

Watermarking Security Part One: Theory

François Cayre^a, Caroline Fontaine^b, and Teddy Furon^{a *}

^a INRIA, TEMICS project, Rennes, France

^b CNRS, LIFL, Université des sciences et des technologies de Lille, France

ABSTRACT

This article proposes a theory of watermarking security based on a cryptanalysis point of view. The main idea is that information about the secret key leaks from the observations, for instance watermarked pieces of content, available to the opponent. Tools from information theory (Shannon's mutual information and Fisher's information matrix) can measure this leakage of information. The security level is then defined as the number of observations the attacker needs to successfully estimate the secret key. This theory is applied to common watermarking methods: the substitutive scheme and spread spectrum based techniques. Their security levels are calculated against three kinds of attack.

Keywords: Watermarking, Security, Equivocation, Fisher information matrix.

1. INTRODUCTION

Digital watermarking studies have always been driven by the improvement of *robustness*. Most of articles of this field deal with this criterion, presenting more and more impressive experimental assessments. Some key events in this quest are the use of spread spectrum [1], the invention of resynchronization schemes [2], the discovery of side information channel [3,4], and the formulation of the opponent actions as a game [5].

On the contrary, *security* received little attention in the watermarking community. The first difficulty is that security and robustness are neighboring concepts, which are hardly perceived as different. The intentionality behind the attack is not enough to make a clear cut between these two concepts. An image compression is clearly an attack related to robustness, but it might happen intentionally, *i.e.* with the purpose of removing the watermark, or not. *Robust* watermarking is defined in [6] as a communication channel multiplexed into original content in a non-perceptible way, and whose “*capacity degrades as a smooth function of the degradation of the marked content*”. We add that the degradation is due to a classical content processing (compression, low-pass filtering, noise addition, geometric attack ...). The attacker has three known strategies to defeat watermark robustness: to remove enough watermark signal energy, to jam the hidden communication channel, or to desynchronize the watermarked content.

T. Kalker then defines watermarking *security* as “*the inability by unauthorized users to access [i.e. to remove, to read, or to write the hidden message] the communication channel*” established by a robust watermarking. Security deals with intentional attacks whose aims are not only the removal of the watermark signal, excluding those already encompassed in the robustness category since the watermarking technique is assumed to be robust.

Some seminal works have already warned the watermarking community that digital watermarking may not be a secure primitive (*i.e.*, a tool providing information security) despite its robustness. However, they only deal with dedicated attacks relevant to particular applications. The deadlock attack concerns copyright protection and illustrates the impossibility to prevent somebody to watermark content with his own technique and key (by embedding a watermark signal or by creating a fake original) [7]. This ruins the identification of the owner because two watermarking channels interfere in the same piece of content. Multiple problems in the field of copyright protection and authentication stems from the copy attack, where the attacker first copies a watermark and then pastes it in a different piece of content [8].

We do not include these two last attacks in our study because they pertain to the protocol layer, in the sense that it questions the link between the presence (or absence) of watermark and the signification at the application

* Author names appear in alphabetical order. Contact Information: teddy.furon@irisa.fr

The work described in this paper has been supported in part by the French Government through the ACI Fabriano, and by the European Commission through the IST Programme under Contract IST-2002-507932 ECRYPT.

layer. We believe that these attacks stem from a misunderstanding of the watermark designers about the targeted application. In copyright protection, the presence of a watermark has no legal value. The only receivable proof is the belonging of the content to the database of a trusted third party (*ie.*, an author society). The authors must register their works in this database in order to be protected. It is absolutely useless, from a legal point of view, for the authors to watermark their works on their own. If watermarking is used in copyright protection, it will be embedded by the trusted third party during the registration process. Moreover, it will certainly be a function of a registration number. We show here that the deadlock attack is technically feasible, but it has no reality once one knows the framework of the application. In the same way, the copy attack is now a nonsense in authentication application. It is true that the very first watermarking authentication schemes were using a constant watermark. But, nowadays, it is well established that the watermark must depend on the original content like a digital signature in cryptography.

We are more interested here in threats decoupled from the applications. The oracle attack (aka sensitivity attack) is a threat whenever the opponent has access to a watermarking detector (as in copy protection for consumer electronics devices [9], for instance). The attacker first estimates the secret key, testing the detection process on different pieces of content [10]; this disclosure then helps him forging pirated content. Note that in this last case, the number of detection tries is of utmost importance. The watermark designer would like this number to be so huge that the attack lasts too much time. For instance, around N_v tries are necessary for a Direct Spread Spectrum Sequence watermarking scheme [10], whereas around N_v^2 tries are needed for known asymmetric schemes [11].

Articles proposing a complete analysis of robust watermarking security are extremely rare. The authors are only aware of the pioneer work [12], where two digital modulation schemes achieve perfect secrecy, and more recent works sketching a general framework for security analysis [11, 13]. The main idea is here to adapt Shannon's definition of cryptography security to watermarking. At the beginning of the game, the watermarker selects a watermarking technique and picks up randomly a private key. According to the Kerckhoffs's principle, the opponent knows the selected algorithm but not the private key. Then, the watermarker starts producing some marked pieces of content. The opponent has access to some observations and his aim is to estimate the private key. The main idea of Shannon's theory is that information about the private key might leak from the observations. Hence, the *a posteriori* uncertainty of the opponent decreases as he makes more and more observations. However, the above-mentioned works have only translated the cryptanalysis methodology into watermarking terminology.

The goal of this article is to offer a complete and workable theory of watermarking security. It completes Barni's *et al.* approach, assessing for the really first time security levels of substitution, and spread spectrum based techniques. For this purpose, the first section summarizes the methodology and introduces the basic notation. Measurement of the information leakages are based on Shannon's mutual information for a substitutive watermarking method in section 3 and on Fisher's information for a spread spectrum based watermarking method in section 4. This measure is also used for SCS analysis. This yields estimation of security levels for three types of attack. Yet, these information theory tools do not reveal any insight for practical hacking. Part Two tackles this algorithmic issue.

2. METHODOLOGY

2.1. Notation

Let us first list some notational conventions used in this paper. Vectors are sets in bold font, matrices in calligraphic font, and sets in black board font. Data are written in small letters, and random variables in capital ones. The length of the vectors considered in this paper is N_v : $x(i)$ is the i -th component of vector \mathbf{x} . The probability density function of random variable \mathbf{X} (or its probability mass function if \mathbf{X} is discrete) is denoted by $p_{\mathbf{X}}(\cdot)$. Hidden messages have N_c bits and secret keys are usually composed of N_c elements (*e.g.* N_c secret carriers in the spread spectrum case). Finally, N_o vectors are considered: $\mathbf{x}^{N_o} = \{\mathbf{x}_j\}_{j=1}^{N_o}$ represent this collection of vectors and \mathbf{x}_j is the vector \mathbf{x} associated to the j -th observation.

2.2. The cryptanalytic approach

The methodology presented in this section is clearly inspired by the cryptanalysis. It has already been presented in [13], and is based on three key articles: Kerckhoffs [14], Shannon [15] and Diffie-Hellman [16]. We first briefly present these concepts, before formalizing them in the following subsections.

2.2.1. Kerckhoffs' principle.

A. Kerckhoffs stated in 1883 that keeping an encryption algorithm secret for years is not realistic [14], and this principle is now used in any cryptanalytic study. In watermarking, the situation is similar, and it is assumed that the opponent knows the watermarking algorithm. Hence, for a given design and implementation of an algorithm, the security stems from the secrecy of the key. The designer's challenge is: "Am I sure that an opponent will not exploit some weaknesses of the algorithm to disclose the secret key?". In practice, it doesn't mean that watermark designers must disclose their algorithm. It only says that the secrecy by obscurity (non disclosing the algorithm as a defense against hacking) cannot be measured by any means, and it is consequently non reliable. When an expert assesses the security level of a scheme (be it a crypto-system or watermarking technique), he plays the role of a pirate who has somehow disclosed the algorithm.

What does Kerckhoffs' principle imply? Watermarking processes are often split into three functions. The first one extracts some features from content (issued by a classical transform, such as DCT, wavelet, FFT, Fourier Mellin, . . .), which are stored in a so-called extracted vector. The second one mixes the extracted vector with the secret watermark signal, giving a watermarked vector. Then, an insertion function reverses the extraction process to come back in the original world, putting out the watermarked document. Fig. 1 illustrates the embedding process. The detection follows an analogous process as sketched in Fig. 2. According to the Kerckhoffs' principle, the opponent knows all the involved functions. He thus observes the watermarked vectors from contents he has access to, because the extraction function has no secret parameter.

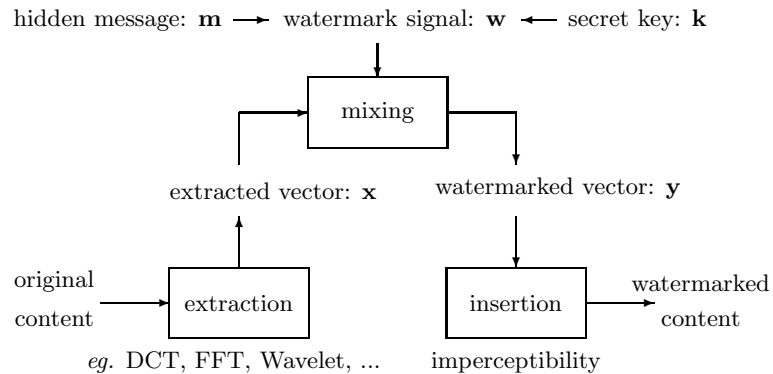


Figure 1. Global point of view of the embedding process

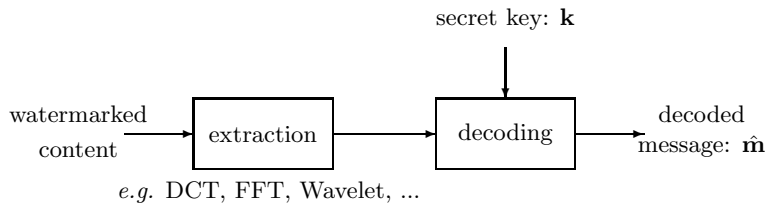


Figure 2. Global point of view of the detection process

2.2.2. Shannon’s approach

The methodology that Shannon exposed for studying the security of encryption schemes is here transposed to watermarking. The embedder has randomly picked up a secret key, and he used it to watermark several pieces of content. The opponent observes these pieces of watermarked content, all related to the same secret key but hiding different messages. The watermarking technique is *perfectly secure* if and only if no information about the secret key leaks from the observations. If it is not the case, the *security level* is defined as the number of observations which are needed to disclose the secret key, or to estimate it with enough accuracy. The bigger the information leakage is, the smaller the security level of the watermarking scheme will be.

2.2.3. Diffie-Hellman’s terminology

Reference [16] is one of the most well known articles in cryptography as it strikes the creation of new directions such as public key cryptography and digital signature. It is also, as far as the authors know, the first time where several contexts of attack are envisaged according to the kinds of data observed by the opponent. In watermarking, the adversary has at least access to watermarked content, but, in some cases, he might also observe the hidden messages (for instance, the name of the author in copyright protection or the status of a movie in copy protection) or the original data (for instance, DVD movies are watermarked for copy protection; but original version of old movies were not protected). This implies that a security level is assessed for a given context. In this article, we study:

- the Watermarked Only Attack (WOA), in which the opponent only has access to N_o watermarked vectors \mathbf{y}^{N_o} ;
- the Known Message Attack (KMA), in which the opponent only has access to N_o watermarked vectors and the associate messages $(\mathbf{y}, \mathbf{m})^{N_o}$;
- the Known Original Attack (KOA), in which the opponent only has access to N_o watermarked vectors and the corresponding original ones $(\mathbf{y}, \mathbf{x})^{N_o}$.

The reader might be surprised that the KOA context deserves any attention. Seemingly, there is no need to attack watermarked content when one has the original version. The pirate does not hack these pieces of content, but his goal is to gain information about the secret key, in order to, later on, hack different pieces of content watermarked with the same key.

Other contexts, not studied in this article, will certainly deserve a proper study in the future.

- the Estimated Original Attack, in which the opponent has access to original content but at a lower quality than the watermarked versions. Are small pictures in thumbnail gallery or movies trailers watermarked? Another possibility is that the opponent denoizes watermarked content to estimate its original version.
- the Multiple Embedding Single Original Attack, in which the opponent has access to several watermarked versions of the same content with different hidden messages. Collusion in fingerprinting and tracing traitors applications is tackled here. However, the collusion attack (*ie.*, the process made by the group of colluders) is not reduced here to a simple average of the multiple watermarked version. A proper study must reveal whether a more powerful attack exists in order to assess the security level of fingerprinting schemes.
- the Multiple Embedding Multiple Original Attack, in which the opponent has access to several watermarked versions (each) of some originals. Fingerprinting of movies (*ie.*, several video blocks) is tackled here.

2.3. Perfect covering

Although cryptographic encryption and watermarking are two different security primitives, they might look like the same at first sight. Fig. 3 illustrates this analogy investigated in this subsection. Shannon defined *perfect secrecy* of a crypto-system by the inability of opponents to refine the probability distribution of plaintexts \mathbf{m} by observing related cipher texts, all encrypted by key \mathbf{k} . We adapt this definition to watermarking, stating that

of the equivocation as the remaining uncertainty does not hold when the secret key is regarded as a continuous random variable as in section 4. For instance, the equivocation can take positive or non positive values, ruining the concept of unicity distance.

2.4.2. Fisher's measure

This is the reason why another information measurement is proposed. In statistics, Fisher was one of the first to introduce the measure of the amount of information supplied by the observations about an unknown to be estimated parameter. Suppose observation \mathbf{O} is a random variable with a probability distribution function depending on a parameter vector $\boldsymbol{\theta}$. The *Fisher Information Matrix* (FIM) concerning $\boldsymbol{\theta}$ is defined as

$$\text{FIM}(\boldsymbol{\theta}) = E\boldsymbol{\psi}\boldsymbol{\psi}^T \quad \text{with} \quad \boldsymbol{\psi} = \nabla_{\boldsymbol{\theta}} \log p_{\mathbf{O}}(\mathbf{o}; \boldsymbol{\theta}), \quad (2)$$

where E is the mathematical expectation operator and $\nabla_{\boldsymbol{\theta}}$ is the gradient vector operator defined by $\nabla_{\boldsymbol{\theta}} = (\partial/\partial\theta(1), \dots, \partial/\partial\theta(N_{\theta}))^T$. The Cramér-Rao theorem gives a lower bound of the covariance matrix of an unbiased estimator of parameter vector $\boldsymbol{\theta}$ whenever the FIM is invertible:

$$\mathcal{R}_{\hat{\boldsymbol{\theta}}} \geq \text{FIM}(\boldsymbol{\theta})^{-1}, \quad (3)$$

in the sense of non-negative definiteness of the difference matrix. In our framework, the parameter vector can be the watermark signal or the secret key. (3) provides us a physical interpretation: the bigger the information leakage is, the more accurate the estimation of the secret parameter might be.

The FIM is also an additive measure of the information, provided the observations are statistically independent. For instance, suppose that the watermark signal has been added in N_o pieces of content whose extracted vectors are independent and identically distributed as $\mathbf{X} \sim \mathcal{N}(\mathbf{0}, \mathcal{R}_{\mathbf{X}})$. The observations are N_o watermarked signals. Then,

$$\log p_{\mathbf{O}}(\mathbf{o}; \mathbf{w}) = -1/2 \sum_{j=1}^{N_o} (\mathbf{y}_j - \mathbf{w})^T \mathcal{R}_{\mathbf{X}}^{-1} (\mathbf{y}_j - \mathbf{w}) + \text{const}, \quad (4)$$

$$\boldsymbol{\psi} = -1/2 \sum_{j=1}^{N_o} \mathcal{R}_{\mathbf{X}}^{-1} (\mathbf{y}_j - \mathbf{w}) = -1/2 \sum_{j=1}^{N_o} \mathcal{R}_{\mathbf{X}}^{-1} \mathbf{x}_j, \quad (5)$$

$$\text{FIM}(\mathbf{w}) = N_o/4 \mathcal{R}_{\mathbf{X}}^{-1} E\{\mathbf{x}_j \mathbf{x}_j^T\} \mathcal{R}_{\mathbf{X}}^{-1} = N_o/4 \mathcal{R}_{\mathbf{X}}^{-1}. \quad (6)$$

This models applications which detect presence of (and not decode) watermarks, or also template signals which resynchronize content transformed by a geometric attack.

The mean square error $E\{\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}\|^2\}$ is the trace of $\mathcal{R}_{\hat{\boldsymbol{\theta}}}$, and thus its lower bound decreases in N_o^{-1} . However, the rate $N_o^* = N_o \text{tr}(\text{FIM}(\boldsymbol{\theta})^{-1})$ depends on the statistical model and consequently the kind of observations (see section 4). It means that the estimation is significantly more accurate when the number of independent observations increases of an order of N_o^* . The bigger N_o^* , the more difficult is the disclosure of the secret key. This notion is close to the unicity distance of the above subsection. This is the reason why we use the same notation N_o^* (although absolutely not defined in the same way).

3. SECURITY ANALYSIS OF THE SUBSTITUTIVE METHOD

3.1. Mathematical model

In such a scheme, a binary vector $\mathbf{x} = (x(1) \dots x(N_v))^T$ is extracted from the content. For instance, in the famous Burgett, Koch, and Zao technique [17], N_v pairs of DCT coefficients of an image are compared in absolute value. The message to be hidden is a binary vector $\mathbf{m} = (m(1) \dots m(N_c))^T$. The secret key is a list of N_c integers $\mathbf{k} = [k(1), \dots, k(N_c)]$ with $1 \leq k(\ell) \leq N_v$ and $k(\ell) \neq k(\ell')$ if $\ell \neq \ell'$. The embedding process copies \mathbf{x} in \mathbf{y} and then substitutes the $k(\ell)$ -th bit of \mathbf{y} by the ℓ -th bit of the message to be hidden: $y(k(\ell)) = m(\ell)$. The inverse extraction function maps back the watermarked vector \mathbf{y} into the content. The decoding simply reads the bits whose indices are given by the secret key.

EXAMPLE 1. $N_v = 8$ and $N_c = 4$:

$$\begin{aligned} \mathbf{m} &= (1101) & \mathbf{k} &= [2, 8, 5, 3] \\ \mathbf{x} &= (01001011) & \mathbf{y} &= (01100011) \end{aligned}$$

The uncertainty of the opponent is given by the entropy of the secret key that the embedder has randomly selected among $N_v!/(N_v - N_c)!$ possible keys. Thus:

$$H(\mathbf{K}) = \log_2 \frac{N_v!}{(N_v - N_c)!} \quad (7)$$

3.2. Perfect covering

THEOREM 3.1. *As defined above, a substitutive watermarking scheme provides perfect covering.* Proof: We can model the substitutive watermarking as follows: let \mathbf{X} be a binary N_v -length random vector, whose probability mass function is uniform and equal to 2^{-N_v} , and \mathbf{W} be a binary N_v -length vector whose bits equal to 1 indicates the bits to be flipped. For a given secret key, these ‘flipping’ bits are located at the same indices. If the message to be hidden is a uniformly distributed random variable, \mathbf{W} is finally independent from \mathbf{Y} . Hence, we have $\mathbf{Y} = \mathbf{X} \oplus \mathbf{W}$, giving:

$$\begin{aligned} p_{\mathbf{Y}}(\mathbf{y}) &= \sum_{\mathbf{w} \in \mathbb{W}} p_{\mathbf{Y}|\mathbf{W}}(\mathbf{y}|\mathbf{w})p_{\mathbf{W}}(\mathbf{w}) = \sum_{\mathbf{w} \in \mathbb{W}} p_{\mathbf{X}}(\mathbf{y} \oplus \mathbf{w})p_{\mathbf{W}}(\mathbf{w}) \\ &= 2^{-N_v} \sum_{\mathbf{w} \in \mathbb{W}} p_{\mathbf{W}}(\mathbf{w}) = 2^{-N_v}, \\ p_{\mathbf{Y}}(\mathbf{y}|\mathbf{w}) &= p_{\mathbf{X}}(\mathbf{y} \oplus \mathbf{w}) = 2^{-N_v}. \end{aligned}$$

The Bayes rule, $p_{\mathbf{Y}}(\mathbf{y}|\mathbf{w})p_{\mathbf{W}}(\mathbf{w}) = p_{\mathbf{W}}(\mathbf{w}|\mathbf{y})p_{\mathbf{Y}}(\mathbf{y})$, then gives $p_{\mathbf{W}}(\mathbf{w}) = p_{\mathbf{W}}(\mathbf{w}|\mathbf{y})$.

3.3. Watermarked Only Attack

The substitutive method providing perfect covering, it is then very easy to show that $I(\mathbf{Y}; \mathbf{W}) = 0$, which implies that $I(\mathbf{Y}; \mathbf{K}) = 0$. There is no information leakage, and the equivocation is equal to $H(\mathbf{K})$ whatever the number of observations. In a way, one can say that security level $N_o^* = +\infty$.

However, note the utmost importance of the random messages assumption. If the message to be hidden is fixed and \mathbf{X} a random vector uniformly distributed, then the bits of \mathbf{Y} stuck to the same value will disclose the positions selected by the secret key. This might, for instance, indicate a security threat in fingerprinting (the hidden message is the serial number of a buyer) of videos (a succession of several original video blocks, almost independent) with a substitutive watermarking method.

3.4. Known Message Attack

If the opponent observes only one watermarked content \mathbf{y}_1 and its hidden message \mathbf{m}_1 , the indices i such that $y_1(i) = m_1(\ell)$ are possible values of $k(\ell)$. Denote $\mathbb{S}_1(\ell)$ this set. As $P(y_1(i) = m_1(\ell)|i \neq k(\ell)) = 1/2$, there are in expectation $1 + (N_v - 1)/2$ elements in this set.

Now assume that the opponent observes several contents \mathbf{y}^{N_o} and their hidden messages \mathbf{m}^{N_o} . Set $\mathbb{S}_{N_o}(\ell)$ is now defined by $\mathbb{S}_{N_o}(\ell) = \{i : y_j(i) = m_j(\ell) \forall j, 1 \leq j \leq N_o\}$. The probability that $y_j(i) = m_j(\ell) \forall j$ knowing that $i \neq k(\ell)$ is $1/2^{N_o}$. Thus, in expectation, $|\mathbb{S}_{N_o}(\ell)| = 1 + (N_v - 1)/2^{N_o}$, and the equivocation about $k(\ell)$ is equal to $\log_2(1 + 2^{-N_o}(N_v - 1))$. However, there might be some overlapping between the N_c sets $\mathbb{S}_{N_o}(\ell)$, and the total equivocation is smaller than the sum of the equivocations about $k(\ell)$. As the calculus is quite complex, we stay with this approximation:

$$H(\mathbf{K}|\mathbf{Y}, \mathbf{M})^{N_o} \lesssim N_c \log_2(1 + 2^{-N_o}(N_v - 1)). \quad (8)$$

Shannon approximated this equivocation by $N_c(\log_2(N_v - 1) - N_o)$ when $N_o \ll \log_2(N_v - 1)$, and by $2^{-N_o} N_c(N_v - 1)/\log(2)$ when $N_o \gg \log_2(N_v - 1)$ (see Fig. 4). He also approximated the unicity distance by $N_o^* = \log_2 N_v$ [15, Sect. 14].

3.5. Known Original Attack

If the opponent observes only one watermarked content \mathbf{y}_1 and its original version \mathbf{x}_1 , the indices i such that $x_1(i) \neq y_1(i)$ are possible values for the key samples. There are in expectation $N_c/2$ of such indices, as $p(x_1(k(\ell)) = m_1(\ell)) = 1/2$. When the opponent observes j pairs, the set $\mathbb{S}_j = \{\ell : \exists j', 1 \leq j' \leq j, x_{j'}(\ell) \neq y_{j'}(\ell)\}$ grows up. However, the event that an index revealed by a new pair was already known happens with a probability $|\mathbb{S}_{j-1}|/N_c$. This leads to the following series:

$$|\mathbb{S}_j| = |\mathbb{S}_{j-1}| + N_c(1 - |\mathbb{S}_{j-1}|/N_c)/2 = N_c(1 - 2^{-j}). \quad (9)$$

Yet, it is not possible to assign a key sample to one of these indices. The equivocation is then the sum of two terms: one is due to the $N_c - |\mathbb{S}_{N_o}|$ undisclosed indices to be picked up randomly among the remaining candidates, the second one is due to the $N_c!$ possible permutations of the chosen indices:

$$H(\mathbf{K} | (\mathbf{Y}, \mathbf{X})^{N_o}) = \log_2 \left(\frac{(N_v - \lceil |\mathbb{S}_{N_o}| \rceil)!}{(N_v - N_c)! (N_c - \lceil |\mathbb{S}_{N_o}| \rceil)!} \right) + \log_2(N_c!). \quad (10)$$

The security level (in the unicity distance sense) is not defined as the equivocation is always greater than zero. This is due to the term $\log_2(N_c!)$ reflecting the ambiguity in the order of the estimated key samples. We preferably consider that within a number of observations greater than $N_o^* = \log_2 N_c$, the opponent learns all the indices store in the secret key. This information is helpful for watermark jamming. He can also notice if two hidden messages are the same. Yet, the ambiguity prevents him reading the hidden messages (he cannot put the hidden bits in the right order), and writing hidden messages.

Fig. 4 gives a good synthesis of the results. In the WOA case, the opponent cannot get any information on the key, and then cannot do anything. In the KMA case, he is able to completely disclose the key, and then he will be able to read, erase, write or modify hidden messages. In the KOA case, he is able to recover the components of the key but up to a permutation, and then he will be able to erase the hidden message, but not to read or write a proper one.

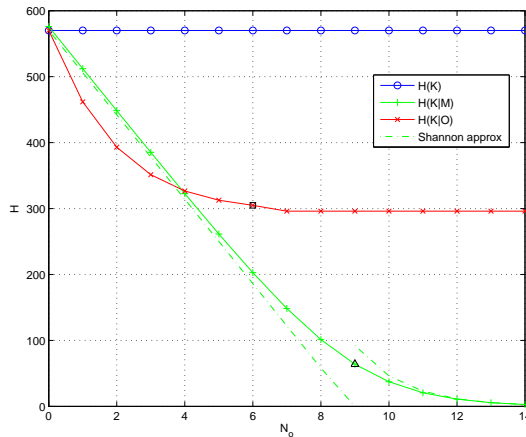


Figure 4. Substitutive watermarking: equivocations for WOA, KMA and KOA, against the number of observations. $N_c = 64$, $N_v = 512$. The triangle and the square respectively mark the security levels for the KMA and KOA.

4. SECURITY ANALYSIS OF SPREAD SPECTRUM BASED TECHNIQUES

Spread spectrum is a military communication scheme invented during World War II [18]. It was designed to be good at combatting interference due to jamming, hiding a signal by transmitting it at low power, and achieving secrecy. These properties make spread spectrum very popular in nowadays digital watermarking. Theoretical studies [5] and practical implementations [19] focus on the optimization of operational capacity-robustness functions for a given embedding distortion.

4.1. Mathematical model

Denote by \mathbf{x} a vector of N_v samples extracted from original content. The embedding is the addition of the watermark signal which is the modulation of N_c private carriers \mathbf{u}_ℓ :

$$\mathbf{w} = \frac{\gamma}{\sqrt{N_c}} \sum_{\ell=1}^{N_c} a(\ell) \mathbf{u}_\ell, \quad (11)$$

where $\gamma > 0$ is a small gain fixing the embedding strength, and $\|\mathbf{u}_\ell\| = 1$, $1 \leq \ell \leq N_c$. The Watermark to Content power Ratio (WCR) equals $\gamma^2 \sigma_a^2 / \sigma_x^2$ (or $10 \log_{10}(\gamma^2 \sigma_a^2 / \sigma_x^2)$ if expressed in dB). An inverse extraction function puts back vector $\mathbf{y} = \mathbf{x} + \mathbf{w}$ into the media to produce the watermarked content.

Symbol vector \mathbf{a} represents the message to be hidden/transmitted through content. In the case of a Direct Sequence Spread Spectrum (DSSS), the modulation is a simple BPSK: $a(\ell) = (-1)^{m(\ell)}$, $1 \leq \ell \leq N_c$ and $\sigma_a^2 = 1$. Yet, the scope of this model is far broader than the sole case of DSSS. Spread spectrum is a very common process used to increase the signal to noise ratio by projecting signals on a smaller subspace of dimension $N_c < N_v$. This also covers some side-informed watermarking techniques (sometimes called spread transform) [4,20–22]. Symbols $a(\ell)$ are then continuous real values (see Part Two).

For security reason, the carriers are private and issued by a pseudo-random generator fed by a seed. Many people think the secret key is the seed. This is not false as the disclosure of the seed obviously gives the carriers and allows the access to the watermarking channel. However, the knowledge of the carriers is sufficient and the pirate has no interest in getting back to the seed. Hence, in this article, the secret key, defined as the object the opponent is keen on revealing, is constituted by the carriers.

In the sequel, the security analysis considers several watermarked vectors \mathbf{y}_j , $1 \leq j \leq N_o$, with different embedded symbols $\mathbf{a}_j = (a_j(1) \dots a_j(N_c))^T$ being linearly mixed by the $N_v \times N_c$ matrix $\mathcal{U} = (\mathbf{u}_1 \dots \mathbf{u}_{N_c})$. To cancel inter-symbol interferences at the decoding side, carriers are two-by-two orthogonal vectors: $\mathcal{U}^T \mathcal{U} = \mathcal{I}_{N_c}$, where \mathcal{I}_N is the $N \times N$ identity matrix. Index i denotes the i^{th} samples of a signal, whereas j indices the different signals. Thus, there are N_o watermarked vectors given by:

$$\mathbf{y}_j = \mathbf{x}_j + \frac{\gamma}{\sqrt{N_c}} \mathcal{U} \mathbf{a}_j, \quad (12)$$

or, equivalently, concatenating N_o vectors \mathbf{x}_j (resp. \mathbf{y}_j or \mathbf{a}_j) column-wise in the $N_v \times N_o$ matrix \mathcal{X} (resp. \mathcal{Y} or the $N_c \times N_o$ matrix \mathcal{A}):

$$\mathcal{Y} = \mathcal{X} + \frac{\gamma}{\sqrt{N_c}} \mathcal{U} \mathcal{A}. \quad (13)$$

4.2. Perfect covering

Assume that $\mathbf{X} \sim \mathcal{N}(\mathbf{0}, \mathcal{R}_{\mathbf{X}})$ and that \mathbf{w} is picked up randomly among sequences distributed as $\mathcal{N}(\mathbf{0}, \mathcal{R}_{\mathbf{w}})$. Then, $p_{\mathbf{Y}} = \mathcal{N}(\mathbf{0}, \mathcal{R}_{\mathbf{X}} + \mathcal{R}_{\mathbf{w}})$ and $p_{\mathbf{Y}|\mathbf{w}=\mathbf{w}} = \mathcal{N}(\mathbf{w}, \mathcal{R}_{\mathbf{X}})$. The Bayes rule shows that spread spectrum based watermarking does not provide perfect covering. Even if the attacker has only access to watermarked pieces of content, some information about the watermark signal is leaking from these observations. The following subsections investigate whether the opponent can, thanks to this leakage on the watermark signal, gain some knowledge about the secret carriers.

4.3. Known Message Attack

In this subsection, the opponent has access to (watermarked signals/hidden messages) pairs. Moreover, only the DSSS technique (*i.e.*, a BPSK modulation) is considered. Our attack may not work with side information embedding because the opponent still ignores symbols \mathbf{a} , as they also depend on the original signal. Formally, the observations considered in this subsection are $(\mathbf{y}, \mathbf{a})^{N_o}$.

Assume, for simplicity reason, that each occurrence of random vector \mathbf{X} is independently drawn from $\mathcal{N}(\mathbf{0}, \sigma_x^2 \mathcal{I}_{N_v})$. The following theoretical derivations can be adapted to colored original signals and even non stationary original signals [23]. Another motivation is that, according to the Power Spectrum Constraint [24],

watermark signals usually adopt the statistical structure of host signals in order to increase their robustness, *i.e.* $\mathcal{R}_{\mathbf{w}} = \gamma^2 \mathcal{R}_{\mathbf{x}}$. Hence, the Karhunen-Loève Transform simultaneously whitens both signals.

The likelihood is the probability of observing the data \mathbf{y}^{N_o} , while knowing the model:

$$L(\mathbf{y}^{N_o}) = \frac{1}{(\sqrt{2\pi}\sigma_x)^{N_o N_v}} e^{\left(-\frac{1}{2\sigma_x^2} \sum_{j=1}^{N_o} \|\mathbf{y}_j - \frac{\gamma}{\sqrt{N_c}} \mathcal{U} \mathbf{a}_j\|^2\right)}, \quad (14)$$

and the log-likelihood is $\log L = K - \frac{1}{2\sigma_x^2} \sum_{j=1}^{N_o} \|\mathbf{y}_j - \frac{\gamma}{\sqrt{N_c}} \mathcal{U} \mathbf{a}_j\|^2$. The opponent wants to estimate the private carriers \mathbf{u}^{N_c} . So, the derivative implied in the FIM is $\boldsymbol{\psi} = \partial \log L / \partial (\mathbf{u}_1^T \dots \mathbf{u}_{N_c}^T)^T$ with

$$\frac{\partial \log L}{\partial \mathbf{u}_\ell} = \frac{\gamma}{\sigma_x^2 \sqrt{N_c}} \sum_{j=1}^{N_o} a_j(\ell) \mathbf{x}_j. \quad (15)$$

The expectation of the products gives the following $N_v \times N_v$ sub-blocks:

$$E \left(\frac{\partial \log L}{\partial \mathbf{u}_\ell} \right) \left(\frac{\partial \log L}{\partial \mathbf{u}_k} \right)^T = \frac{\gamma^2}{N_c \sigma_x^2} (\mathcal{F}_{uu})_{\ell,k} = \frac{\gamma^2}{N_c \sigma_x^2} \sum_{j=1}^{N_o} a_j(\ell) a_j(k) \mathcal{I}_{N_v}. \quad (16)$$

The FIM is then the following block matrix:

$$\text{FIM} = \frac{\gamma^2}{N_c \sigma_x^2} \begin{bmatrix} (\mathcal{F}_{uu})_{1,1} & \dots & (\mathcal{F}_{uu})_{1,N_c} \\ \vdots & & \vdots \\ (\mathcal{F}_{uu})_{N_c,1} & \dots & (\mathcal{F}_{uu})_{N_c,N_c} \end{bmatrix} = \frac{\gamma^2}{N_c \sigma_x^2} \mathcal{F}_{uu} \xrightarrow{N_o \rightarrow +\infty} N_o \frac{\gamma^2 \sigma_a^2}{N_c \sigma_x^2} \mathcal{I}_{N_v N_c}. \quad (17)$$

With a BPSK modulation, $\sigma_a = 1$. The information leakage is linear with the number of observations, thanks to the assumption of independence, and the rate is given by the Watermark to Content power Ratio per carrier $\gamma^2 / N_c \sigma_x^2$. The security level of spread spectrum based watermarking techniques against KMA is $N_o^* = N_c \sigma_x^2 / \gamma^2$ of (watermarked signals/hidden messages) pairs.

4.4. Known Original Attack

The opponent observes $(\mathbf{y}, \mathbf{x})^{N_o}$. The vector difference of each observation j gives the source signals \mathbf{a}_j being linearly mixed by the $N_v \times N_c$ matrix \mathcal{U} :

$$\mathbf{d}_j = \mathbf{y}_j - \mathbf{x}_j = \frac{\gamma}{\sqrt{N_c}} \mathcal{U} \mathbf{a}_j. \quad (18)$$

Assume that $N_o \geq N_c$ and that there are at least N_c linearly independent messages. The difference matrix $\mathcal{D} = \mathcal{Y} - \mathcal{X} \propto \mathcal{U} \mathcal{A}$ is then full rank, and $\text{Span}(\mathcal{D}) = \text{Span}(\mathcal{U})$. The observation of difference vectors discloses the secret subspace $\text{Span}(\mathcal{U})$, provided symbol matrix \mathcal{A} is full rank. However, this doesn't reveal the private carriers. Denote by \mathcal{E} a matrix whose columns constitute an orthonormal basis of the subspace $\text{Span}(\mathcal{D})$. We have $\mathcal{E} = \mathcal{U} \mathcal{P}^T$, with \mathcal{P} a unitary $N_c \times N_c$ matrix. *A priori*, there is no reason for which $\mathcal{P} = \mathcal{I}_{N_c}$. Hence, decoding the symbols with matrix \mathcal{E} gives the following mixture $\mathbf{v} = \sqrt{N_c} \mathcal{E}^T \mathbf{d} / \gamma = \mathcal{P} \mathbf{a}$. This is a blind source separation (BSS) problem with a square mixing matrix. Comon proved that it is possible to identify \mathcal{P} (and thus \mathcal{U}), but up to a permutation and scale ambiguity, only if at most one source is Gaussian [25]. The scale ambiguity is indeed a sign ambiguity in our problem, as we set $\mathcal{U}^T \mathcal{U} = \mathcal{I}$. In conclusion, at best, the mixing matrix is identified by $\hat{\mathcal{U}} = \Pi \Sigma \mathcal{U}$ with Π a permutation matrix and Σ a diagonal matrix whose elements are ± 1 . At best for the opponent, the secret carriers are identified up to a signed permutation (*i.e.*, matrix $\Pi \Sigma$) ambiguity.

The likelihood to observe \mathbf{v} for a given matrix \mathcal{P} is $p(\mathbf{v}; \mathcal{P}) = |\det \mathcal{P}|^{-1} p_{\mathbf{A}}(\mathcal{P}^{-1} \mathbf{v})$, and its score is:

$$\frac{\partial}{\partial \mathcal{P}} \log p(\mathbf{v}; \mathcal{P}) = -\mathcal{P}^{-T} + \mathcal{P}^{-T} \boldsymbol{\chi}(\mathcal{P}^{-1} \mathbf{v}) \mathbf{v}^T \mathcal{P}^{-T}, \quad (19)$$

with $\chi(\mathbf{x}) = -\frac{\partial}{\partial \mathbf{x}} \log p_{\mathbf{A}}(\mathbf{x})$ [26]. The asymptotic accuracy of the estimations is known to be only dependent on the symbols distribution, and especially on its non-Gaussianity. As, in our case, symbols are i.i.d., denote by $\chi(\cdot)$ the score function of $a_j(i)$, and by $\chi_n(\cdot)$ the score function of a Gaussian random variable sharing the same variance (*i.e.*, $\chi_n(x) = x/\sigma_a^2$). The trace of the Cramér-Rao Bound is then shown to be proportional to $(g^{-1} + 1/2)/2N_o$ for large N_o [27], with g defined as:

$$g = \frac{E\{(\chi(a) - \chi_n(a))^2\}}{E\{\chi_n(a)^2\}}. \quad (20)$$

However, g is not above bounded and tends to $+\infty$ when the symbols tend to have a discrete or bounded support. This is typically the case in watermarking, as the embedder would not allow the use of unbounded symbols for a perceptual distortion reason. In the case of discrete symbols, error free mixing matrix recovery is possible within a finite number of observations. For instance, [28] shows a workable algorithm needing $N_o > N_c^2$ observations for BPSK symbols. In the case of bounded support symbols, the trace of CRB decreases at a faster rate than $1/N_o$ [27, 29].

4.5. Watermarked Only Attack

In this section, the sources are unknown and can then be regarded as nuisance parameters [30, 31]. Vector ψ equals then $\partial \log L / \partial (\mathbf{u}_1^T \dots \mathbf{u}_{N_c}^T \mathbf{a}_1^T \dots \mathbf{a}_{N_o}^T)^T$, with the following $N_c \times 1$ vectors:

$$\frac{\partial \log L}{\partial \mathbf{a}_j} = \frac{\gamma}{\sigma_x^2 \sqrt{N_c}} \mathcal{U}^T \mathbf{x}_j \quad \forall j \in \{1, \dots, N_o\}. \quad (21)$$

The expectations of the products give the following sub-blocks:

$$\begin{aligned} E \left(\frac{\partial \log L}{\partial \mathbf{a}_j} \frac{\partial \log L}{\partial \mathbf{a}_k} \right) &= \frac{\gamma^2}{N_c \sigma_x^2} (\mathcal{F}_{aa})_{j,k} = \frac{\gamma^2}{N_c \sigma_x^2} \mathcal{I}_{N_c} \delta_{j,k} \\ E \left(\frac{\partial \log L}{\partial \mathbf{u}_\ell} \frac{\partial \log L}{\partial \mathbf{a}_j} \right) &= \frac{\gamma^2}{N_c \sigma_x^2} (\mathcal{F}_{ua})_{\ell,j} = \frac{\gamma^2}{N_c \sigma_x^2} (\mathcal{F}_{au})_{j,\ell}^T, \end{aligned}$$

where $\delta_{i,j}$ is the Kronecker function. We write with explicit notation:

$$\text{FIM} = \frac{\gamma^2}{N_c \sigma_x^2} \begin{bmatrix} \mathcal{F}_{uu} & \mathcal{F}_{ua} \\ \mathcal{F}_{au} & \mathcal{F}_{aa} \end{bmatrix}. \quad (22)$$

Note that $\mathcal{F}_{aa} = \mathcal{I}_{N_o N_c}$. The Cramér-Rao Bound for estimated $\text{Vect}(\mathcal{U}) = (\mathbf{u}_1^T, \dots, \mathbf{u}_{N_c}^T)^T$ is $\text{CRB}(\text{Vect}(\mathcal{U})) = \frac{N_c \sigma_x^2}{\gamma^2} \tilde{\mathcal{F}}_{uu}^{-1}$, with $\tilde{\mathcal{F}}_{uu} = (\mathcal{F}_{uu} - \mathcal{F}_{ua} \mathcal{F}_{aa}^{-1} \mathcal{F}_{au}) = (\mathcal{F}_{uu} - \mathcal{F}_{ua} \mathcal{F}_{au})$. It is known that, in the general case, $\tilde{\mathcal{F}}_{uu}^{-1} \geq \mathcal{F}_{uu}^{-1}$ (*i.e.* $\tilde{\mathcal{F}}_{uu}^{-1} - \mathcal{F}_{uu}^{-1}$ is non negative definite). In other words, nuisance parameters render the estimation of \mathcal{U} less accurate [26]. But, the situation is even worse here as the FIM becomes singular. Indeed:

$$(\mathcal{F}_{ua} \mathcal{F}_{au})_{\ell,k} = \sum_{j=1}^{N_o} (\mathcal{F}_{ua})_{\ell,j} (\mathcal{F}_{au})_{j,k} = \sum_{j=1}^{N_o} a_j(\ell) a_j(k) \mathcal{U} \mathcal{U}^T, \quad (23)$$

therefore $\tilde{\mathcal{F}}_{uu} = \mathcal{A} \mathcal{A}^T \otimes (\mathcal{I}_{N_o} - \mathcal{U} \mathcal{U}^T)$. As $(\mathcal{I}_{N_o} - \mathcal{U} \mathcal{U}^T) \mathbf{u}_k = \mathbf{0}$, $\tilde{\mathcal{F}}_{uu}$ is singular.

This problem stems from two facts. First, we did not integrate some constraints during our derivation. Especially, we know that $\mathbf{u}_\ell^T \mathbf{u}_k = \delta_{\ell,k}$. [30] gives an alternative expression for the bound in the case where the unconstrained problem is unidentifiable and the FIM non invertible.

However, the integration of the above-mentioned constraints in the derivation of the FIM is not sufficient for $N_c > 1$. The second fact is that an ambiguity remains about the order and ‘phase’ of the carriers. The system is only identifiable up to a signed permutation. The case $N_c = 1$ is interesting, as constraint integration removes the FIM singularity because the ambiguity of the permutation does not exist.

4.5.1. One carrier

The parameter vector to be estimated is composed of the unique carrier and the hidden symbols as nuisance parameters: $(\mathcal{U}^T \mathcal{A})$. Please, note that \mathcal{U}^T and \mathcal{A} are row vectors in this case. The constraint on \mathbf{u}_1 is: $(\|\mathbf{u}_1\|^2 - 1)/2 = 0$. The sequel is only the strict application of [30]. The $1 \times (N_v + N_o)$ gradient matrix of the constraint is equal to $\mathcal{G} = (\mathbf{u}_1^T \mathbf{0}_{N_o}^T)$, where $\mathbf{0}_N$ is a N zero vector. There exists a matrix $\mathcal{H} \in \mathbb{R}^{(N_v+N_o) \times (N_v+N_o-1)}$ whose columns form a basis for the nullspace of \mathcal{G} , that is, such that $\mathcal{G}\mathcal{H} = \mathbf{0}$. In our case, one particular choice of \mathcal{H} is readily verified to be:

$$\mathcal{H} = \begin{bmatrix} \mathcal{U}^\perp & \mathbf{0} \\ \mathbf{0} & \mathcal{I}_{N_o} \end{bmatrix}, \quad (24)$$

with \mathcal{U}^\perp being a basis of the complementary subspace of $\text{Span}(\mathbf{u}_1)$ in \mathbb{R}^{N_v} . Then, according to [30, Th. 1], the Cramér-Rao Bound under the above-mentioned constraint is $\text{CRB}(\mathcal{U}^T \mathcal{A}) = \mathcal{H}(\mathcal{H}^T \text{FIM} \mathcal{H})^{-1} \mathcal{H}^T$. With our choice of \mathcal{H} , this yields:

$$\text{CRB}(\mathcal{U}^T \mathcal{A}) = \frac{\sigma_x^2}{\gamma^2} \begin{bmatrix} (\mathcal{A}\mathcal{A}^T)^{-1} \mathcal{U}^\perp \mathcal{U}^{\perp T} & \mathbf{0} \\ \mathbf{0} & \mathcal{I}_{N_o} \end{bmatrix}, \quad (25)$$

and we finally get:

$$\text{CRB}(\mathcal{U}^T) = \frac{\sigma_x^2}{\gamma^2} (\mathcal{A}\mathcal{A}^T)^{-1} \mathcal{U}^\perp \mathcal{U}^{\perp T} \xrightarrow{N_o \rightarrow +\infty} \frac{\sigma_x^2}{N_o \sigma_a^2 \gamma^2} \mathcal{U}^\perp \mathcal{U}^{\perp T}. \quad (26)$$

4.5.2. N_c carriers ($N_c > 1$)

The ambiguity renders the Fisher Information Matrix singular, even when considering the constraints. However, Part Two shows that, in practice, the opponent builds noisy estimation of the carriers up to a signed permutation. A possibility in [31], is to pretend that the opponent knows N_m messages (for instance $\{\mathbf{a}_\ell\}_{\ell=1}^{N_m}$), in order to *artificially* remove the ambiguity. This adds $N_m N_c$ constraints of the type: $\hat{a}_j(\ell) = a_j(\ell)$. At the end, calculation leads to:

$$\text{CRB}(\text{Vect}(\mathcal{U})) = \frac{N_c \sigma_x^2}{\gamma^2} \mathcal{H}_{uu} \mathcal{B}^{-1} \mathcal{H}_{uu}^T, \quad (27)$$

with \mathcal{B} the $N_c(N_v - N_m) \times N_c(N_v - N_m)$ matrix whose $(N_v - N_m) \times (N_v - N_m)$ blocks are $(\mathcal{B})_{\ell,k} = (\mathcal{A}\mathcal{A}^T)_{\ell,k} \mathcal{U}_\ell^\perp \mathcal{U}_k^\perp - (\mathcal{A}_{N_m:N_o} \mathcal{A}_{N_m:N_o}^T)_{\ell,k} \mathcal{U}_\ell^{\perp T} \mathcal{U}_k^{\perp T}$, and \mathcal{H}_{uu} the $N_c N_v \times N_c(N_v - 1)$ diagonal matrix whose $N_v \times (N_v - 1)$ blocks on diagonal are $(\mathcal{H}_{uu})_{\ell,\ell} = \mathcal{U}_\ell^\perp$. In these expressions, the columns of \mathcal{U}_ℓ^\perp form an orthonormal basis of the complementary subspace of $\text{Span}(\mathbf{u}_\ell)$, and $\mathcal{A}_{N_m:N_o} = (\mathbf{a}_{N_m+1} \dots \mathbf{a}_{N_o})$. However, the minimal number N_m to remove the ambiguity depends on the symbols' pdf [31].

Facing the difficulty of finding the right parameter N_m and the cumbersome calculus, we prefer to approximate the information leakage about a carrier by (26), where γ^2 is replaced by the power per carrier γ^2/N_c . The security level is then $N_o^* = N_c \sigma_x^2 / \sigma_a^2 \gamma^2$ which is, by the way, coherent with (27). This result is quite surprising because the security level is the same against KMA and WOA. Yet, the estimation of the secret carriers remains up to a signed permutation in the WOA.