Fundamentals of Data Hiding Security and their Application to Spread-Spectrum Analysis

Pedro Comesaña, Luis Pérez-Freire and Fernando Pérez-González *

Signal Theory and Communications Department
University of Vigo, Vigo 36310, Spain
{pcomesan, lpfreire, fperez}@gts.tsc.uvigo.es

Abstract. This paper puts in consideration the concepts of *security* and *robustness* in watermarking, in order to be able to establish a clear frontier between them. A new information-theoretic framework to study data-hiding and watermarking security is proposed, using the mutual information to quantify the information about the secret key that leaks from the observation of watermarked objects. This framework is applied to the analysis of a Spread-Spectrum data-hiding scheme in different scenarios. Finally, we show some interesting links between a measure proposed in previous works in the literature, which is based on Fisher Information Matrix, and our proposed measure.

1 Introduction

Although a great amount of the watermarking and data-hiding[1] literature deals with the problem of robustness, little has been said about security, and even in this time of relative maturity of watermarking research no consensus has been reached about its definition, and robustness and security continue to be often seen as overlapping concepts. The purpose of this first section is to give an overview of the evolution of research on watermarking security.

First, the notation and a general model for the evaluation of watermarking security will be introduced. The model is depicted in Figures 1-a and 1-b: a message $\mathbf{M}$ will be embedded in an original document $\mathbf{X}$ (the *host*), yielding a watermarked vector $\mathbf{Y}$. The embedding stage is parameterized by the embedding key $\mathbf{\Theta}_e$, and the resulting watermark is $\mathbf{W}$. In the detection/decoding stage, the detection key $\mathbf{\Theta}_d$ is needed;[2] $\hat{\mathbf{M}}$ denotes the estimated message in the case of

* This work was partially funded by *Xunta de Galicia* under projects PGIDT04 TIC322013PR and PGIDT04 PXIC32202PM; MEC project DIPSTICK, reference TEC2004-02551/TCM; FIS project IM3, reference G03/185 and European Comission through the IST Programme under Contract IST-2002-507932 ECRYPT. ECRYPT disclaimer: The information in this paper is provided as is, and no guarantee or warranty is given or implied that the information is fit for any particular purpose. The user thereof uses the information at its sole risk and liability.

[1] In this paper we will use these both terms with no distinction.

[2] In symmetric watermarking $\mathbf{\Theta}_e = \mathbf{\Theta}_d$

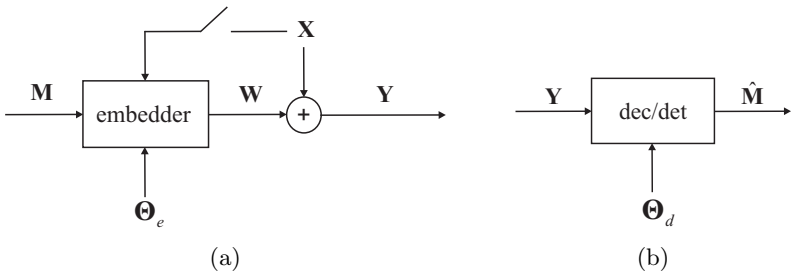

Fig. 1. General model for security analysis: embedding (a) and decoding/detection (b)

decoding, and the decision whether the received signal is watermarked or not in the case of detection. Capital letters denote random variables, and bold letters denote vectors.

During the first years, most of the literature deals with the problem of robustness, overlooking the meaning of security, in such a way that, at most, there was the notion of *intentional* and *non-intentional* attacks [1]. It could be said that the sensitivity attack [2] raised up the problem of security in watermarking, showing that a watermarking system could be broken in a number of iterations which is linear with the dimensionality of the host signal, but the first attempt at proposing a theoretical framework for assessing the security of a general watermarking scenario was [3]. The two main issues of this paper are the *perfect secrecy* (concept directly borrowed from the seminal work on cryptanalysis by Shannon in [4]) of the embedded message and the robustness of the embedding, characterizing both in terms of mutual information. However, this approach does not take into account that some information about the secret key may leak from the observations, giving advantage to the attacker. Thereafter, Kalker [5] shed some light on the concept of security in watermarking, giving definitions for *robust watermarking* and *security*, but perhaps they have the problem of being too general.

Another framework for watermarking security was proposed in [6], modeling watermarking as a game with some rules that determine which information (parameters of the algorithm, the algorithm itself, etc.) is public. According to this rules, attacks are classified as fair (the attacker only exploits publicly available information) or unfair (the attacker tries to access all the possible information which can be of help for him/her). The authors also define the *security level* as *"the amount of observation, the complexity, the amount of time, or the work that the attacker needs to gather in order to hack a system"*.

To the best of our knowledge, the most recent paper dealing with security is [7]. We agree with the authors about the difficulty of distinguishing between security and robustness. Kerckhoff's principle is also translated from cryptography to watermarking (it was introduced for the first time in [8]): all functions (encoding/embedding, decoding/detection, ...) should be declared as public except for a parameter called the secret key. An important contribution of [7] is the proposal of a security measure based on Fisher's Information Matrix [9]. In

Section 4.2 it will be shown that the proposed measure is somewhat questionable since it is neglecting some important parameters as the uncertainty (differential entropy) in the secret key or in the watermarked signal. Finally, in [7] the security analysis of spread spectrum is performed and some practical methods for hacking systems are introduced.

After this brief overview the rest of the paper is organized as follows: In Sect. 2, definitions of *security* and *robustness* are proposed, and related issues are studied. In Sect. 3, a new information-theoretic measure is proposed for data-hiding security; this is applied to the study of Spread Spectrum watermarking security analysis in Sect. 4. Finally, in Sect. 5, the conclusions of this work are presented.

2 Fundamental definitions

In this section, some thoughts about the concept of watermarking security are expounded and some definitions are proposed. First, in order to establish a clear line between robustness and security, the following definitions are put forward for consideration:

Definition 1. *Attacks to robustness are those whose target is to increase the probability of error of the data-hiding channel.*

Definition 2. *Attacks to security are those aimed at gaining knowledge about the secrets of the system (e.g. the embedding and/or detection keys).*

At first glance, in the definition of attacks to robustness we could have used the concept of channel capacity instead of the probability of error, but this entails some potential difficulties: for instance, an attack consisting on a translation or a rotation of the watermarked signal is only a desynchronization, thus the capacity of the channel is unaffected, but depending on the watermarking algorithm, the detector/decoder may be fooled. Another considerations about security, taking into account the above definitions, are the following:

About the intentionality of the attacks : attacks to security are obviously intentional, but not all intentional attacks are threats to security. For instance, an attacker may perform a JPEG compression to fool the watermark detector because he knows that, under a certain JPEG quality factor, the watermark will be effectively removed. Notice that, independently of the success of his attack, he has learned nothing about the secrets of the system. Hence, *attacks to security imply intentionality, but the converse is not necessarily true.*

About the blindness *of the attacks* : *blind* attacks are those which do not exploit any knowledge of the watermarking algorithm. Since attacks to security will try to disclose the secret parameters of the watermarking algorithm, it is easy to realize that they can not be blind. On the other hand, a *non-blind* attack is not necessarily targeted at learning the secrets of the system; for instance, in a

data-hiding scheme based on binary scalar Dither Modulation (scalar DM), if an attacker adds to each watermarked coefficient a quantity equal to a quarter of the quantization step, the communication is completely destroyed because the bit error probability will be 0.5, although the attacker has learned nothing about the secrets of the systems. Hence, *security implies non-blindness, but the converse is not necessarily true.*

About the final purpose *of attacks* : many attacks to security constitute a first step towards performing attacks to robustness. This can be easily understood with a simple example: an attacker can perform an estimation of the secret pseudorandom sequence used for embedding in a spread-spectrum-based scheme (attack to security); with this estimated sequence, he can attempt to remove the watermark (attack to robustness).

About the distinction *between security and robustness* : a watermarking scheme can be extremely secure, in the sense that it is (almost) impossible for an attacker to estimate the secret key(s), but this does not necessarily affect the robustness of the system. For instance, the boundary of the detection region of watermarking algorithms whose decisions are based on linear correlation can be complicated by using, as a decision boundary, a fractal curve [10]; this way, security is highly improved since, for example, sensitivity-like attacks are no longer effective because the boundary of the detection region is extremely hard to describe. However, this countermeasure against security attacks does not improve anyway the robustness of the method. *Therefore, higher security does not imply higher robustness.*

About the measure of security itself : security must be measured separately from robustness. The following analogy with cryptography may be enlightening in this sense: in cryptography, the objective of the attacker is to disclose the encrypted message, so the security of the system is measured assuming that the communication channel is error-free; otherwise it makes no sense to measure security, since the original message was destroyed both for the attacker and fair users. By taking into account the definition of robustness given at the beginning of this section, the translation of this analogy to the watermarking scenario means that security must be measured assuming that no attacks to robustness occur.

The measure of security proposed here is a direct translation of Shannon's approach [4] to the case of continuous random variables, which was already hinted for watermarking by Hernández *et al.* in [11]. Furthermore, we will take into account Kerckhoff's principle [12], namely that the secrecy of a system must depend only on the secret keys. Security can be evaluated in the two scenarios of Figure 1.

1. For the scenario depicted in Figure 1-a, security is measured by the mutual information between the observations $\mathbf{Y}$ and the secret key $\mathbf{\Theta}$

$$\begin{aligned} I(\mathbf{Y}^1, \mathbf{Y}^2, \ldots, \mathbf{Y}^{N_o}; \mathbf{\Theta}) &= h(\mathbf{Y}^1, \mathbf{Y}^2, \ldots, \mathbf{Y}^{N_o}) - h(\mathbf{Y}^1, \mathbf{Y}^2, \ldots, \mathbf{Y}^{N_o} | \mathbf{\Theta}) \\ &= h(\mathbf{\Theta}) - h(\mathbf{\Theta} | \mathbf{Y}^1, \mathbf{Y}^2, \ldots, \mathbf{Y}^{N_o}), \end{aligned} \tag{1}$$

where $h(\cdot)$ stands for differential entropy, and $\mathbf{Y}^n$ denotes the n-th observation.[3] Equivocation is defined as the remaining uncertainty about the key after the observations:

$$h(\mathbf{\Theta}|\mathbf{Y}^1, \mathbf{Y}^2, \ldots, \mathbf{Y}^{N_o}) = h(\mathbf{\Theta}) - I(\mathbf{Y}^1, \mathbf{Y}^2, \ldots, \mathbf{Y}^{N_o}; \mathbf{\Theta}). \tag{2}$$

This scenario encompasses attacks concerning the observation of watermarked signals, where it is possible that additional parameters like the embedded message $\mathbf{M}$ or the host $\mathbf{X}$ are also known by the attacker. The model is valid for either side-informed and non-side-informed watermarking/data-hiding schemes.

2. The scenario depicted in Figure 1-b covers the so-called *oracle attacks*. In this case, the attacker tries to gain knowledge about the secret key $\mathbf{\Theta}$ by observing the outputs $\hat{\mathbf{M}}$ of the detector/decoder corresponding to some selected inputs $\mathbf{Y}$, so the information leakage is measured by

$$I(\hat{\mathbf{M}}^1, \ldots, \hat{\mathbf{M}}^{N_o}, \mathbf{Y}^1, \ldots, \mathbf{Y}^{N_o}; \mathbf{\Theta}),$$

where, in this case, $\mathbf{Y}^n$ are not necessarily watermarked objects but any arbitrary signal, for instance the result of the iterations of an attacking algorithm.

The translation of Shannon's approach to the continuous case is straightforward; we only must be careful with the concept of differential entropies, in order to redefine properly the *unicity distance* for continuous random variables: in this case, an attacker will have perfect knowledge of the key when $h(\mathbf{\Theta}|\mathbf{Y}^1, \mathbf{Y}^2, \ldots, \mathbf{Y}^{N_o}) = -\infty$. Hence, the security level is the number N_o of observations required to reach the unicity distance. However, since this number is ∞ in general, the security level could be measured by the growth-rate of mutual information with the number of observations N_o; another possibility is the establishment of a threshold in the value of the equivocation, which is directly related to the minimum error variance in the estimation of the key:

$$\sigma_E^2 \geq \frac{1}{2\pi e} e^{2h(\mathbf{\Theta}|\mathbf{Y})}, \tag{3}$$

where σ_E^2 is the estimation error variance. For an attack based on the key estimate, its probability of success is given by the variance of the estimation error. This way, we can give the following definition:

Definition 3. *Given a required probability of success of an attack P_s, let σ_E^2 be the resulting variance of the secret key estimation error. Then, the security level is the minimum number of observations N_o^* needed to satisfy inequality (3).*

For the measure of security to be well defined, at least two of the three quantities involved in (2) must be given, because important information about the security of the system may be masked when only one of those quantities is available:

[3] The observations are independent signals watermarked with the same secret key $\boldsymbol{\Theta}$.

– The value of $h(\mathbf{\Theta})$ is only the a priori uncertainty about the key, so it does not depend on the system itself.
– The value of $I(\mathbf{Y}^1, \mathbf{Y}^2, \ldots, \mathbf{Y}^{N_o}; \mathbf{\Theta})$ shows the amount of information about the key that leaks from the observations, but a smaller information leakage does not necessarily imply a higher security level: notice that, for example, a deterministic key would yield null information leakage, but the security is also null.
– The value of the equivocation $h(\mathbf{\Theta}|\mathbf{Y}^1, \mathbf{Y}^2, \ldots, \mathbf{Y}^{N_o})$ is indicative of the remaining uncertainty about the key, but it does not reflect what is the a priori uncertainty.

3 Theoretical evaluation of security

In this section some theoretical measures about the residual entropy will be presented. The notation is borrowed from [7]: N_v will denote the length of the vectors (number of samples in each observation), N_o the number of observations, and N_c the number of carriers (or hidden symbols). After some modifications in the nomenclature described in [7], the following attacks will be analyzed:

– Known Message Attack (KMA): In this case the mutual information between the received signal and the secret key, when the sent message is known by the attacker, should be computed:

$$\begin{aligned} I(\mathbf{Y}^1, \cdots, \mathbf{Y}^{N_o}; \mathbf{\Theta}|\mathbf{M}^1, \cdots, \mathbf{M}^{N_o}) &= h(\mathbf{Y}^1, \cdots, \mathbf{Y}^{N_o}|\mathbf{M}^1, \cdots, \mathbf{M}^{N_o}) \\ &\quad - h(\mathbf{Y}^1, \cdots, \mathbf{Y}^{N_o}|\mathbf{\Theta}, \mathbf{M}^1, \cdots, \mathbf{M}^{N_o}), \end{aligned}$$

so the residual entropy will be

$$\begin{aligned} h(\mathbf{\Theta}|\mathbf{Y}^1, \cdots, \mathbf{Y}^{N_o}, \mathbf{M}^1, \cdots, \mathbf{M}^{N_o}) &= h(\mathbf{\Theta}) - h(\mathbf{Y}^1, \cdots, \mathbf{Y}^{N_o}|\mathbf{M}^1, \cdots, \mathbf{M}^{N_o}) \\ &\quad + h(\mathbf{Y}^1, \cdots, \mathbf{Y}^{N_o}|\mathbf{\Theta}, \mathbf{M}^1, \cdots, \mathbf{M}^{N_o}). \quad (4) \end{aligned}$$

– Watermarked Only Attack (WOA): The mutual information between the observations and the secret key is

$$I(\mathbf{Y}^1, \cdots, \mathbf{Y}^{N_o}; \mathbf{\Theta}) = h(\mathbf{Y}^1, \cdots, \mathbf{Y}^{N_o}) - h(\mathbf{Y}^1, \cdots, \mathbf{Y}^{N_o}|\mathbf{\Theta})$$

and the residual entropy will be

$$\begin{aligned} h(\mathbf{\Theta}|\mathbf{Y}^1, \cdots, \mathbf{Y}^{N_o}) &= h(\mathbf{\Theta}) - h(\mathbf{Y}^1, \cdots, \mathbf{Y}^{N_o}) + I(\mathbf{Y}^1, \cdots, \mathbf{Y}^{N_o}; \mathbf{M}^1, \cdots, \mathbf{M}^{N_o}|\mathbf{\Theta}) \\ &\quad + h(\mathbf{Y}^1, \cdots, \mathbf{Y}^{N_o}|\mathbf{\Theta}, \mathbf{M}^1, \cdots, \mathbf{M}^{N_o}). \end{aligned}$$

– Estimated Original Attack (EOA): In this case the following will be computed

$$\begin{aligned} I(\mathbf{Y}^1, \cdots, \mathbf{Y}^{N_o}; \mathbf{\Theta}|\hat{\mathbf{X}}^1, \cdots, \hat{\mathbf{X}}^{N_o}) &= h(\mathbf{Y}^1, \cdots, \mathbf{Y}^{N_o}|\hat{\mathbf{X}}^1, \cdots, \hat{\mathbf{X}}^{N_o}) \\ &\quad - h(\mathbf{Y}^1, \cdots, \mathbf{Y}^{N_o}|\mathbf{\Theta}, \hat{\mathbf{X}}^1, \cdots, \hat{\mathbf{X}}^{N_o}), \quad (5) \end{aligned}$$

where $\hat{\mathbf{X}}^i \triangleq \mathbf{X}^i + \tilde{\mathbf{X}}^i$ is an estimate of $\mathbf{X}^i$ and $\tilde{\mathbf{X}}^i$ is the estimation error; $\tilde{\mathbf{X}}^i$ is assumed to have power E and to be independent of $\mathbf{X}^i$. The Known Original Attack (KOA) proposed in [7] can be regarded to as a particular case of EOA, where the variance of the original host estimation error is set to 0. On the other hand, when the original host estimation error is σ_X^2, we are in the WOA case, so it can be also seen as particular case of EOA. The attacker could obtain this estimate by averaging several versions of the same host watermarked with different keys, but in order to ensure independence between the key and the estimate, the watermarked version with the to-be-estimated key should not be included in the averaging. Other alternative could be to filter the watermarked signal to compute the estimate of the original host (assuming the resulting signal is independent of the watermark). Taking into account (5), it is possible to write

$$h(\boldsymbol{\Theta}|\mathbf{Y}^1,\cdots,\mathbf{Y}^{N_o},\hat{\mathbf{X}}^1,\cdots,\hat{\mathbf{X}}^{N_o}) = h(\boldsymbol{\Theta}) - h(\mathbf{Y}^1,\cdots,\mathbf{Y}^{N_o}|\hat{\mathbf{X}}^1,\cdots,\hat{\mathbf{X}}^{N_o}) + h(\mathbf{Y}^1,\cdots,\mathbf{Y}^{N_o}|\boldsymbol{\Theta},\hat{\mathbf{X}}^1,\cdots,\hat{\mathbf{X}}^{N_o}).$$

Finally, note that, depending on the method, the secret key could be related to the watermarking scheme parameters (i.e. the spreading sequence in spread-spectrum, the dither sequence in SCS or the codebooks in Costa schemes with random codebooks) through a deterministic function, constructing a Markov chain, in such a way that the attacker could be interested just in estimating the result of this function and not in the secret key itself. When $N_o = 1$, the superscript denoting the observation will be obviated for notation simplicity.

4 Security Analysis of Spread Spectrum Watermarking

For these methods, N_c random vectors (the spreading sequences), denoted by $\mathbf{U}_i$ are generated depending on the secret key $\boldsymbol{\Theta}$. In this way, the embedding function can be written as:

$$\mathbf{Y}^j = \mathbf{X}^j + \frac{1}{\sqrt{N_c}} \sum_{i=1}^{N_c} \mathbf{U}_i (-1)^{M_i^j}, \;\; 1 \leq j \leq N_o, \tag{6}$$

with $\mathbf{Y}^j$, $\mathbf{X}^j$ and $\mathbf{U}_i$ N_v-dimensional vectors and $U_{i,j}$ is the j-th component of the i-th of the spreading sequence. The host is modeled as an i.i.d. Gaussian process, $\mathbf{X}^j \sim \mathcal{N}(0, \sigma_X^2 \mathbf{I}_{N_v})$, and the message letters $M_i^j \in \{0,1\}$, being $Pr\{M_i^j = 0\} = Pr\{M_i^j = +1\} = 1/2$. All of these quantities are assumed to be mutually independent. Since (6) is related with $\boldsymbol{\Theta}$ only through the $\mathbf{U}_i$'s, we will measure the security with respect to the $\mathbf{U}_i$'s.

4.1 Known Message Attack

To compute $I(\mathbf{Y}; \mathbf{U}_1, \mathbf{U}_2, \ldots, \mathbf{U}_{N_c}|\mathbf{M})$ (so $N_o = 1$) for a generic distribution of $\mathbf{U}_i$ numerical integration must be used. In Fig. 2 and Fig. 3 the results of this

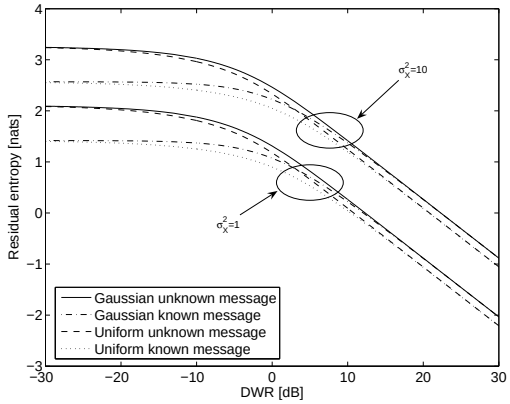

Fig. 2. Results of numerical integration for the equivocation $h(\mathbf{U}_1|\mathbf{Y})$ and $h(\mathbf{U}_1|M,\mathbf{Y})$ in spread-spectrum for Gaussian and uniform distributions of $\mathbf{U}_1$ and $N_v = 1$.

numerical integration are shown for $N_c = 1$ and both Gaussian and uniform distributions of $\mathbf{U}_1$ in the scalar case. Those figures show that the information the attacker can not learn (i.e., $h(\mathbf{U}_1|\mathbf{Y})$) is larger if $\mathbf{U}_1$ is chosen to be Gaussian. Taking this into account, we will focus on the case $\mathbf{U}_i \sim \mathcal{N}(0, \sigma_U^2 \mathbf{I}_{N_v})$. When the sent symbol is known to the attacker, the following result is derived in Appendix A.1 for $N_v > 1$, $N_c > 1$ and $N_o = 1$,

$$I(\mathbf{Y}; \mathbf{U}_1, \mathbf{U}_2, \ldots, \mathbf{U}_{N_c}|\mathbf{M}) = \frac{N_v}{2} \log\left(1 + \frac{\sigma_U^2}{\sigma_X^2}\right), \tag{7}$$

yielding

$$h(\mathbf{U}_1, \mathbf{U}_2, \ldots, \mathbf{U}_{N_c}|\mathbf{Y}, \mathbf{M}) = \frac{N_v}{2} \log\left[\left(2\pi e \frac{\sigma_U^2}{N_c}\right)^{N_c} \cdot \frac{\sigma_X^2}{\sigma_X^2 + \sigma_U^2}\right].$$

The result in (7) says that the information that an attacker can obtain is the same whatever the number of carriers, although the entropy of the key is a linear function of this parameter (this result applies to a great variety of pdf's for the key, since by the central limit theorem, the sum of the carriers tends to a Gaussian). This result is also a consequence of the power normalization performed in (6); independently of the number of carriers, the power of the watermark stays constant.

In App. A.2, we analyze the case of one sent bit ($N_c = 1$), $N_v = 1$, when there are several available observations ($N_o > 1$), all of them watermarked with the same secret key. If $N_v > 1$ and the components are independent, the result is also valid, after multiplying it by N_v, so we can write

$$I(\mathbf{Y}^1, \cdots, \mathbf{Y}^{N_o}; \mathbf{U}_1|\mathbf{M}^1, \cdots, \mathbf{M}^{N_o}) = \frac{N_v}{2} \log\left(1 + \frac{N_o \sigma_U^2}{\sigma_X^2}\right), \tag{8}$$

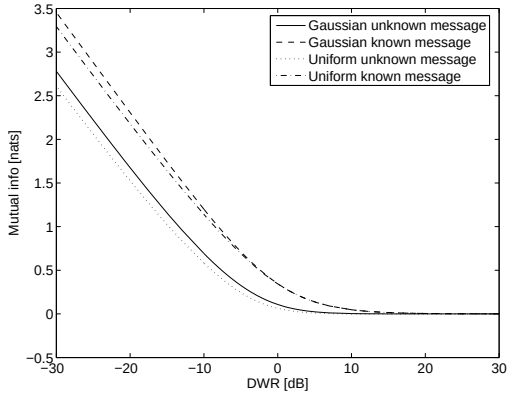

Fig. 3. Results of numerical integration for $I(\mathbf{Y};\mathbf{U}_1)$ and $I(\mathbf{Y};\mathbf{U}_1|M)$ in spread-spectrum for Gaussian and uniform distribution of $\mathbf{U}_1$ and $N_v = 1$.

which yields

$$h(\mathbf{U}_1|\mathbf{Y}^1,\cdots,\mathbf{Y}^{N_o},\mathbf{M}^1,\cdots,\mathbf{M}^{N_o}) = \frac{N_v}{2}\log\left(2\pi e\frac{\sigma_U^2\sigma_X^2}{N_o\sigma_U^2+\sigma_X^2}\right) \tag{9}$$

This result shows that $I(\mathbf{Y}^1,\cdots,\mathbf{Y}^{N_o};\mathbf{U}_1|\mathbf{M}^1,\cdots,\mathbf{M}^{N_o})$ grows non-linearly with the number of observations, although for large Document to Watermark Ratios[4] (DWR $>>$ 0) and low values of N_o it grows almost linearly. Moreover, (8) coincides with the capacity of a Gaussian channel with signal power σ_U^2 and noise power σ_X^2/N_o. This suggests that the best method the attacker should follow for estimating $\mathbf{U}_1$ is just to average the observations $\mathbf{Y}^i$ (at least this is the case when both the host signal and the watermark are Gaussian distributed). In Fig. 4 the mutual information is compared with its linear version when DWR = 30 dB.

4.2 Comparison with the result in [7]

In [7], the security level is defined as $O(N_o^\star)$, where $N_o^\star \triangleq N_o\text{tr}(\text{FIM}(\boldsymbol{\theta})^{-1})$ with FIM($\boldsymbol{\theta}$) the Fisher Information Matrix. In this section we try to link the result obtained in that paper with the one obtained here for spread-spectrum KMA when $N_c = 1$.

It is shown in App. B that the FIM obtained when a constant multiple (vectorial) parameter is estimated in the presence of i.i.d. Gaussian noise, taking into account N_o independent observations in the estimate, is $\frac{N_o}{\sigma_X^2}\mathbf{I}_{N_v}$, where σ_X^2 is the power of the interfering signal (the original host in our case). This is the only term considered in [7]. Nevertheless, an additional term should be taken

[4] The Document to Watermark Ratio is defined as DWR$\triangleq 10\log_{10}\left(\frac{\sigma_X^2}{\sigma_U^2}\right)$.

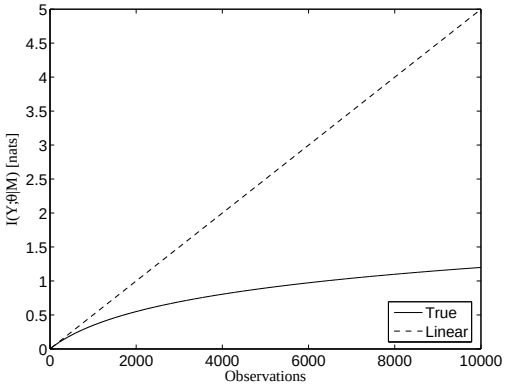

Fig. 4. $I(\mathbf{Y}^1, \cdots, \mathbf{Y}^{N_o}; \boldsymbol{\theta} | \mathbf{M}^1, \cdots, \mathbf{M}^{N_o})$ for spread-spectrum and Known Message Attack. DWR = 30 dB.

into account, due to the random nature of the secret key (see [13]):

$$J_{P_{ij}} = \mathrm{E}\left[\frac{\partial \log f_{\mathbf{U}_1}(\mathbf{u}_1)}{\partial u_{1,i}} \cdot \frac{\partial \log f_{\mathbf{U}_1}(\mathbf{u}_1)}{\partial u_{1,j}}\right]. \tag{10}$$

If $\mathbf{U}_1$ is an i.i.d. Gaussian vector, it is easy to prove that $\mathbf{J}_P = \frac{1}{\sigma_U^2}\mathbf{I}_{N_v}$, so $\mathrm{FIM}(\mathbf{U}_1) = \left(\frac{N_o}{\sigma_X^2} + \frac{1}{\sigma_U^2}\right)\mathbf{I}_{N_v}$, yielding

$$N_o^\star = N_v \frac{\sigma_X^2 \sigma_U^2}{\sigma_U^2 + \sigma_X^2/N_o},$$

which is obviously related with the proposed information-theoretic approach, since (9) is the differential entropy of a i.i.d. Gaussian random vector with covariance matrix $N_o^\star/(N_o N_v)\mathbf{I}_{N_v}$.

On the other hand, if we considered only the FIM obtained when estimating a constant multiple parameter, the obtained $N_o^\star$ is $N_v\sigma_X^2$, which is obviously related with $h(\mathbf{Y}^1, \cdots, \mathbf{Y}^{N_o} | \mathbf{U}_1, \mathbf{M}^1, \cdots, \mathbf{M}^{N_o}) = \frac{N_v N_o}{2}\log(2\pi e \sigma_X^2)$; this was the methodology followed in [7]. Therefore, it does not take into account the entropy of the secret key neither the entropy of the watermarked signal. As stated in Sect. 2, both terms are relevant for the analysis of the system, so they should be considered. In fact, $h(\mathbf{Y}^1, \cdots, \mathbf{Y}^{N_o} | \mathbf{U}_1, \mathbf{M}^1, \cdots, \mathbf{M}^{N_o})$ for the KMA case grows linearly with the number of observations, while the mutual information will not increase linearly due to the dependence between observations. The linear approximation is actually an upper-bound; the larger the number of observations, the worse this approximation is.

4.3 Watermarked Only Attack

Due to the symmetry of the pdf's, it is possible to conclude that the components of the vector $\mathbf{Y}$ are still mutually independent, so for $N_c = 1$ and a single

observation, we can write

$$I(\mathbf{Y};\mathbf{U}_1) = N_v I(Y_i;U_{1,i}) = N_v\left(h(Y_i) - h(Y_i|U_{1,i})\right) \tag{11}$$
$$= N_v\left(h(Y_i|\mathbf{M}=\mathbf{0}) - h(Y_i|U_{1,i})\right). \tag{12}$$

In order to determine this for a generic distribution of $\mathbf{U}_1$, numerical integration should be used, whose results are plotted in Fig. 2. Once again, the information the attacker can not learn ($h(\mathbf{U}_1|\mathbf{Y})$) is larger for the shown cases when $\mathbf{U}_1$ is chosen to be Gaussian. Therefore, assuming $\mathbf{U}_1$ to be Gaussian, we can write

$$I(\mathbf{Y};\mathbf{U}_1) = N_v\left(\frac{1}{2}\log\left(2\pi e(\sigma_X^2+\sigma_U^2)\right) - h(Y_i|U_{1,i})\right). \tag{13}$$

The rightmost term of (13) must still be numerically computed. When DWR $<<$ 0 we can easily analyze the asymptotic behavior of the mutual information taking into account $h(\mathbf{Y}) \approx h(\mathbf{U}_1)$ and $h(\mathbf{Y}|\mathbf{U}_1) \approx h(\mathbf{X}) + \log(2)$, yielding

$$I(\mathbf{Y};\mathbf{U}_1) \approx h(\mathbf{U}_1) - h(\mathbf{X}) - \log(2), \tag{14}$$
$$I(\mathbf{Y};\mathbf{U}_1|\mathbf{M}) \approx h(\mathbf{U}_1) - h(\mathbf{X}). \tag{15}$$

This explains and quantifies the gap between the WOA and KMA cases, which is exactly $\log(2) = 0.69$ nats. Nevertheless, note that a very small DWR is not practical, since it would yield unuseful watermarked images. This case has been introduced here only to shed some light into the general behavior of the mutual informations. On the other hand, to compute the gap between a Gaussian and a uniform distribution for $\mathbf{U}_1$, $h(\mathbf{U}_1)$ will be determined in both cases for a constant variance σ_U^2,

$$h(\mathbf{U}_{Gauss}) - h(\mathbf{U}_{unif}) = \frac{1}{2}\log(2\pi e\sigma_U^2) - \frac{1}{2}\log(12\sigma_U^2) = \frac{1}{2}\log\left(\frac{\pi e}{6}\right) = 0.1765,$$

which will be the asymptotic gap (in residual entropy terms) between the Gaussian and uniform cases for both known and unknown messages (see Fig. 2) when DWR $>>$ 0, since for a large DWR both $I(\mathbf{Y};\mathbf{U}_1)$ and $I(\mathbf{Y};\mathbf{U}_1|\mathbf{M})$ are approximately 0.

For N_c carriers and $N_o = 1$ we have, similarly to the KMA case, the following mutual information:

$$\begin{aligned} I(\mathbf{Y};\mathbf{U}_1,\mathbf{U}_2,\ldots,\mathbf{U}_{N_c}) &= N_v I(Y_i;U_{1,i},U_{2,i},\ldots,U_{N_c,i}) \\ &= N_v\left(h(Y_i) - h(Y_i|U_{1,i},\ldots,U_{N_c,i})\right) \\ &= N_v\left[\frac{1}{2}\log(2\pi e(\sigma_x^2+\sigma_u^2)) - h(Y_i|U_{1,i},\ldots,U_{N_c,i})\right], \end{aligned}$$

where the second term of the last equality must be numerically computed again.

The case of one sent bit ($N_c = 1$), $N_v = 1$, and several available observations ($N_o > 1$) needs very expensive numerical computations. Practical computations demand the reduction of the number of available observations to a very small value; in that case, the mutual information will be in the linear region, so no knowledge is available about the growth of the mutual information for large values of N_o.

4.4 Estimated Original Attack

In this case, the attacker will have access to an estimate of the original host signal, with some estimation error denoted by $\tilde{\mathbf{X}}$, which is assumed to be i.i.d. Gaussian with variance E, in such a way that for $N_o = 1$ we can write $I(\mathbf{Y}; \mathbf{U}_1, \cdots, \mathbf{U}_{N_c} | \mathbf{X} + \tilde{\mathbf{X}}) = N_v \left[h(Y_i | X_i + \tilde{X}_i) - h(Y_i | X_i + \tilde{X}_i, U_{1,i}, \cdots, U_{N_c,i}) \right]$. Assuming $\sigma_X^2 >> E$, $\tilde{X}_i$ will be almost orthogonal (and therefore independent) to $X_i + \tilde{X}_i$, so

$$I(\mathbf{Y}; \mathbf{U}_1, \cdots, \mathbf{U}_{N_c} | \mathbf{X} + \tilde{\mathbf{X}}) \approx N_v \left\{ h\left(\frac{1}{\sqrt{N_c}} \sum_{j=1}^{N_c} U_{j,i} (-1)^{M_j} - \tilde{X}_i \right) - h\left(\frac{1}{\sqrt{N_c}} \sum_{j=1}^{N_c} U_{j,i} (-1)^{M_j} - \tilde{X}_i | U_{1,i}, \cdots, U_{N_c,i} \right) \right\}.$$

This situation is equivalent to that described in 4.3, but replacing σ_X^2 by E, so when $N_c = 1$ it is possible to use Fig. 2 for obtaining numerical results, using the *Estimation error to Watermark Ratio* (EWR), defined as $10 \log_{10} \left(\frac{E}{\sigma_U^2} \right)$, instead of the DWR, in the horizontal axis. When the estimate is perfect, i.e. $\sigma_{\tilde{x}}^2 = 0$, the mutual information approaches infinity.

5 Conclusions

In this paper, an overview of watermarking security has been introduced, showing the evolution of this concept in the last years. The frontier between *security* and *robustness* is rather fuzzy, so we have proposed some definitions in order to make a clear distinction between these two concepts, which in turn allows the isolation of the security analysis from the robustness issue. Based on these definitions, a new information-theoretic framework to evaluate watermarking security has been introduced based on the use of mutual information to measure the secret key leakage; this measure has been shown to be more complete than the measure proposed in [7], which was based on the FIM and did not take into account the term related with the variability of the secret key. Security of Spread Spectrum watermarking has been analyzed in different scenarios classified by the amount of information available to the attacker, quantifying the information leakage about the key as a function of the number of observations and the DWR.

A Calculation of mutual information for spread spectrum

A.1 Known Message Attack (KMA) for a single observation

For a single observation ($N_o = 1$) and $N_c = 1$, we have

$$I(\mathbf{Y};\mathbf{U}_1|\mathbf{M}) = \sum_{i=1}^{N_v}\sum_{j=1}^{N_v} I(Y_i;U_{1,j}|\mathbf{M},Y_{i-1},\ldots,Y_1,U_{1,j},\ldots,U_{1,1}) \tag{16}$$

$$= \sum_{i=1}^{N_v} I(Y_i;U_{1,i}|\mathbf{M},Y_{i-1},\ldots,Y_1) \tag{17}$$

$$= \sum_{i=1}^{N_v} I(Y_i;U_{1,i}|\mathbf{M}) = N_v I(Y_i;U_{1,i}|\mathbf{M}), \tag{18}$$

where (17) follows from the fact that Y_i and $U_{1,j}$ are independent $\forall\ i \neq j$; (**??**) follows from the independence between the components of $\mathbf{Y}$ given the message, and (18) follows from the fact that $\mathbf{Y}$ and $\mathbf{U}_1$ are i.i.d. processes. The theoretical expression for (18) is easy to calculate:

$$I(Y_i;U_{1,i}|\mathbf{M}) = I(Y_i;U_{1,i}|\mathbf{M}=\mathbf{0}) = h(Y_i|\mathbf{M}=\mathbf{0}) - h(Y_i|\mathbf{M}=\mathbf{0},U_{1,i}),$$

where $h(Y_i|\mathbf{M}=\mathbf{0})$ will obviously depend on the distribution of $U_{1,i}$. Assuming $\mathbf{U}_1$ to be Gaussian, i.e. $\mathbf{U}_1 \sim \mathcal{N}(0,\sigma_U^2\mathbf{I}_{N_v})$, we can write

$$I(Y_i;U_{1,i}|\mathbf{M}) = h(\mathcal{N}(0,\sigma_X^2+\sigma_U^2)) - h(\mathcal{N}(0,\sigma_X^2)) = \frac{1}{2}\log\left(1+\frac{\sigma_U^2}{\sigma_X^2}\right).$$

Next, the case of multiple carriers is analyzed. When $N_c > 1$, we can write

$$\begin{aligned} I(\mathbf{Y};\mathbf{U}_1,\mathbf{U}_2,\ldots,\mathbf{U}_{N_c}|\mathbf{M}) &= N_v I(Y_i;U_{1,i},U_{2,i},\ldots,U_{N_c,i}|\mathbf{M}) \\ &= N_v\left\{h(Y_i|\mathbf{M}) - h(Y_i|U_{1,i},\ldots,U_{N_c,i},\mathbf{M})\right\} \\ &= N_v\left\{h\left(X_i+\textstyle\sum_{j=1}^{N_c}(N_c)^{-1/2}U_{j,i}\right) - h(X_i)\right\} \\ &= N_v\left\{h(\mathcal{N}(0,\sigma_X^2+\sigma_U^2)) - h(\mathcal{N}(0,\sigma_X^2))\right\} \\ &= \frac{N_v}{2}\log\left(1+\frac{\sigma_U^2}{\sigma_X^2}\right). \end{aligned} \tag{19}$$

A.2 Known Message Attack (KMA) for multiple observations

When $N_v = 1$, there are several available observations ($N_o > 1$) watermarked with the same secret key and there is one bit to be sent in each observation ($N_c = 1$) which we will assume without loss of generality to be the same for all the observations, it can be seen that the covariance matrix of $(\mathbf{Y}^1,\cdots,\mathbf{Y}^{N_o})$, denoted by $R_{\mathbf{Y}}$, becomes

$$R_{\mathbf{Y}} = \begin{pmatrix} \sigma_X^2+\sigma_U^2 & \sigma_U^2 & \cdots & \sigma_U^2 \\ \sigma_U^2 & \sigma_X^2+\sigma_U^2 & \cdots & \sigma_U^2 \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_U^2 & \sigma_U^2 & \cdots & \sigma_X^2+\sigma_U^2 \end{pmatrix},$$

so its entropy is (see [14])

$$h(\mathbf{Y}^1,\cdots,\mathbf{Y}^{N_o}) = \frac{1}{2}\log\left((2\pi e)^{N_o}|R_{\mathbf{Y}}|\right) = \frac{1}{2}\log\left((2\pi e)^{N_o}\left[\frac{N_o\sigma_U^2}{\sigma_X^2}+1\right]\sigma_X^{2N_o}\right),$$

and we can write $I(\mathbf{Y}^1,\cdots,\mathbf{Y}^{N_o};\mathbf{U}_1|\mathbf{M}^1,\cdots,\mathbf{M}^{N_o}) = \frac{1}{2}\log\left(1+\frac{N_o\sigma_U^2}{\sigma_X^2}\right)$.

B Fisher Information Matrix for SS-KMA

In this section we will compute the Fisher Information Matrix of the estimate of the constant multiple parameter $\boldsymbol{\theta}$ taking into account the observations $\mathbf{Y}^1,\cdots,\mathbf{Y}^{N_o}$. Let us consider $\mathbf{Y}^j = \mathbf{X}^j + \boldsymbol{\theta}$, with $\mathbf{X}^j \sim \mathcal{N}(0,\sigma_X^2\mathbf{I}_{N_V})$, and the $\mathbf{X}^j$'s to be mutually independent for $1 \le j \le N_o$ [5]. Following the definition of Fisher Information Matrix ([13]), we can write

$$\mathrm{FIM}_{ii}(\boldsymbol{\theta}) = \int f(\mathbf{y}^1,\cdots,\mathbf{y}^{N_o}|\boldsymbol{\theta})\left(\frac{\partial}{\partial\theta_i}\log f(\mathbf{y}^1,\cdots,\mathbf{y}^{N_o}|\boldsymbol{\theta})\right)^2 d\mathbf{y}^1\cdots d\mathbf{y}^{N_o},$$

where $f(\mathbf{y}^1,\cdots,\mathbf{y}^{N_o}|\boldsymbol{\theta}) = \prod_{k=1}^{N_v}\prod_{j=1}^{N_o}\frac{1}{\sqrt{2\pi\sigma_X^2}}e^{\frac{-(y_k^j-\theta_k)^2}{2\sigma_X^2}}$, in such a way that

$$\frac{\partial}{\partial\theta_i}\log f(\mathbf{y}^1,\cdots,\mathbf{y}^{N_o}|\boldsymbol{\theta}) = \sum_{j=1}^{N_o}\frac{y_i^j-\theta_i}{\sigma_X^2} = \frac{\sum_{j=1}^{N_o}x_i^j}{\sigma_X^2},$$

and, finally, after a variable change,

$$\mathrm{FIM}_{ii}(\boldsymbol{\theta}) = \int\left(\frac{\sum_{j=1}^{N_o}x_i^j}{\sigma_X^2}\right)^2\prod_{j=1}^{N_o}\frac{1}{\sqrt{2\pi\sigma_X^2}}e^{\frac{-(x_i^j)^2}{2\sigma_X^2}}dx_i^1\cdots dx_i^{N_o} = \frac{N_o}{\sigma_X^2},\ 1\le i\le N_v.$$

On the other hand,

$$\begin{aligned}\mathrm{FIM}_{ik}(\boldsymbol{\theta}) &= \int f(\mathbf{y}^1,\cdots,\mathbf{y}^{N_o}|\boldsymbol{\theta})\left(\frac{\partial}{\partial\theta_i}\log f(\mathbf{y}^1,\cdots,\mathbf{y}^{N_o}|\boldsymbol{\theta})\right)\\ &\quad\left(\frac{\partial}{\partial\theta_k}\log f(\mathbf{y}^1,\cdots,\mathbf{y}^{N_o}|\boldsymbol{\theta})\right)d\mathbf{y}^1\cdots d\mathbf{y}^{N_o}\\ &= \left(\int\frac{\sum_{j=1}^{N_o}x_i^j}{\sigma_X^2}\prod_{j=1}^{N_o}\frac{1}{\sqrt{2\pi\sigma_X^2}}e^{\frac{-(x_i^j)^2}{2\sigma_X^2}}dx_i^1\cdots dx_i^{N_o}\right)\\ &\quad\cdot\left(\int\frac{\sum_{l=1}^{N_o}x_k^l}{\sigma_X^2}\prod_{l=1}^{N_o}\frac{1}{\sqrt{2\pi\sigma_X^2}}e^{\frac{-(x_k^l)^2}{2\sigma_X^2}}dx_k^1\cdots dx_k^{N_o}\right) = 0,\ \text{for all } i\ne k,\end{aligned}$$

so, we can conclude $\mathrm{FIM}(\boldsymbol{\theta}) = \frac{N_o}{\sigma_X^2}\mathbf{I}_{N_v}$.

[5] Be aware that this is the case described in Sect. 4.1 for $N_c = 1$, after multiplying the j-th observation by $(-1)^{M_1^j}$. In that case, the parameter to be estimated is $\mathbf{U}_1$.

References

1. Cox, I.J., Linnartz, J.P.M.G.: Some general methods for tampering with watermarks. IEEE Journal on Selected Areas in Communications **16** (1998) 587–593
2. Linnartz, J.P.M.G., van Dijk, M.: Analysis of the sensitivity attack against electronic watermarks in images. In Aucsmith, D., ed.: 2nd Int. Workshop on Information Hiding, IH'98. Volume 1525 of Lecture Notes in Computer Science., Portland, OR, USA, Springer Verlag (1998) 258–272
3. Mitthelholzer, T.: An information-theoretic approach to steganography and watermarking. In Pfitzmann, A., ed.: 3rd Int. Workshop on Information Hiding, IH'99. Volume 1768 of Lecture Notes in Computer Science., Dresden, Germany, Springer Verlag (1999) 1–17
4. Shannon, C.E.: Communication theory of secrecy systems. Bell system technical journal **28** (1949) 656–715
5. Kalker, T.: Considerations on watermarking security. In: IEEE Int. Workshop on Multimedia Signal Processing, MMSP'01, Cannes, France (2001) 201–206
6. Barni, M., Bartolini, F., Furon, T.: A general framework for robust watermarking security. Signal Processing **83** (2003) 2069–2084
7. Cayre, F., Fontaine, C., Furon, T.: Watermarking security part one: theory. In Edward J. Delp III, Wong, P.W., eds.: Security, Steganography, and Watermarking of Multimedia Contents VII. Volume 5681., San Jose, California, USA, SPIE (2005) 746–757
8. Furon, T., et al.: Security Analysis. European Project IST-1999-10987 CERTIMARK, Deliverable D.5.5 (2002)
9. Fisher, R.A.: On the mathematical foundations of theoretical statistics. Philosophical Transactions of the Royal Society **222** (1922) 309–368
10. Barni, M., Bartolini, F.: Watermarking Systems Engineering. Signal Processing and Communications. Marcel Dekker (2004)
11. Hernández, J.R., Pérez-González, F.: Throwing more light on image watermarks. In Aucsmith, D., ed.: 2nd Int. Workshop on Information Hiding, IH'98. Volume 1525 of Lecture Notes in Computer Science., Portland, OR, USA, Springer Verlag (1998) 191–207
12. Kerckhoff, A.: La cryptographie militaire. Journal des sciences militaires **9** (1883) 5–38
13. van Trees, H.L.: Detection, Estimation, and Modulation Theory. John Wiley and Sons (1968)
14. Cover, T.M., Thomas, J.A.: Elements of Information Theory. Wiley series in Telecommunications (1991)