The following paper is reproduced here with copyright permission from the

Charles Babbage Institute
Center for the History of Information Technology
University of Minnesota

Signature Simulation and Certain Cryptographic Codes

Carl Hammer, Ph. D.
Director, Computer Sciences
UNIVAC
Washington, D.C.

Invited Paper

Third Annual Simulation Symposium
14 January 1970
Tampa, Florida

Signature Simulation and Certain Cryptographic Codes

Carl Hammer, Ph. D.
Director, Computer Sciences
UNIVAC, Wahington, D.C.

Abstract

Three cyphers allegedly authored by Thomas Jefferson Beale in 1822 have been the subject of intensive study for over one hundred years. Generations of cryptanalysts have expended untold man-years, thus far without success, attempting to decode them; vast armies of fortune hunters and treasure seekers have devoted Herculean labors to digging up the rolling hills of Virginia trying to locate the promised bonanza. The history of pertinent activities would fill volumes yet serious students of cryptography have always had nagging doubts about the cyphers' authenticity. It has been alleged that the "known solution" to Cypher Number Two: 115, 73, 24, 818, 37, 52, 49, … ("I have deposited in the County of Bedford about four miles from Buford's in an excavation or vault…") with the aid of an unsanitized version of the Declaration of Independence was merely a superb, imaginative and grandiose hoax perpetrated ages ago for whatever reasons.

Modern computer technology could obviously perform signature analyses on the Beale cyphers and could also, in fact, simulate the process of encoding itself so as to yield new clues and deeper insights into their construction. For the benefit of the uninitiated, the encoding method used in the second cypher employs a specified document whose words are simply numbered consecutively and first letters of these words are sought out at random to match the letters of the cleartext or message. The sequence of numbers corresponding to these matches is then written down as the final code. While primitive, the process has the advantage of relative security until the source document becomes known; at that moment the cypher can be decoded even by second graders.

The work now completed with the help of our UNIVAC 1108 includes numerous analytical studies of the Beale cyphers and various types of simulations. For example, we have turned the entire process of simulated encoding by various schemes over to the machine and analyzed the signatures of these synthetic codes; we have also encoded various messages by hand, using different texts and a variey of methods to obtain their signatures. These simulations provide convincing evidence that the signatures are both process and data dependent; they indicate also very strongly that Mr. Beale's cyphers are for real and that it is merely a matter of time before someone finds the correct source document and locates the right vault in the Commonwealth of Virginia.

Table of Contents

1. Introduction

These cyphers allegedly authored by one Thomas Jefferson Beale in 1822 have been subject of intensive study for over one hundred years. Generations of cryptanalysts have expended untold man-years attempting to decode them while treasure hunters have spent an equal amount of time and effort in digging through the hills and caves of Virginia in an attempt to locate Beale's treasure. During the summer of 1968, several members of the American Cryptogram Association (ACA) decided that a concentrated group study might be successful where individual efforts had thus far failed. in response to several inquiries, eleven persons indicating an interest in this cypher convened in Washington on Saturday, 20 September 1968, to discuss present knowledge, pool talents and resources, and formulate plans for future work. It was unanimously agreed that modern computers should be used to analyse the content of these cyphers in depth, to develop their "signatures," and to simulate the encoding process allegedly used by Beale in his three messages. The group suggested numerous modifications of already existing analytical computer programs and the ideas proposed then were eventually translated into real and working programs.

This report summarizes the work done since then. Naturally, we cannot include all the detailed computer printouts which have accumulated. These printouts, however, can be made available for inspection in our Washington office at any time. We hope that interested parties will avail themselves of this opportunity and that during the year 1970 joint efforts will bring us closer to the solution of this very interesting project.

As previously mentionned, the Beale Cyphers are three numerical codes allegedly constructed during the second decade of the past century by Thomas Jefferson Beale for the purpose of identifying the site of a treasure buried by him. The three codes are shown in Appendix 9. The method used by Beale to encode the second of his three messages was "broken" by a James B. Ward several decades later. It is very much like that already described by Arthur Conan Doyle in "The Valley of Fear." Taking any readily available source document, such as the Declaration of independence, each word in this keytext is numbered sequentially: (1) When (2) in (3) the (4) course (5) of (6) human (7) events … The letters of the message to be encoded are then selected at random from appropriate starting letters of these words. The final code consists thus only of a string of numbers, as in Beale Cypher Number 2: 115, 73, 24, 818, 37, 52, … Correlating these numbers against the keytext, we find that they represent consecutively the letters I, H, A, V, E, … and this combination of keytext and code reveals quite readily the entire message contained in B2:

> I have deposited in the County of Bedford about four miles from Buford's in an excavation or vault six feet below the surface of the ground the following articles belonging jointly to the parties whose names are given in number three herewith. The first deposit consisted of ten hundred and

> fourteen pounds of gold and thirty eight hundred and twelve pounds of silver deposited November 1819. The second was made December 1821 and consisted of nineteen hundred and seven pounds of gold and twelve hundred and eighty eight pounds of silver, also jewels obtained in St. Louis in exchange to save transportation and valued at thirteen thousand dollars. The above is securely packed in iron pots with iron covers. The vault is roughly lined with stones, and the vessels rest on solid stones and are covered with others. Paper number one describes the exact locality of the vault so that no difficulty will be had in finding it.

This is not a tutorial on cryptography and we shall discuss here only the methodology employed in this particular encoding/decoding process. First of all, it is obvious that even if the exact methodology were known, decoding without an exact specification of the keytext may be a very difficult process. Even with the help of advanced cryptographic methods it can introduce obstacles of enormous magnitude. Secondly, the encoding method indicated by Sir Arthur Conan Doyle and Beale's B2 is only one of several possible variations. For example, instead of the first letter of the numbered words, their second letters could be chosen, or their last letter, etc. Then, instead of numbering the words of the basic document, the letters could be numbered sequentially counting or not counting blanks and/or punctuations. The encoder may also introduce a lead or lag function λ such that the code element N actually refers to word or letter number $N \pm \lambda$. At this point it becomes clear that we have at least a major data processing problem on our hands when we encounter cyphers of this type. In fact, it is more than likely that we also have a major cryptographic problem if very little is known about the source of the cypher. In the case of historical cyphers, there is the additional difficulty of locating the authentic documents or, what may even be worse, unsanitized or specialized versions thereof that an author of centuries past may have used. More will be said later about this problem in connection with our own work on Beale Cypers 1 and 3.

2. Three Computer Programs

The power of the Univac 1108 machine was tapped with the aid of several computer programs specially developed for our purpose. The tasks which these programs carry out fall into three categories. The first CRYPTA Program is basically analytical; it takes a string of numbers (i.e. a numerical code) and analyses it with the help of many mathematical-statistical tools. The second CRYPTT Program involves list processing and various decoding attempts at obtaining a concordance between a given numerical code and an alphabetical keytext. The third CRYPTS Program is a computer simulation of the human process of encoding some cleartext with the help of a given alphabetical keytext.

CRYPTA Program Description

This Fortran program performs a number of analytical tasks on a given numerical cypher. In its present form, the inputs are punched cards, the last of which carries a special punch to signify the end of the data deck. The data deck is preceded by one informational BCD card which is used to construct the heading of the outputs.

The program outputs begin with a summary of the data statistics which includes the title (from the header card), the number of entries in the cypher, their numerical average, their root-mean-square, as well as a listing of the raw data. Next, runs-up and runs-down are enumerated, followed by a sort of the data and their first differences. The original data are then reduced modulo 26 (an attempt to correlate them with the English alphabet) and the resultant frequency table is printed out and tested for all possible 26 cyclical permutations but results are only printed out if they submit to a Chi-Square statistical significance test. The data are also cross-summed and the resultant array is printed out. Again, the cross-summed frequency table is compared with English letter frequencies and the results are printed out if they are statistically significant.

Next, the data are subjected to an autoregressive analysis which looks for statistically significant cycles "hidden away" in the raw data. Significant frame sizes are printed out for autoregressive lags ranging from 2 to 30. Respective averages and standard deviations are given for each frame position in statistically significant frame sizes. Finally, a Kasiski-like analysis is performed on the raw data elements by examining their differenced position values in the cypher. This analysis is summarized by listing frequencies of the divisors ranging from 2 to 36. Typical running time of this program with 500 data points, including compilation, is about eleven seconds on the Univac 1108.

2.2 CRYPTS Program Description

This Fortran program encodes a given alphabetical cleartext (which may not contain any numbers or special symbols) with the help of another alphabetical keytext by the concordance or matching process allegedly employed in the Beale cypher. Program output is a listing and a deck of punched cards in the same format used for input into the CRYPTA and CRYPTT programs described elsewhere. This program now has three options.

The first option searches the keytext sequentially, always beginning with its first word, until a match between a given cleartext letter and the first letter of some word in the keytext is obtained; the position number of that word is then recorded. Encoding of the cleartext will thus produce a string of the lowest valued position numbers in the keytext. If no match is found during the complete search of the keytext, dummy numbers 1, 2, 3, .. are successively inserted into that string.

The second option searches the keytext sequentially, beginning with its first word, until a match between a given cleartext letter and the first letter of some word in the keytext is found; the position number of that word is then recorded. However, when the same type letter comes up again for encoding the search for a match resumes at the position last recorded and this process is continued to the end of the keytext before returning to its beginning. If a letter to be matched does not occur in the first letters of the keytext words, the next letter in the alphabet is chosen cyclically, i.e., Z is followed by A.

The third option uses a rectangular random number generator to select matching first letters from the keytext words in the process of encoding the cleartext letters. If a letter required in the encoding process has frequency zero among the first letters of the keytext words, the next letter in the cyclical alphabet, A, B, C, … Z, A, B, .. is chosen to replace any unencodable letter.

2.3 CRYPTT Program Description

This Fortran program performs a number of list processing tasks on a given alphabetical keytext under control of a given numerical cypher. In its present form two punched card input decks are required. The first deck begins with a text header card from which later program outputs are constructed; it is followed by a keytext data deck and an extra control card which signifies its end. The second data deck begins also with a text header card which is followed by cards containing the numerical cypher data. The last card of the numerical cypher data deck carries a control punch signifying its end.

The program outputs first list the two headers to identify the source of the alphabetical keytext and of the numerical cypher data. They are followed by a summary giving the number of words and literals in the textual data and a count of the elements in the numerical cypher. The alphabetical keytext is then printed out with word counts indicated over the first letter of every fifth word. It is printed out again with a letter count indicated over every tenth letter; spaces are not included in this count. Then follows a digram analysis of the text which tabulates frequencies of the digrams in a 26x26 matrix ranging from AA to ZZ with row and column totals; the digram list is also printed out in order of descending frequencies. A letter frequency analysis of the keytext is supplied. Finally, a listing of the numerical cypher elements is given in array form.

At this point the program develops several letter and word concordances from the alphabetical text and the numerical cypher. The first parameterized approach matches the numerical entries of the cypher against corresponding first letters of the search text; it also introduces integer lags into this matching process and the key cypher numbers are systematically incremented by these lags. The outputs from this first approach are printed as pseudotext where blanks replace impossible word or character numbers. In the second approach, the

parameterized matching process creates a pseudotext by reversing the first process and printing the last characters of the words matching the given cypher numbers. Again, an arbitrary lag is introduced both in the word and character counts. the resultant pseudotext is then printed out, allowing for blanks in impossible word or character assignments. In the third approach, the key elements of the cypher are taken to be word position counts. Pseudotext is again created and printed out with an arbitrary incremental lag imposed on the numerical cypher. In the fourth approach, the program matches letter position numbers with arbitrary lags against the numerical cypher. In this approach blanks in the text are not counted. Finally, in the fifth approach, a match is established between the numerical cypher elements and character counts which include blanks. Typical running times of the program with 8000 literals, 500 numerical cypher key elements, and with lags ranging from 0 to 10 are less that one minute on the Univac 1108 system.

3. Simulation Studies and Synthetic Codes

One fundamental question which has permeated the history of the Beale Cypher has been a determination of its authenticity. There are some who believe that it is nothing more than a grandiose hoax, while others firmly believe it is legitimate and will be cracked sooner or later. In order to arrive at an answer to this question, one might proceed from the assumption (statistical hypothesis) that Beale Cyphers 1 and 3 are random doodles while Beale Cypher 2 was constructed to create the basis for a hoax. If it now could be proven that Cyphers 1 and 3 were purely random numbers with a "signature" significantly different from that of Cypher 2, the weight of the evidence would tip the scales toward abandoning hope of ever obtaining legitimate solutions to Cyphers 1 and 3. Several studies along these lines have been conducted and are discussed below.

3.1 Code Simulation with Rectangular Random Numbers

The most primitive assumption which we could make is that Beale Cyphers 1 and/or 3 were written down as sequences of pure random numbers. For example, they could be rectangular random numbers with a given range. In order to test such an hypothesis, we wrote a short Fortran Program to produce strings of rectangular random numbers which are then punched into data cards having the same format as those acceptable to the CRYPTA and CRYPTT Programs. By way of defining rectangular random numbers, note that they have a uniform distribution of such a nature that the occurrence of any number within the stated range is equally likely. As a consequence, the sorted array of these numbers will have no mode and their first differences also tend to be uniformly distributed. The subject program was used to generate one output deck RS (for Rectangular Simulation) which we used to test the earlier stated hypothesis (cf. Section 4.4).

3.2 Code Simulation with Poisson Random Numbers

Even a most cursory inspection of the three Beale Cyphers B1, B2, and B3 reveals that the numbers are not uniformly distributed. Therefore, a second random number generator program was developed whose output is Poisson distributed. This distribution occurs widely in psychological tests, communications, and other natural and engineering phenomena. According to this distribution law, smaller numbers will occur more frequently than larger numbers and one such set of data was generated. The output deck PS (for Poisson Simulation) is arranged in a format acceptable to the CRYPTA and CRYPTT Programs. The results obtained with PS are described in Section 4.5.

3.3 Hammer's Simulation of the Beale Process

The data decks RS and PS described above are strictly statistical simulations of some assumed random number distribution law, permitting us only to test the Beale Cyphers against these hypothetical distributions. It was felt that we should go one step further and actually simulate the Beale process itself.

We chose a paragraph of text (randomly from a speech recently published by this writer) and proceeded to encode it by the alleged Beale Process. A listing of Beale's Version of the Declaration of Independence was used as the keytext for encoding this cleartext. The annotated keytext carried word numbers for every fifth word throughout and was printed on typical computer output paper. As a matter of fact, it was the by-product of one of the earlier CRYPTT runs.

We scanned the text sequentially to find the required letters, writing down word position numbers as their first letters were found to match the looked-for letters. We would proceed for a while in this fashion then turn to another section of the keytext and continue the sequential search-and-match process. While carrying out this work, we noted the mental strain in searching for rare or non-existent letters. We were also tempted to "memorize" position numbers for certain letters and we developed a resistance to turning to later sections of the keytext. However, we made a conscious effort to switch to different areas of the text, not previously used, and we worked the text always in a forward-search mode, never in reverse. Naturally, all these psychological factors would be reflected in the "signature" of the produced string of numbers, as indicated by the analysis in Section 4.6.

3.4 Caldwell's Random Data Code

One member of the study team who submitted a "personalized" random data code was Mr. Robert Caldwell. Nothing is known (on purpose) about the method which he used to produce these CS data. Upon receipt of his manuscript, data

cards were punched, using the format acceptable to our CRYPTA and CRYPTT Programs. These decks were then subjected to computer analysis.

3.5 Nelson's Random Data Code

Mr. Carl Nelson, another member of the study team, has also submitted a "personalized" random data code NS about whose source we did not inquire. His code, too, was punched up and subjected to analysis by the CRYPTA Program.

3.6 Three Synthetic Codes Generated with CRYPTS

It was a simple matter to take the same text used in Hammer's simulation HS and encode it with the three CRYPTS options. The three optional outputs S1, S2, and S3 could then be analyzed with the help of CRYPTA (Section 4.9) and could be validated by running them against CRYPTT. Naturally, there would be no startling results forthcoming from this last experiment since the three codes were known to have a solution. They would serve only as a benchmark in the analysis of codes B1 and B3 of unknown origin.

4. Analytical Studies with the CRYPTA Program

Including the three original Beale Cyphers, our stockpile of real or simulated codes now contains eleven sets of data (cf. Section 3) allowing exhaustive analysis with the CRYPTA program. Of maximum interest, of course, is a comparison of measurable statistical parameters for simulated and real codes. We shall highlight significant differences in these parameters reflecting structural or signature information in the following.

4.1 Beale Cypher No. 1

The B1 cypher has 520 entries with a mean of 273 ranging from 1 to 2906 (cf. Appendix 9.1). The distribution of runs-up-and-down varies significantly from random data; there is an excess of runs-down of length one which is compensated for by a shortage of runs-down of lengths three and higher. This provides a possible clue to the construction of the cypher as the author might have "jumped back" more frequently while engaged in the encoding process. Autoregressive analysis reveals only one significant pattern (of length sixteen) but by itself this fact does not give rise to any suspicions about the nature of the code. Modulo reductions of the code numbers yields twelve significant variations from randomness indicative of the difficulties which the author might have experienced in selecting keytext letters to encode his message. Significantly, most of these parameters are even; this fact might indicate that the author numbered only every other word of his keytext and then fell for the psychological preference for numbered over unnumbered words.

4.2 Beale Cypher No. 2

The B2 cypher is longer than B1. It has 763 entries ranging from 1 to 994 with a mean of 162 (cf. Appendix 9.2). It has, of course, a known solution with Beale's Version of the Declaration of Independence as keytext. Again runs-down of length one dominate and are compensated for by too few runs-down of length three and up. Autoregression reveals three significant cycles of lengths three, five, and seventeen which is about right for a hand-coded job. Modulo reduction indicates again a significant deviation from randomness with 21 parameters but this time only half of them are for even numbers, indicating that the method of encoding chosen was "better balanced".

4.3 Beale Cypher No. 3

The length of cypher B3 falls between B1 and B2. It has 618 entries ranging from 1 to 975 with a mean of 153. However, runs-up-and-down extend far beyond the range to be expected from random numbers with three runs-up of length nine and also a shortage of runs-down of length three and greater. In view of this we might suspect that earlier "practice" or a change in the mental approach to the encoding task could account for this unusual pattern. Also, letter-counting instead of word-counting might produce such a pattern. Autoregression produces four significant cycles of lengths three, five, eleven, and seventeen. By comparison with B1 and B2 we find that the longest of these cycles reflects the personal "signature" of the author. Modulo reductions of the cypher entries yield six significant values without predominance of odd or even; this time multiples of five dominate, possibly indicating that the numbering scheme used for encoding of this cypher differs from that used in B1 or B2.

4.4 Rectangular Random Number Code

This first of the computer-generated benchmarks has 500 entries whose mean is similar to that of B1 (by design). While there is nothing interesting about the runs-up-and-down, the logarithmic sort fit yields parameters significantly different from the three Beale Cyphers, indicating clearly that the latter come from a non-random source. Autoregression yields only two significant frequencies of 2 and 3; these can be traced to the periodic nature of the random number generator itself. Modulo reduction, likewise, yields only one value of significance against many values for the real cyphers. At this point it becomes thus obvious that none of the Beale Cyphers was constructed with the help of early random number tables, or by tossing coins or rolling dice.

4.5 Poisson Random Number Code

The generation of Poisson distributed numbers yields several surprises. While their runs-up-and-down are similar to those obtained from the uniform random numbers, the logarithmic sort fit resembles much more the data obtained from

the three Beale Cyphers. Autoregression yields three significant values of 3, 7, and 23. The individual values can be traced to the idiosyncracies of our random number generator. The fact that three of them show up yields a good benchmark for comparison with other simulation schemes. Finally, modulo reduction yields no significant values which indicates once more the non-randomness of the Beale Cyphers.

4.6 Hammer's Simulated Beale Type Data

In addition to having a known solution, these data correspond most closely to B3. There is a preponderance of runs-down of length one, compensated for by too few runs-down of lengths greater than three. This is evidence for the manner in which the encoding process was carried out, namely by running forward along the keytext, and jumping back whenever the urge struck us. Evidently, at least this encoder was unable to control his tendencies in that direction, producing long runs-up against short runs-down. Autoregression yields also five significant cycles, including two of great length, as did B3. On the other hand, modulo reduction produced only two significant values, due probably to the fact that we tried very hard to avoid a preference for the numbered elements of the comparison text.

4.7 Caldwell's Random Data Code

The source of these data and the method of their generation is not known. Comparison with several benchmarks of computer-generated or man-made simulated codes would indicate that they correspond most nearly to RS or S3, which suggests that their basis is indeed a set of true random numbers. Runs-up-and-down, autoregressive parameters, and modulo reduction all point in that direction.

4.8 Nelson's Random Data Code

Here we have an entirely different pattern from the one generated by the Caldwell data. While runs-up-and-down (except for one run-down of length 6) resemble most clearly a random pattern (say, of the Poisson type), autoregressive analysis reveals the non-randomness of the data to be very much like Hammer's simulation or the original Beale Cyphers. Also, modulo reduction produces seven values of significance, distributed very much like those of B3. All this evidence leads us to suspect that this cypher is based upon a real text and contains a real message.

4.9 Three Codes Generated with CRYPTS

As mentionned in Section 3.6, this program has three options and it simulates (on the computer) the encoding process that a human being might want to employ. The three codes produce no spectacular or unexpected results. Runs-

up-and-down for option one (after a hit, return to beginning of keytext) indicate a dominance of runs-down for length one, as expected; frequencies of longer runs adjust themselves accordingly. Only opion two (scan entire text and then return to beginning) yields significant cycles for autoregression. The logarithmic sort fit for option three (uniform selection of matching letters) produces constants which differ significantly from those obtained by the other options. Again, option one yields a large number of significant parameters (22) during modulo reduction as compared with just one such parameter for the other two options. These results proved to be very valuable when we tried to "bracket" the unknown cyphers B1, B3, CS, and NS (cf. Section 6).

5. Decoding Studies with CRYPTT

As mentionned earlier, this computer program eliminates the rote and drudgery connected with setting up concordances between a given numerical cyper (such as the Beale Codes) and a keytext which might yield the cleartext. The CRYPTT Program provides additional statistics about the chosen keytext, of value in further analytical studies. The CRYPTT output subroutines yield only interesting garbage unless the chosen keytext happens to be the "correct" one. However, even the "signature" of this garbage can be rather revealing as we discovered.

5.1 Beale Cyphers Nos. 1 though 3

Beale's Version of the Declaration of Independence (See Appendix 9.4) as keytext produces alphabetical strings but little intelligence for any of the available CRYPTT methods or their variants. For example, if we correlate the numbers of the cypher with the initial letters of correspondingly numbered words in the keytext, we obtain SCS ETFA GSDOTTUCWOTWTAATWDBIIDTTWTTAABBP LAABWCT… The spaces result from code numbers exceeding the number of words in the keytext; the computer program maintains a count of these words and inserts blanks where code numbers exceed the upper limit of numbered text words. Letter frequency counts indicate that this pseudotext does not agree with the letter frequency counts of ordinary English. Similarly, if we number the letters in the basic text, rather than the words, CRYPTT produces another string of seemingly random letters: EONECTONBOESOOHCDELSCSTBOSSHWHEERS CLCSDEESTAILMCAREHD… There may occur chance digraphs and trigraphs in such a string of letters; the partial "pseudotext" shown above contains such words as one, ton, so, boss, rail, care, etc. The letter frequency distribution in this pseudotext matches that of the English language much more closely; also, there are no blanks since the keytext has 6527 literals (not counting spaces) and the highest code number in Beale Cypher No. 1 is 2906.

No further progress is made by applying lags or leads to this cypher. For example, its beginning of: 71, 194, 38, 1701, … could refer to the first letters of words 70, 193, 37, 1700, … or to letters 73, 196, 40, 1703, … etc. However, analysis of the pseudotexts obtained with such lags or leads indicates a much

better match with English letter frequencies for the method of letter-counting over the method of word-counting. In this connection, remember that the method of word-counting is the one that "breaks" Beale Cypher No. 2. Nevertheless, letter-counting seems to provide a better option for decoding B1.

Beale Cypher No. 2 can be "broken" by applying the method of unlagged word-counts against the keytext of the Declaration of Independence. All other methods (with or without lags and leads) available under CRYPTT produce interesting benchmarks for letter frequencies against which we can test other decoding methods. For example, if another keytext produces letter frequencies that are statistically less significant than those produced by the Declaration of independence plus an inapplicable method, we may take this clue to mean that we have the wrong method and/or the wrong keytext. More about this interesting facet of our signature analysis in Section 6.

Beale Cypher No. 3 does not yield anything new and startling. Neither word-counting nor letter-counting produces anything but gibberish but the relevant letter frequencies are again better for the letter-counting method.

5.2 Hammer's Simulated Beale Type Data

These data are based upon a cleartext of modern English writing; their submission to CRYPTT with the Declaration of Independence produces, of course, the correct solution. It furnishes us also with benchmarks for the distortion of letter frequencies resulting from introduction of lags or leads or an incorrect encoding method. For example, decoding by the correct method will produce the cleartext: COMPUTERSHAVESTARTEDTODOAMAZINGTHINGS THEYDESIGNAIRPLANESGUIDEMISSILES… Introduction or a lead of one unit produces the following pseudo-cleartext: AFIESOLEEIAXTRCEOXAOPIPNEM ISEOOATHOAXNEVET… As before, code numbers in excess of the available word-count in the keytext are translated by the computer into blanks. Very few recognizable digraphs or trigraphs result from this simple shift; similarly, letter frequencies are immediately distorted from that expected for ordinary English.

5.3 Caldwell's Random Data Code

Application of the Declaration of Independence against these data produces meaningless letter strings for all methods and relevant lags or leads. For the unlagged letter-count method, we have BOMHHLOLNS T B WP PJTTA TDA NDOFWN TAIPIAD T IPG R S … Both the frequency of recognizable digraphs or trigraphs is very low as is the correlation of resultant letter frequencies against those of the English language alphabet. Our earlier stated suspicion thus receives support: In all likelihood these were true random numbers. They are certainly not derived from a cleartext encoded against the Declaration of Independence by the Beale method.

5.4 Nelson's Random Data Code

The Nelson Cypher, unlike Caldwell's Random Data Code, produces strings of letters with frequencies similar to those obtained in other attempts where we deliberately matched data against the "wrong" keytext. For example, the word-counting method, without lag or lead, yields against the Declaration of Independence: OTTSO SOWOC RTRNHSAWAOTI THTBAATLAONIPAANR… Letter frequency counts of the pseudotexts tend to confirm that we have here a legitimate cypher but the wrong document. More about that under Section 6.

5.5 Codes Generated with CRYPTS

Application of a computer-simulated method of encoding will, of course, yield the desired cleartext for the right choice of method. Thus the several options of CRYPTS will lead us to the correct solution, say for the unlagged word-counting method: COMPUTERSHAVESTARTEDTODOAMAZINGTHINGS… which converts for a lag of one into: OFAENHVEEUNAVEHNEHVIHFIF… Letter frequencies here indicate clearly the transition from solution to non-solution even though we have the right text. Again, this result will prove of interest when we attempt to summarize our results in Section 6.

5.6 The Magna Carta

During the latter part of 1965 we had a visitor who stated in no uncertain terms that he had "broken" Beale Codes 1 and 3 but needed the assistance of our computer to complete his work. We ignored the "minor" contradiction contained is his statement and pressed him for further information.

He told us, reluctantly, to "try the Magna Carta" -- and disappeared. We syllogized the following conclusions: (i) He was telling us the simple truth; it is the Magna Carta, (ii) He was leading us down the garden path; It is not the Magna Carta, (iii) He was being super-devious; it is the Magana Carta but he thinks that we think he gave us a false lead, thus we won't try it.

At least this document would provide us with another simulation tool which we had not tried before. In what language was the source document written? If we could make any positive satements by additional, statistical analyses, we would indeed have another piece of information needed to solve the Beale mystery. We therefore studied in depth decoding attempts with CRYPTT, using several versions of the Magna Carta as possible source documents.

There is no need to relate in detail the extensive literature available on the Magna Carta. There exist many versions authored by John in AD 1215 and by Henry in AD 1225. Both documents are in Latin and both begin with a lengthy Preamble which constitutes the King's authority and lists his fiefs and advisors. Authorized translations were used by Thomas Jefferson and his friends in the

drafting of our own Constitution. If the Beale legend is not myth, the author of these cyphers would have known these documents. A short, sanitized version (both in Latin and English) of John's Preamble can also be found in the Encyclopaedia Britannica.

Table 1 lists some details about the four Latin and the eight English versions of the Magna Carta which were used as keytexts in relevant decoding attempts of Beale Cyphers 1 and 3, discussed in Section 6.2. For the sake of completeness, we have also listed two additional documents, the Declaration of Independence and a modern English text. These were used to establish further benchmarks and to allow us a better assessment of the results of the other CRYPTT runs. The indicated run numbers in the first column of that table refer respectively to runs made against Beale Cyphers 1, 2, or 3, as shown in the second column of the table. Some data about these keytexts, such as number of words or number of letters, are provided in later columns. The table also indicates the source documents and any variants chosen for a particular run. The results obtained from these and additional runs will be discussed in Section 6.2.

6. Analysis of Results

Our studies are largely predicated upon the two programs CRYPTA and CRYPTT and upon computer runs made with the several data decks described earlier. Having already mentionned briefly some of our findings, we shall now try to consolidate our position and review pertinent data in detail.

6.1 Summary of CRYPTA Data

Table 2 provides all of the data obtained for the three Beale Cyphers (B1, B2, B3), Rectangular and Poisson Distributed Random Data Simulations (RS, PS), Simulations contributed by Hammer, Caldwell, and Nelson (HS, CS, NS), and Computer Generated Simulations obtained from the three CRYPTS Program Options (S1, S2, S3).

The first segment of Table 2 gives the number N of data elements in the respective codes, their mean, standard deviation, and range. The second segment of the table summarizes the number of runs-up-and-down. The third segment indicates parameters obtained for the logarithmic sort fit: Intercept A, slope B, and standard deviation S. The fourth segment lists significant cycles detected for autoregressive analysis by length and number; thus 3, 5, 17(3) for Beale Cyper B2 indicated 3 significant cycles of lengths three, five and seventeen. The fifth segment of the table indicates results of the Kasiski analysis; it lists excesses for significant divisors. For example, under Beale Cypher 2 we find only one such divisor, namely 7, which occurs to a greater degree that can be expected if the data were random. The sixth and final segment of the table summarizes results obtained from modulo reduction; significant moduli are listed

TABLE 1

CRYPTT COMPUTER RUNS MADE WITH THREE BEALE CYPHERS AND VARIOUS BASIC TEXTS

Runs	Beale	Language	Source Document	Author	Date	Document Source	Words	Letters	Overall Description, Comments
1-21-41	1-2-3	Latin	Magna Carta	John	1215	Encyclopaedia	180	1253	Preamble Only
2-22-42	1-2-3	Latin	Magna Carta	John	1215	Stubbs, p. 292	363	2337	Preamble and Text
3-23-43	1-2-3	Latin	Magna Carta	Henry	1225	Stubbs, p. 350	284	1725	Preamble and Text
4-24-44	1-2-3	Latin	Magna Carta	John	1215	Stubbs, p. 284	344	1990	Preamble Omitted
5-25-45	1-2-3	English	Magna Carta	John	1215	Encyclopaedia	228	1140	Preamble Only
6-26-46	1-2-3	English	Magna Carta	Henry	1225	Swindler	676	2852	Preamble and Text; Italics, no brackets
7-27-47	1-2-3	English	Magna Carta	John	1215	Swindler	692	3010	Preamble and Text; Brackets, no italics
8-28-48	1-2-3	English	Magna Carta	Henry	1225	Swindler	671	2806	Preamble Omitted; Bold, italics
9-29-49	1-2-3	English	Magna Carta	Henry	1225	Swindler	456	2280	Preamble Omitted; no brackets
10-30-50	1-2-3	English	Magna Carta	John	1215	Swindler	576	2390	Preamble Omitted; Bold, no italics
11-31-51	1-2-3	English	Magna Carta	John	1215	Swindler	355	1439	Preamble Omitted; Bold, brackets, italics
12-32-52	1-2-3	English	Magna Carta	John	1215	Swindler	269	1100	Preamble Omitted; III Omitted, "Forever" one word; brackets, italics; variation on runs 11-31-51
13-33-53	1-2-3	English	Declaration	Beale	1789	Committee	1322	6527	Beale's version; Declaration of Independence
14-34-54	1-2-3	English	Essay	Hammer	1968	UNIVAC	522	2656	Modern Text; Calibration

and counted. For example, Beale Cypher 3 has six significant moduli, namely 5, 10, 15, 20, 21, and 25, which occur too frequently and must be attributed to some property of the data.

A grandstand view of this table allows us to compare codes with similar properties and to classify these codes as falling between other codes with known properties. If we look at runs-up-and-down, we observe that all Beale Cyphers have a significant deficiency of runs-up of length one, the expected numbers of length two, and a significant deficiency of longer runs-down. Now we note that a similar pattern obtains for HS and S1, indicating that in constructing these codes Beale and Hammer must have acted very much alike. Recalling our earlier comments (Section 4.6) about the psychological forces which become active when encoding messages by this method, we suspect that Beale must have worked, as we did, with a long document but that he jumped backward more often than forward. In other words, he worked his way down the document, then jumped to another spot, worked down again, but probably never reversed his process by working concordances backwards or up.

The logarithmic sort fit indicates that the Beale Cyphers resemble much more closely Poisson-distributed Random Numbers (PS) that rectangular ones. The S2 technique (which scans the entire keytext for concordances before returning to the starting point) falls short of all three Beale Cyphers; the S1 technique (which returns after each hit to the starting point) rests firmly above them, there being little significant difference between the observed slopes. Again, we have here an indication that Beale proceeded in jumps, like a human, and less efficient that our patient computer.

Autoregression indicates that Mr. Beale was rather successful in Cypher 1 as he chose his matches without producing significant short cycles. However, B2 and B3 do contain hidden cycles which resemble very much those of Hammer's simulation HS and are bracketed by computer simulations S1 and S2, as before. Here we also find that Nelson's data (NS) look very much like those of a real code, while Caldwell (CS) might have used a very efficient computer random number generator or good tables.

Kasiski analyses were only included on the outside chance that Beale Cyphers 1 and 3 had been constructed by an entirely different method than B2. The absence of significantly abundant divisors in B1 and the two divisors of 14 and 19 under B3 may be considered as spurious, just as the several divisors detected in the simulations. This finding rules out a large class of encoding methods commonly employed during that time, including Beaufort, Gronsfeld, Porta, and other periodic cyphers.

Finally, the results of modulo reductions indicate once more rather strongly the non-random nature of the Beale Cyphers. All three exhibit a significant

TABLE 2

COMPARATIVE STATISTICS: CRYPTA - SIGNATURE ANALYSIS

IDENTIFICATION		B1	B2**	B3	RS	PS	HS**	CS	NS	S1**	S2**	S3**
N		520	763	618	500	500	424	500	504	424	424	424
Mean		273	162	153	270	285	430	928	433	39	295	630
Sigma		356	202	177	157	298	353	624	492	1120	276	387
Range		1-2906	1-994	1-975	1-546	1-1878	1-1291	1-2028	1-1999	1-818	1-1313	2-1322
Runs*	1	79/99	135/156	78/113	115/116	116/115	58/67	93/116	78/86	76/98	86/92	89/96
Up/Down	2	44/50	70/69	41/41	39/46	45/39	28/29	52/38	58/48	46/32	43/43	46/33
	3	24/7	30/13	17/9	14/9	11/14	17/8	11/7	16/15	12/7	11/8	6/11
	4	9/2	7/5	14/2	4/2	1/4	5/7	6/3	3/4	4/1	4/0	3/3
	5	2/0	2/0	6/1	1/0	0/1	3/2	3/1	0/0	2/0	---	0/1
	6	1/0	0/1	4/1	---	---	1/1	---	0/1	0/1	---	---
	7	---	---	4/0	---	---	0/0	---	---	---	---	---
	8	---	---	0/0	---	---	2/1	---	---	---	---	---
	9	---	---	3/0	---	---	---	---	---	---	---	---
Log.	A	0.108	0.074	0.327	0.603	0.228	0.147	0.872	0.061	0.041	0.435	1.989
Sort	B	1.340	1.236	1.035	1.105	1.255	1.456	1.251	1.528	1.117	1.188	1.071
Fit	S	0.462	0.453	0.370	0.107	0.266	0.195	0.063	0.495	0.761	0.236	0.063
Auto-regression		16(1)	3,5, 17(3)	3,5,11, 17(4)	2,3(2)	3,7, 23(3)	2,3,5, 17,19(5)	18, 23(2)	3,4,13, 19(4)	5(1)	3,11,14, 16,17(5)	3(1)
Kasiski		-(0)	7(1)	14,19(2)	-(0)	20,29, 33(3)	11,31(2)	13(1)	29(1)	-(0)	28(1)	7(1)
Modulo Reductions		2,4-6,8, 10,12,16, 20,22,24, 25(12)	2,4-8, 10,11, 13-25 (21)	5,10,15, 20,21, 25(6)	25(1)	-(0)	9,25(2)	25(1)	9,11,15, 18,20, 22,25(7)	4-25 (22)	25(1)	25(1)

Notes: * Run Up/Down of Length 7 or greater indicate significant deviations from a controlled or random process.

** Known solution with Beale's version of "Declaration of Independence" which has 1332 text words and 6527 letters.

preference for certain numbers. B1 abounds in even numbers, while B3 prefers numbers divisible by 5. No such clearcut separation exists for B2 which strengthens our conviction that B1 and B3 were written is a slightly different manner from that used to encode B2. Note that S1 and S2 bracket the three Beale Cyphers which gives us once more an indication that he used a process whose severity lies between the maxims used in these two computer simulated codes.

6.2 Summary of CRYPTT Data

Table 3 lists the results obtained from a Chi-Square analysis for letter frequencies of the various pseudotexts compared with standard English. The table has four major headings for the code data resulting from the three Beale Cyphers and the Poisson-generated random data. Each of the fourteen runs listed first in this table was set up identically to produce various pseudotexts from the stated keytext and the respective Beale Cypher, or the random code PS. Under Method I we created output using the word-counting method and went though a range of lags (or leads) from -5 to +5, for a total of eleven pseudotexts from the first letters of these words. We also produced outputs for second, third, etc. letters but they did not add anything new and are not shown. Under Method II, we used last letters and leads ranging from 0 to 2 for a total of three pseudotexts; again, second-last and other letters are not shown here although we did produce them. Under Method III we used the letter-counting method, ignoring spaces and punctuations, with lags (or leads) ranging from -10 to +10 for a total of 21 pseudotexts per run. Finally, under Method IV we used again the letter-counting method, including also spaces between words but ignoring punctuations, with a lag (or lead) ranging from -2 to +2 for a total of five pseudotexts. Thus all entries in Table 3 reflect averages over the number of pseudotexts produced by any of these four methods.

For reference purposes, the left side of the table also lists the languages of the keytext; the respective run numbers were also shown in Table 1. The first fourteen lines of the table show individual results for each text; later segments of the table provide information on certain averages. For example, line 1 of Table 3, referenced against Table 1 can be partially interpreted as follows: Run 1 against Beale Cypher 1 with the Latin Preamble to John's version of the Magna Carta, as listed in the Encyclopaedia Britannica produced eleven pseudotexts by Method I with an average Chi-Square deviation from standard English letter frequencies of 346. Method II with an average of over three pseudotexts yielded a Chi-Square value of 698. The same document yielded 21 pseudotexts for Method III with a Chi-Square average of 175. A Chi-Square of 177 was finally obtained by Method IV for five pseudotexts.

TABLE 3

CHI-SQUARE SIGNIFICANCE TESTS ON LETTER FREQUENCIES FOR FOUR DECODING METHODS

Line	Lan-guage	Beale Cypher No. 1				Beale Cypher No. 2				Beale Cypher No. 3				Poisson Random Data			
		I	II	III	IV	I	II	III	IV	I	II	III	IV	I	II	III	IV
1	L*	346	698	175	177	555	1082	374	308	487	970	248	249	252	451	151	124
2	L*	380	775	173	142	590	1213	372	348	519	1015	237	240	321	646	145	154
3	L*	362	658	192	150	617	1154	368	386	571	946	248	245	333	498	153	163
4	L	721	587	165	171	1329	905	299	449	951	826	193	182	564	558	146	123
5	E*	370	479	128	103	676	868	302	190	545	724	193	146	285	348	93	83
6	E*	368	511	99	104	634	757	266	210	525	659	145	148	349	416	71	59
7	E*	461	532	126	118	810	997	297	209	614	872	192	180	405	496	88	68
8	E	289	439	74	61	461	719	160	172	341	529	106	91	270	365	65	59
9	E	338	448	81	77	644	804	172	195	491	571	113	93	302	395	71	66
10	E	269	360	79	68	465	561	180	138	368	507	120	105	278	360	67	63
11	E	367	365	74	62	725	667	165	183	549	464	92	84	310	329	55	41
12	E	346	356	75	62	730	609	168	184	539	449	93	82	301	278	54	40
13	E	355	347	54	62	514**	624	147	147	462	418	71	68	278	315	50	58
14	E	305	236	61	45	532	464	166	149	416	388	92	84	286	248	43	42
1-3	AV	363	710	180	156	587	1150	371	347	526	977	244	245	302	532	150	147
5-7	AV	400	507	118	108	707	874	288	203	561	752	177	158	346	420	84	70
8-12	AV	322	394	77	66	605	672	171	174	458	504	105	91	292	345	62	54
13-14	AV	330	292	58	54	523*	544	156	148	439	403	82	76	282	282	46	50
1-4	AV	452	680	176	160	773	1088	353	372	632	939	232	229	368	538	149	141
5-12	AV	347	407	85	76	619	707	203	178	485	558	122	108	306	355	66	58
1-14	AV	377	485	111	100	663**	816	246	233	527	667	153	143	324	407	90	82

Notes: * With Preamble

** Known Solution in this Group with $X^2=73$

Later segments of this table provide further averages on these Chi-Square values. AV 1-3 takes care of Latin versions of the Magna Carta with preamble. These averages are readily compared with Line 4 where we have a Latin text with the preamble omitted. AV 5-7 provides information about English versions of the Magna Carta including its preamble, while AV 8-12 summarizes results obtained for English versions under omission of the preamble. In this segement of Table 3, AV 13-14 summarizes results obtained from other English texts, namely Beale's version of the Declaration of Independence and a modern English text; both of these texts were chosen on purpose to provide a basis for further benchmarks.

All Latin basic texts are averaged once more under AV 1-4 while all English texts are averaged out under AV 5-12. The last line of this table provides a grand average of all observed Chi-Square values. While there may be some question about normalization of the results of these Chi-Square tests between various keytexts. i.e., reading Table 3 from left to right, vertical comparisons can always be made without introducing any correction factors. Statistically speaking, this correction factor would involve division by the number of actual (e.g., not blank) data points in each cypher. A similar problem of normalization arises between methods.

Even a cursory examination of Table 3 reveals at once striking differences between Latin and English texts, AV 1-4 and AV 5-12, respectively. In all cases, Latin texts produce significantly poorer letter frequencies in the pseudotexts. It turns out that the derived letter frequencies are equally poor with or without the preamble. Therefore, we can conclude with a very high degree of confidence that the keytext was not Latin. In itself, this answers one of our earlier questions and contributes somewhat to the inferential knowledge we now possess about the Beale Cyphers.

Now we shall turn to a comparison of methods. Basically, Methods I and II refer to word-counts while III and IV refer to letter-counts. Let us first rule out Method II which uses last letters of numbered words. Not shown in this table are results obtained for second-last, third-last, etc. letters but they exhibit a like pattern. With few exceptions, Method II produces consistently results which are significantly worse than Method I, allowing us to discontinue any further consideration of Method II.

Now, there remain the original Beale word-counting Method I and two letter-counting Methods III and IV, without or with spaces between the words counted. As a benchmark, we have a Chi-Square level of 73 for the known solution to Beale Cypher 2, as indicated in a footnote to the table. We also have a benchmark from line 14 which introduces, on purpose, a keytext which could not possibly have been known to Mr. Beale. Focussing our attention on AV 5-12 for all English Magna Carta Texts, we find that Method I over Method III provides a 3.05 reduction in Chi-Square averages which is due to the change in number of

observable data elements, for Beale Cypher 2. However, for B1 and B3 comparable reductions of 4.09 and 3.97 are obtained, while the PS random data yield a ratio of 4.64. Significant changes in respective ratios are observed for AV 1-14, but not for AV 1-4 and some others. Thus we have detected a structural difference between B1 and B3 on the one hand and B2 on the other. This difference is both language and data dependent. Using the random data PS as a benchmark, we find that B1 and B3 differ significantly from the expected reduction values for AV 13-14 which contains two English texts but not the Magna Carta. Therefore, we now look at line 13 and find that B2 reduces by 3.5 while B1 and B3 reduce by 6.58 and 6.50 respectively, while the random data PS yield only a coefficient of 5.55. This is the desired clue: For at least one document (Beale's version of the Declaration of Independence) Method III produces significantly better letter frequencies (albeit still garbled in the pseudotext) than expected, for Beale Cyphers B1 and B3, by comparison with Method II for B2.

A similar analysis of Method IV does not yield the same kind of significant change in comparisons between the random data PS and B2 versus B1 and B3. This observation rules out Method IV which we had only included for the sake of completeness. When setting up our programs we had never assumed that the cyphers' author would go to all the trouble of counting letters and spaces and/or punctuation marks. Such techniques have only recently assumed a dominant position because electronic data processing devices can handle such chores more efficiently than man.

7. Conclusions

The solutions to Beale Cyphers 1 and 3 have remained undiscovered despite time and the tenacity of many analysts. We would certainly have gone into deep shock if one of our CRYPTT runs had come out in cleartext and provided us with the geographical coordinates of the alleged treasure! In fact, had such been the case it is doubtful that this paper would have been written. Rather, the entire group responsible for this project would have shouldered pick and shovel and taken off for the Virginia hills. Nevertheless, the results obtained from this simulation study have contributed greatly to a better understanding of these cyphers and produced what we consider significant results.

By way of summarizing our findings, the following are statements of facts, as of this date:

(i) Beale Cyphers 1 and 3 are "for real." They are not random doodles but do contain intelligence and messages of some sort. Further attempts at decoding are indeed warranted.

(ii) The method used for encoding cyphers 1 and 3 is similar to that used for cypher 2. It is very probable that a letter-counting, rather than word-counting method was used. If it was indeed a letter-

count, then spaces between letters and punctuation marks were not included in the count.

(iii) The basic text for encoding cyphers 1 and 3 is not Latin. There is enough evidence to assume it was English. Other languages have not been subjected to the type analysis indicated here.

Thus we return to the drawing board, as it were, hoping that sooner or later someone will find the right text(s) with which to decode Beale Cyphers 1 and 3. We doubt very much that it was the Magna Carta (English version) as suggested by our Pennsylvania visitor. Others before us have already ruled out the Declaration of Independence. However, with the help of our Univac 1108 CRYPTT programs it should be a simple matter to establish a systematic procedure and to test routinely any and all documents of relevance. If and when the mystery is finally resolved, we may have to eat much crow in view of the statements made above. Much time has elapsed since Beale buried his treasure and many people must have passed THE spot. It is quite likely that upon locating the vault it will be found empty. Whether located accidentally or by breaking the code, the first successful treasure hunter is not likely to reveal his find. It is only the second successful treasure hunter that will surface and cause a great deal of disappointment … except in the circles of professional cryptanalysts!

8. Bibliography

1. The Cryptogram, Official Publication of the American Cryptogram Association, 9504 Forest Road, Bethesda, Maryland 20014 (Sesame, May-June, 1968.)
2. Sir Arthur Conan Doyle, The Complete Sherlock Holmes, Doubleday & Company Inc., New York, 1953 (The Valley of Fear, pp. 903 ff.)
3. Encyclopaedia Brittanica, Declaration of Independence and Magna Carta.
4. Helen Fouché Gaines, Cryptanalysis, Dover Publications Inc., New York, 1956
5. Carl Hammer, Private Communications, Beale Cypher Study Committee, 2121 Wisconsin Avenue, N. W., Washington, D. C. 20007.
6. Frances Beale (Smith) Hodges, The Genealogy of the Beale Family; 1399-1956, Ann Arbor, Michigan, 1956
7. P. B. Innis, The Beale Fortune, Argosy, August 1964.
8. David Kahn, The Codebreakers, The MacMillan Company, New York, 1967
9. Al Masters, Has the Beale Treasure Code Been Solved? True Treasure, September-October 1968
10. William Stubbs, Select Charters of English Constitutional History, Oxford at the Clarendon Press.
11. William F. Swindler, Magna Carta - Legend and Legacy, the Bobbs-Merrill Company Inc., New York.

9. Appendix

In the interest of brevity, the appendices are not reproduced in this pdf version. They are the following: