segarra

A THESIS

in

Electrical Engineering

Presented to the Faculties of the University of Pennsylvania in Partial Fulfillment of the Requirements for the Degree of Master of Science in Engineering

2014

Dr. Alejandro Ribeiro - Supervisor of Thesis

Dr. Saswati Sarkar - Graduate Group Chairperson

Contents

Acknowledgements

I would like to express my sincere gratitude to my advisor Dr. Alejandro Ribeiro for his useful comments, thoughtful remarks, motivating discussions and constant support. Moreover, I would like to thank Dr. Gabriel Egan, Professor of Shakespeare Studies at De Montfort University. His vast literature knowledge and kind disposition were fundamental for the development of the applied portion of this thesis. Finally, I thank my colleague Mark Eisen for his invaluable contributions towards the implementation of the method here described. His help in generating networks from texts and running multiple experiments was essential in the construction of the present thesis.

Abstract

A method for identifying the author of a written text by using function word adjacency networks (WANs) is developed. Function words express grammatical relationships between other words but do not carry meaning on their own. In the WANs, nodes are function words and directed edges represent the likelihood of directed co appearance between two of these words. For each candidate author, a profile WAN is constructed against which the WAN of a text to attribute is compared. By reinterpreting WANs as Markov chains where the state space corresponds to the set of function words, they can be compared in terms of 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 authorial style. Attribution accuracy is observed to exceed the one achieved by methods that rely on word frequencies alone. Furthermore, combining WANs with methods that rely on word frequencies, results in larger attribution accuracy, indicating that both sources of information encode different aspects of stylometric fingerprints. In the second part of this thesis, we use WANs to analyze authorial controversies among Shakespeare and his contemporaries. After validating the accuracy of WANs in this dataset, we test existing hypothesis about authors of anonymous texts and the partition among authors of collaborative plays. Finally, the impact of genre on attribution accuracy is examined revealing that the genre of a play partially conditions the choice of the function words used in it.

Chapter 1

Introduction

With the Sumerian invention of writing in ancient Mesopotamia over 5,000 years ago, mankind achieved partial independence from oral communication. This implied that words could be separated from their author in both space and time. This paradigm-shifting event was key for the further development of human beings, allowing the spread of ideas, the organization of society, the documentation of events and more. Furthermore, with written language a new problem appeared: authorship attribution.

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 [1], defense intelligence [2] and plagiarism [3]. 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 [4–6].

Authorship attribution dates at least to more than a century ago with a work that proposed distinguishing authors by looking at word lengths [7]. This was later improved by [8] 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 [9] 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 [9], 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., [10–13]. Other attribution methods include the stylometric techniques in [14], the use of vocabulary richness as a stylometric marker [15–17] – see also [18] for a critique –, the use of stable words defined as those that can be replaced by an equivalent [19], and syntactical markers such as taggers of parts of speech [20]. Markov chains have also been used as a tool for authorship attribution [21,22]. 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.

Our method for attributing texts also measures function word usage to distinguish au-

thor styles. Rather than only considering word frequencies, however, we consider a more complex relational structure between an author's usage of function words. The rest of the thesis is organized as follows. We first provide a formal description of the problem of authorship attribution and its variants (Chapter 2). The thesis is then divided into two parts: the description of a method to solve authorship attribution problems (Chapters 3 to 7) and the application of this method to attribute Early Modern English plays (Chapters 8 to 12).

In the first part of the thesis, we begin by describing word adjacency networks (WANs), which are directed networks that store information of co-appearance of two function words in the same sentence. 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. 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 (Chapter 3). 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 (Chapter 4). 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 texts of known and unknown authorship and the similarity of styles between candidate authors (Chapter 5). 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. We also show that WANs can be used to detect collaboration between several authors (Chapter 6). We further demonstrate that our classifier performs better than techniques based on function word frequencies alone. 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 (Chapter 7). The results in this first part were published in [23, 24].

In the second part of the thesis, we begin by discussing the main playwrights analyzed in our work and the construction of their profile networks as well as a measure of similarity between profiles (Chapter 8). We then perform a stylometric analysis of the complete canons of our five primary playwrights, followed by a summary of results that shows the value of WANs to classify texts from that period (Chapter 9). An analysis of a set of plays published anonymously or without a clear author is performed (Chapter 10). We then examine the use of WANs in determining authorship of plays known to be written by multiple authors in collaboration (Chapter 11). This is first done by analyzing entire plays and then through extensive interplay analysis of a set of particularly controversial plays. Our results largely corroborate existing theories regarding these plays as well as, in some cases, propose new authorship breakdowns. We conclude by providing a brief analysis of the use of WANs in distinguishing between the three most common dramatic genres of the era: comedy, tragedy, and history (Chapter 12). The results in this second part were published in [25].

Chapter 2

The Authorship Attribution Problem

Authorship attribution can be loosely described as the endeavor to extract the characteristics of an author of a text. Although, this broad definition allows various interpretations, three main problems in authorship attribution have been pointed out in the literature [5]. The first problem type, called *closed class*, is where you are given a text which was written by someone within a pool of candidate authors and you want to determine who the rightful author is. The second problem type, called *open class*, is similar to the previous one but there exists the possibility that the text does not belong to any of the candidate authors. The third type, sometimes called *profiling*, deals with extracting extra information about the author apart from his identity, e.g. gender or nationality. In the present work, we focus our interest mainly in the closed class problem and present some results on author profiling as well.

The closed class authorship attribution problem can be formally defined as follows. 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.

$$T^{(i)} = \{t \mid t \in T, r_T(t) = a_i\}. \tag{2.1}$$

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 Chapter 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. Define the classification accuracy 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

$$\rho(\hat{r}_U) = \frac{1}{k} \sum_{u \in U} \mathbb{I}\left\{\hat{r}_U(u) = r_U(u)\right\}. \tag{2.2}$$

Throughout this work, we use $\rho(\hat{r}_U)$ to gauge performance when classifying a set of texts.

Notice that the described framework can be modified to accommodate the description of the open class problem by adding an additional author a_0 to the set A representing the option of any author other than a_1 through a_n. For this to be consistent, we should have $|T^{(0)}| = 0$, i.e. we have no sample text about this artificial author. Accuracy of the classifier

built can still be computed as in (2.2).

Author profiling can also be thought in terms of the closed problem formulation where the set A contains author profiles instead of individual authors. For example, if we are trying to infer the gender of an author, there are two possible profiles – male a_1 and female a_2 – and the set T contains known texts written by authors of both genders and our task is to decide if an unknown text $u \in U$ was written by a man a_1 or a woman a_2.

Part I

Description and Analysis of WANs

Chapter 3

Function Word Adjacency Networks

In order to solve the authorship attribution problems formalized in the previous chapter, we construct word adjacency networks (WANs) 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 $0 < i(\omega_2) - i(\omega_1) \leq D$ is

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 3.1: Most common function words in analyzed texts.

defined as

$$d(\omega_1, \omega_2) := \alpha^{i(\omega_2)-i(\omega_1)-1}. \tag{3.1}$$

In every sentence there are two kind of words: function and non-function words [26]. While in (3.1) the words w_1 and w_2 need not be function words, in this work 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 3.1. 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 $D = 4$.

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. 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 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 Chapter 4.

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

$$Q_t(f_i, f_j) = \sum_{h,e} \mathbb{I}\{s_t^h(e) = f_i\} \sum_{d=1}^{D} \alpha^{d-1} \, \mathbb{I}\{s_t^h(e+d) = f_j\}, \tag{3.2}$$

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 (3.2) is the sum of the directed proximities from f_i to f_j defined in (3.1) 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 (3.2) 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

$$Q_t = \begin{array}{c} \\ \text{a} \\ \text{in} \\ \text{of} \\ \text{but} \end{array} \begin{array}{c} \begin{array}{cccc} \text{a} & \text{in} & \text{of} & \text{but} \end{array} \\ \begin{pmatrix} 0 & 3 \times 0.8^1 & 0.8^1 & 0 \\ 2 \times 0.8^3 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 1 & 0.8^2 & 0 & 0 \end{pmatrix} \end{array}. \tag{3.3}$$

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

$$Q_c = \sum_{t \in T^{(c)}} Q_t. \tag{3.4}$$

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

$$\hat{Q}_c(f_i, f_j) = \frac{Q_c(f_i, f_j)}{\sum_k Q_c(f_i, f_k)}, \tag{3.5}$$

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 (3.5) 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 (3.5) to the similarity function in Example 2, we obtain the following normalized similarity matrix

$$\hat{Q}_t = \begin{array}{c} \\ \text{a} \\ \text{in} \\ \text{of} \\ \text{but} \end{array} \begin{array}{c} \begin{array}{cccc} \text{a} & \text{in} & \text{of} & \text{but} \end{array} \\ \begin{pmatrix} 0 & 0.75 & 0.25 & 0 \\ 1 & 0 & 0 & 0 \\ 0.25 & 0.25 & 0.25 & 0.25 \\ 0.61 & 0.39 & 0 & 0 \end{pmatrix} \end{array}. \tag{3.6}$$

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.

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

$$H(P_1, P_2) = \sum_{i,j} \pi(f_i) P_1(f_i, f_j) \log \frac{P_1(f_i, f_j)}{P_2(f_i, f_j)}, \tag{3.7}$$

where π is the limiting distribution of 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 [27]. Relative entropy (3.7), also called Kullback-Leibler divergence rate [28], is a common dissimilarity measure among Markov chains and is used in a variety of applications such as face recognition [29] and gene analysis [30]. Notice that the limit distribution π in (3.7) 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 [31]. This is unrelated to their use here as measures of the relational information captured in function WANs.

Using (3.7), we generate the attribution function $\hat{r}_U(u)$ by assigning the text u to the

author with the most similar relational structure

$$\hat{r}_U(u) = a_p, \text{ where } p = \underset{c}{\operatorname{argmin}}\, H(P_u, \hat{P}_c). \tag{3.8}$$

Notice that the relative entropy in (3.8) takes an infinite value when any word transition that appears in the unknown text does not appear in the profile. In practice we compute the relative entropy in (3.7) by summing only over the non-zero transitions in the profiles,

$$H(P_1, P_2) = \sum_{i,j|P_2(f_i,f_j)\neq 0} \pi(f_i) P_1(f_i, f_j) \log \frac{P_1(f_i, f_j)}{P_2(f_i, f_j)}. \tag{3.9}$$

Because the calculation of relative entropy in (3.9) only adds relative entropy for nonzero transitions, profiles built from fewer total words will on average contain less nonzero transitions and will thus sum together fewer terms than larger profiles. When attributing an unknown text among profiles of varying size, we avoid this potential biasing for smaller profiles by summing only over transitions that are nonzero in every profile being considered,

$$H(P_1, P_2) = \sum_{\substack{i,j|P_c(f_i,f_j)\neq 0 \\ \text{for all } a_c\in A}} \pi(f_i) P_1(f_i, f_j) \log \frac{P_1(f_i, f_j)}{P_2(f_i, f_j)}. \tag{3.10}$$

We proceed to specify the selection of function words in F as well as the choice of the parameters α and D after the following remark.

Remark 1 For the relative entropies in (3.8) 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 (3.7), (3.9) and (3.10) exist. This is true if the texts that generated P_u are sufficiently long. If this is not true for a particular network, 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 given by the word frequencies in the text.

Chapter 4

Function Words and WAN Parameters

The classification accuracy of the function WANs introduced in Chapter 3 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 [26]. 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

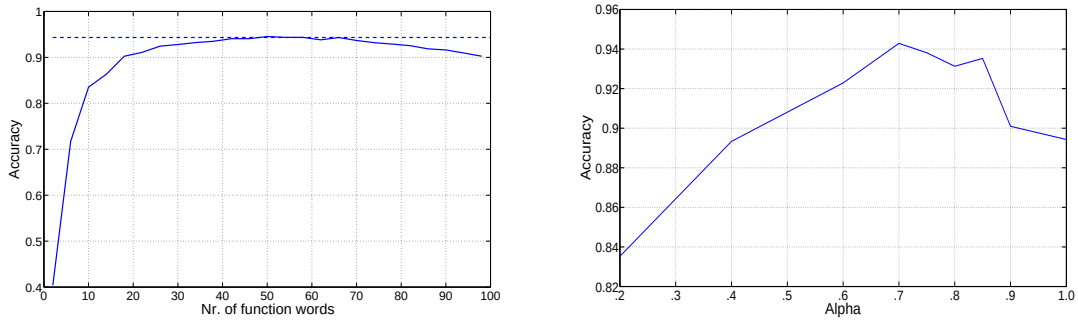
higher computational complexity without increase in accuracy since grammatical relations of words more than 10 positions apart are rare.

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 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. As an illustration, we consider in Fig. 4.1a the attribution of 1,000 texts of length 10,000 words among 7 authors chosen at random from our pool of 19th century authors [32] for a fixed value of $\alpha = 0.75$ and profiles of 100,000 words – see also Chapter 5 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 method is not sensitive to the choice of n. 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 of n 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. 4.1a 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 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



(a) (b)

Figure 4.1: 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) Attribution accuracy as a function of the network size. 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) Attribution accuracy as a function of the discount factor α. Accuracy is maximized for values of the discount factor α in the range between 0.70 and 0.85.

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. 4.1b 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 [32] using profiles with 100,000 words. 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 $\alpha = 0.7$ 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. 4.1b 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.

Chapter 5

Attribution Accuracy

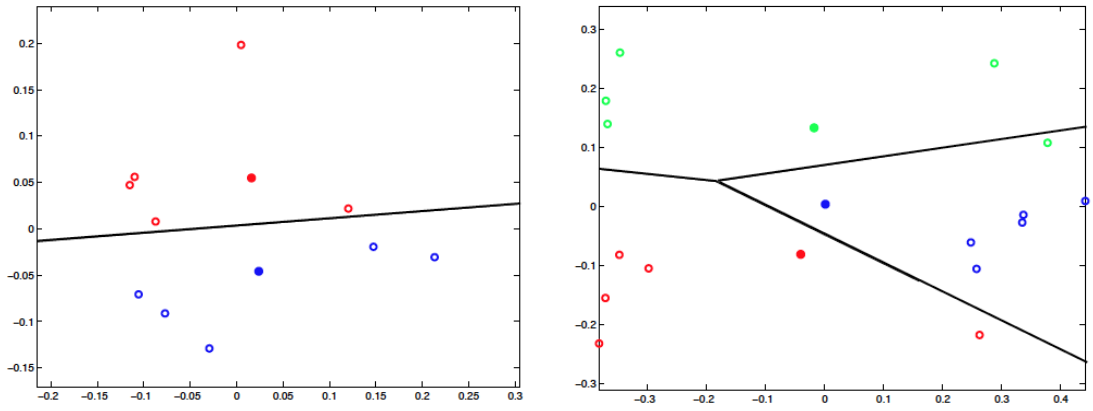
Henceforth, we fix the WAN generation parameters to the optimal values found in Chapter 4, i.e., the set of sentence delimiters is $\{ \ . \ ? \ ! \ ; \ \}$, the discount factor is $\alpha = 0.75$, and the window length is $D = 10$. The set of function words F is picked adaptively for every attribution problem by performing $M = 10$ cross validation rounds.

The text corpus used for the simulations consists of authors from two different periods [32]. 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 corre-

sponds 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 [32]. 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 midsize edition. The five known texts from each author are used to generate corresponding profiles as described in Chapter 3. Relative entropies in (3.9) 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) [33] 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. 5.1a. Twain's and Melville's profiles are depicted as red and blue filled circles, respectively. Unknown texts are depicted as empty circles, where the color indicates the real author, i.e. red for Twain and blue for



(a) MDS representation for two authors. (b) MDS representation for three authors.

Figure 5.1: (a) Perfect accuracy is attained for two candidate authors. Every empty circle falls in the half plane corresponding to the filled circle of their color. (b) One mistake is made for three authors. One green empty circle falls in the region attributable to the blue author.

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. 5.1a, 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 circles fall in the half plane closer to the filled blue circle. 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. 5.1a 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. 5.1b where filled circles represent profiles and empty circles represent texts. Different colors are used to distinguish Twain (red), Melville (blue), and Austen (green). 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. 5.1a, attribution is not perfect since one of Austen's texts is mistakenly attributed to Melville. This is represented in Fig 5.1b by the green empty circle that appears in the section of the Voronoi tessellation that corresponds to the blue 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, the length of the texts of unknown authorship, and the similarity of writing styles as captured by the relative entropy dissimilarities between profiles. We study these effects in sections 5.1,5.2, and 5.3, respectively.

5.1 Varying Profile Length

The profile length is defined as the total number of words, function or otherwise, used to construct the profile. To study the effect of varying profile lengths we fix $\alpha = 0.75$, $D = 10$,

Table 5.1: Profile length vs. accuracy for different number of authors (text length = 25,000)

Nr. auth.	Number of words in profile (thousands)										Rand.
	10	**20**	**30**	**40**	**50**	**60**	**70**	**80**	**90**	**100**	
2	.927	.964	.984	.985	.981	.979	.981	.986	.992	.988	.500
3	.871	.934	.949	.962	.968	.975	.982	.978	.974	.978	.333
4	.833	.905	.931	.949	.948	.964	.963	.968	.969	.977	.250
5	.800	.887	.923	.950	.945	.951	.953	.961	.961	.969	.200
6	.760	.880	.929	.932	.937	.941	.948	.952	.950	.973	.167
7	.755	.851	.909	.924	.937	.943	.937	.957	.960	.957	.143
8	.722	.841	.898	.911	.932	.941	.938	.947	.952	.955	.125
9	.683	.855	.882	.905	.915	.931	.932	.944	.952	.955	.111
10	.701	.827	.882	.910	.923	.923	.934	.935	.943	.935	.100

Table 5.2: Profile length vs. accuracy for different number of authors (text length = 5,000)

Nr. auth.	Number of words in profile (thousands)										Rand.
	10	**20**	**30**	**40**	**50**	**60**	**70**	**80**	**90**	**100**	
2	.863	.930	.932	.945	.928	.952	.942	.907	.942	.967	.500
3	.821	.884	.886	.890	.910	.901	.943	.912	.911	.914	.333
4	.728	.833	.849	.862	.892	.867	.888	.905	.882	.885	.250
5	.698	.819	.825	.839	.862	.884	.859	.865	.882	.893	.200
6	.673	.754	.789	.798	.832	.837	.863	.870	.896	.878	.167
7	.616	.754	.806	.838	.812	.848	.859	.832	.873	.868	.143
8	.600	.720	.748	.820	.805	.831	.831	.854	.857	.850	.125
9	.587	.718	.767	.781	.796	.809	.833	.850	.843	.847	.111
10	.556	.693	.737	.753	.805	.827	.829	.824	.843	.846	.100

Table 5.3: Profile length vs. accuracy for different number of authors (text length = 1,000)

Nr. auth.	Number of words in profile (thousands)										Rand.
	10	**20**	**30**	**40**	**50**	**60**	**70**	**80**	**90**	**100**	
2	.738	.788	.747	.823	.803	.803	.802	.800	.812	.793	.500
3	.599	.698	.690	.737	.713	.744	.724	.726	.757	.698	.333
4	.528	.638	.640	.672	.658	.663	.656	.663	.651	.707	.250
5	.491	.561	.598	.627	.686	.621	.633	.661	.674	.632	.200
6	.469	.549	.578	.593	.626	.594	.598	.617	.606	.582	.167
7	.420	.469	.539	.551	.583	.564	.603	.593	.583	.598	.143
8	.392	.454	.544	.540	.572	.551	.583	.589	.563	.599	.125
9	.385	.449	.489	.528	.519	.556	.551	.580	.560	.576	.111
10	.353	.410	.466	.480	.506	.536	.529	.542	.556	.553	.100

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, and 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 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 [32] 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 5.1 - 5.3 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 5.1. 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 Chapter 7. 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. Observe that accuracy does not monotonically decrease when increasing the candidate authors due to the noise introduced by the random selection of authors and texts.

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 5.2 for the same profile lengths and number of candidate authors considered in Table 5.1. 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 5.3, 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 re-

Table 5.4: Text length vs accuracy for different number of authors (profile length = 100,000)

Nr. auth.	Number of words in texts (thousands)												Rand.
	1	**2**	**3**	**4**	**5**	**6**	**8**	**10**	**15**	**20**	**25**	**30**	
2	.840	.917	.925	.938	.940	.967	.958	.977	.967	.989	.988	.986	.500
3	.789	.873	.890	.919	.913	.932	.936	.956	.952	.979	.979	.975	.333
4	.736	.842	.870	.902	.906	.933	.937	.952	.965	.970	.973	.974	.250
5	.711	.797	.858	.874	.891	.906	.924	.925	.955	.971	.980	.964	.200
6	.690	.796	.828	.886	.884	.911	.919	.922	.944	.957	.969	.961	.167
7	.633	.730	.814	.855	.874	.890	.910	.911	.928	.947	.956	.951	.143
8	.602	.740	.811	.846	.882	.887	.915	.910	.930	.944	.957	.963	.125
9	.607	.721	.774	.826	.845	.870	.889	.890	.918	.948	.951	.953	.111
10	.578	.731	.792	.816	.842	.855	.872	.893	.921	.933	.942	.961	.100

Table 5.5: Text length vs. accuracy for different number of authors (profile length = 20,000)

Nr. auth.	Number of words in texts (thousands)												Rand.
	1	**2**	**3**	**4**	**5**	**6**	**8**	**10**	**15**	**20**	**25**	**30**	
2	.812	.850	.903	.912	.913	.912	.938	.945	.918	.964	.964	.969	.500
3	.760	.797	.858	.899	.887	.918	.920	.918	.919	.938	.929	.928	.333
4	.670	.747	.813	.852	.868	.887	.889	.906	.918	.915	.900	.913	.250
5	.621	.721	.749	.813	.823	.819	.859	.878	.876	.887	.889	.893	.200
6	.557	.681	.754	.782	.799	.831	.852	.866	.871	.879	.881	.872	.167
7	.493	.610	.674	.706	.731	.770	.798	.807	.828	.862	.867	.858	.143
8	.467	.623	.675	.721	.741	.769	.790	.826	.822	.857	.841	.857	.125
9	.474	.574	.656	.672	.710	.734	.781	.783	.813	.845	.837	.841	.111
10	.433	.535	.612	.663	.684	.706	.752	.772	.836	.840	.851	.848	.100

sults can be of use as circumstantial evidence in support of attribution claims substantiated by further proof.

5.2 Varying Text Length

In this section we analyze the effect of text length in attribution accuracy for varying profile lengths 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 text lengths 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. The

Table 5.6: Text length vs. accuracy for different number of authors (profile length = 5,000)

Nr. auth.	Number of words in texts (thousands)												Rand.
	1	**2**	**3**	**4**	**5**	**6**	**8**	**10**	**15**	**20**	**25**	**30**	
2	.672	.740	.747	.707	.803	.823	.788	.848	.820	.802	.827	.832	.500
3	.547	.623	.626	.653	.744	.669	.712	.757	.736	.764	.734	.733	.333
4	.452	.487	.528	.597	.652	.623	.623	.662	.682	.661	.632	.694	.250
5	.403	.510	.535	.538	.505	.573	.618	.592	.681	.606	.638	.570	.200
6	.372	.457	.480	.485	.529	.518	.545	.577	.605	.631	.599	.601	.167
7	.349	.382	.460	.469	.475	.504	.522	.539	.528	.568	.588	.562	.143
8	.302	.390	.453	.440	.473	.510	.465	.517	.541	.530	.534	.549	.125
9	.296	.347	.370	.427	.477	.439	.485	.492	.506	.530	.557	.532	.111
10	.254	.337	.373	.405	.413	.427	.455	.487	.480	.460	.443	.463	.100

fine resolution for short texts permits estimating the shortest texts that can be attributed accurately. As in Section 5.1, for every combination of number of authors and text length, enough independent attribution experiments were performed to ensure that every accuracy value in tables 5.4 - 5.6 is based on at least 600 attributions.

For profiles of length 100,000 words, the results are reported in Table 5.4. As done in tables 5.1-5.3, 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.

If we reduce the length of the profiles to 20,000 words, reasonable accuracies are attained for small pools of candidate authors; see Table 5.5. 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 5.6. For shorter texts or larger number of candidate authors, WANs can provide supporting evidence but not definitive proof.

In Sections 5.1 and 5.2, 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

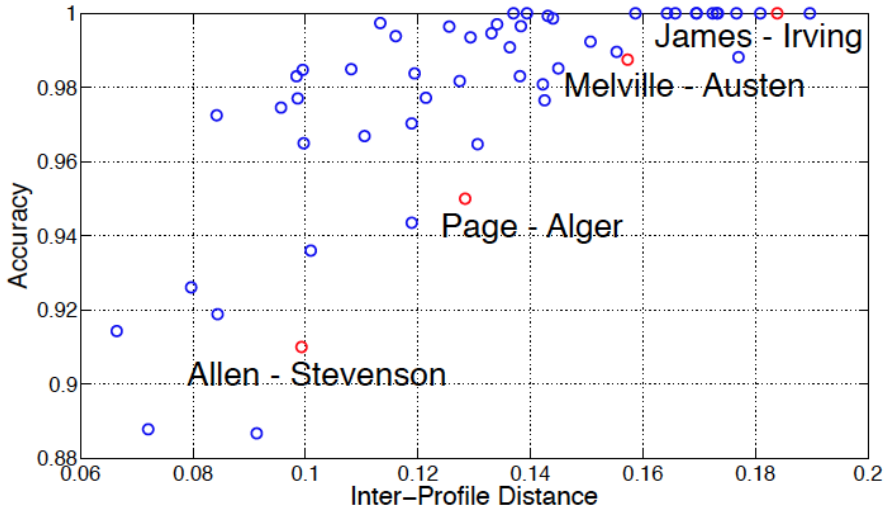


Figure 5.2: Binary attribution accuracy as a function of the inter-profile dissimilarity. Higher accuracy is attained for attribution between authors which are more dissimilar.

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.

5.3 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. (3.9)]. 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. 5.2. 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. 5.2. 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. 5.2 shows that classification is perfect for 11 out of 12 instances where the inter-profile dissimilarity exceeds 0.16. Errors are rare for 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.

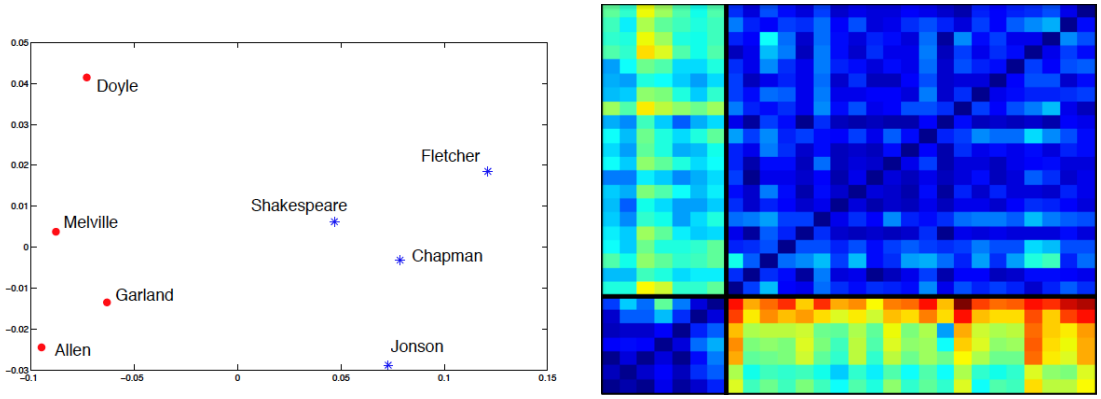
Chapter 6

Meta Attribution Analysis

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.

6.1 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 5.3, 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, this values are reasonable given that Shakespeare and Chapman were contemporaries but Melville lived more than two centuries



(a) (b)

Figure 6.1: (a) MDS plot for authors of different time periods. 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) Heat map of inter-profile relative entropies. High inter-profile relative entropies are illustrated with warmer colors. Two groups of authors with small inter-profile relative entropies are apparent: the first seven correspond to 17th century authors and the rest to 19th century authors.

after them.

To further illustrate this point, in Fig. 6.1a 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 [32]. 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 relative entropies in Fig. 6.1b where bluish colors represent smaller entropies. Since heat maps allow the representation of asymmetric data, we directly plot the relative entropies instead of the symmetrized inter-profile dissimilarities. The first 7 rows and columns correspond to authors of the 17th century whereas

Table 6.1: Inter-profile dissimilarities (x100) between authors of different genres.

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

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.

6.2 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. 6.2, 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. 6.1b, bluish colors in Fig. 6.2 represent smaller relative entropies. Two blocks along the diagonal can be distinguished that coincide with the plays of the two different genres. Indeed, if we sequentially extract one text from

Table 6.2: Relative entropies from *Two Noble Kinsmen* to different profiles (x100).

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

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 6.1 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 6.1. 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.

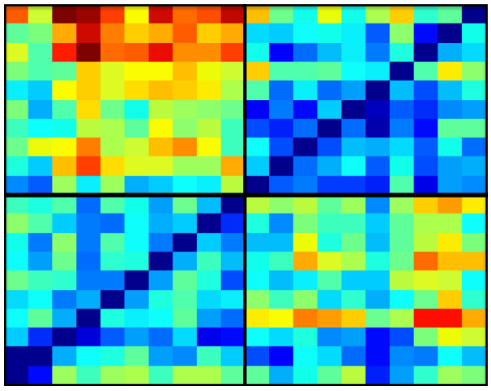


Figure 6.2: Heat map of relative entropies between 20 Shakespeare extracts. The first 10 texts correspond to history plays while the last 10 correspond to comedy plays. Relative entropies within texts of the same genre are smaller than across genres.

Table 6.3: Relative entropies from *Two Noble Kinsmen* to hybrid profiles (x100).

	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

6.3 Gender

Word usage can be used for author profiling [34] and, in particular, to infer the gender of an author from the written text. To illustrate this, we divide the 21 authors from the 19th century [32] 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 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,

Table 6.4: Relative entropies from *Eastward Ho* to hybrid profiles composed of two authors (x100).

	Sh.	Jon.	Fle.	Mid.	Cha.	Marl.
Sh.	17.6	16.8	17.3	16.7	17.1	18.2
Jon.	16.8	**16.8**	17.0	**16.5**	**16.7**	17.3
Fle.	17.3	17.0	18.7	17.6	17.4	17.9
Mid.	16.7	**16.5**	17.6	17.6	16.9	17.1
Cha.	17.1	**16.7**	17.4	16.9	**17.5**	17.8
Marl.	17.4	17.1	17.6	17.3	17.4	18.1

we obtain a mean accuracy of 0.63. Although this accuracy is lower than state-of-the-art gender profiling methods [35], the difference with random attribution, i.e. accuracy of 0.5, validates the fact that WANs carry gender information about the authors.

6.4 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 [32]. 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 [36]. In Table 6.2, 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 6.3 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 6.3 corresponds to profiles of sole authors and, thus, coincides with Table 6.2. The smallest relative entropy in Table 6.3 is achieved by the hybrid profile composed by Fletcher and Shakespeare, which is consistent with the accepted attribution of the play.

The attribution between hybrid profiles is not always accurate. For example, consider the play *Eastward Ho* which is a collaboration between three authors, two of which are Chapman and Jonson [36]. If we repeat the above procedure and compute the relative entropies between the play and the different pure profiles, we see that in fact the two smallest entropies are achieved for Jonson and Chapman; see the diagonal in Table 6.4. However, the smallest entropy in the whole table is achieved by the hybrid profile composed by Jonson and Middleton. The hybrid profile of Jonson and Chapman, the real authors, achieves an entropy of 16.7, which is the second smallest among all profiles in Table 6.4.

Chapter 7

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 [37]. 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 [38, Chapter 8], nearest neighbors (NN) [38, Chapter 2], decision trees (DT) [38, Chapter 14], and support vector machines (SVM) [38, Chapter 7].

In Table 7.1 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 [32] and attribute ten excerpts of each of them

Table 7.1: 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.

Nr. auth.	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

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) [39].

The WANs achieve a lower attribution error than frequency based methods; see Table 7.1. 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 7.1 we inform the error rate of majority voting between WANs and the two best performing frequency based methods, namely, naive Bayes and SVMs. For the voting method, the 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 current state of the art.

Although the error rates presented in Table 7.1 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.

Part II

WANs applied to Research in Literature

Chapter 8

Early Modern English Literature

The stylometric analysis in this part of the thesis focuses on the attribution of plays written during the English Early Modern period stretching from the late 16th century to the early 17th century. William Shakespeare is the most prominent playwright active in this period but there are several other authors that were also active during this time. Due to the rudimentary documentation and the frequent collaborations among playwrights, there are plenty of authorship controversies around the plays written in this period. We can divide these controversies into two categories: finding the real authors of anonymous plays and determining which author wrote which part in collaborative plays. By using WANs, we tackle both types of controversies and in some cases we support existing theories whereas in others we propose new theories of our own.

8.1 Author Profiles

We focus on the authors listed below where we also detail the number of plays that are currently attributed to each of them and the period during which they are presumed to have written said plays. The information is compiled from the Database of Early English Playbooks (DEEP) [36] and the database of catalogued plays in Chadwyck-Healey Literature Online (LION) [40]. Whenever inconsistencies in authorship information arise, we consider [36] as accurate.

(1) George Chapman (1559-1634), active circa 1596-1620. Considered sole author of a total of 13 plays plus 2 collaborations.

(2) John Fletcher (1579-1625), active circa 1605-1625. Supposed to have written 47 plays, being sole author in 22 of them while Francis Beaumont and Phillip Massinger were his main collaborators in the rest.

(3) Ben Jonson (1572-1637), active circa 1596-1637. Presumed sole author of 17 plays plus 1 collaboration.

(4) Christopher Marlowe (1564-1593), active circa 1586-1593. Putative sole author of 6 plays and 1 collaboration.

(5) Thomas Middleton (1580-1627), active circa 1603-1625. Believed to have written 26 plays, 14 of them as sole author and 12 in collaboration.

(6) William Shakespeare (1564-1616), active circa 1589-1614. A total of 38 plays are attributed to Shakespeare and collaborators.

In the above list, we do not consider as plays minor dramatical compositions such as masques, entertainments and pageants. Chapman, Fletcher, Jonson, Marlowe, Middleton, and Shakespeare are included in our analysis since they posses large and well studied canons compared with their contemporaries.

The WAN attribution algorithm developed in the first part of this work uses known texts of a given author to construct a profile against which unknown texts are compared. Since profiling accuracy increases with the length of the texts considered when building the profile, we build profiles from all texts of sole authorship for a given author that have little or no history of authorship dispute. The full list of plays used to build the six profiles is reported in Table 8.1. When building profiles for a given author, we generally subscribe to the information provided in [36] to determine texts of sole authorship. An exception to this is Middleton, who's profile is built using the texts attributed to him in the 2007 Oxford Collected Works of Middleton [41], which contains a more recent and accepted study of his canon. Two plays included in Middleton's corpus in [41]–*The Revenger's Tragedy* and *The Second Maiden's Tragedy*–were published anonymously and have long history of disputed authorship [42–45]. To be safe, we do not include these plays in Middleton's profile but provide an analysis of their authorship in Chapter 10.

Notice that each profile is built from a different number of texts. Marlowe, the least prolific writer of the ones here considered, is accepted as the sole author of 6 plays that totalize 103,160 words. Shakespeare, the most prolific sole author, is the undisputed sole author of 28 plays, totaling 679,256 words. Due to this difference, we compute the relative entropy between the WAN of an unknown text $u \in U$ and each profile using (3.10) rather than (3.9).

In order to generate faithful representations of authors' styles, we remove artifacts introduced by modern transcriptions by using the earliest editions available of each text in

Table 8.1: Plays used to build author profiles

William Shakespeare		Thomas Middleton
Antony and Cleopatra (ANT)	The Merchant of Venice (MV)	Your Five Gallants (FIV)
All's Well that Ends Well (AWW)	The Merry Wives of Winsdor (WIV)	A Game at Chess (GAC)
As You Like It (AYL)	A Midsummer Night's Dream (MDB)	A Mad World My Masters (MAD)
The Comedy of Errors (ERR)	Much Ado About Nothing (ADO)	A Chaste Maid in Cheapside (MAC)
Coriolanus (COR)	Othello (OTH)	Hengist King of Kent (HEN)
Cymbeline (CYM)	Richard II (R2)	Michaelmas Term (MIC)
Hamlet (HAM)	Richard III (R3)	More Dissemblers Besides Women (DIS)
1 Henry IV (1H4)	Romeo and Juliet (ROM)	No Wit, No Help Like a Woman's (NOW)
2 Henry IV (2H4)	The Taming of the Shrew (SHR)	The Phoenix (PHO)
Henry V (H5)	The Tempest (TMP)	The Puritan Widow (PUR)
Julius Caesar (JC)	Troilus and Cressida (TRO)	A Trick to Catch the Old One (TCO)
King John (JN)	Twelfth Night (TN)	The Widow (WID)
King Lear (LR)	The Two Gentlemen of Verona (TGV)	The Witch (WTH)
Love Labour's Lost (LLL)	The Winter's Tale (WT)	Women Beware Women (BEW)
Ben Jonson	**John Fletcher**	**George Chapman**
Alchemist (ALC)	Bonduca (BON)	All Fools (ALL)
Bartholomew Fair (BAR)	Chances (CHA)	Sir Giles Goosecap (SGG)
Catiline's Conspiracy (CAT)	The Faithful Shepherdess (TFS)	Bussy Dambois (BDA)
Cynthia's Revels (CYN)	The Humorous Lieutenant (HUM)	Caesar and Pompey (CAP)
The Devil is an Ass (DIA)	The Island Princess (ISL)	The Conspiracy of Charles Duke of Byron (CDB)
Epicoene (EPI)	The Loyal Subject (LOY)	The Tragedy of Charles Duke of Byron (TDB)
Every Man in His Humour (MIH)	The Mad Lover (TML)	The Gentlemen Usher (GEN)
Every Man Out of His Humour (MOH)	Monsieur Thomas (THO)	A Humorous Day's Mirth (HDM)
The Magnetic Lady (MAG)	The Pilgrim (PIL)	May Day (MAY)
The New Inn (NEW)	Rule a Wife and Have a Wife (RAW)	Monsieur D'Olive (MDO)
Poetaster (POE)	Valentinian (VAL)	The Blind Beggar of Alexandria (BBA)
The Sad Shepherd (SAD)	Wife for a Month (WFM)	The Revenge of Bussy Dambois (RBD)
Sejanus's Fall (SEJ)	The Wild Goose Chase (WGC)	The Widow's Tears (WID)
The Staple of News (SON)	The Woman's Prize (WPR)	**Christopher Marlowe**
A Tale of a Tub (TUB)	Women Pleased (WPL)	Dr Faustus (DRF)
Volpone (VOL)		Edward II (E2)
		The Jew of Malta (JEW)
		The Massacre at Paris (MAS)
		1 Tamburlaine (T1)
		2 Tamburlaine (T2)

the LION database [40], with the exception of Shakespeare's First Quarto editions. Although Shakespeare's canon was first published in full in 1623, there exist earlier editions for a number of his plays known as First Quartos. As there is currently no scholarly consensus on which editions are more authoritative, to be consistent we use 1623 editions for all Shakespeare texts. When using original transcriptions we have to account for the fact that many words had multiple accepted spellings during the Early Modern era. E.g., the word 'of' is also spelled as 'off', 'offe', or 'o' whereas the word 'with' may also appear as 'wid', 'wyth', 'wytt', 'wi', 'wt', and 'wth'. Many of these alternate spellings are used infrequently and thus do not contribute highly to the WAN of a text. Nevertheless, we correct the WANs so that the occurrence of any of the alternative spellings is treated as the same word. We emphasize that spelling preferences carry little information about the authorship of a play. Indeed, spellings in printed editions were not necessarily those of authors as they were often selected by printers to accommodate the fixed length of lines in printing presses [46]. In addition, we remove speech prefixes, or the character name preceding each speech, to avoid cases in which character names are abbreviated to function words (e.g. Anne abbreviated to 'An').

For the WANs in this work we use the optimal parameters determined in Chapter 4, $\alpha = 0.75$ and $D = 10$. Because punctuation marks were often added by publishers rather than the authors themselves [47], we instead delimitate sentences at the end of character speeches. The WANs are built with the 100 most common function words in the Early Modern period, listed in Table 8.2. This number is chosen based on an adaptive training

Table 8.2: List of function words used in WANs

a	both	in	no	past	this	while
about	but	into	none	shall	those	who
after	by	it	nor	should	though	whom
against	can	like	nothing	since	through	whose
all	close	little	of	so	till	will
an	could	many	off	some	to	with
and	dare	may	on	such	until	within
another	down	might	once	than	unto	without
any	enough	more	one	that	up	would
as	every	most	or	the	upon	yet
at	for	much	other	them	us	
away	from	must	our	then	what	
bar	given	need	out	therefore	when	
because	hence	neither	over	these	where	
before	if	next	part	they	which	

Table 8.3: Common alternative spellings for function words

Conventional	Alternative					
it	yt	t				
of	off	offe	o			
that	thatt	thate	yat	yt		
with	wid	wyth	wytt	wi	wt	wth

period as described in Chapter 4 based on all the texts with undisputed authorship, i.e. those plays listed in Table 8.1. A list of the most common Early Modern period alternative spellings is given in Table 8.3. For the cases where one alternative spelling can be assigned to multiple conventional spellings, e.g. ‘yt’ can be associated with ‘it’ and ‘that’, we assign every appearance of the alternative spelling to the most common usage.

Table 8.4: Relative entropy between profiles.

	Shakespeare	Fletcher	Jonson	Marlowe	Middleton	Chapman
Shakespeare		8.9	4.7	8.9	6.8	4.8
Fletcher	7.4		7.3	14.7	8.0	8.4
Jonson	4.1	7.9		11.1	6.7	5.4
Marlowe	10.1	17.4	13.0		16.5	12.9
Middleton	5.8	8.2	6.3	14.1		6.6
Chapman	4.7	9.6	5.8	11.4	7.3	

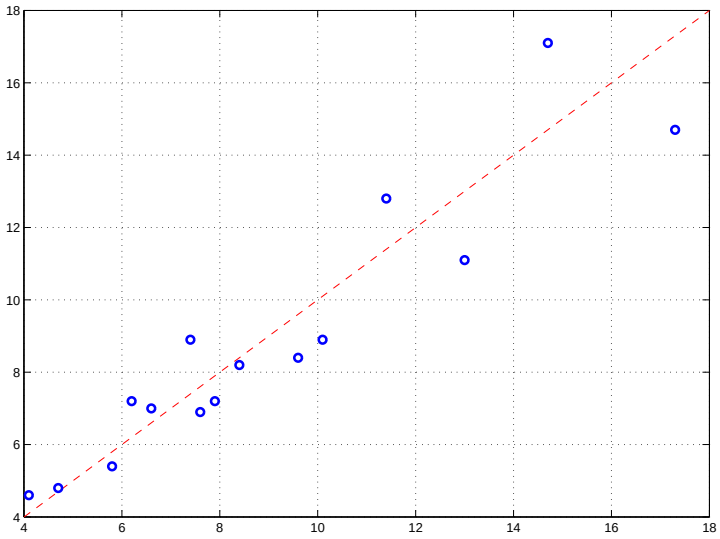


Figure 8.1: Asymmetry of dissimilarities in Table 8.4.

8.2 Similarity of profiles

We compute the relative entropy between every pair of author profiles for the six authors introduced in Section 8.1 using expression (3.10); see Table 8.4. Every entry in the table represents the relative entropy between the corresponding authors in the rows and columns. In this table, as well as in the remaining of the thesis, relative entropies are multiplied by 100 to facilitate their display. We use the term centinats, or cn for short, to denote the resultant unit of measure of relative entropy. The 4.7 in the Chapman row entry and Shakespeare

column entry indicates a relative entropy of $4.7cn$ between Chapman's and Shakespeare's profiles. Note that, as expression (3.10) is not symmetric, the values in the table are also not symmetric, although they are similar in most cases. E.g. the relative entropy between Shakespeare's and Chapman's profiles is $4.8cn$ rather than $4.7cn$ in the opposite direction. In general, dissimilarities between profiles in both directions are highly correlated as can be observed in Fig. 8.1. In this figure, the coordinates of every point correspond to the dissimilarities in both directions for every pair of profiles. The arrangement of the points along the diagonal implies that a high dissimilarity in one direction is associated with a high dissimilarity in the opposite one. Hence, this correlation allows us to speak about the similarity between two authors without specifying a direction.

The entropy-based dissimilarities in Table 8.4 dispel the Marlovian theory of Shakespeare authorship [48]. If Marlowe and Shakespeare were the same person, we should observe the dissimilarities between Marlowe's and Shakespeare's profile to be smaller than the distances between each of the other profiles. However, the relative entropies between Marlowe's and Shakespeare's profiles average $9.5cn$ which is larger than the dissimilarity between Shakespeare and all of the other authors. Shakespeare's profile is, on average, closest to Jonson profile – average relative entropy of $4.4cn$ – although still sufficiently different to assert that they belong to different authors, as verified by the attribution of plays in Chapter 9. The highest dissimilarity among any pair of profiles occurs between Marlowe and Fletcher with a mean of $16.1cn$. As will be seen in Chapter 9, the relative similarity between two profiles affects our ability to distinguish between them when attributing a text.

Chapter 9

Attribution of Undisputed Plays

We attribute the plays written by Jonson, Middleton, Chapman, Marlowe, and Shakespeare among the 6 author profiles introduced in Section 8.1. The attribution of Fletcher's plays is performed in the discussion of collaborations in Chapter 11. When attributing any given play, profiles are built using the plays listed in Table 8.1 excluding the play being attributed. We do not report raw relative entropy values between the play being attributed and the author profiles, but instead subtract from these values the relative entropy between the play and a profile containing all available texts. Intuitively, the profile containing all of the texts represents the writing style of an average playwright from this period. This is done to make the figures easier to view but does not change the results in any way. Each raw relative entropy value is discounted by the same constant value, thus preserving relative distances. As a result, both negative and positive relative entropy values are possible. A negative relative entropy value indicates that the play's WAN is more similar to the author profile

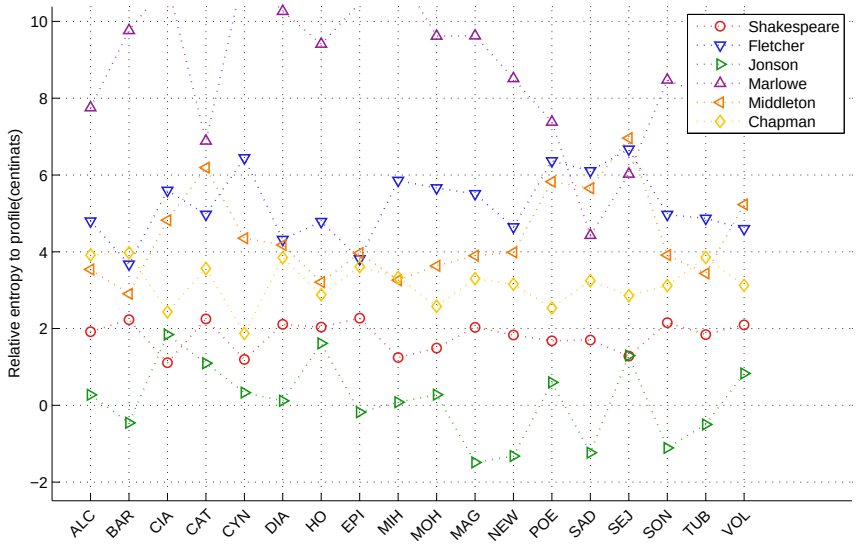


Figure 9.1: Attribution of Jonson plays. We attribute the 16 plays in Table 8.1 plus *The Case is Altered (CIA)* and *Eastward Ho (HO)*. A single misattribution occurs for *The Case is Altered*.

than to the profile of the average playwright while a positive relative entropy indicates the opposite.

9.1 Ben Jonson

In Fig. 9.1 we present the attribution of the 17 plays believed to have been written solely by Jonson, plus one collaboration. In the horizontal axis we present the plays to attribute and the vertical axis represents the entropy distance (3.10) in cn from these plays to the different profiles identified with distinct markers and discounted by the distance to the average playwright.

Among the mentioned 18 total plays, including the collaboration, an accuracy of 94%

Table 9.1: Thomas Middleton plays to be attributed in addition to those listed in Table 8.1.

The Bloody Banquet (BAN)	The Changeling (CHG)
A Fair Quarrel (AFQ)	The Family of Love (FAM)
The Patient Man and The Honest Whore (THW)	Match at Midnight (MAM)
The Old Law (TOL)	The Roaring Girl (TRG)
Anything for a Quiet Life (AGL)	The Spanish Gypsy (TSG)
Wit at Several Weapons (WEA)	

is achieved, correctly attributing 17 of these plays to Jonson, i.e., the entropy distance of every play to the profiles achieves its minimum for Jonson's profile. The play, *Eastward Ho*, is accepted as a collaboration between Jonson and Chapman plus a third author, John Marston, whom we have not profiled. Here Chapman is not well ranked, suggesting that his contributions were minor compared with Jonson's. The relative contributions of both Jonson and Chapman in *Eastward Ho* are analyzed further in Section 11.2.1.

The misattribution in Fig. 9.1 for plays solely written by Jonson occurs for *The Case is Altered*, which is misattributed by a small margin. Mixed authorship has been suggested due to the irregularity in the structure of the last two acts [49]. The play's content has also been compared to *A Comedy of Errors*, written by Shakespeare, who is also here the closest author [50]. Another play, *Sejanus His Fall*, is attributed to Jonson but only by a small margin. It has been pointed out that this play might contain elements of a second author, with both Shakespeare and Chapman a possible candidates [51, 52]. Our analysis indicates that the play is closer to the style of Shakespeare than to the style of Chapman. *Sejanus His Fall* is also one of only two tragedies Jonson published–the other being *Catiline His Conspiracy*–possibly biasing results against a profile built almost entirely with comedies. The relationship between genre and attribution is explored further in Chapter 12.

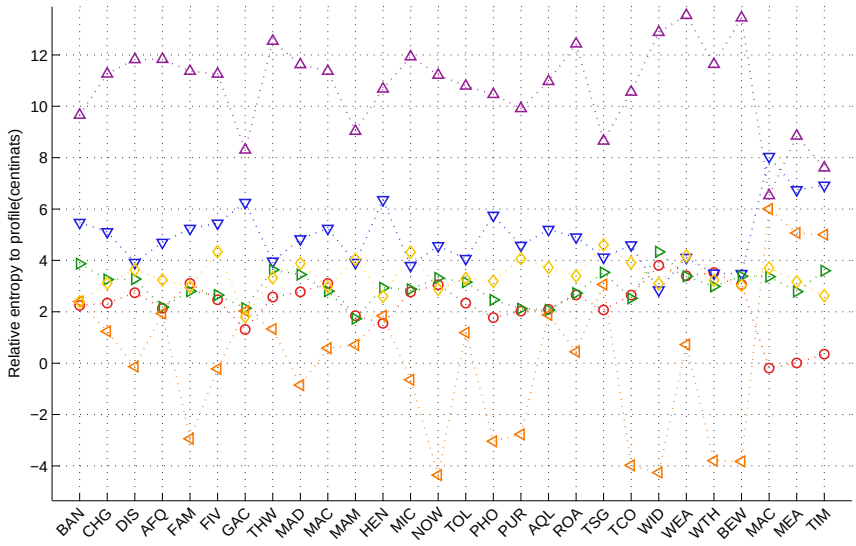


Figure 9.2: Attribution of Middleton plays. We attribute the 14 plays in Table 8.1, the additional 11 plays in Table 9.1 and 3 collaborations with Shakespeare: *Macbeth (MAC)*, *Measure for Measure (MEA)* and *Timon of Athens (TIM)*. Only 2 sole authored plays are misattributed. Also, the attribution of Shakespeare's plays reveal that Middleton's contribution was minor.

9.2 Thomas Middleton

In Fig. 9.2 we present the attribution of 28 plays, 14 of which are generally believed to have been written only by Middleton and 12 in collaboration. We also include in our set two plays originally assigned to Middleton–*The Family of Love* and *Match at Midnight*–but not included in his corpus in [41].

Among the 14 plays believed to be solely written by Middleton, we attribute 12 to Middleton obtaining an accuracy of 85.7%. The first misattributed play, *A Game at Chess*, is attributed to Shakespeare by a very small margin, likely due to random error. This is also true in the case of *Hengist King of Kent*, noted for being the sole history play Middleton

produced. Additionally, although [41] does not find evidence of Middleton in *The Family of Love* or *Match at Midnight*, our results show that he is at least a stronger candidate in these plays than the other five authors. The low relative entropy value of $-3cn$ between Middleton's profile and the WAN of *The Family of Love* adds evidence to the claim that Middleton contributed to this play [53].

Among the 12 collaborative plays, 7 are attributed to Middleton. Thomas Dekker and William Rowley were Middleton's usual collaborators. As neither of these authors are profiled, each of the plays written with these authors is attributed here to Middleton with the exception of *The Bloody Banquet*, which is marginally attributed to Shakespeare over Middleton. Another misattributed play, *The Spanish Gypsy* is usually considered to be a collaboration between Middleton, Dekker, Rowley, and John Ford [53, 54] which may explain why Middleton is ranked second behind another author. We also attribute Middleton's three collaborations with Shakespeare. *Measure for Measure*, *Timon of Athens*, and *Macbeth* are correctly attributed to Shakespeare. Moreover, for these three plays, Middleton is ranked very poorly being the fourth preferred candidate in all of them. This supports the accepted idea that Middleton's contribution in these three plays is minimal [55, 56]. We examine these plays in closer detail in Section 11.2.3.

9.3 George Chapman

Chapman is considered to be the author of 15 plays, 13 as a sole author and 2 in collaboration. In Fig. 9.3, we attribute these plays among the 6 profiles. In total, 10 of the 15

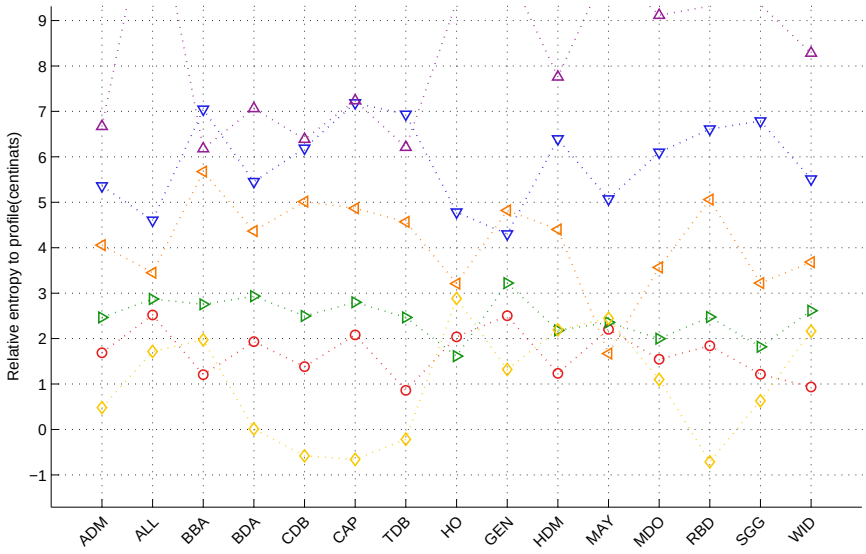


Figure 9.3: Attribution of Chapman plays. We attribute the 13 plays in Table 8.1 plus *The Tragedy of Chabot Admiral of France (ADM)* and *Eastward Ho (HO)*. Out of 15 plays, 10 are attributed to Chapman. Collaboration with Jonson in *Eastward Ho* can also be observed.

plays are attributed to Chapman. In the cases of plays written in collaboration, *The Tragedy of Chabot, Admiral of France* is attributed to Chapman while *Eastward Ho* is attributed to Jonson, as discussed in Section 9.1. Notice that of the four remaining misattributions, three are assigned to Shakespeare with Chapman as the second preferred candidate. This is consistent with the fact that in Table 8.4, Chapman's profile is most similar to Shakespeare. Thus, cases of random error will therefore most likely attribute to Shakespeare.

9.4 Christopher Marlowe

In Fig. 9.4, we present the attribution of 7 plays believed to have been written by Marlowe, where *Dido, Queen of Carthage* is the only collaborative work, with Thomas Nashe as

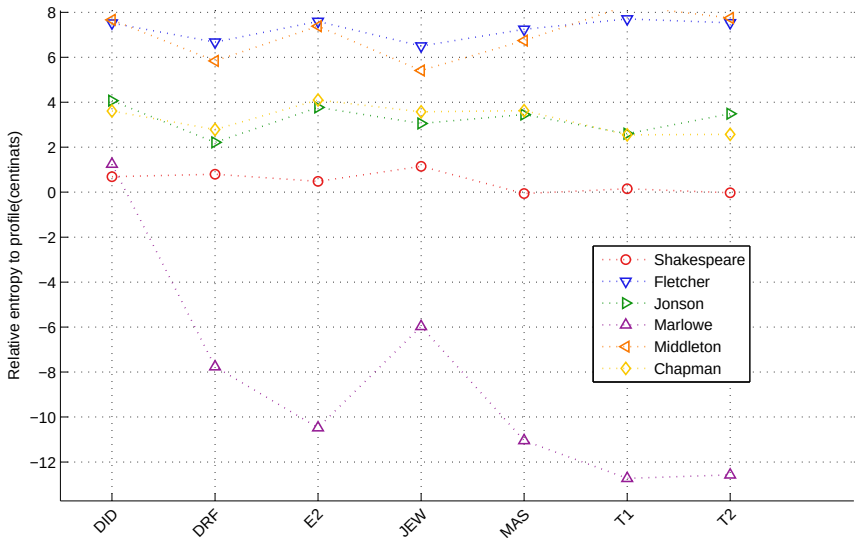


Figure 9.4: Attribution of Marlowe plays. We attribute the 6 plays in Table 8.1 plus *Dido Queen of Carthage (DID)*. A single misattribution occurs for the collaborative play *Dido Queen of Carthage*.

coauthor. We achieve an accuracy of 100% in attributing Marlowe's sole works. *Dido Queen of Carthage* is attributed to Shakespeare by a small margin, with Marlowe as the second best candidate.

In the case of sole authorship plays, each is attributed to Marlowe by a substantial margin and with relative entropies between $-6cn$ and $-13cn$. These large negative values suggest that the plays are much more similar to Marlowe's profile than they are to the profile of an average playwright. This difference may be a result of the fact that Marlowe's plays were written at least a decade before most of the other authors considered, thus indicating a shift in writing style during the one or two decades that separate Marlowe from the rest.

Table 9.2: William Shakespeare plays to be attributed in addition to those listed in Table 8.1.

1 Henry VI (1H6)	2 Henry VI (2H6)
3 Henry VI (3H6)	Henry VIII (H8)
Macbeth (MAC)	Measure for Measure (MEA)
Pericles (PER)	Timon of Athens (TIM)
Titus Andronicus (TIT)	Two Noble Kinsmen (TNK)

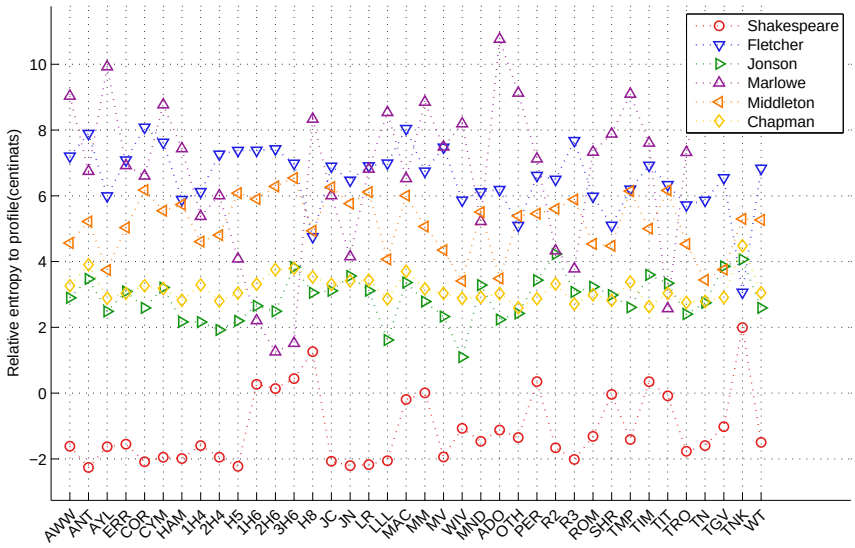


Figure 9.5: Attribution of Shakespeare plays. We attribute the 28 plays in Table 8.1 and the additional 10 plays in Table 9.2. All plays are attributed to Shakespeare. Marlowe's distance to a play is highly dependent on whether the analyzed play is a history play or not, emphasizing the impact of genre in attribution.

9.5 William Shakespeare

In Fig. 9.5 we present the attribution of 38 plays believed to have been written by Shakespeare, 30 of which are attributed solely to Shakespeare in [36]. Note that 2 of the 30 sole authored plays, *2 Henry VI* and *3 Henry VI* are not included in Shakespeare's profile in Table 8.1 because they have a strong history of disputed authorship [57].

All of the 30 plays usually considered to have been written only by Shakespeare are

correctly attributed. However, exceptional situations arise for the plays *1 Henry VI*, *2 Henry VI*, and *3 Henry VI*, in which Marlowe is ranked uncharacteristically high. The fact that Marlowe is ranked second for these plays is noteworthy, since Marlowe's profile is very dissimilar from Shakespeare's in Table 8.4. Consequently, he ranks poorly in the attribution of most plays. In addition, the relative entropy between Marlowe's profile and the WANs of most plays is between $+6cn$ and $+10cn$, while the relative entropy between Marlowe's profile and these plays' WANs is around $+2cn$. Similarly, the relative entropy between Marlowe's profile and the WANs of *Henry V*, *King John*, *Richard II*, and *Richard III* is around $+4cn$. These seven plays have in common that they are history plays, a genre in which Marlowe wrote *Edward II* and *Massacre at Paris*, comprising a third of his profile. Thus, there is a genre bias of history plays towards Marlowe. Focusing on the *Henry VI* saga, where the first part is a known collaboration of Shakespeare with Nashe, we see a particularly strong signature of Marlowe in the three plays compared to Shakespeare's other history plays. Moreover, these plays were written during Marlowe's most fertile years and Marlowe had collaborated with Nashe in 1589 – two years before the *Henry VI* saga – when writing his play *Dido Queen of Carthage*. This supports the hypothesis that there was an unknown collaborator in these plays [58, 59] and points at Marlowe as a probable candidate. These collaborations are covered in greater detail in Section 11.2.4.

Among the 8 plays of accepted collaboration with others, besides the mentioned collaboration in *1 Henry VI*, we can find the three collaborations with Middleton already analyzed in Section 9.2. From the poor ranking of Middleton in the attribution pattern, we

can conclude that Middleton's revisions and contributions were minor. There are also two collaborations with Fletcher, namely *Henry VIII* and *The Two Noble Kinsmen*. We attribute both to Shakespeare, with Fletcher the second preferred author in the latter. In the case of the former, on the other hand, Fletcher is not well ranked and his contribution is not evident from the attribution of the entire play. Shakespeare's collaborations with both Fletcher and Middleton are analyzed further in Sections 11.2.2 and 11.2.3, respectively.

9.6 Summary of Results

In total, we attribute correctly 71 out of the 77 plays we consider that are traditionally attributed to single author and listed in Table 8.1, yielding an accuracy of 92.2%. Furthermore, if we only consider attributions between authors that are more than $5cn$ apart, then we fail only in 3, yielding an accuracy of 96.1%. We utilize the high classification power for plays of sole authorship to shed light on attribution problems of anonymous plays written during the Early Modern period in Chapter 10.

Of the 20 plays we consider that are generally accepted to be collaborations, we attribute 17 to one of the contributing authors, yielding an accuracy of 85%. Collaborative plays are analyzed further in Chapter 11.

Chapter 10

Anonymous Plays

In Fig. 10.1 we present the attribution of 8 anonymous plays written during the English Renaissance. Authorship of some of these plays have been more discussed and studied by scholars than others. E.g., *Edward III* is commonly attributed in part to Shakespeare [60] and our method supports this theory. Indeed, this play was written during the early stages of Shakespeare's career and the Shakespeare profile is the closest. Another play sometimes attributed in part to Shakespeare is *Arden of Faversham* [60]. Again, our method supports this theory. These plays are analyzed further in Section 11.2.4. In addition, the plays *The Revenger's Tragedy*, *The Second Maiden's Tragedy*, and *The Nice Valor* are

Table 10.1: List of texts of unknown authorship.

Arden of Faversham (ARD)	Edward III (E3)
Fair Em (FEM)	Mucedorus (MUC)
The Nice Valor (TNV)	The Revenger's Tragedy (REV)
The Second Maiden's Tragedy (SMT)	Taming of a Shrew (TAS)

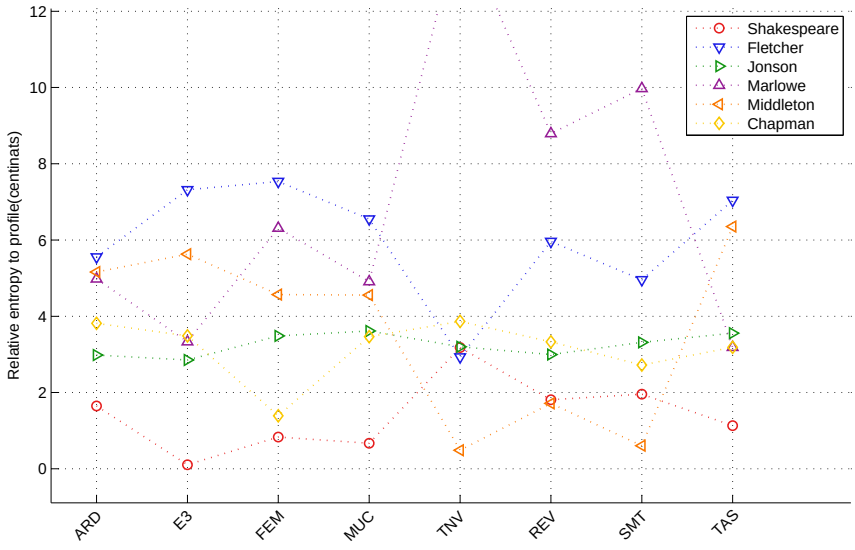


Figure 10.1: Attribution of anonymous plays listed in Table 10.1. Our method supports the usual theories of Shakespeare's hand in *Edward III* and *Arden of Faversham*. Also, Middleton's style in *The Revenger's Tragedy*, *The Second Maiden's Tragedy*, and *The Nice Valor* can be observed, in accordance with current authorship consensus.

usually attributed to Middleton [41], with the former two included in the 2007 Oxford Collected Works. Our method indeed attributes all three works to Middleton. Furthermore, Fletcher is the second attributed author of *The Nice Valor*, a play originally included in the Beaumont and Fletcher folios of 1647 and 1679 [61], leading some to believe that the play is a collaboration between Fletcher and Middleton.

For the remaining plays, definite statements cannot be made, but we can support or undermine existing hypothesis. For example, *Mucedorus* may have been written by Shakespeare as proposed by a number of scholars [62] since he is the first ranked author among the six authors we profile. *Fair Em* has also been assigned to Shakespeare [63] though there is no scholarly consensus, with Robert Wilson, whom we do not profile, often cited

as a likely candidate [62]. *The Taming of a Shrew*, the play generating controversies about the better known Shakespeare play with similar title, is here attributed to the Shakespeare profile. Note, also, that Marlowe is ranked atypically high for this play–second behind Shakespeare. Both Shakespeare and Marlowe have been proposed as candidates for *Taming of a Shrew* [64], in the former case as a possibly early draft of *Taming of the Shrew*. While our analysis points to Shakespeare as a more likely candidate, observe that the attribution of *Taming of the Shrew* in Fig. 9.5 ranks Marlowe as the worst candidate, indicating that much more of his style is evident in the early draft.

Chapter 11

Collaborations

In cases of multiple authors contributing to a single play, we show how our method is still able to detect one or more of the authors present in a full text by identifying the top ranked authors in its attribution.

Table 11.1: Plays used to build profiles for Fletcher & Beaumont and Fletcher & Massinger.

Fletcher & Beaumont	
The Coxcomb (COX)	Philaster (PHI)
The Woman Hater (TWH)	Cupid's Revenge (CUP)
A King and No King (KNK)	Love's Pilgrimage (PIL)
The Maid's Tragedy (TMT)	The Scornful Lady (TSL)
Fletcher & Massinger	
The Custom of the Country (COC)	The Double Marriage (TDM)
The Elder Brother (TEB)	The False One (TFO)
John Van Olden Barnavelt (JVO)	The Little French Lawyer (LFL)
The Lover's Progress (LP)	The Prophetess (PRO)
The Sea Voyage (SEA)	Spanish Curate (TSC)
A Very Woman (TVW)	

Table 11.2: John Fletcher plays to be attributed in addition to those listed in Table 8.1.

Solo	
Beggars' Bush (BB)	The Captain (CAP)
The Fair Maid of the Inn (FAI)	The Noble Gentlemen (TNG)
The Queen of Corinth (QOC)	Wit Without Money (WIT)
Collaborations	
Henry VIII (H8)	The Knight of Malta (KOM)
The Maid in the Mill (MIL)	The Night Walker (NW)
Four Plays in One (FP)	Two Noble Kinsmen (TNK)
Wit at Several Weapons (WEA)	Love's Cure (CUR)
The Bloody Brother (BRO)	Thierry and Theodoret (THI)
Wandering Lovers (WAN)	

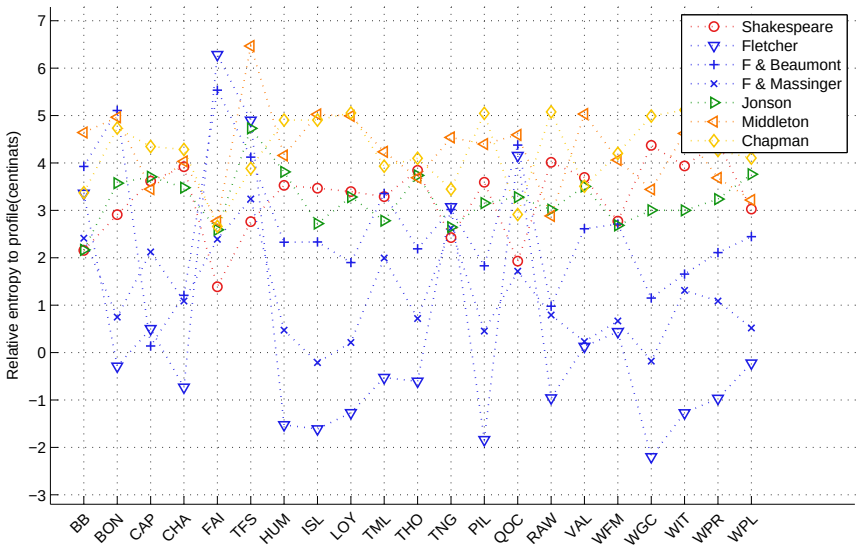


Figure 11.1: Attribution of solo Fletcher plays. We attribute the 15 plays in Table 8.1 and the additional 6 plays in Table 11.2. Six plays are not attributed to the sole Fletcher profile and, among these, two plays are attributed to collaborative profiles including Fletcher.

11.1 John Fletcher and collaborators

John Fletcher wrote numerous plays both by himself and with collaborators. Consequently, his canon is an appropriate text corpus to analyze the attribution of collaborative plays. In addition to the six profiles in the previous section, we include two profiles built from plays written with Fletcher's two most frequent coauthors–Francis Beaumont and Phillip Massinger; see Table 11.1.

The attribution of Fletcher's works are divided into two plots. Fig. 11.1 shows the attribution of plays believed to have been written solely by Fletcher and Fig. 11.2 shows the attribution of plays believed to have been written in collaboration with other authors. The set of plays presented before the first red line include attributions of plays written with Francis Beaumont. The second division shows the attribution of plays written with Phillip Massinger and the third division shows the attribution of plays written with a mix of other authors. In both figures we omit the marker corresponding to Marlowe since he is poorly ranked for every play. This is consistent with Fletcher and Marlowe having the most dissimilar writing styles; see Table 8.4.

In Fig. 11.1, 15 out of 21 plays are attributed to the solo Fletcher profile. Of the six plays attributed to other profiles, two of them, *The Captain* and *Queen of Corinth* are attributed to one of the profiles for Fletcher and a collaborator. *Beggar's Bush* is marginally assigned to Shakespeare and Jonson. *The Faithful Shepherdess*, *The Noble Gentleman*, and *The Fair Maid of the Inn* are mistakenly assigned to Shakespeare as well, with Fletcher and Massinger ranked second. For the latter, existing theories attribute the play to a collabora-

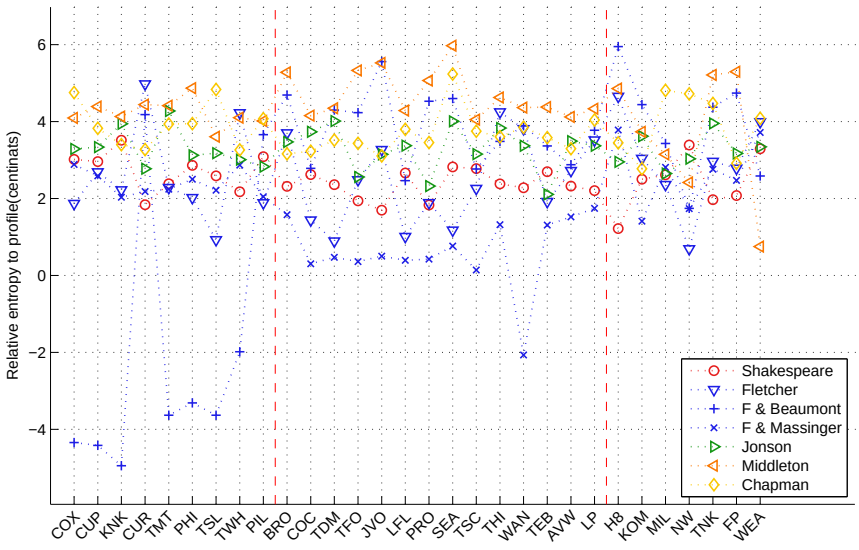


Figure 11.2: Attribution of Fletcher plays written with collaborators. We attribute the plays listed in Tables 11.1 and 11.2. The first division includes plays written with Beaumont and 7 out of 9 are correctly assigned to the Fletcher & Beaumont profile. The second division includes plays written with Massinger and all 14 plays are assigned to the Fletcher & Massinger profile. The third division includes plays written with other collaborators and 3 out of 7 are assigned to a Fletcher profile. Out of 30 plays, a total of 25 are assigned to a Fletcher profile.

tion of four authors, two of which are Fletcher and Massinger, with Fletcher's contribution being minor [65]. This would explain the fact that the Fletcher and Massinger profile is ranked second but the sole Fletcher profile is poorly ranked.

In Fig. 11.2, 7 of the 9 Fletcher and Beaumont plays are attributed to the Fletcher and Beaumont profile, while *Philaster* is assigned to the sole Fletcher profile. A single mistake occurs for *Love's Cure*, a play historically attributed to many different authors [66]. Additionally, all of the 14 Fletcher and Massinger plays are assigned to the Fletcher and Massinger profile. One of the three Fletcher profiles are also listed as the top candidate in 3 out of the 7 plays written by Fletcher with other collaborators. Of the four mistakes,

Table 11.3: Plays used to build profiles for Robert Greene and George Peele.

Robert Greene	
Friar Bacon and Friar Bungay	Orlando Furioso
James IV	Alphonsus, King of Aragon
George Peele	
The Arraignment of Paris	Edward I
The Battle of Alcazar	The Love of King David and Fair Bethsheba
Old Wive's Tale	

Table 11.4: Function words used in the attribution of individual acts, determined in the training process. A total of 76 words are used.

a	both	like	nor	shall	they	when
about	but	little	nothing	should	this	where
against	by	many	of	since	those	which
all	can	may	off	so	though	who
an	could	might	on	some	till	whose
and	for	more	once	such	to	will
any	from	most	one	that	unto	with
as	if	much	or	the	up	without
at	in	must	other	them	upon	would
away	into	no	our	then	us	yet
before	it	none	out	these	what	

two are the plays coauthored with Shakespeare and discussed previously in Section 9.5 and further in Section 11.2.2. These examples demonstrate that our tool remains effective even in cases of mixed authorship and, in many cases, favors profiles built from multiple contributing authors over profiles built from a single contributing author.

11.2 Intraplay analysis

We examine the authorship of collaborative plays through the attribution of its individual acts and scenes. In Section 11.1 we analyzed examples of detecting collaboration in full plays by looking at the top candidate authors. This does not, however, suggest any partic-

Table 11.5: Function words used in the attribution of individual scenes, determined in the training process. A total of 55 words are used.

a	at	from	more	of	shall	the	to	where	yet
all	away	if	most	on	should	them	up	which	
an	but	in	much	one	so	then	upon	who	
and	by	it	must	or	some	these	us	will	
any	can	like	no	our	such	they	what	with	
as	for	may	nor	out	that	this	when	would	

ular breakdown of which sections of the text were contributed by which author. Instead, we may attribute pieces of the play separate from one another to gain deeper insight as to how the play was written. We also see cases where we can detect collaboration through intraplay analysis where we could not when attributing the full text.

In the following sections we attribute plays of known or suggested collaboration between the six original candidate authors as well as two new authors: Robert Greene and George Peele. The plays used to construct Greene's and Peele's profile are listed in Table 11.3. Additionally, we re-train the WAN networks due to the fact that smaller WANS increase the attribution accuracy of shorter texts. This is because shorter texts are less likely to contain less common function words. As a result, larger networks that contain these less common function words are more prone to over-fit to features of specific texts rather than author style. From the training period, we achieve accuracies of 93.4% and 91.5% for acts and scenes, respectively. Note that in the case of scene attribution, this is the accuracy of binary attribution, whereas the act attribution is performed between eight candidate authors. The words used in the resulting networks are listed in Tables 11.4 and 11.5.

The figures display for each act or scene the difference in relative entropy when com-

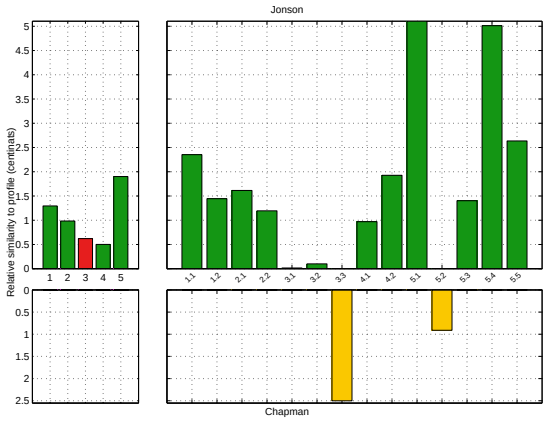


Figure 11.3: Attribution of acts and scenes of *Eastward Ho*. Note that Act 3 is assigned to Shakespeare over both Jonson and Chapman.

paring the two top candidate authors, reflected by both the color of the bars and the titles above and below the plot. A longer bar in a particular direction indicates a larger difference between the entropies of the two candidate authors. For example, in Fig. 11.4, red bars extending upwards indicate an attribution to Shakespeare while blue bars extending downwards indicate an attribution to Fletcher. In the attribution of acts, we identify the two top authors as the two highest ranked, whereas the attribution of scenes we consider the two authors most often cited as candidates. In many cases, the acts and scenes will be attributed between the same pair of authors. Cases in which an act is attributed to a third author are marked in the figure captions.

11.2.1 Jonson and Chapman

We attribute both the individual acts and scenes of the single known collaboration between Jonson and Chapman, *Eastward Ho*, which also includes contributions from a third author, John Marston. Fig. 11.3 displays the results of the act and scene attribution. Every act

is assigned to Jonson, with the exception of Act 3 assigned to Shakespeare. Chapman is ranked either third or forth in all acts except Act 3 in which he is ranked second. These results are similar to the full play attribution from Figs. 9.1 and 9.3, in which Jonson was the top ranked author and Chapman was not well ranked. While these results on their own do not support Chapman's contribution, a look at the scene attribution does reveal some of Chapman's possible contributions. Most of the play is still assigned to Jonson, however Chapman is seen as a more likely candidate in scene 3.3 and 5.2 whereas the attribution of scenes 3.1-2 is too close to make any conclusion. While there is not a scholarly consensus on the scene breakdown, many attribute Marston to Act 1, Chapman to Act 2 and 3, and Jonson to Act 5 [49]. Most scholars agree in particular about scene 3.3 being written by Chapman [58]. Our results support the notion that Chapman did not write Act 1 and Jonson wrote Act 5. We also provide further evidence that Chapman wrote 3.3, as it is, in our analysis, the single scene that is assigned to Chapman by a margin larger than $2cn$. We also, however, find more evidence of Jonson contributing Acts 2 and 4 than Chapman.

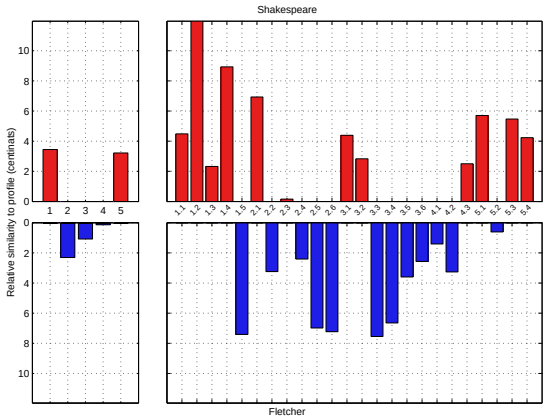


Figure 11.4: Attribution of acts and scenes of *Two Noble Kinsmen*.

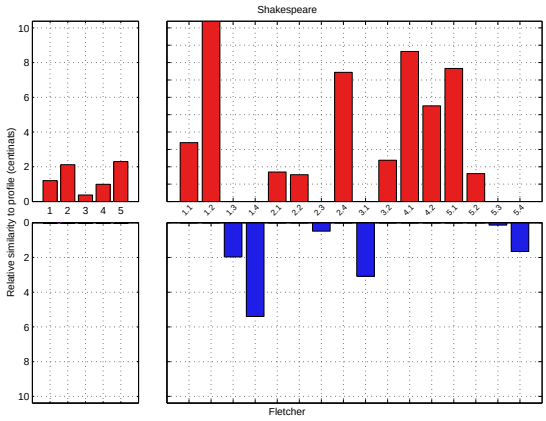


Figure 11.5: Attribution of acts and scenes of *Henry VIII*.

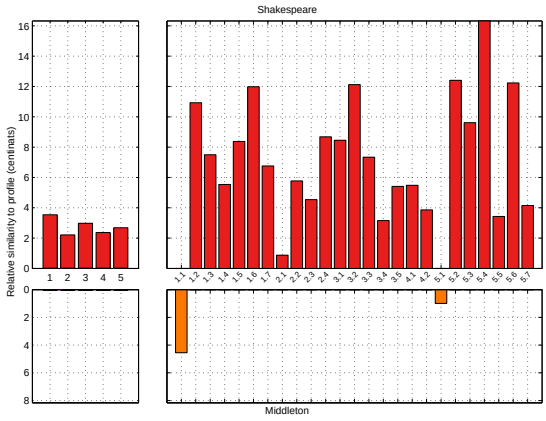


Figure 11.6: Attribution of acts and scenes of *Macbeth*.

11.2.2 Shakespeare and Fletcher

In Fig. 11.4 we show the attribution of individual acts and scenes of *Two Noble Kinsmen*, a known collaboration between Shakespeare and Fletcher. Whereas in Fig. 11.2 the play is assigned to Shakespeare with Fletcher as the second best candidate, here Acts 1 and 5 are assigned to Shakespeare while Acts 2 and 3 are assigned to Fletcher. Act 4 is assigned to Fletcher with Shakespeare and Jonson close behind. A closer look into the scene breakdown reveals more specific assignments. Shakespeare is assigned to scenes 1.1-4, 2.1,

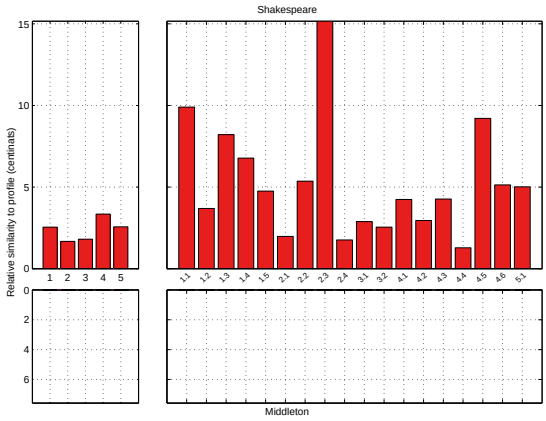


Figure 11.7: Attribution of acts and scenes of *Measure for Measure*.

3.1-2, 4.3, 5.1, and 5.3-4, Fletcher is assigned to scenes 1.5, 2.2, 2.4-6, 3.3-6, and 4.1-2, and close ties in scenes 2.3 and 5.2. The scene breakdown we propose largely supports the one given by Hallet Smith in *The Riverside Shakespeare* [67].

The act and scene analysis of Shakespeare and Fletcher's other collaboration–*Henry VIII*–is displayed in Fig. 11.5. Recall that, when attributing the full play, Shakespeare was the top candidate while Fletcher was in fact ranked fourth, thus revealing no evidence of collaboration; see Fig. 9.5 or Fig. 11.2. We see similar results in Fig. 11.5, in which Shakespeare is assigned every act. Fletcher, again, is ranked poorly in every act. A scene-by-scene analysis between Shakespeare and Fletcher however, does reveal Fletcher to be a stronger candidate than Shakespeare in several individual scenes. In fact, the scene breakdown we observe–in which Shakespeare is assigned scenes 1.1-2, 2.1-2, 2.4, 3.2, 4.1-2, and 5.1-2 and Fletcher is assigned scenes 1.3-4, 3.1, and 5.4, and 2.3 and 5.3 ties between both authors–is aligned to that proposed by Cyrus Hoy [61] and currently accepted by many scholars. The primary area of disparity between the breakdown we propose and the one

given by Hoy is the authorship of Act 4. While Hoy assigns Act 4 to Fletcher, we find that there is greater evidence that Shakespeare contributed this section. Both scenes are attributed to Shakespeare by a significant margin of at least $5cn$. Another point of contention is the assignment of 2.3–given to Shakespeare by Hoy–to Fletcher by a small margin.

The attribution of *Henry VIII* shows a clear example of using intraplay analysis to detect collaboration at the level of scenes that may be undetectable when looking at entire plays or acts. In this play, there are several individual scenes that attribute to Shakespeare by a margin as wide as $7cn$, such as scenes 1.2, 2.4, 4.1, and 5.1, that bias the attribution of complete acts in favor of Shakespeare, while the scene to scene analysis provides a clearer perspective.

11.2.3 Shakespeare and Middleton

We analyze in Figs. 11.6-11.8 Middleton's contributions to Shakespeare's plays, *Macbeth*, *Measure for Measure*, and *Timon of Athens*. The attribution of the full plays in Fig. 9.2 did not suggest that Middleton made any significant contribution to any of these plays. The intraplay analysis of *Macbeth* at the level of acts and scenes, shown in Figure 11.6, supports this conclusion. A total of two scenes are assigned to Middleton over Shakespeare, namely scenes 1.1 and 5.1. Scene 5.1 is attributed to Middleton by only a small margin of $1cn$ while scene 1.1 is assigned by a more substantial margin of $3cn$. Scholars have often flagged scenes 1.2, 3.5, and 4.1 as scenes revised or contributed by Middleton [56], although we do not find evidence of this in our analysis.

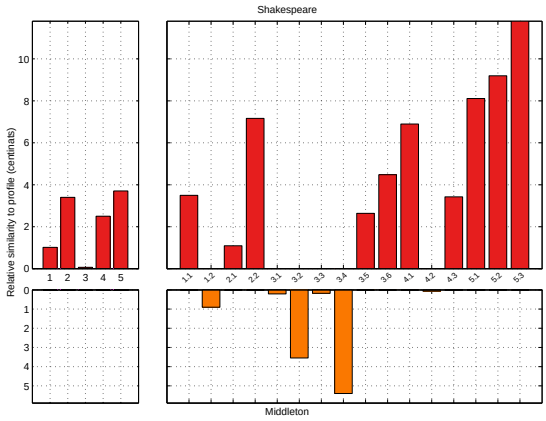


Figure 11.8: Attribution of acts and scenes of *Timon of Athens*.

The case of *Measure for Measure* favors Shakespeare's sole authorship even more; both the act and scene analysis displayed in Fig. 11.7 find Shakespeare to be the sole author of the play. If Middleton had indeed revised the original play as proposed by scholars [56,68], we do not find evidence it was substantial.

Of the three plays, we find that Middleton's contribution was likely largest in *Timon of Athens*. While all five acts attribute to Shakespeare, in Act 3 it is by a margin less than $1cn$ from Middleton; see Fig. 11.8. This is even more evident in the scene analysis. Middleton is a stronger candidate in scenes 1.2, 3.2, and 3.4, with close ties in scenes 3.1, 3.3, and 4.2. This assignment supports much of the claim of authorship provided in [56, 60].

11.2.4 Shakespeare and Marlowe

Although there are no unanimously agreed upon collaborations between Shakespeare and Marlowe, there exist a number of plays with controversial authorship that have been the subject of scholarly treatment regarding Marlowe's contributions. Of these, we examine the

three parts of *Henry VI* as well as the anonymous plays *Arden of Faversham* and *Edward III*.

As suggested by the results in Fig. 9.5, the three parts of *Henry VI* have been considered as possible collaborations between Shakespeare and Marlowe [57], though others such as Greene and Peele have also been suggested. The attribution of the acts of *1 Henry VI*, displayed in Fig. 11.9, suggests that Act 1 could have been written by someone other than Shakespeare. It is here attributed evenly between Shakespeare and Jonson with Marlowe the next preferred candidate. Although Jonson is generally not considered a candidate for this play, it may suggest a similar author we do not profile. The rest of the play is assigned to Shakespeare and, in the case of Acts 3 and 4, by a wide margin from second candidate Marlowe. The scenes are attributed between Shakespeare and Marlowe. In line with the act attribution, three scenes in Act 1 (1.1, 1.5-6) attribute to Marlowe rather than Shakespeare. Other scenes that attribute to Marlowe include 3.2, 3.4, 4.2, 5.1-2. Scene 4.2 in particular is attributed to Marlowe by a large margin of almost $6cn$. These results support parts of the breakdown suggested by Hugh Craig [57], namely the attribution of someone other than Shakespeare in Act 1 as well as Shakespeare in scenes 4.3-7. Although Craig contends that Marlowe likely wrote the scenes involving Joan of Arc, we find only half of the Joan of Arc scenes (1.5-6, 3.2, 5.2) to be more like Marlowe than Shakespeare.

The act and scene attribution of *2 Henry VI* is shown in Fig. 11.10. Act 1 is assigned to Marlowe and the rest is assigned to Shakespeare, with Act 4 being a close tie between them. In the former case, Shakespeare is the third candidate author behind Peele. The

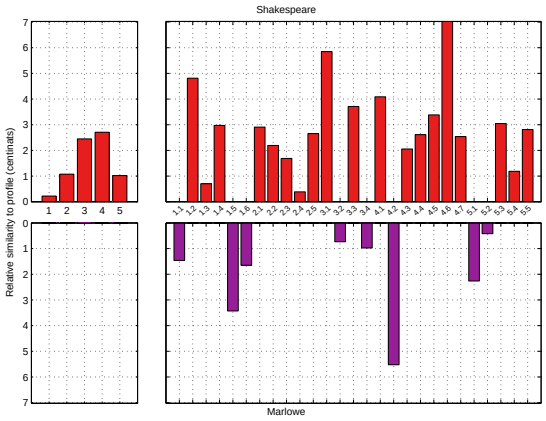


Figure 11.9: Attribution of acts and scenes of *1 Henry VI*.

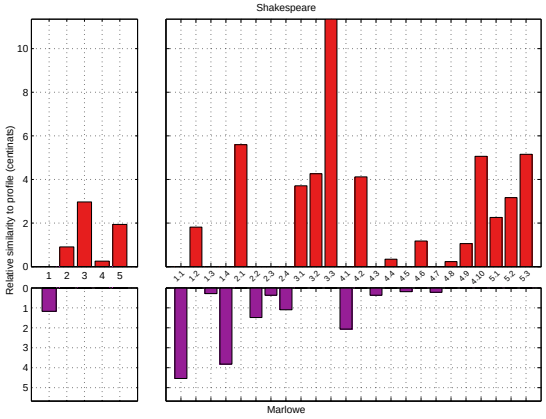


Figure 11.10: Attribution of acts and scenes of *2 Henry VI*.

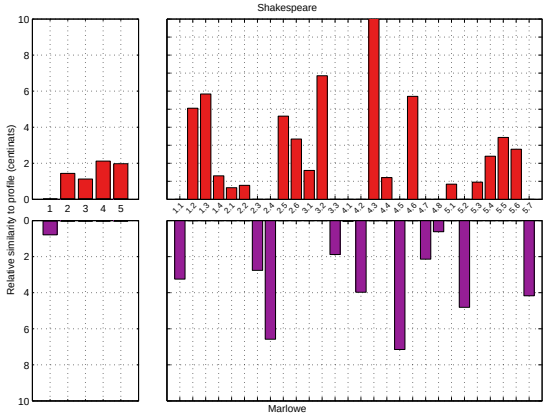


Figure 11.11: Attribution of acts and scenes of *3 Henry VI*. Note that the relative entropy for scene 4.2 extends out of the view of the figure to $+30cn$.

scene analysis assigns to Marlowe scenes 1.1, 1.4, 2.2, 2.4, 4.1, and close ties in scenes 1.3, 2.3, 4.3-5, and 4.7-8. Scenes 1.1 and 1.4, in particular, attribute to Marlowe by a wider margin of $4cn$, increasing the likelihood of his contribution, while the other two scenes in Act 1 show less clear indication of authorship. In comparison to the breakdown offered by Craig, our results support the claims that Shakespeare wrote all of Act 3 and Marlowe possibly wrote scenes involving Jack Cade's rebellion (4.3-9). Act 2, on the other hand, is attributed to Shakespeare in the act analysis but most of the individual scenes are attributed to Marlowe. The WAN of scene 2.1, in particular, has a large relative entropy to Marlowe's profile and indicates a strong likelihood it was written by Shakespeare.

The intraplay analysis of *3 Henry VI* in Fig. 11.11 attributes Act 1 to Marlowe and the rest to Shakespeare. Although Craig has suggested that the part of the text most likely written by other authors is Act 4, the act analysis alone here suggests otherwise. However, the attribution of individual scenes shows a different pattern. Here, Marlowe is assigned four of the eight scenes in Act 4, while Shakespeare is attributed scene 4.3 by a very wide margin of $30cn$–caused by the presence of a rare transition–which likely skewed the entire act in Shakespeare's favor. In addition to scenes 2, 5, 7, and 8 in Act 4, Marlowe is selected as the more likely candidate in scenes 1.1, 2.3-4, 3.3, 5.2, and 5.7. Shakespeare, meanwhile, is assigned scenes 1.2-4, 2.1-2, ,2.5-6, 3.1-2, 4.3-4, 4.6, 5.1, and 5.3-6. Scene 4.1 is a close tie between authors.

We also perform in Fig. 11.12 the intraplay analysis on the play *Arden of Faversham*, attributed to Shakespeare in Fig. 10.1. Every act is attributed here to Shakespeare. Although

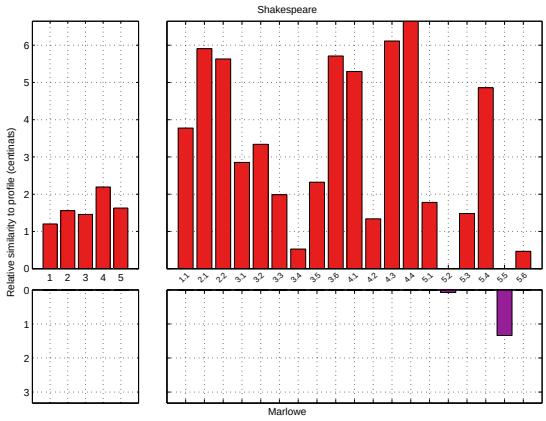


Figure 11.12: Attribution of acts and scenes of *Arden of Faversham*.

not shown in the figure, the second preferred candidate in all acts except Act 5 is Jonson, who is not typically considered a potential author due to the year it was written. The other commonly considered candidates for authorship are Thomas Kyd and Marlowe [57, 69]. The former is not profiled due to a lack of a sufficient number of texts to build a profile and the latter is not well ranked in Acts 1-4 but is close to the second preferred candidate in Act 5. For this reason, we attribute the scenes between Shakespeare and Marlowe rather than Shakespeare and Jonson. The scene-by-scene analysis shows Shakespeare as the most likely candidate for almost the entire play, with many scenes attributed to Shakespeare by a margin of at least $4cn$. The exception to this is scene 5.5, which is assigned to Marlowe, and scene 5.2, a tie between candidates. Our results support existing claims by MacDonald P. Jackson [70] that Shakespeare at the very least wrote the middle of the play (Act 3), however we also find him to be a likely candidate in at least Acts 1, 2, and 4 as well.

An analysis is performed for *Edward III*, attributed to Shakespeare in Fig. 10.1. As before, the two most commonly cited candidates for co-authorship with Shakespeare are

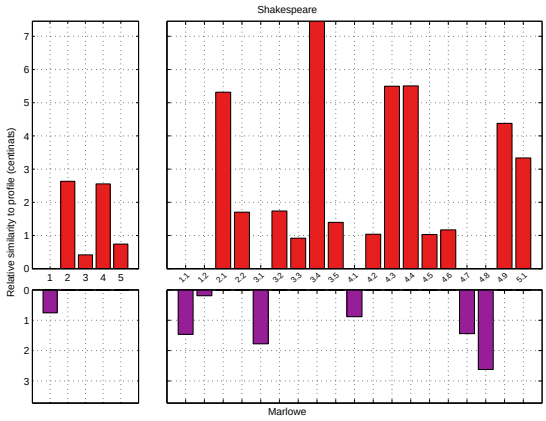


Figure 11.13: Attribution of acts and scenes of *Edward III*.

Kyd and Marlowe [57, 71]. The act attribution of *Edward III* in Fig. 11.13 shows Act 1 assigned to Marlowe. Acts 2, 4, and 5 are attributed to Shakespeare, as well as Act 3 by a small margin of less than $0.5cn$. A look into the scene by scene attribution, however, shows that in addition to 1.1, Marlowe is also assigned scene 3.1 by a clear margin of $2cn$. Marlowe is also assigned scenes 4.1 and 4.7-8, while the attribution of scene 1.2 does not provide a clear candidate. While not shown in Fig. 11.13, the relative entropy values in attribution of scene 4.3 is large between both profiles ($+2cn$ and $+6cn$ between Shakespeare's and Marlowe's profile, respectively), suggesting neither Shakespeare nor Marlowe, but possibly a third author contributed the scene.

Timothy Irish Watt has suggested that Shakespeare wrote scenes 1.2 and 2.1 while someone other than Shakespeare, Marlowe, or Peele wrote scenes 3.1-4.3 [57]. Our results point to Shakespeare as a likely candidate for scene 2.1, with his profile being almost $4cn$ closer to the WAN of *Edward III* than Marlowe's profile. Additionally, along with scene 4.3, we find scenes 3.2-3 and 4.1-2, 4.5 and 4.9 to be possibly written by a third author due

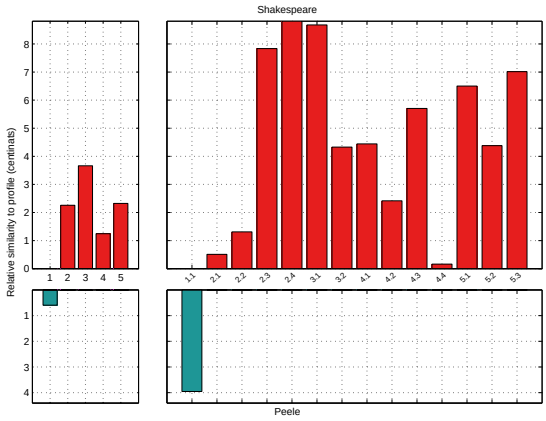


Figure 11.14: Attribution of acts and scenes of *Titus Andronicus*. Note that here the comparative relative entropies for Act 1 and its sole scene, 1.1, differ. The plot of scene 1.1 reports the difference in relative entropy between Peele and Shakespeare while the plot of Act 1 reports the difference in relative entropy between Peele and the second ranked author, Marlowe.

Table 11.6: Relative entropies between scene 3.2 of *Titus Andronicus* and author profiles.

Shakespeare	Fletcher	Jonson	Marlowe	Middleton	Chapman	Peele	Greene
0.47	5.69	2.76	**0.27**	3.72	2.73	4.80	1.12

to comparatively large distance between the scenes' WANs and both profiles. Not displayed in Fig. 11.13, the closest profile between Shakespeare and Peele for each of these scenes has a relative entropy between $+0.1cn$ and $+1.7cn$, whereas all other scenes range from $-0.3cn$ and $-3.5cn$ from the closest profile.

11.2.5 Shakespeare and Peele

Shakespeare's play, *Titus Andronicus*, is commonly cited to include additions by Peele [60], and is attributed act by act and scene by scene in Figure 11.14. Act 1 is assigned to Peele while the rest of the play is attributed to Shakespeare. In the scene attributions scenes 2.1 and 4.4 are attributed to Shakespeare by a small margin of less than $1cn$, evidencing

possible contributions of Peele. Typical attributions of this play, such as the one performed by Brian Vickers [60], assign to Peele Act 1 as well as scenes 2.1 and 4.1.

Another scene of interest in *Titus Andronicus* in the context of attribution studies is scene 3.2, also known as the "Fly" scene. This particular scene is present in the 1623 Folio but not earlier additions, suggesting it was a later addition to the play and possibly added by another author. The relative entropies for this scene are compared in Table 11.6. The two top candidates here are Shakespeare and Marlowe. However, the scene only appeared in editions published long after Marlowe's death so our top candidate for this scene remains Shakespeare.

Chapter 12

Genre Analysis

Based on the results in Section 6.2, we use WANs to distinguish between plays at the level of genre. We build three profiles for each of the three primary genres–comedy, tragedy, and history–using plays that were not written by the six main playwrights studied in this part of the thesis. The complete list of texts used in the genre profiles is displayed in Table 12.1. The profiles use at most one play from any particular author to avoid biasing the results based on author similarity rather than genre similarity.

In Fig. 12.1, the results are shown from the attribution of ten comedy, tragedy, and history plays between the genre profiles. A total of seven of the ten comedy plays–displayed to the left of the first red line–correctly attribute to the comedy profile. Note also that all three misattributions are attributed to the tragedy profile. The attribution of ten tragedy plays, displayed to the right of the first red line, results in only three plays being assigned to the tragedy profile, with *Hamlet* a close three way tie between all profiles. From the

Table 12.1: Plays used to build profiles for genre profiles.

Comedy	
A Shoemaker a Gentleman (William Rowley)	Fair Maid of the West (Thomas Heywood)
City Madam (Phillip Massinger)	Humor Out of Breath (John Day)
Heir (Thomas May)	Orlando Furioso (Robert Green)
Tragedy	
Atheist's Tragedy (Cyril Tourneur)	Rape of Lucrece (Thomas Heywood)
Cleopatra (Samuel Daniel)	Fleire (Edward Sharpham)
Broken Heart (John Ford)	Spanish Tragedy (Thomas Kyd)
History	
Duchess of Suffolk (Thomas Drue)	Edward IV (Thomas Heywood)
Sir John Oldcastle (Robert Wilson)	Thomas Lord Cromwell (S.W.)
Perkin Warbeck (John Ford)	Fuimos Troes (Jasper Fisher)

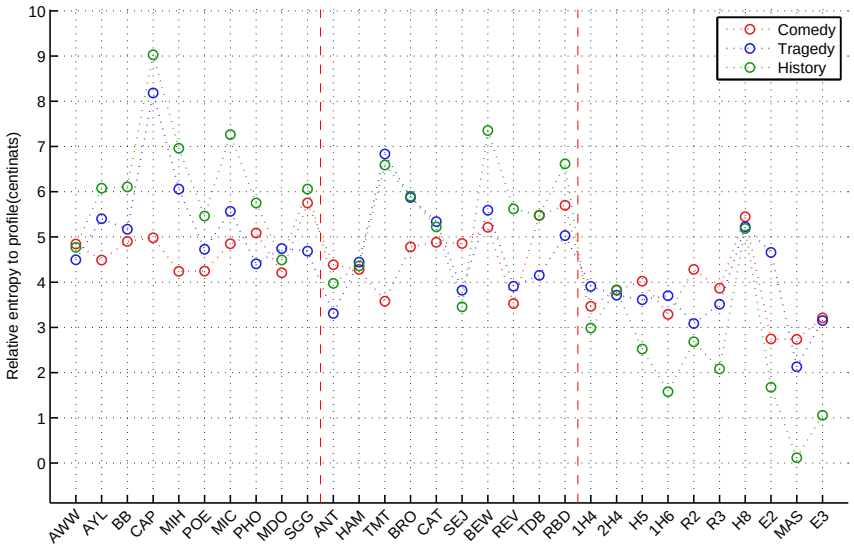


Figure 12.1: Attribution of plays between genre profiles. The plays to left of the first red line include comedy plays. The plays to right of the first red line include tragedy plays. The plays to right of the second red line include history plays.

remaining six plays, five are assigned to the comedy profile. However, the attribution of history plays results in 90% accuracy; shown to the right of the second red line. In our results, we find that distinguishing between history and the other genres is easier than distinguishing between comedy and tragedy. This is interesting because it is common to consider history and tragedy more thematically similar than comedy and tragedy. Our results, by contrast, suggest that the writing styles of comedy and tragedy are more closely linked than the writing styles of either comedy and history or tragedy and history.

Chapter 13

Conclusion

We presented a novel method to solve closed class authorship attribution problems and to answer author profiling questions. This method is based on relational data between function words that we represented using normalized word adjacency networks (WANs). 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 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.

After showing the value of WANs in solving attribution problems, we applied them to analyze the authorship of texts written by popular playwrights during the Early Modern English period. The existence of several authorship controversies during that period made it an attractive dataset for the application of WANs. We validated the method for this dataset by attributing plays of undisputed authorship. After showing high classification accuracy, a selection of anonymous plays were attributed among the author profiles. The classification power was then further evaluated with respect to plays written by multiple authors, both through the attribution of an entire play as well as its individual act and scene components. The acts and scenes were individually analyzed in a set of plays with highly disputed co-authorship, in which we both corroborate existing breakdowns and provide evidence of new assignments. The influence of genre in the choice of function words was also examined.

Bibliography

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

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

[3] Sven Meyer zu Eissen, Benno Stein, and Marion Kulig, "Plagiarism detection without reference collections," in *Advances in Data Analysis*, Reinhold Decker and Hans-J. Lenz, Eds., Studies in Classification, Data Analysis, and Knowledge Organization, pp. 359–366. Springer Berlin Heidelberg, 2007.

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

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

[6] 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.

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

[8] 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.

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

[10] 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.

[11] 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.

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

[13] 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.

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

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

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

[17] 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.

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

[19] 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.

[20] 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.

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

[22] 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.

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

[24] Santiago Segarra, Mark Eisen, and Alejandro Ribeiro, "Authorship attribution through function word adjacency networks," *Transactions on Signal Processing*, vol. (submitted), 2014. Available at http://arxiv.org/abs/1406.4469.

[25] Santiago Segarra, Mark Eisen, Gabriel Egan, and Alejandro Ribeiro, "Stylometric analysis of early modern period english plays," *Preprint*, 2014. Available at https://fling.seas.upenn.edu/∼ssegarra/wiki/index.php?n=Research.Publications.

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

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

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

[29] Hao 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.

[30] M. Vidyasagar, S.S. Mande, C.V.S.K. Reddy, and V.V.R. 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.

[31] Ying Zhao, Justin Zobel, and Phil Vines, "Using relative entropy for authorship attribution," in *Information Retrieval Technology*, HweeTou Ng, Mun-Kew Leong, Min-Yen Kan, and Donghong Ji, Eds., vol. 4182 of *Lecture Notes in Computer Science*, pp. 92–105. Springer Berlin Heidelberg, 2006.

[32] 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.

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

[34] Shlomo Argamon, Moshe Koppel, James W. Pennebaker, and Jonathan Schler, "Automatically profiling the author of an anonymous text," *Commun. ACM*, vol. 52, no. 2, pp. 119–123, Feb. 2009.

[35] Moshe Koppel, Shlomo Argamon, and Anat Rachel Shimoni, "Automatically categorizing written texts by author gender," *Literary and Linguistic Computing*, vol. 17, no. 4, pp. 401–412, 2002.

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

[37] Ying Zhao and Justin Zobel, "Effective and scalable authorship attribution using function words," in *Information Retrieval Technology*, GaryGeunbae Lee, Akio Yamada, Helen Meng, and SungHyon Myaeng, Eds., vol. 3689 of *Lecture Notes in Computer Science*, pp. 174–189. Springer Berlin Heidelberg, 2005.

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

[39] Leo Breiman, Jerome Friedman, Charles J Stone, and Richard A Olshen, *Classification and regression trees*, CRC press, 1984.

[40] Chadwyck-Healey. ProQuest Information and Learning, "Literature Online," *http://lion.chadwyck.com*.

[41] Gary Taylor and John Lavagnino, *Thomas Middleton: The Collected Works*, Oxford University Press, 2007.

[42] Richard H Barker, "The authorship of the second maiden's tragedy and the revenger's tragedy," *The Shakespeare Association Bulletin*, vol. 20, pp. 52–62, 1945.

[43] Cyril Tourneur and Lawrence J Ross, *The revenger's tragedy, ed*, 1967.

[44] Anne Begor Lancashire and Thomas Middleton, *The second maiden's tragedy*, Manchester University Press, 1978.

[45] MWA Smith, "The authorship of the'revengers tragedy'+ tourneur, cyril or middleton, thomas," 1991.

[46] Philip Gaskell, *A new introduction to bibliography*, Clarendon Press Oxford, 1972.

[47] Edwin J Howard, "The printer and elizabethan punctuation," *Studies in Philology*, pp. 220–229, 1930.

[48] Archie Webster, "Was marlowe the man?," *National Review*, pp. 81–6, 1923.

[49] T. P. Logan and D. S. Smith, *The New Intellectuals*, University of Nebraska Press, 1977.

[50] James Loxley, *The complete critical guide to Ben Jonson*, Psychology Press, 2002.

[51] A. Gurr, *The Shakespeare Company, 1594-1642*, Cambridge University Press, 2004.

[52] A. Barton, *Ben Jonson: Dramatist*, Cambridge University Press, 1984.

[53] David J Lake, *The canon of Thomas Middleton's plays: internal evidence for the major problems of authorship*, Cambridge University Press, 1975.

[54] H Dugdale Sykes, "John ford, the author of" the spanish gipsy"," *The Modern Language Review*, vol. 19, no. 1, pp. 11–24, 1924.

[55] MacDonald Pairman Jackson, *Studies in Attribution: Middleton and Shakespeare*, Institut für Anglistik und Amerikanistik, Universität Salzburg Salzburg, 1979.

[56] Stanley Wells, *Shakespeare and Co.: Christopher Marlowe, Thomas Dekker, Ben Jonson, Thomas Middleton, John Fletcher and the Other Players in His Story*, Random House LLC, 2009.

[57] D Hugh Craig and Arthur F Kinney, *Shakespeare, computers, and the mystery of authorship*, Cambridge University Press, 2009.

[58] Edmund Kerchever Chambers, *The Elizabethan Stage*, vol. 3, Clarendon Press Oxford, 1923.

[59] Patrick Cheney, *The Cambridge Companion to Christopher Marlowe*, Cambridge University Press, 2004.

[60] B. Vickers, *Shakespeare, Co-Author: A Historical Study of the Five Collaborative Plays*, Oxford University Press, 2002.

[61] Cyrus Hoy, "The shares of fletcher and his collaborators in the beaumont and fletcher canon (v)," *Studies in Bibliography*, pp. 77–108, 1960.

[62] Terence P Logan and Denzell S Smith, *The predecessors of Shakespeare*, vol. 1, University of Nebraska Press, 1973.

[63] CF Brooke, "Tucker, the shakespeare apocrypha: Being a collection of fourteen plays which have been ascribed to shakespeare," 1908.

[64] Stephen Roy Miller, *The Taming of a Shrew: the 1594 quarto*, Cambridge University Press, 1998.

[65] Terence P Logan and Denzell Stewart Smith, *The Later Jacobean and Caroline Dramatists*, vol. 4, University of Nebraska Press, 1978.

[66] Ernest Henry Clark Oliphant, *Plays of Beaumont and Fletcher*, Yale University Press, 1927.

[67] William Shakespeare, Gwynne Blakemore Evans, and John Joseph Michael Tobin, *The Riverside Shakespeare*, vol. 1, Houghton Mifflin Boston, 1974.

[68] Gary Taylor and John Jowett, *Shakespeare Reshaped, 1606-1623*, Cambridge Univ Press, 1993.

[69] W. W. Greg, "Shakespeare and arden of feversham," *The Review of English Studies*, vol. 21, no. 82, pp. 134–136, 1945.

[70] MacDonald P Jackson, "Shakespeare and the quarrel scene in arden of faversham," *Shakespeare Quarterly*, vol. 57, no. 3, pp. 249–293, 2006.

[71] T. Merriam, "Marlowe's hand in edward iii," *Literary and linguistic computing*, vol. 8, no. 2, pp. 59–72, 1993.