Notes for Lecture 7

Wed, 1/28/2026

Historical example: substitution cipher

Example 42. (substitution cipher) In a substitution cipher, the key k is some permutation of the letters $A, B, \ldots, Z$. For instance, $k = FRA\ldots$. Then we encrypt $A \rightarrow F$, $B \rightarrow R$, $C \rightarrow A$ and so on. How large is the key space?

Solution. Key space has size $26! \approx 10^{26.6} \approx 2^{88.4}$, so a key can be stored using 89 bits. That's actually a fairly large key space (for instance, DES has a key size of 56 bits only). Too large to go through by brute force.

However, still easy to break. Since each letter is always replaced with the same letter, this cipher is susceptible to a **frequency attack**, exploiting that certain letters (and, more generally, letter combinations!) occur much more frequently in, say, English text than others. For instance, Lewand's book on Cryptology lists the following frequencies:

E: 12.7%, T: 9.1%, A: 8.2%, O: 7.5%, I: 7%, N: 6.7%, S: 6.3%, H: 6.1%, R: 6%, D: 4.3%, L: 4%, C: 2.8%, ...

The rarest letters are Q and Z with a frequency of about 0.1% only. (The exact frequencies and precise ordering various between different sources and the body of text that the frequencies were obtained from.)

The most common letter pairs (digrams) are TH HE AN RE ER IN ON AT ND ST ES EN OF TE ED OR TI HI AS TO.

More information at: https://en.wikipedia.org/wiki/Letter_frequency

Comment. Note that the frequencies and even the ranking depend considerably on the source of text. For instance, using government telegrams, a military resource lists EN followed by RE, ER as the most frequent digrams. That same manual suggests SENORITA as a mnemonic to remember the most frequent letters.

http://www.umich.edu/~umich/fm-34-40-2/ (Field Manual 34-40-2, Department of the Army, 1990)

Example 43. It seems convenient to add the space as a 27th letter in the historic encryption schemes. Can you think of a reason against doing that?

Solution. In most texts, the space occurs more frequently and more regularly than any other letter. Adding it to the encryption schemes would make them even more susceptible to attacks.

Example 44. (bonus challenge!) You intercept the following message from Alice:

WHCUHFWXOWHUQXOMOMQVSQWAMWHCUHFXOLNWXQMQVSQWAWMQLN

Your experience tells you that Alice is using a substitution cipher. You also know that this message contains the word "secret". Can you crack it?

Note. In modern practice, it is not uncommon to know (or suspect) what a certain part of the message should be. For instance, PDF files start with "%PDF" (0x25504446).

See https://en.wikipedia.org/wiki/Magic_number_(programming) for more such instances.

(To collect a bonus point, send me an email within the next week with the plaintext and how you found it.)

One-time pad (continued)

Example 45. A ciphertext only attack on the one-time pad is entirely hopeless. Explain why!

Solution. The attacker only knows $c = m \oplus k$. The attacker is unable to get any information on m, because every other message m' (of the right length) could have resulted in the same ciphertext c.

Indeed, the key $k' = m' \oplus c$ encrypts m' to c as well (because $m' \oplus k' = m' \oplus (m' \oplus c) = c$). Moreover, every plaintext m' is equally likely because it corresponds to a unique key.

The next example highlights the importance of only using the key once.

Example 46. (attack on the two-time pad) Alice made a mistake and encrypted the two plaintexts m_1, m_2 using the same key k. How can Eve exploit that?

Solution. Eve knows the two ciphertexts $c_1 = m_1 \oplus k$ and $c_2 = m_2 \oplus k$.

Hence, she can compute $c_1 \oplus c_2 = (m_1 \oplus k) \oplus (m_2 \oplus k) = m_1 \oplus m_2$.

This means that Eve knows $m_1 \oplus m_2$, which is information about the original plaintexts (no key involved!). That's a cryptographic disaster: Eve should never be able to learn *anything* about the plaintexts.

In fact. If the plaintexts are, say, English text encoded using ASCII then Eve very possibly can (almost) reconstruct both m_1 and m_2 from $m_1 \oplus m_2$. The reason for that is that the messages are expressed in ASCII, which means 8 bits per character of text. However, the **entropy** (a measure for the amount of information in a message) of (longer) typical English text is frequently below 2 bits per character.

Some details and beautiful graphical illustrations are given at:

http://crypto.stackexchange.com/questions/59/taking-advantage-of-one-time-pad-key-reuse

We saw in Example 45 that ciphertext only attacks on the one-time pad are entirely hopeless. What about other attacks?

Attacks like known plaintext or chosen plaintext don't apply if the key is only to be used once.

Yet, the one-time pad by itself provides **little protection of integrity**. The next example shows how tampering is possible without knowledge about the key.

Example 47. Alice sends an email to Bob using a one-time pad. Eve knows that and concludes that, per email standard, the plaintext must begin with `To: Bob`. Eve wants to tamper with the message and change it to `To: Boo`, for a light scare.

- Eve wants to change the 7th letter of the plain text m from b to o.
- Since b is $0x62$ and o is $0x6F$, we have $b \oplus o = 0x0D$. Hence, $b \oplus 0x0D = o$.
- Therefore, if $e = 0x\underbrace{000000000000}_{\text{6 characters}}0D00\ldots$, then $\underbrace{\text{“TO: Bob...”}}_{m} \oplus e = \underbrace{\text{“TO: Boo...”}}_{m'}$.
- Alice sends $c = m \oplus k$. If Eve changes the ciphertext c to $c' = c \oplus e$, then Bob receives c' and decrypts it to $c' \oplus k = \overbrace{\underbrace{m \oplus k}_{=c} \oplus e}^{c'} \oplus k = m \oplus e = m'$, which is what Eve intended.

Using the one-time pad presents several challenges, including:

- keys must not be reused (see Example 46)
- while perfectly protecting against eavesdropping (if done correctly), the one-time pad is not secure against tampering (see Example 47)
- key distribution and management

  Alice and Bob have to somehow exchange huge amounts of keys, so that, at a later time, they are able to communicate securely.
- for perfect confidentiality, the key must be perfectly random

  But how can we produce huge amounts of random bits?

  Especially, how to teach a deterministic machine like a computer to do that? Think about it! This is much more challenging that it may seem at first...

These issues make one-time pads difficult to use in practice.

Historic comment. During the Cold War, the "hot line" between Washington and Moscow apparently used one-time pads for secure communication.

Example 48. One thing that makes the one-time pad difficult to use is that the key needs to be the same length as the plaintext. What if we have a shorter key and just repeat it until it has the length we need?

That's essentially the Vigenere cipher (in a different alphabet).

Solution. Assuming the attacker knows the length of our key (if she doesn't she can just try all possibilities), this is equivalent to using the one-time pad several times with the same key. That should never be done! Even using a key twice means that we become susceptible to a ciphertext only attack (see Example 46).

So, repeating the key is a terrible idea. However, the idea to create a longer (random) key out of a shorter (random) key is good (we will discuss pseudorandom generators next).

Let us emphasize that, in order to be perfectly confidential, the key for a one-time pad must be chosen completely at random (otherwise, an attacker can make assumptions on the used keys).

Indeed, the need to generate random numbers shows in every modern cipher.

Stream ciphers

Once we have a way to generate **pseudorandom numbers**, we can use the idea of the one-time pad to create a **stream cipher**.

Start with key of moderate size (say, 128 bits).

Use the key k and a PRG (**pseudorandom generator**) to generate a much longer **pseudorandom keystream** $\mathrm{PRG}(k)$. Then encrypt $E_k(m) = m \oplus \mathrm{PRG}(k)$.

We lost perfect confidentiality. Security relies on choice of PRG (must be unpredictable).

As with the one-time pad, we must never reuse the same keystream! That does not mean that we cannot reuse the key: we can do that using a **nonce**: $E_k(m) = m \oplus \mathrm{PRG}((\mathrm{nonce}, k))$, where the seed is produced by combining the nonce and k (for instance, just concatenating them).

The nonce is then passed (unencrypted) along with the message.

To make sure that we never reuse the same keystream, we must never use the same nonce with the same key.

Remark. A nonce can only be used once, as is in its name. Apparently, according to Urban Dictionary, it is also common as a British insult, roughly equivalent to wanker.

How to generate random numbers?

Natural randomness is surprisingly difficult to harness.

You can for instance play around with a Geiger counter but our department is short on these and getting lots of random numbers is again challenging.

Linear congruential generators

(linear congruential generator) Let a, b, m be chosen parameters.
From the seed x_0, we produce the sequence $x_{n+1} \equiv a x_n + b \pmod{m}$.

The choice of a, b, m is crucial for this to generate acceptable pseudorandom numbers.

For instance, glibc uses $a = 1103515245$, $b = 12345$, $m = 2^{31}$. (This is one of two implementations.) In that case, each x_i is represented by precisely 31 bits. [Note that the choice of m makes this very fast.]

https://en.wikipedia.org/wiki/Linear_congruential_generator

Linear congruential generators (LCG) are easy to predict and must not be used for cryptographic purposes. More generally, all polynomial generators are cryptographically insecure. They are still used in practice, because they are fast and easy to implement and have decent statistical properties. (For instance, our online homework is generated using random numbers, and there is no need for crypto-level security there.)

Statistical trouble. Can you see why the sequences produced by the glibc LCG alternate between even and odd numbers? (Similarly, other low bits are much less "random" than the higher bits.) Because of this defect, some programs (and other implementations of `rand()` based on LCGs) throw away the low bits entirely.

Comment. The particular choices of a and b in glibc are somewhat mysterious. See, for instance:

https://stackoverflow.com/questions/8569113/why-1103515245-is-used-in-rand

Example 49. Generate values using the linear congruential generator $x_{n+1} \equiv 5x_n + 3 \pmod{8}$, starting with the seed $x_0 = 6$.

Solution. $x_1 \equiv 1$, $x_2 \equiv 0$, $x_3 \equiv 3$, $x_4 \equiv 2$, $x_5 \equiv 5$, $x_6 \equiv 4$, $x_7 \equiv 7$, $x_8 \equiv 6$. This is the value x_0 again, so the sequence will now repeat. Note that we went through all 8 residues before repeating. Period 8.

Note. Because $8 = 2^3$ we can represent each x_i using exactly 3 bits. Then $x_1, x_2, x_3, \ldots = 1, 0, 3, \ldots$ corresponds to the bit stream $(001\ 000\ 011\ \ldots)_2$.

Example 50. (extra) Observe that the sequence produced by the linear congruential generator $x_{n+1} \equiv a x_n + b \pmod{m}$ must repeat, at the latest, after m terms. (Why?!)

One can give precise conditions on a, b, m to achieve a full period m. Namely, this happens if and only if $\gcd(b, m) = 1$ and $a - 1$ is divisible by all primes (as well as 4) dividing m.

(a) Generate values using a linear congruential generator $x_{n+1} \equiv 2x_n + 1 \pmod{10}$, starting with the seed $x_0 = 5$. When do they repeat? Is that consistent with the mentioned condition?

(b) What are possible values for a so that the LCG $x_{n+1} \equiv a x_n + 11 \pmod{100}$ has period 100?

(c) glibc uses $a = 1103515245$, $b = 12345$, $m = 2^{31}$. After how many terms will the sequence repeat?

Solution.

(a) $x_1 \equiv 1$, $x_2 \equiv 3$, $x_3 \equiv 7$, $x_4 \equiv 5$. This is the value x_0 again, so the sequence will repeat. Period 4.
[The period is less than 10. This is as predicted by the mentioned condition, because $a - 1$ is not divisible by 2 and 5.]

(b) We need that $a - 1$ is divisible by 4 and 5. Equivalently, $a \equiv 1 \pmod{20}$. Hence, possible values are $a = 1, 21, 41, 61, 81$.

(c) Clearly, $\gcd(b, m) = 1$. Also, $a - 1$ is divisible by 4 (and no primes other than 2 divide m). Hence, for every seed, values repeat only after going through all 2^{31} residues.

Example 51. Let's use the PRG $x_{n+1} \equiv 5x_n + 3 \pmod{8}$ as a stream cipher with the key $k = 4 = (100)_2$. The key is used as the seed x_0 and the keystream is $\text{PRG}(k) = x_1 x_2 \ldots$ (where each x_i is 3 bits). Encrypt the message $m = (101\ 111\ 001)_2$.

Solution. We first use the PRG with seed $x_0 = k$ to produce the keystream $\text{PRG}(k) = 7,\ 6,\ 1,\ \ldots = (111\ 110\ 001\ \ldots)_2$.

We then encrypt and get $c = E_k(m) = m \oplus \text{PRG}(k) = (101\ 111\ 001)_2 \oplus (111\ 110\ 001)_2 = (010\ 001\ 000)_2$.

Decryption. Observe that decryption works in the exact same way:
$D_k(c) = c \oplus \text{PRG}(k) = (010\ 001\ 000)_2 \oplus (111\ 110\ 001)_2 = (101\ 111\ 001)_2$.

Note. The keystream continues as $\text{PRG}(k) = 7, 6, 1, 0, 3, 2, 5, 4, \ldots$ At this point it repeats itself because we obtained the value 4, which was our seed. Since the state of this PRG only depends on the value of x_n, and there are 8 possible values for x_n, the period 8 is the longest possible. The previous (extra) example gave conditions on the PRG that guarantee that the period is as long as possible.

Example 52. Can you think of a way in which the numbers produced by a linear congruential generator differ from truly random ones?

Solution. An easy observation for our small examples is the following: by construction, $x_{n+1} \equiv ax_n + b \pmod{m}$, individual values don't repeat unless a full period is reached and everything repeats. Truly random numbers do repeat every now and then (however, if m is large, then this observation is not exactly practical).

Of course, knowing the parameters $a,\ b,\ m$, the numbers generated by the PRG are terribly **predictable**. Knowing just one number, we can produce all the next ones (as well as the ones before). A PRG that is safe for cryptographic purposes should not be predictable like that! (See next example.)

The next example illustrates the vulnerability of stream ciphers, based on predictable PRGs.

Recall that it is common to know or guess pieces of plaintexts; for instance, every PDF begins with $\%\text{PDF}$.

Example 53. Eve intercepts the ciphertext $c = (111\ 111\ 111)_2$. It is known that a stream cipher with PRG $x_{n+1} \equiv 5x_n + 3 \pmod{8}$ was used for encryption. Eve also knows that the plaintext begins with $m = (110\ 1\ldots)_2$. Help her crack the ciphertext!

Solution. Since $c = m \oplus \text{PRG}$, we learn that the initial piece of the keystream is $\text{PRG} = m \oplus c = (110\ 1\ldots)_2 \oplus (111\ 1\ldots)_2 = (001\ 0\ldots)_2$. Since each x_n is 3 bits, we conclude that $x_1 = (001)_2 = 1$.

Because the PRG is predictable, we can now recreate the entire keystream! Using $x_{n+1} \equiv 5x_n + 3 \pmod{8}$, we find $x_2 \equiv 0$, $x_3 \equiv 3$, ... In other words, $\text{PRG} = 1, 0, 3, \ldots = (001\ 000\ 011\ \ldots)_2$.

Hence, Eve can decrypt the ciphertext and obtain $m = c \oplus \text{PRG} = (111\ 111\ 111)_2 \oplus (001\ 000\ 011)_2 = (110\ 111\ 100)_2$.