# Stylometric Techniques for Multiple Author Clustering

## Shakespeare's Authorship in *The Passionate Pilgrim*

David Kernot[1 3]
[1]Joint and Operations Analysis Division
Defence Science Technology Group
Edinburgh, SA, Australia

Terry Bossomaier
[2]The Centre for Research in Complex Systems
Charles Sturt University
Bathurst, NSW, Australia

Roger Bradbury
[3]National Security College
The Australian National University
Canberra, ACT, Australia

***Abstract*—In 1598-99 printer, William Jaggard named Shakespeare as the sole author of *The Passionate Pilgrim* even though Jaggard chose a number of non-Shakespearian poems in the volume. Using a neurolinguistics approach to authorship identification, a four-feature technique, RPAS, is used to convert the 21 poems in The Passionate Pilgrim into a multi-dimensional vector. Three complementary analytical techniques are applied to cluster the data and reduce single technique bias before an alternate method, seriation, is used to measure the distances between clusters and test the strength of the connections. The multivariate techniques are found to be robust and able to allocate nine of the 12 unknown poems to Shakespeare. The authorship of one of the Barnfield poems is questioned, and analysis highlights that others are collaborations or works of yet to be acknowledged poets. It is possible that as many as 15 poems were Shakespeare's and at least five poets were not acknowledged.**

***Keywords*—Authorship Identification; Principal Component Analysis; Linear Discriminant Analysis; Vector Space Method; Seriation***

## I. Introduction

William Jaggard first printed The Passionate Pilgrim in 1598-99, and the authorship of the 21 poems within it was attributed to William Shakespeare [1]. However, Bartholomew Griffin's 1596, Fidessa More Chaste Than Kind, already contained poem 11 [2]. Another, poem 19, appeared anonymously in Anne Cornwallis' 1580 personal notebook alongside works from Sir Philip Sidney, Sir Walter Raleigh, Sir Edward Dyer and Edward de Vere, 17th Earl of Oxford [3]. The list grows, and in 1598, Jaggard's brother John printed Richard Barnfield's,

The Encomion of Lady Pecunia, containing poems 8 and 11 [1]. By 1609, only five had been confirmed as Shakespeare's (poems 1, 2, 3, 5, and 17) having appeared in The Sonnets, or his play, Love's Labour's Lost [4]. Then, England's Helicon also printed a version of poem 20, attributing it to Christopher Marlowe, although its reply (signed Ignato) was later said to be by Sir Walter Raleigh [2]. Jaggard persisted with his claim, and in the 1612 third edition added a number of poems from Thomas Heywood, however, after complaints, Jaggard removed Shakespeare's name from the title [1]. By then, the authorship of 12 unknown poems lay in doubt, something that has remained for over 400 years.

Modern scholars are divided on the authorship of the remaining unknown twelve. Reference [5] suggests Jaggard used Shakespeare's name because the majority of the poems were Shakespeare's, including 12 unidentified poems in *The Passionate Pilgrim* said to be his earlier quality work and never meant for publishing. She also adds there is some doubt surrounding the authorship of the Barnfield and Griffin poems. Reference [6] disputes Shakespeare's authorship, while [7] suggest eight, not 12 of the anonymous poems are Shakespeare's. However, [2] suggest poems 7, 10, 13, 14, 15, 16, and 19 use a similar six-line stanza format to Shakespeare's *Venus and Adonis*, and poems 4, 6, and 9 are *about* Venus and Adonis and have Shakespearian similarities, but [5] says poems 7 and 13 resemble Robert Greene's poems.

It is interesting to note that unknown poem 12 gets little attention, even though it appears in Thomas Delany's *The Garland of Goodwill*, and entered into the Stationers Register ledger during 1592-3 [8]. When chosen by Jaggard, Delaney was living with an arrest warrant over his head because of his insightful writing during the London riots and in no position to complain [8], but what is strange are the few references in the literature to Delaney as the author until recently. Either way, Jaggard cannot be asked about the true authorship of the 21 poems, and today, the 12 poems, for the most part, remain unidentified.

Stylometric analysis, the quantitative analysis of a text's linguistic features has been extensively used to determine the authorship of the undocumented collaborations of the playwrights from the Elizabethan period, including Shakespeare [9]. There appears dissension among leading Shakespearean authorship attribution scholars about an agreed method [10], but the most successful and robust methods are based on low-level information such as character n-grams or auxiliary words (function word, stop words such as articles and prepositions) frequencies [11]. The premier work in evaluating authorship in the 16$^{th}$ to mid-17$^{th}$ centuries includes MacDonald P. Jackson, Brian Vickers, and Hugh Craig and Arthur Kinney [9]. Jackson [12] uses common low-frequency word phrases, repetition of phrases, collocation, and images to link word groups to other works. Vickers [13] uses a tri-gram, or n-gram, approach, while Hirch and Craig [14] use function word frequency and other methods, that includes ones based on word probabilities and the Information Theoretic measure Jensen-Shannon divergence (JSD) and unsupervised graph

partitioning clustering algorithms [15]. However, there are other techniques used in this period of Shakespearean analysis, including simple function words [16, 17] and word adjacency networks (WANs) [9]. However, the meaning-extracting method (MEM) from the field of psychology to extract themes from commonly used adjectives and describe a person from their personality, or self is very different [18, 19]. The authors offer a new and alternative approach to authorship identification using personality.

### A. An Approach Using RPAS

In this paper, a methodology is employed that adopts a multi-faceted approach to text analysis and reveal details about a person's personality; their sense of self, from subtle characteristics hidden in their writing style [20-22]. The techniques draw on biomarkers for creativity and known psychological states [23-24] to identify characteristics within *The Passionate Pilgrim* poems. It uses a series of four indicators (**RPAS**) identified in [25] to create a stylistic signature from a person's writing: **Richness (R)** [26], the number of unique words used by an author; **Personal Pronouns (P)** [27-30], the pronouns used, closely aligned to gender and self; **Referential Activity Power (A)** [31-32], based on function words, or word particles derived from clinical depression studies; and **Sensory (S)** [33-36], five sensory measures (V-visual A-auditory H – haptic O – olfactory G - gustatory) corresponding to the senses.

**RPAS** is used to create individual stylistic signatures of the 21 *The Passionate Pilgrim* poems and the known works of William Shakespeare, Christopher Marlowe and Sir Walter Raleigh, Richard Barnfield, and Bartholomew Griffin are labelled. Three clustering techniques are then applied to identify the likely authorship of the 12 unknown poems within *The Passionate Pilgrim*.

## II. Methodology

*The Passionate Pilgrim* contained within the complete works of Shakespeare [37] is used to process the data with the Stanford Parts Of Speech Tagger [38] to remove all punctuation and symbols and then aggregate the works by word frequency. *The Passionate Pilgrim* is further broken down into chunks that represent each known poem, and a decision made to follow the modern approach by editors [2], and divide poem 14 into two poems (labelled as 14 and 15) with a subsequent renumbering of the remaining poems so that there are twenty-one and not twenty poem chunks (refer to Table 1).

The 3,190-word data ends up as an aggregated matrix of 1,032 distinct word types across 21 poems, and the size of each varies between 96 and 377 words (average = 152). Putting this into perspective, they are slightly larger than a Shakespearian sonnet which varies between 91 and 132 words (average = 116).

TABLE I. The List of the Poems by Shakespeare, Barnfield, Griffin, Marlowe Including the 12 Unknown Authored Poems in the Passionate Pilgrim Poems by Author and Abbreviated ID

| ID | Abbreviated | Author |
|---|---|---|
| 1 | 1S | William Shakespeare |
| 2 | 2S | William Shakespeare |
| 3 | 3S | William Shakespeare |
| 4 | 4U | Unknown |
| 5 | 5S | William Shakespeare |
| 6 | 6U | Unknown |
| 7 | 7U | Unknown |
| 8 | 8B | Richard Barnfield |
| 9 | 9U | Unknown |
| 10 | 10U | Unknown |
| 11 | 11G | Bartholomew Griffin |
| 12 | 12U | Unknown (Thomas Delaney) |
| 13 | 13U | Unknown |
| 14 | 14U | Unknown |
| 15 | 15U | Unknown |
| 16 | 16U | Unknown |
| 17 | 17S | William Shakespeare |
| 18 | 18U | Unknown |
| 19 | 19U | Unknown |
| 20 | 20M | Christopher Marlowe and Walter Raleigh |
| 21 | 21B | Richard Barnfield |

A 1613 play written after Shakespeare ceased writing is used to provide an independent author perspective and clustering technique. *The Tragedy of Mariam, the Fair Queen of Jewry* by English poet and dramatist, Elizabeth Cary [39], was published 14 years after *The Passionate Pilgrim*, and stylistically very different to Shakespeare's work.

A nine-dimensional array is created from the data using RPAS before applying three complementary techniques to reduce any single bias and overlay the results against Richness (R) and Personal Pronoun (P) to determine the possible authorship of the 12 unknown poems. As a final measure, seriation, an exploratory combinatorial data analysis technique, is used to visualise the nine-dimensional array as a one-dimensional continuum and test the strength of the co-located cluster edges by adding random noise to the data vector.

### A. Three Complementary Techniques

Principal Component Analysis (PCA) of the 21 poems (threshold set to 0.30 to ignore any non-significant contributions) determines the variance explained through eigenvalues and identifies any significant factors, known as components, from within the data. Four components are then aggregated to examine the clusters.

Linear Discriminant Analysis (LDA) is used as an alternate classification technique to PCA [29-30]. The unknown works are removed, and all of the individual known authors' poems are numbered from 1 to 4 before training the model and reintroducing the unknown poems. Using the

resultant coefficients from the three Canonical Discriminant Functions, functions 1-2 and 1-3 are aggregated to visually compare the clusters.

The Vector Space Method (VSM) technique [42-43] is used with Elizabeth Cary's, *The Tragedy of Mariam, the Fair Queen of Jewry* as an imposter [44]. Pair-wise comparisons of each of the 21 *Passionate Pilgrim* poems is made against Elizabeth Carey's play (42 pair-wise comparisons) using both cosine and minmax similarity detection, to highlight the clusters that form based on their distance from Cary's play.

### *B. Seriation*

According to [45] "Seriation is an exploratory combinatorial data analysis technique to reorder objects into a sequence along a one-dimensional continuum so that it best reveals regularity and patterning among the whole series." Seriation is the process of placing a linear ordering on a set of N multi-dimensional quantities. The total number of possible orderings is N! (factorial). This grows extremely quickly with N. 5! = 120, 10! = 3.6 million and 20! = $2.4x10^{18}$, or 2.4 billion billion (or quintillion). Thus, even for quite small N, it is not possible to calculate the shortest path by calculating all possible paths. A heuristic or approximation is needed. Inevitably any given approximation will work better with some data than others. Thus, for a robust estimation of the shortest path, it might be necessary to try a range of different estimators and look for consistency among them.

Using the free software environment for statistical computing and graphics, R, and its seriation package [46], and provide the seriation package with the 9x21 matrix consisting of the nine RPAS values for each of the 21 poems of *The Passionate Pilgrim*. Using the Euclidean distance option, seriation attempts to minimise the Hamiltonian path length (the Hamiltonian path on a graph is a path which visits all the nodes just once). The results of the six Hamiltonian path-length calculations produced by the seriation package are evaluated (TSP: *Travelling Salesperson*, Chen: *Rank two ellipse Seriation*, ARSA: *Anti-Robinson Simulated Annealing*, HC: *Hierarchical Clustering*, GW: *Hierarchical Clustering (Gruvaeus Wainer heuristic)*, and OLO: *Hierarchical Clustering (Optimal Leaf Ordering))*. While seriation gives a one-dimensional continuum, Dendrogram branch and leaf visualization are also provided, and clusters can be separated by their Hamiltonian path distances [47]. The technique that provides the shortest Hamiltonian path is selected, and noise introduced into the matrix to examine the strength of the connected groups by using the jitter function in R. The function adds random noise to the vector by drawing samples from the uniform distribution of the original data [48].

## III. Analysis

Using **RPAS** Personal Pronouns (P) is plotted against Richness (R) (PtoR) for the 21 *The Passionate Pilgrim* poems (see Fig. 1). PtoR discriminates the unknown poems 14 and 16 with Shakespeare (poems 2 and 3), and they have a low feminine gendered style (P > 10), while all of Shakespeare's known poems have a lower feminine gendered style (P > 30), contrasting this is the group consisting of the cluster with unknown poems 7 and 19 that are similar in style to Griffin (poem 11) and Barnfield (poem 21) who all have a higher masculine style (P >50). The Shakespeare (poem 1) and the Marlowe and Walter Raleigh (poem 20) are similar, as are Barnfield (poem 8) and Shakespeare (poem 5). The unknown poem 12 (from Delaney) has a low Richness score is separate from the main body of poems.

### *A. Principal Component Analysis (PCA)*

The findings show that many PCA correlations are in excess of 0.30. A visual indication of the correlation matrix highlights 24 coefficients are around 0.30 or higher and some are as high as 0.77, and Bartlett's test is significant (p = 0.001) meaning there is some correlation between variables indicating that PCA is worthwhile. Four components are extracted and account for 81.95% of the variance.

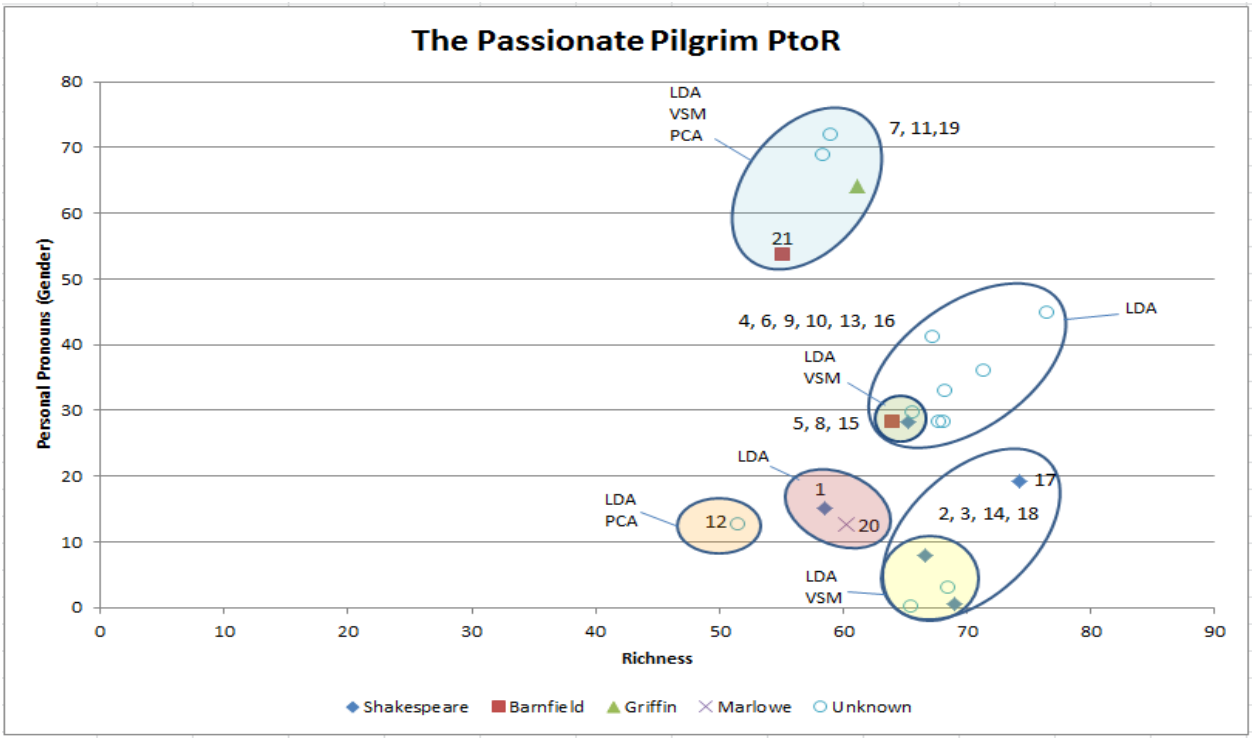


Fig. 1. In this The Passionate Pilgrim gendered Personal pronouns (P) versus Richness (R) diagram, the overlays of the results of LDA, VSM, and PCA analysis highlight the consistency of other results. A Barnfield / Griffin series of poems can be seen (7, 11, 19, and 21) with greater than 50% gendered personal pronouns. This is supported by LDA, VSM and PCA Analysis. A Shakespeare series of poems can be observed (2, 3, 14, 17, and 18), also supported by LDA and VSM analysis. A Shakespeare / Marlowe / Raleigh series is observed (1 and 20) to have less than 20% gendered personal pronouns supported by LDA analysis. Clearly, Delany's poem 12 is supported by LDA, and PCA analysis as a standalone work also has the lowest Richness. In the range of 25-50%, gendered personal pronouns are the Shakespeare / Barnfield poems (5, 8, and 15) supported by LDA and VSM analysis, and these alongside the unknown poems (4, 6, 9) (and 10, 13, 16 supported by LDA analysis). Further, the ellipses are a visual clustering assignment

In Fig. 1, the two common clusters are overlaid. A Barnfield / Griffin group (11 and 21) is found to sit with unknown poems 7 and 19. While unknown poem 12 (Thomas Delaney) was close to Shakespeare (1) and Marlowe and Raleigh (20), it is the furthest poem from the Shakespeare cluster on the Factor 1 and 2 scale that accounts for ~55% of the variance. Additionally, the results highlight all of the known Shakespeare poems cluster (poems 1, 2, 3, 5, 17 with 6, 14, 15, and 16). Poem 4 is close to Barnfield (8), and poems 6, 9, 15, and 16 are close to Shakespeare (5).

### B. *Linear Discriminant Analysis (LDA)*

Three functions were extracted, and the first two accounted for 99.6% of the variance (1 = 95.9 and 2 = 3.7). The Wilks' Lambda test of functions 1 through 3 was significant (p=0.009) which highlights that the null hypothesis can be rejected and suggests that all three functions together have a discriminating ability. The second and third functions together are not significant (p=0.190), neither is function 3 on its own (p=0.453). Functions 1-2 and functions 1-3 are plotted to generate six common clustering results (see Fig. 1). It is found that the unknown poems 10 and 13 are again close to Shakespeare (5) and Barnfield (8), as is 15. Unknown poems 7 and 19 are closer to Griffin (11) this time and further from Barnfield (21). Unknown poem 12 (Thomas Delaney) is again closest to Shakespeare (1) and Marlowe and Raleigh (20) but stands alone. Poem 14 is again close to Shakespeare (2 and 3).

While poem 18 is also close to Shakespeare (1, 2, and 3), poem 4 is far from all the poems but closest to Griffin (11). Poem 6 is closest to Shakespeare (17). Poem 16 is closest to Shakespeare (5), and poem 9 is in the middle of Shakespeare (5), Barnfield (21) and Griffin (11). Again, there is some consistency with these results, but there seems to be a lack of clarity with poems 4, 6, 9 and 16.

### C. *The Vector Space Method (VSM)*

Pair-wise comparisons of each of the 21 *Passionate Pilgrim* poems against Elizabeth Carey's play, *The Tragedy of Mariam, the Fair Queen of Jewry* (42 pair-wise comparisons) using both cosine and minmax similarity detection, highlights the clusters that form based on their distance from Cary's play. Fig. 1, indicates the three common clustering results. Here, unknown poems, 7 and 19 are in a cluster with Griffin (11). Unknown poem 14 is in a cluster with Shakespeare (1, 2, and 3) and Marlowe / Raleigh (20) and poems 12 and 18, and closest to Shakespeare (1), while Delaney's poem 12 and 14 are closest to Shakespeare (2), but furthest away. Unknown poems 4, 6, 9, 10, 13, 15, and 16 are in a cluster with Shakespeare (5 and 17) and Barnfield (8). In this cluster Barnfield (8) is very close to Shakespeare (5), and poems 10 and 13 have an almost identical score.

Throughout these different analysis techniques, there is a consistency in three to four clusters forming with common

poems in them, but many of the techniques have been dependent on an arbitrary visual clustering size. Therefore, to add further reliability to the results, the data is clustered using seriation to measure cluster distances.

### D. Seriation

The R seriation package is fed a 9x21 matrix of the data, and using Euclidean distance seriation of the data minimizes the Hamiltonian path length. Results of the six seriation techniques available highlight that Hierarchical Clustering with Optimal Leaf Ordering (OLO) outperforms the Travelling Salesperson technique (path lengths 214.63 vs. 228.92). Incorporating the clustering of the OLO Dendrogram at a height of 25, the order of the 21 chunks with clusters highlighted is [21 19 7 11] [4 9 6] [5 8 10 13 15 16 17] [20 12 1 3 2 14 18] and it highlights some susceptibility between poems 11-4, 6-5, and 17-20. When the distances between each poem are compared, and either side of poems 11-4 (7-11-4-9), 6-5 (9-6-5-8), and 17-20 (16-17-20-12), the ordering sequence and distance information is important (refer Table 2).

TABLE II. HAMILTONIAN PATH DISTANCES BETWEEN THE 21 *THE PASSIONATE PILGRIM* POEMS. THE OLO DENDROGRAM EDGE CLUSTERS THAT FORM AT A DENDROGRAM HEIGHT OF 25 HIGHLIGHTS A CONSISTENCY IN TWO OF THE THREE SEPARATION POINTS. IN THE CLUSTER SPLIT AT POEMS 11-4, 7-11 AND 4-9 ARE CLOSER THAN 11-4 (27.3 VERSUS 11.8 AND 9.6). IN THE CLUSTER SPLIT AT POEMS 6-5, 9-6 AND 5-8 ARE CLOSER THAN 6-5 (10.61 VERSUS 7.7 AND 3.4), BUT IN THE 17-20 CLUSTER SPLIT, WHILE 16-17 AND 20-12 ARE CLOSER THAN 17-20, THE DIFFERENCES BETWEEN 16-17 AND 17-20 ARE MARGINAL (15.8 AND 12.6 VERSUS 16.8)

| Poem edges | | Path length |
|---|---|---|
| 21 | 19 | 16.60488 |
| 19 | 7 | 24.69437 |
| 7 | 11 | 9.561261 |
| 11 | 4 | 27.27893 |
| 4 | 9 | 11.78111 |
| 9 | 6 | 7.683108 |
| 6 | 5 | 10.61323 |
| 5 | 8 | 3.444387 |
| 8 | 10 | 4.88489 |
| 10 | 13 | 3.22249 |
| 13 | 15 | 3.455063 |
| 15 | 16 | 4.449576 |
| 16 | 17 | 15.8412 |
| 17 | 20 | 16.75323 |
| 20 | 12 | 12.6397 |
| 12 | 1 | 14.13468 |
| 1 | 3 | 11.68744 |
| 3 | 2 | 8.28891 |
| 2 | 14 | 13.00578 |
| 14 | 18 | 6.162732 |

Further, when examining the OLO dendrogram edge clusters that form at a dendrogram height of 25 and find consistency in two of the three separation points. In the cluster split at poems 11-4, it can be seen that 7-11 and 4-9 are closer than 11-4 (27.3 versus 11.8 and 9.6). In the cluster split at poems 6-5, 9-6 and 5-8 are closer than 6-5 (10.61 versus 7.7 and 3.4), but in the 17-20 cluster split, while 16-17 and 20-12 are closer than 17-20, the differences between 16-17 and 17-20 are marginal (15.8 and 12.6 versus 16.8).

To see how stable the results are, in particular, the stability of the clusters connected at the poems 17-20 split, noise is inserted into the initial 9x21 RPAS-poem matrix and recalculate Euclidean distances with various amounts of noise (noise 1 – 8000). An examination of the scene chunk order after seriation (refer Table 3) highlights the high level of stability within the seriation and OLO clustering results. The different OLO seriation results are showing changes in order when noise is added to the RPAS poem matrix. At around noise levels of 500, poems 15 and 16 switch positions, but then revert back with further noise. At noise levels 800 and above, the Barnfield – Griffin cluster (7, 11, 19, and 21) move internally within the cluster but no poems leave. At noise levels 800 and higher the Shakespeare – Marlowe cluster (1, 2, 3, 12, 14, 18, and 20) move internally, and at no point does poem 20 moves out of the cluster and join with poem 17.

TABLE III. THE DIFFERENT OLO SERIATION RESULTS ARE SHOWING CHANGES IN ORDER WHEN NOISE IS ADDED TO THE RPAS POEM MATRIX. AT AROUND NOISE LEVELS OF 500, POEMS 15 AND 16 SWITCH POSITIONS, BUT THEN REVERT WITH FURTHER NOISE. AT NOISE LEVELS 800 AND ABOVE, THE BARNFIELD – GRIFFIN CLUSTER (7, 11, 19, AND 21) MOVE INTERNALLY WITHIN THE CLUSTER BUT NO POEMS LEAVE. AT NOISE LEVELS 800 AND HIGHER THE SHAKESPEARE – MARLOWE CLUSTER (1, 2, 3, 12, 14, 18, 20) MOVE INTERNALLY. THIS SUGGEST A HIGH LEVEL OF STABILITY IN THE SERIATION OLO ORDER AND OLO CLUSTERING RESULTS ([21 19 7 11] [4 9 6] [5 8 10 13 15 16 17] [20 12 1 3 2 14 18])

| **Noise** | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| ***Order*** | ***0*** | ***100*** | ***500*** | ***800*** | ***1000*** | ***2000*** | ***4000*** | ***8000*** |
| 1 | 21 | 21 | 21 | 7 | 7 | 7 | 7 | 7 |
| 2 | 19 | 19 | 19 | 11 | 11 | 11 | 11 | 11 |
| 3 | 7 | 7 | 7 | 19 | 19 | 19 | 19 | 19 |
| 4 | 11 | 11 | 11 | 21 | 21 | 21 | 21 | 21 |
| 5 | 4 | 4 | 4 | 4 | 4 | 4 | 4 | 9 |
| 6 | 9 | 9 | 9 | 9 | 9 | 9 | 9 | 6 |
| 7 | 6 | 6 | 6 | 6 | 6 | 6 | 6 | 4 |
| 8 | 5 | 5 | 5 | 5 | 5 | 5 | 5 | 5 |
| 9 | 8 | 8 | 8 | 8 | 8 | 8 | 8 | 8 |
| 10 | 10 | 10 | 10 | 10 | 10 | 10 | 10 | 10 |
| 11 | 13 | 13 | 13 | 13 | 13 | 13 | 13 | 13 |
| 12 | 15 | 15 | 16 | 15 | 15 | 15 | 15 | 15 |
| 13 | 16 | 16 | 15 | 16 | 16 | 16 | 16 | 16 |
| 14 | 17 | 17 | 17 | 17 | 17 | 17 | 17 | 17 |
| 15 | 20 | 20 | 20 | 14 | 20 | 20 | 14 | 14 |
| 16 | 12 | 12 | 12 | 18 | 12 | 12 | 18 | 18 |
| 17 | 1 | 1 | 1 | 20 | 1 | 1 | 20 | 20 |
| 18 | 3 | 3 | 3 | 12 | 3 | 3 | 12 | 12 |
| 19 | 2 | 2 | 2 | 1 | 2 | 2 | 1 | 1 |
| 20 | 14 | 14 | 14 | 3 | 14 | 14 | 3 | 3 |
| 21 | 18 | 18 | 18 | 2 | 18 | 18 | 2 | 2 |

## IV. DISCUSSION

Overall, the techniques were generally consistent, and seriation was useful because it was able to provide clustering and distance measures that appeared stable even with a relatively high level of introduced noise. Therefore, the basis of these findings lies in a rigorous multivariate approach to analysis and not a single technique. However, one of the biggest concerns is the influence of the publisher. While Jaggard or his associates cannot be discounted from having a hand in adding their own touches to some of these unknown poems, blending them as it were so they appear as part collaborations, it is an unknown factor. It is known that Jaggard was able to get hold of some of Shakespeare's unpublished work, and both he and his brother John had access to a wide number of Elizabethan works. What cannot be known is how much of this was early unpublished works.

Of the 12 anonymous poems, two are likely Shakespeare's, possibly from his earlier unpublished works (poems 14 and 18 are similar to Shakespeare's poems 2 and 3 and a lesser extent poem 1). However, if they were not earlier Shakespearian poems, then they are from another poet entirely, one that has not been examined. Two other poems (7 and 19) have a blended style similar to Griffin (11) and Barnfield (21), and there is more of Griffin's style (similar to poem 11) in them than Barnfield's, and they are more likely to be Griffin's unpublished work. Again, if they are not an unpublished Griffin poem, then they too are a poet that has not been examined in this paper. Poem (12) has a blended style similar to Shakespeare (1) and Marlowe / Raleigh (20) but consistently shows itself to be different enough to be an independent poet and be the work of Thomas Delaney whose other poems were outside of this analysis.

The remaining seven unknown poems (4, 6, 9, 10, 13, 15, and 16) are all similar in style to a blended Shakespeare (5 and 17) and Barnfield (8). All of these, as are all of Shakespeare's poems here, have a Richness score over 65%. They all have a Personal Pronoun score below 50%, which would be deemed as a feminine writing style which fits Shakespeare. Poems 4, 6, and 9 are very similar in style to each other and closer to Shakespeare's (5) style than Barnfield (8). Poems 10, and 13 are closer to Barnfield's (8) style than Shakespeare (5, 17). Poems 15 and 16 have a higher Shakespeare (5) style than Barnfield's (8) and are higher overall from the Shakespeare poems (5 and 17).

This close style of Barnfield's poem (8) to Shakespeare's (5) is an anomaly, and if it were not for the work sitting in the Shakespeare cluster between 5 and 17, then it could be easily be said that all the poems (4, 6, 9, 10, 13, 15, and 16) are Shakespeare's. The literature around Richard Barnfield is examined more closely. While Barnfield and Shakespeare were certainly friends [49] and could have collaborated, these poems are likely to be Shakespeare's because the style of Barnfield's poem (8) is very similar to Shakespeare's poem (5). It has been suggested, that the 1598 version of Barnfield's manuscript obtained by William Jaggard's brother John was of insufficient length (indicated by the sparse printing layout), and William Jaggard provided his brother two poems from the yet unpublished *The Passionate Pilgrim* to extend Barnfield's *Lady Pecunia* publication. In the 1605 reprint of Richard Barnfield's *Lady Pecunia*, the two poems from the 1598 first edition (poems 8 and 21 from *The Passionate Pilgrim*) were not included [50-51]. According to [52], Barnfield is said to have claimed authorship of only *one* of the two poems (stylistically likely poem 21). If this is true, then it explains the striking similarities between the Shakespeare and Barnfield poems (5 and 8), and a good indication that Shakespeare wrote both 5 and 8, and therefore poems 4, 6, 9, 10, 13, 15, and 16 are Shakespeare's poems. While it further reinforces Jaggard's approach to borrowing from other author's works, from the analysis it is believed that Shakespeare wrote nine of the twelve unknown poems (4, 6, 9, 10, 13, 14, 15, 16, and 18) including 1, 2, 3, 5, 17, and 8.

## V. Conclusion

Given Shakespeare's signature in almost three-quarters of the poems, Jaggard may have adopted shrewd marketing tactics in using Shakespeare's name as the sole author. Indeed, when he expanded the third edition with a collection of nine of Heywood's poems, he did not remove Shakespeare's name from the title, nor did he add Heywood as co-author, but in his collection of assorted verses. Jaggard merely adopted what was a standard convention by publishers in the day [53]. The analysis would suggest that the five authors, Barnfield, Delaney, Griffin, Marlowe, and Raleigh were not acknowledged, and several poems may well be collaborative works between Shakespeare and others but this also was common [54]. It is also possible that several poems (7, 14, 18,19) are not early work or collaborations, but other writer's poems not studied here. This failing to acknowledge all author's poems would seem, at least by today's standards, to be an injustice. However, as it can be seen with Jaggard's publication of *The Passionate Pilgrim* and his later publication of Shakespeare's first folio, Jaggard focussed on promoting Shakespeare's work above all others.

In this paper, authors have demonstrated an alternate stylometric technique that can identify self and cluster multiple authors using RPAS. It includes the use of sensory-based adjectives and words that are strong in concreteness and imageability that reflect known psychological states in an individual's personality. They believe that further research is warranted to see if RPAS can identify changes in an individual's stylometric fingerprint over time.

## Acknowledgment

The authors thank D. Crone and C. van Antwerpen for critical discussions and reading of the manuscript. This research supported by the Defence Science Technology Group, the Australian Government's lead agency dedicated to providing science and technology support for the country's defence and security needs.

## References

[1] Erne, L. (2013). *Shakespeare and the book trade*. Cambridge University Press. Pp. 56-86.

[2] Devington, D. (2007) The Poems by William Shakespeare. Bantam Books, . New York.

[3] Woudhuysen, H. R. (1996). *Sir Philip Sidney and the circulation of manuscripts, 1558-1640*. Oxford University Press.

[4] Connor, F. X. (2014). Shakespeare, poetic collaboration and The Passionate Pilgrim. pp119-130, in Holland, P. (Ed.). (2014). *Shakespeare Survey: Volume 67, Shakespeare's Collaborative Work* (Vol. 67). Cambridge University Press.

[5] Chiljan, K. (2012). Reclaiming The Passionate Pilgrim for Shakespeare. *Oxfordian* , 2012, Vol. 14, p74-81

[6] Bednarz, J.P. (2007) "Canonizing Shakespeare: The Passionate Pilgrim, England's Helicon and the Question of Authenticity," Shakespeare Survey 60 (2007): 255-58,260,262.

[7] Elliott, W. E., & Valenza, R. J. (1991). A Touchstone for the Bard. *Computers and the Humanities*, *25*(4), 199-209.

[8] Korp, C. (2015). Shoemakers, Clowns, and Saints: The Narrative Afterlife of Thomas Deloney. Available at: http://escholarship.org/uc/item/8hk20311

[9] Segarra, S., Eisen, M., Egan, G., & Ribeiro, A. (2015). Stylometric analysis of early modern period English plays. *Digital Scholarship in the Humanities, vol.(submitted)*.

[10] Rudman, J. (2016). Non-Traditional Authorship Attribution Studies of William Shakespeare's Canon: Some Caveats. *Journal of Early Modern Studies*, *5*, 307-328.

[11] Stamatatos, E. (2009). A survey of modern authorship attribution methods. *Journal of the American Society for information Science and Technology*, *60*(3), 538-556.

[12] Jackson, M. P. (2006). Shakespeare and the quarrel scene in arden of faversham. *Shakespeare Quarterly*, *57*(3), 249-293.

[13] Vickers, B. (2011). Shakespeare and authorship studies in the twenty-first century. *Shakespeare Quarterly*, *62*(1), 106-142.

[14] Hirsch, B. D., & Craig, H. (2014). " Mingled Yarn": The State of Computing in Shakespeare 2.0.

[15] Arefin, A. S., Vimieiro, R., Riveros, C., Craig, H., & Moscato, P. (2014). An information theoretic clustering approach for unveiling authorship affinities in Shakespearean era plays and poems. *PloS one*, *9*(10), e111445.

[16] Matthews, R. A., & Merriam, T. V. (1993). Neural computation in stylometry I: An application to the works of Shakespeare and Fletcher. *Literary and Linguistic Computing*, *8*(4), 203-209.

[17] Merriam, T. V., & Matthews, R. A. (1994). Neural computation in stylometry II: An application to the works of Shakespeare and Marlowe. *Literary and Linguistic Computing*, *9*(1), 1-6.

[18] Boyd, R. L., & Pennebaker, J. W. (2015). Did Shakespeare write Double Falsehood? Identifying individuals by creating psychological signatures with text analysis. *Psychological science*, 0956797614566658.

[19] Chung, C. K., & Pennebaker, J. W. (2008). Revealing dimensions of thinking in open-ended self-descriptions: An automated meaning extraction method for natural language. *Journal of research in personality*, *42*(1), 96-132.

[20] Argamon, S., Koppel, M., Pennebaker, J. W., & Schler, J. (2009). Automatically profiling the author of an anonymous text. Communications of the ACM, 52(2), 119-123.

[21] Iqbal, F., Binsalleeh, H., Fung, B., & Debbabi, M. (2013). A unified data mining solution for authorship analysis in anonymous textual communications. Information Sciences, 231, 98-112.

[22] Northoff, G., Heinzel, A., de Greck, M., Bermpohl, F., Dobrowolny, H., & Panksepp, J. (2006). Self-referential processing in our brain—a meta-analysis of imaging studies on the self. *Neuroimage*, *31*(1), 440-457.

[23] Rosenstein, M., Foltz, P. W., DeLisi, L. E., & Elvevåg, B. (2015). Language as a biomarker in those at high-risk for psychosis. *Schizophrenia research*.

[24] Zabelina, D. L., O'Leary, D., Pornpattananangkul, N., Nusslock, R., & Beeman, M. (2015). Creativity and sensory gating indexed by the P50: Selective versus leaky sensory gating in divergent thinkers and creative achievers. *Neuropsychologia*, *69*, 77-84.

[25] Kernot, D., Bossomaier, T., & Bradbury, R. (2017). Novel Text Analysis for Investigating Personality: Identifying the Dark Lady in Shakespeare's Sonnets, Journal of Quantitative Linguistics (Accepted 18 Jan, 2017).

[26] Tweedie, F. J., & Baayen, R. H. (1998). How variable may a constant be? Measures of lexical Richness in perspective. Computers and the Humanities, 32(5), 323-352

[27] Argamon, S., Koppel, M., Fine, J., Shimoni, A.R. (2003). Gender, genre, and Writing Style in Formal Written Texts. Text, Volume 23, Number 58, August 2003.

[28] Kernot, D. (2016) *Can Three Pronouns Discriminate Identity in Writing* in Data. In Sarker, R., Abbas, H., Dunstall, S., Kilby, P., Davis, R. Young, L. (eds) Data and Decision Sciences in Action: Proceedings of the Australian Society for Operations Research Conference 2016, Springer.

[29] Pennebaker, J. W. (2011). The secret life of pronouns. *New Scientist*, *211*(2828), 42-45.

[30] Pennebaker, J. W., Mehl, M. R., & Niederhoffer, K. G. (2003). Psychological aspects of natural language use: Our words, our selves. *Annual review of psychology*, *54*(1), 547-577.

[31] Bucci, W. (2002). The referential process, consciousness, and the sense of self. *Psychoanalytic Inquiry*, *22*(5), 766-793.

[32] Bucci, W., & Maskit, B. (2004). Building a weighted dictionary for referential activity. In *Spring Symposium of the American Association for Artificial Intelligence in Palo Alto, CA, March.*

[33] Kernot, D. The Identification of Authors using Cross Document Co-Referencing. The University of New South Wales. Nov 2013. Available at: http://www.unsworks.unsw.edu.au/primo_library/libweb/action/dlDisplay.do?vid=UNSWORKS&docId=unsworks_12072

[34] Lynott, D., & Connell, L. (2009). Modality exclusivity norms for 423 object properties. *Behavior Research Methods*, *41*(2), 558-564.

[35] Miller, G. A. (1995). *The science of words*. New York: Scientific American Library.

[36] van Dantzig, S., Cowell, R. A., Zeelenberg, R., & Pecher, D. (2011). A sharp image or a sharp knife: Norms for the modality-exclusivity of 774 concept-property items. *Behavior research methods*, *43*(1), 145-154

[37] Farrow, J. M. (1993) The Collected Works of Shakespeare. http://sydney.edu.au/engineering/it/~matty/Shakespeare/

[38] Toutanova, K., & Manning, C. D. (2000, October). Enriching the knowledge sources used in a maximum entropy part-of-speech tagger. In *Proceedings of the 2000 Joint SIGDAT conference on Empirical methods in natural language processing and very large corpora: held in conjunction with the 38th Annual Meeting of the Association for Computational Linguistics-Volume 13* (pp. 63-70). Association for Computational Linguistics.

[39] Mark, M. (2014) A Celebration of Women Writers. Available at: http://digital.libttrary.upenn.edu/women/cary/Mariam/Mariam.html Accessed 27 October 2014.

[40] Balakrishnama, S., & Ganapathiraju, A. (1998). Linear discriminant analysis-a brief tutorial. *Institute for Signal and information Processing*.

[41] Ye, J., Janardan, R., & Li, Q. (2004). Two-dimensional linear discriminant analysis. In *Advances in neural information processing systems* (pp. 1569-1576).

[42] Koppel, M., & Winter, Y. (2014). Determining if two documents are written by the same author. *Journal of the Association for Information Science and Technology*, *65*(1), 178-187.

[43] Voorhees, E. M. (1998). Using WordNet for text retrieval. *Fellbaum (Fellbaum, 1998)*, 285-303.

[44] Seidman, S. (2013). Authorship verification using the impostors method. In *CLEF 2013 Evaluation Labs and Workshop-Online Working Notes*.

[45] Liiv, I. (2010). Seriation and matrix reordering methods: An historical overview. *Statistical analysis and data mining*, *3*(2), 70-91.

[46] Buchta, C., Hornik, K., & Hahsler, M. (2008). Getting things in order: an introduction to the R package seriation. *Journal of Statistical Software*, *25*(3), 1-34.

[47] Earle, D., & Hurley, C. B. (2015). Advances in dendrogram seriation for application to visualization. *Journal of Computational and Graphical Statistics*, *24*(1), 1-25.

[48] Stahel, W., Maechler, M. (2011). 'Jitter' (Add Noise) to Numbers. R Documentation (1995 – 2011) available at: http://stat.ethz.ch/R-manual/R-devel/library/base/html/jitter.html. Accessed: 2 August 2016.

[49] Sauer, M. M. (2008). *The Facts on File Companion to British Poetry Before 1600*. Infobase Publishing.

[50] Barnfield, R. (1598). Lady Pecunia, Or, The Praise of Money: Also A Combat Betwixt Conscience and Covetousnesse ; Together with The Complaint of Poetry for the Death of Liberality. In Volume 1, Issue 7 of Illustrations of old English literature. pp 1-49. Digitized 25 Oct 2012. Available at: https://books.google.com.au/books?id=0J1TAAAAcAAJ. Accessed on: 11 Nov 2015.

[51] Barnfield, R. (1605). Lady Pecunia, Or, The Praise of Money: Also A Combat Betwixt Conscience and Covetousnesse ; Together with The Complaint of Poetry for the Death of Liberality. In Volume 1, Issue 4 of Illustrations of old English literature. pp 1-38. Digitized 25 Oct 2012. Available at: https://books.google.com.au/books?id=y51TAAAAcAAJ. Accessed on: 11 Nov 2015.

[52] Britannica, E. (2008). Richard Barnfield The Project Gutenberg EBook of Encyclopaedia Britannica, 11th edition, Volume 3, Part 1, Slice 3. Published 10 December, 2008. Page 415.

[53] Reid, L. A. (2012). "Certaine Amorous Sonnets, Betweene Venus and Adonis": fictive acts of writing in The Passionate Pilgrime of 1612. *Études Épistémè. Revue de littérature et de civilisation (XVIe–XVIIIe siècles.*

[54] Thomas, M. W. (2000). Eschewing credit: Heywood, Shakespeare, and plagiarism before copyright. *New Literary History*, *31*(2), 277-293.