CSL7480 Cryptography **06 Jan 2023**

Lecture 2: Cryptography

Lecturer: Somitra Sanadhya *Scribe: Diksha Jena* $\|B20CS013$

2.1 Recap

- The Basic setting of communication is having two parties, a sender and a receiver and the aim is that the messages being shared between them maintains
  1. Confidentiality : the message should not be understood by the adversary (eavesdropper) i.e. there should not be any leak of information.
  2. Integrity : data should not be modifiable at any stage (while being communicated or during encryption/decryption).
- Kerckhoff's principle : states that the security lies completely and only on the secret key, i.e. even if the attackers knows everything except the secret key, he/she would still not be able to make any harm. In communication we require the algorithm to be 100 % accurate (not even a single bit of information should be revealed as it could be harmful : eg manipulating the largest bit in a check)
- Toy Cipher
  1. Shift (Caesar) cipher : Enc(k,m) = (m+k) mod 26
     computation required to crack is o(n) n=26, easily achievable on computer.
     we observe that a large key space should be maintained to avoid brute force attacks else any cipher could be broken and there would be no protection.
  2. Story of dancing man : in Sherlock Holmes, substitution cipher was used using figures of dancing characters.
     Substitution (permutation) cipher
     m = {a,b,c... z}
     k = { π a,b... z}
     we observe that default correlations existing due to the english language (eg words like the, are etc are very frequently used, e is the most used character) needs to be broken.
     As otherwise, some info of plain text is known (english language constraints) hence possible to break.

Toy ciphers are mono-alphabetic ciphers
one plain text character $\rightarrow$ specific cipher text characters (breakable)
Limitation : Information known is the length of the massage (equal to the length of cipher text), 26 characters possible and length known would make is possible to brute force.

Poly alphabetic Substitution : (solution)
minimal conditional requirement is that the cipher length is greater than the message length.
$\|C\| \geq \|M\|$
main idea is to disguise the plaintext letter frequency to interfere with a straightforward application of frequency analysis.

2.2 Vigenere Cipher

In vigenere cipher each letter has a different shift count. We have a key of some length, which may or may not be extended to make its length equal to that of message and then used to encode.
Usually preferred to have longer key length instead of repetition to increase length else it is easier to decode but having a key of such large length randomly and equi-probably generation itself poses as a problem (difficult to achieve).

$$\text{m} = \{x_1, x_2, x_3 ... x_n\} \text{ where } x_c \in \{\text{a,b,c...,z}\}$$
$$\text{Enc (k,m)} = (k_1 + m_1)\ (k_2 + m_2)\ ...\ (k_n + m_n)$$
$$\text{where each : } k_i + m_i = (k_i + m_i) \bmod 26$$

Key space : —K— $= 26^n$, each character can belong to any of the English alphabets, multiple mapping of keys is allowed (same key to multiple alphabets)

Attack of Vigenere Cipher
Condition : cipher text is reasonably long (statistical methods work only then)
Otherwise properties become sparse instead of being broken. (meaning that the chance of repetition of a character in both key and message is very low making it difficult to perform any analysis or observe any pattern.)

1. Assume Known n
2. Now attack : also a weakness that knowing the length n makes it easy to perform crypt analysis

   form n groups (where n is the length of key) it reduces to shift cipher.

If n is known : guess and determine attack (guess n, retrieve the most possible value of key and message followed by simply verifying if it correct)
Note : the space (character) in messages is assumed to be known.

Three methods to determine n :

2.2.1 Brute force

Take all possible values of n to break patterns of commonly known letters of the language. (eg the, of...)
Use of knowledge of plaintext or recognizable word as a key.

2.2.2 Kasiski's Test

While encrypting in vigenere cipher there is a possibility for repeating words to by chance be encrypted using the same keys.
This follows the conclusion that, if same letters of English alphabet repeats, than the gap between them is a factorial of the actual key (divide and guess).

Look for separate patterns in cipher text and break it, eg $|K| = 6 \longrightarrow |K| = 12$
Repeating pattern gap will also have a gap of its factorial.
Reduced no of K values to be checked.

2.2.3 Index of Coincidence

Also known as the Friedman's test or Kappa test. We know that in vigenere cipher matching characters in each columns would remain same before and after encryption.

Now, if we compare the probabilities of :

P(two random characters matching (in cipher)) : $\kappa_r = 1/26 = 0.039$
P(two characters from English to match) : $\kappa_p = \Sigma f_c^2 \ / \ 26^n$ = some constant 0.67

The two probabilities are fairly distinct because the frequency of usage of English characters in communication is different, while that of an encoded cipher is uniformly random (desired for 0 information retrieval).
* The frequency pattern of English characters in known (to the attacker / universally).

Key length is approximated as :

$$\frac{\kappa_p - \kappa_r}{\kappa_o - \kappa_r}$$

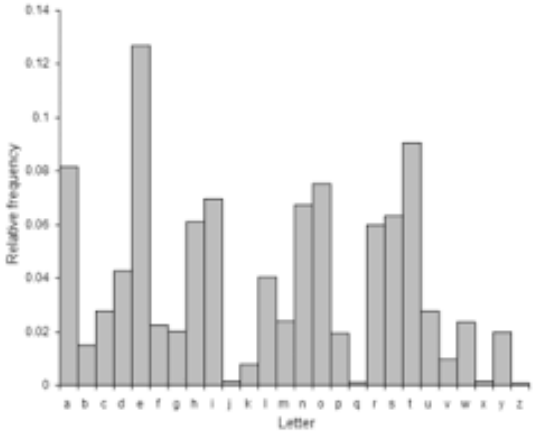
from the observed coincidence rate

$$\kappa_o = \frac{\sum_{i=1}^{c} n_i(n_i - 1)}{N(N-1)}$$

in which c is the size of the alphabet (26 for English), N is the length of the text and n_1 to n_c are the observed ciphertext letter frequencies, as integers. If the index of coincidence is equal/closer to 0.67 than correct n received, else check for the next possible value.

E	11.1607%	56.88	M	3.0129%	15.36
A	8.4966%	43.31	H	3.0034%	15.31
R	7.5809%	38.64	G	2.4705%	12.59
I	7.5448%	38.45	B	2.0720%	10.56
O	7.1635%	36.51	F	1.8121%	9.24
T	6.9509%	35.43	Y	1.7779%	9.06
N	6.6544%	33.92	W	1.2899%	6.57
S	5.7351%	29.23	K	1.1016%	5.61
L	5.4893%	27.98	V	1.0074%	5.13
C	4.5388%	23.13	X	0.2902%	1.48
U	3.6308%	18.51	Z	0.2722%	1.39
D	3.3844%	17.25	J	0.1965%	1.00
P	3.1671%	16.14	Q	0.1962%	(1)

(a) English letter frequency distribution



(b) Probability distribution Graph

Figure 2.1: Reference : https://www3.nd.edu/ busiforc/handouts/cryptography/letterfrequencies.html

The friedman's test is followed by frequency analysis. Knowing the possible key values, we divide the text into those many columns and break the shift ciphers to generate the key.

2.3 Good and Bad Ciphers

The ultimate Goal in cyber security is to derive a cipher (encryption algorithm) that cannot be broken as we demand 100% accuracy in our algorithm.
Requirement : the adversary who observes the cipher text can't find out what it is (plaintext).

- It is required that not even a single bit info should be leaked. (all bits protected)
  Even if we have multiple messages we should not be able to modify / identify one, based on the information acquired from the other.

  $c_1 = Enc_k(m_1)$
  $c_2 = Enc_k(m_2)$

  Bad cipher :

  1. If c_1, c_2, m_1 are known then m_2 can be figured as the algo/key being used is the same in both the encryption. (CFA attack)
  2. Using the same algorithms for multiple users makes it very easy for them to gain information of others which is against the aim.
- Same key should not be used multiple times (make it completely unusable for the next communication)
- Also plain texts should not be comparable based on cipher texts, as some information on the bounds of our trail (eg $m_2 > m_1$) is being revealed making it not completely secure.

Design Scheme for a good (secure) cipher ::: one time pad (next class)