CHAPTER 3

INFORMATION

THEORITIC APPROACH
3. INFORMATION THEORITIC APPROACH

3.1 INTRODUCTION

Cryptographic system attains secrecy if plaintext can be recovered from ciphertext only by those who are authorized for it. Two types of secrecy can be achieved, theoretical secrecy and practical secrecy. In theoretical secrecy there is always a measure of uncertainty, regardless of what method of analysis is used. In this strong model, one assumes that the adversary is having infinite computing power, but restricted to a ciphertext-only attack. The goal of modern cryptographers is to build systems such that one key can be used many times. Such systems are unconditionally secure, by Shannon’s theorem, and must be at best only computationally secure. There is a need to develop the information theory needed to deal with these computationally secure systems.

In practical secrecy, sufficient information is available to break the cipher. Shannon invented a theoretical model using information theory and described the relationship between the amount of intercepted ciphertext and the likelihood of a successful attack. It is not sure that for a system which is not ideal, there is a unique solution for sufficiently large $N$ requires a large amount of work to
break with every method of analysis. There are two approaches to this problem: (1) Study the possible methods of solution available for cryptanalysis and attempt to describe them in sufficiently general terms to cover any method that can be used. Then construct the system to resist this “general” method of solution. (2) Construct the cipher in such a way that breaking it is equivalent to finding the solution of some problem known to be laborious. Thus, if one can show that solving a certain system requires at least as much work as solving a system of simultaneous equations with a large number of complex unknowns, it is possible to have a lower bound for the work characteristic.

3.2 ROLE OF ENTROPY AND REDUNDANCY IN CRYPTANALYSIS

Information theory defines the amount of information in a message as the minimum number of bits required to encode all possible meanings of that message. The assumption is that all messages have equal probability. The entropy of a natural language is a statistical parameter that measures, how much information is produced on an average for every letter of a text in the language. The complex properties of natural languages play an important part in cryptography. A natural language in addition to statistical properties has another fundamental property, which is loosely referred to as meaning. For a meaningful message it is possible to shorten it without
destroying the meaning. This property of a natural language is called redundancy. Redundancy measures the amount of constraint imposed on a text in the language due to its statistical structure.

In calculating the entropy and redundancy, long range statistics, extending over phrases, sentences, etc. has to be taken into account. For this, a study of predicting the next letter of a text be predicted when the preceding \( N \) letters are known is necessary. The experimental and theoretical results are combined to estimate upper and lower bounds for the entropy and redundancy. From this analysis it seems that, the long range statistical effects reduce the entropy to a value of the order of one bit per letter and the corresponding redundancy of around 75% is observed with English. The redundancy is still higher when paragraphs, chapters, etc. are included in the structure.

According to Shannon, Entropy is calculated using a series of approximations \( F_0, F_1, F_2, \ldots \ldots \) which considers more statistics of the language. The \( n \)-gram entropy \( F_n \) measures the entropy due to statistics extending over \( n \) adjacent letters. \( F_n \) is given by the expression

\[
F_n = -\sum_{i,j} p(m_i, j) \log_2 p_{m_i}(j)
\]

\[
= -\sum_{i,j} p(m_i, j) \log_2 p(m_i, j) + \sum_i p(m_i) \log_2 p(m_i) \ldots \ldots \ldots \ldots \ldots \ldots (3.1)
\]
Where

$m_i$ is a block of $n-1$ successive letters

$p(m_i, j)$ is the probability of the $n$-gram $m_i, j$

$p_{mi}(j)$ is the conditional probability of letter $j$ after the block $m_i$

$j$ is any arbitrary number following the block $m_i$

$F_n$ in the above expression measures the average uncertainty of letter $j$ when $n-1$ proceeding letters are known. When long range statistics are considered the entropy $H$ is given by

$$H = \lim_{n \to \infty} F_n$$

The $n$-gram entropies for small values of $n = 1, 2, 3$ are calculated by using standard unigram, bigram and trigram probabilities. The value of $F_0$ is equal to $\log_2 (A)$ where $A$ represents the size of the alphabets of the language under consideration. The value of $A$ for Telugu is assumed to be 64 as some of the characters are not in usage and it is 26 for text in English with out spaces.

To calculate the entropy values for Telugu text the corresponding $n$-gram probabilities of the Telugu text are to be computed. The process is to be repeated to find unigram, bi gram and tri gram probabilities and to calculate the corresponding entropies. To find the unigram probabilities, the statistical behavior of all characters in Telugu expressed as a percentage of the letters in a sample of over 32,00,000 characters is evaluated.
They show quite clearly, that English text is likely to be dominated by a small number of letters. When text in Telugu is considered, the Table 3.1 shows the probabilities of the character code points of the alphabet in a sample of over 3,200,000 characters taken from passages from numerous newspapers, novels, stories, songs, sports
and literature etc. The reason for probabilities of certain code points in the above table to be zero is that they are the deprecated characters in the usage of the language. The zero frequencies are observed for the numbers from 0 to 9 in Telugu language which are not used in colloquial language.

Using these probabilities, the entropy values of unigram($F_1$), bigram($F_2$) and trigram($F_3$) of Telugu text are calculated.

$$F_1 = - \sum_{i=1}^{64} p(i) \log_2 p(i)$$

$$= 4.7 \text{ bits/letter}$$

$$F_2 = - \sum_{i,j} p(i,j) \log_2 p(i,j) + \sum_{i} p(i) \log_2 p(i)$$

$$= 8.67 - 4.70$$

$$= 3.97 \text{ bits/letter}$$

$$F_3 = - \sum_{i,j,k} p(i,j,k) \log_2 p(i,j,k) + \sum_{i,j} p(i,j) \log_2 p(i,j)$$

$$= 12.10 - 8.67$$

$$= 3.43 \text{ bits/letter}$$

The entropy values of unigram, bi gram and tri gram are also evaluated for Kannada, Hindi and English and are listed in Table 3.2.
### Table 3.2 n-gram Entropy values of different Languages

<table>
<thead>
<tr>
<th>S.No.</th>
<th>n-gram</th>
<th>Entropy of n-gram model</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Telugu</td>
</tr>
<tr>
<td>1</td>
<td>1-gram</td>
<td>4.70</td>
</tr>
<tr>
<td>2</td>
<td>2-gram</td>
<td>3.97</td>
</tr>
<tr>
<td>3</td>
<td>3-gram</td>
<td>3.43</td>
</tr>
</tbody>
</table>

The entropy of unigram for Telugu, Kannada, Hindi, English are 4.70, 4.76, 5.09, 4.14 respectively. Information content of any language is dependent on the amount of statistical information with regard to various units of the language model. Uncertainty is the reciprocative behaviour of information. This is the reason why the entropy values gradually decrease with increase in n of n-gram value for any language as shown in Table 3.2.

Successive letters in the language are not independent. Relations among successive letters further reduce the entropy. Thus the average information content of the language decreases. The structure of the language itself is going to introduce some redundancy. A language has a particular value of redundancy if for any N it is possible to show that the possible sequences of N letters are not all equally probable. In a redundant language, the messages of N letters can be divided into two sets, meaningful and not meaningful. As N increases, the ratio of
meaningful to meaningless messages approaches zero. The cryptogram has a unique solution. The redundancy $R_L$ is defined to be

$$R_L = 1 - \frac{H_L}{\log_2 |P|} \quad \text{.........................(3.2)}$$

Where $H_L$ is the entropy of the language and $|P|$ is the size of the message alphabet. This $R_L$ is a measure of the fraction of excess characters.

According to Schenier [SCH 2000], the redundancy of English is found to be 3.4 bits per letter i.e. approximately 75%. The redundancy of Telugu is evaluated using maximum entropy and the rate of the language and is found to be approximately 19% which is in agreement [BHA et al. 2002] with the results of Rajeev Sanghal et al. An encrypted message is protected against cryptanalysis, the less redundancy it contains. Thus language characteristics play a vital role in cryptanalysis and in providing strength to the cryptographic system. While performing encryption, these language characteristics need to be considered so that for more complex languages, strength can be improved with less key size.

### 3.3 MESSAGE EQUIVOCATION AND KEY EQUIVOCATION

According to Shannon’s mathematical theory of communication systems, information can be conveniently measured by means of entropy. Assume that the $p1; p2; \ldots; p_n$ be the probabilities of the set of possibilities, then the entropy $H$ is given by
\[H = \sum_{i=1}^{n} p_i \log p_i\]

The uncertainty is associated with the choice of the message and choice of the key. The amount of information produced when a message is chosen is given by

\[H(m) = \sum_{i=1}^{n} p(m_i) \log p(m_i)\]

The amount of information produced when a key is chosen is given by

\[H(k) = \sum_{i=1}^{n} p(k_i) \log p(k_i)\]

For a perfect secrecy system, the maximum amount of information in the message is at most \(\log_2 n\) (when all messages are equiprobable). So, the information can be concealed completely if the key uncertainty is at least \(\log_2 n\).

Let \(m_1, m_2, \ldots, m_n\) be the possible finite messages and \(P(m_1), P(m_2), \ldots, P(m_n)\) be the corresponding apriori probabilities. These messages are transformed into \(c_1, c_2, \ldots, c_n\) by using a transformation \(T_i\) which is represented as \(c = T_i m\). In the process of cryptanalysis, for the intercepted message \(c\) the posterior probability \(p(m|c)\) is calculated. The intercepting message gives no information about the plain text if the posterior probability is equal to the apriori probability, which is called the perfect secrecy condition. According to Shannon’s Communication theory of secrecy systems, a
necessary and sufficient condition to achieve perfect secrecy is that
\[ P(m | c) = P(m) \] for all \( m \) and \( c \).

Assume that a simple substitution cipher is used on any natural
language text and that \( N \) letters of the cipher text is intercepted. For
sufficiently large value of \( N \) always there is a unique solution for the
cipher. With a smaller value of \( N \), the chance of more than one
solution is greater. Before cipher text is intercepted, calculate the
apriori probabilities of the various possible messages and also that of
the keys. As the cipher text is intercepted, calculate the posterior
probabilities. As \( N \) increases, the probabilities of certain messages
increase, and of most decrease, until only one is left, whose
probability is nearly one, while the sum of probabilities of all others is
nearly zero.

It is desirable to use the equivocation as a theoretical index for
measuring secrecy. There are two significant equivocations, one that
of the key and other that of the message. These are denoted by
\( H(k | c) \) and \( H(m | c) \) respectively. The message and key equivocation
are given by
\[
H(m | c) = \sum_{m,c} P(m,c) \log P(m | c)
\]
\[
H(k | c) = \sum_{k,c} P(k,c) \log P(k | c)
\]
Where \( m, c, k \) are plain text, cipher text and key respectively. \( P(m|c) \), \( P(k|c) \) are the posterior probabilities of message, key respectively if cipher text \( c \) is intercepted. \( P(m,c) \) and \( P(k,c) \) are the probabilities of message and cipher text, key and cipher text respectively.

The summations \( H(k|c) \) and \( H(m|c) \) are over all possible cryptograms of length say \( N \) letters, over all keys and all messages respectively. Thus both key and message equivocations are functions of the number of intercepted letters \( (N) \). A zero equivocation requires that one message (or key) have unit probability, all other have zero probability. The gradual decrease of equivocation corresponds to increasing knowledge of the original key or message. Equivocation can be considered as a function of \( N \). It is possible to construct secrecy systems with a finite key in which the equivocation does not approach zero as \( N \to \infty \). In this case, no matter how much material is intercepted, it is not possible to obtain a unique solution to the cipher. Such systems are called ideal systems.

### 3.4 Unicity Distance

Let \((M,K,C,E,D)\) is a crypto system where \(M,K,C\) are plain text message space, key space and cipher text space respectively. \(E\) and \(D\) represent encryption and decryption transformations. Then using the entropy relations the following expression holds good

\[
H(K|C) = H(K) + H(P) - H(C)
\]
Where $H(K|C)$ is the entropy of key given the cipher text and $H(K)$, $H(P)$, $H(C)$ represent entropy of key, plain text and cipher text respectively.

Now consider a string of plain text symbols $m_1, m_2, ... , m_n$ on encryption with a key produces cipher text $c_1, c_2, ... , c_n$. Assume that the attack is a cipher text only attack and there are infinite computational resources and the language is a natural language. Certain keys are ruled out, but certain keys remain out of which only one key is the correct key and all other keys are called spurious keys.

Assuming that size of $P$ and size of $C$ are equal and all keys are equiprobable. Let $R_L$ is the redundancy of the underlying language. Then for a string of sufficiently large cipher text of length $n$, the expected number of spurious keys, $s_n$, satisfies the relation

$$S_n = \frac{|K|}{|P|^n R_L} - 1$$

(3.3)

As $n$ increases the quantity on the right side of the above expression approaches zero exponentially. This is the condition where the number of spurious keys is zero. The value $n_0$ of $n$ at which the number of spurious keys $S_n$ is zero is called unicity distance $U$.

Substitute $S_n = 0$ in the above expression,

$$n_0 \approx \log_2 |k| / (R_L \log_2 |P|)$$

Therefore $U \approx H(K) / (R_L \log_2 |P|)$ ..............(3.4)
For natural languages and the usual type of ciphers the unicity distance is approximately \( H(K)/D \) where \( H(K) \) is a measure of the “size” of the key space. When all keys are equally likely \( H(K) \) is the logarithm of the number of possible keys. \( D \) is the redundancy of the language in bits and is a measure of the amount of “statistical constraint” imposed by the language. The unicity distance of a cryptosystem is defined to be the average amount of cipher text needed for an opponent to uniquely determine the key. The unicity distance is directly proportional to key uncertainty and inversely proportional to the redundancy. As redundancy approaches zero, unicity distances tend to infinity. Then a trivial cipher is unbreakable with a cipher text-only attack. Unicity distance do not make any deterministic predictions, rather gives probabilistic results. Unicity distance is not a measure of cipher text required for cryptanalysis, but the ciphertext required for only one reasonable solution for cryptanalysis. In general, longer the unicity distance, better the cryptosystem in terms of the strength.

Figure 3.1 shows the Shannon’s message and key equivocation characteristics for varying cipher text length. The message equivocation curve follows Shannon’s prediction that is rising and then falling. Because short ciphers have relatively small number of solutions and the average uncertainty is less. As the cipher length is increased, the message equivocation rises. At some point, the message
equivocation decreases, as the cipher begins to disclose its secret through repeated patterns.

**Figure 3.1 Shannon’s Equivocation characteristics for simple substitution on English, for human-level language model**

According to Shannon, by analytical means the unicity distance in terms of n-gram entropies is given by the expression

\[
\text{Unicity Distance } (U) = \frac{H(K)}{(A-B)} \quad \text{..........................(3.5)}
\]

Where \( H(K) = \text{Entropy of Key} \)

\( A = \text{Entropy of 0-gram model} \)

\( B = \text{Entropy of n-gram model used} \)

Using the n-gram entropies of English and Telugu in Table 3.2, the unicity distance of English, Telugu, Kannada, Hindi are calculated using the expression (3.5) and are listed in Table 3.3. The unicity
distance value for Telugu, Kannada, Hindi and English for Unigrams are 228, 248, 328, 167 respectively. These values decrease with increase in n-gram value which can be concluded from the results. The unicity distance of Indic scripts is much higher than English for any n-gram value. It is evident from the results that the unicity distance value decreases as higher n-gram language models are considered which are shown in Figure.3.2. The results indicate that, more the language statistics available, easier in revealing the message contents. It is also evident that the algorithm strength in the form of unicity distance is more for Indic scripts than English and Deciphering is more complex for Indic scripts than English.

<table>
<thead>
<tr>
<th>S.No.</th>
<th>Language</th>
<th>Unicity Distance</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>1-gram</td>
</tr>
<tr>
<td>1</td>
<td>Telugu</td>
<td>228</td>
</tr>
<tr>
<td>2</td>
<td>Kannada</td>
<td>248</td>
</tr>
<tr>
<td>3</td>
<td>Hindi</td>
<td>328</td>
</tr>
<tr>
<td>4</td>
<td>English</td>
<td>167</td>
</tr>
</tbody>
</table>

Table 3.3 Unicity Distance of Telugu and English for n-grams of n=1,2,3.
Redundancy measures the amount of constraint imposed on a text in the language because of its statistical nature of letters. Based on the entropy of the language and the size of n-gram space the redundancy $R_L$ is calculated using the expression (3.2) for four different languages Telugu, Kannada, Hindi and English which are listed in Table 3.4. English is more redundant with 75% redundancy and for Telugu, Kannada, Hindi these values are 19, 19.2, 34.6 respectively.

**Table 3.4 Redundancy percentage for Four different Languages**

<table>
<thead>
<tr>
<th>S.No.</th>
<th>Language</th>
<th>Redundancy</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Telugu</td>
<td>19</td>
</tr>
<tr>
<td>2</td>
<td>Kannada</td>
<td>19.2</td>
</tr>
<tr>
<td>3</td>
<td>Hindi</td>
<td>34.6</td>
</tr>
<tr>
<td>4</td>
<td>English</td>
<td>75</td>
</tr>
</tbody>
</table>
Based on the size of the key in bits, the entropy of the key is calculated. Unicity distance is calculated using the redundancy of different languages and the entropy of the key and by using expression (3.4). The procedure is repeated for Telugu, Kannada, Hindi and English languages with different key sizes in the range of 40 to 256 bits and the computed unicity distance values are listed in Table 3.5.

<table>
<thead>
<tr>
<th>S.No.</th>
<th>Key Length (in bits)</th>
<th>Unicity Distance (in characters)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>ENGLISH</td>
</tr>
<tr>
<td>1</td>
<td>40</td>
<td>6</td>
</tr>
<tr>
<td>2</td>
<td>56</td>
<td>9</td>
</tr>
<tr>
<td>3</td>
<td>64</td>
<td>10</td>
</tr>
<tr>
<td>4</td>
<td>80</td>
<td>12</td>
</tr>
<tr>
<td>5</td>
<td>128</td>
<td>19</td>
</tr>
<tr>
<td>6</td>
<td>256</td>
<td>38</td>
</tr>
</tbody>
</table>

The unicity distance values of English are in the range of 6 to 38 characters for key in the range of 40 to 256 bits. For the same range of key these values are varying from 13 to 82, 14 to 87 and 11 to 68 characters respectively for Telugu, Kannada and Hindi. Unicity distance for English with a 128 bit key is observed as 19, where as the
same measure is found between key sizes of 56 and 64 for Indic scripts viz. Telugu, Kannada and Hindi. The similar strength of the crypto system which uses message units of English with a help of 128 bit key size can be achieved with less than 64 bit key on Indic scripts. Figure 3.3 illustrates the variation of unicity distance for varying key length over English, Telugu, Kannada and Hindi. It is clear from the figure that for any language increase in key size results in increase in unicity distance.

![Figure 3.3 Unicity Distance of different languages for different key sizes](image)

The unicity distance which is a metric for the strength of the cryptographic system is more for Telugu than English for a particular key size. But the techniques that are available in the literature for ASCII based Latin texts are applied directly with the same key size on any application irrespective of its language. This results in increased
hardware and software complexity. It is easy to conclude from the results that, while applying any cryptographic algorithm, depending on the language of the application and corresponding unicity distance appropriate key size is to be selected.

The above probabilistic measure is based on the assumption that the language units are equi-probable nature. In the real world scenario it is necessary to understand and realize the behavioral patterns of message units of every language to strengthen the probabilistic prediction as described by Shannon. The present work addresses this issue using cipher text only attack in decipherment processes.