Watermarking security: theory and practice

Teddy Furon, François Cayre, Caroline Fontaine

Watermarking Security: Theory and Practice

François Cayre, Caroline Fontaine, and Teddy Furon

Author names appear in alphabetical order.

F. Cayre and T. Furon are with INRIA in TEMICS project. TEMICS / IRISA, Campus de Beaulieu, 35042 Rennes cedex, France. tel: +33 2 99 84 71 98. fax: +33 2 99 84 71 71. email: francois.cayre@irisa.fr, teddy.furon@irisa.fr.

C. Fontaine is with CNRS. LIFL, Université des sciences et des technologies de Lille. 59655 Villeneuve d'Ascq cedex, France. tel: +33 3 28 77 85 69. fax: +33 28 77 85 37. email: caroline.fontaine@lifl.fr.

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

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 two common watermarking methods: the substitutive scheme and the spread spectrum based techniques. Their security levels are calculated against three kinds of attack. The experimental work illustrates how Blind Source Separation (especially Independent Component Analysis) algorithms help the opponent exploiting this information leakage to disclose the secret carriers in the spread spectrum case. Simulations assess the security levels derived in the theoretical part of the article.

Index Terms

Watermarking, Security, Equivocation, Fisher information matrix, Blind source separation.

I. 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], [3], the discovery of side information channel [4], [5], and the formulation of the opponent actions as a game [6].

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 [7] 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) [8]. This ruins the identification of the owner because two watermarking channels interfere in the same piece of content. The collusion attack (*i.e.*, the mixing of several watermarked versions of the same content) is related to the fingerprinting application. 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 [9]. The oracle attack is a threat whenever the opponent has access to a watermarking detector as in copy protection for consumer electronics devices [10]. The attacker first estimates the secret key, testing the detection process on different pieces of content [11]; this disclosure then helps him forging pirated content. The number of detection tries is here of utmost importance.

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 [13], [14]. 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 secret key. According to the Kerckhoffs's principle, the opponent knows the selected algorithm but not the secret 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. Shannon's main idea 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, especially, spread spectrum based watermarking methods. 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 III and on Fisher's information for a spread spectrum based watermarking method in section IV. This yields estimation of security levels for three types of attack. Yet, these information theory tools do not reveal any insight for practical hacking of spread spectrum based watermarking. Section V tackles this algorithmic issue. Tools from the blind source separation (BSS) field appear to be extremely helpful for the attacker, especially Principal Component Analysis (PCA) and Independent Component Analysis (ICA).

II. METHODOLOGY

A. 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 $p_{\mathbf{X}}(.)$. Hidden messages have N_c bits and secret keys are usually composed of N_c elements, *e.g.* several carriers: $\mathbf{u}_\ell$ the ℓ-th carrier. Finally, N_o vectors are considered: $\mathbf{x}^{N_o}$ represent this collection of vectors and $\mathbf{x}_j$ is the vector $\mathbf{x}$ associated to the j-th observation.

B. The cryptanalytic approach

The methodology presented in this section is clearly inspired by the cryptanalysis. It has already been presented in [14], and is based on three key articles: Kerckhoffs [15], Shannon [16] and Diffie-Hellman [17]. We first briefly present these concepts, before formalizing them in the following subsections.

Kerckhoff's principle. It has been stated in 1883 that keeping an encryption algorithm secret for years is not realistic, and this principle is now used in any cryptographic 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?". 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

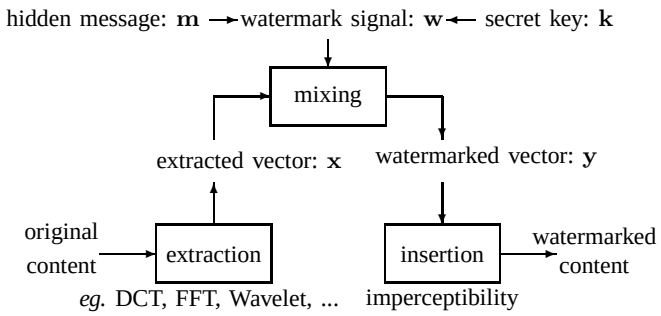

Fig. 1. Global point of view of the embedding process

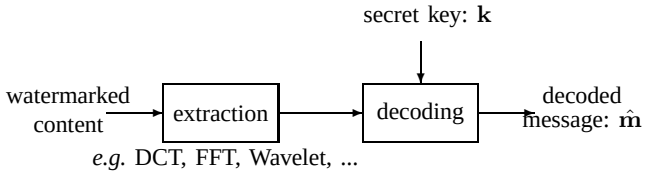

Fig. 2. Global point of view of the detection process

embedding process. The detection follows an analogous process as sketched in Fig. 2. According to the Kerckhoff's 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.

Shannon's approach. The methodology for studying the security of encryption schemes is here transposed to watermarking. The embedder has randomly picked up a secret key, and 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. The bigger the information leakage is, the smaller the security level of the watermarking scheme will be.

Diffie-Hellman's terminology. According to the context of the attack, the opponent may have access to several kinds of data. The opponent 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

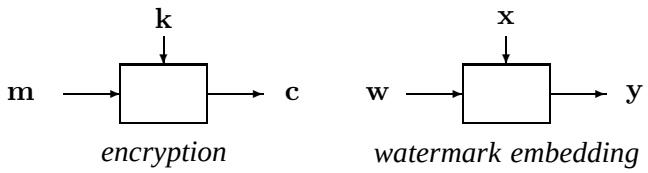

Fig. 3. An analogy with cryptography: plaintext $\mathbf{m} \rightarrow$ watermark $\mathbf{w}$, key $\mathbf{k} \rightarrow$ original $\mathbf{x}$, ciphertext $\mathbf{c} \rightarrow$ watermarked content $\mathbf{y}$.

the status of a movie in copy protection) or to the original data (for instance, imagine DVD movies are watermarked for copy protection; 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 N_o watermarked vectors $\mathbf{y}^{N_o}$;
- the Known Message Attack (KMA), in which the opponent only has 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 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.

C. Perfect covering

Although 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 the most important thing to be hidden is the watermark signal, and not the original content. The equivalent of the plaintext is, here, the watermark signal.

Definition 1: A watermark embedding makes a *perfect covering* if $p_{\mathbf{W}}(\mathbf{w}) = p_{\mathbf{W}}(\mathbf{w}|\mathbf{y})$ for any $(\mathbf{y}, \mathbf{w})$.

This means that in a perfect covering scheme, the observations of only watermarked pieces of contents will never reveal any information on the watermark signal: $I(\mathbf{Y}; \mathbf{W}) = 0$. If $\mathbf{K} \rightarrow \mathbf{W} \rightarrow \mathbf{Y}$ is a Markov chain, $I(\mathbf{Y}; \mathbf{W}) \geq I(\mathbf{Y}; \mathbf{K})$ holds. Consequently, perfect covering implies perfect secrecy.

Shannon easily found a necessary condition to get perfect secrecy, by using his information theory tools: $H(\mathbf{M}) \leq H(\mathbf{K})$, where $H(.)$ denotes the entropy, that is, $H(\mathbf{M}) = -\sum_{\mathbf{m}} p(\mathbf{m}) \log p(\mathbf{m})$. Yet, the same proof yields the following necessary condition to get perfect covering: $H(\mathbf{W}) \leq H(\mathbf{X})$. This deeply reveals the difference between cryptography and watermarking. As suggested by the greek word $\kappa\rho\upsilon\pi\tau\omega$ (meaning "I hide"), the role of the secret key is, in encryption, to hide the meaning of the plaintext. Hence, its entropy should be greater or equal to the one of the plaintext. Whereas steganography ($\sigma\tau\epsilon\gamma\alpha\nu\omega$ means "I cover") hides the watermark covered by the host signal.

D. Information leakages and physical interpretation

If a watermarking scheme does not provide perfect secrecy, then one would like to measure the information leakage on the secret key. For this purpose, this subsection presents several tools from information theory, which will later be useful to analyze classical watermarking schemes.

1) Shannon's measure: In the case where the secret key $\mathbf{K}$ is a discrete variable, and more usually a binary word, the entropy $H(\mathbf{K})$ measures the uncertainty of the opponent on the true value of $\mathbf{k}$. When he makes some observations[1] $\mathbf{O}^{N_o}$, his uncertainty is now evaluated through a conditional entropy, which Shannon named *equivocation*: $H(\mathbf{K}|\mathbf{O}^{N_o}) = H(\mathbf{K}) - I(\mathbf{K}; \mathbf{O}^{N_o})$. The information leakage is measured by the mutual information between the observations and the secret key. The bigger the information leakage

[1] *e.g.* observations can be "cipher texts", or "pairs of plain/cipher texts".

is, the smaller the uncertainty of the opponent is. Equivocation is a non increasing function with N_o. It goes from $H(\mathbf{K})$, ideally down to 0. When it becomes null, this means that the opponent has enough observations to uniquely determine the secret key. Shannon defined the *unicity distance* the first value of N_o for which the equivocation becomes null, , meaning that the set of all possible keys is now reduced to only one element. This is a way to measure the security level $N_o^\star$ of a primitive.

Unfortunately, these tools are not suitable for any watermarking scheme. It is well known that entropy (or conditional entropy) of a continuous random variable does not measure a quantity of information. Mutual information $I(\mathbf{K};\mathbf{O}^{N_o})$ is always pertinent as a measure of information leakages; but the physical interpretation of the equivocation as the remaining uncertainty does not hold when the secret key is regarded as a continuous random variable as in section IV. For instance, the equivocation can take positive or non positive values, ruining the concept of unicity distance.

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 parameter to be estimated. 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

$$\text{FIM}(\boldsymbol{\theta}) = E\boldsymbol{\psi}\boldsymbol{\psi}^T \quad \text{with} \quad \boldsymbol{\psi} = \nabla_{\boldsymbol{\theta}} \log p_{\mathbf{O}}(\mathbf{o};\boldsymbol{\theta}), \tag{1}$$

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

$$\mathcal{R}_{\hat{\boldsymbol{\theta}}} \geq \text{FIM}(\boldsymbol{\theta})^{-1}, \tag{2}$$

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. (2) provides us a physical interpretation: the bigger the

information leakage is, the more accurate the estimation of the secret parameter is.

The FIM is also an additive measure of the information, provided the observations are statistically independent. 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, $\log p_{\mathbf{O}}(\mathbf{o}; \mathbf{w}) = -1/2 \sum_{j=1}^{N_o} (\mathbf{y}_j - \mathbf{w}) \mathcal{R}_{\mathbf{X}}^{-1} (\mathbf{y}_j - \mathbf{w})^T + const.$ Calculation readily gives $\text{FIM}(\mathbf{w}) = N_o \mathcal{R}_{\mathbf{X}}^{-1}$. 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 $N_o^\star = N_o \text{tr}(\text{FIM}(\boldsymbol{\theta})^{-1})$ depends on the statistical model and consequently the kind of observations (see section IV). The estimation is significantly more accurate when the number of independent observations increases of an order of $N_o^\star$. The bigger $N_o^\star$, 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 $N_o^\star$ (although absolutely not defined in the same way).

III. Security analysis of the substitutive method

A. 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 [18], 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 $N_c = 4$:

$$\mathbf{m} = (1101) \qquad \mathbf{k} = [2,8,5,3]$$

$$\mathbf{x} = (01001011) \qquad \mathbf{y} = (0\mathbf{11}0\mathbf{0}11\mathbf{1})$$

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:

$$H(\mathbf{K}) = \log_2 \frac{N_v!}{(N_v - N_c)!} \tag{3}$$

B. Perfect covering

Theorem 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. Hence, we have $\mathbf{y} = \mathbf{x} \oplus \mathbf{w}$, giving:

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

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 $p_{\mathbf{W}}(\mathbf{w}) = p_{\mathbf{W}}(\mathbf{w}|\mathbf{y})$.

C. 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 $N_o^\star = +\infty$.

D. 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.

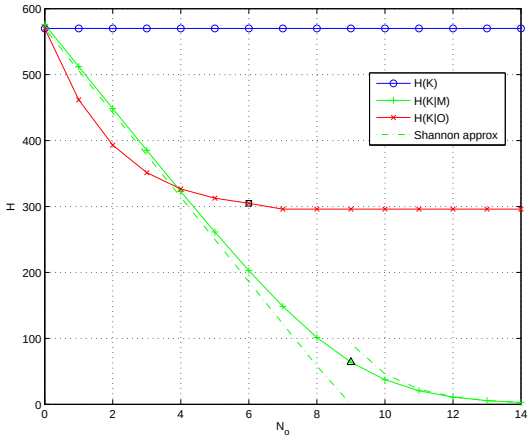

Fig. 4. Substitutive watermarking: equivocations for WOA, KMA and KOA, against the number of observations. $N_c = 64$, $N_v = 512$. The triangle and the square respectively mark the security levels for the KMA and KOA.

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, $|\mathbb{S}_{N_o}| = 1 + (N_v - 1)/2^{N_o}$, 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:

$$H(\mathbf{K}|(\mathbf{Y},\mathbf{M})^{N_o}) \lesssim N_c \log_2(1 + 2^{-N_o}(N_v - 1)). \tag{4}$$

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 $N_o^\star = \log_2 N_v$ [16, Sect. 14].

E. 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:

$$|\mathbb{S}_j| = |\mathbb{S}_{j-1}| + N_c(1 - |\mathbb{S}_{j-1}|/N_c)/2 = N_c(1 - 2^{-j}). \tag{5}$$

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:

$$H(\mathbf{K}|(\mathbf{Y}, \mathbf{X})^{N_o}) = \log_2 \left(\frac{(N_v - \lceil |\mathbb{S}_{N_o}| \rceil)!}{(N_v - N_c)!(N_c - \lceil |\mathbb{S}_{N_o}| \rceil)!} \right) + \log_2(N_c!). \tag{6}$$

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 $N_o^\star = \log_2 N_c$, 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.

IV. Security analysis of spread spectrum based techniques

Spread spectrum is a military communication scheme invented during World War II [19]. 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 [6] and practical implementations [20] focus on the optimization of operational capacity-robustness functions at given embedding distortions.

A. 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 $\mathbf{u}_\ell$:

$$\mathbf{w} = \frac{\gamma}{\sqrt{N_c}} \sum_{\ell=1}^{N_c} a(\ell)\mathbf{u}_\ell, \tag{7}$$

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). The inverse extraction puts back vector $\mathbf{y} = \mathbf{x} + \mathbf{w}$ into the media producing 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) [5], [21]–[23]. Symbols $a(\ell)$ are then continuous real values (see subsection V-D).

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 watermarking channel access. 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 the carriers.

In the sequel, the security analysis considers several watermarked vectors $\mathbf{y}_j$, $1 \leq j \leq N_o$, with different embedded messages $\mathbf{a}_j = (a_j(1) \dots a_j(N_c))^T$ 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:

$$\mathbf{y}_j = \mathbf{x}_j + \frac{\gamma}{\sqrt{N_c}}\mathcal{U}\mathbf{a}_j, \tag{8}$$

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 $\mathcal{A}$):

$$\mathcal{Y} = \mathcal{X} + \frac{\gamma}{\sqrt{N_c}}\mathcal{U}\mathcal{A}\,. \tag{9}$$

B. 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.

C. 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 (as well as the algorithm used in experiments in section V) can be adapted to colored original signals and even non stationary original signals [24].

Another motivation is that, according to the Power Spectrum Constraint [25], 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:

$$L(\mathbf{y}^{N_o}) = \frac{1}{(\sqrt{2\pi}\sigma_x)^{N_o N_v}} e^{\left(-\frac{1}{2\sigma_x^2}\sum_{j=1}^{N_o} \|\mathbf{y}_j - \frac{\gamma}{\sqrt{N_c}}\mathcal{U}\mathbf{a}_j\|^2\right)}, \tag{10}$$

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

$$\frac{\partial \log L}{\partial \mathbf{u}_\ell} = \frac{\gamma}{\sigma_x^2 \sqrt{N_c}} \sum_{j=1}^{N_o} a_j(\ell)\mathbf{x}_j. \tag{11}$$

Product expectation gives the following $N_v \times N_v$ sub-blocks:

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

The FIM is then the following block matrix:

$$\begin{aligned} \text{FIM} &= \frac{\gamma^2}{N_c\sigma_x^2}\begin{bmatrix} (\mathcal{F}_{uu})_{1,1} & \dots & (\mathcal{F}_{uu})_{1,N_c} \\ \vdots & & \vdots \\ (\mathcal{F}_{uu})_{N_c,1} & \dots & (\mathcal{F}_{uu})_{N_c,N_c} \end{bmatrix} \\ &= \frac{\gamma^2}{N_c\sigma_x^2}\mathcal{F}_{uu} \overset{N_o\to+\infty}{\longrightarrow} N_o\frac{\gamma^2\sigma_a^2}{N_c\sigma_x^2}\mathcal{I}_{N_vN_c}. \end{aligned} \tag{12}$$

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 $N_o^\star = N_c\sigma_x^2/\gamma^2$ of (watermarked signals/hidden messages) pairs.

D. 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 $\mathcal{U}$:

$$\mathbf{d}_j = \mathbf{y}_j - \mathbf{x}_j = \frac{\gamma}{\sqrt{N_c}}\mathcal{U}\mathbf{a}_j. \tag{13}$$

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 [26]. 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:

$$\frac{\partial}{\partial\mathcal{P}}\log p(\mathbf{v};\mathcal{P}) = -\mathcal{P}^{-T} + \mathcal{P}^{-T}\boldsymbol{\chi}(\mathcal{P}^{-1}\mathbf{v})\mathbf{v}^T\mathcal{P}^{-T}, \tag{14}$$

with $\boldsymbol{\chi}(\mathbf{x}) = -\frac{\partial}{\partial\mathbf{x}}\log p_{\mathbf{A}}(\mathbf{x})$ [27]. 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 $\chi(.)$ the score function of $a_j(i)$, and by $\chi_n(.)$ 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 [28], with g defined as:

$$g = \frac{E\{(\chi(a)-\chi_n(a))^2\}}{E\{\chi_n(a)^2\}}. \tag{15}$$

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, [29] 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 $1/N_o$ [28], [30].

E. Watermarked Only Attack

In this section, the sources are unknown and can then be regarded as nuisance parameters [31], [32]. Vector $\boldsymbol{\psi}$ 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:

$$\frac{\partial \log L}{\partial \mathbf{a}_j} = \frac{\gamma}{\sigma_x^2\sqrt{N_c}}\mathcal{U}^T\mathbf{x}_j \quad \forall j \in \{1,\dots,N_o\}. \tag{16}$$

Product expectation gives the following sub-blocks:

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

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

$$\text{FIM} = \frac{\gamma^2}{N_c\sigma_x^2}\begin{bmatrix} \mathcal{F}_{uu} & \mathcal{F}_{ua} \\ \mathcal{F}_{au} & \mathcal{F}_{aa} \end{bmatrix}. \tag{17}$$

Note that $\mathcal{F}_{aa} = \mathcal{I}_{N_oN_c}$. The Cramér-Rao Bound for estimated $\text{Vect}(\mathcal{U}) = (\mathbf{u}_1^T,\dots,\mathbf{u}_{N_c}^T)^T$ is given by:

$$\text{CRB}(\text{Vect}(\mathcal{U})) = \frac{N_c\sigma_x^2}{\gamma^2}\widetilde{\mathcal{F}}_{uu}^{-1}, \tag{18}$$

with $\widetilde{\mathcal{F}}_{uu} = (\mathcal{F}_{uu} - \mathcal{F}_{ua}\mathcal{F}_{aa}^{-1}\mathcal{F}_{au}) = (\mathcal{F}_{uu} - \mathcal{F}_{ua}\mathcal{F}_{au})$. It is known that, in the general case, $\widetilde{\mathcal{F}}_{uu}^{-1} \geq \mathcal{F}_{uu}^{-1}$ (*i.e.* $\widetilde{\mathcal{F}}_{uu}^{-1} - \mathcal{F}_{uu}^{-1}$ is non negative definite). In other words, nuisance parameters render the estimation of $\mathcal{U}$ less accurate [27]. But, the situation is even worse here as the FIM becomes singular. Indeed:

$$(\mathcal{F}_{ua}\mathcal{F}_{au})_{\ell,k} = \sum_{j=1}^{N_o} (\mathcal{F}_{ua})_{\ell,j}(\mathcal{F}_{au})_{j,k} = \sum_{j=1}^{N_o} a_j(\ell)a_j(k)\mathcal{U}\mathcal{U}^T, \tag{19}$$

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

This problem stems from two facts. First, we did not integrate some constraints during our derivation. Especially, we know that $\mathbf{u}_\ell^T\mathbf{u}_k = \delta_{\ell,k}$. [31] 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.

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 [31]. 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:

$$\mathcal{H} = \begin{bmatrix} \mathcal{U}^\perp & \mathbf{0} \\ \mathbf{0} & \mathcal{I}_{N_o} \end{bmatrix}, \tag{20}$$

with $\mathcal{U}^\perp$ a basis of the complementary subspace of Span$(\mathbf{u}_1)$ in $\mathbb{R}^{N_v}$. Then, according to [31, Th. 1], the Cramér-Rao Bound under the above-mentioned constraint is 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:

$$\text{CRB}(\mathcal{U}^T\mathcal{A}) = \frac{\sigma_x^2}{\gamma^2}\begin{bmatrix} (\mathcal{A}\mathcal{A}^T)^{-1}\mathcal{U}^\perp\mathcal{U}^{\perp T} & \mathbf{0} \\ \mathbf{0} & \mathcal{I}_{N_o} \end{bmatrix}, \tag{21}$$

and we finally get:

$$\text{CRB}(\mathcal{U}^T) = \frac{\sigma_x^2}{\gamma^2}(\mathcal{A}\mathcal{A}^T)^{-1}\mathcal{U}^\perp\mathcal{U}^{\perp T} \overset{N_o\to+\infty}{\longrightarrow} \frac{\sigma_x^2}{N_o\sigma_a^2\gamma^2}\mathcal{U}^\perp\mathcal{U}^{\perp T}. \tag{22}$$

2) N_c carriers ($N_c > 1$): The ambiguity renders the FIM singular, even when considering the constraints. However, section V shows that, in practice, the opponent builds noisy estimation of the carriers up to a signed permutation. A possibility in [32], 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_mN_c constraints of the type: $\hat{a}_j(\ell) = a_j(\ell)$. At the end, calculation leads to:

$$\text{CRB}(\text{Vect}(\mathcal{U})) = \frac{N_c\sigma_x^2}{\gamma^2}\mathcal{H}_{uu}\mathcal{B}^{-1}{\mathcal{H}_{uu}}^T, \tag{23}$$

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 $(\mathcal{B})_{\ell,k} = (\mathcal{A}\mathcal{A}^T)_{\ell,k}\mathcal{U}_\ell^{\perp T}\mathcal{U}_k^\perp - (\mathcal{A}_{N_m:N_o}{\mathcal{A}_{N_m:N_o}}^T)_{\ell,k}\mathcal{U}_l^{\perp T}\mathcal{U}\mathcal{U}^T\mathcal{U}_k^\perp$, and $\mathcal{H}_{uu}$ the $N_cN_v \times N_c(N_v-1)$ 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 $\mathcal{U}_\ell^\perp$ form an orthonormal basis of the complementary subspace of 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 [32].

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 (22), where γ^2 is replaced by the power per carrier γ^2/N_c. The security level is then $N_o^\star = N_c\sigma_x^2/\sigma_a^2\gamma^2$ which is, by the way, coherent with (23). This is quite surprising because the security level is the same against KMA. Yet, the estimation of the secret carriers remains up to a signed permutation in the WOA.

F. Possible Hacks

The conclusion of this security analysis stands in the different possibilities to forge pirated content.

- The pirate discloses secret subspace Span (U). He can now focus attack's noise in this subspace to jam the communication far more efficiently. He can also nullify the watermarked signals projection in this subspace to remove the watermark.
- The pirate discloses the secret carriers up to a signed permutation. The above-mentioned hacks are still possible. Besides, he can detect whether two watermarked pieces of content share the same hidden message. He can also flip some randomly chosen bits. Moreover, the accidental knowledge of hidden messages in few watermarked pieces of content might remove this ambiguity. This extra security analysis indeed pertains to subsection III-D.
- The pirate discloses the secret carriers. He has a full access to the watermarking channel to read, write or erase hidden message.

Of course, the quality of the pirated pieces of content depends on the accuracy of his estimation. The authors focus on this aspect in [33].

V. Algorithms for spread spectrum based techniques

Section III not only gives security levels of the substitutive method, but also contains almost practical implementations of workable algorithms. On the contrary, section IV only presents theoretical assessment of security levels. Hence, this section deals with practical algorithms useful to hack spread spectrum based watermarking schemes. For each attack, an algorithm is presented, and tested on synthetic data as supposed by the model of (8), with BPSK symbols and gaussian host vectors. These algorithms are then applied on spread transform side information methods and one still image technique.

This section has an intensive use of PCA and ICA algorithms, which is completely new in watermarking security analysis, as the only other papers mentioning PCA/ICA in the watermarking community have different purposes. [34] and [35] used ICA to design a watermarking embedder. [36] presented a technique for estimating the watermark by observing only one image. Their purpose is the simple erasure of the

whole watermark signal and not the disclosure of the secret parameters, whereas the approach here allows a complete access to the watermarking communication channel to remove, read or write hidden data [2].

The following average normalized correlation measures the efficiency of our attack:

$$\eta = \frac{1}{N_c} \sum_{\ell=1}^{N_c} \frac{\hat{\mathbf{u}}_\ell^T \mathbf{u}_\ell}{\|\hat{\mathbf{u}}_\ell\|}. \tag{24}$$

Although the normalization renders estimators $\hat{\mathbf{u}}_j / \|\hat{\mathbf{u}}_j\|$ biased [38], the normalized correlation is preferred because it is an extremely popular measure in the watermarking community. $\eta \lesssim 1$ means that the opponent discloses vectors almost collinear with the secret carriers. When existing, we manually removed the ambiguity of the signed permutation. Measures of η are done averaging $N_t = 128$ experimental results.

The relation with the theoretical security levels is not difficult to find out. (24) is in expectation the cosine of the angle between $\mathbf{u}_\ell$ and $\hat{\mathbf{u}}_\ell = \mathbf{u}_\ell + \mathbf{n}$, $\mathbf{n}$ being the estimation noise (orthogonal to $\mathbf{u}_\ell$ and whose norm is $\sqrt{\text{tr}(\text{CRB}(\text{Vect}(\mathcal{U})))/N_c}$, with $\text{tr}(A)$ the trace of matrix A.) The following relation holds:

$$\eta \approx \frac{\|\mathbf{u}_\ell\|}{\sqrt{\|\mathbf{u}_\ell\|^2 + \text{tr}(\text{CRB}(\text{Vect}(\mathcal{U})))/N_c}}. \tag{25}$$

A. Known Message Attack

Observing $(\mathbf{y}, \mathbf{a})^{N_o}$, the opponent can use the Maximum Likelihood Estimator (MLE) related to (10). This estimator is also defined by $\frac{\partial \log L}{\partial \mathbf{u}_\ell} = \mathbf{0} \quad \forall \ell \in \{1, \ldots, N_c\}$, which gives:

$$\hat{\mathcal{U}} = \frac{\sqrt{N_c}}{\gamma} (\mathcal{Y}\mathcal{A}^T)(\mathcal{A}\mathcal{A}^T)^{-1}. \tag{26}$$

The MLE is known to be unbiased and consistent, *i.e.* it asymptotically achieves the CRB derived in subsection IV-D. Fig. (5) shows experimental values of η against N_o and WCR $= \gamma^2/\sigma_x^2$ for the DSSS case. The locus of points such that $\eta = const$ are projected on the plane $\eta = 0$. They appear to be parallel with the curve $N_o = N_c \sigma_x^2 / \gamma^2$. Tests done with different N_v confirm that the efficiency of the

[2]We discovered after submission a similar approach uniquely devoted to watermark removal and only based on PCA in [37].

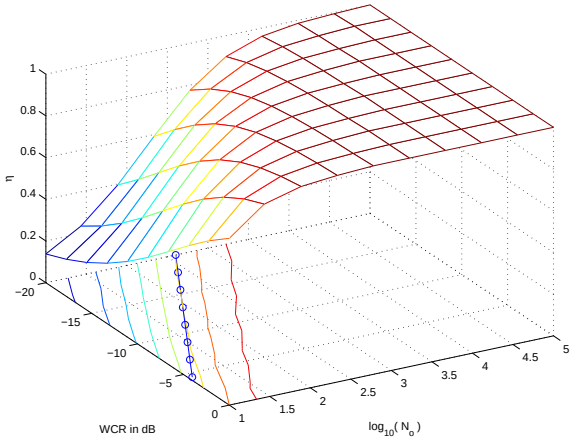

Fig. 5. KMA for DSSS ($N_c = 4$, $N_v = 512$). η against $\log_{10}(N_o)$ and WCR in dB. The curve $N_o = N_c\sigma_x^2/\gamma^2$ is plotted with small circles.

attack does not depend on the vector length. This asserts the theoretical security level of subsection IV-C.

B. Known Original Attack

In this case, the opponent observes several instances of $\mathbf{d}_j = (\mathbf{y}_j - \mathbf{x}_j) \propto \mathcal{U}\mathbf{a}_j$. As seen in subsection IV-D, this is related to the well known problem of signal processing called Blind Source Separation (BSS), with no noise. A lot of papers have already been written on BSS, and we just recall here its most common algorithms. Note that spread spectrum corresponds to the BSS over-determined case (*i.e.*, $N_v \geq N_c$).

The most classical algorithm in BSS is the Principal Component Analysis (PCA). Denote $\mathcal{D} = \mathcal{Y} - \mathcal{X}$. This technique makes an eigendecomposition of the matrix $\mathcal{D}\mathcal{D}^T = \gamma^2\mathcal{U}\mathcal{A}\mathcal{A}^T\mathcal{U}^T/N_c$. This corresponds to a Gram-Schmidt orthogonalization of vectors $\mathbf{d}^{N_o}$. Please, note that $\rho \stackrel{\Delta}{=} \text{Rank}\,(\mathcal{A})$ is also the rank of $\mathcal{D}\mathcal{D}^T$. Hence, the decomposition outputs ρ orthonormal vectors lying in $\text{Span}\,(\mathcal{U})$. In the best case, the opponent has $\rho = \min(N_o, N_c)$. Nevertheless, in reality, he may have $\rho \leq \min(N_o, N_c)$ if the N_o symbol vectors are linearly dependent.

When successful (*i.e.*, when $\rho = N_c$), the PCA technique yields a orthonormal basis of the secret subspace $\text{Span}\,(\mathcal{U})$. The possibilities to hack watermarked pieces of content when $\text{Span}\,(\mathcal{U})$ is disclosed

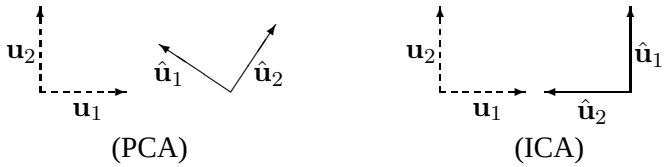

Fig. 6. PCA *vs.* ICA. PCA finds the secret carriers up to a rotation, whereas ICA succeeds to align the estimated carriers $\hat{\mathbf{u}}^{N_c}$ with $\mathbf{u}^{N_c}$ (Here, $N_c = 2$). An ambiguity remains in their order (permutation) and orientation (sign).

are summarized in subsection IV-F. Yet, the vectors of this basis are not necessary collinear with the private carriers. This is due to the unitary matrix $\mathcal{P}$ mentioned in subsection IV-D. The opponent cannot decode, as projection of watermarked signals onto this basis gives a mixture of the hidden symbols. This is illustrated by Fig. 6. The same reason prevents him transmitting information in the hidden channel. Nevertheless, under the assumption that the symbol vectors are *statistically* independent, the opponent can resort to a more powerful tool: the Independent Component Analysis (ICA). It is an extension of PCA, constraining the output estimated symbol vectors to be independent [26]. Good tutorials on ICA and on its links with BSS are [28], [39]. A very general ICA algorithm named FastICA [40] has been preferred to algorithms dedicated to specific symbol distribution [29], [30].

In short, ICA algorithms usually work in the basis recovered by a PCA. This basis describes exactly the secret subspace (provided that $\rho = N_c$). The problem is now reduced to the estimation of the $N_c \times N_c$ matrix $\mathcal{P}$. Hence, parameter N_v has absolutely no influence on the attack. Then, in an iterative process, the ICA 'rotates' the basis until it nullifies an objective function (often called a constrast function) of the estimated sources $\hat{\mathbf{a}}^{N_o}$. This function can be an approximation of the mutual information of the estimated sources. Contrast functions depend on the distribution of the symbol sources. However, this measure reflects statistical independence only for large N_o. For a finite number of observations, ICA algorithms usually search for a minimum of the contrast function with the help of a gradient descent technique.

When successful, ICA reduces the set of ambiguity matrices $\mathcal{P}$ to the one of signed permutations. This is illustrated by Fig. 6. Subsection IV-F lists the possibilities to hack watermarked pieces of content

when the carriers are disclosed up to a signed permutation.

C. Watermarked Only Attack

The WOA case is quite similar to KOA, as it is related to BSS in a noisy environment. The covariance matrix $\mathcal{R}_{\mathbf{y}}$ has the following expression:

$$\mathcal{R}_{\mathbf{y}} = \mathcal{R}_{\mathbf{x}} + \frac{\gamma^2}{N_c}\mathcal{U}\mathcal{R}_{\mathbf{a}}\mathcal{U}^T = \sigma_x^2\mathcal{I} + \frac{\gamma^2\sigma_a^2}{N_c}\mathcal{U}\mathcal{U}^T. \tag{27}$$

Its diagonalization leads to N_c eigenvalues equaling $\sigma_x^2 + \frac{\gamma^2\sigma_a^2}{N_c}$, and $N_v - N_c$ eigenvalues equaling σ_x^2. Hence, the eigenvectors related to the N_c biggest values constitute a basis of Span $(\mathcal{U})$, which is also known as the signal space in blind equalization for digital communications.

PCA estimates covariance matrix $\mathcal{R}_{\mathbf{y}}$ by $\mathcal{Y}\mathcal{Y}^T/N_o$, and outputs N_c eigenvectors whose eigenvalues are the biggest ones. Due to this rough estimation, these vectors do not live exactly in Span $(\mathcal{U})$. Compared to Fig. 6, these noisy estimation vectors would not lie in the plan of the page, regarded as subspace Span $(\mathcal{U})$ in this simple example. However, ICA will still try to rotate them in order to render the decoded symbols independent. Fig. 7 shows the locus of points such that $\eta = const$ for different values of N_c and N_o, with the DSSS method (*i.e.*, a BPSK modulation). The ICA algorithm meets the theoretical limit only for large N_o, and high energy of watermark signal per carrier: $\gamma^2 N_v/N_c$. Note that, for $N_c = 4$, the gap between experimental performances and theoretical limit gets larger.

D. Spread transform side information watermarking

This subsection presents experiments with side information watermarking using the process on spread spectrum. In these methods, the symbols $a_j(\ell)$ depend on the host signal in the following way:

$$a_j(\ell) = f(m_j(\ell), \mathbf{u}_\ell^T\mathbf{x}_j) \tag{28}$$

Three techniques were investigated: Improved Spread Spectrum (ISS) [23], Scalar Costa Scheme (SCS) [21], and Maximized Robustness Embedding (MRE) [22]. Two implementations of SCS have been done.

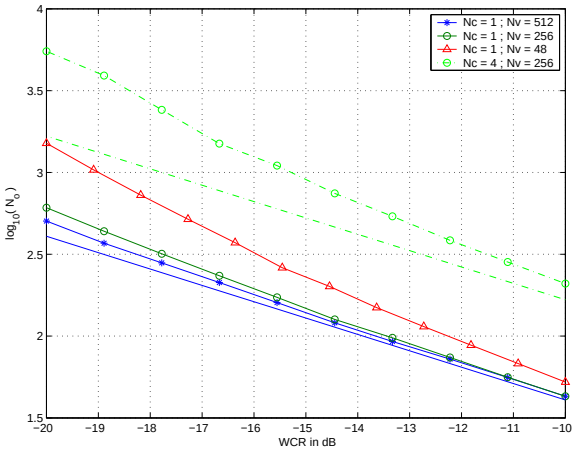

Fig. 7. WOA for DSSS. Operating points achieving $\eta = 0.8$ for different parameters N_c and N_v. The solid line is the theoretical limit for $N_c = 1$, and curves with stars, circles and triangles are the experimental results. They capture the efficiency of the PCA, as only one carrier is used. The dashed line is the theoretical limit for $N_c = 4$ (*i.e.* the solid line translated of $\log_{10}(N_c)$), the dashed curve with circles is the experimental results with the FastICA algorithm [40].

The carriers have disjoint supports in the first one, which is a possible interpretation of [21]: $\mathbf{u}_1 = (\mathbf{u}^T \mathbf{0}_\tau^T \dots \mathbf{0}_\tau^T)^T$, $\mathbf{u}_2 = (\mathbf{0}_\tau^T \mathbf{u}^T \dots \mathbf{0}_\tau^T)^T$, and so on with $\tau N_c = N_v$. The second implementation is called SCS with Subspace Projection (SSP) [41]: the carriers have a full support and are orthonormal. The embedding distortion, the vector length and the number of hidden bits are the same for a fair comparison.

The KMA case has not been investigated. The knowledge of the messages does not usually imply the disclosure of the symbols. In SCS, function $f(.)$ of (28) is private and depends on a secret key (*i.e.*, a dithering vector). However, information about the symbols may leak from the message. Symbols are Gaussian variables centered on $\gamma(-1)^{m_j(\ell)}$ for the ISS technique:

$$a_j(\ell) = \gamma(-1)^{m_j(\ell)} - \lambda \mathbf{u}_\ell^T \mathbf{x}_j. \tag{29}$$

We foresee that the MLE algorithm could easily be tuned to exploit this information leakage.

The KOA is simpler, as the basic assumption is still valid: $\mathbf{u}_\ell^T \mathbf{x}_j$ and $\mathbf{u}_k^T \mathbf{x}_j$ ($k \neq \ell$) are Gaussian distributed and non correlated; thus, the symbols are statistically independent. Yet, the efficiency of BSS depends on the symbols distribution, so that we expect different performances. Once again, in our

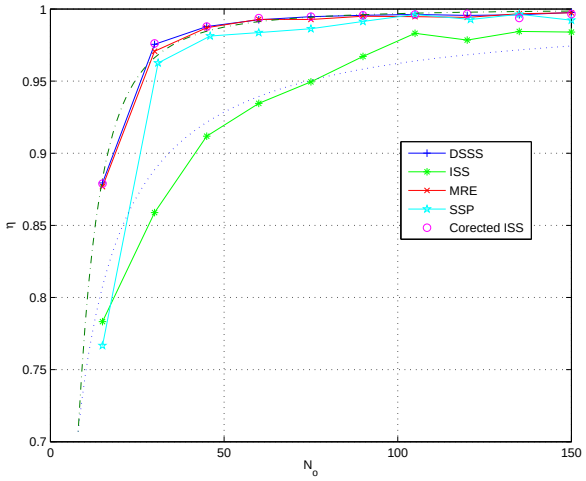

Fig. 8. KOA for four different watermarking techniques($N_c = 4$, $N_v = 512$). Dotted line: $\eta = (1 + k/N_o)^{-1}$; Dash-dotted line: $\eta = (1 + (k/N_o)^2)^{-1}$.

simulation, the opponent always uses the same generic ICA algorithm. No fine tuning according to the expected symbols distribution is done. Fig. (8) shows the results, except for SCS[3]. Surprisingly, the rate of the noise estimation variance is in ${1/N_o}^2$ for DSSS, SSP and MRE. This seems to be due to the bounded support feature of the symbols in these methods, despite of the use of a generic algorithm. For ISS, the rate is in $1/N_o$. Please, note that, according to (29), the KOA for ISS is similar to a WOA for the SS method, with a watermark to host power ratio of $\gamma^2/\lambda^2\sigma_x^2$. A smarter attack on ISS stems from this remark. First, difference vectors are used to disclose the secret subspace with a PCA. Then, they are corrected in adding the projection of the original vectors scaled by a factor λ. We are now in a situation similar to a KOA with DSSS. Finally, ICA finishes the job working on the corrected vectors. The last curve named 'Corrected ISS' in Fig. (8) shows the dramatic improvement. The security level of ISS is in practice as low as the DSSS one.

The WOA is also straightforward as we applied the same ICA algorithm for DSSS, ISS, MRE, and SSP. For SCS, the observed watermarked vectors are split by chunks of τ samples. Thus, the opponent

[3]For SCS, $N_o = 1$ is enough to disclose small length carrier $\mathbf{u}$ up to a sign.

has $N_o{}' = N_o\tau$ vectors whose length is $N_v{}' = N_v/\tau$, watermarked with $N_c{}' = 1$ secret carrier. The algorithm is thus a simple PCA in this case. Fig. (9) shows the results. SCS (or more precisely the way we have implemented it) is obviously the less secure. But the simple change brought in the implementation of SSP is sufficient to correct this security flaw [4]. The other techniques share the same security level. ISS seems to be slightly more secure; however, remember that we did not tune the contrast function of the ICA algorithm. In the same way, the embedding parameters (γ, λ) play a big role in the symbols distribution, and the attack might thus perform differently. This is the reason why we prefer to look at the global shape of the curves, rather than to draw erroneous conclusions from these meager differences.

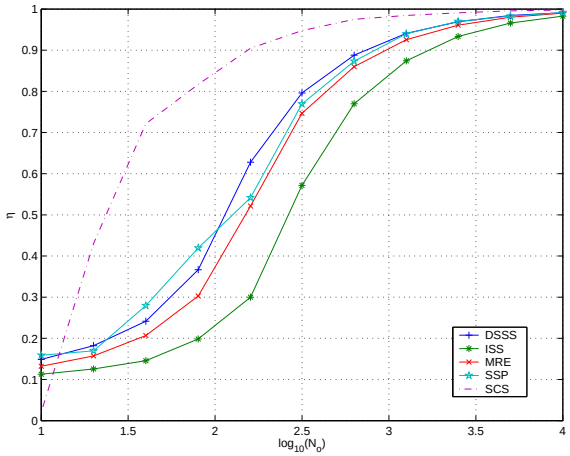

Fig. 9. WOA for five different watermarking methods ($N_v = 512$, $N_c = 4$, WCR=-15dB). $\tau = 128$ for SCS. For SCS, SSP and ISS, the embedding parameters are optimal for an expected noise attack whose distortion equals the embedding distortion: WNR=0 dB.

[4]We only analyze here the security of the spreading transform. Yet, the dithering vector in SCS-like technique constitute a second barrier, which will be the subject of a future work.

E. Application to a robust watermarking technique

The goal of this last subsection is to demonstrate the power of 'smart' attacks based on secret carriers estimation. So far, this article has investigated the first phase of the attack: the secret disclosure. Now, in a second phase, the opponent uses this *a posteriori* information to hack pieces of content, which were watermarked with the same secret key. To this end, the subsection deals with real still images. The robust watermarking technique from [20] has been chosen.

A challenge is proposed to two opponents: they attack a watermarked image with an increasing attack distortion, until an oracle warns them that the decoded message is different from the embedded message ($N_c = 8$ bits, PSNR=38dB). Pirate A uses *blind* attacks (*i.e.*, pertaining to the robustness issue – except any geometric attack). For instance, in this article, he scales the size of the image by a quarter, JPEG compresses it with a decreasing quality factor, and finally scales back the image. Pirate B uses *smart* attacks. He has estimated the secret carriers by a WOA, with $N_o \sim 1000$ images such that $\eta = 0.5$[5], and he tries to remove the hidden information for one carrier. Details of algorithm adaptations to real images may be found in [33]. Fig. (10) shows the result of the challenge for the Lena image. For a panel of 50 pictures (512×512 pixels), pirate B on average produces an attack distortion 15dB smaller than pirate A to successfully hack watermarked pictures.

VI. CONCLUSION

As in cryptanalysis, measurement of information leakages is the fundamental principle underlying the theoretical framework for robust watermarking security assessment presented in this article. A watermarking technique, even robust, is not secure if the opponent can refine his knowledge on the presumably secret key while pieces of content are watermarked with the same key. The security level is then defined

[5]The opponent cannot know this last value. However, nothing prevents him to run simulations with his own private carriers in order to estimate η.

(a) Pirate A (b) Pirate B

Fig. 10. Comparison between the two pirated Lena images. This is their best quality for a successful attack. Pirate A: PSNR=21.8 dB, Pirate B: PSNR=35.8 dB.

by the number of observations the opponent needs in order to accurately estimate the secret key.

The conclusion of this article is not that spread spectrum based watermarking techniques or substitutive schemes are broken. The goal is to warn the watermarking community that security is a crucial issue. Designers should not only control the imperceptibility and the robustness of their schemes but also assess their security levels. Depending on the application designers are targeting (and especially on the observations available to the pirate), watermarking several pieces of content with the same key might bring threats. This potentially arises difficulties on the key management. For instance, it is not clear how a blind watermarking decoder will be informed of the secret key, if this later one is to be changed according to the security levels assessed in this article.

REFERENCES

[1] I. Cox, M. Miller, and J. Bloom, *Principles and Practice*, Morgan Kaufmann Publisher, 2001.

[2] J. O'Ruanaidh and T. Pun, "Rotation, scale and translation invariant spread spectrum digital image watermarking," *Signal Processing*, vol. 66, no. 3, pp. 303–17, 1998.

[3] S. Pereira and T. Pun, "Fast robust template matching for affine resistant image watermarks," in *Proc. IHW*, A. Pfitzmann, Ed., Dresden, Germany, Sept. 1999, pp. 199–210, Springer Verlag.

[4] I. Cox, M. Miller, and A. McKellips, "Watermarking as communication with side information," *Proc. IEEE*, vol. 87(7), pp. 1127–1141, July 1999.

[5] B. Chen and G. Wornell, "Quantization index modulation: A class of provably good methods for digital watermarking and information embedding," *IEEE Trans. Inform. Theory*, vol. 47, pp. 1423–1443, May 2001.

[6] P. Moulin, "The role of information theory in watermarking and its application to image watermarking," *Signal Processing*, vol. 81, pp. 1121–1139, 2001.

[7] T. Kalker, "Considerations on watermarking security," in *Proc. MMSP*, Cannes, France, Oct. 2001, pp. 201–206.

[8] S. Craver, N. Memon, B.-L. Yeo, and M.M. Yeung, "On the invertibility of invisible watermarking technique," in *Proc. ICIP*, Washington, DC, USA, Oct. 1997, IEEE, pp. 540–543.

[9] M. Kutter, S. Voloshynovskiy, and A. Herrigel, "Watermark copy attack," in *Security and Watermarking of Multimedia Contents II*, P.W. Wong and E. Delp, Eds., San Jose, Cal., USA, Jan. 2000, vol. 3971.

[10] I. Cox and J.-P. Linnartz, "Some general methods for tampering with watermarks," *IEEE J. Select. Areas Commun.*, vol. 16, no. 4, pp. 587–93, May 1998, Special issue on copyright and privacy protection.

[11] J.P. Linnartz and M. van Dijk, "Analysis of the sensitivity attack against electronic watermarks in images," in *Proc. IHW*, D. Aucsmith, Ed., Portland, Oregon, USA, Apr. 1998, vol. 1525 of *Lecture Notes in Computer Science*, Springer Verlag.

[12] T. Mittelholzer, "An information-theoritic approach to steganography and watermarking," in *Proc. IHW*, A. Pfitzmann, Ed., Dresden, Germany, Sept. 1999, pp. 1–17, Springer Verlag.

[13] T. Furon and P. Duhamel, "An asymmetric watermarking method," *in [42]*, pp. 981–995.

[14] M. Barni, F. Bartolini, and T. Furon, "A general framework for robust watermarking security," *Signal Processing*, vol. 83, no. 10, pp. 2069–2084, Oct. 2003, Special issue on Security of Data Hiding Technologies, invited paper.

[15] A. Kerckhoffs, "La cryptographie militaire," *Journal des sciences militaires*, vol. 9, pp. 5–38, janvier 1883.

[16] C.E. Shannon, "Communication theory of secrecy systems," *Bell system technical journal*, vol. 28, pp. 656–715, Oct. 1949.

[17] W. Diffie and M. Hellman, "New directions in cryptography," *IEEE Trans. Inform. Theory*, vol. 22, no. 6, pp. 644–54, Nov. 1976.

[18] S. Burgett, E. Koch, and J. Zhao, "Copyright labelling of digitized image data," *IEEE Commun. Mag.*, vol. 36, no. 3, pp. 94–100, Mar. 1998.

[19] D. Kahn, "Cryptology and the origins of spread spectrum," *IEEE Spectr.*, pp. 70–80, Sept. 1984.

[20] S. Pateux and G. Le Guelvouit, "Practical watermarking scheme based on wide spread spectrum and game theory," *Signal Processing: Image Communication*, vol. 18, pp. 283–296, Apr. 2003.

[21] J.Eggers, R. Baüml, R. Tzschoppe, and B.Girod, "Scalar costa scheme for information embedding," *in [42]*, pp. 1003–1019.

[22] M. Miller, I. Cox, and J. Bloom, "Informed embedding: exploiting image and detector information during watermark insertion," in *Proc. ICIP*, Vancouver, Canada, Sept. 2000.

[23] H.S. Malvar and D.A.F. Florêncio, "Improved spread spectrum: A new modulation technique for robust watermarking," *in [42]*, pp. 868–905.

[24] D.T. Pham and J.F. Cardoso, "Blind separation of instantaneous mixtures of non stationary sources," *IEEE Trans. Signal Processing*, vol. 49, no. 9, pp. 1837–1848, 2001.

[25] J. Su, J. Eggers, and B. Girod, "Analysis of digital watermarks subjected to optimum linear filtering and additive noise," *Signal processing*, vol. 81, pp. 1141–1175, 2001.

[26] P. Comon, "Independent component analysis, a new concept?," *Signal Processing*, vol. 36, no. 3, pp. 287–314, 1994.

[27] S.-I. Amari and J.F. Cardoso, "Blind source separation; semiparametric statistical approach," *IEEE Trans. Signal Processing*, vol. 45, no. 11, 1997, Special issue on neural networks.

[28] J.-F. Cardoso, "Blind signal separation: statistical principles," *Proc. IEEE*, vol. 86, no. 10, pp. 2009–2025, Oct. 1998.

[29] A.-J. van der Veen, "Blind separation of BPSK sources with residual carriers," *Signal Processing*, vol. 73, no. 10, pp. 67–79, Jan. 1999.

[30] F. Gamboa and E. Gassiat, "Source separation when the input sources are discrete or have constant modulus," *IEEE Trans. Signal Processing*, vol. 45, no. 12, pp. 3062–3072, Dec. 1997.

[31] P. Stoica and B.C. Ng, "On the Cramér-Rao bound under parametric constraints," *IEEE Signal Processing Lett.*, vol. 5, no. 7, pp. 177–179, 1998.

[32] Y. Yao and G.B. Giannakis, "On regularity and identifiability of blind source separation under constant-modulus constraints," *IEEE Trans. Signal Processing*, 2004, To appear.

[33] F. Cayre, C. Fontaine, and T. Furon, "Watermarking attack: Security of wss techniques," in *Proc. IWDW*, Seoul, Corea, Oct. 2004, Springer-Verlag.

[34] F.J. González-Serrano and J.J. Murillo-Fuentes, "Independent component analysis applied to image watermarking," in *Proc. ICASSP*, 2001.

[35] S. Bounkong, B. Toch, D. Saad, and D. Lowe, "ICA for watermarking digital images," *Journal of Machine Learning Research*, vol. 1, pp. 1–25, 2002.

[36] J. Du, C.-H. Lee, H.-K. Lee, and Y. Suh, "Watermark attack based on blind estimation without priors," in *Proc. IWDW*. 2002, Lecture Notes in Computer Science, Springer-Verlag.

[37] G. Doërr and J.-L. Dugelay, "Danger of low-dimensional watermarking subspaces," in *Proc. ICASSP*, Montreal, Canada,

may 2004, vol. 3.

[38] P. Stoica and B. Ng, *Signal Processing Advances in Wireless and Mobile Communications*, vol. 1, chapter Performance Bounds for Blind Channel Estimation, pp. 41–62, Prentice Hall, 2001.

[39] A. Hyvärinen and E. Oja, "Independent component analysis: a tutorial," *Neural Networks*, vol. 13, no. 4-5, pp. 411–430, 2000.

[40] A. Hyvärinen, "Fast and robust fixed-point algorithms for independent component analysis," *IEEE Trans. Neural Networks*, vol. 10, no. 3, pp. 626–634, 1999.

[41] R. Fischer, R. Tzschoppe, and R. Bäuml, "Lattice costa schemes using subspace projection for digital watermarking," *European Trans. Telecommunications*, vol. 15, no. 4, pp. 351–362, Aug. 2004.

[42] A. Akansu, E. Delp, T. Kalker, B. Liu, N. Memon, P. Moulin, and A. Tewfik, "Special issue on signal processing for data hiding in digital media and secure content delivery," *IEEE Trans. Signal Processing*, vol. 51, no. 4, Apr. 2003.