Breaking Vigenère Cipher

Adilson Medronha
adilson.medronha@edu.pucrs.br
Pontifícia Universidade Católica do Rio Grande do Sul
Porto Alegre, Brazil

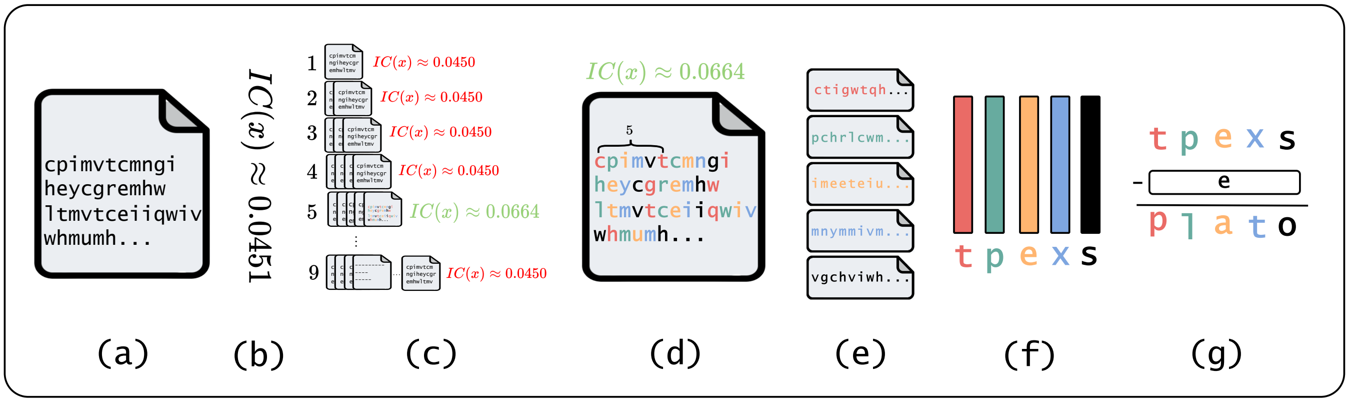

Figure 1: (a) Load ciphered text (b) compute IC (c) compute IC for each split (d) obtain the split that has the closest IC when compared to the IC table of languages (e) create five sub strings (f) get the letter distribution of five strings (g) compute the distance between the letters that appear most frequently and the most commonly used letter in the corresponding IC language.

ABSTRACT

In this study, I'll provide an overview of the Vigenère cipher, including its history, how it works, and how it can be broken.

1 INTRODUCTION

Blaise de Vigenère introduced the Vigenère cipher in the 16th century as a polyalphabetic substitution cipher that uses a series of interwoven Caesar ciphers to encrypt messages. The cipher employs a different substitution alphabet for each letter, with the alphabets based on a repeated keyword. This keyword makes the cipher much more secure than a simple substitution cipher that uses one fixed substitution alphabet for the entire message. As a result, the Vigenère cipher is considered to be one of the earliest examples of modern cryptography [1].

The Vigenère cipher's complexity makes frequency analysis and other cryptanalysis techniques less effective against it. Despite being thought of as unbreakable for over three centuries, Charles Babbage and Friedrich Kasiski discovered a method to break it in the mid-19th century. Nevertheless, the Vigenère cipher remains an important historical cipher that has contributed significantly to the development of modern cryptography.

2 BACKGROUND

The index of coincidence (IC) is a measure of how likely it is to draw two matching random letters given a text. For a text of length N, the index of coincidence is defined as:

$$IC = \frac{\sum_{i=1}^{c} f_i(f_i - 1)}{N(N-1)} \tag{1}$$

where f_i is the frequency of the i-th letter in the text and c is alphabet set size, $c = 26$ or $c = 36$ if digits $[0-9]$ is included.

2.1 index of coincidence

The IC of an English text is about 0.066, while the IC of a random sequence of letters is about 0.0385. The IC of a text encrypted with the Vigenère cipher depends on the length of the keyword used for encryption. If the length of the keyword is k, then the IC of the encrypted text is about 0.065 for every k letter.

Language	Index of Coincidence
English	0.0667
Portuguese	0.0727
Italian	0.0738
Random	0.0385

Table 1: IC for English, Portuguese, Italian and Random

2.2 Encoder

The encryption function is given by:

$$E(p) = (p_i + k_{i \bmod |k|}) \bmod 26, \tag{2}$$

where p_i is the i-th letter of the plaintext and ith mod $|k|$ letter of k, key used for encryption. The operator mod returns the remainder of the division by 26, ensuring that the ciphertext letters are in the range of 0 to 25 (A to Z in alphabetical order).

2.3 Decoder

The decryption function is given by:

$$D(e) = (e_i - k_{i \bmod |k|}) \bmod 26, \tag{3}$$

where e_i is the ith letter of the ciphered text and ith mod $|k|$ letter of k, key used for encryption.

3 METHOD

To decipher the Vigenère cipher, I followed the steps outlined in Figure 1. The detailed process to discover the keyword used:

(a) Load the ciphered text.

(b) Compute IC. If the IC is close to 0.065, it suggests that a straightforward mapping between the most frequently used letters and the known frequency distribution of the English language might effectively crack the cipher. In such cases, it is the Caesar cipher with a single-character keyword.

(c) Since $IC(x) \approx 0.0451$, the keyword size is not one. To determine the length of the keyword used for encryption, split the text into sub-strings of lengths 1 to 9, and compute their ICs. Each sub-string is obtained by taking every i-th letter of the text. The sub-string containing the closest IC to 0.065 suggests the length of the keyword used for encryption is i.

(d) The text split into 5 sub-strings has the closest $IC(x) \approx 0.066$ to the expected value of 0.065.

(e) Save the five sub-strings: the first sub-string contains every 5th letter of the ciphered text, the second sub-string contains every 5th letter starting from the second letter, and so on.

(f) Compute the letter distribution of all five sub-strings. Obtain the most used letter in each sub-string: "**TPEXS**".

(g) The distance between "**TPEXS**" and the most frequently used letter in English ("**E**") suggests the corresponding letter position, the letter in the keyword: "**PLATO**".

The Table 2 presents indexes of alphabet: "ABCDEF...", ciphered text:"CPIMVT..." and deciphered text: "NEITHER...". Using the decoder described in Equation 3, it is possible to reveal the plain-text. Here is an example of how to compute it once we have the keyword:

$$\begin{aligned}
D(C) &= (C - k_{0 \bmod 5}) \bmod 26 = (2 - 15) \bmod 26 = 13 = N \\
D(P) &= (P - k_{1 \bmod 5}) \bmod 26 = (15 - 11) \bmod 26 = 4 = E \\
D(I) &= (I - k_{2 \bmod 5}) \bmod 26 = (8 - 0) \bmod 26 = 8 = I \\
D(M) &= (M - k_{3 \bmod 5}) \bmod 26 = (12 - 19) \bmod 26 = 19 = T \\
D(V) &= (V - k_{4 \bmod 5}) \bmod 26 = (21 - 14) \bmod 26 = 7 = H \\
D(T) &= (T - k_{5 \bmod 5}) \bmod 26 = (19 - 15) \bmod 26 = 4 = E \\
D(C) &= (C - k_{6 \bmod 5}) \bmod 26 = (2 - 11) \bmod 26 = 17 = R
\end{aligned}$$

0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25
A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z
2	15	8	12	21	19	2	12	13	6	8	7	4	24	2	6	17	4	12	7	22	11	19	12	21	19
C	P	I	M	V	T	C	M	N	G	I	H	E	Y	C	G	R	E	M	H	W	L	T	M	V	T
15	11	0	19	14	15	11	0	19	14	15	11	0	19	14	15	11	0	19	14	15	11	0	19	14	15
P	L	A	T	O	P	L	A	T	O	P	L	A	T	O	P	L	A	T	O	P	L	A	T	O	P
13	4	8	19	7	4	17	12	20	18	19	22	4	5	14	17	6	4	19	19	7	0	19	19	8	4
N	E	I	T	H	E	R	M	U	S	T	W	E	F	O	R	G	E	T	T	H	A	T	T	H	E

Table 2: deciphered text given discovered keyword

Since the first and second most frequent letters in Portuguese are very close to each other, if the plain-text language is Portuguese, it may be necessary to compute the distance between the most frequent letters of sub-strings using the second, third, or next most common letter (known a priori) in Portuguese. This behavior is related to the step presented in Figure 1.**e-g**.

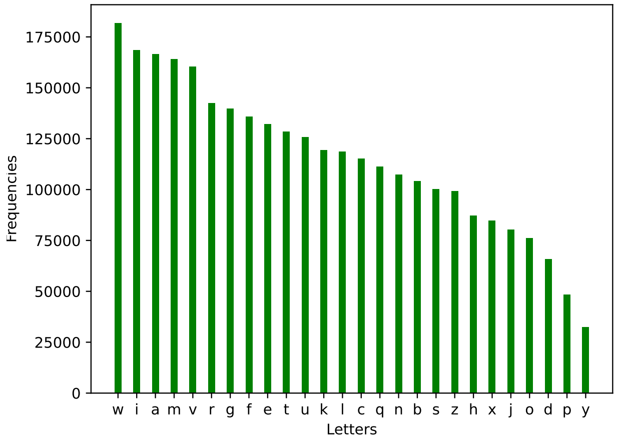

Figure 2: Portuguese letter distribution

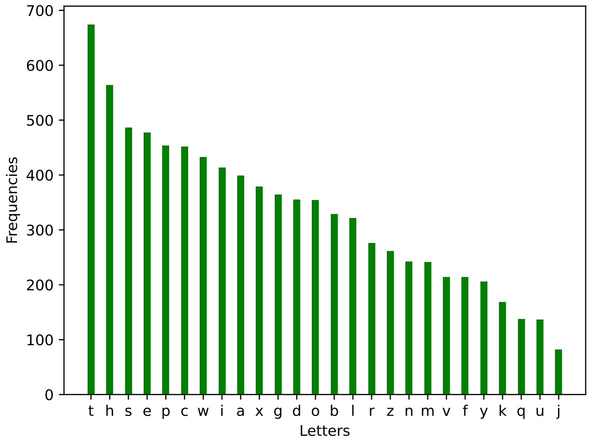

Figure 3: English letter distribution

4 CONCLUSION

Using the presented method, the Vigenère cipher was successfully decrypted, revealing the plain text to be about the Republica of Plato. The method is based on the analysis of the index of coincidence and the frequency of letters in the substrings of the ciphered text.

REFERENCES

[1] Claude Shannon. "Communication Theory of Secrecy Systems". In: *Bell System Technical Journal* 28.4 (1949), pp. 656–715. DOI: 10.1002/j.1538-7305.1949.tb00928.