Private Key Encryption - Introduction

Q1: A Private Key Encryption scheme can be defined using three probabilistic polynomial time algorithms and a message space.

a Explain input, output, and the purpose of each algorithm (Key Generation, Encryption, Decryption).

b What are the key space, the message space, and the ciphertext space?

c Formally define the correctness requirement of an encryption scheme.

A1:

a We have 3 functions in our scheme: Key Generation, Encryption and Decryption, to which I will refer to as *Gen*, *Enc* and *Dec* for brevity.

Gen function takes a series of bits as a parameter (mostly shown as 1^λ for lambda bits) and simply returns a key from the uniformly distributed keys of the key space. The range of this function defines the key space K. *Enc* function takes two parameters, a key k from K and a message m from the message space M. $Enc(k, m)$ returns a ciphertext c, as $c := Enc(k, m)$. The range of *Enc* function also defines the ciphertext space C. *Dec* function takes two parameters, a key k from K and a ciphertext c from C, as $m := Dec(k, c)$. The result is expected to be a message m from M.

b As explained above, key space is the set of all possible keys generated by our key generation function. Message space is the set of all possible messages that *Enc* might take. Ciphertext space is set of all possible ciphertexts generated by *Enc*. In a scheme, just defining the message space would suffice, as the other two are defined by the functions and the message space.

c The formal definition of the correctness of an encryption scheme is that: $Dec(k, Enc(k, m)) = m$. In plain English: using the same key, encrypting a message and decrypting the obtained ciphertext should yield that same message.

Q2: Think about how to define security. Write down your proposed definition. Explain how your definition complies with the modern cryptography principles?

A2: My definition for security is as follows: For any communication of parties defined by a certain group, separated in space or time; a transmission of message among these parties should be conducted in a way such that an adversary with access to the medium and does not initially belong to this group should not obtain any information regarding the message being transmitted.

Here, the group can be a single person separated in time (similar to the example in Fig. 1.2 in the book) , or multiple people separated in space, or even multiple devices of a single person. The adversary does not initially belong to this group. The keyword initial is used because the adversary can join to the group somehow, or open multiple accounts, thereby increasing it's attack power from a mere ciphertext-only attack to chosen plaintext attack, or even chosen ciphertext attack.

In relation to the principles, I believe this is fairly formal but lacks the mathematical rigor in it's definitions, which would require more time and thought to write. Furthermore, it might be problematic for the 3rd principle in particular, as it may be hard to prove formally given that the definitions are not mathematically formulated.

Q3: Briefly explain each of the following treat models for an encryption scheme.

a Ciphertext-only attack
b Known-plaintext attack
c Chosen-plaintext attack
d Chosen-ciphertext attack

A3:

a In a ciphertext-only attack, the adversary just has the access to the ciphertext (usually one, but can be more than one) which was obtained from eavesdropping on the transmission.

b In a known-plaintext attack, the adversary obtains information regarding one or more plaintext-ciphertext pairs. An example is for example the observation that two parties say hello to each other every morning, which the adversary may guess.

c Chosen-plaintext attack means that the adversary can obtain the ciphertext of a message of it's choice. As a result, they can obtain plaintext-ciphertext options per their choice.

d In a chosen-ciphertext attack, the attacker can also learn information regarding the decrypted ciphertexts, i.e.: the plaintext corresponding to a ciphertext of adversaries choice. From a to d, the power of the adversary increases.

Q4: For the following historical ciphers, define their key generation, encryption, and decryption functions, and their key, message, and ciphertext spaces. Discuss their security. Show how they can be easily broken using a chosen plaintext attack.

a Caesar cipher
b Shift ciphers
c Substitution ciphers

A4: Both a and b are actually a type of substitution cipher. Shift cipher is a form of mono-alphabetic substitution cipher, where each character is mapped to another character, and the mapping stays constant throughout the scheme. In shift cipher, every character corresponds to the k'th character starting from it. This k is bounded by the set of characters, (e.g. 26 for English alphabet) and the boundaries are circular, i.e. after z you wrap back to a. The Caesar cipher is actually an instance of shift cipher, where $k = 3$.

All these can be broken per se, but with a chosen plaintext attack the adversary can very easily break the scheme. For all three cases, the adversary can basically choose the plaintext as the characters of the set of characters (i.e. letters of the alphabet).

For example: 'abcdefghijklmnopqrstuvwxyz' would yield how every English letter is mapped, thereby rendering this scheme useless for English messages!

Q5: What is entropy? Min-entropy? Mutual information? Define these formally and also discuss. Can one talk about entropy of some number or outcome?

A5: Entropy can be though of a quantification of order (or disorder) of a system. In physics, this is measure as the thermal energy (e.g. cold = low thermal energy, low entropy | hot = high thermal energy, high entropy), however from a more mathematical perspective and more related to our domain would be the information theoretical aspect of entropy, which we first attribute to Claude Shannon.

Given a stream of characters (or symbols, letters, etc.), how can we guess a character from that stream, asking as little questions as we can? In the video material given to us, we were guessing a letter from a group of letters (e.g. all of them are same (least entropy), or all of them are different letters (highest entropy)) , asking as little as possible 'yes/no' questions. The average number over many experiments yielded the entropy.

As another answer, suppose you have a random variable X with outcomes $x_1, x_2, ..., x_n$ with probabilities $p(x_1), p(x_2), ..., p(x_n)$.

- Entropy is H as defined by:
$$H(X) = -\sum_{i=1}^{n} p(x_i) log(p(x_i))$$
This $H(x)$ is a measure of "randomness".
- Min-entropy is when $\Pr[X = x] < 2^{-n}$ for any outcome x.
- Mutual Information I is defined for two random variables X and Y:
$$I(X,Y) = \sum_{x \in X, y \in Y} P_{(X,Y)}(x,y) log \left(\frac{P_{(X,Y)}(x,y)}{P_X(x) P_Y(y)} \right)$$
Note here that we considered these as discrete random variables, and used PMF (Probability Mass Function) of X and Y as $P_{(X,Y)}$ and the marginal probability mass functions of X and Y as P_X and P_Y. $P_X(x) = Pr[X = x]$ and $P_{(X,Y)}(x,y) = Pr[X = x \text{ and } Y = y]$.

Q6: If we choose q elements $y_1, \dots, y_q$ uniformly at random from a set of size N , what is the probability that there exist distinct $i, j (i \neq j)$ with $y_i = y_j$? Explain your answer.

A6: The question formally asks $Pr[\exists i \neq j : y_i = y_j]$, which is equal to $Pr[\text{there is a duplicate}]$. If $q > N$, the answer is 1, due to the pigeonhole principle. If $q \leq N$ we can calculate the probability, using the beginning steps done in the proof of the Birthday Paradox. First, we negate the statement:

$$Pr[\exists i \neq j : y_i = y_j] = 1 - Pr[\forall i \neq j : y_i = y_j]$$

Next, we unroll the probabilty statement:

$$Pr[\exists i \neq j : y_i = y_j] = 1 - 1 \times \left(\frac{N-1}{N} \right) \left(\frac{N-2}{N} \right) \dots \left(\frac{N-(q-1)}{N} \right)$$

Here, the first term is the probability y_1 will not collide with anything yet. Since y_1 is the first, it will not collide with anything for sure. Then, we ask the same question for y_2. At this point, y_1 is one of the q possible values, therefore y_2 can only be one of the remaining $N-1$, so and so forth until we get to y_q which can have values from the remaining $N-(q-1)$ values. We can write the final answer using the product symbol on this equation:

$$Pr[\exists i \neq j : y_i = y_j] = 1 - \prod_{i=1}^{q-1} \left(1 - \frac{i}{N} \right)$$

1. Exercises

E1[1]: Show that the shift, substitution, and Vigenere ciphers are all trivial to break using a chosen-plaintext attack. How much chosen plaintext is needed to recover the key for each of the ciphers?

A1: A plaintext that is basically the alphabet itself would be enough to break all of them. To find the key, just one letter would suffice for shift ciphers.

E2[2]: The shift, substituion and Vigenere ciphers can also be defined over the 128-character ASCII alphabet rather than the 26-character English alphabet.

a Provide a formal definition of each of these schemes in this case.

b Discuss how the attacks we have shown in this chapter can be modified to break each of these modified schemes.

A2: For the Vigenere cipher:

a *KeyGen* chooses a period $t \overset{R}{\leftarrow} \{1, \ldots, t_{\max}\}$ for some known maximum period. Then, for $i \in \{0, \ldots, t-1\}$ a key is uniformly chosen $k_i \in \{1, \ldots, 127\}$. The encryption of a plaintext $m = m_1, \ldots, m_l$ (l is length) is $c = c_1, \ldots, c_l$ where $c_i = [[m_i + k_i \bmod t] \bmod 128]$. Decryption is done as $m_i = [[c_i - k_i \bmod t] \bmod 128]$.

b The frequency attacks are still possible, but now instead of English alphabet the ASCII table should be taken into consideration.

[1] KL Ed.2 Ch.1 Exc. 1.5

[2] KL Ed.2 Ch.1 Exc. 1.8