COMP 8920: Cryptography Winter 2025

Lecture 4 — Jan 16, 2025

Prof. Curtis Bright *Scribe: Aaron Barnoff*

1 Overview

This lecture covers material from chapter 2.2 of the text.

2 Cryptanalysis

Cryptanalysis refers to the strategies used by adversaries to break cryptosystems.

Kerckhoff's Principle states that an adversary knows the cryptosystem being used.

Attack Models There are four primary attack models:

1. **Ciphertext-only**: The adversary possesses a string of ciphertext, $\mathbf{y}$.
2. **Known plaintext**: The adversary possesses both a string of plaintext $\mathbf{x}$ and its corresponding ciphertext $\mathbf{y}$.
3. **Chosen plaintext**: The adversary can choose a plaintext string $\mathbf{x}$ and construct its corresponding ciphertext $\mathbf{y}$.
4. **Chosen ciphertext**: The adversary can choose a ciphertext string $\mathbf{y}$ and construct its corresponding plaintext $\mathbf{x}$.

3 Cryptanalysis of the Affine Cipher

The Affine Cipher cryptosystem is vulnerable to a **ciphertext-only** attack.

Some characters in the English language, such as E, T, A, occur more frequently than others like X, Q, Z.

- Certain bigrams and trigrams also appear more frequently, such as $ER, IN, AN, THE, ING, AND$.

This means that the frequency of characters in plaintext $\mathbf{x}$ is not evenly distributed.

- If $\mathbf{y}$ is long enough, an adversary can analyze the frequency of characters in the ciphertext and deduce information about the plaintext.

Example: Suppose R occurs most frequently in a given ciphertext and D second most:
We can guess that R decrypts to e and D to t, leading to:

$$e_k(4) = 17, \quad e_k(19) = 3$$

$$4a + b = 17, \quad 19a + b = 3$$

- Solving this system of equations in $\mathbb{Z}_{26}$ gives $(a, b) = (6, 19)$.
- However, since $\gcd(6, 26) = 2 > 1$, a does not have an inverse, making decryption impossible.

Next, if K is the next most frequent ciphertext character, we guess that K decrypts to t, leading to:

$$e_k(4) = 17, \quad e_k(19) = 10$$

- Solving this system gives $(a, b) = (3, 5)$, which is a valid key.
- Applying the decryption function with $k = (3, 5)$, we get:

$$d_k(y) = a^{-1}(y - b) = 9(y - 5) = 9y + 7$$

4 Cryptanalysis of the Substitution Cipher

The Substitution Cipher is more complex, as character frequencies may be more evenly distributed.

- Working with unigrams might be too difficult, so we analyze **bigrams** instead.

For example, suppose Z occurs most frequently, and we guess it decrypts to e.

- Look for frequent bigrams of the form $-Z$ or $Z-$.
- Frequent bigrams involving e include: $he, er, re, en, es, te, ed$.

By making successive guesses, the entire ciphertext can eventually be decrypted, revealing the key.

5 Cryptanalysis of the Vigenère Cipher

The keyword length (m) can be determined via brute force.

- If m is known, plaintext characters encrypted with the same key character can be grouped together.
- We partition the ciphertext into blocks of length m.

There are two methods for determining m, discussed below.

Method 1: Repetition Distance

Identical plaintext sections will encrypt to the same ciphertext when separated by a multiple of m.

Search the ciphertext for repeated segments and guess that m is the gcd of their distances.

Method 2: Index of Coincidence

The **index of coincidence** $I_c(x)$ is the probability that two randomly chosen elements of the plaintext $\mathbf{x} = x_1 x_2 \dots x_n$ are identical.

Suppose $|\mathbf{x}| = n$ and let f_i be the frequency of character i in $\mathbf{x}$.

- There are $\binom{n}{2}$ ways to choose two elements of $\mathbf{x}$.
- For each character i, there are $\binom{f_i}{2}$ ways to choose both elements to be i.

Thus, we define:

$$I_c(\mathbf{x}) = \frac{\sum_{i=0}^{25} \binom{f_i}{2}}{\binom{n}{2}} = \frac{\sum_{i=0}^{25} f_i(f_i-1)}{n(n-1)} \approx \frac{\sum_{i=0}^{25} f_i^2}{n^2}.$$

For English text, we approximate:

$$I_c(\mathbf{x}) \approx \sum_{i=0}^{25} p_i^2 \approx 0.065,$$

where p_i is the probability of letter i occurring.

For any monoalphabetic cipher (or each block of a Vigenère cipher), since the letters are only permuted, the individual probability terms in $\sum p_i^2$ are permuted similarly, but $I_c(x)$ remains unchanged.

Suppose $\mathbf{y} = y_1 y_2 \dots y_n$ has been encrypted using a Vigenère cipher. To determine the keyword length m, we define m substrings of $\mathbf{y}$:

$$\mathbf{y}_1 = y_1 y_{m+1} y_{2m+1} \cdots$$

$$\mathbf{y}_2 = y_2 y_{m+2} y_{2m+2} \cdots$$

$$\vdots$$

$$\mathbf{y}_m = y_m y_{2m} y_{3m} \cdots$$

where each substring consists of characters encrypted using the same letter of the keyword.

If m is the correct keyword length, then each substring $\mathbf{y}_i$ behaves like a monoalphabetic cipher, and we expect:

$$I_c(\mathbf{y}_i) \approx 0.065.$$

If m is incorrect, the substrings $\mathbf{y}_i$ will appear more random. For a completely random string, the expected index of coincidence is:

$$I_c \approx 26 \left(\frac{1}{26}\right)^2 = \frac{1}{26} \approx 0.038.$$

Since 0.065 and 0.038 are significantly different, we can determine the correct keyword length by testing various values of m and checking which one gives $I_c(\mathbf{y}_i) \approx 0.065$.

Finding k_i from $\mathbf{y}_i$

Let $n' = \frac{n}{m}$ be the length of each substring $\mathbf{y}_i$, and let f_c be the frequency of character c in $\mathbf{y}_i$.

- The probability distribution of letters in $\mathbf{y}_i$ is:

$$\frac{f_0}{n'}, \frac{f_1}{n'}, \ldots, \frac{f_{25}}{n'}.$$

- After shifting by k_i, the distribution becomes:

$$\frac{f_{k_i}}{n'}, \frac{f_{1+k_i}}{n'}, \ldots, \frac{f_{25+k_i}}{n'}.$$

- The shifted distribution should match the ideal letter distribution $p_0, p_1, \ldots, p_{25}$.

To estimate k_i, define:

$$M_g = \sum_{i=0}^{25} \frac{p_i f_{i+g}}{n'}.$$

- If $g = k_i$, then we expect:

$$M_g \approx \sum_{i=0}^{25} p_i^2 = 0.065.$$

- If $g \neq k_i$, then M_g should be around 0.05.

Cryptanalysis of the Hill Cipher

Breaking the Hill Cipher is difficult with a ciphertext-only attack but is easily broken using a **known-plaintext** attack.

If we have m known $(\mathbf{x}, \mathbf{y})$ pairs, we define two $m \times m$ matrices:

$$X = (x_{i,j}), \quad Y = (y_{i,j}),$$

giving the matrix equation:

$$Y = XK,$$

where K is the unknown key matrix.

- If X is invertible, then we can compute:

$$K = X^{-1}Y$$

  and recover the key.
- If X is not invertible, we must try additional plaintext-ciphertext pairs.

Under a **chosen-plaintext** attack, we can always ensure an inverse exists.

Since the block size m is relatively small, it can be determined via brute force.