Index of Conincidence

A. Garsia - 1988

The estimation of period in a substitution cipher.

Let us suppose that we are given a ciphertext

$$C_1, C_2, \ldots, C_N$$

which is known to have been encrypted by a Vigenere substitution with keyword of length p. For convenience let us assume that N is a multiple of p and set $M = \frac{N}{p}$.

This means that if the original plaintext was

$$X_1, X_2, \ldots, X_N$$

then for each $k = 1, 2, \ldots, M$ the k^{th} block of ciphertext

$$C_{(k-1)p+1}, C_{(k-1)p+2}, \ldots, C_{(k-1)p+p}$$

is obtained from the corresponding k^{th} block of plaintext

$$X_{(k-1)p+1}, X_{(k-1)p+2}, \ldots, X_{(k-1)p+p} \tag{1}$$

by the p substitutions

$$C_{(k-1)p+j} = \phi_j(X_{(k-1)p+j}) \quad (j = 1, 2, ..., p).$$

For a Vigenere cipher each of the substitutions ϕ_j circularly shifts the alphabet. In the case of a general periodic substitution cipher each of the ϕ_j's permutes the alphabet in an arbitrary manner.

We shall present here a statistic which may be used to estimate the period p of a Vigenere substitution. It may be used as well for the most general periodic substitution cipher provided a sufficiently large ciphertext sample is available.

This statistic usually referred to as the *index of coincidence* can be verbally defined as follows

$$I_C = \frac{\text{number of pairs of equal letters in ciphertext}}{\text{the total number of pairs of letters}}. \tag{2}$$

Note that if N_A denotes the total number of A's in the ciphertext then the binomial coefficient $\frac{N_A(N_A-1)}{2}$ gives the number of $A - A$ pairs. More generally, if N_A denotes the

number of A's in the ciphertext then $\frac{N_A(N_A-1)}{2}$ gives the number of $A-A$ pairs. Thus if we set $D_C = \frac{1}{2}\sum_{\alpha=A}^{Z} N_\alpha(N_\alpha-1)$, the index of coincidence is given by the formula

$$I_C = \frac{D_C}{\frac{N(N-1)}{2}} = \frac{1}{N(N-1)}\sum_{\alpha=A}^{Z} N_\alpha(N_\alpha-1). \tag{3}$$

Once I_C has been calculated, the period of the substitution cipher may be estimated by the formula

$$p = \frac{.027N}{(N-1)I_C + 1 - .038N}. \tag{4}$$

Of course this may not be an integer and the actual period should be found within the closest integers to this number.

A word of caution should be added here. Statistics carried out on too small samples may lead to grossly erroneous conclusions. The estimate in equation (3) cannot be relied upon on ciphertexts of less than 500 letters.

It is good to give here the reasoning that leads to the expression in (3) since it exhibits an interesting use of probabilities. Roughly speaking, the idea is to obtain another expression for I_C which involves the period p, equate this expression to that given in (2) and then solve for p .

Let us rewrite the ciphertext in M rows, the k^{th} row containing the k^{th} block of ciphertext

$$C_{(k-1)p+1}, C_{(k-1)p+2}, \ldots, C_{(k-1)p+p}.$$

Then we see that the i^{th} column of the resulting arrangement will consist of all the ciphertext letters encrypted by the substitution ϕ_i. For convenience let $M_\alpha^{(i)}$ denote the number of letters equal to α in the i^{th} collumn of this arrangement. This given it is not difficult to see that we have

$$2D_C = \sum_{i=1}^{p}\sum_{\alpha=A}^{Z} M_\alpha^{(i)}(M_\alpha^{(i)}-1) + 2\sum_{i=1}^{p}\sum_{j=i+1}^{p}\sum_{\alpha=A}^{Z} M_\alpha^{(i)}M_\alpha^{(j)}. \tag{5}$$

The term $(I) = \sum_{i=1}^{p}\sum_{\alpha=A}^{Z} M_\alpha^{(i)}(M_\alpha^{(i)}-1)$ here represents the contribution to $2D_C$ coming from pairs of letters that are in the same column, likewise the term

$$(II) = 2\sum_{i=1}^{p}\sum_{j=i+1}^{p}\sum_{\alpha=A}^{Z} M_\alpha^{(i)}M_\alpha^{(j)}.$$

gives the contribution to $2D_C$ coming from pairs of letters that are in different columns.

Note now that for each i we have

$$\sum_{\alpha=A}^{Z} M_\alpha^{(i)} = M,$$

(this sum gives the total number of letters in column i). Thus

$$(I) = \sum_{i=1}^{p} \sum_{\alpha=A}^{Z} (M_\alpha^{(i)})^2 - pM$$

Let us then set for each $i = 1, 2, \ldots, p$ and each letter α

$$p_\alpha^{(i)} = \frac{M_\alpha^{(i)}}{M}$$

so we may write

$$(I) = M^2 \sum_{i=1}^{p} \sum_{\alpha=A}^{Z} (p_\alpha^{(i)})^2 - pM$$

If our sample is sufficiently large, $p_\alpha^{(i)}$ should approximate the probability of the letter α in the i^{th} column of ciphertext. Since each substitution ϕ_i permutes the letters of the alphabet, a moment of thought should reveal that the sum

$$\sum_{\alpha=A}^{Z} (p_\alpha^{(i)})^2$$

should (for each i) be very close to the sum

$$\sum_{\alpha=A}^{Z} (p_\alpha)^2$$

where p_α gives the probability of the letter α in the original plaintext. From our tables of english letter frequencies we obtain that

$$\sum_{\alpha=A}^{Z} (p_\alpha)^2 \approx .065$$

so we may conclude that

$$(I) \approx p \times M^2 \times .065 - pM$$

To estimate (II) we reason as follows. Note first that we may write

$$(II) = M^2 \sum_{i=1}^{p} \sum_{j=i+1}^{p} \sum_{\alpha=A}^{Z} p_\alpha^{(i)} p_\alpha^{(j)}.$$

Now for fixed i and j let X denote the random variable obtained by selecting a letter at random from the i^{th} column and Y be the random variable obtained by selecting a letter at random from the j^{th} column. This given we see that the sum

$$\sum_{\alpha=A}^{Z} p_\alpha^{(i)} p_\alpha^{(j)}$$

may be simply interpreted as *the probability that* $X = Y$ However, if we assume that the two substitutions ϕ_i and ϕ_j, used in the i^{th} and j^{th} columns respectively, are entirely unrelated, it is not unreasonnable to expect that

$$P[X = Y] \approx \frac{1}{26} \approx .038.$$

Substituting this in equation (4) we obtain

$$(II) \approx p(p-1)M^2.038.$$

Combining the two approximate values of (I) and(II) gives

$$I_C = \frac{2D_C}{N(N-1)} \approx \frac{N\frac{N}{p} \times .065 - N + N(N - \frac{N}{p}) \times .038}{N(N-1)}.$$

Cancelling the common factor N multiplying by $N-1$ and solving for p yields our desired estimate (3) for the period.

You may practice on the computer on some Vigenere encrypted files to see how close the estimate gets in practice to the real period. The computer component is essential here since, as we said, this statistic is meaningless for small samples.