CRYPTOGRAPHY AND INFORMATION

We shall work here with a *Random Cryptographic System*. More precisely, the ingredients are as follows:

a) **A message space**

$$M = \{m_1, m_2, \ldots, m_n\}.$$

These are the plaintexts we want to be able to send.

b) **A cipher space**

$$C = \{c_1, c_2, \ldots, c_n\}.$$

These are the ciphertexts that we may receive.

c) **A key space**

$$K = \{k_1, k_2, \ldots, k_m\}.$$

These are the keys that may be used.

d) A set of **injections** (one-to-one maps of M onto C)

$$c = E_k(m) : M \to C \quad (k \in K)$$

These are the encrypting transformations of the system.

e) Two sets of **Probabilities**

$$\{p_1, p_2, \ldots, p_n\}.$$

and

$$\{q_1, q_2, \ldots, q_m\}.$$

which are respectively the probabilities with which the **messages** and the **keys** are selected.

This given, a cryptographic transaction in such a system takes place as follows:

1. The sender produces a message M which is a random variable with

$$P[M = m_i] = p_i.$$

2. The sender and receiver decide on a key K by a mechanism (or device) which is independent of that which produces the message and so that

$$P[K = k_i] = q_i.$$

3. The sender encrypts M into $C = E_k(M)$ and sends it to the receiver.

To be precise, we may assume that M and K are produced spinning two fortune wheels with arc-lengths given by the $p_i's$ and $q_j's$ respectively. In summary, our cryptographic transaction here may be viewed as an *experiment* producing three random variables

$$M, K \text{ and } C = E_k(M),$$

with the above set of probabilities, with C a function of M and K and K independent of M.

The first aim of these notes is to present a quantitative approach for determining the amount of ciphertext the opponent needs to reconstruct the key.

More precisely if N denotes the number of characters in the ciphertext, then we define as the *unicity distance* and denote it by UD, the least value of N for which the opponent can uniquely reconstruct the key from the given ciphertext. More precisely, in the *cyphertext only* attack, we need to determine when K is a function of C , while in the *known plaintext* attack we need to determine when K is a function of M and C.

The basic identities can be stated as follows

Theorem 1

$$H(K|C) = H(K) - H(C) + H(M) \tag{1}$$

Theorem 2

$$H(K|M, C) = H(K) - H(C|M) \tag{2}$$

Theorem 3

$$H(M|C) = H(K|C) - H(K|M, C) \tag{3}$$

Meaning:

The quantity $H(K|C)$ gives the uncertainty about the key given the ciphertext. Similarly, $H(K|M, C)$ gives the uncertainty of the key given both message and cyphertext. This means that if K is to be a function of C we must have $H(K|C) = 0$. Similarly, if K is to be a function of M and C we must have $H(K|M, C) = 0$.

EXAMPLE:

Let us suppose that we are enciphering english messages which have N characters each. Let $X, X_1, \ldots, X_N$ denote the successive letters in a message. If we adopt the model of english language which assumes that successive characters are independent, that is we ignore interletter dependencies, then we get

$$H(M) = H(X_1) + H(X_2) + \ldots + H(X_N) = N\ H(X_1)$$

The last equality holding if we assume that all letters in a message have the same distribution. The statistics of the english language give that

$$H(X_1) \approx 4.16.$$

from which we derive that

$$H(M) \approx 4.16N. \tag{4}$$

Taking account of biletter statistics, the experimental findings give

$$H(M) \approx 3.2N \tag{5}$$

Finally, for N larger than 15 it has been shown that

$$H(X|X_1, X_2, \ldots, X_N) \approx 1.2$$

For this reason, for large messages we should take a smaller value as an approximation of $H(M)$. We shall take here

$$H(M) \approx 2N \tag{6}$$

Which of (4), (5) or (6) we should use depends essentially on the length of the message and on the specific cryptographic system we are to study. Generally for short messages, ($N \leq 8$ characters), we shall use (4) , for $9 \leq N \leq 15$ we shall use (5), and for $N \geq 16$ we shall use (6).

This given let us assume that we are encrypting N characters of english text by a simple substitution (the same for each letter) but rather than Caesar we just take any random permutation of the letters of the alphabet. This gives us a key space with 26! elements. Assuming that the keys are chosen with a fortune wheel with equal slots gives

$$H(K) = \log_2 26! \approx 91.69$$

Assuming that all ciphers are equally likely [1] gives

$$H(C) = \log_2 26^N = N \log_2 26 \approx 4.7N$$

Formula (1) gives then that (using (6))

$$H(K|C) \approx 91.69 - N4.7 + 2N$$

Thus setting $H(K|C) = 0$ gives

$$91.69 = 2.7N \text{ or } N \approx 33.96$$

In other words, a cryptogram of 34 letters encrypted by a monoalphabetic substitution completely determines the key! Actually a slightly more refined argument leads to the conclusion that only 23 characters are sufficient to decipher such a cryptogram. The reason being that we do not need to find out what the actual key was since not all 26 letters of the alphabet will necessarily occur in the cryptogram. In fact we have the following table

$N =$	5	10	5	20	30	40	50
#diff chars	4	8	11	12	14	16	18

For instance this says that a message with 30 characters will have on the average only 14 different letters in it. Thus, using this information and the previous result, we can see that as far as the cryptoanalyst is concerned, the key space has only $26 \times 25 \times \ldots \times 13$ elements. That gives

$$H(K) = \log_2 26 \times \ldots \times 13 \approx 59.54$$

So setting $H(K|C) = 0$ now gives

$$59.54 = 2.7N \text{ or } N \approx 22.05$$

EXERCISES:

[1] An assumption that is valid just before it becomes impossible to derive the key from the ciphertext

1. Determine the unicity distance for a Vigenere encryption which uses an 8 character keyword. Assume all keys and ciphers are equally likely and the entropy of an English letter is 3.2 bits.

2. Determine the unicity distance for a Rectangular Transposition which uses a key of length 7. Assume all keys and ciphers are equally likely and the entropy of an English letter is 3.2 bits.

3. Determine the unicity distance for the Playfair encipherment scheme. Assume all keys and ciphers are equally likely and the entropy of an English letter is 2 bits. (Hint: There are less than 25! different keys.)

4. Determine the unicity distance for the ADFGVX encipherment scheme with a permutation of length 6, assuming that all keys are equally likely, all cipher characters are equally likely and the entropy of an English letter 1.2 bits.

5. Determine the unicity distance for the Hill encipherment scheme using a 2×2 key mod 29. Assume all keys and ciphers are equally likely and the entropy of an English letter is 3.2 bits. (Hint: The 2×2 matrix must be invertible. In other words, neither row may consist of all 0's and the second row cannot be a multiple of the first row.)