# Scribe: Cryptography and Network Security (Class.5.A)

Akash Tiwari

29-Sep-2020

## 1 Unicity Distance and Dependencies

In this lecture, we define a Unicity distance in words and then define the definitions/calculations of some of the dependencies we need to Quantify the unicity distance.

- Section 2 - Defining Unicity Distance
- Section 3 - Pre-requisites from the prev lecture used
- Section 4 - Spurious keys
- Section 5 - Entropy of a Language - $H_L$
- Section 6 - Redundancy of a Language - $R_L$
- Section 7 - A lower Bound of equivocation of key
- Section 8 - Expected number of spurious keys
- Section 9 - An upper bound of equivocation of key
- Section 10 - Quantifying Unicity Distance

## 2 Unicity Distance

Under the assumption of a boundless attacker,It is defined as the least amount of plain text can can be determined uniquely from the corresponding cipher text.
Note that if the key is changed within the unicity difference, even a brute force approach will not guarantee a unique key to the attacker due to the lack of data available to the attacker.

# 3 Pre-requisites from last lecture

## 3.1 Entropy

Let us take a random variable X which takes values from a finite set. Then the entropy is defined as -

$$H(X) = -\sum_{x \in X} P[x] * log_2(P[x])$$

## 3.2 Key Equivocation

Uncertainty of the key given the cryptogram and is defined as(H represent the entropy)-

$$H(K|C) = H(K) + H(P) - H(C)$$

# 4 Spurious Keys

Let the correct key of a cipher be$K_0$. Now suppose an attacker guesses all the keys and keeps the keys that make meaningful decryption in a set K. Then the Keys in the set $K - \{K_0\}$ are called as spurious Keys.

# 5 Entropy of a Language - $H_L$

It quantifies the amount of information per letter of meaningful words of plain text.
For eg- Let us assume the set $Z_{26}$ for English alphabets and let us assume each letter occurs with the same probability i.e 1/26.
Then a random string, by the definition, will have the entropy-

$$\begin{aligned} & -\sum_{x \in Z_{26}} \frac{1}{26} * log_2(\frac{1}{26}) \\ & = \sum_{x \in Z_{26}} \frac{1}{26} * log_2(26) \\ & = log_2(26) \\ & = 4.76(approx) \end{aligned}$$

But the English language does not have a uniform distribution for the letters and thus this value when calculated comes approximately 4.19. This entropy calculation can be done for bi-grams, trig-rams and n-grams as well.

Now, to define $H_L$, we define the random variable $P_L$ = probability distribution of N-grams of plain-text. Now, $H_L$ is defined as -

$$H_L = \lim_{n \to \infty} \frac{H(P_n)}{n}$$

## 6 Redundancy of a Language - $R_L$

It quantifies the fraction of excess symbols we're using as encoding.
We saw for the English language in the previous section that in a uniform distribution of a language the entropy came out to be $log_2|P|$ where $|P|$ represent the number of symbols in the language(26 for English). Then we can define the redundancy as -

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

## 7 A lower Bound of equivocation of key

In this we use the definitions of Redundancy,Equivocation and Entropy of a language to get a lower bound on the equivocation.
$P^n$ : Random Variable representing an n-gram plain text
$C^n$ : Random Variable representing an n-gram cipher text
Now, From the definition of equivocation discussed in pre-reqs-

$$H(K|C^n) = H(K) + H(P^n) - H(C^n)$$

Now, From the definition of entropy of a language -

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

For large n -

$$H(P^n) = n * H_L$$

Using this result and the definition of Redundancy -

$$R_L = 1 - \frac{H_L}{log_2|P|}$$
$$R_L = 1 - \frac{H(P^n)}{n * log_2|P|}$$
$$1 - R_L = \frac{H(P^n)}{n * log_2|P|}$$
$$H(P^n) = n * (1 - R_L) * log_2|P|$$

Now, we know that the max value of entropy exists at uniform distribution and is equal to $log_2|P|$ where $|P|$ represent the number of symbols in the language. However, in other distribution it decreases, therefore-

$$\frac{H(C^n)}{n} \leq log_2(|C|)$$
$$H(C^n) \leq n * log_2(|C|)$$

Now if we place these results into the equivocation definition, we will get the lower bound

$$H(K|C^n) \geq H(K) + n * (1 - R_L) * log_2|P| - n * log_2(|C|)$$

Assuming $|P| = |C|$

$$H(K|C^n) \geq H(K) - n * R_L * log_2|P|$$

# 8 Expected number of spurious keys

Expectation defined as -

$$\sum x_i * P(X = x_i)$$

Expected number of spurious keys is calculated as -

$$s_n = \sum_{y \in C^n} P(y)(|K(y)| - 1)$$

$$s_n + 1 = \sum_{y \in C^n} P(y)|K(y)|$$

# 9 Computing the upper bound of equivocation of key

First let us define a Concave function, a function is called Concave if -

$$f(\frac{x+y}{2}) \geq \frac{f(x) + f(y)}{2}$$

We will now show that $log_2(x)$ is a concave function. Given $x > 0$ and $y > 0$, We know that the arithmetic mean of x and y is greater than equal to Geometric mean i.e -

$$\frac{x+y}{2} \geq \sqrt{x * y}$$

$$log_2(\frac{x+y}{2}) \geq log_2(\sqrt{x * y})$$

$$log_2(\frac{x+y}{2}) \geq \frac{log_2(x * y)}{2}$$

$$log_2(\frac{x+y}{2}) \geq \frac{log_2(x) + log_2(y)}{2}$$

$$f(\frac{x+y}{2}) \geq \frac{f(x) + f(y)}{2}$$

Now, according to Jensen's inequality, For a concave function f(x) if-

$$sum_{i=1}^{n} a_i = 1$$

then -

$$sum_{i=1}^{n} a_i f(x_i) \leq f(sum_{i=1}^{n} a_i * x_i)$$

Now, after proving $log_2(x)$ is concave, we will use this inequality later. We will now find the upper bound of equivocation, By the definition of conditional entropy, we can write-

$$H(K|C^n) = \sum_{y \in C^n} P(y) H(K|y)$$
$$H(K|C^n) = \sum_{y \in C^n} P(y) H(K(y))$$

Now, we know max value of H(K(y)) is $log_2(|K(y)|)$, substituting this value gives -

$$H(K|C^n) \leq \sum_{y \in C^n} P(y) * log_2(|K(y)|)$$

The conditions for Jensen's Inequality is satisfied -

$$H(K|C^n) \leq log_2( \sum_{y \in C^n} P(y) * |K(y)|)$$
$$H(K|C^n) \leq log_2(s_n + 1)$$

# 10 Quantifying Unicity Distance

Now we combine both the lower bounds and upper bounds of equivocation to get-

$$H(K) - n * R_L * log_2|P| \leq H(K|C^n) \leq log_2(s_n + 1)$$
$$H(K) - n * R_L * log_2|P| \leq log_2(s_n + 1)$$

Now, assuming keys are chosen with equal probability, we get $H(K) = log_2(|K|)$, substituting this -

$$log_2(|K|) - n * R_L * log_2|P| \leq log_2(s_n + 1)$$
$$log_2(|K|) - log_2(|P|^{n*R_L}) \leq log_2(s_n + 1)$$
$$log_2(\frac{|K|}{|P|^{n*R_L}}) \leq log_2(s_n + 1)$$
$$\frac{|K|}{|P|^{n*R_L}} \leq s_n + 1$$
$$s_n \geq \frac{|K|}{|P|^{n*R_L}} - 1$$

Now, if we equate Sn = 0 and find the corresponding $n_0$,

$$s_n \geq \frac{|K|}{|P|^{n_0 * R_L}} - 1 = 0$$
$$\frac{|K|}{|P|^{n_0 * R_L}} - 1 = 0$$
$$\frac{|K|}{|P|^{n_0 * R_L}} = 1$$
$$|K| = |P|^{n_0 * R_L}$$
$$log_2(|K|) = n_0 * R_L * log_2(|P|)$$
$$n_0 = \frac{log_2(|K|)}{R_L * log_2(|P|)}$$

So Unicity distance Will be defined as $n \geq n_0$ such that the spurious keys are reduced to 0.

Example of calculation of Unicity distance for a substitution cipher -

$$|P| = 26$$
$$|K| = 26! = 40, 32, 91, 46, 11, 26, 60, 56, 35, 58, 40, 00, 000$$
$$R_L = 0.75$$
$$n_0 = \frac{log_2(|K|)}{R_L * log_2(|P|)}$$
$$n_0 = \frac{log_2(26!)}{0.75 * log_2(26)}$$
$$n_0 = 25.07$$

therefore unicity distance is 25 and given cipher-text of length 25 its possible to predict the correct key uniquely.

# 11 Conclusion

Thus we have successfully defined and quantified unicity distance which is one of the key parameters we should look at while looking at a cipher.