1 Codebreaking

Continued from previous section.

1.1 Breaking Rectangular Transposition

Suppose you're given a long passage of ciphertext (with 2808 characters) that is known to be encrypted using rectangular transposition. How do we break the code? We'll talk about a strategy for breaking the code.

1. First, start by making an arbitrary guess for the "period," i.e., the length of the key word. We know that the period has to be a *divisor* of the length of the ciphertext. Note that 2808 has 32 possible divisors:

   $$\{1, 2, 3, 4, 6, 8, 9, 12, 13, 18, 24, 26, 27, \ldots, 234, 312, 351, 468, 702, 936, 1404, 2808\}.$$

   Since there are only 26 characters in the English alphabet, the period can be at most 26. This means that the period must be one of the following:

   $$\{1, 2, 3, 4, 6, 8, 9, 12, 13, 18, 24, 26\}.$$

   Suppose we guess that the period is 6.

2. Next, we can arrange our ciphertext into a rectangle of length 6 (the period we guessed). Note that our rectangle will have height $N = \frac{2808}{6} = 468$, so for the sake of being concise only the first few rows will be shown:

   ```
   OIPWMJ
   ALWSLE
   LJLYEA
   MENUAB
   IHSDAC
   ESRTIE
   EMKHAO
   AMNPAI
   IELNAP
      .
      .
      .
   ```

3. For every pair of numbers $i \neq j$ between 1 and 6 (the period we guessed), we consider the tall column of width 2 we would get by placing the ith and jth column of the above rectangle next to each other. For example, if $i = 4$ and $j = 2$, we would get the following 468×2 rectangle:

   ```
   WI
   SL
   YJ
   UE
   DH
   TS
   HM
   PM
   NE
   .
   .
   .
   ```

4. We can think of this as 468 observations of a pair of English letters *if* the columns i and j were consecutive in the plaintext.In particular, for every pair of letters α and β, we count the number of times that we see the sequence $\alpha\beta$ appearing in this column. Let $O_{\alpha\beta}^{(i,j)}$ be this number. In our truncated example above (in step 3), notice that $O_{\mathrm{WI}}^{(4,2)}$, $O_{\mathrm{SL}}^{(4,2)}$, etc. are all at least 1.

   On the other hand, we can use a large sample of English to calculate the probability $p_{\alpha\beta}$ of the pair $\alpha\beta$ occurring in the English text. We can use these to calculate the expected counts $E_{\alpha\beta} = Np_{\alpha\beta} = 468p_{\alpha\beta}$ and then calculate a corresponding value of G using the observed counts $O_{\alpha\beta}^{(i,j)}$. We can call this $G^{(i,j)}$; in other words,

$$G^{(i,j)} = \sum_{\alpha\beta} O_{\alpha\beta}^{(i,j)} \ln\left(\frac{O_{\alpha\beta}^{(i,j)}}{E_{\alpha\beta}}\right).$$

   We can then assemble all of these values[1] of $G^{(i,j)}$ as $i \neq j$ varies into a box of numbers:

$$\begin{bmatrix} \infty & G^{(1,2)} & G^{(1,3)} & G^{(1,4)} & G^{(1,5)} & G^{(1,6)} \\ G^{(2,1)} & \infty & G^{(2,3)} & G^{(2,4)} & G^{(2,5)} & G^{(2,6)} \\ G^{(3,1)} & G^{(3,2)} & \infty & G^{(3,4)} & G^{(3,5)} & G^{(3,6)} \\ G^{(4,1)} & G^{(4,2)} & G^{(4,3)} & \infty & G^{(4,5)} & G^{(4,6)} \\ G^{(5,1)} & G^{(5,2)} & G^{(5,3)} & G^{(5,4)} & \infty & G^{(5,6)} \\ G^{(6,1)} & G^{(6,2)} & G^{(6,3)} & G^{(6,4)} & G^{(6,5)} & \infty \end{bmatrix}.$$

   If we guessed the period correctly, then we should find that every row except *one of them* has *one* number that's much smaller than all the others. This tells us something about how to permute the letters to find the plaintext. For example, if we find in the first row that $G^{(1,4)}$ is *much* smaller than the other numbers, that tells us that rows 1 and 4 are likely to be *consecutive* in the plaintext, because the frequency distribution of the pairs that occur in the long 468×2 rectangle displayed earlier is close to the frequency distribution of pairs that occur in the English plaintext.

   Note that there are *many* calculations to do by hand. Therefore, we will make use of a compute to do these calculations for us.

(Example.) Suppose our "G-box" is

$$\begin{bmatrix} \infty & 1151.3 & 1090.2 & \underline{\mathbf{485.5}} & 1069.3 & 1005.0 \\ 1234.4 & \infty & 1228.3 & 1049.6 & \underline{\mathbf{440.2}} & 1148.6 \\ \underline{\mathbf{437.5}} & 1044.1 & \infty & 1004.1 & 1164.5 & 933.4 \\ 1154.7 & 1088.6 & 977.3 & \infty & 1115.7 & 1023.6 \\ 1137.2 & 1221.9 & \underline{\mathbf{425.9}} & 1100.0 & \infty & 1070.0 \\ 1003.7 & \underline{\mathbf{442.3}} & 944.9 & 1021.6 & 1086.1 & \infty \end{bmatrix}.$$

The numbers themselves are not very important. *However*, what's important is how every row except one has a number that's significantly smaller than the other numbers on that row. The numbers that are smaller than the others on the same row are bolded and underlined. Notice that every row except the fourth row has a bolded/underlined entry. Now,

- the fact that, in row 1, the 4th column is much smaller than the other entries in that row suggests that columns 1 and 4 in our 468×6 rectangle are consecutive.
- notice that, in row 2, the 5th column is much smaller than the other entries suggests that columns 2 and 5 are consecutive.

[1] Note that all the diagonal entries of this box are set to ∞ since we only compute $G^{(i,j)}$ when $i \neq j$. This is an arbitrary convention and the diagonal entries should just be ignored.

- the 4th row not having an entry that's much smaller than the others corresponds to the fact that the 4th column gets reordered to the end.

Observing all the relations this way, and then putting them together, we find that the above G-box leads us to think that the ordering of the columns is $\boxed{6, 2, 5, 3, 1, 4}$.

To clarify how the ordering was obtained, notice how

- in row 1, the smallest number is in column 4.
- in row 2, the smallest number is in column 5.
- in row 3, the smallest number is in column 1.
- in row 4, no number is significantly smaller, so we can assume that the 4th column was reordered to the end.
- in row 5, the smallest number is in column 3.
- in row 6, the smallest number is in column 2.

With this in mind, notice how we have pairs $(1, 4)$, $(2, 5)$, $(3, 1)$, $(5, 3)$, and $(6, 2)$. If we "connect" the pairs, we end up with

$$(6, 2), (2, 5), (5, 3), (3, 1), (1, 4).$$

Removing the connecting duplicate numbers yields

$$6, 2, 5, 3, 1, 4.$$

(Exercise.) Suppose that, when trying to break rectangular transposition, you find "G-boxes" of the following forms, where the exclamation mark indicates an entry that is much smaller than every other entry on its row. Write down the corresponding decrypting permutation (i.e., the ordering of the columns in the plaintext) that this configuration of values suggests.

(a) $$\begin{bmatrix} . & . & . & ! & . \\ . & . & . & . & ! \\ . & . & . & . & . \\ . & ! & . & . & . \\ . & . & ! & . & . \end{bmatrix}$$

Notice how

- in row 1, the exclamation mark is in column 4.
- in row 2, the exclamation mark is in column 5.
- in row 3, no exclamation mark exists, implying that 3 will be at the end of the ordering.
- in row 4, the exclamation mark is in column 2.
- in row 5, the exclamation mark is in column 3.

With this in mind, we have the pairs $(1, 4)$, $(2, 5)$, $(4, 2)$, and $(5, 3)$. If we "connect" the pairs, we end up with

$$(1, 4), (4, 2), (2, 5), (5, 3).$$

Joining the pairs (and removing the consecutive equal numbers) yields

$$1, 4, 2, 5, 3.$$

(b) $\begin{bmatrix} . & ! & . & . & . \\ . & . & . & . & . \\ ! & . & . & . & . \\ . & . & ! & . & . \\ . & . & . & ! & . \end{bmatrix}$

We have the pairs $(1, 2)$, $(3, 1)$, $(4, 3)$, and $(5, 4)$. "Connecting" them gives us

$$(5, 4), (4, 3), (3, 1), (1, 2).$$

Joining the pairs, removing the consecutive equal numbers, yields

$$5, 4, 3, 1, 2.$$

(c) $\begin{bmatrix} . & . & . & ! & . & . & . \\ . & . & . & . & . & . & ! \\ . & ! & . & . & . & . & . \\ . & . & ! & . & . & . & . \\ . & . & . & . & . & . & . \\ . & . & . & . & ! & . & . \\ . & . & . & . & . & ! & . \end{bmatrix}$

We have the pairs $(1, 4), (2, 7), (3, 2), (4, 3), (6, 5), (7, 6)$. Connecting them yields

$$(1, 4), (4, 3), (3, 2), (2, 7), (7, 6), (6, 5)$$

Joining the pairs, removing the consecutive equal numbers, yields

$$1, 4, 3, 2, 7, 6, 5.$$