THE VERNAM TWO TAPE SYSTEM

In an early (1926) paper on secret communication by wire and telegraph G. S. Vernam proposed an encryption system based on a pseudo random one time pad. The pad itself was to be constructed from two relatively short random keys.

More precisely, the Vernam system can be described as follows:

1) We have two key sequences

$$U = (u_1, u_2, ..., u_p) \ , \ V = (v_1, v_2, ..., v_q)$$

of $0's$ and $1's$ of lengths p and q respectively, where p and q are chosen relatively prime .

2) These sequences are then extended to arbitrary length by setting

$$u_{i+p} = u_i \quad , \quad v_{i+q} = v_i \ ,$$

that is, the extended U and V are made periodic of periods p and q respectively.

3) A long sequence $\mathbf{R}$

$$\mathbf{R} = (r_1, r_2, ..., r_n, ...)$$

is then constructed by setting

$$r_i = u_i + v_i \qquad (mod\ 2) \tag{1}$$

It can be shown that the sequence $\mathbf{R}$ will be periodic of period no longer than pq. Moreover, if U and V are randomly selected then with very high probability the period will be exactly pq. For instance for $p = 63$ and $q = 71$ the sequence $\mathbf{R}$ can be made to have period 4473. Thus we can see that a long non-repeating sequence can be produced by two quite short sequences. This fact suggested Vernam that such a sequence as $\mathbf{R}$ could replace the random sequences used as one-time pads in a perfect secrecy system.

4) Once $\mathbf{R}$ has been constructed, a $(0, 1)$-plaintext message

$$X = (x_1, x_2, ..., x_{pq})$$

is encrypted into the cyphertext

$$Y = (y_1, y_2, ..., y_{pq})$$

by setting

$$y_i = x_i + r_i \qquad (mod\ 2) \tag{2}$$

5) The receiver, given the keys U, V calculates the r_i by means of (1) and recovers the original message from the formula

$$x_i = y_i - r_i \ = \ y_i + r_i \qquad (mod\ 2)$$

Now it develops that the opponent can recover the sequence $\mathbf{R}$ and thus the original message as well from the knowledge of relatively few corresponding cyphertext-plaintext pairs

$$(x_i, y_i)$$

We shall show that this is so also in a more general setting than we have described above. Indeed, with no additional effort we can proceed under the assumption that the characters used in $\mathbf{U}$, $\mathbf{V}$, $\mathbf{X}$ and $\mathbf{Y}$, are the integers in $[1, m-1]$ and all operations are carried out *mod* m.

Under this setting we have the following basic result

Theorem 1

The opponent needs no more than $p+q-1$ *plaintext-cyphertext pairs to recover the original key. Infact, the knowledge of any sequence of pairs*

$$(x_i, y_i) \qquad (i = s+1, s+2, ..., s+n) \tag{3}$$

yields the original keys as long as $n \geq p+q-1$.

We shall establish this result by giving an algorithm for recovering $\mathbf{R}$ from the given pairs. However, before doing so it is good to make some observations of purely theoretical nature to see why such a result must be true. To this end observe that as long as the modulus m is a prime number, the integers *mod* m behave very much as the rational numbers and thus all the results of elementary linear algebra hold true also for vector spaces of n-tuples of integers *mod* m.

Now observe that the opponent's knowledge of a pair

$$(x_i, y_i)$$

yields him r_i through the relation

$$r_i = y_i - x_i \qquad (mod\ m)$$

This given, he obtains information about $\mathbf{U}$ and $\mathbf{V}$ from the equations

$$u_i + v_i = r_i (mod\ m) \tag{4}$$

We are then to find out how many of these equations are necessary to recover the original $\mathbf{U}$ and $\mathbf{V}$. Of course, we have pq equations altogether. However, it would be a useless exercise for the opponent to recover $\mathbf{U}$ and $\mathbf{V}$ from the knowledge of the full plaintext, since the whole *one time pad* $\mathbf{R}$ is going to be used only once! So it is of paramount importance that considerably less that the full text yields the rest of it.

$$\mathbf{U} = (u_1, u_2, ..., u_p)$$

The basic question then is how many independent equations do we have in (4). This question is easily answered if we look at the situation in the following manner. First of all let us consider both

U and V as vectors of indefinite lengths whose entries are periodic of periods p and q respectively. The vector **R** then being the sum of these two vectors lies in the vector space which consists of the vectors which are expressible as the sum of two such sequences. More precisely, let **U** and **V** denote the vector spaces consisting of the sequences which are periodic of periods p and q respectively. and let

$$\mathbf{W} = \mathbf{U} + \mathbf{V}$$

denote the direct sum of these vector spaces. In other words **W** represents the space of vectors which are expressible as the sum of a vector from **U** plus a vector from **V** . Thus another way of expressing our previous statements is that our vector **R** belongs to **W** . Our problem can be stated as follows then

How many of the components of a vector in **W** determine the rest of the vector?

Now, from linear algebra we get that the least number of components needed to determine a vector from a given vector space is exactly equal to the dimension of that space. From linear algebra we also get that the dimension of a direct sum is given by the formula

$$dim\ \mathbf{U} + \mathbf{V} = dim\ \mathbf{U} + dim\ \mathbf{V} - dim\ \mathbf{UV} \tag{5}$$

Now clearly the dimensions of **U** and **V** in our case are given by p and q respectively. Thus to find our answer we need only calculate the dimension of the intersection of **U** and **V**, that is the dimension of the space of sequences that are both periodic of period p and q. To determine this dimension, let us suppose we have such a sequence, call it **D**, that is

$$\mathbf{D} = (d_1, d_2, ..., d_n, ...)$$

and

$$d_{i+p} = d_i, d_{i+q} = d_i$$

holds for all i. Now since p and q are relatively prime we can find (by the Euclidean algorithm) two integers h and k so that

$$1 = hp - kq$$

this done we note that for any s we must have

$$d_s = d_{s+hp} = d_{s+1+kq} = d_{s+1}$$

In other words, what we have derived is the following remarkable fact.

Any sequence which has two relatively prime periods is necessarily constant!

The consequence of this fact for our considerations here is that the intersection **U** **V** of our two spaces is the one dimensional space consisting of vectors whose components are all equal to some constant. Thus from formula (5) we derive that

$$dim\ U + V = p + q - 1$$

and this is precisely our assertion that only $p+q-1$ components of $\mathbf{R}$ are needed to recover $\mathbf{R}$.

Now it develops that in our case here any $p+q-1$ components of $\mathbf{R}$ are sufficient to recover the rest of $\mathbf{R}$. This fact follows from the following simple identity satisfied by $\mathbf{R}$ and holding for any s whatever

$$r_{s+1} + r_{s+2} + ... + r_{s+p} = r_{s+q+1} + r_{s+q+2} + ... + r_{s+q+p} \tag{6}$$

This means that if we know any $p+q-1$ successive values of r_i, this relation can be used to find the preceding as well as the following value of r_i. For instance suppose that $p=3$ $q=5$ $m=26$ and the first $3+5-1=7$ values of r_i are

$$7 \quad 0 \quad 2 \quad 4 \quad 10 \quad 17 \quad 22$$

then from (6) we obtain (setting $s=0$) that

$$7+0+2 = 17+22+r_8 \pmod m$$

and this gives $r_8 = 22$. But now of course we can use (6) with s=1 and recover r_9... We can see that recursively we can recover all values of r_i without ever having to find $\mathbf{U}$ or $\mathbf{V}$!

It might be worthwhile seeing why formula (6) holds at all. For this we need only look at an example. Let $p=4$ and $q=7$. Then for the first $p+q(=11)$ values of r_i we have

$$\begin{aligned}
r_1 &= u_1 + v_1 \\
r_2 &= u_2 + v_2 \\
r_3 &= u_3 + v_3 \\
r_4 &= u_4 + v_4 \\
r_5 &= u_1 + v_5 \\
r_6 &= u_2 + v_6 \\
r_7 &= u_3 + v_7 \\
r_8 &= u_4 + v_1 \\
r_9 &= u_1 + v_2 \\
r_10 &= u_2 + v_3 \\
r_11 &= u_3 + v_4
\end{aligned} \tag{7}$$

Now note that in the sum of the right hand sides of the first $p(=4)$ of these equations we will have

$$u_1, u_2, u_3, u_4, v_1, v_2, v_3, v_4$$

appearing. However, if we do the same for the last p equations we see that the same quantities do appear! Thus (6) must necessarily follow in this case. We can easily see that the same idea works in the general case as well.

Using this fact we can easily decrypt any message encrypted by the Vernam system with parameters p and q and $m = 26$ if we know that a certain word or sentence containing $p + q - 1$ characters or more appears somewhere in the original plaintext. We can locate this segment in the cyphertext very easily using (6). An example will suffice.

Let $p = 2$ and $q = 3$ and the cyphertext be

SYTPF HLYVG WQLYM G

Assume that the original plaintext was encrypted by converting the letters into the integers $0, 1, \ldots, 25$ and then all arithmetic was carried out mod 26. Let us suppose that the original sentence started with the letters

NO MORE

we then get the following values of r_i

$$r_1 = S - N = 18 - 13 = 5$$

$$r_2 = Y - O = 24 - 14 = 10$$

$$r_3 = T - M = 19 - 12 = 7$$

$$r_4 = P - O = 15 - 14 = 1$$

now from our theory we get that

$$r_5 = 5 + 10 - 1 = 14$$

this gives that the 5th letter of the plaintext must have been

$$x_5 = y_5 - r_5 = F - 14 = 5 - 14 = 17 = R$$

and as we see this checks beautifully! The reader is urged to continue this calculation and determine the rest of the plaintext.

We should note that the equations in (7) determine U and V only up to an additive constant. So that it may be impossible to recover **U** and **V** entirely from the given data. However, for long texts we can efficiently do our decryption if we can make use of the values of U and V that can be derived from (7) **under the assumption** that $u_1 = 0$. This may not be the case for the actual **U** used in encryption, but no matter, the resulting **U** and **V** can be used nevertheless to reconstruct the entire sequence **R**, with a lesser number of operations than needed in the method we described above. The idea is that setting $u_1 = 0$ yields v_1 from the first equation. then the 8th equation in (7) yields u_4, and so from the 4th we get v_4, thereby from the 11th we get u_3 , then the third gives v_3 and finally the 10th yields u_2. Now that all the values of u_i within a period p are obtained the remaining values of v_i can be obtained from the rest of the equations in (7). With these "fake" values of **U** and **V** the whole vector **R** can be reconstructed simply using the original formula (1) as if they were the true ones.

Exercises

1) Find the plaintext from the given cyphertext given that it was encrypted by Vernam cipher with a keys of lengths 3 and 4 and that the word ARTIST appears starting at the third letter of of the text:

HJDWU ZGSJG DYYNH FMCME EPQMO AVMJJ KNNM