Stylometric analysis of Early Modern period English plays

Mark Eisen and Alejandro Ribeiro
Department of Electrical and Systems Engineering, University of Pennsylvania, Philadelphia, USA

Santiago Segarra
Institute for Data, Systems, and Society, Massachusetts Institute of Technology, Cambridge, USA

Gabriel Egan
School of Humanities, De Montfort University, Leicester, UK

Abstract

Function word adjacency networks (WANs) are used to study the authorship of plays from the Early Modern English period. In these networks, nodes are function words and directed edges between two nodes represent the relative frequency of directed co-appearance of the two words. For every analyzed play, a WAN is constructed and these are aggregated to generate author profile networks. We first study the similarity of writing styles between Early English playwrights by comparing the profile WANs. The accuracy of using WANs for authorship attribution is then demonstrated by attributing known plays among six popular playwrights. Moreover, the WAN method is shown to outperform other frequency-based methods on attributing Early English plays. In addition, WANs are shown to be reliable classifiers even when attributing collaborative plays. For several plays of disputed co-authorship, a deeper analysis is performed by attributing every act and scene separately, in which we both corroborate existing breakdowns and provide evidence of new assignments.

Correspondence:
Mark Eisen, Department of Electrical and Systems Engineering, University of Pennsylvania, Philadelphia, PA 19104, USA.
E-mail:
maeisen@seas.upenn.edu

1 Introduction

Stylometry involves the quantitative analysis of a text's linguistic features to gain further insight into its underlying elements, such as authorship or genre. Along with common uses in digital forensics (De Vel *et al.*, 2001; Stamatatos, 2009) and plagiarism detection (Meuschke and Gipp, 2013), stylometry has also become the primary method for evaluating authorship disputes in historical texts, such as the Federalist papers (Mosteller and Wallace, 1964; Holmes and Forsyth, 1995) and the Mormon scripture (Holmes, 1992), in a field called authorship attribution. Such disputes exist regarding the collection of dramatic works produced in the England during the Early Modern era, covering the 16th through mid-17th century. Due to factors such as inaccurate publication information on title pages and undocumented collaborations, the precise authorship of many of these plays—including works by William Shakespeare and John Fletcher—remains highly contested.

Stylometric analysis of the work from this time period dates as far back as the 19th century in F. G. Fleay's analysis of verse features in Shakespeare's

Digital Scholarship in the Humanities, Vol. 33, No. 3, 2018. © The Author 2017. Published by Oxford University Press on behalf of EADH. All rights reserved. For Permissions, please email: journals.permissions@oup.com
doi:10.1093/llc/fqx059 Advance Access published on 8 December 2017

plays (Fleay, 1878). Similar analyses based on the manual counting of linguistic features continued throughout the early to late 20th century (Timberlake, 1931; Oras, 1960; Tarlinskaja *et al.*, 1987). Computer-based techniques for counting the frequency of various stylistic features, such as rare words or phrases, have become very common over the past few decades. The most recent work done in evaluating authorship in Early Modern era drama includes that by MacDonald P. Jackson (Jackson, 2003, 2006), Brian Vickers (Vickers, 2002), and Hugh Craig and Arthur Kinney (Craig and Kinney, 2009), each of whom studied the works of Shakespeare and his contemporaries extensively using computational stylometry techniques.

The techniques used in modern authorship attribution began almost a century ago by examining sentence lengths in texts to determine authorship (Yule, 1939). Mosteller and Wallace (1964) were the first to consider function words as important stylistic markers in stylometric analysis, producing unprecedented results. As such, function words have continued to be common in analysis techniques (Argamon and Levitan, 2005; Juola, 2006) due to their context independence and ubiquity at high rates of occurrence in English language texts. These methods rely mainly on the frequency of usage of function words. Numerous other stylistic features have since been used in authorship attribution studies, including vocabulary richness (Holmes, 1991; Hoover, 2003) and parts of speech (Cutting *et al.*, 1992).

Our method for attributing texts, developed in (Segarra *et al.*, 2015), also measures function word usage to distinguish author styles. Rather than only considering word frequencies, however, we consider a more complex relational structure in an author's usage of function words. We construct word adjacency networks (WANs) with function words as nodes, and edges containing information regarding the use of two function words within a certain distance (measured in intervening words) from one another. We interpret each WAN as a Markov chain (MC) that assigns transition probabilities to the appearance of two function words in succession, derived from their actual occurrences in succession at varying distances within the securely attributed texts. Thus, these probabilities stand for the author's expressed preference for following one particular word with another. We can then quantify similarity between WANs by using a measure of relative entropy. MCs have previously been used in (Khmelev and Tweedie, 2001) and (Sanderson and Guenter, 2006) for the purposes of authorship attribution, though neither consider the use of function words. Results in (Segarra *et al.*, 2015) show an increase in attribution accuracy compared to frequency-based methods for general texts of English literature. In this work we perform further validation of the method's performance specifically on plays from the Early Modern period and compare this performance to that of word frequency-based methods previously used in Shakespeare attributional studies. We then employ this new technique to comment on authorship disputes concerning Early Modern English dramatic works.

We first present an overview of the construction and comparison of WANs in Section 2. We discuss in Section 3 the main playwrights used in our analysis as well as the construction of their profile networks, and in Section 4 we present a measure of similarity between profiles. As a validation of the method, in Section 5 we perform a stylometric analysis of the complete undisputed works of our six primary playwrights, followed by a comparison with existing methods in Section 5.1. We are able to demonstrate high attribution accuracy in discriminating between six candidate authors. We then examine the use of WANs in determining authorship of plays known to be written by multiple authors in collaboration. This is first done by analyzing entire plays in Section 6 and then through extensive interplay analysis of a set of particularly controversial plays in Section 7. Our results largely corroborate existing theories regarding these plays and, in some cases, propose new divisions of labor.

2 Word Adjacency Networks

When doing authorship attribution, we are given a set of candidate authors $A = \{a_1, a_2, \ldots, a_n\}$ and a set of known texts written by each of these authors, and the objective is to correctly attribute a collection of texts of unknown authorship among the authors. In (Segarra *et al.*, 2013, 2015), we propose an authorship attribution method based on function WANs. For each text, we can construct a WAN of function words. These include prepositions, conjunctions, pronouns, auxiliary verbs, and articles that convey only grammatical relationships between the so-called lexical words that carry meaning. Formally, from a given text t we construct the network $W_t = (F, Q_t)$ where $F = \{f_1, f_2, \ldots, f_f\}$ is the set of nodes composed by a collection of function words, and Q_t is a similarity measure between ordered pairs of function words.

The similarity function Q_t measures the directed co-appearance of two function words. Once we encounter a particular function word, Q_t indicates the likelihood of encountering another one in the few words following the first one. More precisely, to compute Q_t we first divide the text t into units of consecutive words (e.g. sentences, speeches) s_t^h where h ranges from 1 to the total number of units. We denote by $s_t^h(e)$ the word in the e-th position within unit h of text t. Moreover, we consider that two words in the same unit are related if they are at most $D \in \mathbb{N}$ positions apart and the relation between words decays with their position difference according to a discount factor $\alpha \in (0, 1)$. In this way, with $\mathbf{I}\{\cdot\}$denoting the indicator function, we define:

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

for all $f_i, f_j \in F$. The selection of the decay parameter α, the window size D, and the delimiting units s_t^h, in general, may vary based on the texts and authors being considered. In this work, we select $\alpha = 0.75$ and $D = 10$, determined in (Segarra *et al.*, 2015) to be generally optimal and robust parameter choices. However, because punctuation marks were often added by publishers rather than the authors themselves (Howard, 1930), and because dramatic characters do not necessarily speak in sentences, when applying our method to Early Modern plays (rather than novels), we use individual speeches (rather than clauses or sentences) as the units into which we break our texts.

We then generate a profile network $W_c = (F, Q_c)$ for every author a_c using the WANs from those texts known to have been written by the corresponding author a_c. Formally, if we denote by $T^{(c)}$ the set of texts written by author a_c, then the similarity function Q_c of the profile is computed as:

$$Q_c = \sum_{t \in T^{(c)}} Q_t \, . \tag{2}$$

The similarity function Q_c depends on the number and length of the texts written by author a_c. This is a problem since we aim to compare profiles of different authors whose canons will be of differing sizes. Thus, we apply the following normalization to the similarity measures:

$$\hat{Q}_c(f_i, f_j) = \frac{Q_c(f_i, f_j)}{\sum_j Q_c(f_i, f_j)}, \tag{3}$$

for all $f_i, f_j \in F$. In Equation (3) we assume that the combined length of the texts written by author a_c is long enough to guarantee a non-zero denominator for a given number of function words $|F|$. 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. In this way, we achieve normalized networks $P_c = (F, \hat{Q}_c)$ for each author a_c. The network P_c provides an estimate of the potentially discriminative word selection preferences of author a_c. Since the similarities out of every node sum up to 1 in the network P_c, it can be interpreted as a discrete time MC. Thus, the normalized similarity $\hat{Q}_c(f_i, f_j)$ 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 for texts written by author a_c. Similarly, we can use normalization Equation (3) to build an MC P_u for each unknown text.

To perform the attribution, we need a way of comparing the generated MCs. By construction, every MC has the same state space F, facilitating the comparison. Indeed, we use the relative entropy $H(P_1, P_2)$ as a dissimilarity measure between any two chains P_1 and P_2. The relative entropy is given by:

$$H(P_1, P_2) = \sum_{i,j} \pi(f_i) P_1(f_i, f_j) \log \frac{P_1(f_i, f_j)}{P_2(f_i, f_j)}, \tag{4}$$

where π is the limiting distribution of P_1, and we consider 0log0 to be equal to 0. From (Kesidis and Walrand, 1993), if we denote as w_1 a realization of the MC P_1, then $H(P_1, P_2)$ 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 of probabilities is 1, meaning that a given realization of P_1 has the same probability of being observed in both MCs. Thus, H is a reasonable dissimilarity measure between MCs. Utilizing Equation (4) we construct the attribution function $\hat{r}_U$ by assigning the text u to the author with the MC most similar to P_u, i.e.:

$$\hat{r}_U(u) = a_{c^*}, \text{ where } c^* = \operatorname*{argmin}_c H(P_u, P_c). \quad (5)$$

Notice that the relative entropy in Equation (5) takes an infinite value when any word-pair collocation that appears in the unknown text does not appear in the profile. In practice we compute the relative entropy in Equation (4) by summing only over the non-zero transitions in the profiles:

$$H(P_1, P_2) = \sum_{i,j|P_2(f_i,f_j)\neq 0} \pi(f_i) P_1(f_i, f_j) \log \frac{P_1(f_i, f_j)}{P_2(f_i, f_j)}. \quad (6)$$

Because the calculation of relative entropy in Equation (6) only adds relative entropy for non-zero transitions, profiles built from fewer total words will on average contain fewer non-zero transitions and will thus sum together fewer terms than larger profiles. When attributing an unknown text among profiles of differing sizes, we avoid this potential biasing for smaller profiles by summing only over transitions that are non-zero in every profile being considered:

$$H(P_1, P_2) = \sum_{\substack{i,j\,|\,P_2(f_i,f_j)\neq 0 \\ \textit{for all } a_c \in A}} \pi(f_i) P_1(f_i, f_j) \log \frac{P_1(f_i, f_j)}{P_2(f_i, f_j)}. \quad (7)$$

In the following sections, Equation (7) is used to compare the MC representations of WANs when performing attributions following rule in Equation (5).

3 Author Profiles

The stylometric analysis in this article 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. For most of the article, we focus on just six playwrights, whose life spans (in brackets) and assumed career spans are:[1]

(1) George Chapman (1559–1634), active circa 1596–1620.
(2) Christopher Marlowe (1564–93), active circa 1586–93.
(3) William Shakespeare (1564–1616), active circa 1589–1614.
(4) Ben Jonson (1572–1637), active circa 1596–1637.
(5) John Fletcher (1579–1625), active circa 1605–25.
(6) Thomas Middleton (1580–1627), active circa 1603–25.

We focus on plays for the professional public theater and disregard works commissioned for one-off and/or private performance, such as masques, entertainments, and pageants. Chapman, Marlowe, Shakespeare, Jonson, Fletcher, and Middleton are central to our analysis, since they created large and well-studied canons compared to their contemporaries.

The WAN attribution algorithm developed in (Segarra *et al.*, 2015) and briefly reviewed in Section 2 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 used in building the profile, we use all texts of sole authorship that have little or no history of authorship dispute. The full list of plays used to build the six profiles is reported in Table 1. When building profiles for a given author, we generally use the DEEP database (Farmer and Lesser, 2007) to determine texts of sole authorship. An exception to this is Middleton, for whom the 2007 Oxford Collected Works of Middleton (Taylor and Lavagnino, 2007) is taken

Table 1 Plays used to create sole-authorship canons

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

(continued)

Table 1 Continued

The Prophetess (PRO)	The Sea Voyage (SEA)
Spanish Curate (TSC)	A Very Woman (TVW)
Thomas Middleton	
Your Five Gallants (FIV)	A Game at Chess (GAC)
A Mad World My Masters (MAD)	A Chaste Maid in Cheapside (MAC)
Hengist King of Kent (HEN)	Michaelmas Term (MIC)
More Dissemblers Besides Women (DIS)	No Wit, No Help Like a Woman's (NOW)
The Phoenix (PHO)	The Puritan Widow (PUR)
The Revenger's Tragedy (REV)	The Second Maiden's Tragedy (SMT)
A Trick to Catch the Old One (TCO)	The Widow (WID)
The Witch (WTH)	Women Beware Women (BEW)
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 Bathsheba
Old Wives Tale	

as the most reliable source. 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 six plays containing 103,160 words. Shakespeare, the most prolific writer, is the undisputed sole author of 28 plays, containing 679,256 words. Due to this difference, we compute the relative entropy between the WAN of an unknown text and each profile using the size-corrected expression for relative entropy in Equation (7) rather than the original expression in Equation (6).

To prevent distortions introduced by different editions handling modernization differently—Shakespeare typically being more heavily modernized than other writers—we rely on the earliest editions available of each text in the Literature Online (LION) database (Chadwyck-Healey/ProQuest), with the exception of Shakespeare plays for which multiple early editions exist. About half of Shakespeare's plays were first published during his lifetime in single-play editions known (from their paper format) as quartos, and some of these plays went through multiple quarto editions. Seven years after his death, a collection of thirty-six of his plays was published in a book now known as the First Folio (1623), forming the foundation of his canon. Thus for many plays we have multiple quarto editions and the Folio edition to choose from, and in most cases scholars have reached no general consensus about which of these editions best reflect Shakespeare's own intentions for his works. Because the First Folio edition was manufactured by one team of workmen in one printshop over a relatively short period of time (1622–23), we choose, in those cases where there is a choice of editions to be made, the Folio text over any preceding quarto. When using original transcriptions we have to account for the fact that many words had multiple accepted spellings during the Early Modern era. In general, spelling preferences in printed editions are a poor guide to authorship because printers were free to alter spellings whenever doing so assisted in producing fully justified lines of type (Gaskell, 1972). However, the alternative spellings occur infrequently (relative to the high frequency of function words, in general) and do not affect the conclusions made by our method, and are therefore ignored. In addition, we remove speech prefixes, meaning the character name preceding each speech, to avoid cases in which character names are abbreviated to function words (such as Anne abbreviated to 'An').

The WANs for each play and author profile are built using up to 100 of the most common function words from the Early Modern period, listed in Table 2. The number of function words chosen from the full set of 100 varies for each experiment and is

Table 2 Function words used to build WANs

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

Note: Only the words designated with an *a* or *s* are used in the networks used to attribute acts and scenes, respectively.

Table 3 Relative entropy between profiles

Author	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	8.4
Jonson	4.1	7.9		11.1	6.7	5.4
Marlowe	10.1	17.4	13		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	

determined by a training process in which we measure the power of each word in helping to discriminate between the particular authors under consideration.

4 Similarity of Profiles

We compute the relative entropy between every pair of author profiles for the six authors introduced in Section 3 using Equation (7); see Table 3. 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 article, relative entropies are multiplied by 100 to scale the figures up and thereby give results that are more easily compared by eye. 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 4.7*cn* between Chapman's and Shakespeare's profiles. Note that, as Equation (7) is not symmetric, the values in the table are asymmetric, although they are similar in most cases. Thus the relative entropy between Shakespeare's and Chapman's profiles is 4.8*cn* when Shakespeare's profile is taken first and Chapman's second, but only 4.7*cn* when Chapman's profile is taken first and Shakespeare's second. This is an inevitable consequence of the comparison method's non-commutativity. In general, dissimilarities between profiles in both directions are highly correlated as can be observed in Fig. 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 3 dispel the Marlovian theory of Shakespeare authorship (Webster, 1923). If Marlowe wrote the works traditionally attributed to Shakespeare, 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 9.5*cn* in both

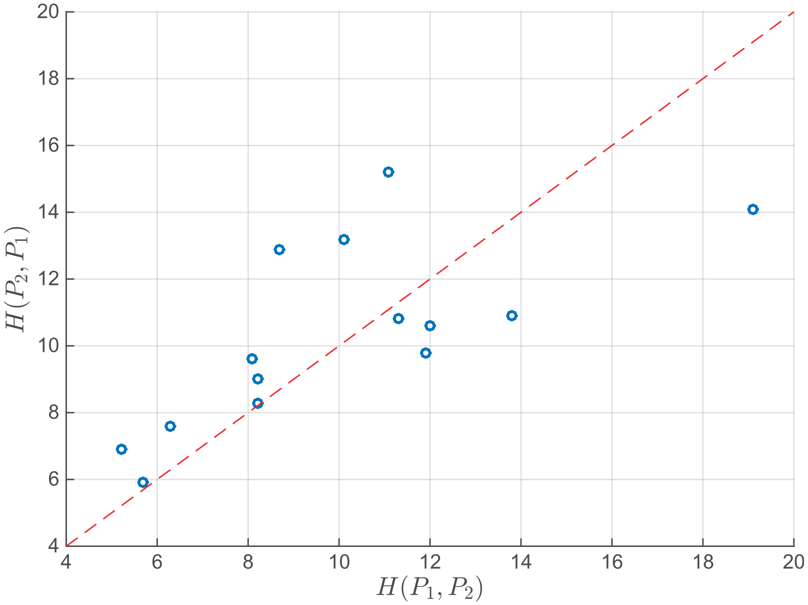

Fig. 1. Asymmetry of dissimilarities in Table 3. The coordinates of each circle represents the relative entropy between two profiles P_1 and P_2 in each direction, i.e. $H(P_1, P_2)$ versus $H(P_2, P_1)$. The dotted line marks the locations where $H(P_1, P_2) = H(P_2, P_1)$

directions 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 4.4*cn*—although still sufficiently different that we can be sure that these are not two names for the same man, as verified by the attribution of plays in Section 5. The highest dissimilarity among any pair of profiles occurs between Marlowe and Fletcher with a mean of 16.1*cn*. As will be seen in Section 5, the relative similarity between two profiles affects our ability to distinguish between them when attributing a text.

5 Attribution of Plays

As a means of validating the accuracy of the WAN method on Early Modern English dramas, we first use it to attribute the undisputed works of Jonson, Middleton, Chapman, Marlowe, Shakespeare, and Fletcher among these six authors. When attributing any given play, profiles are built using the plays listed in Table 1 excluding the one 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 a hypothetical 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 dissimilarities. 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 than to the profile of the average playwright, while a positive relative entropy indicates the opposite.

In Fig. 2 we present our method's attribution of the sixteen plays known to have been written by Ben Jonson in Table 1. In the horizontal axis we present the plays to attribute, and the vertical axis represents the relative entropy from Equation (7) in *cn* from these plays to the different profiles identified with distinct markers and discounted by the distance to

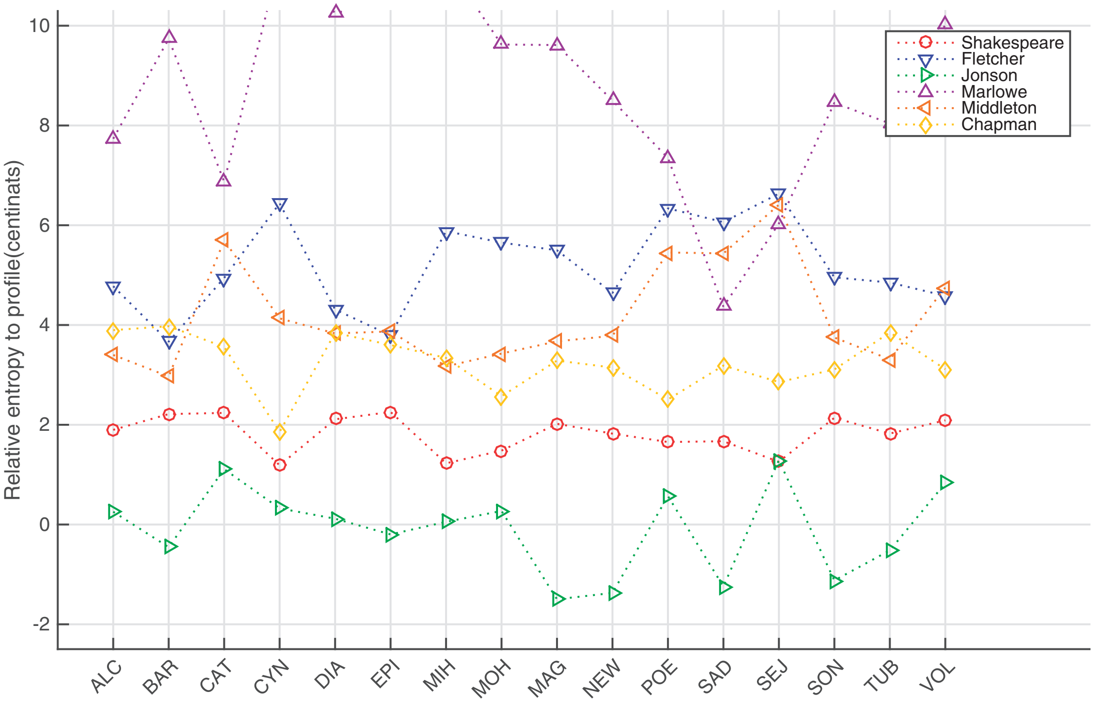

Fig. 2. Attribution of Jonson plays. The sixteen known-to-be Jonson plays in Table 1 are all correctly attributed to Jonson by our method

the average playwright. Observe that we achieve 100% accuracy in attributing the works of Ben Jonson. Note that the play *Sejanus His Fall* is virtually a tie between Jonson and Shakespeare.

In Fig. 3 we present the attribution of sixteen plays written by Thomas Middleton, with fourteen correctly assigned to Middleton by our method. 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 only history play Middleton wrote.

George Chapman is widely accepted as the author of thirteen plays listed in Table 1 and attributed by our method in Fig. 4. In total, nine of the thirteen plays are correctly attributed to Chapman by our method. Of the misattributions, three are assigned to Shakespeare with Chapman as the second nearest candidate. This is consistent with the fact that in Table 3, Chapman's profile is similar to Shakespeare, and hence they are difficult to distinguish with our method. Thus, cases of random error will most likely make our method attribute to Shakespeare plays that were in fact written by Chapman. We are fortunate, however, that theater history gives us no reason to suppose that they ever collaborated, and in practice there are no significant authorship disputes over plays that involve both of them.

In Fig. 5, we present the present method's attribution of six plays known to have been written by Christopher Marlowe. Our method achieves an accuracy of 100% in attributing Marlowe's solo works. Observe that, in the case of sole-authorship plays, each is attributed to Marlowe by a substantial margin of between 7*cn* and 13*cn*. The large negative values (between −6*cn* and −13*cn*) for the relative entropy between these plays and Marlowe's profile show that they are much closer in style to his profile than they are to the profile of an average playwright. This difference may be due in part to the fact that Marlowe's plays were written at least a decade before

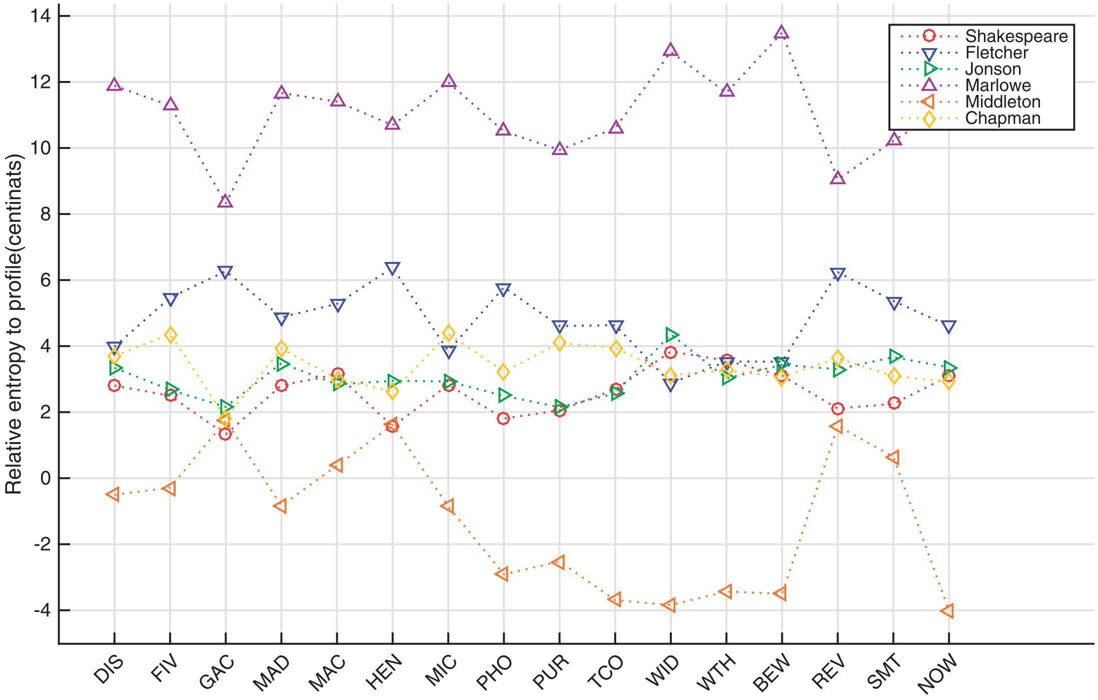

Fig. 3. Attribution of Middleton plays. Of the sixteen known-to-be Middleton plays in Table 1, two sole-authored plays are misattributed by our method

most of the other authors considered, thus possibly indicating a shift in writing style during the one or two decades that separate Marlowe from the rest.

In Fig. 6 we show the attribution of twenty-eight plays generally accepted to be written solely by William Shakespeare, and they are correctly attributed to him by our method. Strikingly, there is a well-defined distinction between plays that are widely believed to be sole-authored by Shakespeare, all having a relative entropy to Shakespeare's profile of below $-0.5cn$, and the set of plays above that threshold: *1, 2, 3 Henry VI*, *Henry VIII*, *Macbeth*, *Measure for Measure*, *Pericles*, *The Taming of the Shrew*, *Timon of Athens*, *Titus Andronicus*, and *The Two Noble Kinsmen*. Aside from *The Taming of the Shrew*, this is precisely the set of plays that multiple independent studies have recently confirmed as co-authored by Shakespeare and other writers (Taylor and Loughnane, 2017). And, indeed, independently of the present study the New Oxford Shakespeare recently presented strong reasons to suspect that *The Taming of the Shrew* is also co-authored (Taylor and Loughnane, 2017, p. 499–503). The results presented here strengthen that suspicion.

It is interesting to observe, however, an exceptional situation in the case of Marlowe. Marlowe's profile is generally very dissimilar from Shakespeare's in Table 3 and, consequently, he ranks poorly in the attribution of most of Shakespeare's plays, being consistently near the top of Fig. 6. However, the relative entropy between Marlowe's profile and the plays *Henry V*, *King John*, *Richard II*, and *Richard III* is around $+4cn$, an uncharacteristically small value compared to the rest of Shakespeare's canon. These four works are all history plays, a genre in which Marlowe wrote *Edward II* and *Massacre at Paris*, comprising a third of his profile. This suggests a potential for genre to confound attributions of authorship, although it may well be a problem confined to the particular genre of history plays (Arefin *et al.*, 2014; Taylor and Loughnane, 2017, p. 435–6).

In Fig. 7, our method attributes the fifteen plays of John Fletcher listed in Table 1. Only the play *The Faithful Shepherdess* is misattributed by our method,

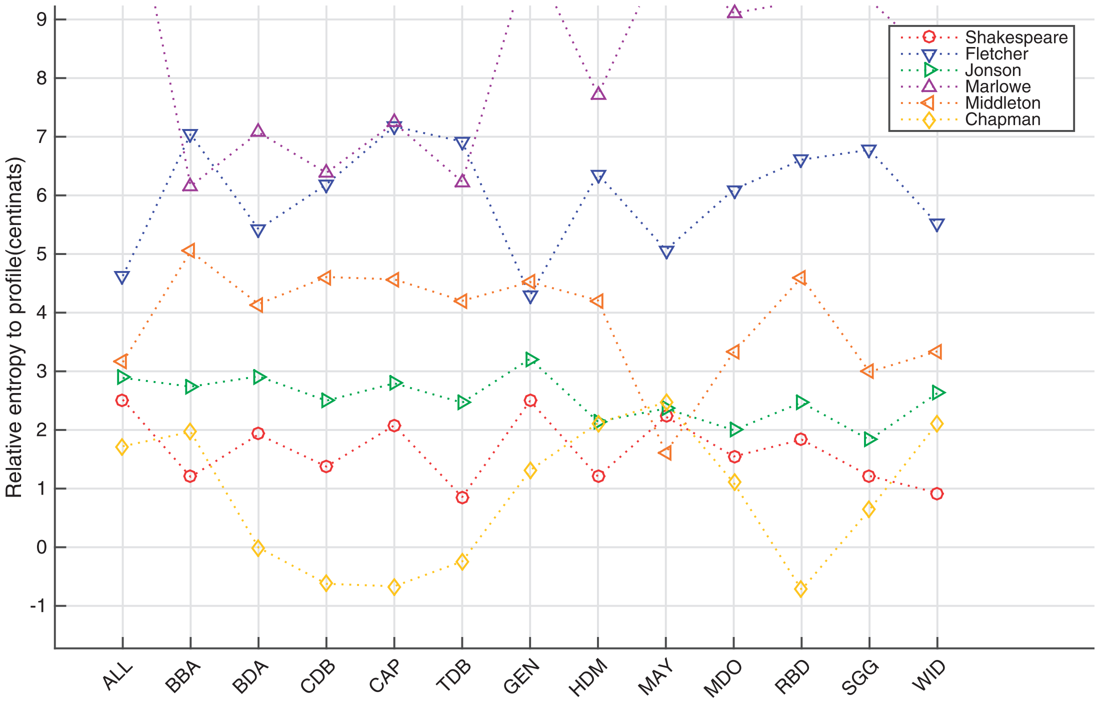

Fig. 4. Attribution of Chapman plays. Of the thirteen plays known-to-be Chapman plays in Table 1, four sole-authored plays are misattributed by our method

with Fletcher notably ranked behind Shakespeare, Chapman, and Jonson. This case is unusual in that the relative entropy between the play and Fletcher's profile is around +5*cn*, with all other Fletcher plays obtaining scores less than +1*cn*. This finding is consistent with that of Cyrus Hoy in his comprehensive study of the authorship of plays attributed to Fletcher (Hoy, 1956). Hoy concluded that although *The Faithful Shepherdess* is undoubtedly Fletcher's play, 'linguistically at least it has nothing in common with any other of his unaided works' because he wrote it in the archaic style of pastoral poetry using forms such as 'hath' and 'doth', which he 'seldom or never uses in his other unaided plays, while all the most distinguishing of his colloquial forms are either completely absent, or present in only a negligible degree' (Hoy, 1956, p. 142). Thus, Hoy writes 'Nothing could be more misleading than to regard the language of *The Faithful Shepherdess* as typically Fletcherian'. It is salutary to note that when a writer departs markedly from his usual style, he can confound studies that attribute authorship by analysis of style, but nonetheless comforting that our method corroborates a judgment made long ago by the world's leading expert on this author's style, using methods quite different from ours.

Finally, it is noticeable that by the method presented here Shakespeare is the writer to whom the greatest number of plays gets wrongly attributed, but none of his plays are wrongly attributed to someone else. This may suggest an 'averageness' to Shakespeare's writing, a phenomenon previously explored by others (Rosso *et al.*, 2009; Craig, 2011).

5.1 Comparison with existing methods

In total, our method attributes correctly eighty-seven of the ninety-four single-authored and elsewhere reliably attributed plays in Table 1, yielding an accuracy of 92.6%. Furthermore, if we consider only the attributions in which we have the greatest confidence, i.e. among authors that are more than 5*cn* apart in Table 3, then we fail only in 4, yielding an accuracy of 95.7%. To compare the power of our

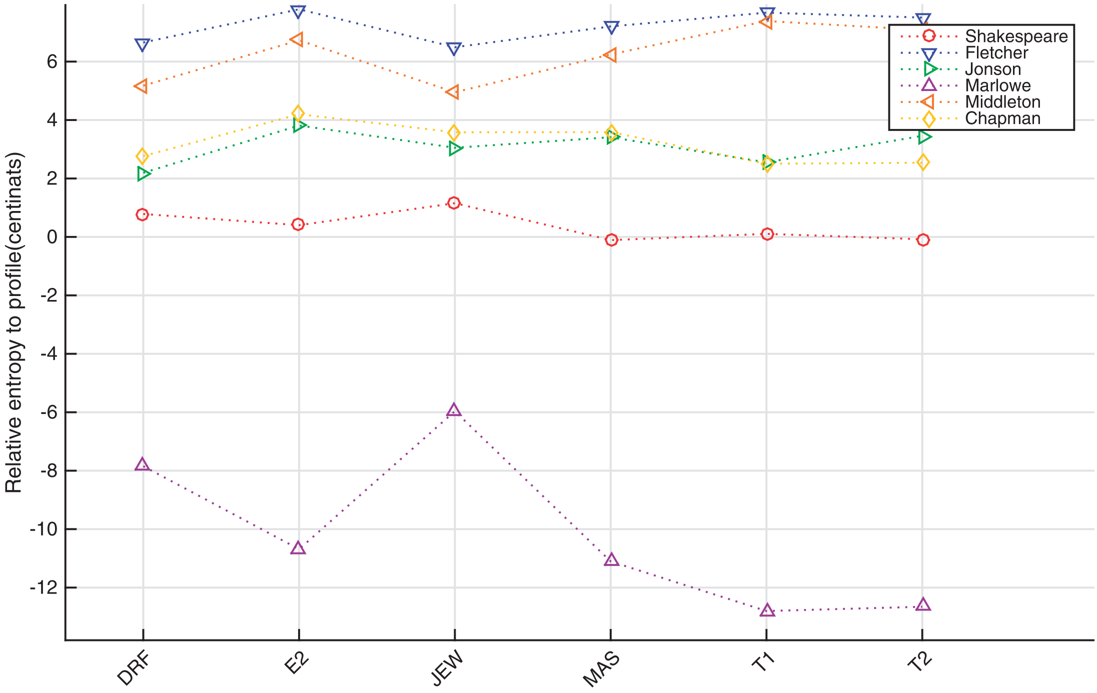

Fig. 5. Attribution of Marlowe plays. Our method correctly attributes all six known-to-be Marlowe plays in Table 1 by a large margin

method in attributing plays from this era with the power of other methods commonly used in Shakespeare attributional studies, we run the same validation tests on the ninety-four plays using two common frequency-based methods. The first, known as Burrows's Delta method (Burrows, 2002), involves measuring word frequency vectors for each play. The frequency vectors are normalized to Z-scores and compared with one another using a distance metric. The mean distance of the Z-score vector of an unknown play to the Z-score vectors of the plays of a candidate author determines the distance of a play to the candidate author. Various metrics are used, the most common being Manhattan distance, Euclidean distance, and cosine similarity. The Manhattan distance computes the sum of the absolute value of the difference in each component of the Z-score vectors. Alternatively, the Euclidean distance computes the sum of the squared difference in each component of Z-score vectors. The cosine similarity, on the other hand, computes the similarity as the cosine of the angle between the two Z-score vectors, which rises from 0 to 1, as the size of the angle between them increases from 0° (when the vectors lie on top of one another, showing maximum similarity) to 90° (when the vectors are orthogonal, showing maximum dissimilarity).

We additionally compare the accuracy of the WAN method against the principal component analysis (PCA) method used in (Craig and Kinney, 2009). Word frequency vectors are again constructed, but now reduced to principal component (PC) score vectors. This method reduces the dimensionality of the word frequency vectors to contain only the components with highest variation, known as PCs. The play is attributed to the author whose PC score vectors have the smallest mean distance to the target play's PC score vector. In this case, the method can be varied by using more or fewer PCs.

In Table 4 we compare the accuracy of the WAN method against the Delta and PCA methods when attributing whole plays among the six authors. For each method, we choose the number of function

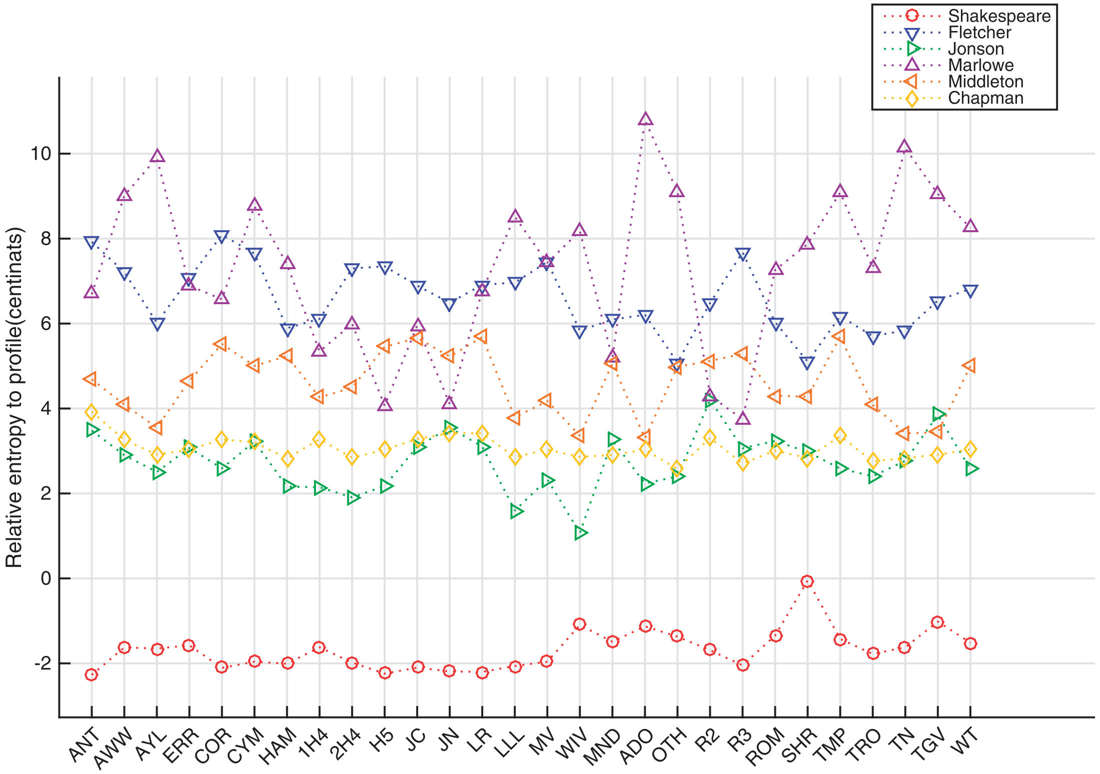

Fig. 6. Attribution of Shakespeare plays. Our method correctly attributes the twenty-eight known-to-be Shakespeare plays in Table 1. The distance between Marlowe's profile and each play is smallest for Shakespeare's history plays, suggesting an impact of genre in attribution

words that maximized attribution accuracy. Observe that the WAN method achieves an accuracy of 92.6%, outperforming five variations of the aforementioned methods. The closest competing strategy is the Delta method with Manhattan distance, which achieved an accuracy of 91.3%. All other methods achieve accuracies lower than 82%. We stress the high classification power of the WAN method for plays of sole authorship relative to other popular methods.

6 Collaborations

In cases of multiple authors contributing to a single play, we can show how our method is able to detect one or more of the authors present. We illustrate this ability in two ways: (1) by attributing collaborative plays to profiles built from other collaborations, and (2) by attributing collaborative plays to profiles built from sole-authored plays for each contributing author.

6.1 John Fletcher and collaborators

John Fletcher wrote numerous plays both by himself and with collaborators, making him a suitable case study for how our method copes with co-authorship. In addition to the six profiles built for sole-authored plays in the previous section, we now include two profiles built from plays written by Fletcher in collaboration with his two most frequent co-authors: Francis Beaumont and Phillip Massinger; see Table 1.

The attribution of Fletcher's collaborative works with Beaumont is shown in Fig. 8, while the attribution of Fletcher's collaborative works with Massinger is shown in Fig. 9. In both figures we omit the marker corresponding to Marlowe, since

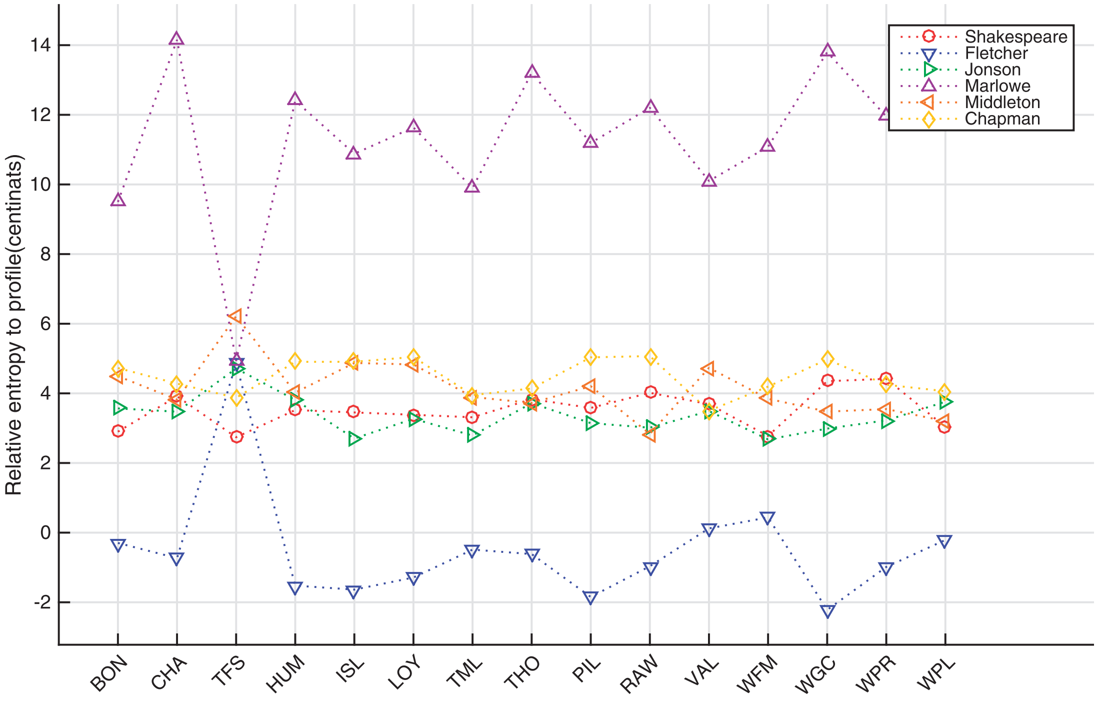

Fig. 7. Attribution of Fletcher plays. Of the fifteen known-to-be Fletcher plays in Table 1, our method misattributes only *The Faithful Shepherdess*

Table 4 Accuracies of various attribution methods on full plays between six candidate authors

Method	WAN	PCA (four PCs)	PCA (sixteen PCs)	Delta (Manhattan)	Delta (Euclidean)	Delta (Cosine)
Accuracy	92.6	72.8	81.5	91.3	79.3	81.5

he is poorly ranked for every play. This is consistent with Fletcher and Marlowe having the most dissimilar writing styles; see Table 3.

In Fig. 8, seven of the eight Fletcher and Beaumont plays are correctly attributed by our method to the Fletcher and Beaumont canon. A single mistake occurs for *Love's Pilgrimage*, although even for that play, the solo Fletcher profile and Fletcher and Massinger profile are ranked nearest and second nearest, respectively. Additionally, ten Fletcher and Massinger plays are correctly attributed by our method to the Fletcher and Massinger canon, as can be seen in Fig. 9. Observe that, in many cases, the solo Fletcher profile is ranked second behind the correct collaborative profile. These results demonstrate a case in which the WAN method is not only able to distinguish between single author in plays but is able to distinguish between an author's collaborations with two different authors. While this is not a comprehensive study of the method's discriminative power, it suggests that multiple authorial styles can be encoded in the WAN structure simultaneously.

6.2 Other collaborations

A shortcoming of the attribution method used to attribute the collaborations of Fletcher with Beaumont and Massinger is that it requires multiple collaborations between two authors, so that a reliable profile of each collaboration can be built. We also examine, therefore, the case in which we attribute a collaborative play using only profiles built

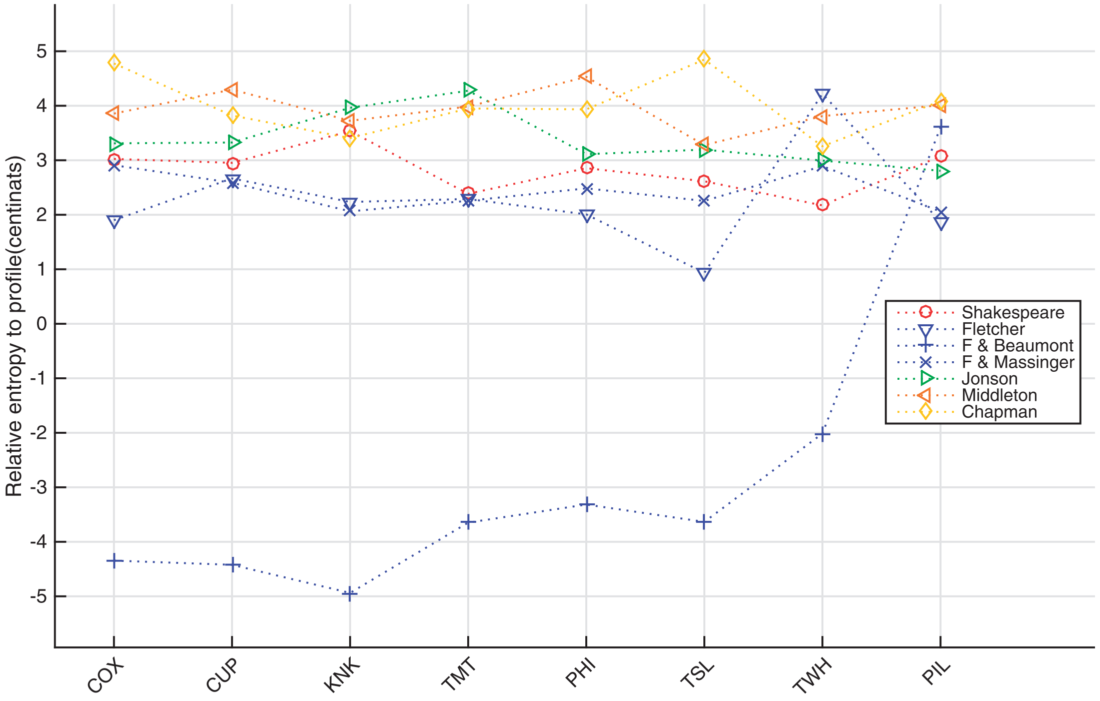

Fig. 8. Attribution of Fletcher and Beaumont plays. A single play, *Love's Pilgrimage*, is here wrongly attributed by our method to the solo Fletcher canon instead of the Fletcher and Beaumont canon where it belongs

from single authors. In Table 5 we list a set of plays with either undisputed or speculated collaboration between multiple authors previously profiled. These plays are attributed by our method by observing which of the six original author profiles are closest to each play's WAN, with the results shown in Fig. 10.

First, we attribute the play *Eastward Ho*, generally accepted as a collaboration between Jonson and Chapman and a third author, John Marston, whom we have not profiled. By our method, Jonson and Chapman are indeed ranked first and third, respectively. We also attribute two well-known collaborations between Shakespeare and 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; we cannot account for this. The attribution of Shakespeare's collaborations with Middleton, on the other hand, does not suggest the presence of both authors. While all three plays, *Measure for Measure*, *Macbeth*, and *Timon of Athens* are correctly attributed to Shakespeare, Middleton is ranked very poorly being the fourth closest candidate in all of them. This is consistent with the accepted idea that Middleton's contribution to the first two plays is minimal, but the most recent study of *Timon of Athens* attributes to him about a third of the lines, and we cannot explain his poor showing here, although there is more to be said on this topic below (Jackson, 1979; Taylor and Lavagnino, 2007, p. 467; Wells, 2009).

We also perform an attribution of a set of plays often considered to be collaborations between Shakespeare and Marlowe, though with less scholarly consensus than the previous examples. In the full play attribution of *1 Henry VI*, *2 Henry VI*, and *3 Henry VI*, however, our method shows that Shakespeare is the strongest presence with Marlowe ranked second. This is notable because Marlowe is generally ranked very poorly when

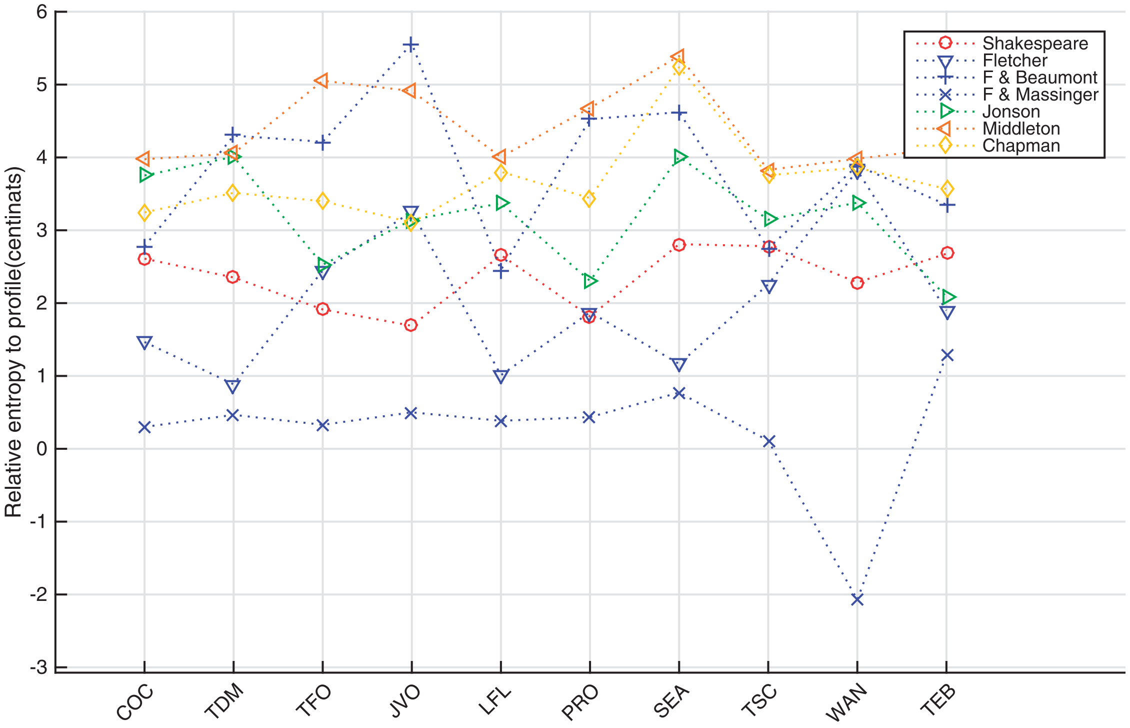

Fig. 9. Attribution of Fletcher and Massinger plays. All plays are correctly attributed to the Fletcher and Massinger canon by our method, with the solo Fletcher canon often ranked second

Table 5 Plays used in co-authorship attributions

Jonson and Chapman		Shakespeare and Fletcher	
Eastward Ho (HO)		Henry VIII (H8)	Two Noble Kinsmen (TNK)
Shakespeare and Middleton		Shakespeare and Marlowe	
Macbeth (MAC)	Measure for Measure (MEA)	1 Henry VI (1H6)	2 Henry VI (2H6)
Timon of Athens (TIM)		3 Henry VI (3H6)	Arden of Faversham (ARD)
		Edward III (E3)	

attributing Shakespeare's other works as shown in Fig. 6. While a genre bias toward Marlowe was evident with Shakespeare's other history plays, the relative closeness to Marlowe's profile in these cases is so striking that this bias cannot be the full explanation: Marlowe's hand is also definitely present here. The WAN method is used to further examine Marlowe's role in this trilogy in (Segarra *et al.*, 2016). The attribution of the anonymously published *Arden of Faversham* and *Edward III* both support theories of Shakespearean co-authorship (Vickers, 2002).

We continue by providing a more detailed study of the collaboration in these works in Section 7 by breaking each play into smaller components.

7 Collaborations—Intraplay Analysis

We examine the authorship of collaborative plays through the attribution of its individual acts and scenes. In Section 6 we attempted to detect collaborations in full plays by looking at the top candidate

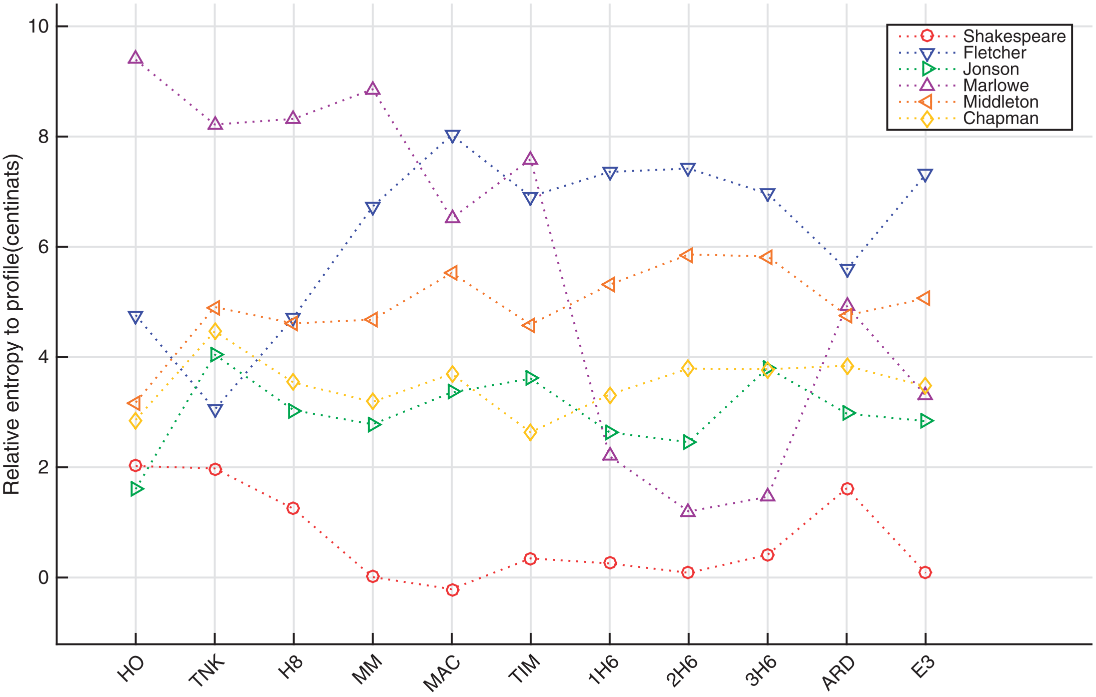

Fig. 10. Attribution by our method of collaborative plays listed in Table 5. All plays here are attributed to one of the commonly proposed contributing authors, with the other contributing author often ranking second or third

authors. This does not, however, suggest any particular breakdown of which sections of the text were contributed by which author. To pursue that topic, we divide plays into acts and scenes and attempt to attribute these individually 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 eight candidate authors: the six previously introduced plus Robert Greene and George Peele. These two additional authors were not included previously because their canons are small but are included here because they have been suggested as candidates for collaboration for some of the plays that we are considering. The plays used to build Greene's and Peele's author profiles are listed in Table 1.

We first re-train the WAN networks due to the fact that, counterintuitively, smaller WANs may increase the attribution accuracy when working with shorter texts. This is because shorter texts are less likely to contain uncommon function words. As a result, larger networks that contain these uncommon function words are more prone to overfit to features of specific texts rather than the features of the broader authorial style. We must here also point out that, due to the short length of scenes, there are fewer word transitions available for our method to characterize author style, and it thus has less distinguishing power. We therefore only seek to distinguish the more plausible of the two most commonly cited candidates when working with scenes, rather than selecting among all eight candidates as we do when working with acts. To re-train, we divide each text in Table 1 into acts and scenes and find the network size that correctly attributes the greatest number of these units to its known author. Note that, when attributing a particular act or scene of a play, the *entire* play is removed from the corresponding author profile to avoid bias.

Our training procedure finds that using the seventy-six and fifty-five most common functions

Table 6 Accuracies of various attribution methods on acts and scenes among eight and two authors, respectively

Method	WAN	PCA (four PCs)	PCA (sixteen PCs)	Delta (Manhattan)	Delta (Euclidean)	Delta (Cosine)
Accuracy (Act)	93.4	62.6	71.4	74.3	67	68.1
Accuracy (Scene)	91.5	69.1	71.5	70.1	69.3	69.8

words optimally achieve accuracies of 93.4 and 91.5% for attributing acts and scenes, respectively, of the plays in Table 1. The words used in these reduced-size networks are listed in Table 2. As a point of comparison, we include in Table 6 the attribution accuracies achieved by the Delta and PCA-based methods on the same act and scene divisions used by the WAN method. In this case, the WAN method outperforms all the other methods by significant margins for both acts and scenes. The largest accuracies achieved by the alternative methods are 74.3 and 71.5% for acts and scenes, respectively. We stress the high classification power of the WAN method relative to the other attribution schemes when attributing individual acts and scenes.

John Burrows' and Hugh Craig's investigations usually divide plays into blocks of equal size, typically of 2,000 words, and report on the likely authorship of each block. Instead, we divide plays into acts and scenes. There is almost no historical evidence for how dramatists divided up the task of collaboration (Vickers, 2002, p. 27–34), if indeed there was even a standard way, but the 'scene' was a natural unit of play construction, and its centrality to the way that dramatists conceived of their dramatic effects is thoroughly documented (Jones, 1971). The 'act' became an equally natural unit of work after 1609 when performances at outdoor as well as indoor theaters began to observe four intervals (Taylor and Jowett, 1993, p. 3–50). Before acts were formally marked in performance, the 'act' was already a conceptual unit of composition because dramatists were generally educated in the classical traditions of Greek and Roman drama, which used them. In the absence of prior knowledge of how collaborative work was divided, scene-wise, act-wise, and arbitrary-block-wise analyses are equally likely to miss the real boundaries where one writer stopped and another started.

In the following subsections, the figures display the difference in relative entropy for acts and scenes when comparing the two top candidate authors, reflected by the titles above and below the plot. The longer the bar in a particular direction, the larger the difference between the entropies of the two top candidate authors. For example, in Fig. 12, bars extending upward indicate an attribution to Shakespeare, while bars extending downward indicate an attribution to Fletcher. The attribution of acts is performed between eight candidate authors, though we only plot the distances to the two most highly ranked by our method for ease of viewing. In the attribution of scenes, on the other hand, we consider only the two authors most often cited as candidates. In many cases, the acts and scenes are attributed among the same pair of authors. Cases in which an act is attributed to a third author—because our method proposes a candidate not previously considered by most other investigators—are marked in the figure captions.

7.1 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 displays the results of the act (left) and scene (right) attribution. In the eight author comparison, every act is assigned to Jonson, with the exception of Act 3 assigned to Shakespeare. Chapman is ranked either third or fourth in all acts except Act 3 in which he is ranked second. These results are similar to the full play attribution from Fig. 10, in which Jonson was the top ranked author and Chapman ranked third. While these results on their own do not highlight Chapman's specific contribution, a look at the scene attribution between just Jonson and Chapman illuminates some of Chapman's possible contributions. Most of the play is still assigned to Jonson; however,

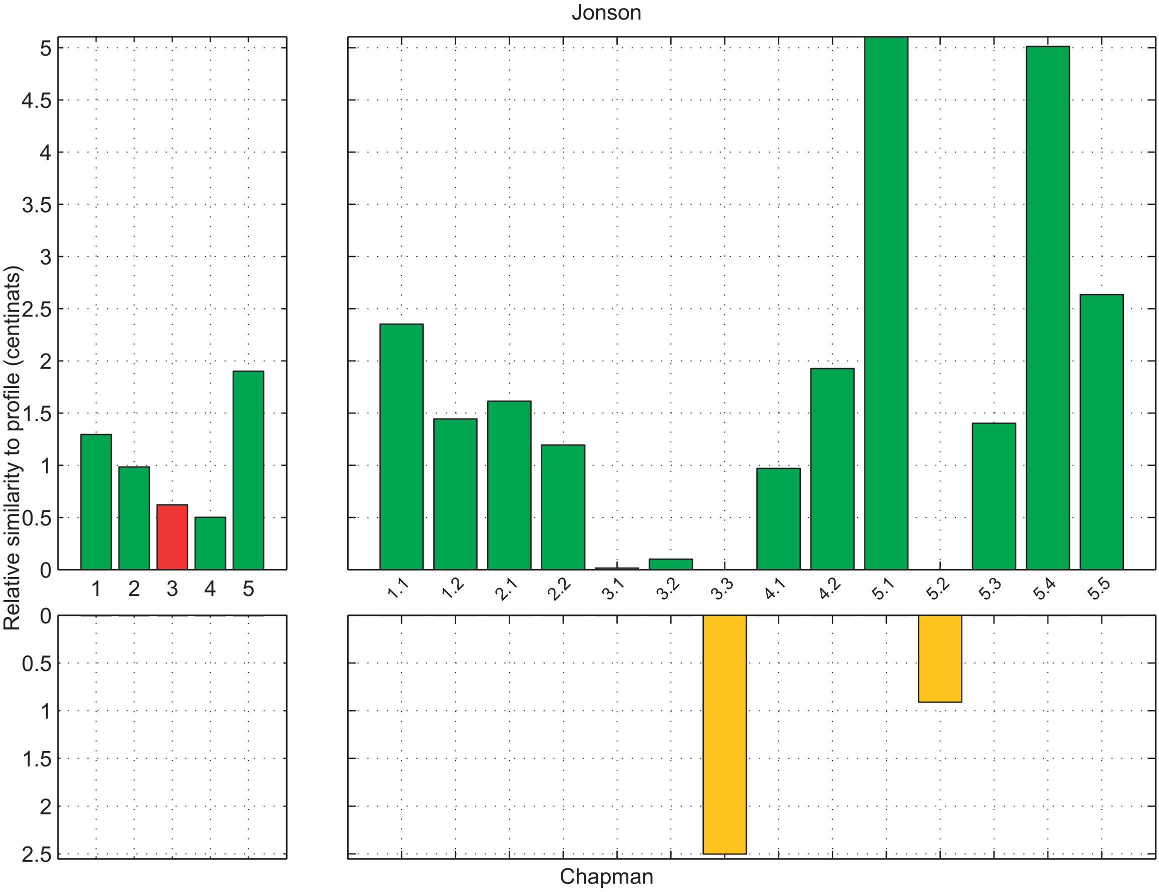

Fig. 11. Attribution of acts and scenes of *Eastward Ho*. Act 3 is assigned to Shakespeare (dark; red) over both Jonson (light; green) and Chapman. This attribution is an exception to the typical case, due to the fact that a third author (Shakespeare) is ranked first for one of the attributions; we nevertheless perform the scene-wise comparison for just Jonson and Chapman because they are the widely accepted co-authors

Chapman is seen as a more likely candidate in Scenes 3.3 and 5.2, whereas the attribution of Scenes 3.1 and 3.2 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 Acts 2 and 3, and Jonson to Act 5 (Logan and Smith, 1977). Most scholars agree in particular about Scene 3.3 being written by Chapman (Van Fossen, 1979). 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 2*cn*. We also, however, find more evidence of Jonson than Chapman contributing Acts 2 and 4.

7.2 Shakespeare and Fletcher

In Fig. 12 we show the attribution of individual acts and scenes of *Two Noble Kinsmen*, a known collaboration between Shakespeare and Fletcher. Whereas in Fig. 10 the play as a whole 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.

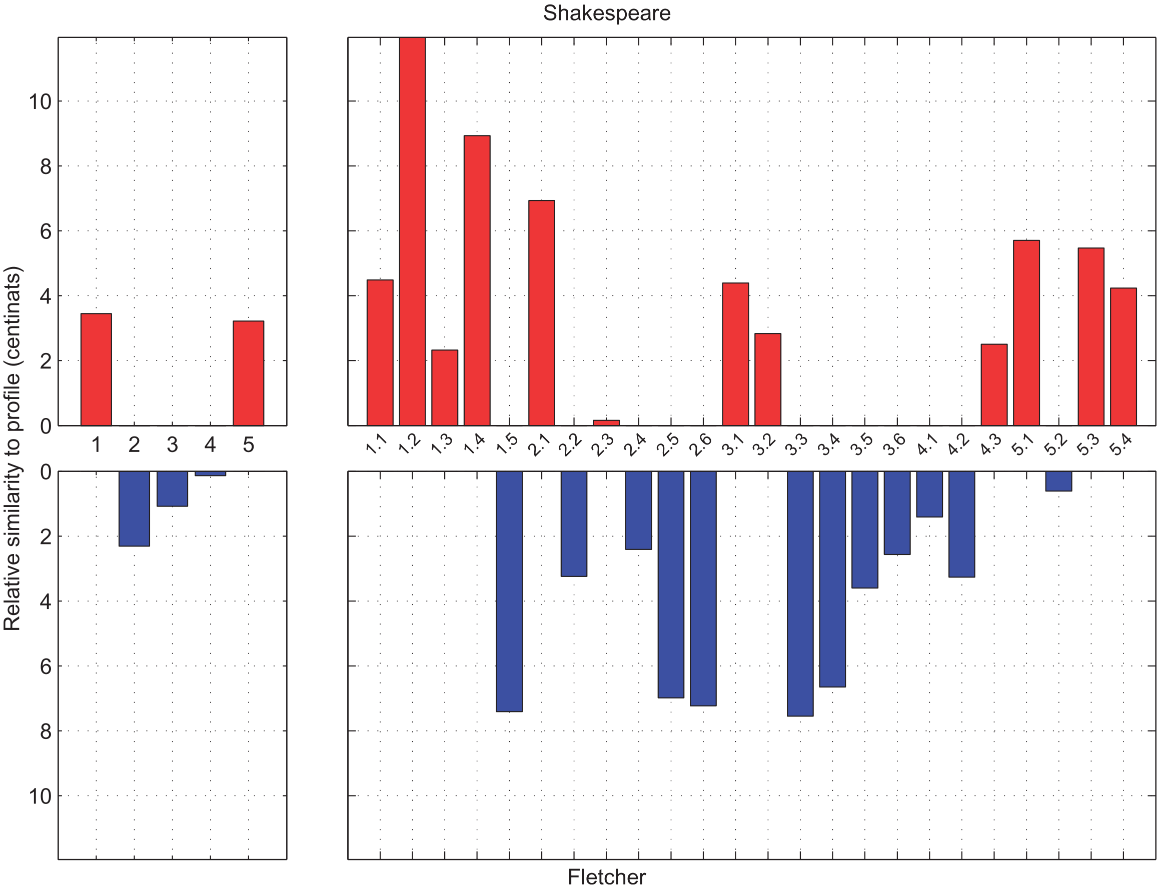

Fig. 12. Attribution of acts and scenes of *Two Noble Kinsmen* between Shakespeare and Fletcher

Act 4 is assigned to Fletcher with Shakespeare and Jonson close behind. Recall that, in Fig. 12 distances to only the two closest candidates are shown for ease of viewing. A closer look into the scene breakdown, where we consider only Shakespeare and Fletcher as candidates, reveals more specific assignments. Shakespeare is assigned Scenes 1.1, 1.2, 1.3, 1.4, 2.1, 3.1, 3.2, 4.3, 5.1, 5.3, and 5.4; Fletcher is assigned to Scenes 1.5, 2.2, 2.4, 2.5, 2.6, 3.3, 3.4, 3.5, 3.6, 4.1, and 4.2; 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* (Shakespeare *et al.*, 1974).

The act and scene analyses of Shakespeare and Fletcher's other collaboration—*Henry VIII*—are displayed in Fig. 13. Recall that, when attributing the full play, Shakespeare was the top candidate, while Fletcher was in fact ranked fourth, thus revealing no substantial evidence of collaboration; see Fig. 10. We see similar results in Fig. 13, in which Shakespeare, in an eight-author act-wise comparison, is assigned every act. Fletcher, again, is ranked poorly in every act. A scene-by-scene analysis involving just 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, 1.2, 2.1, 2.2, 2.4, 3.2, 4.1, 4.2, 5.1, and 5.2; Fletcher is assigned scenes 1.3, 1.4, 3.1, and 5.4; and 2.3 and 5.3 are too close to call—is aligned to that proposed by Cyrus Hoy (Hoy, 1960) and is similar, though not perfectly aligned, with many scholars' breakdowns. The primary area of disparity between the

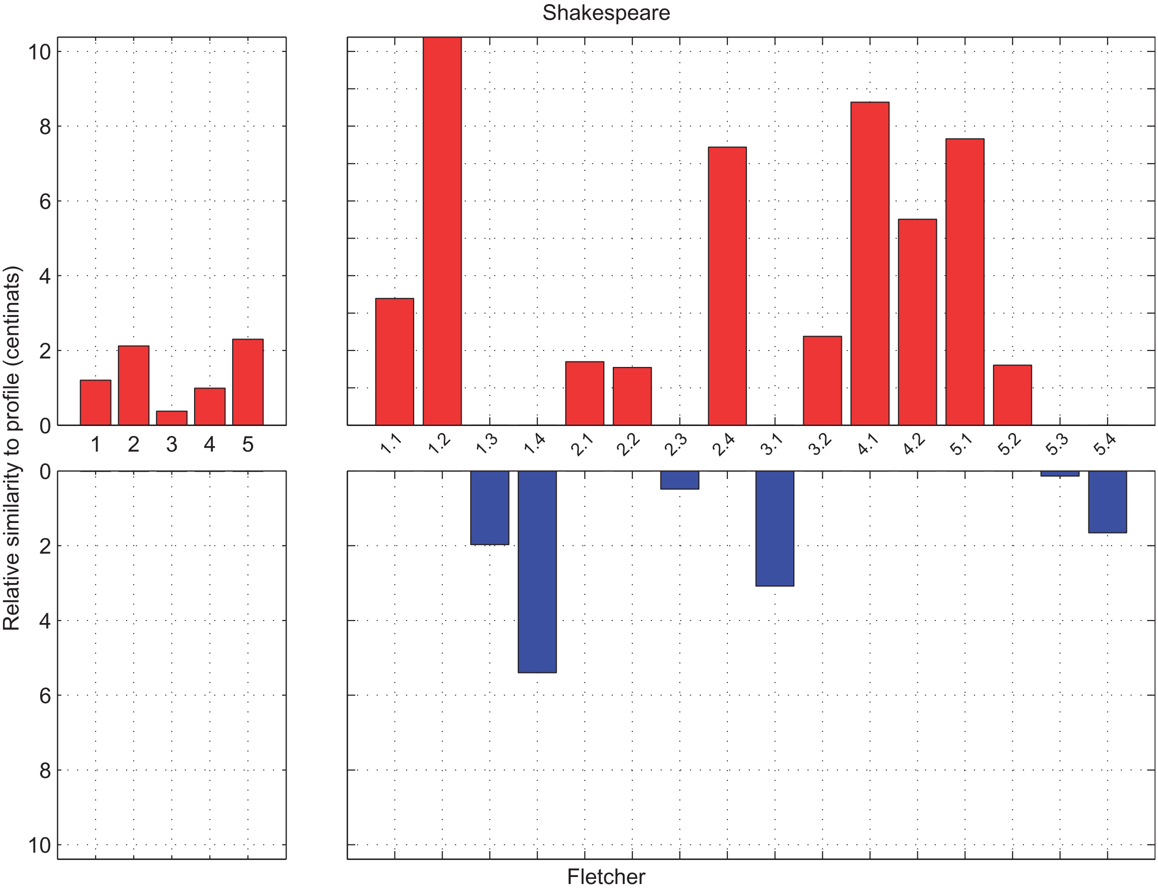

Fig. 13. Attribution of acts and scenes of *Henry VIII* between Shakespeare and Fletcher

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 in Act 4 are attributed by our method to Shakespeare by a significant margin of at least 5*cn*. Another point of contention is that we assign 2.3—given to Shakespeare by Hoy—to Fletcher by a small margin.

The attribution of *Henry VIII* shows that we may 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 we attribute to Shakespeare by a margin as wide as 7*cn*, such as Scenes 1.2, 2.4, 4.1, and 5.1. When an act contains scenes by different authors and some of the scenes have such high scores, this may tend to bias the attribution of complete acts, while the scene-by-scene analysis provides a clearer perspective.

7.3 Shakespeare and Middleton

We analyze in Figs. 14–16 Middleton's contributions to Shakespeare's plays, *Macbeth*, *Measure for Measure*, and *Timon of Athens*. The attribution of the full plays in Fig. 10 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 Fig. 14, 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 1*cn*, while Scene 1.1 is assigned by a more substantial margin of over 4*cn*. Scholars have often flagged

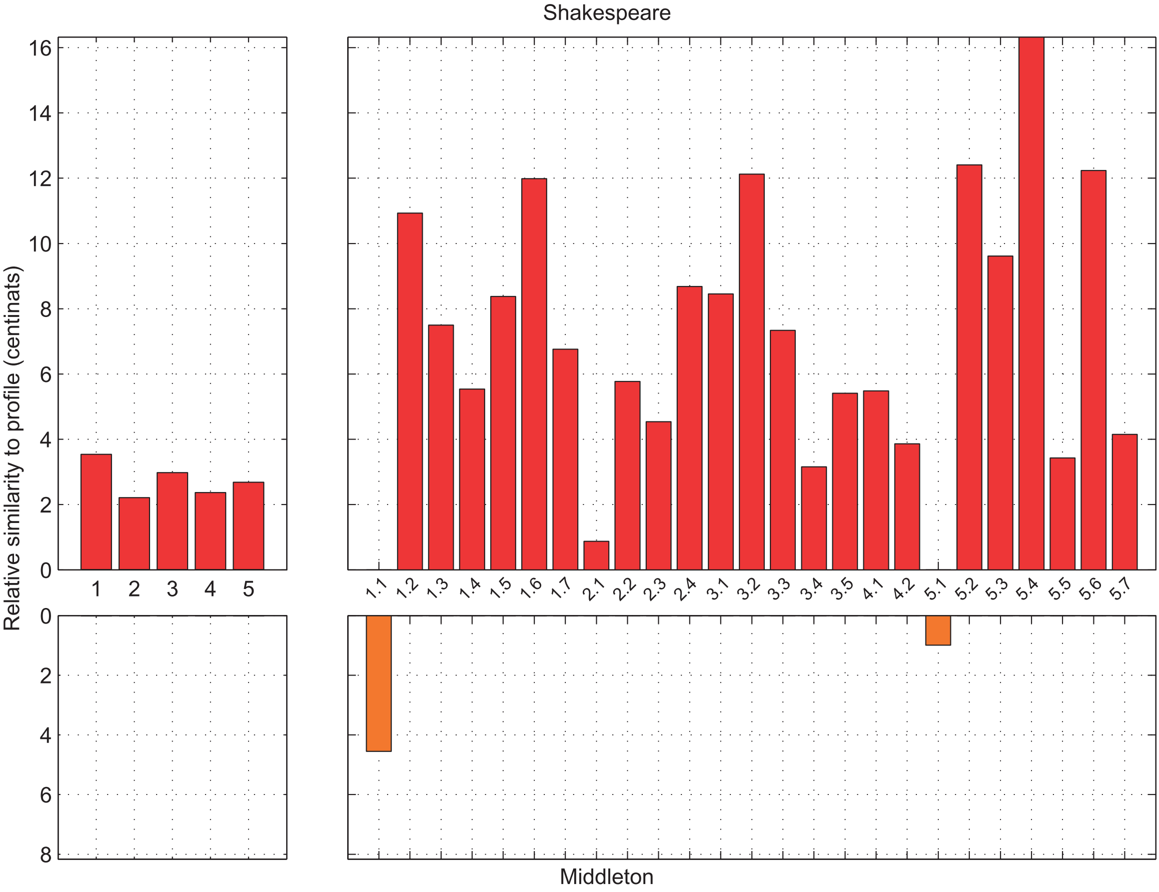

Fig. 14. Attribution of acts and scenes of *Macbeth* between Shakespeare and Middleton

Scenes 1.2, 3.5, and 4.1 as scenes revised or contributed by Middleton (Wells, 2009), although we do not find evidence of this in our analysis.

The case of *Measure for Measure* favors Shakespeare's sole authorship even more; both the act and scene analyses displayed in Fig. 15 find Shakespeare to be the sole author of the play. If Middleton revised the original play as proposed by scholars (Taylor and Jowett, 1993; Wells, 2009), we do not find evidence that it comprised substantial fresh writing.

Of the three plays, we find that Middleton's contribution was likely largest in *Timon of Athens*. While all five acts are attributed by our method to Shakespeare, in Act 3 it is by a margin of less than 1*cn* from Middleton; see Fig. 16. 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 (Vickers, 2002; Wells, 2009), and is broadly consistent with the most thorough analysis that numbers only the scenes, 1 through 19 (Taylor and Lavagnino, 2007, p. 467).

7.4 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 the traditional Shakespeare canon, the three parts of *Henry VI* are the most common points of contention. We analyze the authorship of these plays

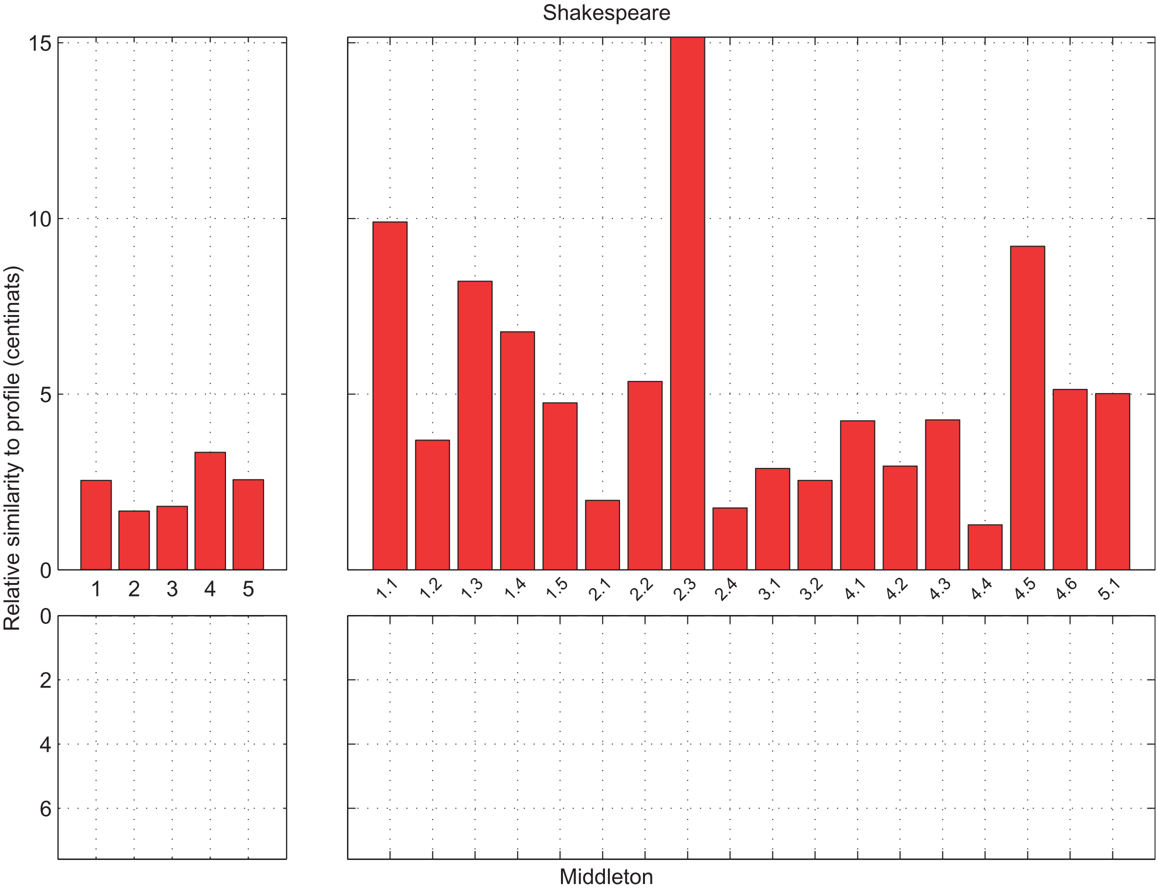

Fig. 15. Attribution of acts and scenes of *Measure for Measure* between Shakespeare and Middleton

using the WAN method in depth in (Segarra *et al.*, 2016). Here we expand this analysis to include two anonymously published plays that have previously been attributed at least in part to Shakespeare and Marlowe, namely, *Arden of Faversham* and *Edward III*.

We perform in Fig. 17 the intraplay analysis on the play *Arden of Faversham*. Every act is attributed here to Shakespeare ahead of our seven other candidates. Although not shown in the figure, the second preferred candidate in all acts except Act 5 is Jonson, who is not generally thought to have started writing plays until the late 1590s, while *Arden of Faversham* was written between 1587 and 1592 (Taylor and Loughnane, 2017, p.497–90). Jonson's immaturity (he was born in 1572) makes him an unlikely candidate unless *Arden of Faversham* was written at the end of its possible date-range. The other commonly considered candidates for authorship are Thomas Kyd and Marlowe (Greg, 1945; Craig and Kinney, 2009) The former is not profiled because his uncontested canon (comprising just *The Spanish Tragedy*) is too small 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 to Shakespeare and Marlowe rather than Shakespeare and Jonson. The scene-by-scene analysis shows Shakespeare as the more likely candidate for almost the entire play, with many scenes attributed to Shakespeare by a margin of at least 4*cn*. The exceptions to this are Scene 5.5, which is assigned to Marlowe, and Scene 5.2, a tie between candidates. Our results are consistent with existing claims by

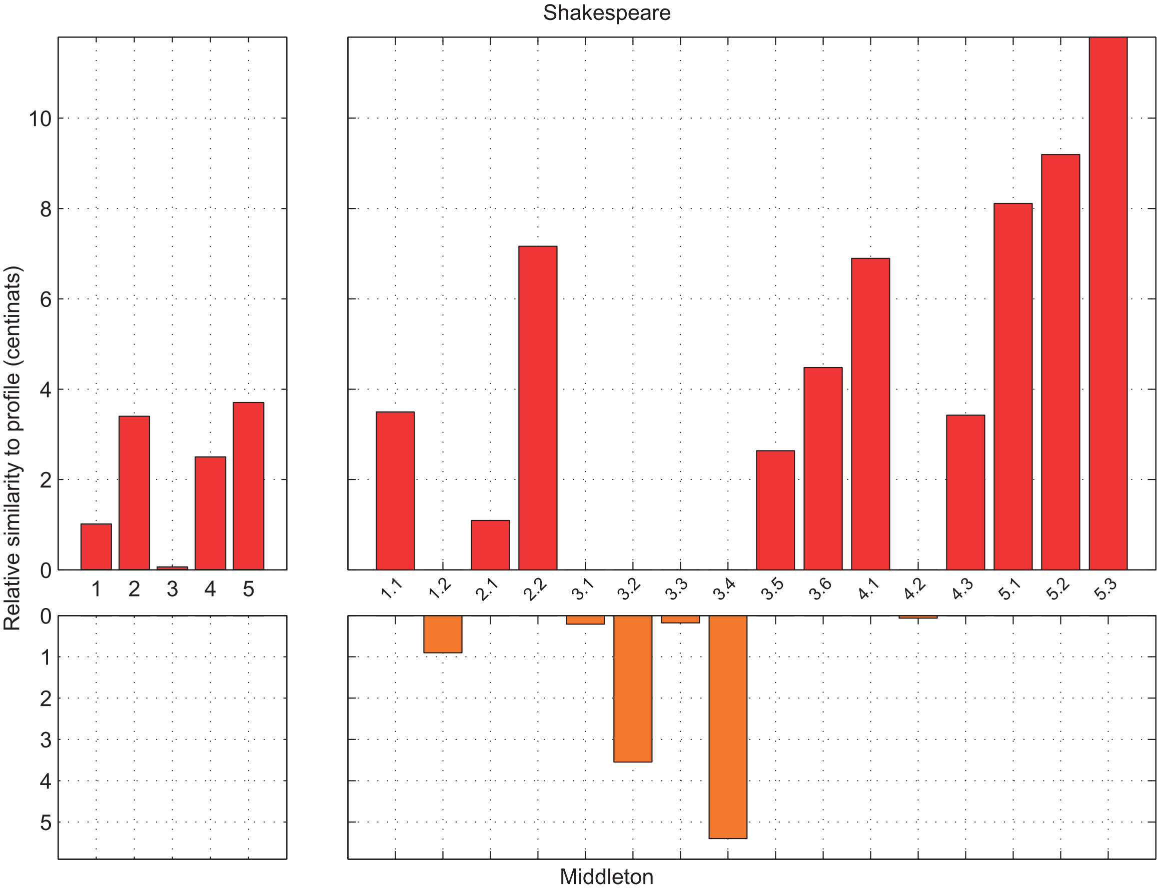

Fig. 16. Attribution of acts and scenes of *Timon of Athens* between Shakespeare and Middleton

MacDonald P. Jackson (Jackson, 2006) that Shakespeare at the very least wrote the middle of the play (Act 3), and of the candidates tested here he is the most likely to have written Acts 1, 2, and 4 as well. The writer(s) of the non-Shakespearian parts of *Arden of Faversham* may, of course, be person(s) entirely unknown to scholarship, and may include Kyd whom we are unable to test for.

An analysis is additionally performed for *Edward III*. As before, the two most commonly cited candidates for co-authorship with Shakespeare are Kyd and Marlowe (Merriam, 1993; Craig and Kinney, 2009). The eight-author act attribution of *Edward III* in Fig. 18 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 0.5*cn*. A look into the scene-by-scene attribution, however, shows that in addition to 1.1, Marlowe is also assigned Scenes 3.1, 4.1, 4.7, and 4.8, while the analysis of Scene 1.2 does not provide a clear result. While not shown in Fig. 18, the relative entropy values in the attribution of Scene 4.3 is large for both profiles—being at a distance of +1.5*cn* from Shakespeare and +7*cn* from Marlowe—suggesting that 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 (Craig and Kinney, 2009). Our results point to Shakespeare as a likely candidate for Scene 2.1, with his profile being more than 5*cn* closer than Marlow's profile to the WAN of *Edward III*. Additionally, along with Scene 4.3, we

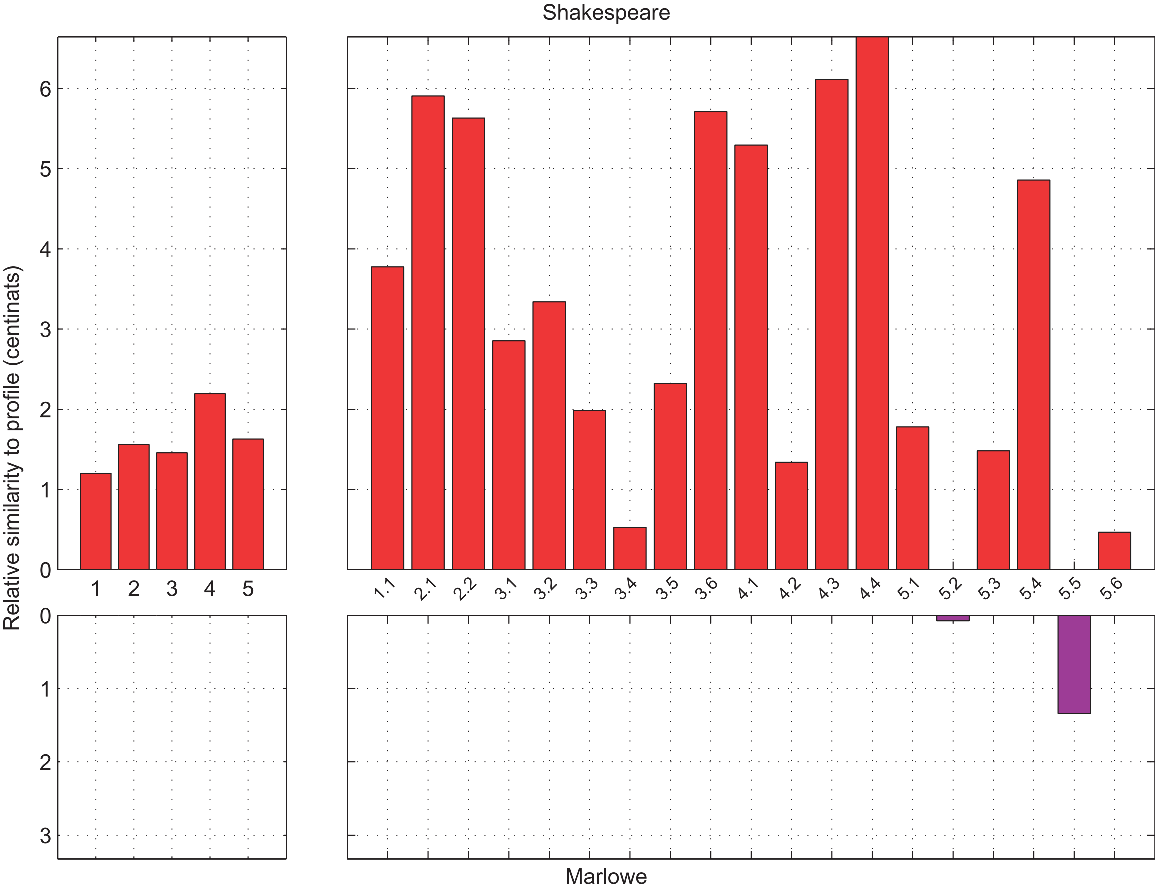

Fig. 17. Attribution of acts and scenes of *Arden of Faversham* between Shakespeare and Marlowe

find Scenes 3.2, 3.3, 4.1, 4.2, 4.5, and 4.9 to be possibly written by a third author due to comparatively large distances between the scenes' WANs and the profiles of Shakespeare and Marlowe. Indeed, although not displayed in Fig. 18, the nearest that each of these scenes comes to Shakespeare's or Marlowe's style is between +0.1*cn* and +1.7*cn*, whereas for all other scenes this distance ranges from −0.3*cn* to −3.5*cn*.

7.5 Shakespeare and Peele

Shakespeare's play, *Titus Andronicus*, is generally agreed to be co-authored with Peele (Vickers, 2002), and is attributed act-by-act and scene-by-scene in Fig. 19. 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 1*cn*. Typical attributions of this play, such as the one performed by Brian Vickers (Vickers, 2002), assign Act 1 to Peele as well as Scenes 2.1 and 4.1. Recently, William W. Weber has cast doubt on Peele's authorship of Scene 4.1 (Weber, 2014), finding strong reasons to give it to Shakespeare, and our method agrees with Weber's conclusion.

The so-called 'Fly' Scene, 3.2, is present in the 1623 Folio but not in quarto editions, suggesting that it was a later addition to the play and possibly added by another author. The relative entropies for this scene are compared in Table 7. 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. Recently,

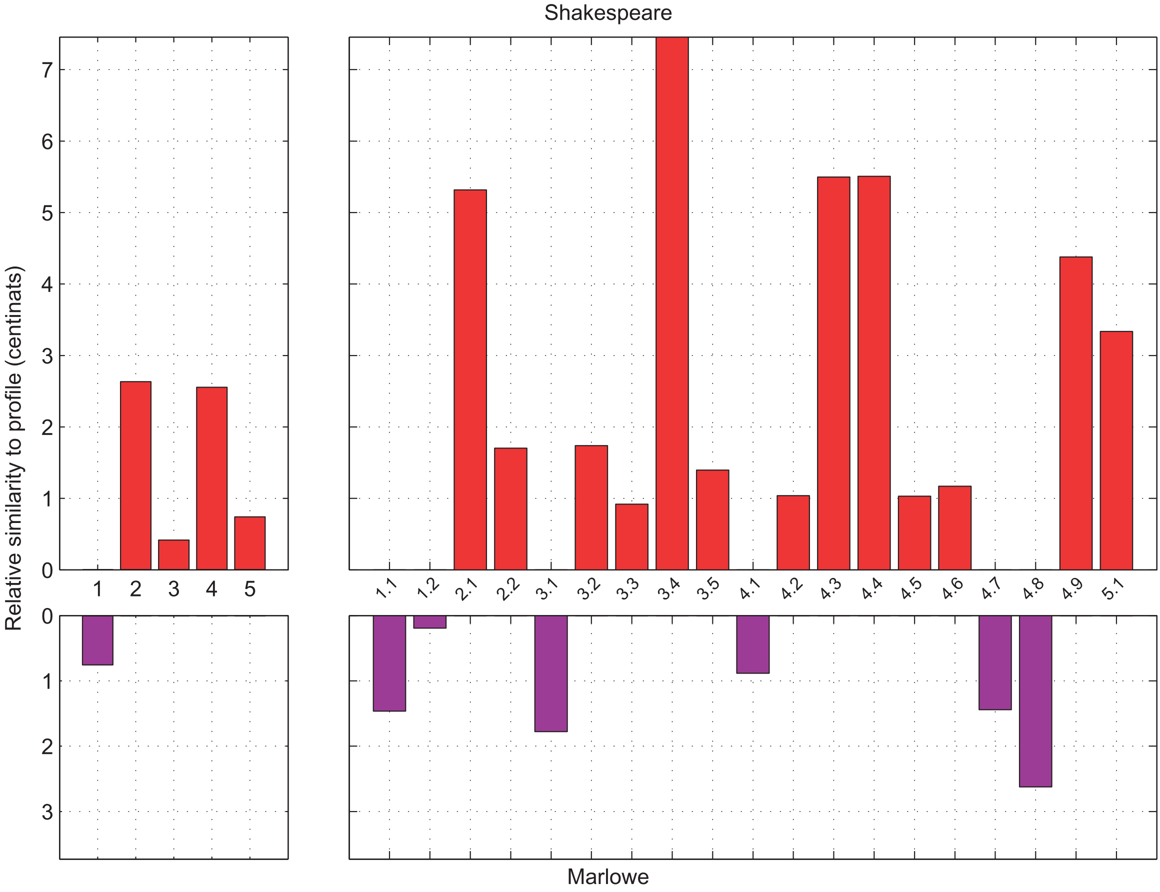

Fig. 18. Attribution of acts and scenes of *Edward III* between Shakespeare and Marlowe

Middleton has also been proposed as a candidate (Taylor and Duhaime, 2017); however, the results in Table 7 do not support this claim.

8 Conclusion

Function WANs were used to analyze the authorship of texts written by popular playwrights during the Early Modern English period. WANs were built for a large set of texts in the corpus of the analyzed authors and were compared via a measure of relative entropy. The networks of every text known to be written by a particular author were aggregated to form a profile network. The profile networks were then compared to one another to determine the general similarity between author styles. Each text in an author's corpus was compared to every profile and attributed to the author whose profile network produced the smallest relative entropy. An attribution accuracy of 92.6% was achieved when attributing amongst all authors. 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. We overall find WANs to be simple yet effective tools in distinguishing between playwrights from the Early Modern era by considering relational structures between function words not previously considered in authorship attribution

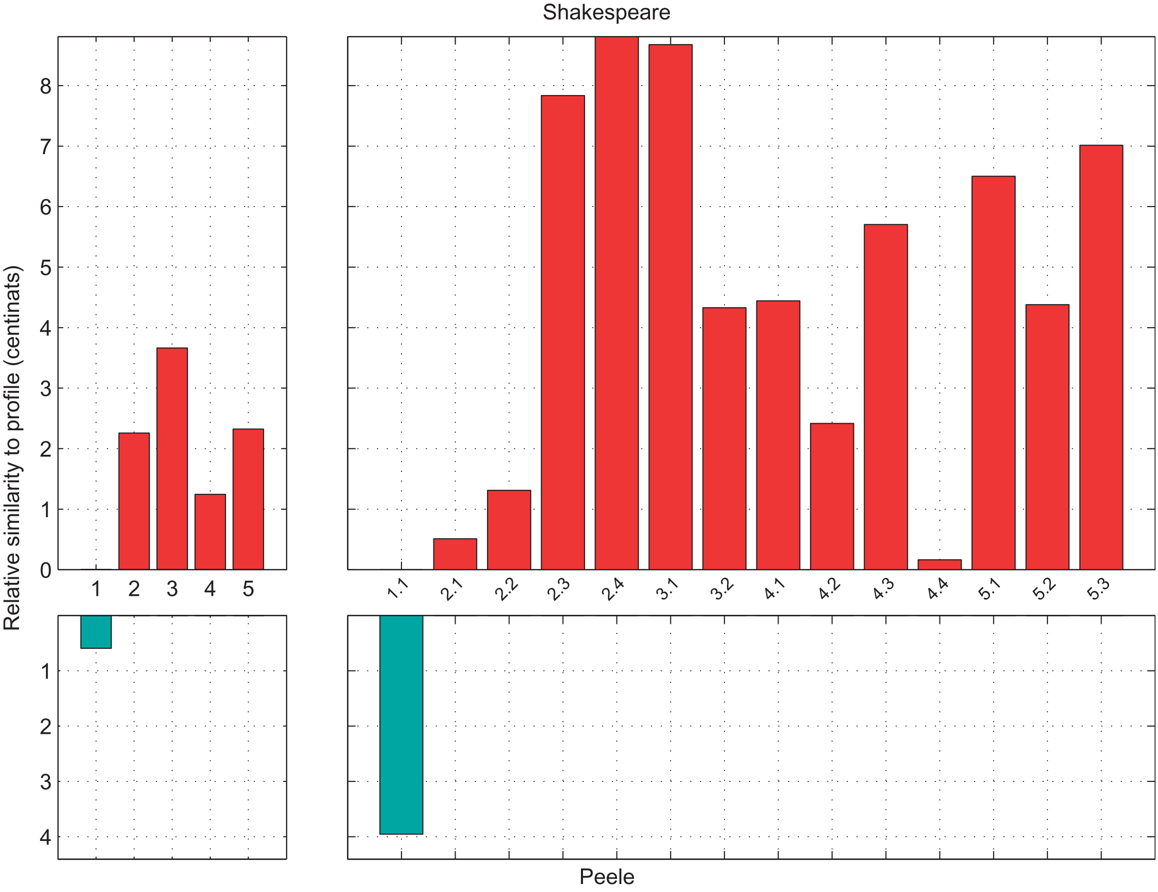

Fig. 19. Attribution of acts and scenes of *Titus Andronicus* between Shakespeare and Peele. Note that here the comparative relative entropies for Act 1 and its sole Scene, 1.1, differ. The plot of Scene 1.1 (right) reports the difference in relative entropy between Peele and Shakespeare, while the plot of Act 1 (left) reports the difference in relative entropy between Peele and Marlowe, the second ranked author.

Table 7 Relative entropies between Scene 3.2 of *Titus Andronicus* and author profiles

Shakespeare	Fletcher	Jonson	Marlowe
0.47	5.69	2.76	**0.27**
Middleton	**Chapman**	**Peele**	**Greene**
3.72	2.73	4.8	1.12

The bolded values are the two lowest entropy values.

studies from this time period. To a considerable extent, the results presented here agree with the general findings of other recent studies of Shakespeare's collaborative writing. We do not always agree on exactly which parts of the plays are by which dramatist, but we agree about which plays are the collaborative ones and that it is no longer tenable to hold the view that Shakespeare very rarely wrote his plays in collaboration with others.

References

Arefin, A. S., Vimieiro, R., Riveros, C., Craig, H., and Moscato, P. (2014). An information theoretic clustering approach for unveiling authorship affinities in Shakespearean Era plays and poems. *PLoS One*, **9**(10): e111445.

Argamon, S. and Levitan, S. (2005). Measuring the usefulness of function words for authorship attribution. In *Proceedings of the 2005 ACH/ALLC Conference.*

Association for Computing and the Humanities, Victoria, BC, 2005.

Burrows, J. (2002). 'Delta': a measure of stylistic difference and a guide to likely authorship. *Literary and Linguistic Computing*, **17**(3): 267–87.

Chadwyck-Healey/ProQuest. Literature Online. http://lion.chadwyck.com/.

Craig, H. (2011). Shakespeare's vocabulary: myth and reality. *Shakespeare Quarterly*, **62**(1): 53–74.

Craig, H. and Kinney, A. F. (2009). *Shakespeare, Computers, and the Mystery of Authorship*. Cambridge: Cambridge University Press.

Cutting, D., Kupiec, J., Pedersen, J., and Sibun, P. (1992). A practical part-of-speech tagger. In *Conference on Applied Natural Language Processing*. Association for Computational Linguistics, pp. 133–40.

De Vel, O., Anderson, A., Corney, M., and Mohay, G. (2001). Mining e-mail content for author identification forensics. *ACM Sigmod Record*, **30**(4): 55–64.

Farmer, E. A. B. and Lesser, Z. (2007). DEEP: database of early english playbooks. http://deep.sas.upenn.edu/.

Fleay, F. G. (1878). *Shakespeare Manual*. London: Macmilan.

Gaskell, P. (1972). *A New Introduction to Bibliography*. Oxford: Clarendon Press.

Greg, W. W. (1945). Shakespeare and *Arden of Faversham*. *The Review of English Studies*, **21**(82): 134–6.

Holmes, D. I. (1991). Vocabulary richness and the prophetic voice. *Literary and Linguistic Computing*, **6**(4): 259–68.

Holmes, D. I. (1992). A stylometric analysis of Mormon scripture and related texts. *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, **155**: 91–120.

Holmes, D. I. and Forsyth, R. S. (1995). The federalist revisited: new directions in authorship attribution. *Literary and Linguistic Computing*, **10**(2): 111–27.

Hoover, D. L. (2003). Another perspective on vocabulary richness. *Computers and the Humanities*, **37**(2): 151–78.

Howard, E. J. (1930). The printer and Elizabethan punctuation. *Studies in Philology*, **27**: 220–9.

Hoy, C. (1956). The shares of Fletcher and his collaborators in the Beaumont and Fletcher canon (I). *Studies in Bibliography*, **8**: 129–46.

Hoy, C. (1960). The shares of Fletcher and his collaborators in the Beaumont and Fletcher canon (V). *Studies in Bibliography*, **12**: 77–108.

Jackson, M. P. (1979). *Studies in Attribution: Middleton and Shakespeare*. Salzburg: Institut für Anglistik und Amerikanistik, Universität Salzburg Salzburg.

Jackson, M. P. (2003). *Defining Shakespeare: 'Pericles' as Test Case*. Oxford: Oxford University Press.

Jackson, M. P. (2006). Shakespeare and the quarrel scene in *Arden of Faversham*. *Shakespeare Quarterly*, **57**(3): 249–93.

Jones, E. (1971). *Scenic Form in Shakespeare*. Oxford: Clarendon Press.

Juola, P. (2006). Authorship attribution. *Foundations and Trends in Information Retrieval*, **1**(3): 233–334.

Kesidis, G. and Walrand, J. (1993). Relative entropy between Markov transition rate matrices. *IEEE Transactions on Information Theory*, **39**(3): 1056–7.

Khmelev, D. V. and Tweedie, F. J. (2001). Using Markov chains for identification of writers. *Literary and Linguistic Computing*, **16**(3): 299–307.

Logan, T. P. and Smith, D. S. (1977). *The New Intellectuals*. Lincoln: University of Nebraska Press.

Merriam, T. (1993). Marlowe's hand in *Edward III*. *Literary and Linguistic Computing*, **8**(2): 59–72.

Meuschke, N. and Gipp, B. (2013). State-of-the-art in detecting academic plagiarism. *International Journal for Educational Integrity*, **9**(1): 50–71.

Mosteller, F. and Wallace, D. (1964). *Inference and Disputed Authorship: The Federalist*. Reading: Addison-Wesley.

Oras, A. (1960). *Pause Patterns in Elizabethan and Jacobean Drama: An Experiment in Prosody*. Gainesville: University of Florida Press.

Rosso, O. A, Craig, H., and Moscato, P. (2009). Shakespeare and other english renaissance authors as characterized by information theory complexity quantifiers. *Physica A*, **388**: 916–26.

Sanderson, C. and Guenter, S. (2006). Short text authorship attribution via sequence Kernels, Markov chains and author unmasking: an investigation. In *Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, pp. 482–91.

Segarra, S., Eisen, M., Egan, G., and Ribeiro, A. (2016). Attributing the authorship of the *Henry VI* plays by word adjacency. *Shakespeare Quarterly*, **67**(2): 232–56.

Segarra, S., Eisen, M., and Ribeiro, A. (2013). Authorship attribution using function words adjacency networks. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, pp. 5563–7.

Segarra, S., Eisen, M., and Ribeiro, A. (2015). Authorship attribution through function word adjacency networks. *IEEE Transactions on Signal Processing*, **63**(20): 5464–78.

Shakespeare, W., Evans, G. B., and Tobin, J. J. M. (eds) (1974). *The Riverside Shakespeare*, vol. **1**. Boston: Houghton Mifflin.

Stamatatos, E. (2009). A survey of modern authorship attribution methods. *Journal of the American Society for Information Science and Technology*, **60**(3): 538–56.

Tarlinskaja, M., Bailey, J., and Wright, G. T. (1987). *Shakespeare's Verse: Iambic Pentameter and the Poet's Idiosyncrasies*. New York: Lang.

Taylor, G. and Duhaime, D. (2017). Who wrote the fly scene (3.2) in *Titus Andronicus*? Automated searches and deep reading. In Taylor, G. and Egan, G. (eds), *The New Oxford Shakespeare Authorship Companion*. Oxford: Oxford University Press, pp. 67–91.

Taylor, G. and Jowett, J. (1993). *Shakespeare Reshaped, 1606-1623*. Cambridge: Cambridge University Press.

Taylor, G. and Lavagnino, J. (2007). *Thomas Middleton: The Collected Works*. Oxford: Oxford University Press.

Taylor, G. and Loughnane, R. (2017). The canon and chronology of Shakespeare's works. In Taylor, G. and Egan, G. (eds), *The New Oxford Shakespeare Authorship Companion*. Oxford: Oxford University Press, pp. 417–602.

Timberlake, P. W. (1931). *The Feminine Ending in English Blank Verse*. Menasha: George Banta.

Van Fossen, R. W. (1979). *Eastward Ho*. In Chapman, G., Jonson, B., and Marston, J. (eds). Manchester: Manchester University Press.

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

Weber, W. W. (2014). Shakespeare after all? The authorship of *Titus Andronicus* 4.1 reconsidered. In Holland, P. (ed.), *Shakespeare Survey, Vol. 67: Shakespeare's Collaborative Work*. Cambridge: Cambridge University Press, pp. 69–84.

Webster, A. (1923). Was Marlowe the Man? *National Review*, 81–6.

Wells, S. (2009). *Shakespeare and Co.: Christopher Marlowe, Thomas Dekker, Ben Jonson, Thomas Middleton, John Fletcher and the Other Players in His Story*. New York: Random House LLC.

Yule, G. U. (1939). On sentence-length as a statistical characteristic of style in prose: with application to two cases of disputed authorship. *Biometrika*, **30**: 363–90.

Note

[1] Information compiled from the Database of Early English Playbooks (DEEP) (Farmer and Lesser, 2007) and the database of catalogued plays in LION (Chadwyck-Healey/ProQuest). Whenever inconsistencies in authorship information arise, we accept the verdicts in (Farmer and Lesser, 2007) unless compelling new research contradicts them.