Cryptography and computer security

Historical cryptography

Cont. No 2

One-Time Pad

- Gilbert Vernam ATT, 1917 One-Time Pad,
  - The only one absolutely secure cipher (the "ideal" frequency distribution of the ciphertext letters)
  - $P=C=K=\{0,1\}$, $|P|=|C|=|K|=\{0,1\}^n$,
  - the key is a stream of n bits, $k=(k_1,k_2,...,k_n)$,
  - $e_k(p_1,p_2,...,p_n)=(p_1+k_1) \bmod 2, ..., (p_n+k_n) \bmod 2= (p_1 \oplus k_1) , ..., (p_n \oplus k_n)$,
  - $d_k(c_1,c_2,...,c_n)=(c_1-k_1) \bmod 2, ..., (c_n-k_n) \bmod 2= (c_1 \oplus k_1) , ..., (c_n \oplus k_n)$

One-Time Pad: Encryption

e=000 h=001 i=010 k=011 l=100 r=101 s=110 t=111

Encryption: Plaintext ⊕ Key = Ciphertext

	h	e	i	l	h	i	t	l	e	r
Plaintext:	001	000	010	100	001	010	111	100	000	101
Key:	111	101	110	101	111	100	000	101	110	000
Ciphertext:	110	101	100	001	110	110	111	001	110	101
	s	r	l	h	s	s	t	h	s	r

One-Time Pad: Decryption

e=000 h=001 i=010 k=011 l=100 r=101 s=110 t=111

Decryption: Ciphertext ⊕ Key = Plaintext

	s	r	l	h	s	s	t	h	s	r
Ciphertext:	110	101	100	001	110	110	111	001	110	101
Key:	111	101	110	101	111	100	000	101	110	000
Plaintext:	001	000	010	100	001	010	111	100	000	101
	h	e	i	l	h	i	t	l	e	r

One-Time Pad

Double agent claims sender used following "**key**"

	s	r	l	h	s	s	t	h	s	r
Ciphertext:	110	101	100	001	110	110	111	001	110	101
"**key**":	101	111	000	101	111	100	000	101	110	000
"Plaintext":	011	010	100	100	001	010	111	100	000	101
	k	i	l	l	h	i	t	l	e	r

e=000 h=001 i=010 k=011 l=100 r=101 s=110 t=111

One-Time Pad

Or sender is captured and claims the "key" is…

	s	r	l	h	s	s	t	h	s	r
Ciphertext:	110	101	100	001	110	110	111	001	110	101
"key":	111	101	000	011	101	110	001	011	101	101
"Plaintext":	001	000	100	010	011	000	110	010	011	000
	h	e	l	i	k	e	s	i	k	e

e=000 h=001 i=010 k=011 l=100 r=101 s=110 t=111

One-Time Pad Summary

- Provably secure…
- Ciphertext provides no info about plaintext
- All plaintexts are equally likely
- …but, only when be used correctly
- Key must be random, used only once
- Key is known only to sender and receiver
- Key is same size as message

Real-World One-Time Pad

- Project https://en.wikipedia.org/wiki/Venona_project
- Encrypted spy messages from U.S. to Moscow in 30's, 40's, and 50's
  - Nuclear espionage, etc.
  - Thousands of messages
- Spy carried one-time pad into U.S.
- Spy used pad (key) to encrypt secret messages
- Repeats within the "one-time" pads made cryptanalysis possible

Homophonic Substitution Cipher

- The Homophonic Substitution cipher is a substitution cipher in which single plaintext letters can be replaced by any of several different ciphertext letters.
- The easiest way to break standard substitution ciphers is to look at the letter frequencies, the letter 'E' is usually the most common letter in english, so the most common ciphertext letter will probably be 'E' (or perhaps 'T').
- If we allow the letter 'E' to be replaced by any of e.g. 5 (8, 12 etc.) different characters, then we can no longer just take the most common letter, since the letter count of 'E' is spread over several characters. As we allow more and more possible alternatives for each letter, the resulting cipher can become very secure.
- The number of characters each letter is replaced by is part of the key, e.g. the letter 'E' might be replaced by any of 5 (8, 12 etc.) different symbols, while the letter 'Q' may only be substituted by 1 symbol.

Homophonic Substitution Cipher

- An Example

Plaintext: HE EATS THEN SLEEPS AND DREAMS **Ciphertext:** ??

Key: The following table is the key.

A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z
11	21	89	58	QP	15	JK	55	BC	B	47	PB	AA	49	50	ZS	A	43	JK	J	90	76	CT	93	30	13
AZ	12	ZA	ND	69	RF	FR	OG	61	71	GC	VC	CG	BB	49	SC	SP	CR	87	77	QQ	VM	59	HR	XT	NE
@			47	??			81											XX	WO						
QA				99			SS											88							
XF				101																					
				DD																					
				YD																					

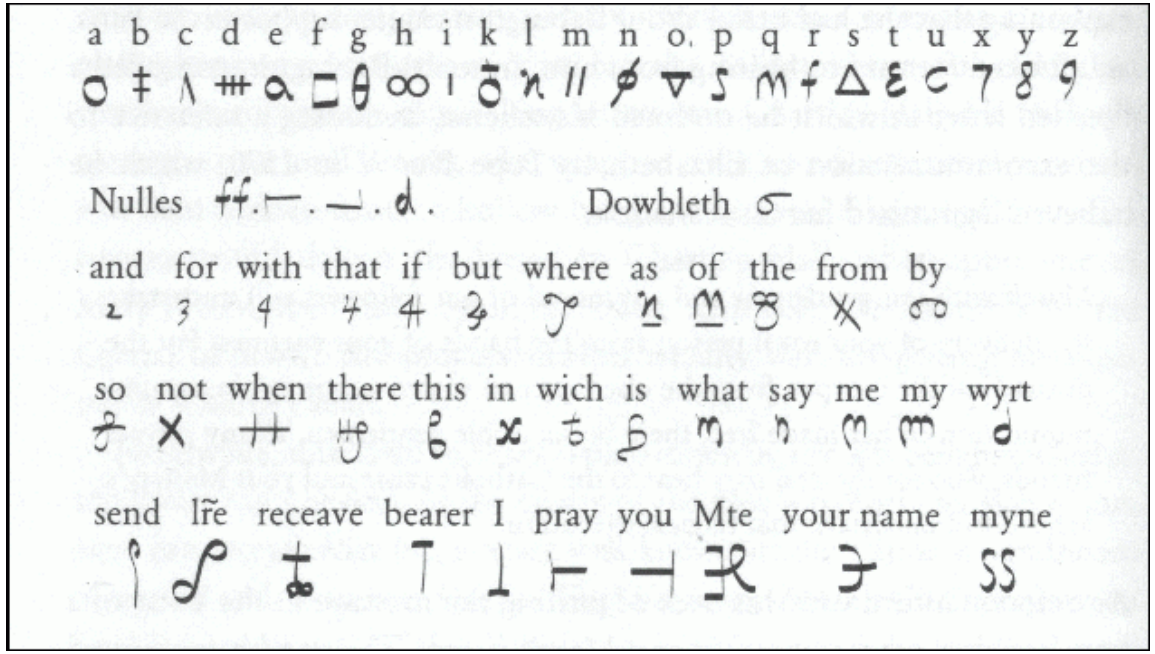
Homophonic Cipher Table

Ciphertext:

55QP69 **11JJK** 77OG?? **4987PB** 99101ZS **XXAZBB** 58ND43 **DD@AA** 88

Nomenclators

- Another variant is the **nomenclator**, where codewords are used to substitute many common words and names. The example below was used by Mary Queen of Scots in 1586.



Codes, Codebooks

- https://en.wikipedia.org/wiki/Code_(cryptography)
- Code is a method used to encrypt a message that operates at the level of meaning; that is, words or phrases are converted into something else. A code might transform "change" into "CVGDK" or "cocktail lounge".

Identifying Unknown Ciphers

- Classes of ciphers
  - Transposition ciphers - involve permuting the positions of the characters, but leaving the identity of the characters unchanged.
  - Monoalphabetic substitution ciphers - each letter is replaced with another.
  - Polyalphabetic ciphers - different alphabets are used to encipher letters depending on their position.
  - Polygram Substitution ciphers - groups of characters are replaced.
- Given that there are so many different ciphers, how can we expect to identify a piece of ciphertext? Different ciphers leave different 'fingerprints' on the ciphertext which we can use. Some of the fingerprints are very faint though.
- We need quite a bit of ciphertext, 1000 or more characters is ideal.
  - If all we have is 20 characters there is not much you can do. Very short ciphers may be unbreakable if their length is less than the Unicity Distance (later …) of the cipher used to encipher them.

Identifying Unknown Ciphers

- How many different characters are there?
  - If there are only 2 different symbols, it is likely the cipher is Baconian (https://en.wikipedia.org/wiki/Bacon%27s_cipher).
  - If there are 5 or 6 it is probably a polybius square cipher of some sort, or it may be ADFGX or ADFGVX.
  - If there are more than 26 characters it is likely to be a code or nomenclator of some sort or a homophonic substitution cipher.
  - If there are 26 characters in the ciphertext, it rules out ciphers based on a 5 by 5 grid such as Playfair. But, if the ciphertext is fairly long and only 25 characters are present, it may indicate a cipher in this class has been used.

Identifying Unknown Ciphers

- English text has a very specific frequency distribution that is not changed by transposition ciphers.
  - All other ciphers change this distribution, so the frequencies can be used to differentiate them.
  - If the frequency distribution looks exactly like a piece of English text but it is still unreadable we can conclude it is probably a transposition cipher, otherwise we move onto the next step.
- The next step is to determine if the cipher is a substitution cipher of some sort.
  - Here we calculate the Index of Coincidence.
  - If the Index of Coincidence is around 0.06 we conclude the cipher is probably a substitution cipher.
  - If it is lower, it is most probably some sort of polyalphabetic, polygram or more complex cipher.

Identifying Unknown Ciphers

- If the cipher is polygram (polygraphic), the length must be a multiple of the n-gram size. E.g. If the ciphertext has an odd number of characters it can't be a bigraphic cipher (replaces pairs of characters) such as Playfair. If the length is not a multiple of 3 it can't be a 3x3 Hill cipher.
- If it is a Vigenere cipher a periodic I.C. calculation will identify large peaks at the length of the keyword. No other ciphers have this property.

Index of Coincidence

- The index of coincidence is a measure of how similar a frequency distribution is to the uniform distribution. The I.C. of a piece of text does not change if the text is enciphered with a substitution cipher. It is defined as:

$$I.C. = \frac{\sum_{i=A}^{i=Z} f_i(f_i - 1)}{N(N-1)}$$

- where f_i is the count of letter i (where i = A,B,...,Z) in the ciphertext, and N is the total number of letters in the ciphertext.

Index of Coincidence

- The index of coincidence of:
  - Randomly generated words 1/26 = 0.03846
  - Czech 0.06027
  - English 0.06689
  - French 0.07460
  - German 0.07667
  - Spanish 0.07661

Index of Coincidence - example

- Example:
- „aaaabbc“
- The Index of Coincidence tells us how likely it is that if we randomly choose two letters from the text to be the same. For example, in "aaaaaa" we have a 100% probability that the selected pair will be the same. In the text "aaaabbc" the probability is 1/3, and the most likely one is that we choose two letters "a".

Index of Coincidence - example

- We have to count the number of the same pairs of letters. How many different pairs of "a" letters exist?
- We have a total f_a option of selecting the letter "a" and we have a total of f_a-1 options for adding "a" to it. Because it does not matter the order, it will be divided by 2. The number of all the different pairs of letters "a" is:

$$(f_a(f_a\text{-}1\ ))/2$$

- For our text:

$$\frac{4 \cdot 3}{2} + \frac{2 \cdot 1}{2} + \frac{1 \cdot 0}{2} = 7$$

- And these are the pairs of the same letters: "aaaabbc", "aaaabbc", "aaaabbc", "aaaabbc", "aaaabbc", "aaaabbc".

Index of Coincidence - example

- To obtain probability, we divide this value by the number of all possible pairs of letters (the same or different) that exist in the text. If we denote the length of the text N, then the number of all pairs will be

$$N\,(N-1)\,/\,2$$

- The reason is again the same: we have a total of N options to select a letter and we have a total of N - 1 options to add another letter to form a pair. Since it does not matter the order, we divide by two. The overall probability, and at the same time the coincidence index, is equal I.C.:

$$I.C. = \frac{\sum_{i=A}^{i=Z} f_i(f_i - 1)}{N(N-1)}$$

- For our example:

$$\frac{4 \cdot 3 + 2 \cdot 1 + 1 \cdot 0}{7 \cdot 6} = \frac{14}{42} = \frac{1}{3}$$

Cryptanalysis of the Vigenere Cipher

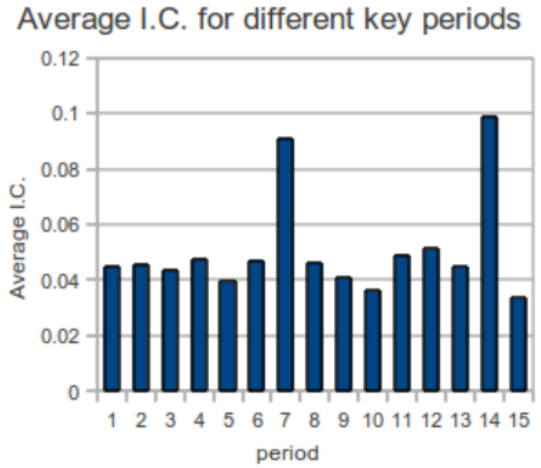
- http://practicalcryptography.com/cryptanalysis/stochastic-searching/cryptanalysis-vigenere-cipher/
- To determine the period of a Vigenere cipher we first assume the key length is 2. We extract the two sequences 1,3,5,7,... and 2,4,6,8,... from the ciphertext. For the example we are working with we get the following result (note that the I.C. is calculated using the whole sequences, not just the part shown)

```
                                                          I.C.
original:    vptnvffuntshtarptymjwzirappljmhhqvsubw...     0.049
if key were length 2:
sequence 1:  v t v f n s t r t m w i a p j h q s b ...     0.049
sequence 2:   p n f u t h a p y j z r p l m h v u w...     0.046
                                                average:  0.048
if key were length 3:
sequence 1:  v  n  f  t  t  p  m  z  a  l  h  v  b ...     0.049
sequence 2:   p  v  u  s  a  t  j  i  p  j  h  s  w...     0.046
sequence 3:    t  f  n  h  r  y  w  r  p  m  q  u  ...     0.046
                                                average:  0.047
```

Cryptanalysis of the Vigenere Cipher

- This procedure of breaking up the ciphertext and calculating the I.C. for each subsequence is repeated for all the key lengths we wish to test. What we are most interested in is the average I.C. for a particular period, for the case of period = 2, the average I.C. is around 0.048. If you were to continue this procedure up to a period of 15 we get the following average I.C. values:

period	avg I.C.
1 :	0.0449443523561
2 :	0.0457833618884
3 :	0.0435885364312
4 :	0.0474962292609
5 :	0.0393612078978
6 :	0.0471437059672
7 :	0.0909922589726
8 :	0.0461858974359
9 :	0.0407804755631
10:	0.0361152882206
11:	0.0491603339901
12:	0.0512663398693
13:	0.0446886446886
14:	0.0988487702773
15:	0.0334554334554



- We have 2 rows that have very high values of average I.C. This indicates the key is probably of length 7, but could also be of length 14. Both of these probabilities should be tested.

Cryptanalysis of Polyalphabetic Ciphers

- in order to break a polyalphabetic (e.g. Vigenere cipher) cipher must
  - first determine how many alphabets were used: Kasiski method & Index of Coincidence help estimate period d
  - then separate ciphertext into d sections
  - then solve each as a monoalphabetic cipher using frequency distribution, common double & triple letters and word boundaries
  - taking care to identify which alphabet each letter in a group comes from

Kasiski Method

- original method developed by Babbage and Kasiski
- use repetitions in ciphertext to give clues as to period
- look for same plaintext an exact period apart, which results in the same ciphertext (!could also be random fluke)
- eg. Plaintext: TOBEORNOTTOBE
- Key: NOWNOWNOWNOWN
- Ciphertext: GCXRCNACPGCXR
- see repeated ciphertext "GCXR"
- since repeats are 9 chars apart, guess period is 3 or 9
- in general find a number of duplicated sequences
- collect all their distances apart, look for common factors
- remembering that some will be random flukes and need to be discarded

- Original plaintext - message in English, all spaces were replaced by letter 'Z', ciphertext obtained by Vigenère cipher with periodical key
- HQEOT FNMKP ELTEL UEZSI KTFYG STNME GNDGL PUJCH QWFEX
  FEEPR PGKZY EHHQV PSRGN YGYSL EDBRX LWKPE ZMYPU EWLFG
  LESVR PGJLY QJGNY GYSLE XVWYP SRGFY KECVF XGFMV ZEGKT
  LQOZE LUIKS FYLXK HQWGI LF
- Solution: repeated digrams an their positions in the ciphertext:

  EL - 11, 14 and 140,
  FY - 23, 119 and 146,
  GN - 31, 64 and 103,
  HQ - 1, 40, 58 and 151,
  LE - 70, 91 and 109,
  YG - 24, 66 and 105.

- in addition: digram GN on positions 64 a 103 is the beginning of sequence GNYGYSLE. Distance between these sequences is 103 - 64 = 39 = 3 * 13
- Distances between bigrams are:

EL: 3 a 126 = 3 * 42,
HQ: 39, 18 a 93,

- .... all are products of 3. That suggests that 3 is by far the most likely key length.

- The next step is to find the key itself. Hypothesis: for encryption three different shifts have been used . The first shift was used on the letters at places 1, 4, 7, 10, etc., the second shift at the places 2, 5, 8, 11, etc., and a third shift to places 3, 6, 9, 12, Therefore, we write the ciphertext into three columns as follows.

```
1 2 3
    H Q E
    O T F
    N M K
    P E L
    T E L
    U E Z
    S I K
    T F Y
    G S T
    N M E
    . . .
```

- The first column has 53 letters, second and third 52 letters. Now we calculate the number of occurrences of letters in each of the three columns and write them into table:

column	A	B	C	D	E	F	G	H	I	J	K	L	M	...	Z
1	0	1	0	0	0	3	13	4	0	0	1	7	1	...	1
2	0	0	0	0	13	6	0	0	3	2	2	1	2	...	1
3	0	0	2	2	4	1	1	1	0	1	5	5	1	...	3

- If the frequency of individual letters are random, we might expect that the number of occurrences of letters in each row are approximately 2.
- But as the numbers in each row correspond to the frequencies of letters in natural language, we expect the numbers from 0 to 10, with the highest incidence is most likely corresponds to the letter in a ciphertext that encrypts the letter Z in an plaintext, replacing the space ' '.
- This is because each column is composed of the letters of the plaintext shifted by the same number of letters.
- In the first column is the most common letter G, which is assumed to be ciphertext of Z. The first shift is thus probably 7 letters. In this case, the most often letter E in a plaintext should correspond to the letter L in the ciphertext, which is actually the second most common letter in the first column. This further supports our hypothesis, that the first shift is 7 letters.
- In the second column is the most common letter E, which is thus a good candidate for encrypting letters Z, (i.e. spaces in plaintext). This means that the second shift is likely to be 5 points. The two most common ciphertext letters in the second column are the F and Q, from which when you move back 5 positions become letters A and L, which are frequent in real plaintexts. By contrast, ciphertext of widely frequented letter in plaintexts, the letter E, by displacement by five positions ahead results in the J, which in the second column appears only twice. Our hypothesis, that the second shift is 5 points ahead, is not very convincingly supported.
- In the third column no ciphertext letter is too frequent, and so we have no reasonable estimate of the third shift. The most common letters in the third column are Y, K and L, one of them will probably replace the open space Z, but we do not know which.
- Therefore, we try to write the ciphertext and the corresponding palintext letters obtained by our estimates of the first and second shift.

```
HQEOTFNMKPELTELUEZSIKTFYGSTNMEGNDGLPUJCHQWFEXFEEPRPGKZY
AL.HO.GH. I.M  .
```

- The first word looks like ALTHOUGH and, if so, then opentext letter T is in the third column replaced by ciphertext E, which means shifting by 11 places ahead.
- In this case, the open space (i.e. letter Z), corresponds to the ciphertext letter K, which is actually one of the most likely candidates for encrypting Z at the third shift.
- We therefore conclude that the encryption key is 7,5,11 and decryption key is 19,21,15. If the decryption key used, we get plaintext

```
ALTHOUGH I AM AN OLD MAN NIGHT IS GENERALLY MY TIME FOR
WALKING IN THE SUMMER I OFTEN LEAVE HOME EARLY IN THE
MORNING AND ROAM ABOUT FIELDS AND LANES ALL DAYS
```

- This is the beginning of one of the novels of Charles Dickens

Cryptanalysis of Row Transposition ciphers

- a frequency count will show a normal language profile
- hence know have letters rearranged
- basic idea is to guess period, then look at all possible permutations in period, and search for common patterns
- use lists of common pairs & triples & other features
- if row transposition then letters rearranged within row
- and must be same rearrangement for each group of letters
- ability to anagram words helps a lot here

Cryptanalysis of Row Transposition ciphers

- given: `LDWOE HETTS HESTR HUTEL OSBED EFIEV NT`
- try successive periods, looking at all rearrangements of first few letters
- `2: LD WO EH ET TS HE ST RH UT EL OS BE DE FI EV NT` - NO
- `3: LDW OEH ETT SHE STR HUT ELO SBE DEF IEV NT` - NO
- `4: LDWO EHET TSHE STRH UTEL OSBE DEFI EVNT` - NO
- `5: LDWOE HETTS HESTR HUTEL OSBED EFIEV NT` - NO
- `6: LDWOEH ETTSHE STRHUT ELOSBE DEFIEV NT` - YES!!!
- note 2nd group suggests "THESET" or "TTHESE"
- 8 keys could give these patterns
- key 5,6,1,4,2,3 recovers English text:
- `WEHOLD THESET RUTHST OBESEL FEVIDE NT` or
- `WE HOLD THESE TRUTHS TO BE SELF EVIDENT`

Cryptanalysis of Columnar Transposition ciphers

- know must be a transposition, guess perhaps a columnar transposition
- try to guess size of matrix by looking at factors of message length
- otherwise simply have to try each size in turn
- write message out in columns
- look for ways of reordering pairs of columns to give common pairs or triples (very much trial & error)
- more generally, use automated tool to try all permutations
- and can suggest displaying only permutations matching a pattern
- pattern could be guessed word
- or could just assume have enough text that common words occur
- eg. the and (preferably repeated)
- process is called anagraming
- Ratio of r vowels/consonants in English is 40:60

- given:

```
EOEYE GTRNP SECEH HETYH SNGND ODDET OCRAE RAEMH TECSE
USIAR WKDRI RNYAC
ANUEY ICNTT CEIET US
```

- Note that this ciphertext has 77 letters. This suggests a block 7x11 or 11x7, although it could be a ragged rectangle 8x10 with the last 3 letters missing. To determine the correct number of rows, we look at all possible values. Since the correct number of rows will tend to keep letters from the same word clumping together, we expect that the variance in vowel/consonant ratios per row will be lower with the correct number of rows. The first rectangle is "better"

```
E E G A E R C - 4 vowels from 7 letters
O C N E U N N - 3 vowels
E E D R S Y T - 3 vowels
Y H D A I A T - 4 vowels
E H D E A R C - 3 vowels
G E D M R A E - 3 vowels
T T E H W N I - 2 vowels
R Y T T K U E - 3 vowels
N H O E D E T - 3 vowels
P S C C R Y U - 1 vowels
S N R S I I S - 2 vowels
```

```
E R H N E R C R N E C - 3 vowels from 11 letters
O N H G T A S W Y Y E - 3 vowels
E P E N O E E K A I I - 8 vowels
Y S T D C M U D R C E - 3 vowels
E E Y D R H S R A N T - 4 vowels
G C H D A T I I N T U - 4 vowels
T E S D E E A R U T S - 5 vowels
```

- The next step should be to try to fit two columns together so as to get good digrams.
- Pair each other column with it on its right-hand side. Look at the digrams thus created.
- Column 7 followed by column 2 produces many good digrams such as TH

```
 7 2        7 2 4        5 7 2 4                3 6 1 5 7 2 4
 C E        C E A        E C E A                G R E E C E A
 N C        N C E        U N C E                N N O U N C E
 T E        T E R        S T E R                D Y E S T E R
 T H        T H A        I T H A                D A Y I T H A
 C H        C H E        A C H E                D R E A C H E
 E E        E E M        R E E M                D A G R E E M
 I T        I T H        W I T H                E N T W I T H
 E Y        E Y T        K E Y T                T U R K E Y T
 T H        T H E        D T H E                O E N D T H E
 U S        U S C        R U S C                C Y P R U S C
 S N        S N S        I S N S  ... -->       R I S I S N S
```

--> The key is 3 6 1 5 7 2 4

- plaintext is

GREECE ANNOUNCED YESTERDAY IT HAD REACHED AGREEMENT WITH TURKEY TO END THE CYPRUS CRISIS