ELSEVIER

Information Processing Letters 60 (1996) 43-47

Information Processing Letters

The unicity distance: An upper bound on the probability of an eavesdropper successfully estimating the secret key

A.Kh. Al Jabri

EE Department, College of Engineering, King Saud University, P.O. Box 800, Riyadh 11421, Saudi Arabia

Received 16 February 1996
Communicated by S.G. Akl

Abstract

The unicity distance, U, of a secret-key cipher is defined by Shannon as the minimum amount of intercepted ciphertext symbols needed, in principle, to uniquely determine the secret key and, therefore, break the cipher. Accordingly, for a ciphertext of size N symbols less than U, the estimated key will have a nonzero probability of error. Of interest is knowing the chance or probability that an eavesdropper, *using the best estimation rule*, successfully estimates the secret key from N ciphertext symbols less than U. An upper bound on this probability is derived in this paper.

Keywords: Secret-key cipher; Probability of success; Unicity distance; Distributed systems; Analysis of algorithms

1. Introduction

In his paper "Communication theory of secrecy systems" [5], Shannon introduced many useful concepts that paved the way for a better understanding of the limits on the performance of secrecy systems. One of these concepts is the *unicity distance* (U) of a secret-key cipher defined as the minimum amount of intercepted ciphertext required, in principle, to determine the key uniquely. In [5], an expression for U was derived for a special kind of cipher known as the random cipher. For other secret-key ciphers, Shannon suggested the possibility of using the same expression but with some corrections [5, p. 693]. An extension to Shannon's work was given by Hellman [3] who used a counting argument, for a given sequence of intercepted ciphertext symbols, to find the number of keys that could have generated that particular sequence. Hellman then rederived the Shannon's expression for the unicity distance of the random cipher and showed that this cipher yields, among the class of ciphers with the same key size and input, the minimum U which is essentially the worst one. Beauchemin and Brassard generalized Hellman results to include ciphers with arbitrary key and message distributions [1].

In designing a cipher, one would like to make U as large as possible. In principle, one should change the secret key after a number of encryption times less than U. In practice, however, the same key is usually used to encrypt much more ciphertext [4]. In this case, the interceptor will, in principle, be able to determine the secret key from these ciphertext symbols. If, however, the number of intercepted ciphertext symbols is less than U, then a reasonable question to be asked is: what is the chance or probability that an eavesdropper, *using the best estimation rule*, successfully estimates the secret key from these symbols? In this paper an upper bound on this probability is derived. To obtain

PII S0020-0190(96)00138-X

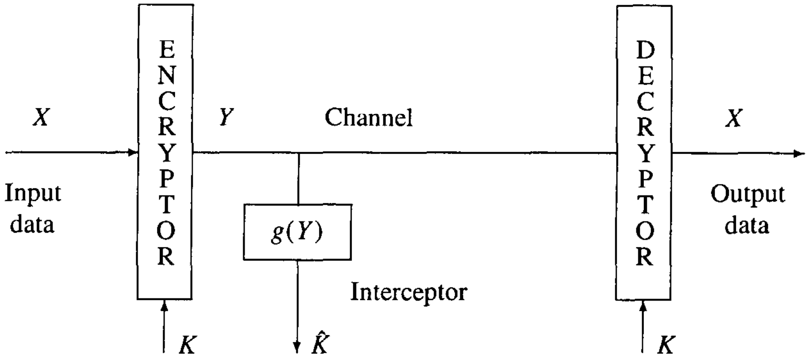

Fig. 1. A schematic of the enciphering process.

this result some preliminaries from information theory are required. This paper also provides a more general definition for the unicity distance. Based on this definition a simple proof of Hellman results is given. In Section 3, the main result of the paper is given where an upper bound on the success probability of an eavesdropper estimating the cipher secret key is derived. Finally, some conclusions are given in Section 4.

2. Preliminaries

A schematic of the enciphering system is shown in Fig. 1. Here, it is assumed that the eavesdropper receives the ciphertext with no errors. Let X, Y and K be random variables denoting the plaintext, ciphertext and the key and taking values in the sets $\mathcal{X}$, $\mathcal{Y}$ and $\mathcal{K}$, respectively. We assume that the cryptosystem is endomorphic, that is, the plaintext and the ciphertext message spaces are the same. In such case, $\mathcal{X}$ and $\mathcal{Y}$ have the same cardinality, or $|\mathcal{X}| = |\mathcal{Y}|$. The output of the encryptor can be expressed as a function of the input and the key, i.e.,

$$Y = f(X, K).$$

For a fixed key $k \in \mathcal{K}$, $f(-, k)$ is a one-to-one mapping of the input alphabet to the output alphabet. For best security, one would like to have a perfect secrecy system. For perfect secrecy the mutual information, $I(X;Y)$, must be zero [4,5]. This is equivalent to saying that Y is independent of X. To realize this, one needs a number of encryption keys greater than or equal to the number of possible messages. Such a requirement is not suitable for most applications. In practice a single secret key is repeatedly used to encipher a certain number of message symbols. This repetition in using the same key for enciphering more than one plaintext block leads to information leakage about the secret key through the ciphertext. It is, therefore, of practical interest to quantify this leakage. One typical measure of this leakage is the unicity distance of the cipher.

The unicity distance can be estimated using the concept of entropy introduced by Shannon. Let W be a discrete random variable defined over the set $\mathcal{W}$ with a probability distribution $P_W(w)$, $w \in \mathcal{W}$. The entropy, $H(W)$, of W is defined as

$$H(W) = \sum_{w \in \mathcal{W},\ P_W(w) \neq 0} P_W(w) \log \frac{1}{P_W(w)},$$

where the log is taken to base 2 [2]. Let Y^N denotes a sequence of N symbols from the cipher output. The uncertainty about the key given N ciphertext symbols is given by [5],

$$\begin{aligned} H(K/Y^N) &= H(KY^N) - H(Y^N) \\ &= H(KX^N) - H(Y^N) \qquad (1) \\ &= H(K) + H(X^N) - H(Y^N), \qquad (2) \end{aligned}$$

where the equality follows in (1) since X^N is a function of K and Y^N and in (2) because X^N and K are assumed independent. Assuming that X^N is an N-symbols segment from a stationary random process, the quantity $H(X^N)/N$, $N = 1, 2, \ldots$, is a nonnegative decreasing function of N and thus with a limit [1, p. 64]. Let this limit be $H_\infty(X)$. That is,

$$H_\infty(X) = \lim_{N\to\infty} \frac{H(X^N)}{N}.$$

From this it follows that

$$H(X^N) \geqslant NH_\infty(X), \quad N = 1,2,\ldots. \tag{3}$$

On the other hand,

$$H(Y^N) \leqslant N\log|\mathcal{Y}| = N\log|\mathcal{X}|, \tag{4}$$

with equality if and only if all the Y sequences of length N are equally probable. In such case, (2) can be rewritten as

$$\begin{aligned} H(K/Y^N) &\geqslant H(K) + NH_\infty(X) - H(Y^N) && (5)\\ &\geqslant H(K) + NH_\infty(X) - N\log|\mathcal{X}|, && (6)\end{aligned}$$

where the inequality in (5) follows from (3) and the inequality in (6) follows from (4). The inequality (6) is true for any secret key cipher as long as the input to the cipher is assumed to be a stationary process. The value of N that makes $H(K/Y^N)$ *approximately* zero is defined by Shannon as the unicity distance of the cipher.

One can, however, define a more general ε-unicity-distance, U_ε, as the minimum N such that $H(K/Y^N) \leqslant \varepsilon$ for some small ε. That is,

$$U_\varepsilon = \min\{N \mid H(K/Y^N) \leqslant \varepsilon\}. \tag{7}$$

Shannon's unicity distance, U, of a secret-key cipher then becomes

$$U = \lim_{\varepsilon\to 0} U_\varepsilon, \tag{8}$$

whenever the limit exists[1]. It follows from (7) and (2) that U_ε can be rewritten as

$$U_\varepsilon = \min\left\{N \;\middle|\; N \geqslant \frac{H(K) - \varepsilon}{H(Y^N)/N - H(X^N)/N}\right\}. \tag{9}$$

For the Shannon's random-cipher, the unicity distance is well approximated by

$$U \approx \left\lceil \frac{H(K)}{\log|\mathcal{X}| - H_\infty(X)} \right\rceil,$$

where $\lceil x \rceil$ denotes the smallest integer greater than or equal to x. Let

$$U_{\mathrm{R}}(X;K) = \left\lceil \frac{H(K)}{\log|\mathcal{X}| - H_\infty(X)} \right\rceil, \tag{10}$$

denote a general function for an arbitrary secret-key cipher. The semicolon is used here instead of a colon to emphasize the fact that this function is not explicit in X and K; this is in similarity to the convention used in defining the mutual information $I(X;Y)$. In what follows U_{R} will be used instead of $U_{\mathrm{R}}(X;K)$.

Because (6) is valid for a general secret key cryptosystem with stationary input, the following can be easily proven.

Theorem 1. *The unicity distance, U, of a general secret-key cryptosystem with stationary input satisfies*

$$U \geqslant U_{\mathrm{R}}.$$

The result asserts that, for a secret-key cipher with a stationary input, the unicity distance will always be greater than or equal to the value obtained from substituting the cipher parameters into Shannon's expression of the random cipher unicity distance.

3. An upper bound on the probability of successfully estimating a cipher secret key from N ciphertext symbols

The problem here is to find an upper bound on the probability that an eavesdropper, using the best estimation rule, successfully estimates the secret key from N intercepted ciphertext symbols where $N < U$. To solve this problem, we propose applying a tool from information theory; namely Fano's inequality [2]. Suppose Y^N is a known random sequence of length N symbols and we want to guess the value of a correlated random variable K. Fano's inequality relates the probability of error in guessing the random variable K to its conditional entropy $H(K/Y^N)$. This probability will be zero if and only if $H(K/Y^N)$ is zero or K is a function of Y^N. For this probability to be small, the conditional entropy $H(K/Y^N)$ must be small.

To estimate a discrete random variable K with a distribution $P_K(k)$, we first observe Y^N which is related to K by the conditional distribution $P_{Y^N/K}(y^n/k)$, then

[1] This implies that U could be infinite which is true for some ciphers when the condition $H(k/Y^N) = 0$ is strictly imposed. For other ciphers this limit exists. For the first case, however, one can use U_ε instead for some small ε.

the function $g(Y^N)$ is calculated (see Fig. 1) which corresponds to the estimate of the secret key K. Now, let $\hat{K}$ be the estimated value of the secret key K based on the observed N intercepted ciphertext symbols and let the probability of estimation error, $P_e(N)$, be

$$P_e(N) = \Pr(\hat{K} \neq K).$$

Theorem 2 (Fano's inequality).

$$h(P_e(N)) + P_e(N)\log(|\mathcal{K}| - 1) \geqslant H(K/Y^N), \tag{11}$$

where $h(p) = -p\log(p) - (1-p)\log(1-p)$ is the binary entropy function.

Because $h(P_e(N)) \leqslant 1$, this inequality can be slightly weakened to

$$P_e(N) \geqslant \frac{H(K/Y^N) - 1}{\log|\mathcal{K}|}. \tag{12}$$

For a proof of this inequality see [2].

The probability, $P_s(N)$, that an eavesdropper successfully estimates the secret key K based on the observing Y^N is

$$P_s(N) = 1 - P_e(N). \tag{13}$$

Theorem 3. *For a number, N, of intercepted symbols less than the unicity distance of a secret-key cipher, the probability that an eavesdropper successfully estimates the cipher secret key satisfies*

$$P_s(N) - \frac{h(P_s(N))}{\log(|\mathcal{K}| - 1)} \leqslant 1 - \frac{H(K/Y^N)}{\log(|\mathcal{K}| - 1)}, \tag{14}$$

which can be weakened to

$$P_s(N) \leqslant \min\Big(1, 1 + \frac{1}{\log|\mathcal{K}|} - \Big(1 - \frac{N}{U_R}\Big)\frac{H(K)}{\log|\mathcal{K}|}\Big). \tag{15}$$

Proof. The first part follows directly from (11) and (13). For the second part, it follows from (12) that

$$P_s(N) \leqslant 1 - \frac{H(K/Y^N) - 1}{\log(|\mathcal{K}|)}. \tag{16}$$

Substituting (6) in the above equation and simplifying, the following is obtained:

$$\begin{aligned} P_s(N) &\leqslant 1 - \frac{H(K) + H(X^N) - H(Y^N) - 1}{\log|\mathcal{K}|}, \\ &= 1 + \frac{1}{\log|\mathcal{K}|} - \Big(1 - \frac{N}{U_R}\Big)\frac{H(K)}{\log|\mathcal{K}|} \\ &\quad 0 \leqslant N < U_R\Big(1 - \frac{1}{H(K)}\Big). \\ &= \min\Big(1, 1 + \frac{1}{\log|\mathcal{K}|} - \Big(1 - \frac{N}{U_R}\Big)\frac{H(K)}{\log|\mathcal{K}|}\Big). \end{aligned}$$

□

For small N, one can get better bounds on $P_s(N)$ by using the different n-gram entropies of X if available. For practical applications, however, such entropies may not be available. As a worst case assumption, one may assume that all the source redundancy is available to the eavesdropper. In such situations the bound in (15) yields the best that an eavesdropper can achieve.

In most ciphers the key is usually selected uniformly from $\mathcal{K}$. In this case, the following corollary directly follows from Theorem 3.

Corollary 4. *If K is uniformly distributed over $\mathcal{K}$, then*

$$P_s(N) \leqslant \frac{N}{U_R} + \frac{1}{\log|\mathcal{K}|}, \quad 0 \leqslant N < U_R\Big(1 - \frac{1}{H(K)}\Big). \quad \square \tag{17}$$

The above inequality can be rewritten as

$$P_s(N) \leqslant \frac{N(\log|\mathcal{X}| - H_\infty(X)) + 1}{\log|\mathcal{K}|}.$$

Therefore, to assure a small probability of success by an eavesdropper, one needs to enlarge the key size and/or increase the source uncertainty by decreasing the quantity $\log|\mathcal{X}| - H_\infty(X)$. There are known techniques to perform this latter task. Examples include data compression, homophnic substitution and text padding [4]. If $\log|\mathcal{K}|$ is $\gg 1$, then the bound on $P_S(N)$ is approximately equal to N/U_R. This shows a linear relationship between the upper bound on $P_s(N)$ and the number of observations with a slope of $1/U$. On the other hand, if U_R is large, then, for small N, N/U_R will be small compared to the term $1/\log|\mathcal{K}|$ and this latter term will be dominant.

Table 1
The values of the two bounds on $P_S(N)$

N	Bound (14)	Bound (15)
1	0.03290	0.04186
2	0.06500	0.07240
3	0.09681	0.10294
4	0.12842	0.13348
5	0.15988	0.16403
6	0.19122	0.19457
7	0.22244	0.22511
8	0.25358	0.25566
9	0.28463	0.28620
10	0.31560	0.31674
11	0.34650	0.34729
12	0.37733	0.37783
13	0.40809	0.40837
14	0.43879	0.43891
15	0.46942	0.46946
16	0.49999	0.50000
17	0.53051	0.53054
18	0.56096	0.56109
19	0.59135	0.59163
20	0.62168	0.62217
21	0.65194	0.65271
22	0.68214	0.68326
23	0.71228	0.71380
24	0.74234	0.74434
25	0.77232	0.77489
26	0.80223	0.80543
27	0.83204	0.83597
28	0.86175	0.86652
29	0.89135	0.89706
30	0.92080	0.92760
31	0.95007	0.95814
32	0.97903	0.98869

To evaluate the above bounds consider the following example.

Example. For a simple substitution cipher with a 26 letter alphabet, the number of possible substitutions or keys is 26!. The unicity distance bound U_R of this cipher with a source entropy of 2 bits per symbol and uniform key distribution is 32. The bounds in (14) and (15) are evaluated for both bounds and the results are summarized in Table 1.

It can be seen from the table that the bound obtained from (14) is close to that obtained from (15). For cases where $\log|\mathcal{K}| \gg 1$ the bound given by (15) can be used as a reasonable approximate upper bound on P_S. For this cipher the term N/U_R is dominant in the bound, leading to an almost linear relation between the probability of success and the number of observed symbols.

4. Conclusions

In this paper the probability that an eavesdropper, using the best estimation rule, successfully estimates the secret key of a cipher from a number of observations less than the unicity distance is investigated. An upper bound on this probability is derived. It is found that this probability is inversely proportional to the logarithm of the key size and directly proportional to the redundancy in the source. Such theoretical bounds are useful in giving more insight to the fundamental limits on the performance of secret-key cryptosystems.

Acknowledgement

The author would like to express his sincere thanks to Professor James Massey (Swiss Federal Institute) for the many useful discussions regarding the subject of this paper.

References

[1] P. Beauchemin and G. Brassard, A generalization of Hellman's extension to Shannon's approach to cryptography, *J. Cryptology* (1988) 129–131.

[2] T. Cover and J. Thomas, *Elements of Information Theory* (John Wiley, new York, 1991).

[3] M. Hellman, An extension of Shannon theory approach to cryptography, *IEEE Trans. Inform. Theory* **23** (1977) 289–294.

[4] J. Massey, Contemporary cryptology: An introduction, in: G. Simmons, ed., *Contemporary Cryptology: The Science of Information Integrity* (IEEE Press, New York, 1992) 1–39.

[5] C. Shannon, Communication theory of secrecy systems, *Bell Systems Tech. J.* **28** (1949) 656–715.