The use of Project Gutenberg and hexagram statistics to help solve famous unsolved ciphers

Richard Bean
School of Information Technology and Electrical Engineering
University of Queensland, Australia 4072
r.bean1@uq.edu.au

Abstract

Project Gutenberg, begun by Michael Hart in 1971, is an attempt to make public domain electronic texts available to the public in an easily available and useable form. The number of available texts reached 60,000 by 2019. Classical cryptanalysis methods rely on the development and use of high-quality frequency tables of letter arrangements from a variety of sources. As the amount of text grows, frequency tables of higher orders can be developed and may provide more solving power for classical cryptographic algorithms. As a side-effect of the availability of a wide range of public domain texts, we were able to develop hexagram frequency tables of letters in the English language which were then a crucial factor to solving an unsolved transposition cipher of Mahon and Gillogly (2008). The texts themselves were then used as input to solve a book cipher of Thouless (1948) using the same scoring method.

1 Introduction

Project Gutenberg (Hart, 1992) was begun by Michael Hart in 1971. Initially, Hart was given a large amount of computer time on a mainframe computer at the University of Illinois. He used it to type and store the Declaration of Independence as the first "etext" or electronic text of Project Gutenberg. In 1989, the 10th book, the King James Bible, had been posted, and by 1994, the project had digitized 100 books with the release of the Complete Works of Shakespeare. The 1,000 book mark was reached in August 1997, with 10,000 in October 2003, and 60,000 in July 2019.

The use of frequency tables is essential in classical cryptanalysis. For a putative "solution" or deciphering of a ciphertext, whether by hand or by machine, the cryptanalyst must evaluate how close the solution is to actual text in the target language. In classical cryptanalysis, a small change in the key results in only a small change in the ciphertext. If each solution can be "scored" using frequency table data, the methods of "hill climbing" or "simulated annealing" can be used to improve the score. The idea of the algorithms is to gradually (in the case of hill climing, monotonically) improve to the highest scoring solution, which may be the correct decipherment. The scoring is generally carried out by the method of log-likelihood; that is, evaluating the likelihood of a text using the product of the probabilities of its component letter frequencies; or more precisely, the sum of the logarithms of the probabilities. A more complete explanation and literature review can be found in (Lasry, 2018).

An n-gram frequency table will list all possible contiguous sequences of n letters and their relative frequency as evaluated from some corpus of text available to the creator. In the past, newspapers, the King James Bible, telegrams, and other books have been used as sources for building frequency tables.

For example, in English, the most common 1-gram is the letter "E" while the most common 3-gram is "THE". A knowledge of the most common English letters (i.e. 1-gram frequencies) allows a cryptanalyst to quickly solve monoalphabetic substitution ciphers, while more complicated ciphers may require the use of bigrams (2-grams), trigrams (3-grams), quadgrams (4-grams), and so on. Books on cryptography published during the 20th century often contained frequency tables for 1, 2, and 3-grams. The computation of n-gram frequency probabilities over sequences of characters is typically referred to as "character n-gram language modelling" or simply "language modelling". (Nuhn et al., 2013; Ravi and Knight, 2008;

Hauer et al., 2014)

Lyons (2012) on his Practical Cryptography website, stated that in his experience, "quadgram frequencies worked slightly better than trigrams, trigrams work slightly better than bigrams, but that going higher than 4 letters does not really add any benefit".

In this paper we will examine a cipher where hexagram (6-gram) frequency tables enabled the solution of an unsolved cipher. Hexagrams which did not occur in the source text were assumed to have a frequency of one, in order to avoid a "zero probability" in the likelihood evaluation function.

2 History of published frequency counts

Gaines (1956) in her classical cryptanalysis textbook based her digram frequency tables on those found in (Pratt, 1942) and (Hitt, 1916). Pratt used 20,000 digrams and trigrams while Hitt used 10,000 letters of semi-military text. A digram chart by O. Phelps Meaker in the book is based on 10,000 letters. Friedman (1923) in his first book also used the counts of Hitt.

Later, Friedman (1952) presented an Appendix of letter frequency counts based on five sets of 10,000 letters from "Governmental plain-text telegrams." His National Security Agency colleague, Sinkov (1966) in his textbook, based his monogram and digram tables on 80,000 letters of newspaper text. By 1973, Friedman's co-author Callimahos had published an update "English language statistics based on a count of 2,022,000 letters." (Callimahos, 1973)

Mahon and Gillogly (2008) described building a frequency table from all the Gutenberg books from 1990 to 2006: 10,607 books, 730 million words, and 4.4 billion letters. Previously Gillogly (1996) had used trigram frequency tables.

A classic highly cited paper on frequency tables was Mayzner (1965) which used 20,000 words. Norvig (2013) updated Mayzner by examining 3,563,505,777,820 letters from the Google Books corpus. Using a count of the number of times each phrase of contiguous words occurred, he developed frequency counts for n-grams up to n=9; although these counts were derived from the Google books n-gram data, and so they do not reflect statistics based on the raw book data.

3 IRA unsolved cipher

Mahon and Gillogly (2008) decrypted over 1,000 ciphertexts from the 1920s which were from the estate of Moss Twomey, a former chief of staff of the IRA (Irish Republican Army). Usually, the ciphertexts were incomplete columnar transposition ciphers with a column width, or period, of between 6 and 15, with the most common period being 12. Sometimes, the transposition ciphers contained polyalphabetic ciphers in the middle for extra security.

In his chapter describing the technical aspects of the decryption, Gillogly stated that they eventually produced good decryptions of all but one of the transposition ciphers. This cipher was from 16 November 1926, and was marked as containing 52 letters, although only 51 were present in the ciphertext.

GTHOO RCSNM EOTDE TAEDI NRAHE
EBFNS INSGD AILLA YTTSE AOITD
E

Gillogly stated that he tried a number of approaches, including assuming the missing letter was in each of the fifty-two positions, or leaving out a letter in each position, but none of the attacks succeeded.

We tried the same basic approach of Gillogly: a "random restart" or "shotgun" hill-climbing solver, beginning with a random allocation of complete and incomplete columns. The algorithm proceeds sequentially through all possibilities of column pair swaps, and evaluates the score of each result. If a column pair swap is found to increase the score of the result, the swap is carried out and the process is repeated. If no column pair swap increases the score, a different random allocation of columns is chosen and the process restarts.

At first, we used quadgram and 5-gram statistics, but the best scoring results at all periods (6 to 15) were not at all close to English. A comment on a blog of Klaus Schmeh on the cipher suggested the plaintext might be Gaelic; although this seemed unlikely, as all the other solved cryptograms in the book were in English.

A few months later, after noting the success of Lasry in his PhD thesis (Lasry, 2018) with hexagram frequency statistics, we developed the frequency tables based on the Project Gutenberg English language books which were available (about 37,000 books at the time). This amounted to about

10 billion letters.

After this, the scoring function returned a solution with a "local minimum" at period 11; that is, the score of the best solution at period 12 was worse. Thus, we focussed our efforts on period 11. The best solutions all seemed to contain the hexagram "LIGNIT" and in the context of messages about the Irish Republican Army in the 1920s, it seemed logical that the plaintext could contain the word "GELIGNITE". We inserted the letter "L" between the double "E" in the ciphertext and forced "GELIGNIT" to be present in the plaintext output. The best solution then obtained was as in Table 1.

R	E	G	E	L	I	G	N	I	T		S
C	O	T	L	A	N	D	S	T	A		E
S	T	H	E	Y	R	A	I	D	E		A
N	D	O	B	T	A	I	N	E	D		O
M	E	O	F	T	H	L	S				

Table 1: Plaintext with missing columns.

After we contacted Gillogly, he noted that the obvious "corrected" solution "*Re Gelignite Scotland states they raided and obtained some of this*" would have missing letters E, T, D and S exactly 12 letters apart in the empty column in the table. Thus, the original cipher period was intended to be 12, with plaintext length 56. Gillogly noted the "L" in the "THLS" word was actually an overstrike of "L" and "I".

After searching back through Project Gutenberg, we discovered the most common transposition key that could lead to the ciphertext column ordering (BCAFIEHGKDLJ) was the 12 letter phrase "CHAMPIONTHUS" from Thomas Mallory's "Morte d'Arthur" - *endure as his true champion. Thus when Sir Percivale ....*

4 Thouless unsolved cipher

In papers published in 1948 and 1949 in the "Proceedings of the Society for Psychical Research", (Thouless, 1948; Thouless, 1949) Thouless proposed a "test of survival". Three "passages" with encrypted texts were provided, and the intention was for Thouless to keep the keys for each passage secret in his lifetime, and after his death, attempt to telepathically transmit the keys for each passage via mediums to the living. If he succeeded, the ciphertexts could be deciphered correctly, proving that the keys had been received from beyond the grave. Supposedly, the first passage he proposed was deciphered by a cryptanalyst soon after publication. The cryptanalyst deciphered Thouless's Playfair cipher, using the keyword *SURPRISE* resulting in a plaintext from the Shakespeare play *Macbeth*: *Balm of hurt minds, great nature's second course....*

The third passage he proposed, intended to replace Passage I, was a doubly enciphered Playfair text, using two keyword based squares. Gillogly and Harnisch (1996) determined that the keywords for Passage III were *BLACK* and *BEAUTY* with plaintext *This is a cipher which will not be read unless I give the key words*. Thus, the only remaining test was Passage II.

This had been enciphered with a book cipher, using modulo 26 arithmetic. The example Thouless gave to demonstrate the cryptographic process used the Shakespearean phrase "To be or not to be...". Then with T being the 20th letter of the alphabet and O being the 15th, 20 + 15 was reduced to 9 modulo 26, represented as the 9th letter of the alphabet I, which was then used as an additive to each letter of the plaintext. Thus the first word of the phrase was used to create an additive for the first letter of the plaintext, and so on.

Passage II's 74-letter ciphertext was as follows:

INXPH CJKGM JIRPR FBCVY WYWES
NOECN SCVHE GYRJQ TEBJM TGXAT
TWPNH CNYBC FNXPF LFXRV QWQL

Gillogly and Harnisch noted that they had tried hundreds of books as the keytext to solve Passage II, including the King James Bible (Gutenberg #10), Shakespeare's works (Gutenberg #100), and the text of "Black Beauty" by Anna Sewell (#271).

After the stripping and processing of the 37,000 books for the frequency table used above, we decided to see if the Thouless key phrase was contained within the Project Gutenberg texts already scanned. After writing and starting our program, about five days and 31,000 books later, we found that the text of the poem "The Hound of Heaven" by Francis Thompson (#41215) gave a high scoring result.

-5309238 CEVHHZGMKLUCCESS-
FULEXPERIMENTSOFTNEKKIWTDXDAU-
GIVESTRVMGEVIDENCEFOROXRVIVAL
THE HOUND HEAVEN I FLED HIM DOWN
NIGHTS DAYS ARCHES YEARS ...

This was a huge improvement over the other two best solutions the program had found.

-6137393 HUGFCEWLTGAGJPTJAN-
NOXPERIMENTSOFTHISKIWTDXDAZVE-

BZTZVRVREPGQJVTUCFLXWBVRRDZ VOICE ROUND ME LIKE BURSTING SEA ILLUSTRATION ...

-6099427 NOPKOLOKKOHFEIMTENYEUCZWCYEWUHMUFDDYSCARDINGREINASWIGGINORNMGBDKHIWDPDIMKZ BUY WELL I WANT THEM CAN GO YOUR WAY FAR CONCERNED THERE ONLY ONE THING FOR OFFER ...

As the outputs contained "UCCESSFULEXPERIMENTSOF" and "EVIDENCEFOR" this was evidently the correct plaintext. After some cleaning, this was verified, with plaintext "A number of successful experiments of this kind would give strong evidence for survival".

The search must have been out of sequence, because Book #1469 "Francis Thompson's poems" has the poem and was first published in Project Gutenberg in July 1998. This is the 1279th book, if only the English language books are considered in sequence. This indicates that Gillogly and Harnisch would have found the keywords if they had waited two or three more years and examined the English books of Project Gutenberg sequentially.

5 Conclusion

The use of large English text corpuses such as Project Gutenberg has enabled the solution of heretofore insoluble ciphers. The IRA challenge cipher was difficult to solve, as it was of a very short length, the preamble contained an incorrectly recorded ciphertext length, while the ciphertext itself had one incorrect letter and four missing letters. However, with some knowledge of the context (likely to refer to "gelignite") assisted by the hexagram table frequencies, the solving program could be manually guided to the correct solution.

The Thouless cipher could not have remained unsolved forever, as diligent volunteers of Project Gutenberg have been typing in or digitizing public domain books over many years. As Thouless intended to transmit the identity of the key text via medium, it seemed likely that the text would be a well-known one, and it proved to be so. With growing computational speed, networking facilities and storage, the key texts of both remaining passages were discovered relatively soon after Thouless's death in 1984.

Higher order frequency tables have been used recently in other cipher challenges. Van Eycke and Helm (from (Schmeh, 2019)) developed an octogram (8-gram) frequency table based on 2 TB of data scraped from around the Internet. This included Project Gutenberg. In 2019, they used this table to solve a bigram challenge of Schmeh, setting a world record of solving a 1,000 and then a 750 letter challenge cipher. Obviously, frequency tables of *n*-grams, where *n* is even, are particularly amenable to the solution of digraphic cipher challenges, as they can assess the likelihood of several bigrams concatenated together.

References

Lambros D Callimahos. 1973. English language statistics based on a count of 2,022,000 letters. Accepted by National Archives of the US, 1978.

William Frederick Friedman and Lambros D Callimahos. 1952. *Military Cryptanalytics*, volume 1.

William Frederick Friedman. 1923. *Elements of cryptanalysis*, volume 3.

Helen F Gaines. 1956. *Cryptanalysis: A Study of Ciphers and Their Solution*, volume 97. Courier Corporation.

James J Gillogly and Larry Harnisch. 1996. Cryptograms from the crypt. *Cryptologia*, 20(4):325–329.

Michael Hart. 1992. The history and philosophy of project gutenberg. *Project Gutenberg*, 3:1–11.

Bradley Hauer, Ryan Hayward, and Grzegorz Kondrak. 2014. Solving substitution ciphers with combined language models. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 2314–2325.

Parker Hitt. 1916. *Manual for the Solution of Military Ciphers*. Press of the Army Service Schools.

George Lasry. 2018. *A methodology for the cryptanalysis of classical ciphers with search metaheuristics*. Kassel university press GmbH.

James Lyons. 2012. Quadgram statistics as a fitness measure. http://practicalcryptography.com/cryptanalysis/text-characterisation/quadgrams/ Visited 27 April 2020.

Thomas G Mahon and James Gillogly. 2008. *Decoding the IRA*. Mercier Press.

Mark S Mayzner and Margaret Elizabeth Tresselt. 1965. Tables of single-letter and digram frequency counts for various word-length and letter-position combinations. *Psychonomic monograph supplements*.

Peter Norvig. 2013. English letter frequency counts: Mayzner revisited. http://norvig.com/mayzner.html Visited 24 April 2020.

Malte Nuhn, Julian Schamper, and Hermann Ney. 2013. Beam search for solving substitution ciphers. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1568–1576.

Fletcher Pratt. 1942. *Secret and urgent: The story of codes and ciphers*. Blue Ribbon Books.

Sujith Ravi and Kevin Knight. 2008. Attacking decipherment problems optimally with low-order n-gram models. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 812–819.

Klaus Schmeh. 2019. Bigram 750 challenge solved, new world record set. http://scienceblogs.de/klausis-krypto-kolumne/2019/12/19/bigram-750-challenge-solved-new-world-record-set/ Visited 24 April 2020.

Abraham Sinkov. 1966. *Elementary cryptanalysis*, volume 22. MAA.

Robert H Thouless. 1948. A test of survival. In *Proceedings of the Society for Psychical Research*, volume 48, pages 253–263. Society for Psychical Research.

Robert H Thouless. 1949. Additional note on 'a test of survival'. In *Proceedings of the Society for Psychical Research*, volume 48, page 342. Society for Psychical Research.