Lecture Notes 2

8 Compromising a cryptosystem

We've already mentioned one kind of compromise of a cryptosystem—learning the secret message m. But actually there are many different degrees of compromise, some more serious than others, depending on the application. Here is a list of some of the kinds of compromise which one should be aware of and ideally be able to prevent:

Complete break Eve finds the key k. Eve now has the same knowledge as both Alice and Bob, so she can impersonate Alice when talking to Bob and impersonate Bob when talking to Alice. She can also passively read all communications between Alice and Bob.

Complete message recovery Eve can recover any message m even without knowing k. She can thus passively read all communications between Alice and Bob, but she can't impersonate one to the other since she is unable to produce encrypted messages herself.

Selected message recovery Eve is only able to recover certain messages from their ciphertexts. Namely, there is a set $M \subseteq \mathcal{M}$ of messages which Eve is able to read; others remain secure. The larger the set M is, the more serious this kind of compromise.

Uncertain message recovery Eve is able to recover m from c only when certain keys $k \in K \subseteq \mathcal{K}$ are in use. The larger the set K is, the more serious this kind of compromise. Because the key is generally chosen randomly and uniformly from $\mathcal{K}$, Eve has probably $|K|/|\mathcal{K}|$ of compromising Alice's messages.

Partial message recovery In each of the above kinds of compromise, Eve either succeeds or not in recovering m. In partial message recovery, she obtains partial information about m. Initially, Eve knows only the probability distribution from which m is drawn. After seeing the ciphertext $c = E_k(m)$, Eve may be able to determine that m lies in a subset $M \subseteq \mathcal{M}$. If $|M| = 1$, then Eve has narrowed down the possibilities for m to a single message, so she has learned m. But if $|M| \geq 2$, she has obtained only partial information about M. Whether or not that partial information is useful to Eve depends on what partial information she receives and of course on the application.

Perfect secrecy (information-theoretic security) This is the other extreme in which Eve learns nothing at all about m from seeing the ciphertext c. Whatever she knew about m before seeing c she still knows, but she learns nothing new from seeing c. (Note that if Eve already knew m beforehand, then she would learn nothing new about m from seeing c, so in that special case, all cryptosystems are totally secure.)

9 Probabilistic model

When we start talking about the probability of Even learning the message, we must ask, "Where does randomness enter into the discussion of attacks on a cryptosystem?" We assume there are three such places:

1. The message itself is drawn at random from some probability distribution over the message space $\mathcal{M}$. That distribution is not necessarily uniform or even anything close to uniform[1]. However, whatever the distribution is, we assume that it is part of Eve's *a priori* knowledge.

2. The secret key is chosen uniformly from the key space $\mathcal{K}$.

3. The attacker (Eve) has access to a source of randomness which she may use while attempting to break the system. For example, Eve can choose an element $k' \in \mathcal{K}$ at random. With probability $p = 1/|\mathcal{K}|$, her element k' is actually the correct key k.

We further assume that these three sources of randomness are *statistically independent*. This means that Eve's random numbers do not depend on (nor give any information about) the message or key used by Alice. It also means that Alice's key does not depend on the particular message or vice versa.

These multiple sources of randomness give rise to a *joint probability distribution* that assigns a well-defined probability (i.e., a real number in the interval $[0, 1]$) to each triple (m, k, z), where m is a message, k a key, and z is the result of the random choices Eve makes during her computation. The independence requirement simply says that the probability of the triple (m, k, z) is precisely the product of $p_m \times p_k \times p_z$, where p_m is the probability that m is the chosen message, p_k is the probability that k is the chosen key, and p_z is the probability that z represents the random choices made by Eve during the course of her computation.

When talking about random choices, we often use the term *random variable* to denote the experiment of choosing an element from a set according to a particular probability distribution. For example, we sometimes use m to denote the random variable that describes the experiment of Alice choosing a message $m \in \mathcal{M}$ according to the assumed message distribution. Ambiguously, we often use m to denote a particular element in the set $\mathcal{M}$, perhaps the outcome of the experiment described by the random variable m. While these two uses may sometimes be confusing, we will try to make which usage is intended clear from context.

For example, if we let p_m denote the probability that the message distribution assigns to the particular message $m \in \mathcal{M}$, then there is no experiment under consideration and it is clear that m is being used to denote an element of $\mathcal{M}$. On the other hand, if we talk about the event $m = m_0$, then we are using m in the sense of a random variable, and the event $m = m_0$ means that the outcome of the experiment described by m is m_0. We denote the probability of that event occurring by $\mathrm{prob}[m = m_0]$, which in this case is simply p_{m_0}. See section 2.2 of Stinson for a more thorough review of basic probability definitions.

10 Perfect secrecy

Perfect secrecy is defined formally in terms of statistical independence. In our probabilistic model, we can consider both m and k to be random variables. Together, they induce a probability distribution on the ciphertext $c = E(k, m)$. Here we see the power of the random variable notation, for c itself can be considered to be a random variable which is a function of k and m. To define c formally, we must define the probability of the events $c = c_0$ for each $c_0 \in \mathcal{C}$. These are easily derived from the random variables k and m as follows:

$$\mathrm{prob}[c = c_0] = \sum_{k_0, m_0 : c_0 = E(k_0, m_0)} \mathrm{prob}[k = k_0 \wedge m = m_0].$$

[1]The *uniform* distribution over a finite set Ω assigns probability $1/|\Omega|$ to each point $x \in \Omega$. In other words, each element $x \in \Omega$ has equal probability of being chosen.

Because of the assumed independence of k and m, this is the same as

$$\text{prob}[c = c_0] = \sum_{k_0, m_0 : c_0 = E(k_0, m_0)} \text{prob}[k = k_0] \times \text{prob}[m = m_0].$$

The reason for the summation is that the same ciphertext c_0 can arise from many different key-message pairs.

Now, what it means for the cryptosystem to have perfect secrecy (be information-theoretically secure) is that the random variables c and m are statistically independent, that is,

$$\text{prob}[m = m_0 \wedge c = c_0] = \text{prob}[m = m_0] \times \text{prob}[c = c_0]$$

for all $m_0 \in \mathcal{M}$ and $c_0 \in \mathcal{C}$.

This independence is often expressed more intuitively in terms of conditional probabilities. We define

$$\text{prob}[m = m_0 \mid c = c_0] = \frac{\text{prob}[m = m_0 \wedge c = c_0]}{\text{prob}[c = c_0]}$$

assuming the denominator of the fraction is non-zero. This is the probability that $m = m_0$ given that $c = c_0$. We then say that c and m are statistically independent if

$$\text{prob}[m = m_0 \mid c = c_0] = \text{prob}[m = m_0]$$

for all $m_0 \in \mathcal{M}$ and $c_0 \in \mathcal{C}$ such that $\text{prob}[c = c_0] \neq 0$. Hence, even after Eve receives the ciphertext c_0, her opinion of the likelihood of each message m_0 is the same as it was initially, so she has learned nothing about m_0.

11 Caesar cipher

To make these ideas more concrete, we look at the Caesar cipher, which is said to go back to Roman times. It is an alphabetic cipher, that is, used to encode the 26 letters of the Roman alphabet $A, B, \ldots, Z$. For convenience, we will represent the letters by the numbers $0, 1, \ldots, 25$, where $A = 0$, $B = 1$, …, $Z = 25$.

11.1 Encrypting single letters

The base cipher handles just single letters. The message space is $\mathcal{M} = \{0, \ldots, 25\}$, and the ciphertext space and key space are the same, so $\mathcal{M} = \mathcal{C} = \mathcal{K}$.

To encrypt, E_k replaces each letter m by the letter k places further on in the alphabet when read circularly, with A following Z. To decrypt, D_k replaces each letter c by the letter k places earlier in the alphabet when read circularly. Mathematically, the encryption and decryption functions are given by

$$E_k(m) = (m + k) \bmod 26$$

$$D_k(c) = (c - k) \bmod 26.$$

(Notation: $x \bmod 26$ is the remainder of x divided by 26.) Thus, if $k = 3$, then $E_k(P) = S$ and $D_k(S) = P$.

11.2 Encrypting strings

Of course, one is generally interested in encrypting more than single letter messages, so we need a way to extend the encryption and decryption functions to strings of letters. The simplest way of extending is to encrypt each letter separately using the same key k. If $m_1 \dots m_r$ is an r-letter message, then its encryption is $c_1 \dots c_r$, where $c_i = E_k(m_i) = (m_i + k) \bmod 26$ for each $i = 1, \dots, r$. Decryption works similarly, letter by letter.

What we have really done by the above is to build a new cryptosystem from the base cryptosystem on single letters. In the new system, the message space and ciphertext space are

$$\mathcal{M}^r = \mathcal{C}^r = \underbrace{\mathcal{M} \times \mathcal{M} \times \dots \mathcal{M}}_{r},$$

that is, length-r sequences of numbers from $\{0, \dots, 25\}$. The encryption and decryption functions are

$$E_k^r(m_1 \dots m_r) = E_k(m_1) \dots E_k(m_r)$$
$$D_k^r(c_1 \dots c_r) = D_k(c_1) \dots D_k(c_r).$$

Here, we use the superscript r to distinguish these functions from the base functions $E_k()$ and $D_k()$.

12 Probabilistic analysis of a simplified Caesar cipher

To make the ideas from Section 11 concrete, we work through a probabilistic analysis of a simplified version of the Caesar cipher restricted to a 3-letter alphabet, so $\mathcal{M} = \mathcal{C} = \mathcal{K} = \{0, 1, 2\}$, $E_k(m) = (m + k) \bmod 3$, and $D_k(m) = (m - k) \bmod 3$.

Assume for starters that the message distribution is given in Figure 12.1. Because the keys

m	p_m
0	1/2
1	1/3
2	1/6

Figure 12.1: A priori message probabilities.

are always chosen uniformly, each key has probability 1/3. Hence, we obtain the joint probability distribution shown in Figure 12.2.

m \ k	0	1	2
0	1/6	1/6	1/6
1	1/9	1/9	1/9
2	1/18	1/18	1/18

Figure 12.2: Joint probability distribution.

Next, we calculate the conditional probability distribution $\text{prob}[m = 1 \mid c = 2]$. To do this, we consider the pairs (m, k) in our joint sample space and compute $c = E_k(m)$ for each. Figure 12.3

shows the result, where each point is labeled by a triple (m, k, c), and those points for which $c = 2$ are shown in bold.

$$\begin{array}{ccc}
(0,0,0) & (0,1,1) & (\mathbf{0},\mathbf{2},\mathbf{2}) \\
\cdot & \cdot & \bullet \\
(1,0,1) & (\mathbf{1},\mathbf{1},\mathbf{2}) & (1,2,0) \\
\cdot & \bullet & \cdot \\
(\mathbf{2},\mathbf{0},\mathbf{2}) & (2,1,0) & (2,2,1) \\
\bullet & \cdot & \cdot
\end{array}$$

Figure 12.3: Sample space.

The probability that $c = 2$ is the sum of the probabilities that correspond to the bold face points, i.e., $\mathrm{prob}[c = 2] = 1/6 + 1/9 + 1/18 = 6/18 = 1/3$. The only one of these points for which $m = 1$ is $(1, 1, 2)$, and it's probability is $1/9$, so $\mathrm{prob}[m = 1 \wedge c = 2] = 1/9$. Hence,

$$\begin{aligned}
\mathrm{prob}[m = 1 \mid c = 2] &= \frac{\mathrm{prob}[m = 1 \wedge c = 2]}{\mathrm{prob}[c = 2]} \\
&= \frac{1/9}{1/3} = \frac{1}{3}.
\end{aligned}$$

This is the same as the initial probability that $m = 1$. Repeating this computation for each possible message m_0 and ciphertext c_0, one concludes that

$$\mathrm{prob}[m = m_0 \mid c = c_0] = \mathrm{prob}[m = m_0].$$

Hence, our simplified Caesar cipher is information-theoretically secure.

Now let's make a minor change in the cipher and see what happens. The only change is that we reduce the key space to $\mathcal{K} = \{0, 1\}$. The a priori message distribution is still given by Figure 12.1, but the joint probability distribution changes as does the sample space as shown in Figure 12.4.

$$\begin{array}{cc|cc}
 & & \multicolumn{2}{c}{\overbrace{\hphantom{1/12 \quad 1/12}}^{k}} \\
 & & 0 & 1 \\
\hline
 & 0 & 1/4 & 1/4 \\
m & 1 & 1/6 & 1/6 \\
 & 2 & 1/12 & 1/12
\end{array}
\qquad\qquad
\begin{array}{cc}
(0,0,0) & (0,1,1) \\
\cdot & \cdot \\
(1,0,1) & (\mathbf{1},\mathbf{1},\mathbf{2}) \\
\cdot & \bullet \\
(\mathbf{2},\mathbf{0},\mathbf{2}) & (2,1,0) \\
\bullet & \cdot
\end{array}$$

Figure 12.4: Joint probability distribution and sample space when $\mathcal{K} = \{0, 1\}$.

Now, $\mathrm{prob}[c = 2] = 1/6 + 1/12 = 3/12 = 1/4$, and $\mathrm{prob}[m = 1 \wedge c = 2] = 1/6$. Hence,

$$\mathrm{prob}[m = 1 \mid c = 2] = \frac{1/6}{1/4} = \frac{2}{3}.$$

While the a priori probability that $m = 1$ is still the same as it was before—1/3, the conditional probability given $c = 2$ is now double what it was. Indeed, there are now only two possibilities for m once Eve sees $c = 2$. Either $m = 1$ (and $k = 1$) or $m = 2$ (and $k = 0$). It is no longer possible that $m = 0$ given that $c = 2$. Hence, Eve narrows the possibilities for m to the set $\{1, 2\}$, and her probabilistic knowledge of m changes from the initial distribution $(1/2, 1/3, 1/6)$ to the new distribution $(2/3, 1/3)$. She has learned at lot about m indeed, even without finding it exactly!