Ciphertext Only Attack of Rectangular Transposition

We shall view rectangular transposition in a slightly different manner than we did originally. We assume that the sender and receiver choose a secret period p and a secret permutation

$$\sigma = (\sigma_1, \sigma_2, \ldots, \sigma_p). \tag{1}$$

This done the plaintext

$$X_1, X_2, X_3, \ldots$$

is broken up into blocks of length p, the $(i+1)^{st}$ block being

$$X_{ip+1}, X_{ip+2}, \ldots, X_{ip+p}.$$

Then each block of plaintext is transposed by the permutation σ given in (1). That is the $(i+1)^{st}$ block is replaced by the p-gram

$$X_{ip+\sigma_1}, X_{ip+\sigma_2}, \ldots, X_{ip+\sigma_p}.$$

This done the resulting ciphertext is broadcast into 5-grams say, to masquerade the period. For simplicity, we assume that the sender adjusts its message to have length a multiple of the chosen period, say by wraparound or by adding gibberish at the end. Let us then denote the resulting ciphertext by

$$C_1, C_2, C_3, \ldots$$

The object of the opponent here is to recover the period p and the permutation σ from a ciphertext only attack. We shall present here a rother striking method which the opponent can use to this effect when sufficient ciphertext is available.

The method is based on simple statistics $Q_{i,j}(p)$ which are calculated as follows. Let $p_{s,t}$ denote the probability of encountering the pair st in the English language. This given, we set for a given pair of integers i, j between 1 and p,

$$Q_{i,j}(p) = \sum_{\substack{s,\,t \\ N_{s,t}(i,j) \neq 0}} p_{s,t} \log N_{s,t}(i,j)$$

where $N_{i,j}(s,t)$ denotes the number of times the pair st occurs in the sequence of ciphertext pairs

$$(C_i, C_j), (C_{i+p}, C_{j+p}), (C_{i+2p}, C_{j+2p}), \ldots.$$

The point of this is that if the j^{th} character of ciphertext did originally immedidately follow the i^{th} in the plaintext, and p is inded the period, then the $(j+kp)^{th}$ would also immediately follow the $(i+kp)^{th}$ for any value of k. This would then force the statistic $N_{s,t}$ to be very close to the value $N \times p_{s,t}$, where fore convenience we let N denote the total number of blocks of length p in the ciphertext. We would then approximately have

$$Q_{i,j}(p) \approx \log N + \sum_{\substack{s,\,t \\ p_{s,t} \neq 0}} p_{s,t} \log p_{s,t}. \tag{2}$$

On the other hand, if the guessed value of p is not the period or if the j^{th} letter did not follow the i^{th} in the plaintext, then we would have instead

$$N_{s,t} \approx N \times q_{s,t}$$

for some different (and most probably flat) probabilities $q_{s,t}$. Our statistic would then approximately be

$$Q_{i,j}(p) \approx \log N + \sum_{\substack{s,\,t \\ p_{s,t} \neq 0}} p_{s,t} \log q_{s,t}. \tag{3}$$

Now a simple theorem (see below) resulting from the convexity of the logarithm function tells us that the quantity in (3) is always less than or equal to that in (2) with equality holding if and only if $q_{s,t} = p_{s,t}$ for all pairs s, t.

Using this fact the opponent guesses a value of p and then constructs the $p \times p$ matrix of values of the statitics $Q_{i,j}(p)$ for all possible pairs i, j between 1 and p. Then one of two cases should arise

1. All the entries in the matrix are approximately the same.
2. Each row except one has a maximum that is much larger than all the other entries in the row and each column except for one has a maximum that is much larger than all the other entries in that column.

In case 1 the opponent concludes that p was a wrong guess for the period and proceeds to repeat the calculation with another value of p. In case 2 the opponent guesses that p is the right value and from the matrix progressively reconstructs the original permutation. To see how to do this it is best to look at an example.

Example:

Say p was actually equal to 5 and the permutation was

4 1 3 5 2

then the first 5 letters of plaintext, that is

$$X_1X_2X_3X_4X_5$$

would appear in the ciphertext in the order

$$X_4X_1X_3X_5X_2$$

This means that the ciphertext letter that used to follow the first letter of ciphertext in the original plaintext is the fourth! And of course this would be the same for each successive block of five letters in the ciphertext. This then (by our pervious reasoning) should suggest that the first five statistics

$$Q_{1,1}(5) \quad Q_{1,2}(5) \quad Q_{1,3}(5) \quad Q_{1,4}(5) \quad Q_{1,5}(5)$$

should turn out as follows

small small small LARGE small

A similar reasoning applied to the second letter of ciphertext gives that since the letter of ciphertext that originally followed that second is the fifth then the second five statistics

$$Q_{2,1}(5) \quad Q_{2,2}(5) \quad Q_{2,3}(5) \quad Q_{2,4}(5) \quad Q_{2,5}(5)$$

should turn out as follows

small small small small LARGE

The reader should conclude that if we place the statistics

$$Q_{i,1}(5) \quad Q_{i,2}(5) \quad Q_{i,3}(5) \quad Q_{i,4}(5) \quad Q_{i,5}(5)$$

in the i^{th} row of our matrix then the total matrix should come out as follows

small	small	small	LARGE	small
small	small	small	small	LARGE
LARGE	small	small	small	small
small	small	small	small	small
small	small	LARGE	small	small

The reason that we should expect the fourth row to have nothing but small values is simply due to the fact that the fourth letter of ciphertext has its immediate follower in the next block! That is none of the letters in its block can yield the proper statistics to give a large value to any of the entries in that row of the matrix. Thus a row of small entries is a giveaway that that particular letter of ciphertext was the last in its block.

Similarly, we should expect the second column to have nothing but small values because the second letter of ciphertext has its predecessor in the previous block!

Given all this information the opponent can easily reconstruct the original permutation. We urge the reader to devise an algorithm for reconstructing the permutation from such a matrix.

A Simple Theorem

Theorem 1 *Let $\{p_i\}$ and $\{q_i\}$ be two sequences of length $n \geq 1$ such that $p_1 + \cdots + p_n = q_i + \cdots + q_n = 1$ and $p_i, q_i \geq 0$ for all i. Then the following inequality must hold:*

$$\sum_{i=1}^{n} p_i \ln q_i \leq \sum_{i=1}^{n} p_i \ln p_i. \tag{4}$$

Proof

Draw a polygon with vertices $(x_i, x_i \ln x_i)$ where $x_i = p_i/q_i$ and $x_1 \leq x_2 \leq \cdots \leq x_n$. Since the function $f(x) = x \ln x$ is concave up for all $x > 0$, the center of mass of a system of n particles located at the aforementioned vertices must necessarily be inside the polygon.

In the event that the i^{th} particle has mass q_i, then the center of mass $(\bar{x}, \bar{y})$ is given by

$$\bar{x} = \sum_{i=1}^{n} q_i x_i = \sum_{i=1}^{n} p_i = 1$$
$$\bar{y} = \sum_{i=1}^{n} q_i x_i \ln x_i = \sum_{i=1}^{n} p_i \ln \frac{p_i}{q_i} = \sum_{i=1}^{n} p_i \ln p_i - \sum_{i=1}^{n} p_i \ln q_i$$

The fact that $(\bar{x}, \bar{y})$ is inside the polygon means that we must have the following inequality

$$0 = f(\bar{x}) \leq \bar{y} = \sum_{i=1}^{n} p_i \ln p_i - \sum_{i=1}^{n} p_i \ln q_i$$

which yields (4)

Exercises:

1. Decrypt the following message that was encoded using rectangular transpostion. The matrices provided should be enough to recover the period and key.

TIAHW TILTL TSDUE OYLUY OGLAL GNAWO
ASDNY IWHSI UOSDY RANTT WWTIO LATIH

$$\begin{bmatrix} 0 & 66 & 57 & 66 \\ 70 & 0 & 55 & 48 \\ 56 & 80 & 0 & 63 \\ 60 & 57 & 64 & 0 \end{bmatrix} \quad \begin{bmatrix} 0 & 47 & 44 & 73 & 51 \\ 72 & 0 & 53 & 48 & 54 \\ 43 & 53 & 0 & 52 & 41 \\ 51 & 42 & 72 & 0 & 47 \\ 42 & 66 & 46 & 43 & 0 \end{bmatrix} \quad \begin{bmatrix} 0 & 47 & 43 & 51 & 41 & 51 \\ 50 & 0 & 43 & 36 & 47 & 41 \\ 38 & 59 & 0 & 47 & 36 & 55 \\ 42 & 38 & 54 & 0 & 40 & 40 \\ 39 & 58 & 41 & 56 & 0 & 50 \\ 37 & 40 & 45 & 41 & 50 & 0 \end{bmatrix}$$

2. What was the permutation used to *encrypt* the message in the previous question?