Информационные процессы, Том 15, № 4, 2015, стр. 460–471.

On the putative watermark in the Shakespearean Sonnets

M. Malyutov and T. Zhang

Mathematics Dept., Northeastern University, 360 Huntington Ave., Boston, MA 02115, USA E-mail: m.malioutov@neu.edu, zhang.tong@husky.neu.edu

Received December, 10, 2015

Abstract—The anomalous frequency difference of a certain **collection of letters** in the Shakespeare Sonnets between its initial and other lines is found. A similar Microsoft Excel analysis shows anomalous frequency of around half of the English alphabet there. The first anomaly suggests possibility of a putative watermark. An outrageous ambiguity of anagrams to literary texts of moderate length is established following from their Shannon modeling as stationary random series.

KEYWORDS: Anagrams, Large deviations probabilities of stationary time series, homogeneity t-test, watermark

1. INTRODUCTION

The controversy concerning authorship of the works ascribed to W. Shakespeare dates back several centuries due to the fact that rare documents related to his life are hard for many to reconcile with his authorship (see e.g. `http://shakespeareauthorship.org/`.)

Many leading figures in poetry, prose and also actors, statesmen, scientists doubt the official authorship version. Around 2000 influential writers, scholars, actors, theater directors, etc. continue to be non-believers and signed the declaration of reasonable doubt to the British government demanding funding for studying the problem, see `http://www.doubtaboutwill.org/declaration`.

A **bibliography** of material relevant to the controversy that was compiled by Prof. J. Galland in 1947 is about **1500 pages** long (see [6]). A comparable work written today might well be at least several times as large. A substantial part of research moved to the Internet, since publishing works contradicting the official version in academic journals is almost prohibited.

A significant complication for heretics is that they do not agree on the alternative author. Presently, the most popular alternative candidate seems to be Christopher (Kit) Marlowe, sudden death of whom is disputed (see more than a dozen large books appeared last few years on this subject). The Hoffman's prize of around 1 000 000 English pounds awaits the person who will prove Marlowe's authorship of substantial part of the SC.

1.1. Preface and Contents outline

Shakespearean Canon (SC) consists of the First Folio (a collection of 36 plays), sonnets and some other poems and plays. The official authorship of the SC is seen as extremely unlikely by many scholars. Search for hidden communications (steganography) in the SC has been popular for centuries in attempts of finding the true author. The Folger Shakespearean Library awarded in 1957 the brilliant review book [6] on early history of futile mining for steganography in the SC. In particular, examples in [6] display the outrageous ambiguity of anagrams to a literary text of moderate

length written in the same language. This notorious ambiguity was formalized in [8, 9] as an Anagram Ambiguity Proposition (AAP) which practically precludes anagrams as steganography tool except for watermarks, i.e. messages known to the readers. Readers only *check their presence*. We propose in section 3 additional arguments to those in [8,9] for the existence of the watermark in the sonnets the anagrammed name of their likely author C. Marlowe. The AAP discussion in section 6 is important in view of an intensive use of alleged anagrams 'deciphered' in [3] and repeated in a forthcoming book [2] for 'recovering' the fascinated C. Marlowe's adventures before and after his untimely official demise. We start displaying our numerical results in section 7 with showing anomalous letter frequency deviations between the Sonnets and current English and continue it with displaying a significant difference of the full anagrammed watermark in the first lines of the sonnets; while all its fractions and name LEDOUX of popular among Marlovians guess for Marlowe in late XVIth century (dismissed recently) prove their insignificance.

The section 1 on letter M frequency analysis in B. Jonson works strengthens recent findings of [7]. Sections 4,5 sketch the history a relevant steganography part. The Microsoft Excel frequency analysis commands displayed in [21] can help future researchers in similar studies.

2. STATISTICAL ANOMALY FOUND BY C. GAMBLE

Absence of the case-insensitive letter 'M' in the attributed to Ben Jonson (BJ) Foreword to the was recently noticed in [7]. This foreword was printed opposite the equally-strange Droeshout engraving, which purportedly depicts the author of the Shakespeare plays (though it is unlikely that Martin Droeshout had ever met the Stratford man, having been just 15 years old when the latter died in 1616 after spending his last several years in his hometown). Also, the engraving itself seems to be a mockery fake, since the right shoulder of the doublet the portrayed person is dressed in is actually the back side of the left shoulder, see[19]. As usual in steganography mining, case-insensitive letter frequency only are considered further indicated as Capitals in our text. The foreword consists of 10 lines of verses, around 300 letters. Our Microsoft EXCEL computations give the frequency 2.35% (close to the contemporary 2.4%) of case-insensitive (as always further in this paper)'M' in BJ and estimated probability of this anomaly to occur under randomness hypothesis is around $(1-0.024)^{300}$ which is approximately $e^{-7.2} = 0.0007 < 1/1000$. Also interesting is atypical distribution of around half of alphabet letters including 'M' in the Shakespeare Sonnets, see our table. Jonson's three verses 'A Celebration of Charis: I. His Excuse for Loving'; 'A Celebration of Charis: IV, Her Triumph'; 'A Fit of Rhyme against Rhyme' contain 46 letters 'M' out of total number 2930. Thus the ratio 46/2930=1.57%; and the corresponding p-value of Gamble's anomaly would be 0.009. However, we checked the 'M' frequency in a larger collection of Jonson's poems and found 431 letters 'M' out of 18320 letters. Thus frequency of 'M' in Jonson is around 431/18320=2.3526% which is not far from the modern value. Simon Singh's statistics of modern English gives frequency 2.4 % of letter 'M'. The deviation of 'M' frequency in the above three verses gives less extreme p-value of homogeneity in the Foreword around 2%.

3. ANAGRAMS AS WATERMARKS

3.1. History of anagrams as watermarks

Book [11] of Stanford historians, p. 253, claims that Greek authors of tragedies used to anagram their names and time of writing in several first lines of their tragedies. Authors of the Greek tragedies constructed the first eight iambic lines so that they not only made sense but also provided letters to make eight other iambic lines, the first two giving the writer's name, the next two the Olympiad, the third a homage to Athena, and the last couplet a warning that the show was about to begin.

We found no independent confirmation to this information which Marlowe could well learn from his best teachers in the King's school, Canterbury, and University of Cambridge. We ask those familiar with this problem to comment especially about known examples (if any) of actual anagrams in Greek plays and whether the rules for watermarking were strict.

A similar tradition was shared by Armenian ancient writers as a protection against plagiarism of copyists, as described in [1]. Announcing discoveries by anagrams was very popular in those times (Galileo, C.Huygens, J. Kepler, the I. Newton's Fundamental Anagram of Calculus [18] which was sent to Leibnitz for claiming Newton's priority). These great authors seem to have overlooked outrageous ambiguity of anagrams; a professional spy Marlowe might have also used anagrams.

3.2. Statistical test for existence of the watermark

The 154 Shakespeare sonnets of 14 lines (7 bi-lines) each (with a single exception) constitute a homogeneous sample appropriate for statistical inference. To test statistically whether Marlowe could follow tradition [11] and hide there his signature, we first evaluated the frequency of his anagrammed name in their first two lines (called the first 'bi-line' further) as compared to the rest 'bi-lines' in [7–9]. A careful choice of an accurate published version of sonnets was suggested by a gifted expert in the field, Dr. D. Khmelev, who had been previously involved in a joint Shakespeare stylometry study with some well-known British linguists. He died suddenly in 2004 being less than thirty years old.

For a given bi-line b, let us introduce the event $A(b)$ which means that b contains the set of case-insensitive letters $\{M, A, R, L, O, W, E\}$. Event A is equivalent to possibility of anagram of this name in this bi-line. Using a specially written code by my request, D.Khmelev found (this has been later confirmed by T. Zhang's straightforward Microsoft Excel calculations):

Proposition 1 [7–9]. *The numbers of first, second, etc. bi-lines in the sonnets for which event $A(\cdot)$ occurs are respectively 111, 112, 88, 98, 97, 101, 102 out of 154 sonnets.*

Our first corollary followed:

Proposition 2. *Let us test the null hypothesis of homogeneity: event A has the same probability for all consecutive bi-lines in sonnets versus the alternative that the first bi-line contains this set of letters more often than subsequent ones. It is also assumed that these events for all bi-lines are independent. Then the P-value of the null hypothesis (i.e. the probability of the frequency deviation to be as large or more under the null hypothesis) is about 3.75%.*

Proof. We apply a standard two-sample test (see e.g. [11] for equality of probabilities based on the normalized difference between frequencies $f_i, i = 1, 2$, of bi-lines containing the case-insensitive set of letters $\{M, A, R, L, O, W, E\}$ inside respectively the first and all other bi-lines respectively which has approximate standard normal distribution for such a large sample; f_1 is nearly 72.1%, f_2 is almost 65%. Thus the approximate normalized difference of frequencies

$$\frac{f_1 - f_2}{\sqrt{\bar{f}(1-\bar{f})(1+1/6)/154}}$$

is around 1.78, where $\bar{f} := (f_1 + 6f_2)/7$, and the normal approximation to the binomial probability of this or larger deviation (P-value) is nearly 3.75% which is a rather unlikely event. A more detailed study of numbers in Proposition 1 ignoring the multiplicity of hypotheses shows that the case insensitive set $\{M, A, R, L, O, W, E\}$ is located anomalously often in the first two bi-lines of the sonnets the homogeneity P-value is around 0.3%. Another popular (according to Ballantine)

signature 'Kit M.' turns out to be found unusually often (the homogeneity P-value is around 5%) in the last two bi-lines concluding the sonnets.

Apparently, this anomaly in homogeneity of bi-lines signals that the first bi-lines were specially designed to include this set of letters as part of an anagram signature. Note that signatures may vary over sonnets which only makes this anomaly stronger since our estimate is an UPPER BOUND for the P-value of bi-lines homogeneity versus several versions of Marlowe's signature in the first bi-lines.

Thus, the existence of anagrams hidden by Marlowe in Shakespeare looks rather likely.

3.3. More statistics about the sonnets

Our Microsoft Excel evaluations show that the frequencies of about half of letters ('M' included) in the Sonnets deviate from the modern language significantly to say the least, see our tables 1-5. Moreover, some frequencies (say, 'M') decline from the first to the rest bi-lines.

Further, having in mind that the full 'Marlowe' anagram was found significantly more often in the first bi-line than in the rest (p-value 3% [7–9]) and even more significantly often in the first TWO bi-lines than in the rest bi-lines (p-value 0.3% [7–9]), we evaluated corresponding p-values for various abbreviated hidden signatures. Remarkably, the full hidden signature was the ONLY one which was significantly more frequent in initial lines of the sonnets, see Table 8. Thus, it can hardly be explained simply by letter frequency anomalies.

4. ANAGRAMS AS A STEGANOGRAPHY TOOL

We describe our attitude to the new type of spiritism — anagrams allegedly discovered in [3] by R. Ballantine (RB), who would become an outstanding fiction writer, given her talent and literary background. Another matter is an intensive use of RB's anagrams by R. Ayres , PhD in Physics, [2] in spite of our heated discussions with the author over my Anagrams Ambiguity Proposition (AAP) formulated in [7–9] with a sketched proof. A detailed proof can be based on the Shannon's celebrated modeling of *meaningful literary texts* in some length range as a regular stationary stochastic process well-approximated by sparse n-Markov chain. However, AAP does not prevent use of the so-called 'watermarks' because the watermark content is KNOWN to potential readers, who only check its presence. Thus the ambiguity problem does not arise for watermarks and we argue that the Sonnets are likely to contain the watermarks– anagrams of the name 'Marlowe'. Some additional 'watermarks' could be used by Marlowe such as a 'torch burning upside down' drawn on his famous (although disputed) portrait. R. Balantine noticed its occurrence in the Italian play L'Ippolito I written by Gregorio de Monti—a secretary in English embassy in Venice.

5. MORE ON HISTORY OF CRYPTOGRAPHY

Mathematical theory of hidden communication (steganography) appeared as a byproduct of joint effort in code-breaking (SIS lab, US Army), headed by the US pioneer in crypto-analysis, US army general William Friedman, and Project X (declassified in 1976) of the Bell Lab and the Cipher School at Bletchley Park, UK. American and English teams included respectively Claude Shannon and Alan Turing. The two teams collaborated. William and Elizabeth Friedman started their crypto-analytic work around 1915 in 'colonel' Fabian's private Riverbank Lab, IL, by dismantling the E. Gallup erroneous discovery of Bacon's Trithemius code in Shakespeare (approved previously by the head of the French military cryptography). Their brilliant review book [5] of 1957 was

awarded by the Folger Shakespeare Library. An amazing example of anagrams ambiguity is shown in [5], pp. 110–111, namely a sample of 3100 different MEANINGFUL lines-anagrams in Latin for the famous short salutation "Ave Maria, gratia plena, Dominus tecum". These were referred to activities of monks collected in a book published in 1711. [5] acknowledges the Shannon's modeling of meaningful literary texts as stationary processes and states that the code-breaking can only be based on the language redundancy in rather lengthy corpus of texts. Nevertheless, further steganography mining in Shakespeare by amateurs such as [3,4] continued. Notably, R. Ballantine wrote referring to her archival and textual discoveries collected in around half a century of hard work: 'My gathered evidence could be viewed only through a glass darkly, spotted with ifs, lacunae. But when I found Marlowe's ciphers, his words within words suddenly made my shadowy outline into a picture of his life graced with undreamed-of detail: new stories appeared, most of them perfectly fitting my bits of evidence.

6. ANAGRAMMING FOR RECOVERY OF LONG STORIES

Unfortunately, a catastrophic ambiguity of anagrams stated in AAP makes the above-written statement an illusion. R. Ayres [2] takes liberty to deviate from the versions written in the anagrams, when he describes Marlowe's adventures in northern Europe under the name Le Doux in 1594-1596 contradicting the R.B.'s version of Marlowe's whereabouts at those times. Sadly, the Le Doux hypothesis popular among Marlovians for considerable time was dismissed recently by P. Farey et al.

One of R.B.'s Venice spectacular stories allegedly deciphered as anagram suspiciously resembles certain story from the Renaissance Italian fiction. Also, her fascinated Micaela Lujan's story based on anagram 'deciphering 'looks very improbable because of Micaela's age and many children from Lope de Vega.

Additional serious doubts about anagrams as appropriate steganography tool consists in their outrageous complexity of encoding and 'decoding'. Only an extraordinary devotion of R. Ballantine to C. Marlowe enabled her many decades of 'decoding' anagrams from many hundreds lines in many Shakespearean works. Her 'deciphered' texts are not in a good English involving rather arbitrary installations because of outrageous complexity of 'deciphering'. R. Ballantine has considered bi-lines as suitable periods for anagramming case-insensitive letters. After deciphering an initial bi-line, she proceeds to the very next one, and so on, until the final signature. In a given play, the first bi-line that begins an anagramming is usually at the beginning of a dialogue, or after an allegedly special, but otherwise meaningless sign, a number of which appear in early editions of Shakespearean works.

A theory of anagram ambiguity can be developed along the lines of the famous approach to cryptography given in C. Shannon's Communication Theory of Secrecy Systems written in 1946 and declassified in 1949. An English literary text is modeled in it as a stationary ergodic sequence of letters with its entropy per letter characterizing the uncertainty of predicting the next letter given a long preceding text. The binary entropy of literary English turns out to be around 1.1 (depending on the author and style), estimated as a result of long experimentation which continues in present studies.

Shannon showed by experimentation that this value of the entropy implies the existence of around $T_N = 2^{1.1N}$ meaningful English texts of large length N. More accurately, this is formulated in terms of the so-called exponential rate:

Proposition 3. $\log_2 T_N/1.1N \to 1$ *as* $N \to \infty$.

Notice that this proposition determines only the coefficient of exponential relation leaving aside much more delicate problem of finding less influential power terms of the dependence which is a common feature of results on the sp-called 'Large Deviation' which we soon touch in arguing for our AAP. Two English texts can be regarded as anagrams to each other, if and only if numbers of all English letters in them coincide.

Due to the approximate ergodicity of long texts, the joint frequency of all English letters in all typical long literary texts is approximately the same, and so all typical texts could be viewed as almost anagrams of each other. Thus, the number of anagrams to a given text grows with the same exponential rate as the number of texts in good English. Since the theory in case of a stationary ergodic approximation is rather involved, we start with a simplified model of an i.i.d. multinomial sequence of symbols to display our main arguments.

Here, the proof of this statement is almost straightforward:

Proposition 4. *Consider an i.i.d. three-nomial N-sequence of three letters A, B, C with rational probabilities $p(A) = L(A)/N, p(B) = L(B)/N, p(C) = L(C)/N$ such that $L(A), L(B), L(C)$ are integers. Then number $N(A, B)$ of N-sequences with $L(A)$ letters A and $L(B)$ letters B satisfies:*

$$\log N(A, B)/N = H(A, B) = -[p(A) \log p(A) + p(B) \log p(B) + p(C) \log p(C)](1 + o(1)).$$

The proof follows immediately from the method of types (see [4]). The fraction above is asymptotically the number of typical N-sequences as we stated above.

A generalization to a general multinomial i. i. d. case without the condition of all probabilities being multiples of $1/N$ is straightforward. The number of meaningful English anagrams for n bi-lines is the n-th power of that for a single bi-line, if deciphering is independent for subsequent bi-lines, and also *exponential* in the length of text under modeling texts as a sequence of independent letters.

A generalization to a model of stationary ergodic model of literary texts can be formulated and proved using its approximation by n-Markov Chain (MC). Moreover, [10] establishes for the Federalist papers that this approximation can be chosen with a sparse transition matrix. Furthermore, the necessary and sufficient criterion for approximating n-MC to be equivalent to a standard MC on the much smaller state space of contexts is in [20]. This opens effective application of the Large Deviations (LD) theory for finite MC.

The continuous rate function for LD is evaluated (see e.g. [13], [16], section 1.15) showing the coefficient of exponential rate of discrimination between different MCs. Thus, not only exponential asymptotics of numbers of typical contexts but also slightly different exponential rate of the number of contexts with a small difference between stationary distributions can be effectively evaluated.

Summarizing, the present asymptotic theory of MC gives similar results to those for independent letters displayed above establishing exponential multiplicity of anagrams in good English to a text written in a good English.

Moreover, the aim of putative anagrams that would become known to an addressee only after the long process of publication is unclear, unless an ESS editor would pass it directly to an addressee.

7. STATISTICAL AND NUMERICAL RESULTS OBTAINED WITH MICROSOFT EXCEL

It turned out that for more than half of letters these frequencies deviate very significantly from the values given in Wikipedia for typical English.

Further on, we decided to continue of 'MARLOWE' s comparative frequencies in the first and other bi-lines of the sonnets with adding a study of the same comparison for 'LEDOUX' —name of popular among many Marlovians guess for Marlowe in late XVIth century (dismissed recently). There are only 3 LEDOUX in the first bi-lines; in the rest bi-lines there are 40 LEDOUX. which make: $f_1 = 3/154 = 0.01948, f_2 = 40/154/6 = 0.04329$. Two-sample t-test found the difference between these frequencies significant. My next step was partitioning sonnets into early (29) and subsequent 125 ones . Our choice is based on some Marlovian online publications claiming that the first around 30 sonnets were created as a present at the 17th birthday of his illegitimate son count William Herbert around 1997, when Marlowe was (most likely) hiding under the name LEDOUX. From 1600 to 1604, first in Spain, then in England, he temporarily accepted his own name. Thus for rather arbitrary (29, 125)-partition, we made the separate testing for 'LEDOUX' in the first 29 sonnets and for 'MARLOWE' in the rest 125 sonnets.

We got : for LEDOUX, $f_1 = 0, f_2 = 8/29/6 = 0.045977, f = 0.039409$, p-value=0.119369 (not significant), For MARLOWE, $f_1 = 96/125 = 0.768, f_2 = 482/125/6 = 0.642667$, p-value=0.003074 (highly significant!)), so their product is $0.119369 \times 0.003074 = 0.00037$. It makes sense that LEDOUX is NOT in the first bi-line since it is a nickname rather than signature. There are only 3 LEDOUX in the first bi-lines; in the rest bi-lines there are 40 LEDOUX. which make: $f_1 = 3/154 = 0.01948, f_2 = 40/154/6 = 0.04329$.

Next is our result on various letter frequency in the sonnets as compared to that in modern English with p-value of standard t-test not more than 0.02:

Tables 1–4 of letters with frequencies $\hat{p}$ of each letter anomalously different from the frequency (p_0) of modern English (from Wikipedia) in respectively first to first four bi-lines of 154 sonnets.

The total number of letters tested is 10309.

Table 1. Modern English test.

Letter	$\hat{p}$	p_0	Test statistic	P-value
A/a	6.944%	8.167%	4.5341	0.0000
C/c	1.823%	2.782%	5.9189	0.0000
D/d	3.695%	4.253%	2.8072	0.0050
G/g	1.552%	2.016%	3.3507	0.0008
H/h	7.264%	6.094%	4.9662	0.0000
I/i	6.352%	6.966%	2.4471	0.0144
M/m	3.026%	2.406%	4.1074	0.0000
N/n	5.596%	6.749%	4.6666	0.0000
O/o	8.108%	7.507%	2.3152	0.0206
P/p	1.358%	1.929%	4.2168	0.0000
T/t	10.222%	9.057%	4.1237	0.0000
U/u	3.433%	2.758%	4.1863	0.0000
V/v	1.319%	0.978%	3.5180	0.0004
W/w	2.987%	2.360%	4.1945	0.0000
X/x	0.039%	0.150%	2.9176	0.0035
Y/y	2.977%	1.974%	7.3238	0.0000

Total number of letters tested is 20884.

Table 2. The first 4 lines test.

Letter	$\hat{p}$	p_0	Test statistic	P-value
A/a	6.813%	8.167%	7.1439	0.0000
C/c	1.748%	2.782%	9.0897	0.0000
D/d	3.811%	4.253%	3.1641	0.0016
H/h	7.024%	6.094%	5.6172	0.0000
I/i	6.325%	6.966%	3.6398	0.0003
M/m	3.002%	2.406%	5.6208	0.0000
N/n	5.846%	6.749%	5.2016	0.0000
P/p	1.412%	1.929%	5.4275	0.0000
S/s	6.813%	6.327%	2.8860	0.0039
T/t	9.949%	9.057%	4.4954	0.0000
U/u	3.304%	2.758%	4.8150	0.0000
V/v	1.269%	0.978%	4.2703	0.0000
W/w	2.954%	2.360%	5.6561	0.0000
X/x	0.081%	0.150%	2.5618	0.0104
Y/y	2.806%	1.974%	8.6404	0.0000
Z/z	0.029%	0.074%	2.4060	0.0161

Total number of letter tested is 31214.

Table 3. The first 6 lines test.

Letter	$\hat{p}$	p_0	Test statistic	P-value
A/a	6.747%	8.167%	9.1637	0.0000
C/c	1.829%	2.782%	10.2360	0.0000
D/d	3.796%	4.253%	4.0000	0.0001
G/g	1.804%	2.016%	2.6645	0.0077
H/h	6.939%	6.094%	6.2389	0.0000
I/i	6.292%	6.966%	4.6801	0.0000
M/m	2.922%	2.406%	5.9444	0.0000
N/n	6.003%	6.749%	5.2514	0.0000
O/o	7.922%	7.507%	2.7840	0.0054
P/p	1.374%	1.929%	7.1253	0.0000
S/s	6.766%	6.327%	3.1842	0.0015
T/t	9.944%	9.057%	5.4612	0.0000
U/u	3.338%	2.758%	6.2575	0.0000
V/v	1.256%	0.978%	4.9868	0.0000
W/w	2.790%	2.360%	5.0074	0.0000
X/x	0.090%	0.150%	2.7529	0.0059
Y/y	2.749%	1.974%	9.8379	0.0000
Z/z	0.032%	0.074%	2.7265	0.0064

Total number of letter tested is 41663.

Table 4. The first eight lines test.

Letter	$\hat{p}$	p_0	Test statistic	P-value
A/a	6.799%	8.167%	10.1925	0.0000
C/c	1.877%	2.782%	11.2339	0.0000
D/d	3.874%	4.253%	3.8361	0.0001
G/g	1.848%	2.016%	2.4318	0.0150
H/h	6.867%	6.094%	6.5929	0.0000
I/i	6.425%	6.966%	4.3372	0.0000
J/j	0.106%	0.153%	2.4752	0.0133
M/m	2.837%	2.406%	5.7399	0.0000
N/n	6.243%	6.749%	4.1198	0.0000
P/p	1.435%	1.929%	7.3272	0.0000
S/s	6.907%	6.327%	4.8669	0.0000
T/t	9.780%	9.057%	5.1485	0.0000
U/u	3.183%	2.758%	5.2912	0.0000
V/v	1.222%	0.978%	5.0536	0.0000
W/w	2.707%	2.360%	4.6700	0.0000
X/x	0.094%	0.150%	2.9744	0.0029
Y/y	2.667%	1.974%	10.1614	0.0000
Z/z	0.038%	0.074%	2.6721	0.0075

Parts of Marlowe signature test:

We use single-tail test, our results are as follows:

take MARL as an example,

f_1: frequencies of MAR happened in first bi-lines of 154 sonnets (there are 123 sonnets with first bi-line containing letters M,A, R and L;)

f_2: frequencies of MAR happened in the rest bi-lines of 154 sonnets (there are 730 bi-lines (excluding the first bi-line) with letter M,A,R and L;)

f: : frequencies of MAR happened in bi-lines of 154 sonnets (there are 853 of sonnets have bi-line with letter M,A, R and L;)

$f_1 = 123/154 = 0.7987$, $f_2 = 730/154/6 = 0.7900$, $f = 853/7/154 = 0.7913$ and z-score: $\frac{f_1 - f_2}{\sqrt{\frac{f(1-f)(1+\frac{1}{6})}{154}}} = 0.2448$

Table 5.

Test	MAR	LOWE	MARL	MARLO	LEDU	LEDOU	LEDOUX
f_1	0.8636	0.8312	0.7987	0.7987	0.7727	0.7727	0.0195
f_2	0.8539	0.7630	0.7900	0.7868	0.7673	0.7641	0.0433
f	0.8553	0.7727	0.7913	0.7885	0.7681	0.7653	0.0399
z-score	0.3181	1.8693	0.2448	0.3349	0.1473	0.2347	1.3978
p-value	0.3752	0.0308	0.4033	0.3688	0.4414	0.4072	0.0811

A more elaborate test for ‘LEDOUX’ in the first 29 sonnets, ‘MARLOWE’ in the rest:

Table 6.

LEDOUX	$N \leq 29$	$N > 29$
LEDOUX	0	8
MARLOWE	96	482
TOTAL	96	490

Table 7.

Test	
f_1	0.6234
f_2	0.5303
f	0.5436
z-score	2.1468
p-value	0.0159

MARLOWE TEST accumulated from the present sonnet to the 154$^{\text{th}}$ sonnet.

Table 8.

	acc.[1] t-score	p-value		acc. t-score	p-value		acc. t-score	p-value
sonnet 1	1.7820	0.0374	sonnet 34	2.5170	0.0059	sonnet 67	2.2046	0.0137
sonnet 2	1.8937	0.0291	sonnet 35	2.4672	0.0068	sonnet 68	2.0810	0.0187
sonnet 3	2.0063	0.0224	sonnet 36	2.4171	0.0078	sonnet 69	1.9885	0.0234
sonnet 4	1.9595	0.0250	sonnet 37	2.3666	0.0090	sonnet 70	1.8616	0.0313
sonnet 5	2.0227	0.0216	sonnet 38	2.4983	0.0062	sonnet 71	1.8000	0.0359
sonnet 6	1.9755	0.0241	sonnet 39	2.4478	0.0072	sonnet 72	1.8052	0.0355
sonnet 7	1.9027	0.0285	sonnet 40	2.3969	0.0083	sonnet 73	1.8105	0.0351
sonnet 8	1.9921	0.0232	sonnet 41	2.3172	0.0102	sonnet 74	1.7140	0.0433
sonnet 9	1.9444	0.0259	sonnet 42	2.2081	0.0136	sonnet 75	1.5816	0.0569
sonnet 10	1.8964	0.0290	sonnet 43	2.3139	0.0103	sonnet 76	1.5169	0.0646
sonnet 11	1.7963	0.0362	sonnet 44	2.2904	0.0110	sonnet 77	1.7468	0.0403
sonnet 12	1.7735	0.0381	sonnet 45	2.2668	0.0117	sonnet 78	1.7524	0.0399
sonnet 13	1.7245	0.0423	sonnet 46	2.4323	0.0075	sonnet 79	1.8809	0.0300
sonnet 14	1.8936	0.0291	sonnet 47	2.3509	0.0094	sonnet 80	1.8168	0.0346
sonnet 15	1.9853	0.0236	sonnet 48	2.3273	0.0100	sonnet 81	2.0187	0.0218
sonnet 16	1.9366	0.0264	sonnet 49	2.2448	0.0124	sonnet 82	1.9902	0.0233
sonnet 17	1.9139	0.0278	sonnet 50	2.2503	0.0122	sonnet 83	1.8902	0.0294
sonnet 18	2.0595	0.0197	sonnet 51	2.1966	0.0140	sonnet 84	2.0609	0.0197
sonnet 19	2.1535	0.0156	sonnet 52	2.1723	0.0149	sonnet 85	2.0323	0.0211
sonnet 20	2.1575	0.0155	sonnet 53	2.0878	0.0184	sonnet 86	1.9672	0.0246
sonnet 21	2.3055	0.0106	sonnet 54	1.9111	0.0280	sonnet 87	2.1424	0.0161
sonnet 22	2.4546	0.0071	sonnet 55	2.0521	0.0201	sonnet 88	2.0772	0.0189
sonnet 23	2.6047	0.0046	sonnet 56	2.0269	0.0213	sonnet 89	2.0113	0.0221
sonnet 24	2.7032	0.0034	sonnet 57	2.2006	0.0139	sonnet 90	1.9819	0.0237
sonnet 25	2.8028	0.0025	sonnet 58	2.1142	0.0173	sonnet 91	1.9523	0.0255
sonnet 26	2.9564	0.0016	sonnet 59	2.2286	0.0129	sonnet 92	1.9225	0.0273
sonnet 27	2.8829	0.0020	sonnet 60	2.3447	0.0095	sonnet 93	1.8543	0.0318
sonnet 28	2.8355	0.0023	sonnet 61	2.2889	0.0110	sonnet 94	1.9608	0.0250
sonnet 29	2.7878	0.0027	sonnet 62	2.2326	0.0128	sonnet 95	1.8917	0.0293
sonnet 30	2.7398	0.0031	sonnet 63	2.3511	0.0094	sonnet 96	1.6640	0.0481
sonnet 31	2.6644	0.0039	sonnet 64	2.2629	0.0118	sonnet 97	1.5511	0.0604
sonnet 32	2.6428	0.0041	sonnet 65	2.1737	0.0149	sonnet 98	1.4362	0.0755
sonnet 33	2.5117	0.0060	sonnet 66	2.0835	0.0186	sonnet 99	1.4015	0.0805

Finally, we display a boxplot of t-scores in Table 8 as the sonnet index goes from 0 to 72.

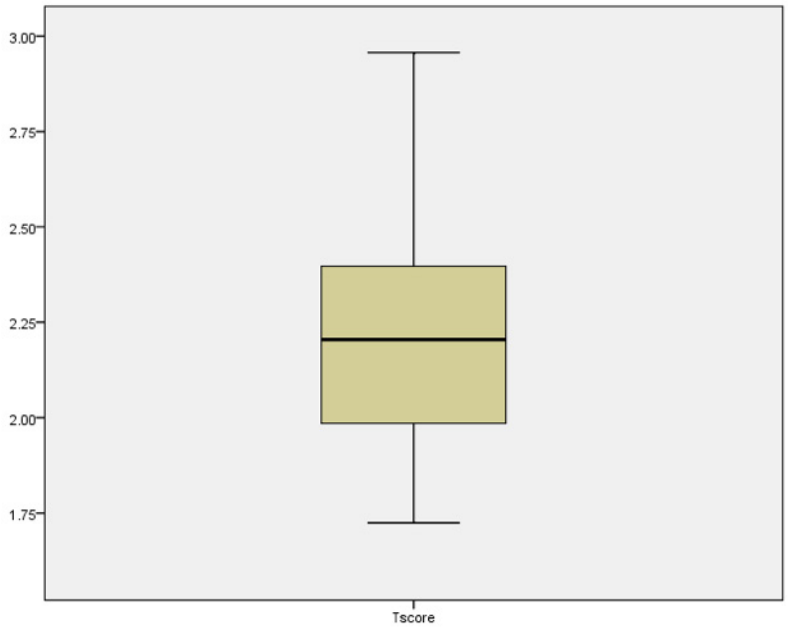

Figure. Boxplot of t-scores in Table 8.

8. MICROSOFT EXCEL COMMANDS USED IN OUR STUDY

A detailed methodology of Microsoft Excel commands use for statistical inference of Sonnets is described in [21] available free from the website indicated there.

8.1. Acknowledgments

The authors are grateful to C. Gamble for discussions.

REFERENCES

1. Abramyan A. *The Armenian Cryptography* (in Armenian), Yerevan University Press, 1974.
2. Ayres R.U. *The Death and Posthumous Life of Christopher Marlowe*, http://robertayres.wordpress.com/2012/08/04/the-death-and-posthumous-life-of-christopher-marlowe-
3. Ballantine R. *Marlowe Up Close: An Unconventional Biography with A Scrapbook Of His Ciphers*, Xlibris, 2007.
4. Cover, T.M. and Thomas, J.A. *Elements of information theory*, second edition, Hoboken: Wiley, 2006.
5. Farey P. *The Stratford Monument: A Riddle And Its Solution*, http://www2.prestel.co.uk/rey/epitaph.htm
6. Friedman, W. and Friedman, E. The Shakespearean Ciphers exposed, Cambridge University Press, 1957.
7. Gamble C.W.H. The Mystery of the Missing 'M', *The Marlowe Society Research Journal* - Volume 10 - 2013.
8. Malyutov M.B. Review of Methods and Examples of Authorship Attribution, *Review of Applied and Industrial Mathematics*, vol. 12, no. 1, 2005, 40-79 (In Russian).

9. Malyutov M.B. Authorship attribution of texts: a review, in *General Theory of Information Transfer and Combinatorics, Springer Lecture Notes in Computer Science, No. 4123, R. Ahlswede et al editors*, 2006, 362–380.

10. Malyutov M., Zhang T., Li Y., and Li X. Time series homogeneity tests via VLMC training. *Information Processes*, 2013, Vol. 13, No. 4, 401–414.

11. Moore D. S., McCabe G. P. and Craig B. *Introduction to the Practice of Statistics, 6th edition*, (Freeman, 2008).

12. Poundstone W. *Fortune's formula*, Hill and Wang, N.Y., 2005.

13. Rassoul-Agha F. and Seppalainen T. *A Course on Large Deviations with an Introduction to Gibbs Measures*, Department of Mathematics, University of Utah, Salt Lake City, 2010.

14. Shannon C., A Mathematical Theory of Communication, *Bell System Tech. J.*, **27**, 379–423, 623–656, 1948.

15. C. Shannon C. Communication Theory of Secrecy Systems. *Bell System Tech. J.*, **28**, 656–715, 1949.

16. Suhov Yu.. and Kelbert M. *Probability and Statistics by Example: Volume 2, Markov Chains: A Primer in Random Processes and their Applications, vol. 2*, Cambridge University Press, 2008.

17. Thompson J.W. and Padover S.K. *Secret diplomacy; espionage and cryptography,1500-1815*, F. Ungar Pub. Co., N.Y, 1963.

18. *The Fundamental Anagram of Calculus*, `http://www.mathpages.com/home/kmath414/kmath414.htm`

19. The "Impossible Doublet" in the Droeshout engraving of William Shakespeare, `https://www.youtube.com/watch?v=gCQt4pOMUqc&feature=youtu.be`

20. MalyutovM., Grosu P. and Zhang T. SCOT Modeling, Training and Statistical Inference, *Proceedings of the 8th workshop on Info. Theoretic methods in Sci. and Engineering, Edited by J. Rissanen et al*, University of Helsinki Department of Computer Science Series of Publications B Report B-2015-1, 31-34., 2015.

21. Zhang. T. A manual of Microsoft Excel commands used for analyzing texts, 2014. `https://dl.dropboxusercontent.com/u/19340569/Shakespeare_Sonnets_Data_Processing.docx`