Index of Coincidence

Hui Chen
Computer Science
Virginia State University, Virginia 23806
E-mail: huichen (AT) ieee.org

Written on September 15, 2014
Lastly revised on September 4, 2015

1 Index of Coincidence

The Index of Coincidence (IC) is the probability that two letters chosen at random from a given text (plain-text or cipher-text) match. Let p_i be the probability of the i-th letter in a language's alphabet. Then the probability that two letters chosen at random match from any text in the language is,

$$\overline{IC} = \sum_{i=0}^{N_A-1} p_i^2 \tag{1}$$

where N_A is the size of the alphabet.

For English, $N_{English} = 26$. Then,

$$\overline{IC}_{English} = \sum_{i=0}^{26-1} p_i^2 = \sum_{i=0}^{25} p_i^2 \tag{2}$$

The frequency (or the probability) of English letters is given in [1] and reproduced in Table 1, i.e., $p_0 = p(A) = 0.080$, $p_1 = p(B) = 0.015$, $\ldots$, $p_{25} = p(Z) = 0.002$. Then, using the frequency of English letters, we obtain,

$$\begin{aligned} \overline{IC}_{English} &= \sum_{i=0}^{25} p_i^2 \\ &= p_0 p_0 + p_1 p_1 \ldots p_{25} p_{25} \\ &\approx 0.080 \cdot 0.080 + 0.015 \cdot 0.015 + \ldots + 0.002 \cdot 0.002 \\ &= 0.065933 \end{aligned} \tag{3}$$

Table 1: Frequency of English Letters

Letter	Frequency
A	0.080
B	0.015
C	0.030
D	0.040
E	0.130
F	0.020
G	0.015
H	0.060
I	0.065
J	0.005
K	0.005
L	0.035
M	0.030
N	0.070
O	0.080
P	0.020
Q	0.002
R	0.065
S	0.060
T	0.090
U	0.030
V	0.010
W	0.015
X	0.005
Y	0.020
Z	0.002

which is the index of coincidence of English plain-text. A mono-alphabetic substitution cipher, e.g., Caesar Cipher, typically preserves the frequency distribution albeit a frequency may now be mapped to a different letter. For the mono-alphabetic substitution cipher that does not alter the frequency distribution, the index of coincidence of cipher-text is the same as that of plain-text.

2 Estimating Index of Coincidence from Given Text

In practice, the frequency distribution of the letters in cipher-text is not conveniently known. We now describe a method estimating the index of coincidence from a given text.

Following [2], the total number of pairs of letters that can be chosen from a given text of length N is,

$$\binom{N}{2} = \frac{N(N-1)}{2} \tag{4}$$

Let F_i be the number of occurrences of the i-th letter of the alphabet of a language. Thus, for a given text, $\sum_{i=0}^{N_A-1} F_i = N$ where N is the length of the text (i.e., the number of characters in the text) and N_A is the size of alphabet of the language.

Note that frequency F_i is actually defined as the number of occurrences of the i-th letter in the text. Then the number of pairs containing just the i-th letter is,

$$\binom{F_i}{2} = \frac{F_i(F_i-1)}{2} \tag{5}$$

Since IC is defined as the probability that two letters chosen at random from the given text match, from equations (4) and (5), we have,

$$IC = \frac{\sum_{i=0}^{N_A-1} \binom{F_i}{2}}{\binom{N}{2}} = \frac{\sum_{i=0}^{N_A-1} F_i(F_i-1)}{N(N-1)} \tag{6}$$

For English,

$$IC_{English} = \frac{\sum_{i=0}^{N_{English}-1} F_i(F_i-1)}{N(N-1)} = \frac{\sum_{i=0}^{26-1} F_i(F_i-1)}{N(N-1)} = \frac{\sum_{i=0}^{25} F_i(F_i-1)}{N(N-1)} \tag{7}$$

which is an estimate of $\overline{IC}_{English}$. If N is *sufficiently large*[1], we expect that,

$$IC_{English} \approx \overline{IC}_{English} \approx 0.065933 \tag{8}$$

[1] It is a good exercise to determine a sufficiently large N.

3 Periodic Poly-Alphabetic Substitution Cipher

Let d be the period of a periodic poly-alphabetic substitution cipher, e.g., a Vigenère Cipher. Denote $\mathcal{C}_1, \ldots, \mathcal{C}_d$ as d cipher alphabets. Let $f_i : \mathcal{A} \to \mathcal{C}_i$ be a mapping from the plain-text alphabet $\mathcal{A}$ to the i-th cipher alphabet $\mathcal{C}_i$ ($1 \leq i \leq d$), i.e.,

$$\mathcal{C}_i = f_i(A) \tag{9}$$

where $1 \leq i \leq d$.

Let M be a plain-text message, i.e.,

$$M = m_1 \ldots m_d m_{d+1} \ldots m_{2d} \ldots \tag{10}$$

Following [2], then the poly-alphabet substitution cipher enciphers the message M by repeating the sequence of mappings $f_1, \ldots, f_d$ every d characters as follows,

$$E_K(M) = f_1(m_1) \ldots f_d(m_d) f_1(m_{d+1}) \ldots f_d(m_{2d}) \ldots \tag{11}$$

Let us rewrite equation (11) as follows,

$$\begin{aligned} E_K(M) &= f_1(m_1) f_2(m_2) \ldots f_d(m_d) f_1(m_{d+1}) f_2(m_{d+2}) \ldots f_d(m_{2d}) \ldots \\ &= x_1 x_2 \ldots x_d x_{d+1} x_{d+2} \ldots x_{2d} \ldots x_{2d+1} x_{2d+2} \ldots x_{3d} \cdots \end{aligned} \tag{12}$$

Following [3], denote X_i as the characters in the cipher-text enciphered using mapping f_i, where $1 \leq i \leq d$. Then,

$$\begin{aligned} X_1 &= x_1 x_{d+1} x_{2d+1} x_{3d+1} \ldots x_{nd+1} \cdots \\ X_2 &= x_2 x_{d+2} x_{2d+2} x_{3d+2} \ldots x_{nd+2} \cdots \\ &\vdots \\ X_d &= x_d x_{d+d} x_{2d+d} x_{3d+d} \ldots x_{(nd+d} \cdots \\ &= x_d x_{2d} x_{3d} x_{4d} \ldots x_{(n+1)d} \cdots \end{aligned} \tag{13}$$

whose corresponding plain-text characters are as follows,

$$\begin{aligned} S_1 &= m_1 m_{d+1} m_{2d+1} m_{3d+1} \ldots m_{nd+1} \cdots \\ S_2 &= m_2 m_{d+2} m_{2d+2} m_{3d+2} \ldots m_{nd+2} \cdots \\ &\vdots \\ S_d &= m_d m_{d+d} m_{2d+d} m_{3d+d} \ldots m_{(nd+d} \cdots \\ &= m_d m_{2d} m_{3d} m_{4d} \ldots m_{(n+1)d} \cdots \end{aligned} \tag{14}$$

Provided that the index of coincidence of S_i $(1 \le i \le d)$ is the same as that of M[2], i.e.,

$$IC(M) = IC(S_1) = IC(S_2) = \ldots = IC(S_d) \tag{15}$$

we now reconstruct $IC(E_K(M))$ as follows. Picking two characters at random from $E_K(M)$, we want to know the probability that the two characters match.

Let N be the number of characters in $E_K(M)$. Now examine two cases in which 2 randomly picked characters from the N characters match.

Case 1. The two characters are in X_i $(1 \le i \le d)$. The number of characters in X_i is N/d. Then the probability that they are in the same X_i is,

$$d\frac{\binom{\frac{N}{d}}{2}}{\binom{N}{2}} = d\frac{\frac{N}{d}(\frac{N}{d}-1)}{N(N-1)} = \frac{N(\frac{N}{d}-1)}{N(N-1)} = \frac{\frac{N}{d}-1}{N-1} = \frac{N-d}{d(N-1)} \tag{16}$$

The intuition behind the above is that (a) the total number of choices of picking 2 characters from N characters is $\binom{N}{2}$; (b) the total number of choices of picking 2 characters from *one* sub-character sequence X_i is $\binom{N/d}{2}$ where the number of sub-character sequences is d.

Case 2. One of the two characters is in X_i and the other in X_j where $i \ne j$, $1 \le i \le d$, and $1 \le j \le d$. Then the probability that they are in X_i and X_j where $i \ne j$, $1 \le i \le d$, and $1 \le j \le d$ is,

$$\frac{\binom{d}{2}\frac{N}{d}\frac{N}{d}}{\binom{N}{2}} = \frac{d(d-1)\frac{N}{d}\frac{N}{d}}{N(N-1)} = \frac{(d-1)N\frac{N}{d}}{N(N-1)} = \frac{(d-1)\frac{N}{d}}{N-1} = \frac{(d-1)N}{d(N-1)} \tag{17}$$

The intuition behind the above is that (a) the number of choices of picking 2 distinct sub-character sequences from d sub-character sequences is $\binom{d}{2}$; (b) for each character of the 2 characters, there are $\frac{N}{d}$ choices, respectively; and (c) the total number of choices of picking 2 characters from N characters is $\binom{N}{2}$.

Let κ_p be the probability that the two characters picked randomly from the cipher-text match and the two characters are picked from the same sub-character sequence X_i, $1 \le i \le d$, i.e., the probability of *Case 1.* Following [3], we argue that if the two characters are in the same X_i then they

[2] This is a reasonable assumption. It is a good exercise to examine English language or any other natural languages to determine that this is a reasonable assumption.

are both enciphered using the same alphabet, so the probability κ_p is the index of coincident of plain-text. Then, according to equation (15) and the definition of the index of coincidence of plain-text, we know,

$$\kappa_p = IC(M) = IC(S_1) = IC(S_2) = \ldots = IC(S_d) \tag{18}$$

Let κ_r be the probability that the two characters picked randomly from the cipher-text and the two characters are from two distinct sub-character sequences, X_i and X_j, $i \neq j$, $1 \leq i \leq d$, and $1 \leq j \leq d$, i.e., the probability of *case 2*. Following [3] again, we argue that if the two characters are in X_i and X_j $i \neq j$, then they are enciphered using different alphabets and we can assume the two characters of the ciphertext are randomly distributed. The probability κ_r is $\frac{1}{N_A}$ where N_A is the alphabet size of the language.

We now have,

$$\begin{aligned} IC(E_K(M)) &= \frac{N-d}{d(N-1)}\kappa_p + \frac{(d-1)N}{d(N-1)}\kappa_r \\ &= \frac{1}{d}\frac{N-d}{N-1}\kappa_p + \frac{d-1}{d}\frac{N}{N-1}\kappa_r \end{aligned} \tag{19}$$

3.1 English Language

For English, $\kappa_p = IC_{English} = 0.065933$ according to equation (3) and $\kappa_r = \frac{1}{N_{English}} = \frac{1}{26} = 0.038462$. Then,

$$\begin{aligned} IC_{English}(E_K(M)) &= \frac{1}{d}\frac{N-d}{N-1}\kappa_p + \frac{d-1}{d}\frac{N}{N-1}\kappa_r \\ &= \frac{1}{d}\frac{N-d}{N-1}0.065933 + \frac{d-1}{d}\frac{N}{N-1}0.038462 \end{aligned} \tag{20}$$

Using equation (21), we compute the Index of Coincidence for a few different periods and the result is in Table 2.

For the case that $N \gg d$, $\frac{N-d}{N-1} \approx 1$ and $\frac{N}{N-1} \approx 1$. We have,

$$\begin{aligned} IC_{English}(E_K(M)) &= \frac{1}{d}\frac{N-d}{N-1}0.065933 + \frac{d-1}{d}\frac{N}{N-1}0.038462 \\ &\approx \frac{1}{d}0.065933 + \frac{d-1}{d}0.038462 \end{aligned} \tag{21}$$

From equation (19), we obtain,

Table 2: Index of Coincidence of English Periodic Poly-Alphabetic Substitution Cipher-Text

Period d	Index of Coincidence $IC_{English}(E_K(M))$
1	0.065933
2	0.052198
3	0.047619
4	0.045330
5	0.043956
10	0.041209
1000	0.038489
∞	0.038462

$$d = \frac{N(\kappa_p - \kappa_r)}{(N-1)IC(E_K(M)) + \kappa_p - N\kappa_r} \tag{22}$$

Since for English, $\kappa_p \approx 0.065933$ and $\kappa_r = 0.038462$,

$$d \approx \frac{0.027471N}{(N-1)IC(E_K(M)) - 0.038462N + 0.065933} \tag{23}$$

which can be used to find key length if $IC(E_K(M))$ is *accurately estimated* and d is *relatively small*[3].

References

[1] Matt Bishop. *Introduction to Computer Security*. Addison-Wesley Professional, 2004.

[2] Robling Denning and Dorothy Elizabeth. *Cryptography and Data Security*. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA, 1982.

[3] Stephen Harris. Expectation value of the index of coincidence, June 2012. `http://crypto.stackexchange.com/questions/3039/expectation-value-of-the-index-of-coincidence`.

[3] It is a good exercise to dermine under what condition $IC(E_K(M))$ can be accurately estimated,and how d affects the estimation.