Stylometric Analysis of Early Modern Period English Plays

Santiago Segarra,[1], Mark Eisen[1], Gabriel Egan[2], and Alejandro Ribeiro[1]

[1]*Dept. of Electrical and Systems Engineering,, University of Pennsylvania, Philadelphia, USA*
[2]*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 likelihood of ordered 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. This high classification power is then used to investigate the authorship of anonymous plays. Moreover, 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. 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.**

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

I. Introduction

Stylometry involves the quantitative analysis of a text's linguistic features in order to gain further insight into its underlying elements, such as authorship or genre. Along with common uses in digital forensics [1], [2] and plagiarism detection [3], stylometry has also become the primary method for evaluating authorship disputes in historical texts, such as the Federalist papers [4], [5] and Mormon scripture [6], in a field called authorship attribution. One of the most notorious collection of texts for which such disputes exist is the collection of dramatic works produced in England during the Early Modern era, covering the 16th through mid-17th century. Due to factors such as inaccurate pressing information and undocumented collaborations, the precise authorship of many of these plays–including works by famous playwrights William Shakespeare and John Fletcher–remains highly contested.

Stylometric analysis of the work from this time period dates as far back as the nineteenth century in F. G. Fleay's analysis of verse features in Shakespeare's plays [7]. Similar analyses based on the manual counting of linguistic features continued throughout the early twentieth century [8]–[10]. 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 premier work done in evaluating authorship in Early Modern era drama includes the work of MacDonald P. Jackson [11], [12], Brian Vickers [13], and Hugh Craig and Arthur Kinney [14], each of whom studied the works of Shakespeare and his contemporaries extensively using computational stylometry techniques.

The techniques used in authorship attribution began almost a century ago by examining sentence lengths in texts to determine authorship [15]. In [4], stylometric analysis first began to consider function words as important stylistic markers, producing unprecedent results. As such, function words have continued to be common in analysis techniques [16], [17] due to their context independence and universal usage 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 the study of vocabulary richness [18], [19] and the use of part of speech taggers [20].

Our method for attributing texts, developed in [21], also measures function word usage to distinguish author styles. Rather than only considering word frequencies, however, we consider a more complex relational structure between 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 the same sentence or phrase. We interpret each WAN as a Markov chain that assigns transition probabilities between the appearance of two function words. We can then measure similarity between WANs by using a measure of relative entropy. Markov chains have previously been used in [22] and [23] for the purposes of authorship attribution, though neither consider the use of function words. Results in [21] show an increase in attribution accuracy compared to the most common frequency-based methods. We employ this new technique then to further analyze and add insight into the authorship disputes of Early Modern English dramatic works.

We first present an overview of the construction and comparison of WANs in Section II. We discuss in Section III the main playwrights used in our analysis and the construction of their profile networks as well as a measure of similarity between profiles in Section IV. In Sections V-A through V-E we perform a stylometric analysis of the complete canons of our five primary playwrights, followed by a summary of results in Section V-F. We are able to demonstrate high attribution accuracy between six candidate authors. An analysis of a set

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

of plays published anonymously or without a clear author is performed in Section VI. 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 VII and then through extensive interplay analysis of a set of particularly controversial plays in Section VIII. Our results largely corroborate existing theories regarding these plays as well as, in some cases, propose new authorship breakdowns. We conclude in Section IX 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.

II. Word Adjacency Networks

When doing authorship attribution, we are given a set of candidate authors A and a set of known texts written by these authors T, and the objective is to correctly attribute a collection of texts U of unknown authorship among the authors in A. More precisely, the idea is to construct a function $\hat{r}_U : U \to A$ relating every text in U with its rightful author in A. In [21], [24], we propose an authorship attribution method based on function word adjacency networks. For each text, a word adjacency network (WAN) of function words, i.e. words that convey only grammatical relationships, can be constructed. 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 and $Q_t : F \times F \to \mathbb{R}_+$ is a similarity measure between ordered pairs of function words.

The similarity function Q_t measures the directed co-appearance of two function words. I.e., 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 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. Moreover, we consider that two words in the same sentence 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, we define

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

for all $f_i, f_j \in F$.

We then generate a profile network $W_c = (F, Q_c)$ for every author $a_c \in A$ using the WANs from those texts in T known to have been written by the corresponding author a_c. Formally, if we denote by $T^{(c)}$ the subset of T written by 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 amount and length of the texts written by author a_c. This is a problem since we aim to compare profiles of different authors. 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 (3) we assume that the total texts written by author a_c are long enough to guarantee a non zero denominator for a given amount 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 $\hat{P}_c = (F, \hat{Q}_c)$ for each author a_c. The network $\hat{P}_c$ estimates an ideal network P_c which captures the stylistic fingerprint of author a_c. Since the similarities out of every node sum up to 1 in the network $\hat{P}_c$, it can be interpreted as a discrete time Markov chain (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 (3) to build a MC P_u for each unknown text $u \in U$.

In order 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 $0 \log 0$ to be equal to 0. From [25], 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 (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.

$$\hat{r}_U(u) = a_p, \text{ where } p = \operatorname*{argmin}_c H(P_u, \hat{P}_c). \tag{5}$$

Notice that the relative entropy in (5) 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 (4) by summing only over the non-zero transitions in the profiles,

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

Because the calculation of relative entropy in (6) 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,

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

In the following sections, expression (7) is used to compare the Markov chain representations of WANs when performing attributions following rule (5).

III. Author Profiles

The stylometric analysis in this paper 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 paper, 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[1]:

(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 [21] and briefly reviewed in Section II 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 I. When building profiles for a given author, we generally subscribe to the information provided in [26] 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 [28], which contains a more recent and accepted study of his canon. Two plays included in Middleton's corpus in [28]–*The Revenger's Tragedy* and *The Second Maiden's Tragedy*–were published anonymously and have long history of disputed authorship [29]–[32]. To be safe, we do not include these plays in Middleton's profile but provide an analysis of their authorship in Section VI.

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 (7) rather than (6).

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 the LION database [27], 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 [33]. 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 [21], $\alpha = 0.75$ and $D = 10$. Because punctuation marks were often added by publishers rather than the authors themselves [34], 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 II. This number is chosen based on a training period to find the optimal number of function words in which we attribute all texts with undisputed authorship, i.e. those plays listed in Table I. A list of the most common Early Modern period alternative spellings is given in Table III. 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.

[1]Information compiled from the Database of Early English Playbooks (DEEP) [26] and the database of catalogued plays in Chadwyck-Healey Literature Online (LION) [27]. Whenever inconsistencies in authorship information arise, we consider [26] as accurate.

TABLE I: Plays used to build author profiles

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)
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)	
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)	
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)
A Trick to Catch the Old One (TCO)	The Widow (WID)
The Witch (WTH)	Women Beware Women (BEW)

IV. Similarity of Profiles

We compute the relative entropy between every pair of author profiles for the six authors introduced in Section III using expression (7); see Table IV. 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 paper, 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 (7) is not symmetric, the values in the table are also not symmetric, although they are similar in most cases. E.g.

TABLE II: 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 III: 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

TABLE IV: 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	

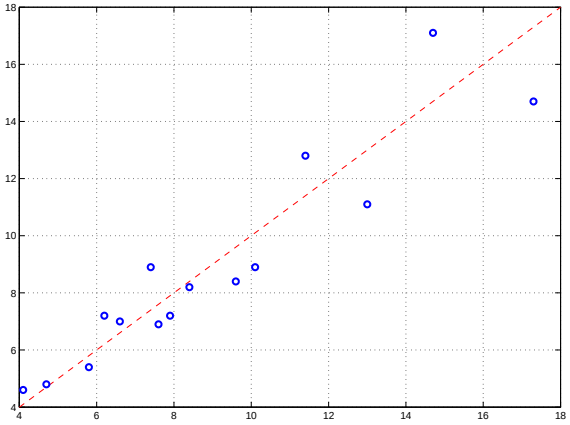

Fig. 1: Asymmetry of dissimilarities in Table IV.

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. 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 IV dispel the Marlovian theory of Shakespeare authorship [35]. 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 $9.5cn$ in both 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 $4.4cn$ – although still sufficiently different so as to assert that they belong to different authors, as verified by the attribution of plays in Section V. The highest dissimilarity among any pair of profiles occurs between Marlowe and Fletcher with a mean of $16.1cn$. As will be seen in Section V, the relative similarity between two profiles affects our ability to distinguish between them when attributing a text.

V. Attribution of Plays

We attribute the plays written by Jonson, Middleton, Chapman, Marlowe, and Shakespeare among the 6 author profiles introduced in Section III. The attribution of Fletcher's plays is performed in the discussion of collaborations in Section VII. When attributing any given play, profiles are built using the plays listed in Table I 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 than to the profile of the average playwright while a positive relative entropy indicates the opposite.

A. Ben Jonson

In Fig. 2 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 (7) 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% 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 VIII-A.

The misattribution in Fig. 2 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 [36]. The play's content has also been compared to *A Comedy of Errors*, written by Shakespeare, who is also here the closest author [37]. 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 [38], [39]. 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 Section IX.

TABLE V: Thomas Middleton plays to be attributed in addition to those listed in Table I.

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)	

B. Thomas Middleton

In Fig. 3 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 [28].

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 [28] 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

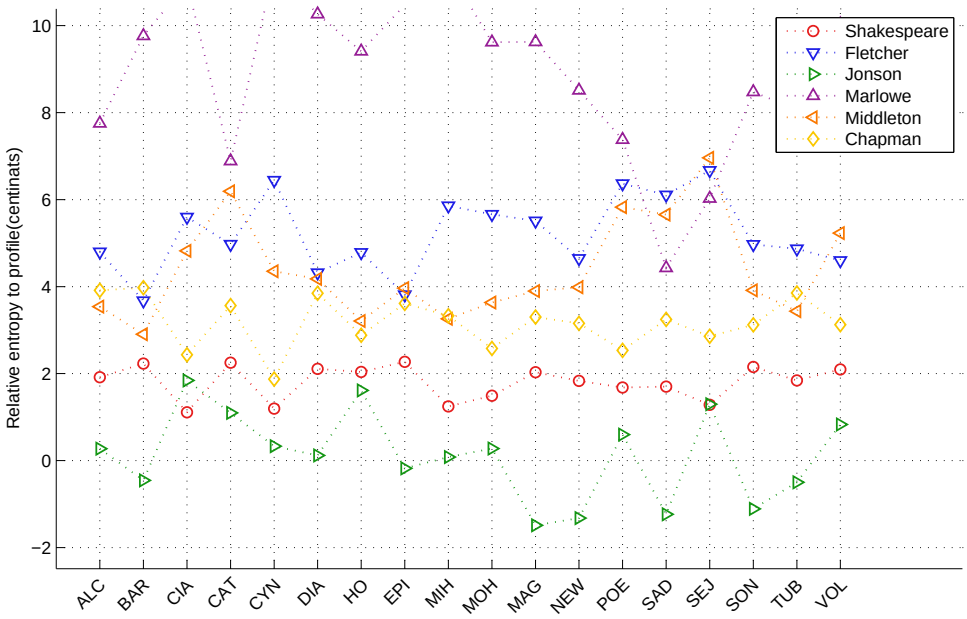

Fig. 2: Attribution of Jonson plays. We attribute the 16 plays in Table I plus *The Case is Altered (CIA)* and *Eastward Ho (HO)*. A single misattribution occurs for *The Case is Altered*.

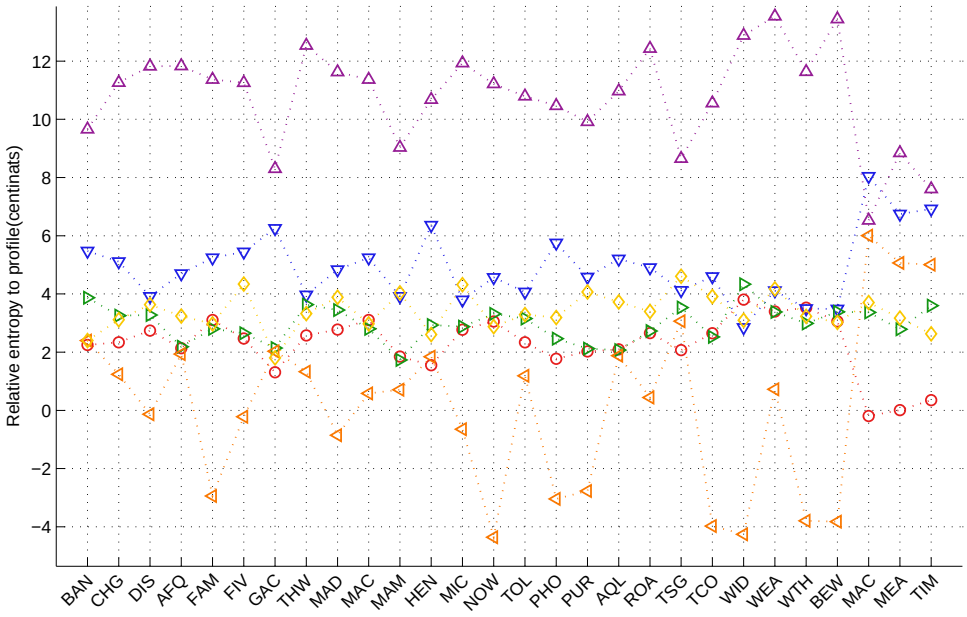

Fig. 3: Attribution of Middleton plays. We attribute the 14 plays in Table I, the additional 11 plays in Table V and 3 collaborations with Shakespeare: *Macbeth (MAC)*, *Measure for Measure (MEA)* and *Timon of Athens (TIM)*. Only 2 sole authored plays are misattributed. Also, the attribution of Shakespeare's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

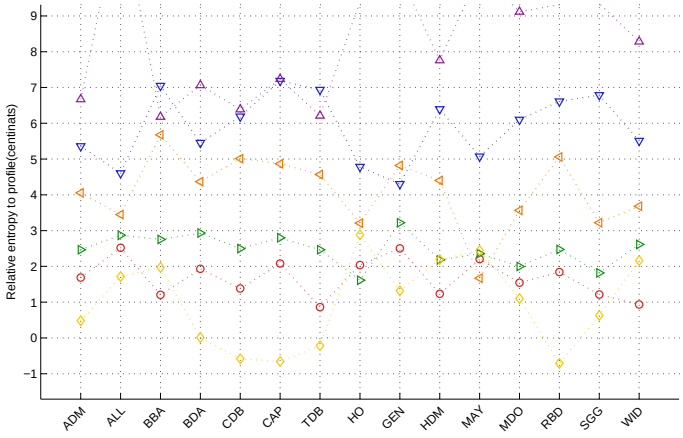

Fig. 4: Attribution of Chapman plays. We attribute the 13 plays in Table I plus *The Tragedy of Chabot Admiral of France (ADM)* and *Eastward Ho (HO)*. Out of 15 plays, 10 are attributed to Chapman. Collaboration with Jonson in *Eastward Ho* can also be observed.

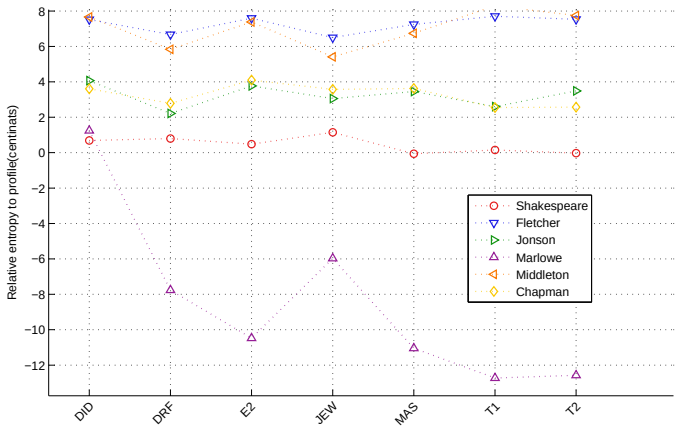

Fig. 5: Attribution of Marlowe plays. We attribute the 6 plays in Table I plus *Dido Queen of Carthage (DID)*. A single misattribution occurs for the collaborative play *Dido Queen of Carthage*.

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 [40].

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 [40], [41] 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 [42], [43]. We examine these plays in closer detail in Section VIII-C.

C. George Chapman

Chapman is considered to be the author of 15 plays, 13 as a sole author and 2 in collaboration. In Fig. 4, we attribute these plays among the 6 profiles. In total, 10 of the 15 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 V-A. 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 IV, Chapman's profile is most similar to Shakespeare. Thus, cases of random error will therefore most likely attribute to Shakespeare.

TABLE VI: William Shakespeare plays to be attributed in addition to those listed in Table I.

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)

D. Christopher Marlowe

In Fig. 5, 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 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.

E. William Shakespeare

In Fig. 6 we present the attribution of 38 plays believed to have been written by Shakespeare, 30 of which are attributed solely to Shakespeare in [26]. 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 I because they have a strong history of disputed authorship [14].

All of the 30 plays usually considered to have been written only by Shakespeare are correctly attributed. However, excep-

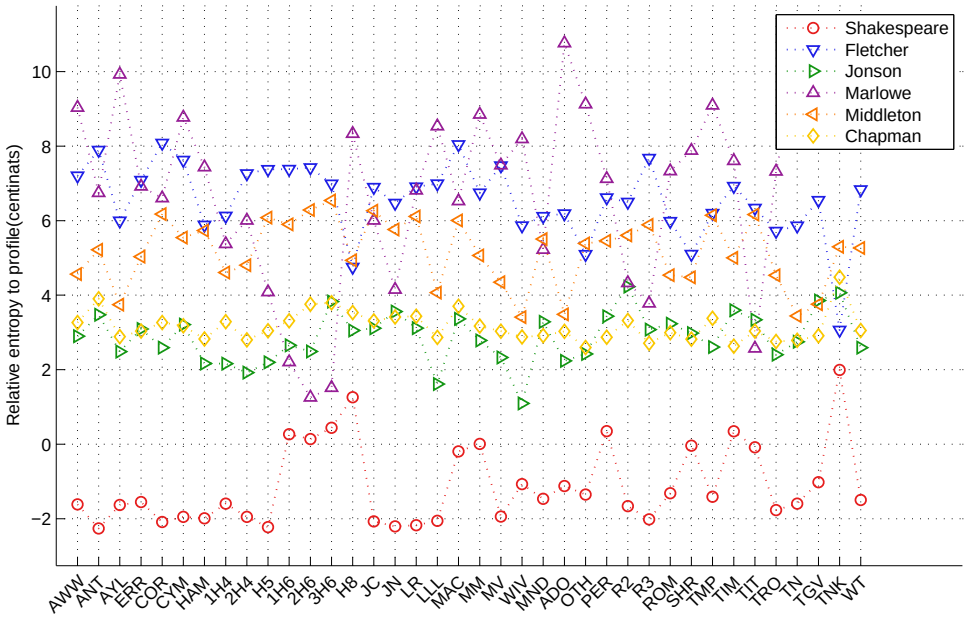

Fig. 6: Attribution of Shakespeare plays. We attribute the 28 plays in Table I and the additional 10 plays in Table VI. All plays are attributed to Shakespeare. Marlowe's distance to a play is highly dependent on whether the analyzed play is a history play or not, emphasizing the impact of genre in attribution.

tional 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 IV. 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 [44], [45] and points at Marlowe as a probable candidate. These collaborations are covered in greater detail in Section VIII-D.

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 V-B. 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 VIII-B and VIII-C, respectively.

F. 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 I, 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 Section VI.

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 Sections VII and VIII.

TABLE VII: 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)

VI. Anonymous Plays

In Fig. 7 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 [13] 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* [13]. Again, our method supports this theory. These plays are analyzed further in Section VIII-D. In addition, the plays *The Revenger's Tragedy*, *The Second Maiden's Tragedy*, and *The Nice Valor* are usually attributed to Middleton [28], 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 [46], 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 [47] since he is the first ranked author among the six authors we profile. *Fair Em* has also been assigned to Shakespeare [48] though there is no scholarly consensus, with Robert Wilson, whom we do not profile, often cited as a likely candidate [47]. *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* [49], 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. 6 ranks Marlowe as the worst candidate, indicating that much more of his style is evident in the early draft.

VII. 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.

A. 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 VIII.

The attribution of Fletcher's works are divided into two plots. Fig. 8 shows the attribution of plays believed to have been written solely by Fletcher and Fig. 9 shows the attribution of plays believed to have been written in collaboration with other authors. The set of plays presented before the first

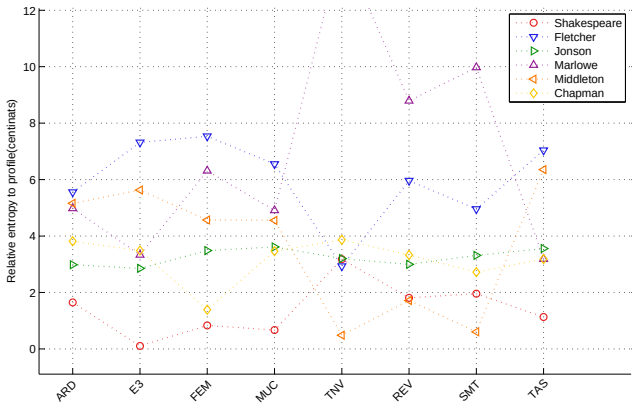

Fig. 7: Attribution of anonymous plays listed in Table VII. Our method supports the usual theories of Shakespeare's hand in *Edward III* and *Arden of Faversham*. Also, Middleton's style in *The Revenger's Tragedy*, *The Second Maiden's Tragedy*, and *The Nice Valor* can be observed, in accordance with current authorship consensus.

TABLE VIII: 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 IX: John Fletcher plays to be attributed in addition to those listed in Table I.

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)	

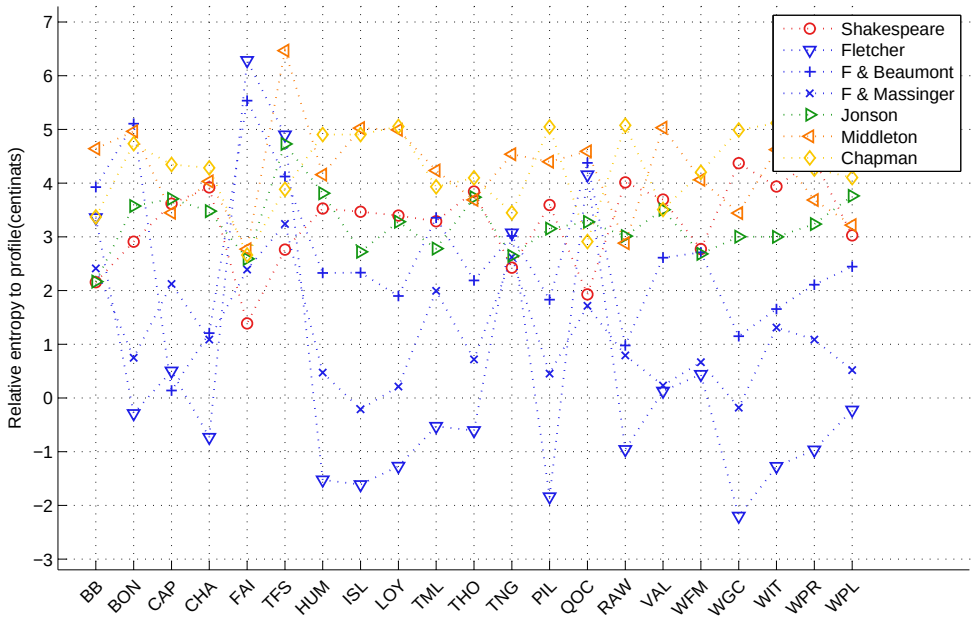

Fig. 8: Attribution of solo Fletcher plays. We attribute the 15 plays in Table I and the additional 6 plays in Table IX. Six plays are not attributed to the sole Fletcher profile and, among these, two plays are attributed to collaborative profiles including Fletcher.

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 IV.

In Fig. 8, 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 collaboration of four authors, two of which are Fletcher and Massinger, with Fletcher's contribution being minor [50]. This would explain the fact that the Fletcher and Massinger profile is ranked second but the sole Fletcher profile is poorly ranked.

In Fig. 9, 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 [51]. 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, two are the plays coauthored with Shakespeare and discussed previously in Section V-E and further in Section VIII-B. 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.

TABLE X: 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	

VIII. Collaborations – Intraplay Analysis

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

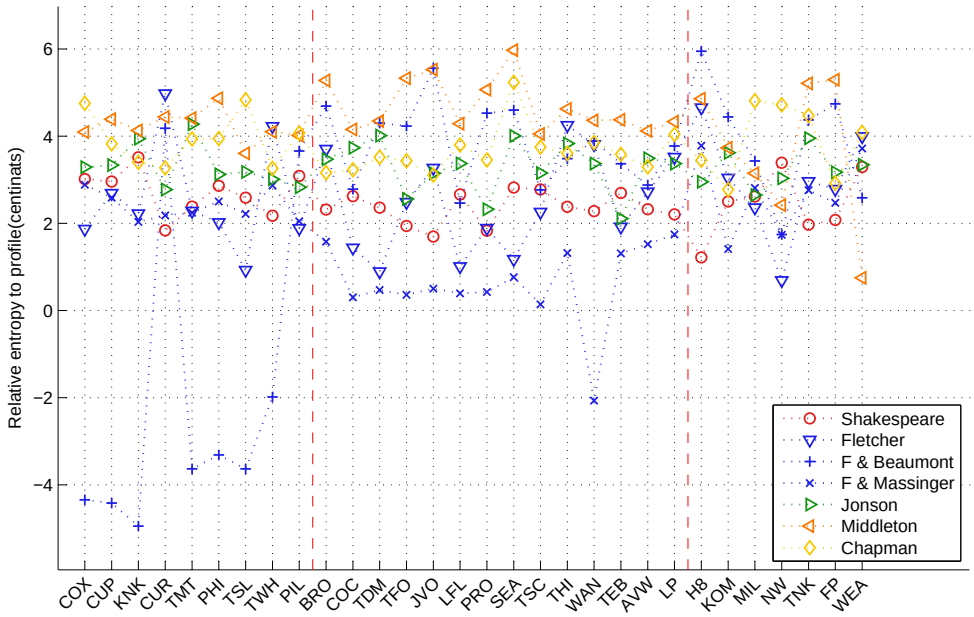

Fig. 9: Attribution of Fletcher plays written with collaborators. We attribute the plays listed in Tables VIII and IX. The first division includes plays written with Beaumont and 7 out of 9 are correctly assigned to the Fletcher & Beaumont profile. The second division includes plays written with Massinger 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.

TABLE XI: 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	

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

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

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 X. 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 XI and XII.

The figures display for each act or scene the difference in relative entropy when comparing 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, 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,

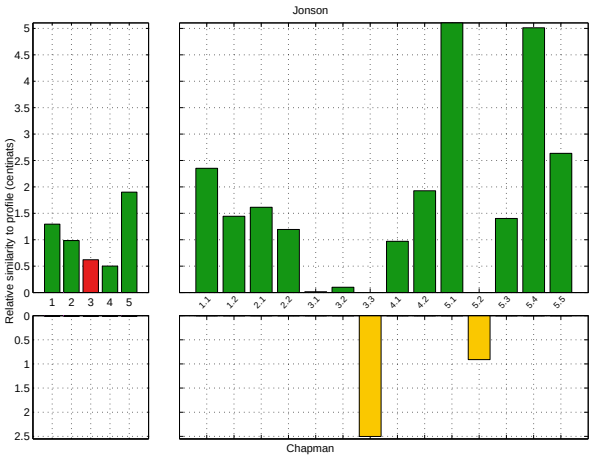

Fig. 10: Attribution of acts and scenes of *Eastward Ho*. Note that Act 3 is assigned to Shakespeare over both Jonson and Chapman.

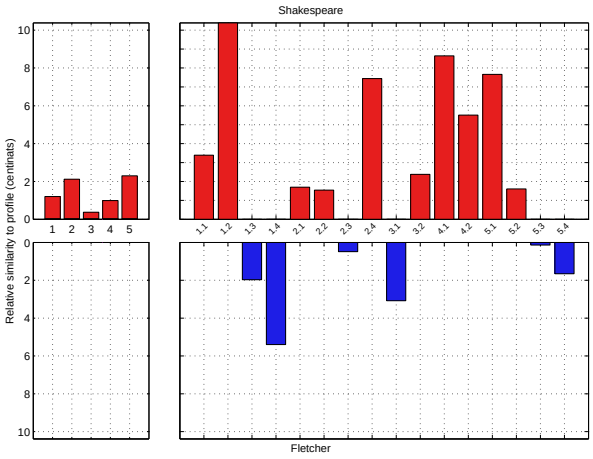

Fig. 12: Attribution of acts and scenes of *Henry VIII*.

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.

A. 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. 10 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. 2 and 4, 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 [36]. Most scholars agree in particular about scene 3.3 being written by Chapman [44]. 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.

B. Shakespeare and Fletcher

In Fig. 11 we show the attribution of individual acts and scenes of *Two Noble Kinsmen*, a known collaboration between Shakespeare and Fletcher. Whereas in Fig. 9 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

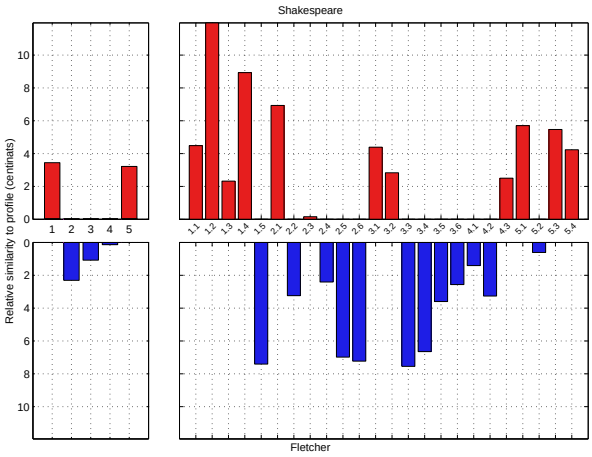

Fig. 11: Attribution of acts and scenes of *Two Noble Kinsmen*.

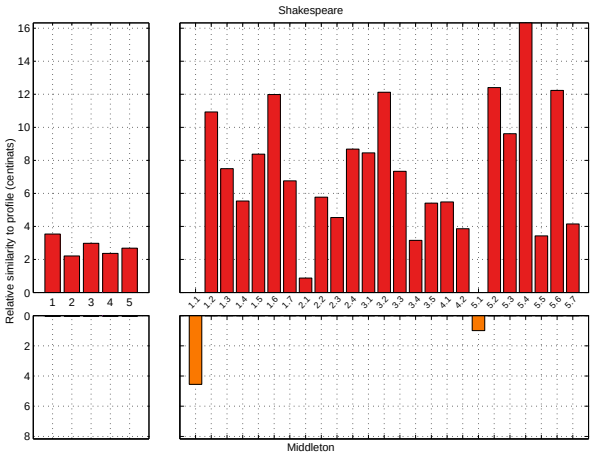

Fig. 13: Attribution of acts and scenes of *Macbeth*.

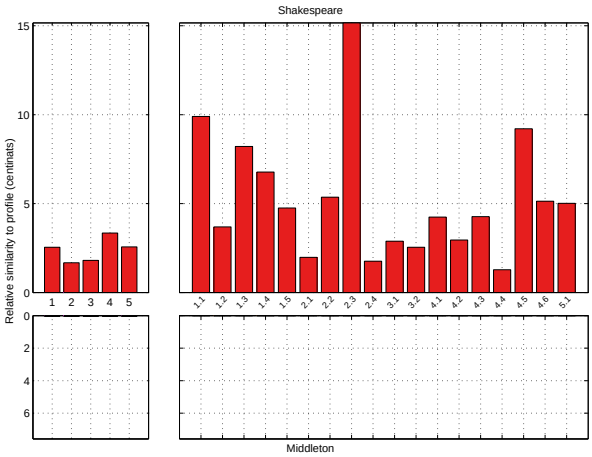

Fig. 14: Attribution of acts and scenes of *Measure for Measure*.

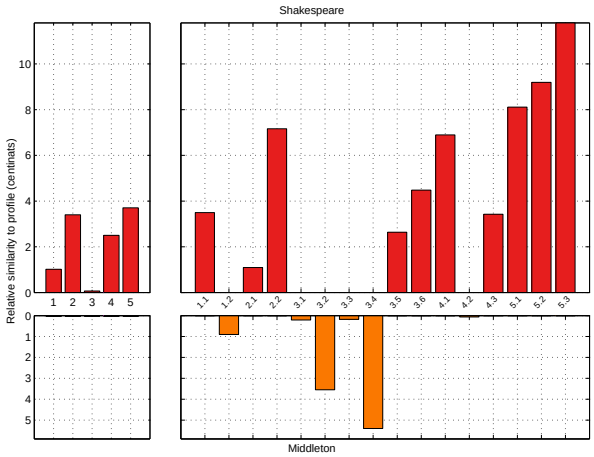

Fig. 15: Attribution of acts and scenes of *Timon of Athens*.

assignments. Shakespeare is assigned to scenes 1.1-4, 2.1, 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* [52].

The act and scene analysis of Shakespeare and Fletcher's other collaboration–*Henry VIII*–is displayed in Fig. 12. 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. 6 or Fig. 9. We see similar results in Fig. 12, 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 [46] 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.

C. Shakespeare and Middleton

We analyze in Figs. 13-15 Middleton's contributions to Shakespeare's plays, *Macbeth*, *Measure for Measure*, and *Timon of Athens*. The attribution of the full plays in Fig. 3 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 13, 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 Shakespeare 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 [43], 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 analysis displayed in Fig. 14 find Shakespeare to be the sole author of the play. If Middleton had indeed revised the original play as proposed by scholars [43], [53], 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. 15. 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 [13], [43].

D. 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. 6, the three parts of *Henry VI* have been considered as possible collaborations between Shakespeare and Marlowe [14], though others such as

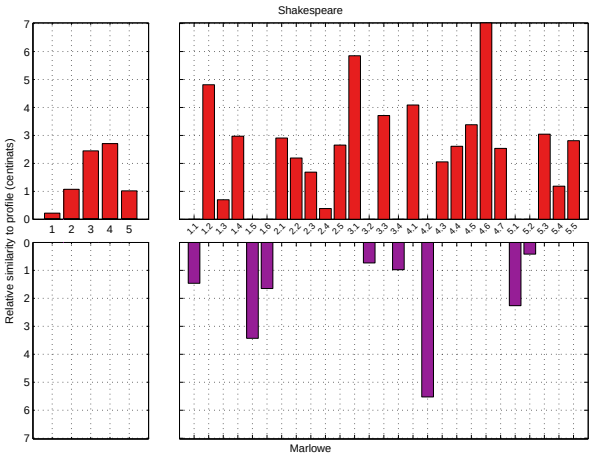

Fig. 16: Attribution of acts and scenes of *1 Henry VI*.

Greene and Peele have also been suggested. The attribution of the acts of *1 Henry VI*, displayed in Fig. 16, 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 [14], 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. 17. 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 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.

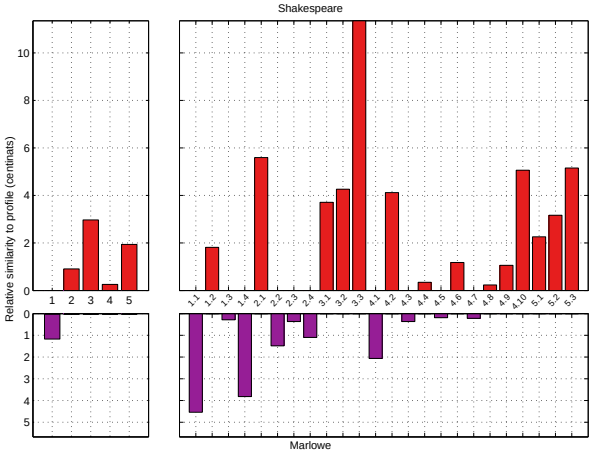

Fig. 17: Attribution of acts and scenes of *2 Henry VI*.

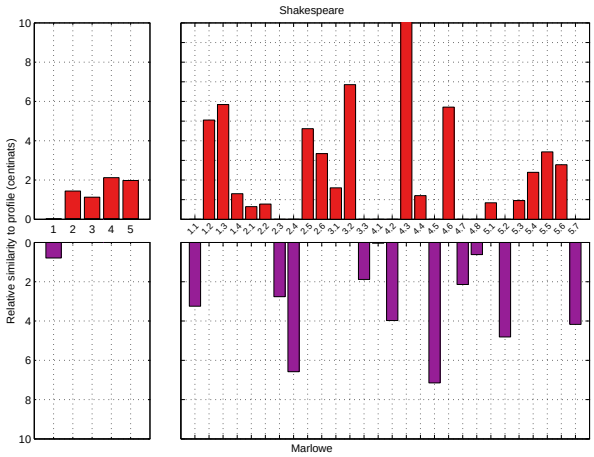

Fig. 18: Attribution of acts and scenes of *3 Henry VI*. Note that the relative entropy for scene 4.2 extends out of the view of the figure to $+30cn$.

The intraplay analysis of *3 Henry VI* in Fig. 18 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. 19 the intraplay analysis on the play *Arden of Faversham*, attributed to Shakespeare in Fig. 7. Every act is attributed here to Shakespeare. Although 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 [14], [54]. 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 [12] 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. 7. As before, the two most commonly cited candidates for co-authorship with Shakespeare are Kyd and Marlowe [14], [55]. The act attribution of *Edward III* in Fig. 20 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. 20, 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 [14]. 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 to comparatively large distance between the scenes' WANs and both profiles. Not displayed in Fig. 20, 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.

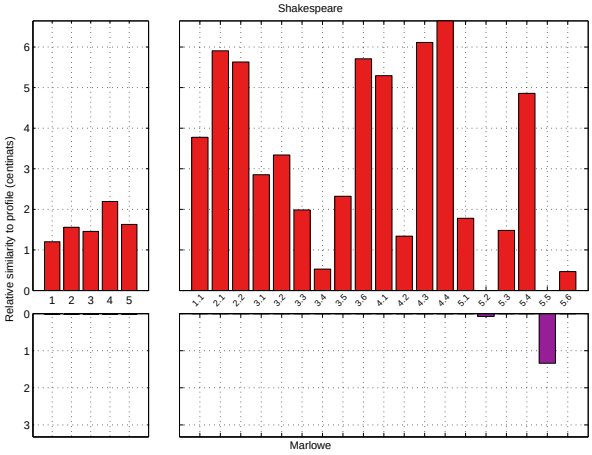

Fig. 19: Attribution of acts and scenes of *Arden of Faversham*.

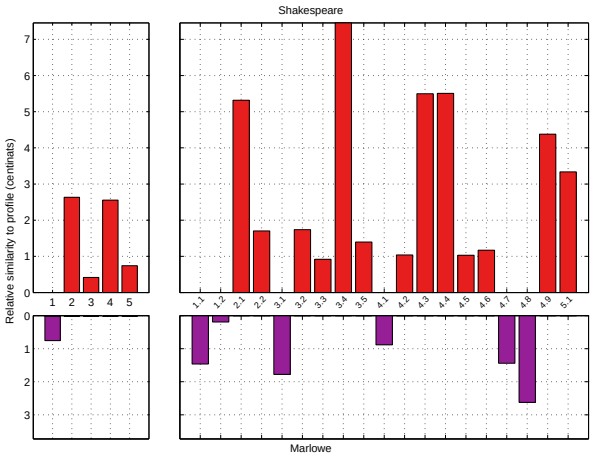

Fig. 20: Attribution of acts and scenes of *Edward III*.

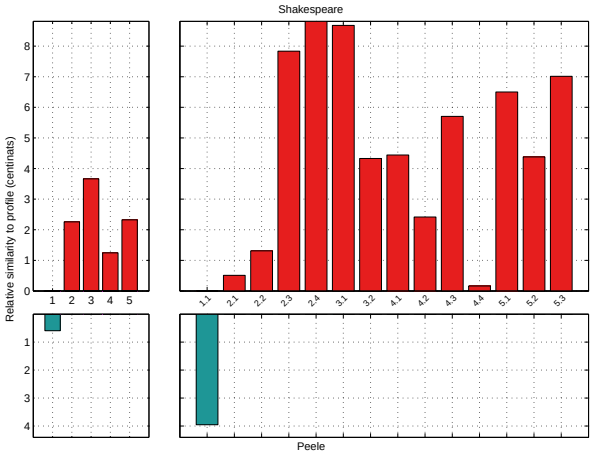

Fig. 21: Attribution of acts and scenes of *Titus Andronicus*. Note that here the comparative relative entropies for Act 1 and its sole scene, 1.1, differ. The plot of scene 1.1 reports the difference in relative entropy between Peele and Shakespeare while the plot of Act 1 reports the difference in relative entropy between Peele and the second ranked author, Marlowe.

E. Shakespeare and Peele

Shakespeare's play, *Titus Andronicus*, is commonly cited to include additions by Peele [13], and is attributed act by act and scene by scene in Figure 21. 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

TABLE XIII: 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	4.80	1.12

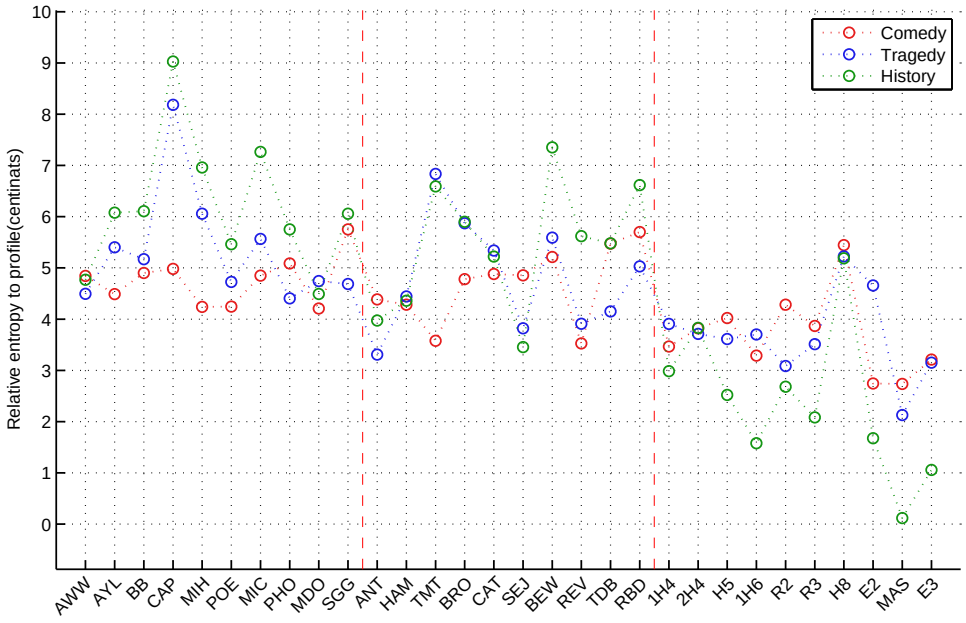

Fig. 22: Attribution of plays between genre profiles. The plays to left of the first red line include comedy plays. The plays to right of the first red line include tragedy plays. The plays to right of the second red line include history plays.

TABLE XIV: 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)

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 [13], 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 XIII. 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.

IX. Genre analysis

In addition to using WANs to distinguish author styles, we may also use them to distinguish between plays more generally at the level of genre. There has been debate among literary scholars as to whether the classification of fiction into genres is something determined solely by the plot or whether it can also be determined by the writing style itself. We demonstrate that, to some extent, it is possible to sort a play into its appropriate genre by considering only its writing style as encoded by function WANs.

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 paper. The complete list of texts used in the genre profiles is displayed in Table XIV. 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. 22, 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 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.

X. Conclusion

Function word adjacency networks (WANs) were used to analyze the authorship of texts written by popular playwrights during the Early Modern English period. Word adjacency networks were built for a large set of texts in the corpus of the analyzed authors and were compared via the relative entropy measure. 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 92.2% was achieved when attributing amongst all authors and an accuracy of 96.1% was achieved when attributing amongst authors that are more dissimilar than $5cn$. With this classification power, a selection of anonymous plays were attributed amongst 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 act and scene components 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 impact of genre on attribution accuracy was also briefly examined to gain insight into the similarity of writing styles with respect to a play's genre. We overall find function word adjacency networks 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 studies from this time period.

References

[1] Olivier De Vel, Alison Anderson, Malcolm Corney, and George Mohay, "Mining e-mail content for author identification forensics," *ACM Sigmod Record*, vol. 30, no. 4, pp. 55–64, 2001.

[2] Efstathios Stamatatos, "A survey of modern authorship attribution methods," *Journal of the American Society for information Science and Technology*, vol. 60, no. 3, pp. 538–556, 2009.

[3] Norman Meuschke and Bela Gipp, "State-of-the-art in detecting academic plagiarism," *International Journal for Educational Integrity*, vol. 9, no. 1, 2013.

[4] Frederick Mosteller and David Wallace, "Inference and disputed authorship: The federalist," 1964.

[5] David I Holmes and Richard S Forsyth, "The federalist revisited: New directions in authorship attribution," *Literary and Linguistic Computing*, vol. 10, no. 2, pp. 111–127, 1995.

[6] David I Holmes, "A stylometric analysis of mormon scripture and related texts," *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, pp. 91–120, 1992.

[7] Frederick Gard Fleay, *Shakespeare Manual*, Ams PressInc, 1878.

[8] Philip Wolcott Timberlake, *The Feminine Ending in English Blank Verse*, George Banta Publishing Company, 1931.

[9] Ants Oras, *Pause patterns in Elizabethan and Jacobean drama: an experiment in prosody*, University of Florida Press, 1960.

[10] Marina Tarlinskaja, James Bailey, and George T Wright, *Shakespeare's verse: Iambic pentameter and the poet's idiosyncrasies*, Lang, 1987.

[11] MacDonald Pairman Jackson, *Defining Shakespeare: Pericles as Test Case*, Oxford University Press, 2003.

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

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

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

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

[16] Shlomo Argamon and Shlomo Levitan, "Measuring the usefulness of function words for authorship attribution," in *ACH/ALLC*, 2005.

[17] Patrick Juola, "Authorship attribution," *Foundations and Trends in information Retrieval*, vol. 1, no. 3, pp. 233–334, 2006.

[18] David I Holmes, "Vocabulary richness and the prophetic voice," *Literary and Linguistic Computing*, vol. 6, no. 4, pp. 259–268, 1991.

[19] David L Hoover, "Another perspective on vocabulary richness," *Computers and the Humanities*, vol. 37, no. 2, pp. 151–178, 2003.

[20] Doug Cutting, Julian Kupiec, Jan Pedersen, and Penelope Sibun, "A practical part-of-speech tagger," in *Proceedings of the third conference on Applied natural language processing*. Association for Computational Linguistics, 1992, pp. 133–140.

[21] Santiago Segarra, Mark Eisen, and Alejandro Ribeiro, "Authorship attribution through function word adjacency networks," *CoRR*, vol. abs/1406.4469, 2014.

[22] Dmitri V Khmelev and Fiona J Tweedie, "Using markov chains for identification of writer," *Literary and linguistic computing*, vol. 16, no. 3, pp. 299–307, 2001.

[23] Conrad Sanderson and Simon Guenter, "Short text authorship attribution via sequence kernels, markov chains and author unmasking: An investigation," in *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2006, pp. 482–491.

[24] 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.

[25] George Kesidis and J Walrand, "Relative entropy between markov transition rate matrices," *IEEE transactions on information theory*, vol. 39, no. 3, pp. 1056–1057, 1993.

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

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

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

[29] 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.

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

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

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

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

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

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

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

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

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

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

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

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

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

[43] 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.

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

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

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

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

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

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

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

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

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

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

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

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