AN INFORMATION-THEORETIC FRAMEWORK FOR ASSESSING SECURITY IN PRACTICAL WATERMARKING AND DATA HIDING SCENARIOS

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

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

ABSTRACT

This paper provides a historical overview of the meaning of security in watermarking, putting special emphasis on some recent works. Inspired by these works, a definition of watermarking security is introduced and a quantitative measure of security is proposed, showing some new results on quantization-based and spread spectrum methods.

1. INTRODUCTION AND STATEMENT OF THE PROBLEM

Although a great amount of the watermarking and data-hiding 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.

During the first years, researchers focused their efforts on the design and study of attacks and countermeasures, overlooking the meaning of security in watermarking. As a result, most of the literature deals with the problem of robustness; at most, there was the notion of *intentional* and *non-intentional* attacks. The work in [1] shows an example of this type of classification, considering separately the so-called *signal transformations* (affine transformations, noise addition, compression) and the *intentional attacks*, introducing at a qualitative level concepts like the *sensitivity attack*, the *collusion attack* and attacks based on the availability of embedding devices. In [2], a complete characterization of the sensitivity attack for spread-spectrum-based methods [3] is given, and even an information-theoretic analysis is performed, measuring the information about the watermark (which depends on a secret key known only by authorized users) that an attacker can gain by each observation of the detector output; later, and following the ideas in [2], a practical method for accomplishing a successful sensitivity attack was proposed in [4], showing alarmingly good results, and raising up the problem of security in watermarking, since this method provided a simple way of fooling any spread-spectrum-based watermarking system, as long as a detector is available to the attacker.

The very first attempt at proposing a theoretical framework for assessing the security of a general watermarking scenario is the work in [5], which considers the problem of security in terms of secrecy of the embedded message, and introduces in watermarking the concept of *perfect secrecy*, directly borrowed from the work on cryptanalysis by Shannon in [6]. However, this approach did not take into account that some information about the secret key may leak from the observations.

The work in [7] came to shed some light on the concept of security in watermarking. In a context of robust watermarking, the following definitions are given:

- *"Robust watermarking is a mechanism to create a communication channel that is multiplexed into original content"*, and whose capacity *"degrades as a smooth function of the degradation of the marked content"*.
- *"Security refers to the inability by unauthorized users to have access to the raw watermarking channel"*. Such an access refers to *"remove, detect and estimate, write and modify the raw watermarking bits"*.

Hence, watermarking security is identified with attacks whose objective is not only the removal of the watermarks, as it is with robustness, but the given definition has the problem of being too general.

In [8], the former definitions of watermarking security are reviewed, identifying now security with intentional non-blind attacks, and robustness with common blind signal processing operations, where *blind* must be understood as *without knowledge of the watermarking technique*. A noticeable contribution of [8] to the study of security is the translation of Kerckhoff's principle from cryptography to the watermarking field: all functions (encoding/embedding, decoding/detection, ...) should be declared as public except for a parameter called the secret key. Furthermore, based on Diffie-Hellman's attacks classification for cryptography, a classification of attacks for watermarking is proposed, based on the amount of information available to the attacker.

Later on, in [9], a new framework to analyze watermarking security is proposed, based on modeling watermarking as a game with some rules; these rules determine which information (parameters of the algorithm, the algorithm itself, etc.) is public. This way, attacks are classified as *fair*, when the attacker only exploits the publicly available information, and *unfair*, when he does not observe the rules of the game. Furthermore, the authors also define in that paper the *security level* as *"the amount of observation,*

[*] 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 G03/185, and European Comission through the IST Programme under Contract IST-2002-507932 ECRYPT.

ECRYPT disclaimer: the information in this document reflects only the authors' views, is provided as is and no guarantee or warranty is given that the information is fit for any particular purpose. The user thereof uses the information at its sole risk and liability.

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 [10], which address the problem of making a clear distinction between robustness and security, but the most remarkable aspect of that paper is that the authors propose to measure the security of a watermarking system by quantifying the information about the secret key that leaks from the observation of watermarked signals, adopting the Fisher Information Matrix (FIM) [11] as measuring tool. The problem with this measure is that it can be shown to neglect some important parameters such as the uncertainty (differential entropy) of the secret key or the watermarked signal, so in the next section the use of a suitable measure will be proposed, and according to that measure, some theoretical results will be presented in Section 3. Finally, the conclusions will be summarized in Section 4.

2. FUNDAMENTAL DEFINITIONS

One of the objectives of this paper is the establishment of a clear distinction between the concepts of robustness and security. To this aim, the following definitions are proposed:

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

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

Note that in the definition of attacks to robustness we use the probability of error instead of channel capacity, because the latter might entail 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 have been fooled. Several implications of the above definitions are the following:

Security implies *intentionality*, but the converse is not necessarily true. For instance, an attacker may intentionally 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.

Security implies *non-blindness*, but the converse is not necessarily true. Bear in mind that 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, an attacker can increase the probability of error to 0.5 in a Dither-Modulation-based scheme simply by adding to each watermarked coefficient a quantity equal to half the quantization step, although he does not learn anything about the secrets of the system.

Many attacks to security constitute a first step towards performing attacks to robustness. For 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).

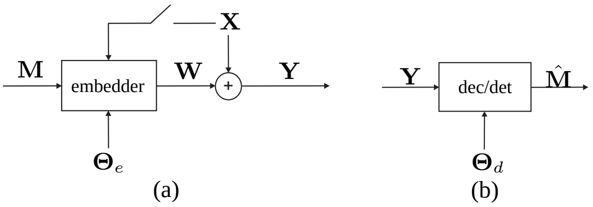

Fig. 1. General model for security analysis: embedding (a) and decoding/detection (b)

Security does not imply robustness at all. 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, those schemes which modify the decision boundary of a spread-spectrum-based scheme by means of fractal curve highly improve the security of the system, but they do not improve in any way the robustness of the method.

For assessing security, we will take into account Kerckhoff's principle, as in [8]. In order to measure the information leakage about the key, we propose a measure which is a direct translation of Shannon's approach [6] to the case of continuous random variables. We will distinguish between two different scenarios for security assessment, depicted in Fig. 1, which also allows us to introduce the notation: 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; $\hat{\mathbf{M}}$ denotes the estimated message in the case of 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. In the following, we will restrict our attention to the case of *symmetric* watermarking, i.e. $\mathbf{\Theta}_e = \mathbf{\Theta}_d = \mathbf{\Theta}$.

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

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

where $h(\cdot)$ stands for differential entropy, and $\mathbf{Y}^n$ denotes the n-th observation. Equivocation is defined as the remaining uncertainty about the key after the N_0 observations:

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

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 or data-hiding schemes.

2. The scenario depicted in Fig. 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

$$I(\hat{\mathbf{M}}^1,\cdots,\hat{\mathbf{M}}^{N_o};\boldsymbol{\Theta}|\mathbf{Y}^1,\cdots,\mathbf{Y}^{N_o}). \tag{3}$$

Clearly, an attacker will need to achieve $h(\boldsymbol{\Theta}|\mathbf{Y}^1,\cdots,\mathbf{Y}^{N_o}) = -\infty$ to completely disclose the secret key. Since the number of observations required to reach the unicity distance is ∞ in a general case, the security level can be measured by establishing a threshold in the value of the equivocation, which is directly related to the minimum error in the estimation of the key:

$$\sigma_E^2 \geq \frac{1}{2\pi e} e^{2h(\boldsymbol{\Theta}|\mathbf{Y})}. \tag{4}$$

For an attack based on the key estimate, its probability of success is given by the variance of the estimation error. This way, the security level could be defined as the minimum number of observations N_o^* needed to achieve the variance of the estimation error which yields the required probability of success. In order not to mask important information about the security of the system, at least two of the quantities in (2) must be given:

- The value of $h(\boldsymbol{\Theta})$ is only the a priori uncertainty about the key, so it does not depend on the system itself.
- The value of $I(\mathbf{Y}^1,\cdots,\mathbf{Y}^{N_o};\boldsymbol{\Theta})$ 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 $h(\boldsymbol{\Theta}|\mathbf{Y}^1,\cdots,\mathbf{Y}^{N_o})$ is indicative of the remaining uncertainty about the key, but it does not reflect what is the a priori uncertainty.

3. THEORETICAL RESULTS

In this section we present several results concerning security analysis in the scenario of Fig. 1-a, borrowing the notation from [10]. We have analyzed the case where the attacker has access to several independent documents watermarked with the same key and he also knows the embedded message on each document (this is the Known Message Attack proposed in [10]).

3.1. Spread Spectrum

In spread spectrum, the embedding function is

$$\mathbf{Y}^j = \mathbf{X}^j + \mathbf{U}(-1)^{M^j}, \;\; 1 \leq j \leq N_o, \tag{5}$$

with $\mathbf{Y}^j$, $\mathbf{X}^j$ and $\mathbf{U}$ (a pseudorandom spreading sequence), N_v-dimensional vectors. Clearly, in this setup, the spreading sequence plays the role of secret key. $\mathbf{X}^j$ and $\mathbf{U}$ are modeled as i.i.d. Gaussian processes, $\mathbf{X}^j \sim \mathcal{N}(0,\sigma_X^2\mathbf{I}_{N_v})$, $\mathbf{U} \sim \mathcal{N}(0,\sigma_U^2\mathbf{I}_{N_v})$, and the message letters $M^j \in \{0,1\}$, being $Pr\{M^j=0\} = Pr\{M^j = 1\} = 1/2$. All of these variables are assumed to be mutually independent. In this case, the mutual information after N_o observations can be shown to be

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

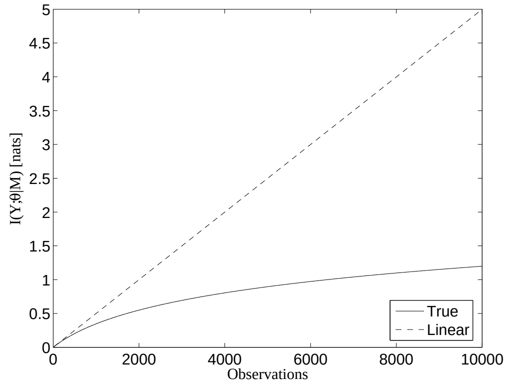

Fig. 2. $I(\mathbf{Y};\mathbf{U}|\mathbf{M})$ for spread-spectrum and Known Message Attack. DWR = 30dB, $N_v = 1$.

and the equivocation reads

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

Fig. 2 shows the mutual information in terms of the number of observations, comparing it with a linear upper bound obtained by assuming that all the observations provide the same amount of information as the first one. Note that the result of Fig. 2 is for DWR = 30 dB, where DWR stands for *Document to Watermark Ratio*, which is defined as DWR $= 10\log_{10}(\sigma_X^2/D_w)$, being D_w the embedding distortion, and in this case $D_w = \sigma_U^2$.

In [10] this same scenario was analyzed using the Fisher Information Matrix. The result obtained there can be shown to be related only to $h(\mathbf{Y}^1,\cdots,\mathbf{Y}^{N_o}|\mathbf{U},M^1,\cdots,M^{N_o})$, so it does not take into account the entropy of the secret key neither the entropy of the watermarked signal, whereas both of them are relevant for the analysis of the system, as it was discussed in Sect. 2. In fact, if only the former term was considered, the growth of the mutual information would be linear with the number of observations. An additional term accounting for the randomness of the secret key (see [12]) should be added to the FIM obtained in [10]; taking into account this modified FIM, the results in [10] can be shown to be equivalent to those obtained in this section.

3.2. DC-DM

DC-DM (Distortion Compensated - Dither Modulation) is a particular implementation of QIM [13]. We will restrict our attention to the case where the embedding lattices are formed by the cartesian product of identical scalar quantizers, thus embedding can be performed in a component-by-component basis:

$$y_k^j = x_k^j + \alpha\left(Q_{\Lambda_{k,j}}(x_k^j+d_k) - x_k^j - d_k\right), \tag{6}$$

where subindex k denotes the k-th component of vector $\mathbf{Y}^j$, α is the distortion compensation parameter, $Q_{\Lambda_{k,j}}$ is a uniform quantizer with its centroids defined by the points in the shifted lattice $\Lambda_{k,j}$, according to the symbol m_k^j

$$\Lambda_{k,j} = \Delta\mathbb{Z} + m_k^j\frac{\Delta}{|\mathcal{M}|},$$

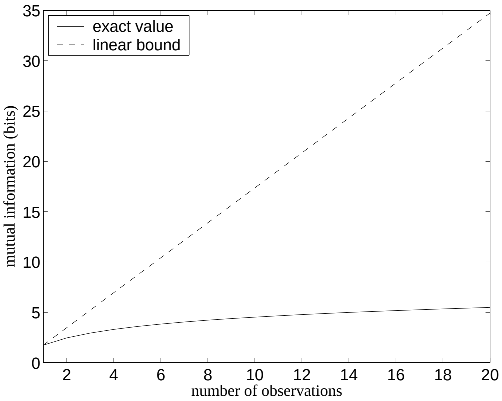

Fig. 3. $I(\mathbf{Y};\mathbf{D}|\mathbf{M})$ for scalar DC-DM and Known Message Attack, for $\alpha = 0.7$ and $N_v = 1$.

and d_k is a pseudorandom dither signal uniformly distributed in the range of a quantization bin, to achieve randomization of the codebook. Therefore, the dither plays the role of secret key in the security analysis. The mutual information when $\alpha > 0.5$ after N_o observations can be shown to be given by

$$I(\mathbf{Y}^1, \cdots, \mathbf{Y}^{N_o}; \mathbf{D}|\mathbf{M}^1, \cdots, \mathbf{M}^{N_o}) = N_v \left(-\log(1-\alpha) + \sum_{i=2}^{N_o} \frac{1}{i}\right) \text{ nats}, \quad (7)$$

and the calculation of the equivocation is straightforward, taking into account that $h(\mathbf{D}) = N_v \log(\Delta)$. Fig. 3 shows the mutual information as a function of the number of observations when $N_v = 1$, comparing it again to the linear upper bound. It is interesting to note that, contrarily to spread spectrum, the behavior of DC-DM is independent of the DWR as long as we can assume that the quantization step is sufficiently small[1]; should this not be true, it can be shown (by means of numerical integration, since the involved pdf's do not allow analytical evaluation) that the information leakage grows when the DWR is decreased, but significant changes only occur for very small values of the DWR (less than 10 dB, for instance), which result unpractical in most applications.

4. CONCLUSIONS AND FURTHER RESEARCH

We have made in this paper a review of the evolution of watermarking security concept. Considering this discussion and inspired by [10], definitions and an information theoretic measure of security have been proposed for watermarking and data-hiding scenarios. We have applied this measure to analyze the security of classical spread spectrum data hiding schemes, establishing a direct link between our measure and that used in [10]. Also, for the first time in the literature, a theoretical security analysis of DC-DM has been presented. We have seen that, in both cases, the information that the attacker can learn is a concave and monotonically increasing function with the number of observations. Open questions now are the extension of the security analysis to other scenarios (oracle attacks, unknown embedded message...) and other watermarking methods, as well as the establishment of proper thresholds in the variance of the estimation error (4).

5. ACKNOWLEDGMENTS

The authors want to thank Dr. T. Furon and members of WAVILA from the European Network of Excellence Ecrypt for their fruitful discussions.

6. REFERENCES

[1] I. J. Cox and J. P. M. G. Linnartz, "Some general methods for tampering with watermarks," *IEEE Journal on Selected Areas in Communications*, vol. 16, pp. 587–593, May 1998.

[2] J. P. M. G. Linnartz and M. van Dijk, "Analysis of the sensitivity attack against electronic watermarks in images," in *2nd Int. Workshop on Information Hiding, IH'98* (D. Aucsmith, ed.), vol. 1525 of *Lecture Notes in Computer Science*, (Portland, OR, USA), pp. 258–272, Springer Verlag, April 1998.

[3] I. J. Cox, J. Killian, T. Leighton, and T. Shamoon, "Secure spread spectrum watermarking for images, audio and video," *IEEE Transactions on Image Processing*, vol. 6, pp. 1673–1687, December 1997.

[4] T. Kalker, J. P. Linnartz, and M. van Dijk, "Watermark estimation through detector analysis," in *IEEE Int. Conf. on Image Processing, ICIP'98*, (Chicago, IL, USA), pp. 425–429, October 1998.

[5] T. Mitthelholzer, "An information-theoretic approach to steganography and watermarking," in *3rd Int. Workshop on Information Hiding, IH'99* (A. Pfitzmann, ed.), vol. 1768 of *Lecture Notes in Computer Science*, (Dresden, Germany), pp. 1–17, Springer Verlag, September 1999.

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

[7] T. Kalker, "Considerations on watermarking security," in *IEEE Int. Workshop on Multimedia Signal Processing, MMSP'01*, (Cannes, France), pp. 201–206, October 2001.

[8] T. Furon *et al.*, "Security Analysis," *European Project IST-1999-10987 CERTIMARK, Deliverable D.5.5*, 2002.

[9] M. Barni, F. Bartolini, and T. Furon, "A general framework for robust watermarking security," *Signal Processing*, vol. 83, pp. 2069–2084, February 2003.

[10] F. Cayre, C. Fontaine, and T. Furon, "Watermarking attack: Security of wss techniques," in *Proc. of Int. Workshop on Digital Watermarking*, (Seoul, Corea), IWDW'04, Springer-Verlag, Oct. 2004.

[11] R. A. Fisher, "On the mathematical foundations of theoretical statistics," *Philosophical Transactions of the Royal Society*, vol. 222, pp. 309–368, 1922.

[12] H. L. van Trees, *Detection, Estimation, and Modulation Theory.* John Wiley and Sons, 1968.

[13] B. Chen and G. Wornell, "Quantization Index Modulation: a class of provably good methods for digital watermarking and information embedding," *IEEE Transactions on Information Theory*, vol. 47, pp. 1423–1443, May 2001.

[1] In this case, $D_w = (\alpha\Delta)^2/12$.