Unicity Distance and Shannon's Theory of Cryptography

Liu Chang, Zhan Dongcheng, Zhan Shihan and Zhang Ziyi

I. Background

Definition I.1 (Cryptosystem)**.** Cryptosystem consists of a 5-tuple: $(\mathcal{P}, \mathcal{C}, \mathcal{K}, \mathcal{D}, \mathcal{E})$. $\mathcal{P}$ is the set of plaintext, which means unencrypted message. $\mathcal{C}$ is the set of ciphertext, which means encrypted message. $\mathcal{K}$ is a string a bits used by cryptography algorithm to encrypt the plaintext to ciphertext. $\mathcal{D}$ means decryption. It's a process that ciphertext transforms to plaintext. $\mathcal{E}$ means encryption. It's a process that plaintext transforms to ciphertext.

Definition I.2 (Entropy)**.** The Shannon Entropy H of a random variable X is

$$H(X) = -\sum p_i \log(p_i).$$

Note that p_i is the probability of element i appearing in the message.

Definition I.3. If a function satisfies

$$f\left(\frac{x+y}{2}\right) > \frac{f(x)+f(y)}{2}$$

where x and y are arbitrary values in an interval I, then the function is strictly concave in I. If a function satisfies

$$f\left(\frac{x+y}{2}\right) < \frac{f(x)+f(y)}{2}$$

where x and y are arbitrary values in an interval I, then the function is strictly convex in I.

Theorem 1. *Jensen's Inequality is related to convexity of function. Let f be an convex function on an interval I and assume $x_1, x_2, x_3, ...x_n$ are inside I, then*

$$f\left(\frac{x_1+x_2+...+x_n}{n}\right) \leq \frac{f(x_1)+f(x_2)+...+f(x_n)}{n}.$$

Note that equality establishes if and only if $x_1, x_2, ..., x_n$ are all equal for strictly convex function.

Property I.1. *Suppose X is a random variable defined by its distribution of probability $p_1, p_2, \cdots, p_n$. The Entropy of X is less or equal to $\log_2 n$, and the condition for equality is that if and only if*

$$p_1 = p_2 = \cdots = p_n = \frac{1}{n}.$$

Proof: Since the function $\log_2 n$ is strictly concave in an interval $(0, +\infty)$, Jensen inequality can be applied here.

$$\begin{aligned} H(X) &= \sum_1^n \log_2\left(\frac{1}{p_i}\right) p_i \\ &\leq \log_2 \sum_1^n \left(\frac{1}{p_i}\right) p_i \\ &= \log_2 n \end{aligned}$$

Property I.2. *If X and Y are two random variables ,$H(X,\ Y) \leq H(X) + H(Y)$ with same occurring if and only if two variables are independent.*

Proof: Let set $X = [x_1, x_2...x_n]$, set $Y = [y_1, y_2...y_n]$, $p_{xi} = p(X = x_i)$, $p_{yi} = p(Y = y_i)$, and $p_{i,j} = p(X = x_i, Y = y_i)$ is a joint probability distribution. It's easy to see that $p_{xi} = \sum_{j=1}^n r_{i,j}$, $p_{yi} = \sum_{i=1}^n r_{i,j}$.

$H(X,Y) = -\sum_{i,j} r_{i,j} \log r_{i,j}$.

$H(X) + H(Y) = -\sum_i p_{xi} \log p_{xi} - \sum_i p_{yi} \log p_{yi} = -(\sum_i \sum_j r_{i,j} \log p_{xi} + \sum_i \sum_j r_{i,j} \log p_{yi})$

Therefore, $H(X,Y) - (H(X) + H(Y)) = -\sum_i \sum_j r_{i,j} \log r_{i,j} + \sum_i \sum_j r_{i,j} \log p_{xi} p_{yi} = \sum_i \sum_j r_{i,j} \log \frac{p_{xi} p_{yi}}{r_{i,j}} \leq \log \sum_i \sum_j p_{xi} p_{yi} = 0$ (Because $r_{i,j}$ is a joint probability distribution, we can apply Jensen's inequality.)

Note that the equality establishes only when X and Y are independent. In other words, $\forall i, j, r_{i,j} = p_{xi} p_{yi}$.

Definition I.4. If two events A and B are independent, then the probability of event B given event A is

$$P(B|A) = \frac{P(A,B)}{p(A)}.$$

Definition I.5. Conditional entropy of two random variables A,B is written by $H(A|B)$, abnd it is easily derived from conditional probability function and Shannon Entropy equation

$$H(A|B) = -\sum \sum p(b) p(a|b) \log(p(a|b)).$$

II. Spurious Keys

The information contained in a given ciphertext would give out certain knowledge about the key distribution. In terms of the conditional entropy, we denote the information about the key $\mathcal{K}$ revealed by the ciphertext $\mathcal{C}$ by $H(\mathcal{K}|\mathcal{C})$, called the *key equivocation*. The key equivocation is given as follows:

Theorem 2. *For a cryposystem $(\mathcal{P}, \mathcal{C}, \mathcal{K}, \mathcal{E}, \mathcal{D})$, the key equivocation is $H(\mathcal{K}|\mathcal{C}) = H(\mathcal{K}) + H(\mathcal{P}) - H(\mathcal{C})$.*

Proof: By the aforementioned properties of conditional entropy, it is clear that $H(\mathcal{K}, \mathcal{P}, \mathcal{C}) = H(\mathcal{C}|\mathcal{K}, \mathcal{P}) + H(\mathcal{K}, \mathcal{P})$.

Notice that the ciphertext can be uniquely determined by a given key and plaintext, *i.e.*, $H(\mathcal{C}|\mathcal{K},\mathcal{P}) = 0$. Since the plaintext and the key are independent, we have $H(\mathcal{K},\mathcal{P},\mathcal{C}) = H(\mathcal{K},\mathcal{P}) = H(\mathcal{K}) + H(\mathcal{P})$. Since $H(\mathcal{P}|\mathcal{K},\mathcal{C}) = 0$, $H(\mathcal{K},\mathcal{P},\mathcal{C}) = H(\mathcal{K},\mathcal{C})$. Hence, $H(\mathcal{K}|\mathcal{C}) = H(\mathcal{K},\mathcal{C}) - H(\mathcal{C}) = H(\mathcal{K},\mathcal{P},\mathcal{C}) - H(\mathcal{C}) = H(\mathcal{K}) + H(\mathcal{P}) - H(\mathcal{C})$.

The key equivocation essentially describes the information about the distribution of key after obtaining a certain ciphertext. For a ciphertext C and a key distribution $\mathcal{K}$, a natural step is to filter out all the keys which are not capable of giving a meaningful plaintext. Therefore, it is important to know the number of keys capable of giving a meaningful plaintext for a given ciphertext. Such keys are defined as follows:

Definition II.1 (Spurious key)**.** A **Spurious Key** is a key, which is incorrect, yet is capable of giving a meaningful plaintext for a given ciphertext.

III. Entropy of a Natural Language and Redundancy

An n-gram is a contiguous sequence containing n consecutive items in a text. For example, "cat and dog" is a 3-gram. Denote n-grams in a plaintext by the random variable $\mathcal{P}^n$. A bound on the expected number of spurious keys for the the n-grams can be found.

For a random sequence, where English alphabets appear equiprobably, the entropy would be $H = \log_2 26 = 4.76$. It is different with natural languages, where alphabets appear with certain probability distributions. For a natural language L, the entropy per letter is the average information carried per letter. This entropy can be applied to roughly approximate the entropy of the language, H_L. However, the contiguous letters in natural languages are usually correlated. The frequencies of different n-grams differ a lot. For example, for digrams (2-grams), we can take the entropy of the probability distribution of all digrams, and divide it by 2 to get a better accuracy in terms of approximating the entropy per letter. For even better approximation of the entropy of the language, we have:

Definition III.1 (Entropy of a natural language)**.** The entropy of a natural language L, denoted by H_L, is given by

$$H_L = \lim_{n\to\infty} \frac{H(\mathcal{P}^n)}{n}.$$

Note that all correlation of contiguous letters in n-grams are taken into account in the definition above by taking the limit of n to ∞.

Take the most common natural language, English, as an example. The first order approximation for the entropy per letter of English is estimated to be 4.19 [**?**], [**?**], which is close to the entropy per letter of a completely random sequence of English alphabets. The second order approximation, which considers the entropy of all bigrams, is estimated to be 1.95 [**?**], [**?**]. Further statistical results show that, the entropy of English is approximately in the range $1 \le H_L \le 1.5$ [**?**], [**?**], [**?**]. It is clear that the correlations among the contiguous letters in n-grams for all n contribute to the decrease in the entropy of English texts.

It leads to the concept of redundancy, which measures the fraction of unnecessary letters in a text which can be deleted without changing the meaning of the text.

Definition III.2 (Redundancy of a natural language)**.** The redundancy of a natural language L is defined as

$$R_L = 1 - \frac{H_L}{\log_2 |\mathcal{P}|}.$$

The ratio of the entropy of the natural language H_L to the entropy of the random sequence $\log_2 |\mathcal{P}|$ is in the range of 0 and 1. If the natural language has little correlations among contiguous letters in n-grams, $i.e.$, closer to a random sequence, the ratio is closer to 1, and the information carried per letter is low. In this case, more letters are necessary to carry given information. Conversely, if the natural language has correlations among n-grams, the ratio is closer to 0, and the information carried per letter is high. Less letters would be needed to carry the same information.

In the aforementioned example of English, we take the entropy of English as 1.25. Thus, the redundancy of English is approximately 75% [**?**], [**?**], which gives an intuitive taste of ideally what fraction of letters can we remove while convey the same meanings in English by carefully encoding.

IV. Unicity Distance

Using the idea of spurious keys, we can measure the security of an encryption method $\mathcal{E}$. The intuitive interpretation is that given a fixed-length ciphertext, the more spurious keys exist in this system, the less likely an attacker can decrypt it back to the original message. In particular, if no spurious key exist, the attacker would be able to uniquely determine the plaintext and cipherkey given unlimited time and sources. We will then elaborate the idea mathematically. From now on we are assuming a ciphertext only attack with a ciphertext $y \in \mathcal{C}^n$ of length n.

y must be encrypted from a valid plaintext $x \in \mathcal{P}^n$ using $\mathcal{E}$. An equivalent statement is that

$$K(y) = \{k \in \mathcal{K} \mid \exists x \in \mathcal{P}^n \ s.t.\ Pr(x) > 0 \text{ and } y = e_k(x)\}$$

The cardinality of this set, $|K(y)|$ is the total number of possible keys and since one of them is the true cipherkey, the number of spurious keys is $|K(y)| - 1$.

Definition IV.1 (Average number of spurious keys)**.** Define $\bar{s}_n$ as the average number of spurious keys in this encryption system,

$$\begin{aligned}\bar{s}_n &= \sum_{y\in\mathcal{C}^n} (K(y) - 1) \cdot Pr(y) \\ &= \sum_{y\in\mathcal{C}^n} K(y)Pr(y) - 1\end{aligned}$$

We have previously proved an identity (theorem 2) regarding $H(\mathcal{K}|\mathcal{C}^n)$ and the link between $\bar{s}_n$ and $H(\mathcal{K}|\mathcal{C}^n)$ is

shown below,

$$\begin{aligned} H(\mathcal{K}|\mathcal{C}^n) &= \sum_{y\in\mathcal{C}^n} H(K|y)Pr(y) \quad \text{by def I.5} \\ &\le \sum_{y\in\mathcal{C}^n} \log_2|K(y)|Pr(y) \quad \text{by property I.1} \\ &\le \log_2 \sum_{y\in\mathcal{C}^n} |K(y)|Pr(y) \quad \text{by theorem 1} \\ &= \log_2(\bar{s}_n+1) \end{aligned}$$

Then we try to manipulate theorem 2 $H(\mathcal{K}|\mathcal{C}^n) = H(\mathcal{K}) + H(\mathcal{P}^n) - H(\mathcal{C}^n)$ to build a bridge between $\bar{s}_n$ and known variables. Consider the following approximation or scaling:

- If the length of the ciphertext is long enough, we can approximate the entropy of $\mathcal{P}^n$ by the language entropy (def III.1) as $H(\mathcal{P}^n) \approx n \cdot H_L$.
- Use the definition of redundancy (def III.2) of a natural language and rearrange the equation, $H_L = \log_2|P| \cdot (1-R_L)$.
- Take the upper bound of ciphertext entropy. Use entropy property I.1 and I.2, $H(\mathcal{C}^n) \le nH(\mathcal{C}) \le n\log_2|C|$.
- Assume $|P| = |C|$, the numbers of elements in plaintext set and ciphertext set are identical.

Bring them into theorem 2,

$$\begin{aligned} H(\mathcal{K}|\mathcal{C}^n) &= H(\mathcal{K}) + H(\mathcal{P}^n) - H(\mathcal{C}^n) \\ &\ge H(\mathcal{K}) + n\cdot\log_2|P|\cdot(1-R_L) - n\log_2|C| \\ &= H(\mathcal{K}) - n\cdot\log_2|P|\cdot R_L \end{aligned}$$

As a summary of everything above, we now have a beautiful inequality

$$H(\mathcal{K}) - n\cdot\log_2|P|\cdot R_L \le H(\mathcal{K}|\mathcal{C}^n) \le \log_2(\bar{s}_n+1)$$

Rearrange the terms to get $\bar{s}_n \ge \frac{2^{H(\mathcal{K})}}{|\mathcal{P}|^{nR_L}} - 1$. It is reasonable to assume all keys are drawn equiprobably and in such case $H(\mathcal{K}) = \log_2(|\mathcal{K}|)$. This completes the proof of the following theorem.

Theorem 3. *If n is large enough, $|P| = |C|$ and keys are chosen equiprobably, the expected number of spurious keys has a lower bound,*

$$\bar{s}_n \ge \frac{|\mathcal{K}|}{|\mathcal{P}|^{nR_L}} - 1.$$

As discussed before, we are interested in the extreme case where $\bar{s}_n \to 0$. The ciphertext length n_0 that achieves this result is called the **Unicity Distance**. If an attacker has a ciphertext whose length exceeds the unicity distance, the encryption system can be uniquely broken theoretically. Using theorem 3, we can calculate n_0 as

$$n_0 = \frac{\log_2|\mathcal{K}|}{R_L\log_2|\mathcal{P}|}$$

V. Interpretation and application

After the proof and some primitive understanding of the unicity distance, we now formally introduce the definition.

Definition V.1 (Unicity Distance)**.** The **unicity distance** of a cipher is the minimum amount of ciphertext (number of characters) required to allow a computationally unlimited adversary to recover the unique encryption key.

This idea is critical in unconditional security in that an encryption system exceeding the limit of its unicity distance is definitely not secure regardless of the implementation details or attack algorithms. The formula of n_0 () also gives us some hints on how to build a more secure (in the sense of unconditional secure but not practical secure) encryption system. To get a large, or infinite if possible, unicity distance, we would like

- A large $|\mathcal{K}|$. One-time pad is a good example that has the property *perfect secrecy* by having an infinite key space.
- A small R_L. Redundancy can be reduced or even eliminated by *compression*[**?**]. So it is always a good idea to deploy compression algorithms prior to encryption.
- A small $|\mathcal{P}|$. This is hard to manipulate since the the plaintext set is usually a meaningful language in practice.

Aside from the discussion of unconditional security, there are several common mistakes when it comes to the unicity distance in applied cryptography.

- "A small unicity distance does not necessarily imply that a block cipher is insecure in practice."[**?**] Take 128-bit AES as an example, the unicity distance of which is roughly 37 characters. So theoretically every AES-encrypted message can be cracked, but at least that is not feasible today.
- Not exceeding the limit of unicity distance does not guarantee security. First the formula is a result of several steps of approximation and scaling. So the proposition "if $n > n_0$ then the cipherkey can be uniquely determined" does not imply "if $n < n_0$ then the cipherkey cannot be uniquely determined". Second even if given the current ciphertext, there indeed exists more than one valid key, one can still decrypt the message in a short time. Because out of the possible, say one thousand, keys, only few of the corresponding plaintexts is meaningful in the sense of a language.

But these common misinterpretations do not mean unicity distance is of no value in applied cryptography. Let us conclude it using an example of simple substitution cipher.

Empirical evidence suggests that, "for essentially any simple substitution cipher on a meaningful message (e.g., with redundancy comparable to English), as few as 25 ciphertext characters suffices to allow a skilled cryptanalyst to recover the plaintext."[**?**] In this case $|\mathcal{K}| = 26! \approx 4\times10^{26}$, so the unicity distance of simple substitution cipher is around $\frac{\log_2(4\times10^{26})}{0.75\times4.7} = 25.1$ characters. The theoretical value is in line with empirical evidence.