# Background & Process

## Cryptography

Cryptography is the practice of securing communications and information by storing and sending it in a particular form, so that only the intended recipients and users can read and use it (Margaret Rouse, 2014). Modern cryptography, especially with the use of computers now mostly revolves around the encryption of plaintext (unencrypted data) to ciphertext (encrypted data) and decrypting ciphertext to plaintext. Encryption is the process of scrambling information so that only those that are authorised can access it (P Christensson, 2014). Decryption is the opposite of this, it is the process of taking data that has been made unreadable through encryption and converting it back to its readable, unencrypted form. Cryptography is important in the modern day as it allows people to communicate and do business electronically with the confidence that their communications are safe from deceit. The constant increase in reliance on electronic transmissions has resulted in an equal increase in the reliance on cryptography (RSA Laboratories).

There were two main ways for encrypting information by early cryptographers:

1. Substitution
2. Transposition

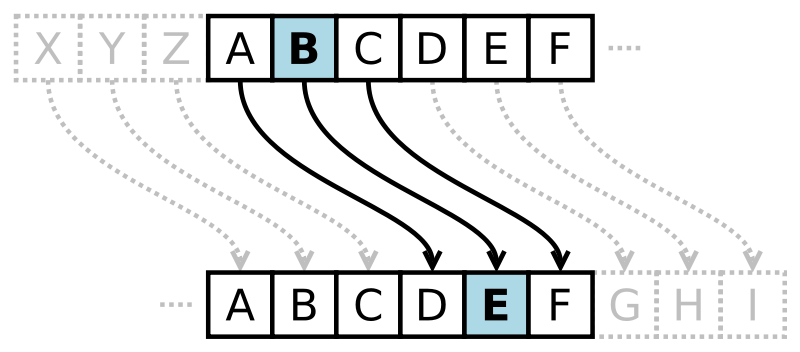
Monoalphabetic substitution ciphers

Substitution ciphers in general are probably the most used technique for encrypting plaintext. Substitution ciphers are an encryption technique where each letter of the plaintext is replaced by a different letter or even a symbol.

A monoalphabetic substitution cipher, also known as a simple substitution cipher, works by simply replacing each letter with a letter or symbol where the substitution is fixed for each letter. For example, if ‘k’ is encrypted to ‘b’, then every occurrence of ‘k’ will be replaced with ‘b’.

The use of simple substitution ciphers can be traced back thousands of years, and were effective for that time, however, in the modern world this kind of encryption technique no longer provides the reliable security that it used to. An example of this kind of cipher is the Caesar shift cipher, developed by Julius Caesar himself. It is a form of substitution cipher, where each letter is replaced by a letter a fixed number of letters in the alphabet away, either up or down. For example, figure 1 shows a section of the Caesar cipher using a right shift of three, where ‘b’ is mapped to ‘e’, ‘c’ is mapped to ‘f’ etc. One of the most effective ways of breaking a simple substitution cipher is to use frequency analysis.

Frequency analysis to decrypt a monoalphabetic substitution cipher encrypted from English involves analysing the occurrences of each symbol and mapping it to a letter in the English language with a similar frequency. In the English language the most frequent letter is ‘e’ **(**Math.cornell.edu, 2017), and so the most common symbol can be mapped to the letter ‘e’. This can be continued for each letter, mapping each symbol to the letter with a similar frequency until a full mapping is done.



*Figure 1, Caesar cipher with a right shift of 3*

Polyalphabetic substitution ciphers

Polyalphabetic substitution ciphers are an improved version of the monoalphabetic ciphers. This cipher involves using a different monoalphabetic substitution as you progress through the text. To encrypt the text a key is needed, which should be the same length as the message. This technique was cryptographer’s response to frequency analysis, as this encryption technique makes it much harder to use decrypt using frequency analysis as the mapping of unencrypted letters to encrypted letters is no longer one-to-one, and is instead one-to-many, where multiple ciphertext letters can represent one plaintext letter.

One of the earliest examples of a polyalphabetic substitution cipher is the Alberti Cipher. It was developed by Leon Battista Alberti around 1467, revolutionising encryption. The basis of his encryption technique was two metal discs, the outer being fixed whilst the inner could be rotated. The outer disc was inscribed with uppercase Latin letters and the numbers 1-4, whilst the inner disc was inscribed with a randomised uppercase Latin alphabet. To encrypt or decrypt the text, there was a mark on the inner disc. Knowledge of the correct letter to match the mark to was needed. To further complicate it, the disc could be turned periodically, resulting in the use of a different alphabet.

Transposition ciphers

Transposition ciphers are the second main way to encrypt text and information. This type of cipher works by changing the order of the letters rather than changing the actual letters themselves.

A simple example of a transposition cipher is the Rail Fence cipher. This works by writing the message on alternate lines and reading each of these lines off in turn. A key for this particular of cipher is usually the number of rows. The process of encrypting a message involves writing the message diagonally until the number of rows in the key is reached and then writing up to the first row repeatedly until the message is fully written, resulting in a zigzag pattern.

One-time pad

A one-time pad is a truly unbreakable cipher where a key, as long as the message intended to be encrypted, is randomly generated from numbers and letter and used just once to encrypt the message. The values are written down on a pad or device and sent to those who are likely to be sending and receiving the message. When sending a message, the sender uses the key to encrypt the message one character at a time. Once it has been used, it can’t be reused, otherwise those intercepting the messages can compare multiple messages together in order to find common words.

Modern cryptography

The invention of the Computer and the rise in its efficiency and processing speed has opened up new doors in cryptanalysis. In the modern age, computers can be used in to decrypt ciphertext many times faster than it would have taken early cryptographers using a pen and paper.

Algorithms have been developed that can be applied to cryptography. These algorithms are designed for finding optimal solutions in a search space, and so can be modified for decrypting ciphers.

Genetic algorithms

A genetic algorithm is a search algorithm that belongs to the larger class of evolutionary algorithms. A genetic algorithm that bases itself on processes which are observable in nature. The algorithm starts with a population of solutions which it individually modifies in an attempt to reach an optimal solution to the problem. Each individual solution is known as a ‘parent’. At each step (also known as a generation), in order to create a better solution, the algorithm employs mainly 3 techniques:

1. Selection rules, this is where the parents are selected based on how good the current solution is.
2. Crossover rules, this is basically combining two parents to form the children to be used in the next generation.
3. Mutation rules involves applying small random changes to get a new solution.

As shown, the main objective of the genetic algorithm is to replicate the elements of nature observable in Darwin’s theory of evolution to come to an optimal solution for problem.

Hill-climbing algorithm

A hill-climbing algorithm is a local search algorithm that starts by generating an arbitrary solution to the problem and then proceeds to incrementally change a single element of the solution based on a number of given factors until the best solution is found. Hill-climbing algorithms traverse the search space by moving up towards the optimal solution with its incremental changes. The algorithm ends when no more optimal moves can be found, however, an issue with hill climbing is that it due to its nature, it may reach a peak in the search space which it interprets to be the best solution but there may be a more optimal solution elsewhere. Figure 2 illustrates this concept, as there is a smaller peak after the biggest one in the search space, the algorithm could potentially get stuck on the smaller peak which is the local maxima when there is a much better solution in the search space that it misses, illustrated by the larger peak before it. One method that can be used to try and solve this, is randomly restarting the algorithm. This is where if the algorithm gets stuck, it can generate a new random solution and work from there in an attempt to reach the global best solution.

State space

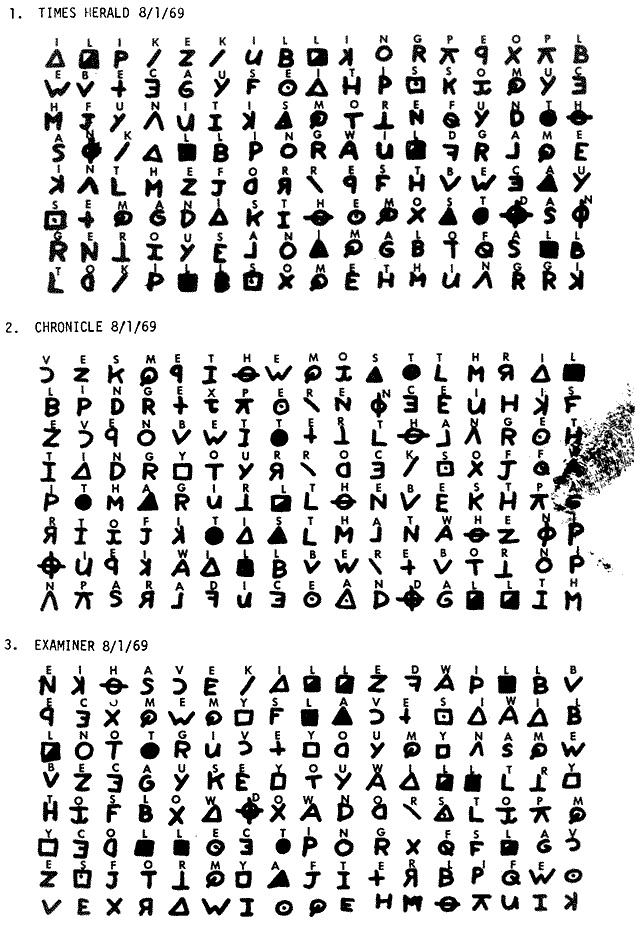
Score

*Figure 2, a diagram illustrating a basic hill climbing algorithm*

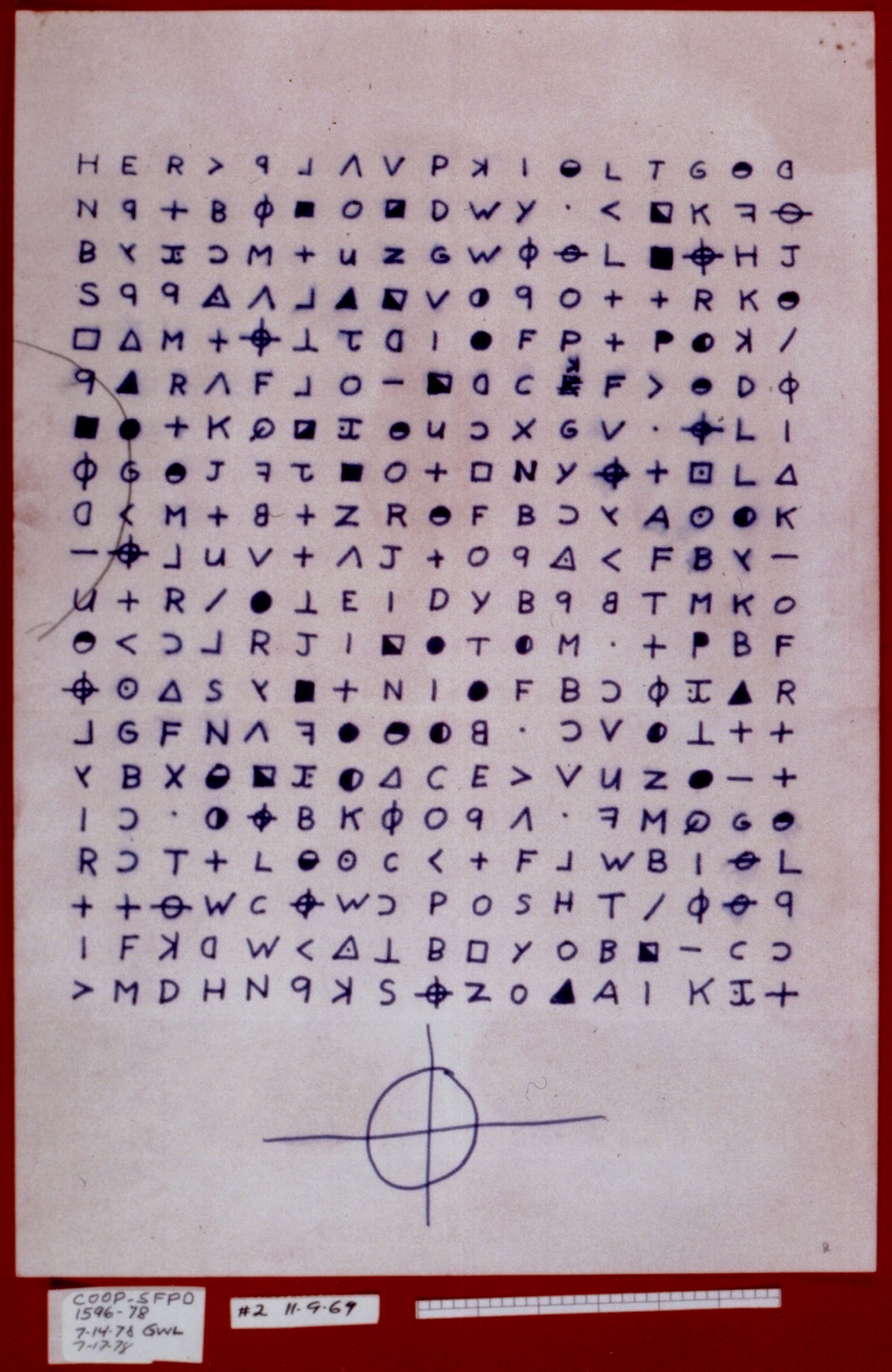
## The Zodiac

The Zodiac was a serial killer operating in the Northern Californian area of the USA during the late 1960’s and possibly into the early 1970’s. He carried out four confirmed attacks totalling 5 murders, with 5 other attacks being linked to him but not confirmed. On July 31 1969, the Zodiac sent 3 letters, one to each of the ‘Vallejo Times-Herald’, ‘San Francisco Examiner’ and ‘San Francisco Chronicle’. These letters were encrypted with various symbols and signed with a symbol that was a circle with a cross through it, known as the Z408 cipher due it being 408 letters in length. Each of the newspapers also received a handwritten letter claiming to be the killer of two teenagers the previous Christmas. The letter also contained threats of further attacks if the encrypted letters he sent to each of them were not printed on the front page. The ‘San Francisco Examiner’ received a follow up letter August 4th 1969, describing the murders in detail and taunting the police for not having cracked his coded letter. However, several days after, this the code was cracked manually by a high school teacher, Donald Harden and his wife, Bettye Harden. This was the only cipher sent by the Zodiac killer to be successfully decoded. The Z408 cipher is an example of a homophonic substitution cipher where each letter can be mapped to multiple symbols but each symbol can only be mapped to one letter. Figure 3 shows the separate parts of the cipher sent to the newspapers, and the Hardens corresponding decryption.

The next cipher to be received was the still unsolved 340 character cipher (Z340), received on the 8th November 1969. As shown in figure 4, like the Z408 cipher the Z340 cipher was signed with the symbol of a circle with a cross through it. The Zodiac also sent two more ciphers, but attempts to solve them have been limited due to the extremely short length of the ciphers and the difficulties this causes. Attempts are still being made to solve the Z340 cipher however. Whilst the main objective of my research is to crack the Z340 cipher, the difficulty and unlikeliness of this is extremely high. For this reason, another aim of my research is to evaluate approaches to solving the cipher.



*Figure 3, the 408 cipher and its decryption*



*Figure 4, 304 cipher*

## Prior work

Attempts have been made throughout the decades to solve the Z340 cipher that the zodiac sent with no success so far. Theories include that the cipher is a type of transposition cipher or is meant to be read in a different order as opposed left-to-right, top-to-bottom. Some theorise that parts of the cipher are just random filler, or that the whole cipher is nonsensical. Many believe that due to the Z408-cipher being a homophonic substitution cipher, then the Z340 cipher is likely to be one as well. This is the assumption that this project is working under. Others have attempted to crack the cipher under the assumption that it is a homophonic substitution cipher, with different techniques including genetic algorithms and hill-climbing algorithms.

## Preparation and Interest

Preparation for this project involved reading into encryption techniques further, with particular focus into the homophonic substitution cipher, but also reading over the alternative angles that other researchers had approached the problem at to get ideas for techniques that could also be used for this projects approach. After deciding on this aspect of the approach, I researched the possible algorithms that could be used in this projects attempt to solve the Z340 cipher by looking at previous attempts to get an idea of which areas have been explored the least as well as what has worked best.

The decision to write the program in Java mainly came from personal preference and competence with the language. The author’s interest in this project stems from two main areas. The first being cipher-text and decryption in general, being able to create something that can decode a complicated piece of obfuscated text and the second is the fact that this particular piece of cipher-text has remained unsolved for almost 50 years and the mystery and background that is behind it.

A large part of the preparation after this was setting up the environment to be as efficient as possible. This included the development environment and source control. Eclipse and GitHub for these respectively, due to personal preference and experience using them. To track progress, an online website/tool called Trello was used, in the form of a scrum board as Agile was the methodology that was decided upon for this project.