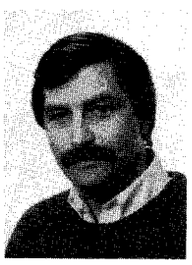
The Pe-security distance as a measure of cryptographic performance

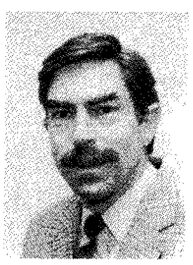
La distance de sécurité-Pe comme une mesure de performance cryptographique



Johan VAN TILBURG

Department of Applied Mathematics, PTT, Dr. Neher Laboratories, PO Box 421, 2260 AK LEIDSCHENDAM, THE NETHERLANDS.

Johan van Tilburg obtained his M.Sc. degree with honours from the Faculty of Electrical Engineering of the Delft University of Technology in 1985. Since 1985 he has been at the Dr. Neher Laboratories as a scientific staff member of the department of applied mathematics. His main interest is in cryptology. He is a member of the IACR.



Dick E. BOEKEE

Information Theory Group, Delft University of Technology, PO Box 5031, 2600 GA DELFT, THE NETHERLANDS.

Dick Boekee obtained his M.Sc. and Ph.D. degrees from the faculty of Electrical Engineering of the Delft University of Technology in 1970 and 1977 respectively. Since 1968 he has been with the Delft University of Technology where he is currently a professor in Information and Communication Theory. From 1977-1980 he has been a visiting professor at the Dept. of Mathematics, Katholieke Universiteit, Leuven, Belgium. His current interest are information theory, cryptology, picture coding and signal processing. He is a member of IACR, KIVI and NERG.

SUMMARY

The unicity distance is frequently rather losely used as a measure of the length of the cryptogram needed to break the enciphered text.

In this paper we discuss some aspects and details of this intuitive idea and introduce the Pe-security distance as a measure of cryptographic performance. In contrast to the classical unicity distance which uses the equivocation we consider the error probability as the fundamental parameter which has to be used. We also show that the classical unicity distance can be seen as a special case of the Pe-security distance.

KEY WORDS

Cryptography, unicity distance, security distance.

RÉSUMÉ

Le présent article est une contribution à la notion intuitive de la distance d'unicité considérée comme une mesure caractéristique de la longueur du texte nécessaire pour briser un cryptogramme.

On situe le travail dans le cadre d'une indice d'études sur la performance cryptographique en utilisant la distance de sécurité-Pe comme mesure de performance. Au contraire de la distance d'unicité utilisant l'équivocation, les auteurs traitent la probabilité d'erreur comme la variable fondamentale dans la partie performance.

On montre ensuite que la distance d'unicité classique peut être considérée comme un cas particulier de la distance de sécurité-Pe.

MOTS CLÉS

Cryptographie, distance d'unicité, distance de la sécurité.

1. Introduction

The use of cipher systems makes it possible to send secret messages via public insecure channels. However, the secrecy of the message depends highly on the cryptografic performance of the cipher system used. When evaluating the theoretical strength of cipher systems with a probabilistic model, it is assumed that the cryptanalyst behaves rationally, that he or she at least knows the set of transformations, the statistics of the message and the key source.

In Shannon's paper [1] it is pointed out that if the cryptanalyst intercepts a cryptogram, that he or she is able to calculate the *a posteriori* probabilities of the various possible messages and keys which might have produced this cryptogram. This set of *a posteriori* probabilities describes how the cryptoanalyst's knowledge of the message and the key gradually becomes more precise as more enciphered text is intercepted. Shannon used as a measure of theoretical strength the equivocation, which deals with a simplified description of the set of *a posteriori* probabilities. Shannon's approach has led to the so-called (classical) unicity distance and will be described in section 2.

Although Shannon's information measure leads to easy manipulation in a natural and intuitive way between different probability distributions, still the underlying relevant parameter is the error probability (or probability of incorrect identification) Pe faced by the cryptanalist.

In cases where determining the error probability in a direct manner is quite involved, bounds on Pe can be considered. By bounding Pe with information measures and/or distance measures, a region is determined in which the actual Pe can be found. The uncertainty in the value of Pe is resolved only in limiting cases where the bounds are tight. In this context it seems to be a natural way to make use of the concept of distance measures since the error probability is actually a distance measure itself. This approach, which can be found in Van Tilburg and Boekee [2], has led to the introduction of the Pe-security distance and will be described in section 3. Finally, in section 4 conclusions are drawn.

2. The classical unicity distance

As it appears from the literature, the Unicity Distance (UD) is often linked to the random cipher model and/or the key equivocation. As a result of this several authors have given definitions of the unicity distance which are incomplete, biased and more restrictive than necessary. As a consequence of this the UD is easily given a wrong interpretation. To clarify this confusion let us first consider the UD as derived by Shannon [1, p. 693]. Shannon defined the (classical) UD for the message based on a ciphertext-only-attack, $UD_{RC}(M^L/E^L)$, by evaluating the key equivocation and the key appearance characteristic in a Random Cipher (RC). As a result he obtained

$$(1) \qquad UD_{RC}(M^L/E^L) = H(K)/D(M^L),$$

where $D(M^L) = \log|\mathcal{M}| - H(M^L)/L$ is the average redundancy per message source symbol in a sequence M^L of L message source symbols, $H(K) = |\mathcal{K}|$ is the entropy of the key source and E^L is the enciphered message of length L. Unfortunately this UD is sometimes confused with the UD for the key based on a ciphertext-only-attack. It trivially holds that

$$(2) \qquad UD_{RC}(K/E^L) \geqq UD_{RC}(M^L/E^L),$$

so that (1) an (2) yields

$$(3) \qquad UD_{RC}(K/E^L) \geqq H(K)/D(M^L),$$

Hellman [3] has proved that the RC-model actually defines a lower bound on the existence of good cipher systems. For this reason (1) and (3) give a worst case indication of the strength of a cipher system. However, these results are not precise and the interpretation depends highly on the size of the key space used and the message source used.

Since $H(M^L/K, E^L) = 0$, it follows directly that a general relation between the key equivocation and the message equivocation is given by

$$(4) \qquad H(K/E^L) - H(M^L/E^L) = H(K/M^L E^L),$$

in which $H(K/M^L E^L)$ is the key appearance equivocation. The left hand side of the equality is based on a ciphertext-only-attack, while the right hand side is based on a knwon-plaintext-attack. Hence Dunham [4] concludes that there is a fundamental trade-off between protecting the key under a known-plaintext-attack and protecting the message under a ciphertext-only-attack when the size of the key space is fixed. And also, when designing a cipher system which is to be strong under a ciphertext-only-attack on the message, (4) suggests that it consequently will be weak under a known-plaintext-attack. From (4) it also follows that

$$(5) \qquad H(K/E^L) \geqq H(M^L/E^L)$$

and thus

$$(6) \qquad UD_M(K/E^L) \geqq UD_M(M^L/E^L),$$

with equality if the key appearance equivocation is zero, which is in agreement with (2), where M is the actual cipher model used. In general, the key equivocation is given by

$$(7) \qquad H(K/E^L) = H(KE^L) - H(E^L).$$

If the message source and the key source are stochastically independent then the key equivocation becomes

$$(8) \qquad H(K/E^L) = H(K) + H(M^L) - H(E^L).$$

Using the inequality $H(E^L) \leqq \log|\mathcal{E}^L| = L.\log|\mathcal{E}|$ and the fact that $|\mathcal{E}| = |\mathcal{M}|$, we easily find that $H(K/E^L)$ in (8) can be lower bounded by

$$(9) \qquad H(K/E^L) \geqq H(K) + H(M^L) - L.\log|\mathcal{E}| = H(K) - L.D(M^L).$$

If we define the unicity distance $UD(K/E^L)$ for the key based on a ciphertext-only-attack as the distance

where $H(K/E^L)$ is zero, then from (9) it follows that

(10) $$UD_M(K/E^L) \geqq H(K)/D(M^L).$$

It is tempting to say that the RC reaches this lower bound, i.e. $UD_M(K/E^L) \geqq UD_{RC}(K/E^L)$. The next Lemma may help to make this statement clear.

Lemma 1: *The average probability of error (or probability of incorrect key identification) in a random cipher model at classical unicity distance is given by*

(11) $$Pe_{RC}(K/E^{UD}) = (|\mathcal{K}| - 1)/|\mathcal{K}|^2.$$

Proof: Suppose that there are $|\mathcal{K}|$ different and independent keys in the RC so that $Pe_{RC}(K/E^L) = \bar{n}_k/|\mathcal{K}|$ in which $\bar{n}_k$ is the average number of spurious key decipherments. According to Hellman [3], Theorem 1, we have $\bar{n}_k = (|\mathcal{K}| - 1) \cdot 2^{-L \cdot D(M^L)}$. Substitution yields

(12) $$Pe_{RC}(K/E^L) = (1 - |\mathcal{K}|^{-1}) \cdot 2^{-L \cdot D(M^L)}.$$

At classical UD it holds that $L = H(K)/D(M^L)$. In addition, the keys are equiprobable so that $H(K) = \log|\mathcal{K}|$. Substitution yields the Lemma. □

Remark: It is important that the assumptions imposed by the RC-model be reasonable for the real secrecy system including the language used.

Lemma 1 tells us that the cryptanalyst is faced with an error probability (unequal to zero) at the classical UD. For this reason $H(K/E^{UD})$ can not be zero and the lower bound (10) does not hold in general. This also shows that Blom's general derivation [5], p. 9, of Hellman's result is not as general as suggested. Actually (10) is restricted to the limiting case where $H(K/E^L) = 0$ can be obtained.

Furthermore, it is illustrated in Van Tilburg, Boekee [2] why the key equivocation (7) itself, when considered as a measure of theoretical security, behaves poorly: it defines an upper bound on the error probability and is usually tight only for large L. Although the key equivocation may be a poor measure of theoretical security in many cases, it certainly does not degrade the use of Shannon's information measure in cryptanalysis. The strength of this measure can be explained by the natural interpretation and accordingly by the convenient way of manipulating between different probability distributions. This has been demonstrated by Lu [7].

Finally, to illustrate the difference between $UD(K/E^L)$ and $UD(M^L/E^L)$ with an extreme example we mention that for a simple substitution cipher using the English language we may obtain $UD(K/E^L) = 1{,}500$ and $UD(M^L/E^L) = 25$ respectively.

To understand the introduction and the interpretation of the Pe-security distance (Pe-SD) as a measure of cryptographic performance it is necessary to formalize the UD.

Definition 1: The unicity distance of a cipher model (including the message source) is the minimal expected length of ciphertext, generated by this model, after which the enciphered text (cryptogram) can be broken on the average. □

This definition of the UD covers at least five important aspects. The first one is that the UD is a minimal expected length. For an accurate interpretation of the UD it might be important to consider higher order statistics too. The second aspect follows from the fact that the cipher model includes the message source also. It is evident that the message source greatly influences the UD. Generally speaking, it is important to know the proces which has generated the enciphered text. The third aspect is inherently related to the plaintext, i.e. the text generated by the message source. If the plaintext is known, then we speak of a UD based on a known-plaintext-attack. If the plaintext is unknown, then we speak of a UD based on a ciphertext-only-attack. The fourth aspect has a strong affinity with the previous one. What is our object: the key or the message? As illustrated in section 2 they might be quite different. Finally, the fifth aspect and this might be the most important one: what is the meaning of "can be broken on the average".

3. The Pe-security distance

Most of the definitions in the open literature approach this problem by introducing the key equivocation and adverbs like almost and nearly. Jürgensen and Matthews [6] were the first who tried to improve the rigour of the notion of security in this respect by using a well-defined probabilistic model and based on this model addressed the problem by defining the β-UD as $MIN\{L \mid H(K/E^L) \leqq \beta\}$. However, as stated before the key equivocation defines an upper bound on the error probability and is usually only tight for large L. Consequently, this contradicts the "minimal expected length" in definition 1 and the worst case approach in general. Moreover the interpretation of the β-UD is not unique and depends highly on the size of the key space used. To avoid these problems one can link a probability function to "can be broken on the average".

For example, if the error probability (or probability of incorrect identification) faced by the cryptanalist is used, then the cryptogram space is divided into equivalence classes, one of which has a unique average error probability Pe for a given cipher model. If we do this, then it follows from (11) that the classicial UD is directly related to an average error probability which is inversely proportional to the cardinality of the key space used. As a result, the meaning of the UD for different sizes of the key space is also different, in the sense of Pe. Actually, that is not what one prefers. For this reason a constant average error probability is taken as a starting point and definition 1 can be restated as a security distance.

Definition 2: The Pe-security distance of a cipher model (including the message source) is the minimal expected length of the ciphertext, generated by this model, necessary in order to be able to break the enciphered text (cryptogram) with an average error probability (or probability of incorrect identification) of at most Pe. □

This definition provides a theoretically attractive measure of cryptographic performance of a cipher

system. In order to give a mathematically suitable definition it is necessary to restrict ourselves to a specific attack, for example as is done in the next definition [2, Definition 4.2].

Definition 3: The Pe-security distance for the key based on a ciphertext-only-attack is defined by

$$(13)\quad \gamma\text{-SD}(K/E^L) = \text{MIN}\{L \mid Pe_m(K/E^L) \leqq \gamma\}$$

where: m, is the actual cipher model, and, γ, is a value of the error probability Pe. □

Remark: Depending on what ones object is i.e. the key or the message, the Pe-SD can be based on $Pe_m(K/E^L)$ or on $Pe_m(M^L/E^L)$. If a known-plaintext-attack is used one may use $Pe_m(K/M^L, E^L)$. From the definition it also follows that the Pe-SD depends on the cipher model m used and the desirable value γ of Pe.

The next corollary [2, Corollary 4.2], shows that the Pe-security distance can be considered as a generalized unicity distance.

Corollary 1: *The Pe-security distance includes the classical unicity distance as a special case.*

Proof: For an RC-model we have (12):

$$Pe_{RC}(K/E^L) = (1 - |\mathcal{K}|^{-1}) \cdot 2^{-L \cdot D(M^L)}.$$

If we choose $\gamma = (|\mathcal{K}| - 1)/|\mathcal{K}|^2$, one easily obtains

$$\text{MIN}\{L \mid L \geqq H(K)/D(M^L)\},$$

which is the classical unicity distance. □

Whereas determining the error probability (and thus the Pe-SD) in a direct manner is quite involved, one can make use of lower bounds only. This is in agreement with the worst case approach. A natural way to obtain lower bounds is to make use of the concept of distance measures, as shown in Van Tilburg and Boekee [2], since the error probability is actually a distance measure itself.

For example, for a pure ciphermodel (PC) using a discrete memoryless source with *a priori* probabilities p and q it holds [2, theorem 2.3], that:

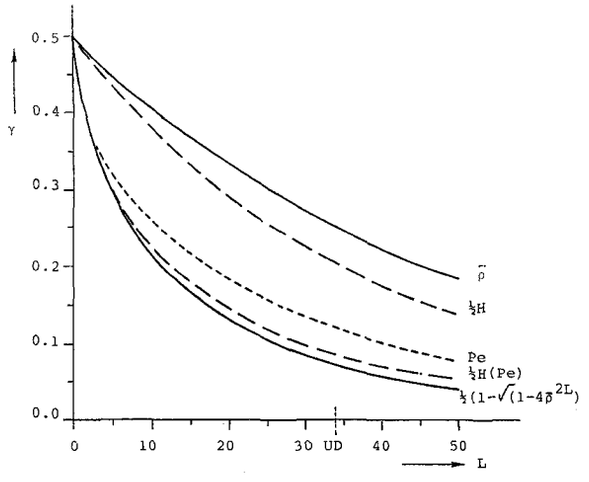
$$(14)\quad \frac{1}{2}(1 - \sqrt{(1 - 4\bar{\rho}^2)}) \leqq Pe_{PC}(K/E^L) \leqq \frac{1}{2}H(K/E^L) \leqq \bar{\rho}$$

where $\bar{\rho} = \frac{1}{2}(\sqrt{4pq})^L$ is the average Bhattacharyya coefficient. From definition 3 and the above inequality it is easily found that:

$$(15)\quad \gamma\text{-SD}_{PC}(K/E^L) \geqq \frac{\log[1 - (1 - 2\gamma)^2]}{\log(4pq)}.$$

However, if the key equivocation is used, than it follows that this measure defines an upper bound on the error probability. For this reason, Fano's inequality $H(K/E^L) \leqq H(Pe) + Pe \cdot \log(|\mathcal{K}| - 1)$ has to be used to obtain the lower bound.

The Pe-SD can be applied in the reverse direction too, i.e. for a given security distance the corresponding expected value γ can be found. By using the same arguments lower bounds on Pe can be considered in order to determine γ. To illustrate the behaviour of the different bounds discussed in this paper we consider a memoryless SSC-model with probability $p = 0.6$. The bounds given in Figure clearly demonstrate that the key equivocation ((1/2) H) is a loose upper bound on Pe at UD. Furthermore, observe that for $\gamma = 0.5$ the cryptanalysts decision is not based on any ciphertext at all ($L = 0$). This is in agreement with the best strategy, i.e. randomly select a key if no ciphertext is available.



Bounds on Pe for a memoryless SSC-model with $p = 0.6$.

In the case that $p = q = 1/2$ it easily follows from (14) that $Pe = 1/2$, and from (15) that $\gamma\text{-SD}_{PC}(K/E^L) \to \infty$ independent of the value of γ.

This is what we intuitively would expect. Since the message source symbols have a uniform probability distribution the redundancy of the message source is zero, so that every message, even the scrambled one, has a meaning. For this reason it does not matter how much ciphertext is intercepted; the cryptanalyst will never obtain a unique solution. On the other hand, if this is the case the cryptanalyst can at best select a key at random, conform (14).

4. Conclusion

Shannon obtained a unicity distance for the message based on a ciphertext-only-attack in a random cipher model, which is referred to as the classical unicity distance. Hellman has shown that Shannon's random cipher result actually defines a lower bound on the existence of good ciphers. Later on, Blom generalized this result in terms of key equivocation. However, Blom's result is not as general as suggested.

After formalizing the unicity distance a potential ambiguity can be found in most definitions in the literature. This ambiguity can be resolved by introducing a probability function.

A natural probability function is one based on the expected error probability (or probability of incorrect identification) faced by the cryptanalyst. As a direct result the Pe-security distance is introduced as the minimal expected amount of enciphered text necessary to make an average probability of incorrect identification of at most Pe.

Finally, if the expected error probability Pe is set equal to $(|\mathcal{K}|-1)/|\mathcal{K}|^2$, then the classical unicity distance is obtained, which shows that the Pe-security distance can be considered as a generalized unicity distance.

Manuscrit reçu le 1er *décembre* 1986.

REFERENCES

[1] C. E. SHANNON, Communication theory of secrecy systems, *Bell Syst. Tech. J.*, 28, 1949, pp. 656-715.

[2] J. VAN TILBURG and D. E. BOEKEE, *Divergence bounds on Key Equivocation and Error probability*, in *Advances in Cryptology-Crypto* 85, Springer Verlag, Berlin, 1986, pp. 489-513.

[3] M. E. HELLMAN, An Extension of the Shannon Theory Approach to Cryptography, *IEEE Trans. Inform. Theory*, IT-23, 1977, pp. 289-294.

[4] J. G. DUNHAM, Bounds on Message Equivocation for Simple Substitution Ciphers, *IEEE Trans. Inform. Theory*, IT-26, 1980, pp. 522-527.

[5] R. BLOM, Bounds on Key Equivocation for Simple Substitution Ciphers, *IEEE Trans. Inform. Theory*, IT-25, 1979, pp. 8-18.

[6] H. JÜRGENSEN an D. E. MATTHEWS, Some result of the information theoretic analysis of cryptosystems, *Proceedings of Crypto*'83, Santa Barbara, California, August 1983, pp. 303-356.

[7] S. C. LU, The existence of good cryptosystems for key rates greater than the message redundancy, *IEEE Trans. Inform. Theory*, IT-25, 1979, pp. 475-477.