Watermarking Security: a Survey

Luis Pérez-Freire, Pedro Comesaña, Juan Ramón Troncoso-Pastoriza,
Fernando Pérez-González

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

Abstract. Watermarking security has emerged in the last years as as a new subject in the watermarking area. As it brings new challenges to the design of watermarking systems, a good understanding of the problem is fundamental. This paper is intended to clarify the concepts related to watermarking security, provide an exhaustive literature overview, and serve as a starting point for newcomers interested in carrying out research on this topic.

1 Introduction

Watermarking security is an emergent topic. A good indicator of the growing interest in this subject is the number of special sessions that have been held in recent conferences [1,2,3,4] and the efforts made in relevant European projects such as Certimark [5] and Ecrypt [6]. Whereas robustness in watermarking has been generally identified with probability of decoding error or resistance against watermark removal, the concept of *watermarking security* is still somewhat fuzzy. In recent works, it has been agreed that attacks to security have a broader scope than attacks to robustness, since the former are not only concerned with a simple impairment of the communication process, but they also consider the achievement of privileges granted by the secret parameters of the system.

The threats that must be faced by a watermarking scheme depend largely on the considered application where it is employed. For instance, there are certain metadata applications [7,8] where the only aim of watermark embedding is to give an "added value" to the asset in consideration, so they are typically not susceptible of being attacked; this characteristic is also shared by other applications, as linking contents to a web or database, or controlling electronic devices (as toys or Personal Video Recorders (PVRs)), where intentional attacks are not expected. On the other hand, applications such as watermarking of medical images, authentication of legal documents, fingerprinting or data monitoring, must face extremely hostile environments where the most harmful attacks are not necessarily those aimed at removing the embedded watermarks. In fact, for those applications somehow related to legal environments, it may be more harmful to accept a forged content as legal than rejecting a legal one, or to read the watermark instead of erasing it. These kind of considerations gave rise in the last years to the watermarking security problem. The purpose of this paper is

to facilitate future research on this topic by providing a thorough review of the existing literature, showing the most relevant achievements so far, and paying attention to the main open problems and challenges. The interested reader is also referred to the previous survey on watermarking security by Furon [9], which covers the subject from a similar viewpoint.

The paper is organized as follows: Section 2 gives an overview of the most popular forms of digital watermarking emphasizing the role of the secret key, which will be seen to be determinant in the security problem. Section 3 gives a classification of attacks on watermarking schemes based on the their treatment of the secret key, and Section 4 reviews the evolution of watermarking security in the literature, introducing relevant definitions which help to link the classification of Section 3 to the concepts of robustness and security. Section 5 is slightly more technical, as it discusses how to measure security in a theoretical, quantitative manner, introducing also the main results so far on this direction. Section 6 gives a bird's-eye view on the works that have performed practical studies of the watermarking security problem, proposing tools for performing successful security attacks. Some countermeasures that can help to improve security are considered in Section 7, discussing their advantages and drawbacks, and finally the main challenges and achievements on this topic are summarized in Section 8. Throughout this paper we will use the terms watermarking/data hiding with no distinction, unless otherwise stated, and the same will apply to the terms detector/decoder. The discussion will be often referred to image watermarking, although it can be straightforwardly extended to digital signals in general.

2 The role of the secret key

In most watermarking methods the key is used to determine certain parameters of the embedding function, such as the domain of embedding, the embedding direction, or the subset of image coefficients that will be watermarked, to list some examples. Key-dependent steganography can be traced to the ancient China, where a method to write a secret message using a paper mask with holes cut in it was developed. Such an idea was rediscovered in the Sixteenth century by Cardan, creating what is now known as Cardan grille [10]. More recently, with the advent of digital watermarking techniques, many different forms of secret communication have been proposed; the most popular of these are recalled in this section.

A number of methods that perform watermark embedding in a secret direction can be found in the literature. Van Schyndel et al. [11] suggested the use of m-sequences [12] commensurate to the image size to create a bipolar sequence that is added to the least significant bit (LSB) of the host image. This technique was later improved by Cox et al. [13] to enhance the invisibility of the watermark by modulating it with a perceptual mask that controls the embedding strength at every DCT coefficient, with the side effect of making it robust to simple LSB flipping attacks. In this method, commonly known as *additive spread-spectrum*, the watermark is a pseudorandom sequence (a.k.a. *spreading*

vector) which is modulated by the bit to be embedded (± 1) in the case of data hiding applications. Cox et al. also recognized the advantages of generating a Gaussian-distributed watermark against collusion attacks as opposed to bipolar sequences. Further benefits of Gaussian watermarks in terms of detection and decoding performance are discussed in [14]. A similar procedure due to Hartung and Girod embeds the watermark in the spatial domain for data-hiding in video applications [15].

The foregoing schemes resemble spread-spectrum communications in that a pseudorandom carrier is chosen so that any other interfering signal (in the particular case of watermarking, the host) will be nearly orthogonal to the former. Other techniques are akin to pulse position modulations in that the secret key is used to select the set of indices of those coefficients that are modified by embedding. For instance, the technique known as *patchwork* [16] pseudorandomly selects two sets of pixels: the luminance is increased for those pixels in the first set, and decreased for those in the second. Detection simply consists in subtracting the sum of pixels in the second set from that corresponding to the first set. Notice that patchwork can be seen as a form of spread-spectrum where the watermark can take the values $\{-1, 0, 1\}$, the 0 value corresponding to those pixels that are left unaltered. The method by Koch *et al.* [17] works in the 8×8-block-DCT domain and selects a subset of blocks in a pseudorandom fashion. Within each block a triple of coefficients in the mid-frequencies range is selected according to the key from a set of 18 triples and modified to encode one bit of information. Hence, Koch et al.'s method can be regarded to as a combination of pulse position modulation (i.e., the selection of the block) and multipulse modulation (i.e., the selection of the triple) of a pseudorandom sequence.

An alternative to the embedding schemes mentioned above is the use of key-dependent transforms, which was proposed by Fridrich [18] to combat sensitivity-like attacks (see Section 6.4), since in this case the attacker would not know in which domain embedding takes place. Fridrich's method constructs a set of pseudorandom-vectors which are smoothed using a low-pass filter and orthogonalized by means of the Gram-Schmidt procedure. Fridrich went on to impose energy compaction constraints on the basis functions. Related techniques are [19], which constructs key-dependent sets of orthogonal or wavelet filters; [20], which pseudorandomly controls the lifting of the Daubechies 9-7 taps used in the JPEG-2000 standard, and [21], where a set of two-channel orthogonal filter banks is pseudorandomly generated using the host image as part of the embedding key.

Yeung and Mintzer's innovative scheme uses the key to select a detection function from a secret lookup table [22]. Embedding in a given pixel proceeds by determining which luminance modification will result in the desired value at the output of the detector. In this sense, this method can be considered a precursor of side-informed algorithms.

A potential problem with pseudorandomly generated watermarks in the spatial domain is their flat spectrum which makes them prone to compression and low-pass filtering attacks. A partial solution is the use of perceptual masking but, as Su and Girod showed in their *power-spectrum condition*, in order to re-

sist filtering attacks, an energy-efficient watermark should match the spectral characteristics of the host signal [23]. Even though this can be achieved by low-pass filtering the watermark prior to its insertion, more sophisticated schemes which suggest new ways of generating the watermark from the key have been proposed. For instance, Voyatzis and Pitas use a one-dimensional chaotic map with a secret initial state [24]. The bipolar watermark is created by thresholding the chaotic sequence. By controlling the frequency of the trajectory oscillations, it is possible to impose the desired low-pass characteristics to the watermark. Finally, a two-dimensional watermark is constructed by using a Peano scan instead of a raster scan to preserve the low-pass characteristic. In a previous work by the same authors, the two-dimensional watermark was generated through a toral automorphism [25].

In the watermarking schemes mentioned so far (with exception of the Yeung-Mintzer scheme [22]) the watermark is independent of the host image, i.e. the latter is completely neglected during watermark generation, except for a possible perceptual masking which takes place a posteriori. The philosophy behind side-informed schemes is the opposite, since the host image is explicitly considered in the computation of the watermark. Nonetheless, in some side-informed methods the generation of the watermark is strongly related to that in spread-spectrum schemes; this is the case of the *Improved Spread-Spectrum* (ISS) technique by Malvar and Florêncio [26] and *Spread Transform - Dither Modulation* (ST-DM) originally proposed by Chen and Wornell [27,28]. In the former, the mapping from key to watermark is essentially the same as in [13], although it also accounts for a partial host interference removal in the secret direction of embedding. Similarly, the watermark in ST-DM is embedded in the subspace spanned by a key-dependent spreading vector, but its value is computed by quantizing the projection of the host image onto this subspace. In *quantization index modulation* (QIM) methods [27], the key is used to generate a secret codebook or set of centroids for quantizing the host image, and the watermark is basically the quantization error between the host and the secret quantizer. In the most popular implementation of QIM, known as dither modulation (DM) [27], as well as in its distortion-compensated version (DC-DM), the codebook consists of a certain lattice which is randomized by means of a key-dependent dither signal which introduces a secret shift in the embedding lattice. A more involved form of randomization of the lattice-based codebook has been recently proposed by Fei *et al* in [29], where only a subset of the lattice points (indexed by a keyed hash function) is valid for embedding. Rational Dither Modulation (RDM) methods [30] are also based on lattice quantization, but the quantization step varies as a certain function of the past watermarked samples; the codebook can be effectively randomized if this function is made key-dependent. The problem brought about by these last two randomization techniques is the difficulty of effectively controlling the embedding distortion. Yet another form of codebook randomization for DC-DM schemes which does not suffer from such drawback consists of a key-dependent rotation of the embedding lattice, as proposed by Moulin and Goteti in [31].

Other popular side-informed methods that are worth mentioning are JANIS (Just Another N-Order Side-Informed Scheme) by Furon *et al.* [32] and the trellis-based algorithm proposed by Miller *et al.* [33]. JANIS belongs to the group of methods that embed the watermark in a secret direction, but the generation of this direction is fundamentally different from the aforementioned methods; in JANIS, the direction of embedding is chosen so as to match the gradient of a key-dependent detection function evaluated in the host. As for the method by Miller *et al.*, the watermark generation is as follows: first, each message is mapped onto a set of paths on a trellis, and each transition in that trellis has associated a number of key-dependent spreading vectors. The decoding is performed by looking for the path on the trellis which maximizes the highest correlation between the received signal and all the spreading sequences related to that path.

3 Attacks on watermarking schemes

As seen in the previous section, the secret key (which hereinafter will be denoted by $\boldsymbol{\Theta}$) is an input to some mapping function $f(\boldsymbol{\Theta})$ that outputs the secret parameters (spreading vector, indices of watermarked coefficients, codebook, etc.) of the embedding and decoding functions. The aim of such parameterization is twofold: first, it is a way of protecting the contents from unauthorized embedding/decoding; second, it makes the watermarked contents more robust to attacks. The latter assertion is easy to see, for instance, in spread-spectrum-based methods: if the attacker ignores the secret subspace where the watermark lives, the best he can do is to perform his attack in a "random" direction of the space. However, if an accurate estimate of the spreading vector is available to the attacker, then he can put all the attacking power on the estimated subspace, so in that case the advantage brought about by spreading vanishes. Similar arguments hold for any method that performs watermark embedding in a secret subspace. As for the methods that rely on the secrecy of the codebook, it can be seen that an estimate of the latter would allow many harmful attacks to the robustness, including the possibility of recovering the original host image. Clearly, when evaluating attacks to watermarking systems it is important to consider the degree of knowledge about the secret key. Based on that amount of knowledge, the following classification of attacks to watermarking systems can be introduced.

1. **Blind watermark removal**. The attacker just tries to erase/modify the watermark without taking care of the secret key, even when the watermarking algorithm could be perfectly known. This is why these attacks are termed *blind*. These are the kind of attacks traditionally considered in the watermarking literature concerned with robustness assessment, and they include addition of noise, compression/filtering attacks, geometric distortions, etc. However, as suggested above, if the attacker manages to gain some knowledge about the secret key, he could devise more harmful attacks. In this sense, these *blind* attacks represent the most optimistic scenario for the watermarker.

2. **Attacks based on key estimation**.When the attacker has knowledge about the used watermarking scheme, he can try to obtain an estimate of $f(\boldsymbol{\Theta})$ through the observation of the outputs of the embedder (i.e., the watermarked images) and/or the decoder. As discussed above, this estimate can help him in succeeding in his task of defeating the system. Notice that we are talking about estimation of $f(\boldsymbol{\Theta})$ instead of $\boldsymbol{\Theta}$ itself; this is so because even when $f(\boldsymbol{\Theta})$ and the mapping function $f(\cdot)$ are perfectly known, it may not be possible to recover the secret key $\boldsymbol{\Theta}$, since $f(\cdot)$ is (or should be) designed so as not to be easily invertible. However, the knowledge of $f(\boldsymbol{\Theta})$ is enough for the attacker's purposes, in general. The computation of the estimate of $f(\boldsymbol{\Theta})$ is a central issue of the attack: similarly to cryptographic scenarios, the watermarker is usually given his own secret key, that he will use repeatedly for watermarking images; hence, all the contents watermarked by the same user will contain information about the same secret key. Typically, a reliable computation of $f(\boldsymbol{\Theta})$ will require a large number of images watermarked with the same secret key, but once an estimate has been obtained it can be used for attacking more contents of the same user without additional effort, i.e., the information learned by the attacker can be reused in subsequent attacks.
3. **Tampering attacks**. If the attacker manages to get perfect knowledge about $\boldsymbol{\Theta}$, this implies a complete break of the watermarking system because the attacker could perform the same actions as any authorized user. As mentioned above, the observation of the outputs of the embedder or the decoder only gives information about $f(\boldsymbol{\Theta})$ but not about $\boldsymbol{\Theta}$; however, the attacker can try other ways for obtaining such information. For instance, when the watermark embedder/decoder is part of an electronic device which is publicly available (such as a DVD player), the attacker can try to tamper with it in order to disclose the secret key. If the detector is thought of as a black box, the attacker would try to break this box, inspect what is inside, and determine the secret key by reverse engineering. One countermeasure at the hardware level against these kind of attacks is the use of tamper-proofing devices, as proposed in the literature [34]. On the "soft" level, other countermeasures based on protocol approaches have been proposed, such as the so-called zero-knowledge schemes [35] (see Section 7).

It is clear that the first category of attacks just introduced (blind watermark removal) is concerned with the classical concept of robustness in watermarking. In the next section we will see how the second category can be related to the concept of *watermarking security*. As for the last one, it also pertains to security, but hardware implementations are out of the scope of this paper; however, zero-knowledge and related concepts will be discussed in Section 7.

4 Review of watermarking security in the literature

In the 90's, when the digital watermarking problem arose, researchers were almost exclusively focused on the robustness of the proposed methods; simple at-

tacks such as additive noise, coarse quantization, or even the interference due to the host signal itself, were already too harmful to the first watermarking schemes, so more elaborated attacks were almost paid no attention at all. At most, there was the distinction between *intentional* and *non-intentional* attacks. An example of this type of classification can be found in [36], where *signal transformations* (affine transformations, noise addition, compression) are distinguished from *intentional attacks*, introducing at a qualitative level concepts like the *sensitivity attack*, the *statistical averaging attack* (which is closely related to the *collusion attack* [37]) and attacks based on the availability of embedding devices. The sensitivity attack belongs to the category of *oracle attacks*, which are those where the attacker exploits his access to a watermark detector. *Statistical averaging attacks* are based on the fact that, if multiple images with the same embedded watermark are available, it is possible to estimate the watermark by averaging all those images: if $\mathbf{x}_i$ denotes the i-th zero-mean host image, $\mathbf{w}$ denotes the watermark, and there are N different watermarked images, then the sum $N\mathbf{w} + \sum_i \mathbf{x}_i$ tends to $N\mathbf{w}$. A similar attack may be performed to estimate the original image when a great number of versions of the same image with different watermarks are available.

Probably the fact that raised the issue of watermarking *security* was the proposal of the sensitivity attack [38]. This attack showed that if a binary-output detector were available, the watermark embedded by means of spread-spectrum [13], which was the most popular watermarking algorithm at that time, could be removed in just $O(n)$ attempts, where n is the dimensionality of the watermarked image. This removal is based on estimating the boundary of the decision region by observing the outputs of the detector. Furthermore, the knowledge of the decision region implies the disclosure of the secret spreading vector, meaning that the attacker could also forge contents at will. Therefore, a watermarking system susceptible of being defeated by this kind of attacks could be barely thought of as being *secure*.

The first attempt at proposing a theoretical framework for analyzing the security of a watermarking scheme was performed by Mitthelholzer in [39], inspired by the works of Cachin [40] and Zöllner *et al.* [41] in the field of steganography. Mitthelholzer studies the trade-off between *secrecy* of the embedded message and robustness from a mutual information approach; in fact, a system is said to achieve perfect secrecy when the mutual information between the watermarked signal and the embedded message is null whenever the secret key is unknown. This definition of perfect secrecy clearly resembles that proposed by Shannon in his seminal work [42], where he established the information-theoretic fundamentals of cryptanalysis.

Although the security issue was becoming more relevant in the watermarking research, the first attempt at clarifying this concept is due to Kalker, who in his work [43] provided the following definitions:

- *"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 trying to *"remove, detect and estimate, write and modify the raw watermarking bits"*.

Notice that, according to these definitions, there is not a clear relationship between the intentionality of the attacks and the security; in fact, they suggest that intentionality and robustness/security can be regarded as independent concepts. Therefore, following Kalker's definitions, both intentional and non-intentional attacks may result in a threat to security.

Based on the above definitions of security and robust watermarking, a classification of watermark attacks according to different criteria is proposed in [43]. The main classification coincides with that given in [7], and it establishes the division in unauthorized watermark removal, detection (estimation), writing, and modification. Furthermore, other attack classifications are also proposed in [43] based on the degree of success of the attacks, the amount of information available to the attacker, the availability of embedding and/or detection engines, the degree of knowledge of the watermarking algorithms, and the degree of *universality* of the attack (ranging from the removal of the watermark from a certain document, to the knowledge of global secrets of the system under attack, such as the secret key).

The definitions by Kalker [43] are reviewed by Furon *et al.* in [44], where the difference between security and robustness is also emphasized; in this sense, it is said that security has a broader scope, since it does not only deal with watermark removal but also with unauthorized embedding and detection. Concerning the intentionality of the attack, [44] argues that intentionality is inherent to security attacks, whereas it is irrelevant to robustness attacks. Furon's work makes also a clear distinction between robustness and security: robustness deals with blind attacks (offering a partial break of the used watermarking technique), whereas security deals with intentional attacks where information about the data hiding scheme is known by the attacker (offering a complete break). This is clearly an evolution of the concept of security from the approach by Kalker in [43].

In [44] Furon *et al.* also translate Kerckhoffs' principle [45] from cryptography to watermarking: all functions (encoding/embedding, decoding/detection ...) should be declared public except for the secret key. The *security level* is said to be the effort required for disclosing this secret key, obtaining that definition as a corollary of Kerckhoffs' principle. Moreover, Furon *et al.* propose a classification of attacks to security based on another classical cryptography paper by Diffie and Hellman [46]. This classification is based on the amount of information relevant for the attack that is revealed to attackers; hence, one can consider the following scenarios (just to mention a few)

– the watermarked images are the only information at hand;
– the pairs original-watermarked are available (this corresponds to clear text - cypher text in cryptography);
– a watermark embedder or decoder is available to the attacker.

Although the above classification does not cover all the possible watermarking applications, it can be extended to account for other important scenarios as

needed. The goodness of a classification like this is that it allows to separate at a great extent the security analysis from the specific watermarking applications. Finally, the authors of [44] adapted Shannon's cryptographic framework to watermarking. It differs from the previous translation by Mitthelholzer [39] in that the secrecy is not measured as the information leakage between the watermarked content and the corresponding message, but as the leakage between the set of available watermarked contents and the secret key, achieving *perfect secrecy* when that information leakage is null.

A different approach to watermarking security was proposed by Barni *et al.* in [47]. The authors consider the watermarking problem as a game with some rules that determine the publicly available information. If the attacker uses only this information, the attack is said to be *fair*; if he tries to learn more information about the system, the attack is said to be *unfair*. The publicly available information can range from *no knowledge*, that clearly collides with Kerckhoffs' principle, *knowledge of embedding and detection algorithms*, *knowledge of the detection key* (for asymmetric schemes), to *knowledge of both embedding and detection keys, and the algorithms*. Similarly to [44], the mutual information is used in [48] to measure the knowledge gained by the attacker. Finally, Barni *et al.* introduce a definition of security level similar to that in [44], although in this case the authors focus on the purpose of removing the watermark, not on disclosing the secret key.

One of the most recent and outstanding works on watermarking security is [49] by Cayre *et al.*. The first point emphasized in [49] is the recognition of the difficulty of distinguishing between security and robustness. A significant evolution from [44] is that in [49] the intentionality of the attack is not enough for deciding if it targets the security or the robustness of the system. In order to define robustness, the authors complete Kalker's definition in [43], establishing that the cause of the degradation of the marked document is a classical content processing. On the other hand, to define security the authors turn again to Kalker's definition in [43], but excluding from the removal attacks *"those already encompassed in the robustness category"*. Similarly, the classification of attacks to security proposed in [44] is the basis for that introduced in [49], where three different categories are introduced, depending on the knowledge available to the attacker:

- Watermarked Only Attack (WOA): the attacker has access only to watermarked contents.
- Known Message Attack (KMA): the attacker has access to pairs of watermarked contents and the corresponding embedded message.
- Known Original Attack (KOA): the attacker designs his attack based on the knowledge of pairs original-watermarked contents.

Probably one of the main contributions of Cayre *et al.* in [49] is the proposal of the Fisher Information Matrix [50] to quantify the security. This topic will be further discussed in Section 5.

This work by Cayre *et al.* [49] is also used as inspiration for [51,52,53,54], where new definitions of security and robustness are proposed. In those works the

authors propose that attacks to robustness are those whose target is to increase the probability of error of the data-hiding channel, whereas attacks to security are those aimed at gaining knowledge about the secrets of the system; obviously, in a model where Kerckhoffs' principle holds the only secret parameters are the key $\boldsymbol{\Theta}$ and the mapping $f(\boldsymbol{\Theta})$. In this sense, attacks to security can be related to the last two categories introduced in Section 3. We will stick to this definition of security in the remaining of this paper. Finally, in [51] and [52] some considerations are made in order to further clarify the boundary between security and robustness, on the basis of the former definitions:

- Attacks to security are intentional, but not all intentional attacks are threats to security.
- Attacks to security are necessarily not blind, but there are non-blind attacks that are not aimed at attacking the security.
- The information gained by means of attacks to security can be used as a first step towards performing attacks to robustness.

5 Tools for measuring security

A consequence of the last definition of attacks to security given in the previous section is that the security of a watermarking system is directly related to the difficulty in estimating the key (or the mapping from the key to the secret parameters) based on the *observations*, where this term refers to all the information made available to the attacker (watermarked signals, embedded messages, etc.), according to the classification of security attacks given in Section 4. Thus, a natural question is how can we quantify the hardness of such estimation problem. In order to obtain fundamental security limits, one can address this question from a theoretical point of view: the first step is to check whether information about the secret key leaks from the observations; if this is the case, the second step is to quantify the amount of information that can be learned from each observation. Intuitively, a large information leakage implies that the system is potentially less secure. Shannon, in his paper on the theory of secrecy systems [42] proposed the mutual information $I(\boldsymbol{\Theta}; \mathbf{Y}_1, \ldots, \mathbf{Y}_N)$ [55] as a measure of information leakage, where $\mathbf{Y}_1, \ldots, \mathbf{Y}_N$ are ciphered texts. Although the application of this approach to the watermarking field had been already suggested by Hernández and Pérez-González in [56], it was applied for the first time by Cayre *et al.* in [49], but relying on the Fisher information [50] instead of the mutual information, arguing that the former is more suited to the watermarking problem. Shannon's approach is finally recovered for watermarking security by Comesaña *et al.* in [52], computing $I(f(\boldsymbol{\Theta}); \mathbf{O}_1, \ldots, \mathbf{O}_N)$, where $\mathbf{O}_i$ stands for the i-th observation. The reason for computing the mutual information between the observations and $f(\boldsymbol{\Theta})$ is that the observations do not provide information about $\boldsymbol{\Theta}$, but about $f(\boldsymbol{\Theta})$, as discussed in Section 3. The work presented in [52] is largely based on [49], and the main difference turns out to be the information leakage measurement, i.e. the mutual information, whose suitability for evaluating watermarking security is justified.

The information-theoretic framework for watermarking security requires a statistical modeling of all the variables involved in the problem: the host image, the secret key, and the embedded messages. Computation of the Fisher information needs the existence and differentiability of the log-likelihood function of the observations given the key, precluding its application to the analysis of some practical methods (dither modulation watermarking, for instance); fortunately, these problems do not appear when using the mutual information for quantifying the security. The adaptation of Shannon's measure to the watermarking security framework is straightforward, but one must be aware of a subtle difference: the paper by Shannon [42] deals with discrete random variables, whereas watermarking deals usually with continuous ones; this implies replacing the entropies in the computation of the mutual information by *differential* entropies [55]. Note that, as opposed to the entropy, the differential entropy can take negative values, yet it is still a useful measure of the uncertainty of a random variable. In the information-theoretic framework proposed by Shannon [42], the *equivocation* is defined as the remaining uncertainty about the secret key after observing the cyphertexts; in our case, the equivocation is redefined as the differential entropy of $f(\boldsymbol{\Theta})$ conditioned on the observations, i.e.,

$$h(f(\boldsymbol{\Theta})|\mathbf{O}_1,\ldots,\mathbf{O}_N) = h(f(\boldsymbol{\Theta})) - I(f(\boldsymbol{\Theta});\mathbf{O}_1,\ldots,\mathbf{O}_N), \tag{1}$$

where $h(f(\boldsymbol{\Theta}))$ is the *a priori* entropy of the secret parameters. The mutual information and the equivocation are the basis for the definition of some fundamental concepts:

Perfect secrecy: a watermarking system is said to achieve perfect secrecy whenever the observations do not provide any information about the secret key, that is, $I(f(\boldsymbol{\Theta});\mathbf{O}_1,\ldots,\mathbf{O}_N) = 0$.[1] This means that all efforts by the attacker for to disclose the secret key will be useless, even if he could afford infinite computational power. Clearly, the construction of watermarking systems complying with this definition may be an extremely difficult task, or lead to unpractical systems (due to complexity requirements or length of the key, for instance).

$\varepsilon - N$ **security**: a watermarking system is said to be $\varepsilon - N$ secure if

$$I(f(\boldsymbol{\Theta});\mathbf{O}_1,\ldots,\mathbf{O}_N) \leq \varepsilon, \tag{2}$$

for a positive constant ε. Anyway, one must be careful with the definition of $\varepsilon - N$ security and perfect secrecy: maybe the information leakage is small (null), but this might be due to a small (null) *a priori* entropy of the secret key; to see this, consider the extreme case where the secret key is deterministic: in this situation, the information leakage is null, but in turn the system completely lacks security, since no secret parameterization takes place. This consideration gives rise to the notion of *security level*, defined next, as a more convenient measure of security.

[1] Notice that this and the subsequent definitions based on the mutual information can be adapted to the measures based on the Fisher information.

γ-security level: for those systems with $I(f(\boldsymbol{\Theta}); \mathbf{O}_1, \ldots, \mathbf{O}_N) \neq 0$, the γ-security level is defined as the number of observations N_γ needed to make

$$h(f(\boldsymbol{\Theta})|\mathbf{O}_1, \ldots, \mathbf{O}_{N_\gamma}) \leq \gamma, \tag{3}$$

where the threshold γ (which can be negative) is established according to some criteria, as discussed below.

Unicity distance: it is defined as the number of observations N_u needed to yield a deterministic key, i.e., $h(f(\boldsymbol{\Theta})|\mathbf{O}_1, \ldots, \mathbf{O}_{N_u}) = -\infty$.[2] In the case of an a priori deterministic key, the unicity distance would be 0; however, it can approach ∞ in a general case, thus making useful the definition of the ε-security level. Furthermore, many attacks to the robustness can be performed without having perfect knowledge of $f(\boldsymbol{\Theta})$; instead, an accurate estimate may be enough for the attacker's purposes.

As mentioned above, it is not possible in general to construct perfectly secure watermarking systems; hence, the question is whether the achievable security levels are good enough for practical scenarios. The required security level will be determined by the specific application and the computational power of the attacker; in video watermarking, for instance, the large number of observations available [57] imposes severe restrictions in terms of security. As pointed out in [58], the information-theoretic models for watermarking security capture the worst case for the watermarker, i.e., they quantify the maximum amount of information about the key that is provided by each observation. An interesting question is the gap between theoretical and practical security, since the complexity of extracting all such information may be unaffordable, in general. In this sense, the mutual information $I(f(\boldsymbol{\Theta}); \mathbf{O}_1, \ldots, \mathbf{O}_N)$ must be regarded to as the *achievable rate* in a hypothetical communications problem, where the information to be transmitted is the secret key, and both the host and the embedded messages play the role of interfering channel. In other words, the information-theoretic analysis will provide a lower bound on the security level, a bound which is achievable by means of an infinite computational power, in general. Of course, the watermarker will be interested in minimizing the achievable rate about the secret key while simultaneously maximizing $I(\mathbf{Y}; M|\boldsymbol{\Theta})$, i.e., the achievable rate about the embedded message for a fair user.[3] For given embedding function and embedding distortion, this can be posed as an optimization problem where the variable to be optimized is the statistical distribution of the secret key.[4] The parameters of the embedding function affect both the security and the robustness of the scheme; their influence and the relation between security and robustness can be made patent by the representation of $I(\mathbf{Y}; M|\boldsymbol{\Theta})$ vs. $I(f(\boldsymbol{\Theta}); \mathbf{O}_1, \ldots, \mathbf{O}_N)$,

[2] The unicity distance was originally defined by Shannon in [42], but dealing with discrete random variables; hence, the N_u in the discrete case is such that $H(\boldsymbol{\Theta}|\mathbf{O}_1, \ldots, \mathbf{O}_{N_u}) = 0$.

[3] $I(\mathbf{Y}; M|\boldsymbol{\Theta})$ denotes the mutual information between the watermarked image $\mathbf{Y}$ and the embedded message M when the secret key $\boldsymbol{\Theta}$ is known.

[4] In this optimization problem, a constraint on the a priori entropy of the secret key must be imposed in order to avoid a trivial solution, such as a deterministic key.

yielding a sort of *achievable regions* similar to those in classical broadcast channels [55].

One of the main criticisms [58] to information-theoretic models for watermarking security is how can they be related to practical security levels, or equivalently, what should be the criteria for establishing the threshold ε in Eq. (3). From a practical point of view, the success of an attack based on the estimate of the secret mapping $f(\boldsymbol{\Theta})$ is closely related to the estimation error attainable by the attacker: the more reliable the estimate, the easier for the attacker to achieve his goals. Thus, it seems natural to fix the threshold ε in accordance with this estimation error. Fortunately, information-theoretic quantities and estimation errors are strongly related; for instance, the well known Cramèr-Rao lower bound [59] used in [49] relates the Fisher Information Matrix (FIM) to the minimum variance σ_E^2 achievable by an unbiased estimator:

$$\sigma_E^2 \geq \mathrm{tr}(\mathrm{FIM}(f(\boldsymbol{\Theta}))^{-1}). \tag{4}$$

Likewise, a similar bound can be arrived at by means of the equivocation [52,54]

$$\sigma_E^2 \geq \frac{1}{2\pi e} e^{\frac{2}{n} h(f(\boldsymbol{\Theta})|\mathbf{O}_1,\ldots,\mathbf{O}_N)}. \tag{5}$$

The strong relation between information-theoretic and statistical measures has been recently reinforced by some works, where exact relations between mutual information and minimum mean squared error are established for a variety of additive channels [60,61,62].

Besides the information-theoretic models, there have been very few additional attempts at establishing theoretical measures of watermarking security: in fact, the authors of this paper are only aware of the so-called *computational* security model proposed in [58], directly inspired by computational models commonly used in cryptography. This model imposes a complexity constraint to the attacker in the sense that only polynomial time computations are allowed, and the security is related to the probability of successfully inferring (after an interaction between watermarker and attacker) which secret key out of two was used for watermarking a certain object. Nevertheless, as recognized in [58], the application of the computational model to existing watermarking schemes may be very difficult, and no results in this direction have been published so far. In contrast, the information-theoretic model has been already applied for assessing the security of the two main classes of watermarking methods: spread-spectrum and quantization-based ones. The main results are summarized below.

5.1 Theoretical results on spread-spectrum methods

Using the Fisher information, the authors of [49] quantified for the first time the security of additive spread spectrum methods under the KMA, WOA and KOA attacks. As mentioned in Section 2, in spread spectrum methods the secret key is mapped to a pseudorandom spreading sequence (the watermark itself). This implies that all the images watermarked with the same key contain the same

pseudorandom pattern, a fact that can be exploited for estimation purposes. The Fisher information measures the information that the observations provide about the spreading sequence. The main conclusions of the analysis presented in [49] are:

- The difficulty of the estimation depends on the relative powers between the host signal and watermark (i.e., the Document to Watermark Ratio, DWR), in such a way that more embedding distortion implies a larger information leakage.
- Perfect estimation of the spreading sequence is only possible in the KMA case; for both the KMA and WOA cases, a sign ambiguity will remain independently of the number of observations.
- The information leakage is linear with the number of observations, i.e. all the observations provide the same amount of information about the spreading sequence.

The former conclusions are completed in [51,52], showing that the information leakage grows as a concave function of the number of observations.

Recently, a modified spread-spectrum embedding function named Natural Watermarking (NW) has been introduced by Bas and Cayre in [63]. The proposed method is shown to achieve perfect secrecy in the WOA scenario by means of information-theoretic tools, also a more intuitive explanation in terms of blind source separation (BSS) theory [64] is also given. However, the advantage of perfect secrecy of NW comes at the price of a significant degradation of the robustness with respect to the original spread-spectrum method. Another version more robust than NW is proposed in the same paper, although it does not preserve the perfect secrecy property.

5.2 Theoretical results on quantization-based methods

The most popular quantization-based methods are those with a codebook constructed by means of lattice quantizers, and they are commonly known as Distortion Compensated - Dither Modulation (DC-DM) [27,65]. In the security analysis carried out so far for this kind of methods, the only form of secret randomization of the codebook that has been considered is by means of a secret dither signal (see Section 2), which on the other hand is the case of virtually all implementations of DC-DM. As noted in [66], the security of these methods is determined by their host-rejecting nature and the boundedness of the support set of the secret dither given the observations. The main conclusions that can be extracted from the security analysis [53,54,66] based on the mutual information between the observations and the secret dither are the following:

- The security depends largely on the distortion compensation parameter α. Values of α close to 1 make the scheme extremely vulnerable to security attacks. However, in certain scenarios such as WOA, DC-DM can be made highly secure by choosing the appropriate value of α (for instance $\alpha \approx 1/2$

in binary transmission schemes with self-similar lattice partitions). This implies the existence of a trade-off between security and achievable rate in information transmission.
- When the embedding distortion is sufficiently small (as it is the case in scenarios of practical interest) the information leakage is virtually independent of the DWR, contrarily to spread-spectrum methods; nonetheless, it is still concave in the number of observations.
- The embedding lattice plays an important role in the security of the DC-DM scheme. The security level can be increased by increasing the dimensionality of the embedding lattice and choosing that lattice with the best mean-squared error quantization properties. In fact, the best security level achievable for DC-DM is conjectured to be given by those lattices whose Voronoi cells are the closest (in the normalized second order moment sense) to hyperspheres.[5]
- The security level of DC-DM schemes has been found to be fairly lower than for spread spectrum methods. Indeed, few tens of watermarked images may be enough for obtaining a sufficiently accurate estimate of the secret dither.

Some closed-form results have been obtained for the Known Message Attack and Watermarked Only Attack scenarios. However, computations with arbitrary lattices require Monte-Carlo integration most of the times.

Costa's theoretical construction [68] is closely related to DC-DM methods. A theoretical security analysis of the former was accomplished in [53], and a comparison between Costa and DC-DM was given in [53] and [66], concluding that the structure imposed to the codebook in DC-DM is responsible for its security weaknesses.

Substitutive schemes [69] can be seen as weakly related to quantization-based methods. They have been theoretically analyzed in [49], where they have been shown to provide perfect secrecy in the WOA scenario.

6 Attacks to security

This section gives an overview of the main practical methods that have been proposed so far for performing security attacks. These methods are aimed at estimating the secret parameters of the system; the usefulness of the obtained estimates is discussed in [49, Sect. IV] and [54, Sect. V] in the context of spread-spectrum and quantization-based methods, respectively. As mentioned in Section 3, the attacker can aspire at disclosing, at most, the mapping $f(\boldsymbol{\Theta})$. Nevertheless, this knowledge may be enough for getting access to the raw watermarking channel [49],[54], i.e., to read the embedded bits, but this may not be enough for reading the actual message if a cryptographic layer is placed upon the watermarking channel (the *physical* layer). One must be aware of these considerations when dealing with security attacks.

[5] The reader interested in a detailed discussion about lattices is referred to the classical text by Conway and Sloane [67].

6.1 Attacks on spread-spectrum methods

Attacks to the security of spread-spectrum methods are aimed at estimating the pseudorandom spreading vector which is derived from the secret key. In detection applications, the watermark signal just consists of this plain spreading vector, although in data hiding applications the spreading vector is modulated by the sign of the embedded message. A consequence of this correspondence between watermark and spreading vector is that most attacks previously proposed for watermark estimation are indeed attacks to security, such as the *Wiener filtering* attack [70] and the statistical averaging attack [36] (which typically needs a large number of watermarked signals to succeed) mentioned in Section 4. Related approaches using denoising techniques besides averaging are discussed in [57]. Another attempt at performing watermark estimation is due to Mıhçak *et al.* in [71], where the authors estimate the watermark based on the fact that the components of the watermark vector take discrete values ($\pm\Delta$), paying special attention to the case where these values are repeated in blocks of a certain length. Under the assumption of Gaussian host, the maximum a posteriori (MAP) estimate of the watermark is computed. The final aim of estimation attacks is to provide the information necessary to perform a *remodulation* attack [72] in order to remove the watermark.

The problem of watermark estimation in general scenarios (continuous-valued watermarks, decoding applications instead of detection) remained unaddressed for some time. A maximum likelihood (ML) watermark estimator (assuming Gaussian-distributed host signals) is proposed in [49] for the KMA case, whereas BSS techniques, namely Principal Component Analysis (PCA) and Independent Component Analyis (ICA) [64,73] are used in more involved scenarios. The rationale behind PCA and ICA-based estimation is that the energy of the watermark is concentrated in one particular subspace; moreover, the latter takes advantage of non-gaussianity of the message distribution and the independence between the embedded messages and the host images. An extension of this approach (focused on the WOA scenario) is considered in [74] using ICA jointly with the Expectation-Maximization (EM) algorithm [75] in order to reduce the computational complexity of the attack when the dimensionality of the spreading vector is very large. It is also pertinent to mention a simultaneous work, [76], in which the subspace generated by the secret key is estimated with PCA in order to remove the watermark. A previous work which used ICA to estimate the watermark signal, although without taking into account security considerations, is given in [77].

6.2 Attacks on quantization-based methods

Two major approaches can be distinguished: those dealing with lattice DC-DM schemes, and those based on Spread Transform - Dither Modulation (ST-DM) [27,28]. Contrarily to spread-spectrum, for this kind of methods the watermark depends both on the secret key and the host signal; thus, a simple watermark

estimation does not necessarily provide information about the secret key or the codebook.

For lattice DC-DM methods, the objective is to estimate the dither signal involved in the codebook parameterization. Dither estimators are addressed by Pérez-Freire *et al.* in [66] from a geometrical point of view, by means of set-theoretic (set-membership) estimators [78,79] that can be applied to generic embedding lattices. These estimators exploit the boundedness of the support set (which is, indeed, a polytope) of the secret dither given the observations. Results about the practical performance of these estimators (up to 8 dimensions) are shown in [54,66], confirming the feasibility of attacks with affordable computational complexity. Set-theoretic estimators turn out to be optimal in certain instances, but the simplifications needed to reduce the complexity of the attack (approximations with ellipsoids, for example [80]) introduce a loss of optimality which is not negligible in general; the performance gap with the optimal dither estimator (as predicted by the information-theoretic analysis) in that case is also considered in [66].

In ST-DM methods, the aim of the attacker is to disclose the secret subspace where quantization takes place. PCA and ICA have been proposed for attacking such schemes in [49] and [81]. Particularly, the good performance of ICA-based estimators was shown in [81], where a large ensemble of natural images watermarked with the same secret key are taken as input to the ICA algorithm, which outputs an estimate of the spreading vector. This estimate is used in a subsequent stage for attacking the robustness of the ST-DM scheme, and the results are compared to other attacks that do not exploit the estimate of the spreading vector at hand.

6.3 Attacks on other methods

A cryptanalytic approach for estimating the secret key was applied to steganographic methods by Fridrich *et al.* in a number of papers. In [82], JPEG steganography was addressed, considering two LSB-like steganographic methods (F5 [83] and OutGuess [84]). In these methods, the secret key determines the DCT coefficients that will be chosen for conveying the *stego message*. The estimate for the *stego key* (or better to say, the subset of coefficients used for embedding) is based on an exhaustive search, taking advantage of the characteristic statistical distribution of the DCT coefficients induced by the LSB embedding. The approach proposed in [82] was extended in [85] to steganography in the spatial domain, focusing again on LSB embedding.

It is worth noting that in the former works only one observation (or *stego image*) is exploited for performing key estimation. This is different from the approach followed in [86], where the Yeung-Mintzer authentication scheme [22] is attacked under the hypotheses given in the present paper: availability of multiple images watermarked with the same secret key (adding in this case the restriction that all the images contain the same watermark). By combining several watermarked images, the secret lookup table used in the Yeung-Mintzer scheme can be easily reconstructed, and consequently the embedded watermark can be

read off. As a countermeasure for invalidating this attack, the use of additional lookup tables that depend on the pixel position is recommended in [86].

6.4 Sensitivity attack

According to the definition given in Section 4, attacks to security are those which try to learn information about the secret key $\boldsymbol{\Theta}$ or the secret parameters $f(\boldsymbol{\Theta})$, which completely determine the decision/decoding regions. Given that there is quite a number of examples in the literature of attacks which try to perform such estimate, and that they have been shown to be effective in removing the watermark from watermarked contents, in view of the discussion in Section 4, it is reasonable to consider whether those attacks are targeted to security.

Probably the best known in this family are the so-called *oracle attacks* where the estimate of the detection region is based on the observation of the detector output, which is available to the attacker. Just a subgroup of oracle attacks are *sensitivity attacks*, originally proposed in [36,38], and further analyzed in [87] and [88]; in those attacks the boundary of the detection region is estimated by modifying a watermarked image component-wise and analyzing the sensitivity of the detector to those changes on its input. The oracle attacks in general and the sensitivity attacks in particular are formulated as iterative processes; this implies that they require a large number of calls to the detector[6] in order to provide a good approximation to the detection region.

The initial proposal of the sensitivity attack, as well as subsequent papers on this topic [36,38,87,88] were concerned with the correlation-based detector used by spread spectrum. In that case the boundary of the detection region is simply given by a hyperplane, so the attacker only needs to estimate n points on that boundary for entirely determining it. This is, for example, the strategy followed by El Choubassi and Moulin in [89]. Once one has estimated the hyperplane, removing the watermark from any watermarked content or creating false positives from any other content is straightforward. Due to such effectiveness, several countermeasures have been proposed:

- Use of non-parametric decision regions: Mansour and Tewfik [90] suggest the use of fractals for complicating the estimate of the detection function.
- Randomization of the detection boundary: Linnartz and van Dijk [87] studied the impact on the sensitivity attack of randomizing the output of the detection function in the closeness of the former decision boundary based on a hyperplane.
- Randomization of the detection output when similar signals are successively fed [91]: Venturini designed a scheme where the detector randomly generates its output whenever a signal with a similar hash has been already input; in this way the attacker can not perform the iterative process inherent to the sensitivity attack.

[6] In fact, $O(n)$ calls are needed for spread-spectrum methods [36], being n the dimensionality of the watermarked image.

– Randomization of the detection function: proposed by El Choubassi and Moulin in [92], it is based on performing detection over a random subset of image coefficients using a *mismatched* detector. The idea of performing detection in random subsets was previously proposed in [93].

Nevertheless, it seems clear that one could always try to estimate the envelope of the actual detection boundary, obtaining a coarse estimate that in most of cases is enough for attacking the watermarking system. As long as this holds true, sensitivity attacks can be considered as security attacks for spread spectrum schemes.

Despite of their impressive performance, sensitivity attacks were not adapted to more generic kinds of detectors until recently. Attempts for doing so are due to El Choubassi and Moulin [89] and Comesaña *et al.* [94]. Both algorithms are based on locally approximating the detection boundary. Nevertheless, given the local nature of the approximation, the impact of these new versions of the sensitivity attack on security is rather limited. Hence, it is perhaps more adequate to say that sensitivity attacks for a general detection function are on the boundary between attacks to security and to robustness; in any case, a formal characterization of these attacks from a security point of view is still a pending question.

7 Countermeasures

In view of the security weaknesses inherent to some watermarking methods, a number of countermeasures against security attacks have been proposed in the literature. As discussed in Section 5.1, the main flaw of spread-spectrum schemes in terms of security is the repeated embedding of the same pseudo-random pattern. Enhanced security can be achieved by adopting the scheme proposed by Doërr and Dugelay in [57] for video watermarking, which consists of randomly alternating between several watermarks in order to prevent averaging attacks; advantages and limitations of this approach are discussed in the same paper. Another possible solution has been suggested by Holliman *et al.* [95] and Fridrich and Goljan [96], who recognized the advantages in terms of security of using image-dependent keys (this is equivalent to the use of a mapping function $f(\mathbf{x}, \boldsymbol{\Theta})$, where $\mathbf{x}$ represents the host image, i.e., the host image is part of the key). In [96], for instance, the authors present a method for generating a Gaussian vector depending on both a secret key and a robust hash function of the host image. However, one major issue for watermarking schemes based on image-dependent keys is that of key synchronization at the decoder side, since the transformations suffered by the watermarked signal may avoid the exact recovery of the mapping, i.e, it may occur that $f(\mathbf{x}, \boldsymbol{\Theta}) \neq f(\mathbf{y}, \boldsymbol{\Theta})$, where $\mathbf{y}$ is the attacked image available to the decoder; thus, it seems clear that the security improvement can be made at the expense of robustness loss. As discussed in the previous sections, the success of security attacks is based on the availability of a number (large, in general) of images watermarked with the same key. The use of a mapping function like $f(\mathbf{x}, \boldsymbol{\Theta})$ is indeed aimed at reducing the number of

images which are useful for the security attack; nevertheless, due to the reasons of synchronization introduced above, the mapping function applied on perceptually similar images must yield exactly the same result, i.e., $f(\mathbf{x}_1, \boldsymbol{\Theta}) = f(\mathbf{x}_2, \boldsymbol{\Theta})$ whenever $\mathbf{x}_1$ and $\mathbf{x}_2$ are perceptually similar. This still implies the existence of a potential security hole in certain applications, such as video watermarking, where the attacker could exploit the existence of a number of perceptually similar frames (in still scenes, for instance).

In side-informed methods the watermark is already host-dependent, thus circumventing some of the security weaknesses of spread-spectrum methods. However, the existing information leakage still makes feasible attacks to security, as those mentioned in Section 6.2. One of the major disadvantages of lattice DC-DM schemes in terms of security is the use of a highly structured codebook: due to the lattice structure imposed, disclosure of one codeword implies disclosure of the whole codebook. One of the aims of the authentication scheme proposed by Fei *et al.* [29], based on lattice-quantization, is indeed the improvement of the security by making the codebook dependent on the host image. As mentioned in Section 2, the codebook in Rational Dither Modulation (RDM) watermarking [30] is also host-dependent, although this dependence is parameterized in a totally different manner.

The above countermeasures are targeted at making more difficult the estimation of $f(\boldsymbol{\Theta})$. However, there is another important group of countermeasures whose primary objective is to protect the secret key $\boldsymbol{\Theta}$. Up to now it has been shown that the most sensitive part of watermarking schemes is the embedding key; once this key is disclosed, the whole system is compromised, so the less information about this key the watermarking scheme leaks, the better for security. Nevertheless, symmetric schemes (as those we discussed up to now in this paper) require the embedding key also for detection/decoding of the inserted watermark, and this represents a security hole. There are two approaches for protecting the embedding key during the detection/decoding process, namely *asymmetric watermarking* and *zero-knowledge watermarking*.

7.1 Asymmetric Watermarking

The goal of asymmetric schemes is to make the process of detection/decoding independent of the embedding, by using different keys in these two steps. Although sometimes the terms public-key and asymmetric watermarking are used indistinctly, they have a different meaning, pointed out in most of the works in this area

- *Asymmetric watermarking*: The keys used for embedding and for extraction are different.
- *Public-key watermarking*: The key used for extraction (public key) holds enough information to accomplish the detection/decoding, while not allowing to remove the watermark or forge illegal contents if the key used for embedding (private key) is kept secret.

In [97], Smith and Dodge differentiate also between strong and weak public key watermarking; in the former, performing the extraction with the public key gives no advantage in stripping the watermark above that provided by the access to a watermark reading oracle, while a weak public key only deals with recovering the original image. They also present a very simple asymmetric method based on periodic watermarks.

Currently, there is no truly public-key watermarking method, although many efforts have been done in order to achieve an asymmetric scheme that fulfils also the requirements of public-key watermarking. In [98], Miller states that key asymmetry is not sufficient to achieve a valid public-key scheme, and he wonders whether it would even be necessary if some scheme applicable in an open-cards scenario existed. Current asymmetric schemes can be classified in two groups: linear and quadratic.

Linear schemes These are schemes based on classical spread-spectrum watermarking techniques, with linear detection functions. The first asymmetric scheme was developed in 1997 by Hartung and Girod [15], as an extension to their symmetric scheme [99]. They obtain the public key by substituting some bits of the private key by a random sequence. This scheme is not secure when disclosing the public key, as it can be used to make the watermark undetectable even when the private part can still be detected.

Two more linear schemes were presented in 2004 by Choi *et al.* [100] and Kim *et al.* [101]. The first one is based on a linear transform on the secret key to generate both the watermark and the public key; it has the same drawback as the previous scheme, but it is also susceptible to erasure of the watermark through the estimation of the linear transform when having a large set of private keys. The scheme by Kim *et al.* uses the same public watermark for a set of private watermarks, using the phase-shift-transform, that allows to have control over the correlation between the public and the secret watermark.

Quadratic schemes The first quadratic scheme was presented in 1999 by van Schyndel *et al.* [102], and it was based on invariance properties of Legendre sequences with respect to the DFT. This method was later improved by Eggers *et al.* [103,104], by using a watermark that is an eigenvector of a given linear transform matrix.

Later, Furon and Duhamel [105,106] presented an scheme that modifies the spectrum shape of an interleaved image to perform embedding of the watermark.

The three previous schemes, as well as the one presented by Smith and Dodge, can be described following the unified approach given by Furon *et al.* [107], that concludes that all these detection functions can be written using a quadratic form; furthermore, all of them have lower efficiency than symmetric schemes for the same DWR, although they are more robust against oracle attacks.

Nevertheless, [107,108] also show a statistical attack that allows to eliminate an embedded watermark relatively easily when the public key is known by the attacker.

This proves that the previous schemes are not really public-key, and their improved security when not publishing any keys comes from the higher complexity of the watermarking regions [48], leading to better security when increasing the order of the detection function [32,109]. This increase produces more complex embedders, what propitiates the use of different regions for embedding and extraction, so that detection regions are much more complicated than embedding ones; an example is the method presented by Mansour and Tewfik [110], that uses fractal theory for building a complex decision region from a simpler embedding region. Taking into account that fractal functions can give an indicator function of the detection region without revealing the boundary [48], these schemes might be seen as a way to achieve a truly public-key watermarking method, although nowadays there is no approach that can be envisaged as securely usable in a public-key scenario.

7.2 Zero-Knowledge Watermarking

Zero-knowledge watermark has arisen as a solution to conceal all the security parameters needed for detection/decoding in symmetric schemes. This way, when using a zero-knowledge watermarking protocol between two parties (Prover and Verifier), only the fact that a watermark is present or not is disclosed to the Verifier, but all the security parameters remain secret. This solves the problem posed by tampering attacks (Section 3), and provides a better protection against sensitivity attacks (Section 6.4), as only blind attacks may succeed.

The concept of zero-knowledge was introduced by Goldwasser *et al.* [111] in 1985. It basically consists in convincing an adversary of an assertion without giving him any knowledge but the assertion whose validity is proven. Zero-knowledge protocols are widely used in cryptography, generally to force a malicious adversary to behave as stated by a determined protocol.

These protocols are based on interactive proofs [111,112] and arguments [113], and especially on proofs of knowledge [114]. All of them are based on the intuitive notion that it is easier to prove a statement through an interaction between both parties (Prover and Verifier), than to write a proof that can be verified by any party without interaction. The concealment of data involved in this interaction is measured in terms of knowledge complexity [115], related to the similarity between random sequences and the sequences produced by the interaction. Zero-knowledge is the result of the indistinguishability of both types of sequences.

The first attempt of application of zero-knowledge to watermark detection was undertaken by Gopalakrishnan [116]; it consists in a protocol that allows to detect an encrypted watermark in an encrypted image, through the use of RSA [117]. Later, Craver [118] proposed several schemes of watermark detection with minimal disclosure, based on permutations using Pitas's scheme [119], or ambiguity attacks to generate a set of watermarks indistinguishable from the real one.

Adelsbach *et al.* [120] proved afterwards that all the preceding works had some flaws that made them non zero-knowledge, as they give information about the embedded watermark when using the detector as an oracle.

The formalization of zero-knowledge watermark detection was given by Adelsbach and Sadeghi [35]; they proposed the use of commitment schemes [121,122] for concealing the secret parameters of the detector; also in this work, they presented a truly zero-knowledge detection protocol for Cox's additive spread spectrum watermarking algorithm [13], as a high level protocol that uses existing zero-knowledge proofs as subblocks; it benefits from the homomorphic properties of some commitment schemes [123,124] for alleviating the communication complexity. Following the same philosophy, Piva *et al.* [125] also presented a zero-knowledge detection protocol for ST-DM.

Nevertheless, there are some security issues that must be taken into account when developing zero-knowledge watermarking protocols; they have been pointed out by Katzenbeisser in [126], and are mainly related to the correct concealing of protocol inputs and the problem of guaranteeing the correct generation of a concealed watermark. To overcome the latter issue, Adelsbach *et al.* [127] proposed several new zero-knowledge protocols that can be used to prove that a given sequence follows a determined probability distribution.

Although zero-knowledge protocols could seem an utopical solution to many security problems, they have advantages and also drawbacks [128]. Their main advantages are their null security degradation when used several times, and their resistance against clear-text attacks; their main drawback is their efficiency, as they commonly produce communication and complexity overheads that are much bigger than those presented by public-key protocols; as an example, a complete complexity study of the zero-knowledge version of Cox's non-blind detection scheme [13] is developed in [129]. Moreover, many techniques that are based on zero-knowledge lack a formal proof of zero-knowledge or even validity, due to the choices of parameters to improve efficiency; actually, many of the concepts related to zero-knowledge are asymptotic and cannot be directly applied to practical protocols.

8 Conclusions and open problems

We have made in this survey a thorough revision of the security problem in the watermarking literature. On the theoretical side, several definitions and measures have been given in order to clarify the concept of security and to establish formal models for security assessment. However, many important problems remain open, such as:

- Quantification of the gap between theoretical and practical security. As mentioned in Section 5, information-theoretic models for security represent the worst-case for the watermarker, and practical security may well be greater.
- Security assessment of a wider variety of watermarking methods, such as [30], [33], and [130]. So far, only the two major groups (spread-spectrum and quantization-based ones) have been analyzed in a few particular scenarios.
- Security assessment of oracle attacks. As said in Section 6.4, the impact of this kind of attacks on watermarking security needs to be clarified.

– The results published so far suggest that there is a trade-off between robustness and security. It is still an open question whether this trade-off is inherent to the considered problem or not.

On the practical side, we have seen that several methods for performing security attacks have been successfully tested. We have introduced also several countermeasures for improving the security level. In this regard, the main research directions appear to be the following:

– Rigorous assessment of the proposed countermeasures. As discussed in Section 7, some of them present serious drawbacks (such as the use of host-dependent keys).
– Development of zero-knowledge protocols with simplified interaction between prover and verifier. Nowadays, zero-knowledge protocols have not yet succeeded as a practical alternative due to their excessive communication complexity.
– Security assessment of watermarking schemes jointly using watermarks and cryptographic primitives, as suggested in [58]. The successful integration of cryptography and watermarking is still a pending issue.
– Development of *true* public-key watermarking schemes. As we have seen, none of the schemes proposed so far can be considered as a truly public-key scheme.

As a final comment, we would like to remark that no special attention was paid in this survey to the steganography scenario because, according to the security definition we have stuck to, steganographic security would be already encompassed by our discussion. This is true whenever secret keys are used in the secret communication process, as shown in Section 6.3, and constitutes a major difference with the survey in [9], where security for steganography has the meaning of *detectability*, according to some previous works [131,132]. Under this last point of view, a steganographic system is secure if it is impossible to distinguish between innocuous and stego images, since the secret of the system is the very existence of the embedded message. In our model, the secrecy of the message is not considered in the definition of security. One possibility for conciliating these two different approaches is to add the existence of the message to the set of secret parameters in our model (so far, only the secret key). Needles to say that this last point must be subject of further discussion.

Acknowledgments

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, and Fundación Caixa Galicia grant for postgraduate studies. 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.

References

1. Bartolini, F., Barni, M., Furon, T.: Security issues in digital watermarking. In: Proc. of 11th European Signal Processing Conference (EUSIPCO). Volume 1., Toulouse, France (2002) 282–302,441–461
2. Barni, M., Pérez-González, F.: Special session: watermarking security. 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) 685–768
3. Barni, M., Pérez-González, F.: Tutorial: Security issues in digital watermarking. In: IEEE International Conference on Image Processing (ICIP), Genova, Italy (2005)
4. Pérez-González, F., Furon, T.: Special session on watermarking security. In Barni, M., Cox, I., Kalker, T., Kim, H.J., eds.: Fourth International Workshop on Digital Watermarking. Volume 3710., Siena, Italy, Springer (2005) 201–274
5. CERTIMARK: Certification of Watermarking Techniques (2000-2002) http://www.certimark.org.
6. ECRYPT: European Network of Excellence in Cryptology (2004-2008) http://www.ecrypt.eu.org.
7. Cox, I.J., Miller, M.L., Bloom, J.A.: Digital watermarking. Multimedia Information and Systems. Morgan Kauffman (2002)
8. Barni, M., Bartolini, F.: Watermarking Systems Engineering. Signal Processing and Communications. Marcel Dekker (2004)
9. Furon, T.: A survey of watermarking security. In Barni, M., ed.: Proc. of Int. Work. on Digital Watermarking. Volume 3710 of Lecture Notes on Computer Science., Siena, Italy, Springer-Verlag (2005) 201–215
10. Petitcolas, F., Anderson, R., Kuhn, M.: Information hiding-a survey. Proceedings of the IEEE **87** (1999) 1062–1078
11. van Schyndel, R.G., Tirkel, A.Z., Osborne, C.F.: A digital watermark. In: Proc. IEEE Int. Conference on Image Processing, Austin, Texas, USA (1994) 86–89
12. Viterbi, A.: CDMA: principles of spread spectrum communication. Addison Wesley Longman Publishing Co., Inc. Redwood City, CA, USA (1995)
13. Cox, I.J., Kilian, J., Leighton, F.T., Shamoon, T.: Secure spread spectrum watermarking for multimedia. IEEE Transactions on Image Processing **6** (1997) 1673–1687
14. Hernández, J.R., Pérez-González, F., Rodríguez, J.M., Nieto, G.: Performance analysis of a 2d-multipulse amplitude modulation scheme for data hiding and watermarking of still images. IEEE J. Select. Areas Commun. **16** (1998) 510–524
15. Hartung, F., Girod, B.: Fast public-key watermarking of compressed video. In: Proc. IEEE ICIP'97. Volume I., Santa Barbara, California, USA (1997) 528–531
16. Bender, W., Gruhl, D., Morimoto, N.: Techniques for data hiding. In: Proc. of the SPIE, San Jose, CA (1995) 2420–2440
17. Koch, E., Rindfrey, J., Zhao, J.: Copyright protection for multimedia data. In: Digital Media and Electronc Publishing. Academic Press (1996) 203–213
18. Fridrich, J.: Key-dependent random image transforms and their applications in image watermarking. In: Proc. International Conference on Imaging Science, Systems, and Technology, Las Vegas, NV, USA (1999) 237–243
19. Dietl, W., Meerwald, P., Uhl, A.: Protection of wavelet-based watermarking systems using filter parametrization. Elsevier Signal Processing **83** (2003) 2095–2116

20. Seo, Y., Kim, M., Park, H., Jung, H., Chung, H., Huh, Y., Lee, J., Center, V., ETRI, T.: A secure watermarking for JPEG-2000. In: Proc. IEEE Int. Conf. Image Processing. Volume 2., Thessaloniki, Greece (2001) 530–533
21. Wang, Y., Doherty, J., Dyck, R.V.: A wavelet-based watermarking algorithm for ownership verificationof digital images. IEEE Trans. on Image Processing **11** (2002) 77–88
22. Yeung, M., Mintzer, F.: An invisible watermarking technique for image verification. In: Proc. IEEE Int. Conf. Image Processing. Volume 2. (1997) 680–683
23. Su, J., Girod, B.: Power-spectrum condition for energy-efficient watermarking. Multimedia, IEEE Transactions on **4** (2002) 551–560
24. Voyatzis, G., Pitas, I.: Chaotic watermarks for embedding in the spatial digital image domain. In: Proc. IEEE Int. Conf. Image Processing. Volume 2., Chicago, IL, USA (1998) 432–436
25. Voyatzis, G., Pitas, I.: Applications of toral automorphisms in image watermarking. In: Proc. IEEE Int. Conf. Image Processing. Volume 2., Laussane, Switzerland (1996) 237–240
26. Malvar, H.S., Florencio, D.A.F.: Improved spread spectrum: a new modulation technique for robust watermarking. IEEE Transactions on Signal Processing **51** (2003) 898–905
27. Chen, B., Wornell, G.W.: Quantization index modulation: A class of provably good methods for digital watermarking and information embedding. IEEE Trans. on Information Theory **47** (2001) 1423–1443
28. Eggers, J.J., Bäuml, R., Tzschoppe, R., Girod, B.: Scalar Costa Scheme for information embedding. IEEE Transactions on Signal Processing **51** (2003) 1003–1019 Special Issue on Signal Processing for Data Hiding in Digital Media and Secure Content Delivery.
29. Fei, C., Kundur, D., Kwong, R.H.: Analysis and design of secure watermark-based authentication systems. IEEE Transactions on Information Forensics and Security **1** (2006) 43–55
30. Pérez-González, F., Mosquera, C., Barni, M., Abrardo, A.: Rational Dither Modulation: a high-rate data-hiding method robust to gain attacks. IEEE Trans. on Signal Processing **53** (2005) 3960–3975 Third supplement on secure media.
31. Moulin, P., Goteti, A.K.: Minmax strategies for QIM watermarking subject to attacks with memory. In: IEEE International Conference on Image Processing, ICIP 2005. Volume 1., Genova, Italy (2005) 985–988
32. Furon, T., Macq, B., Hurley, N., Silvestre, G.: JANIS: Just Another N-order side-Informed watermarking Scheme. In: IEEE International Conference on Image Processing, ICIP'02. Volume 3., Rochester, NY, USA (2002) 153–156
33. Miller, M.L., Doërr, G.J., Cox, I.J.: Applying informed coding and embedding to design a robust high-capacity watermarking. IEEE Transactions on Image Processing **13** (2004) 792–807
34. Anderson, R., Kuhn, M.: Low cost attacks on tamper resistant devices. In: International Workshop on Security Protocols. Volume 1361 of Lecture Notes in Computer Science., Paris, France, Springer Verlag (1997) 125–136
35. Adelsbach, A., Sadeghi, A.R.: Zero-knowledge watermark detection and proof of ownership. In: Information Hiding, Fourth International Workshop. Volume 2137 of Lecture Notes in Computer Science., Springer (2001) 273–288
36. 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

37. Killian, J., Leighton, F.T., Matheson, L.R., Shamoon, T., Tarjan, R.E.: Resistance of watermarked documents to collusional attacks. Technical report, NEC Research Institute, Princeton, NJ (1997)
38. Cox, I.J., Linnartz, J.P.M.G.: Public watermarks and resistance to tampering. In: Proc. IEEE Int. Conf. on Image Processing. Volume 3., Santa Barbara, California, USA (1997) 3–6
39. 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
40. Cachin, C.: An information-theoretic model for steganography. 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) 306–318
41. Zöllner, J., Federrath, H., Klimant, H., Pfitzmann, A., Piotraschke, R., Westfeld, A., Wicke, G., Wolf, G.: Modeling the security of steganographic systems. In Aucsmith, D., ed.: Information Hiding International Workshop. Volume 1525 of Lecture Notes in Computer Science., Portland, OR, USA, Springer (1998) 344–354
42. Shannon, C.E.: Communication theory of secrecy systems. Bell system technical journal **28** (1949) 656–715
43. Kalker, T.: Considerations on watermarking security. In: IEEE International Workshop on Multimedia Signal Processing, Cannes, France (2001) 201–206
44. Furon, T., et al.: Security Analysis. European Project IST-1999-10987 CERTIMARK, Deliverable D.5.5 (2002)
45. Kerckhoffs, A.: La cryptographie militaire. Journal des sciences militaires **9** (1883) 5–38
46. Diffie, W., Hellman, M.: New directions in cryptography. IEEE Transactions on Information Theory **22** (1976) 644–684
47. Barni, M., Bartolini, F., Furon, T.: A general framework for robust watermarking security. Signal Processing **83** (2003) 2069–2084 Special issue on Security of Data Hiding Technologies, invited paper.
48. Barni, M., Bartolini, F., Rosa, A.D.: Advantages and drawbacks of multiplicative spread spectrum watermarking. In Delp III, E.J., Wong, P.W., eds.: Security and Watermarking of Multimedia Contents V. Proceedings of SPIE (2003) 290–299
49. Cayre, F., Fontaine, C., Furon, T.: Watermarking security: Theory and practice. IEEE Trans. Signal Processing **53** (2005) 3976–3987
50. Fisher, R.A.: On the mathematical foundations of theoretical statistics. Philosophical Transactions of the Royal Society **222** (1922) 309–368
51. Comesaña, P., Pérez-Freire, L., Pérez-González, F.: An information-theoretic framework for assessing security in practical watermarking and data hiding scenarios. In: 6th International Workshop on Image Analysis for Multimedia Interactive Services, Montreux, Switzerland (2005)
52. Comesaña, P., Pérez-Freire, L., Pérez-González, F.: Fundamentals of data hiding security and their application to spread-spectrum analysis. In: 7th Information Hiding Workshop, IH05. Lecture Notes in Computer Science, Barcelona, Spain, Springer Verlag (2005) 146–160
53. Pérez-Freire, L., Comesaña, P., Pérez-González, F.: Information-theoretic analysis of security in side-informed data hiding. In: 7th Information Hiding Workshop, IH05. Lecture Notes in Computer Science, Barcelona, Spain, Springer Verlag (2005) 131–145

54. Pérez-Freire, L., Pérez-González, F., Furon, T., Comesaña, P.: Security of lattice-based data hiding against the Known Message Attack. IEEE Transactions on Information Forensics and Security (2006) Accepted for publication.
55. Cover, T.M., Thomas, J.A.: Elements of Information Theory. Wiley series in Telecommunications (1991)
56. Hernández, J.R., Pérez-González, F.: Shedding 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
57. Doërr, G., Dugelay, J.L.: Security pitfalls of frame-by-frame approaches to video watermarking. IEEE Trans. Sig. Proc., Supplement on Secure Media **52** (2004) 2955–2964
58. Katzenbeisser, S.: Computational security models for digital watermarks. In: Workshop on Image Analysis for Multimedia Interactive Services (WIAMIS), Montreux, Switzerland (2005)
59. Cramér, H.: Mathematical methods of statistics. Landmarks on Mathematics. Princeton University Press (1999) Reprint.
60. Guo, D., Shamai, S., Verdú, S.: Mutual information and minimum mean-square error in Gaussian channels. IEEE Transactions on Information Theory **51** (2005) 1261–1282
61. Guo, D., Shamai, S., Verdú, S.: Additive non-Gaussian noise channels: Mutual information and conditional mean estimation. In: IEEE International Symposium on Information Theory (ISIT), Adelaide, Australia (2005) 719–723
62. Palomar, D.P., Verdú, S.: Gradient of mutual information in linear vector Gaussian channels. IEEE Transactions on Information Theory **52** (2006) 141–154
63. P.Bas, Cayre, F.: Natural Watermarking: a secure spread spectrum technique for WOA. In: 8th Information Hiding Workshop, IH06. Lecture Notes in Computer Science, Old Town Alexandria, Virginia, USA, Springer Verlag (2006)
64. Hyvärinen, A., Karhunen, J., Oja, E.: Independent Component Analysis. Adaptive and learning systems for signal processing, communications and control. John Wiley & Sons (2001)
65. Moulin, P., Koetter, R.: Data hiding codes. Proceedings of IEEE **93** (2005) 2083–2126
66. Pérez-Freire, L., Pérez-González, F., Comesaña, P.: Secret dither estimation in lattice-quantization data hiding: a set-membership approach. In Edward J. Delp III, Wong, P.W., eds.: Security, Steganography, and Watermarking of Multimedia Contents VIII, San Jose, California, USA, SPIE (2006)
67. Conway, J., Sloane, N.: Sphere Packings, Lattices and Groups. 3rd edn. Volume 290 of Comprehensive Studies in Mathematics. Springer (1999)
68. Costa, M.H.: Writing on dirty paper. IEEE Trans. on Information Theory **29** (1983) 439–441
69. Burgett, S., Koch, E., Zhao, J.: Copyright labeling of digitized image data. IEEE Communications Magazine **36** (1998) 94–100
70. Su, J., Girod, B.: Power-spectrum condition for energy-efficient watermarking. IEEE Transactions on Multimedia **4** (2002) 551–560
71. Mıhçak, M.K., Venkatesan, R., Kesal, M.: Cryptanalysis of discrete-sequence spread spectrum watermarks. In Petitcolas, F.A.P., ed.: 5^{th} International Workshop on Digital Watermarking, Noordwijkerhout, The Netherlands, Springer-Verlag (2002) 226–246

72. Voloshynovskiy, S., Pereira, S., Iquise, V., Pun, T.: Attack modeling: Towards a second generation benchmark. Signal Processing, Special Issue on Information Theoretic Issues in Digital Watermarking **81** (2001) 1177–1214
73. Hyvärinen, A., Oja, E.: Independent component analysis: Algorithms and applications. Neural Networks **13** (2000) 411–430
74. Cayre, F., Fontaine, C., Furon, T.: Watermarking security: application to a WSS technique for still images. In Cox, I.J., Kalker, T., Lee, H., eds.: Third International Workshop on Digital Watermarking. Volume 3304., Seoul, Korea, Springer (2004)
75. Dempster, A.P., Laird, N.M., Rubin, D.B.: Maximum likelihood from incomplete data via the em algorithm. Journal of the Royal Statistical Society, Series B (Methodological) **39** (1977) 1–38
76. Doërr, G., Dugelay, J.L.: Danger of low-dimensional watermarking subspaces. In: IEEE International Conference on Acoustics, Speech, and Signal Processing. Volume 3., Montreal, Canada (2004) 93–96
77. Du, J., Lee, C., Lee, H., Suh, Y.: Watermark attack based on blind estimation without priors. In: Proc. International Workshop on Digital Watermarking, IWDW. Volume 2613 of Lecture Notes in Computer Science., Seoul, Korea, Springer (2002)
78. Combettes, P.L.: The foundations of set theoretic estimation. Proceedings of the IEEE **81** (1993) 182–208
79. Deller, J.R.: Set membership identification in digital signal processing. IEEE ASSP Magazine **6** (1989) 4–20
80. Cheung, M.F., Yurkovich, S., Passino, K.M.: An optimal volume ellipsoid algorithm for parameter set estimation. IEEE Transactions on Automatic Control **38** (1993) 1292–1296
81. Bas, P., Hurri, J.: Security of DM quantization watermarking schemes: a practical study for digital images. In Barni, M., Cox, I., Kalker, T., Kim, H.J., eds.: Fourth International Workshop on Digital Watermarking. Volume 3710., Siena, Italy, Springer (2005) 186–200
82. Fridrich, J., Goljan, M., Soukal, D.: Searching for the stego-key. In: Proc. of Security and Watermarking of Multimedia Contents VI. Proceedings of SPIE, San Jose, CA, USA (2004) 70–82
83. Westfeld, A.: High capacity despite better steganalysis (F5-A steganographic algorithm). In: Int. Workshop on Information Hiding, IH'01. Volume 2137 of Lecture Notes in Computer Science., New York, NY, USA, Springer Verlag (2001) 289–302
84. Provos, N.: Defending against statistical steganalysis. In: 10th USENIX Security Symposium, Washington DC, USA (2001) 323–336
85. Fridrich, J., Goljan, M., Soukal, D., Holotyak, T.: Forensic steganalysis: determining the stego key in spatial domain steganography. In: Proc. of Security and Watermarking of Multimedia Contents VII. Proceedings of SPIE, San Jose, CA, USA (2005)
86. Fridrich, J., Goljan, M., Soukal, D., Memon, N.: Further attacks on Yeung-Mintzer fragile watermarking scheme. In: Proc. of Security and Watermarking of Multimedia Contents II. Proceedings of SPIE, San Jose, CA, USA (2000) 428–437
87. Linnartz, J.P.M.G., van Dijk, M.: Analysis of the sensitivity attack against electronic watermarks in images. In Aucsmith, D., ed.: 2nd International Workshop on Information Hiding, IH'98. Volume 1525 of Lecture Notes in Computer Science., Portland, OR, USA, Springer Verlag (1998) 258–272

88. Kalker, T., Linnartz, J.P., van Dijk, M.: Watermark estimation through detector analysis. In: IEEE International Conference on Image Processing, ICIP'98, Chicago, IL, USA (1998) 425–429
89. El Choubassi, M., Moulin, P.: New sensitivity analysis attack. In Edward J. Delp III, Wong, P.W., eds.: Security, Steganography and Watermarking of Multimedia contents VII, SPIE (2005) 734–745
90. Mansour, M.F., Tewfik, A.H.: LMS-based attack on watermark public detectors. In: IEEE International Conference on Image Processing, ICIP'02. Volume 3. (2002) 649–652
91. Venturini, I.: Oracle attacks and covert channels. In Barni, M., Cox, I., Kalker, T., Kim, H.J., eds.: Fourth International Workshop on Digital Watermarking. Volume 3710 of Lecture Notes in Computer Science., Siena, Italy, Springer (2005) 171–185
92. Choubassi, M.E., Moulin, P.: On the fundamental tradeoff between watermark detection performance and robustness against sensitivity analysis attacks. In III, E.J.D., Wong, P.W., eds.: Security, Steganography, and Watermarking of Multimedia Contents VIII. Volume 6072., SPIE (2006)
93. Venkatesan, R., Jakubowski, M.H.: Randomized detection for spread-spectrum watermarking: defending against sensitivity and other attacks. In: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Volume 2., Philadelphia, USA (2005) 9–12
94. Comesaña, P., Pérez-Freire, L., Pérez-González, F.: The return of the sensitivity attack. In Barni, M., Cox, I., Kalker, T., Kim, H.J., eds.: International Workshop on Digital Watermarking. Volume 3710 of Lecture Notes in Computer Science., Siena, Italy, Springer (2005) 260–274
95. Holliman, M., Memon, N., Yeung, M.: On the need for image dependent keys for watermarking. In: Proceedings of IEEE Content Security and Data Hiding in Digital Media, Newark, NJ, USA (1999)
96. Fridrich, J., Goljan, M.: Robust hash functions for digital watermarking. In: Proceedings of the International Conference on Information Technology: Coding and Computing, Las Vegas, Nevada, USA (2000) 173–178
97. Smith, J.R., Dodge, C.: Developments in steganography. In: Information Hiding. (1999) 77–87
98. Miller, M.L.: Is asymmetric watermarking necessary or sufficient? In: Proc. XI European Signal Processing Conference, EUSIPCO'02. (2002) 291–294
99. Hartung, F., Girod, B.: Watermarking of uncompressed and compressed video. Signal Processing **66** (1998) 283–301
100. Choi, H., Lee, K., Kim, T.: Transformed-key asymmetric watermarking system. IEEE Signal Processing Letters **11** (2004) 251–254
101. Kim, T.Y., Choi, H., Lee, K., Kim, T.: An asymmetric watermarking system with many embedding watermarks corresponding to one detection watermark. IEEE Signal Processing Letters **11** (2004) 375–377
102. van Schyndel, R.G., Tirkel, A.Z., Svalbe, I.D.: Key independent watermark detection. In: IEEE International Conference on Multimedia Computing Systems (ICMCS99), Florence (1999) 580–585
103. Eggers, J., Su, J., Girod, B.: Public key watermarking by eigenvectors of linear transforms. In: Proceedings of the European Signal Processing Conference, Tampere, Finland (2000)
104. Eggers, J.J., Su, J.K., Girod, B.: Asymmetric watermarking schemes. In: Sicherheit in Mediendaten, Springer Reihe: Informatik Aktuell (2000) Invited paper.

105. Furon, T., Duhamel, P.: An asymmetric public detection watermarking technique. In Pfitzmann, A., ed.: Proc. of the third Int. Workshop on Information Hiding, Dresden, Germany, Springer Verlag (1999) 88–100
106. Furon, T., Duhamel, P.: An asymmetric watermarking method. IEEE Trans. on Signal Processing **51** (2003) 981–995 Special Issue on Signal Processing for Data Hiding in Digital Media and Secure Content Delivery.
107. Furon, T., Venturini, I., Duhamel, P.: An unified approach of asymmetric watermarking schemes. In P.W. Wong, E. Delp, eds.: Security and Watermarking of Multimedia Contents III, San Jose, Cal., USA, SPIE (2001)
108. Furon, T.: Use of watermarking techniques for copy protection. PhD thesis, Ecole Nationale Supérieure des Télécommunications. (2002)
109. Hurley, N.J., Silvestre, G.C.M.: Nth-order audio watermarking. In III, E.J.D., Wong, P.W., eds.: Security and Watermarking of Multimedia Contents IV. Volume 4675 of Proc. of SPIE., San José, CA, USA (2002) 102–109
110. Mansour, M.F., Tewfik, A.H.: Secure detection of public watermarks with fractal decision boundary. In: Proc. XI European Signal Processing Conference, EUSIPCO'02, Toulouse, France (2002)
111. Goldwasser, S., Micali, S., Rackoff, C.: The knowledge complexity of interactive proof-systems. In: Proceedings of the 17th Annual ACM Symposium on the Theory of Computing. (1985) 291–304
112. Goldwasser, S., Micali, S., Rackoff, C.: The knowledge complexity of interactive proof-systems. In: SIAM Journal of Computing. Volume 18. (1989) 186–208
113. Brassard, G., Chaum, D., Crépeau, C.: Minimum disclosure proofs of knowledge. Journal of Computer and System Sciences **37** (1988) 156–189
114. Mihil Bellare, O.G.: On defining proofs of knowledge. In: Proceedings of Crypto'92. Volume 740 of Lecture Notes in Computer Science., Springer-Verlag (1992) 390–420
115. Goldreich, O., Petrank, E.: Quantifying knowledge complexity. Computational Complexity **8** (1999) 50–98
116. Gopalakrishnan, K., Memon, N.D., Vora, P.: Protocols for watermark verification. In: Multimedia and Security Workshop at ACM Multimedia. (1999) 91–94
117. Rivest, R.L., Shamir, A., Adleman, L.: A method for obtaining digital signatures and public-key cryptosystems. Communications of the ACM **21** (1978) 120–176
118. Craver, S.: Zero knowledge watermark detection. In: Information Hiding: Third International Workshop. Volume 1768 of Lecture Notes in Computer Science., Springer (2000) 101–116
119. Pitas, I.: A method for signature casting on digital images. In: Proceedings of ICIP. Volume 3. (1996) 215–218
120. Adelsbach, A., Katzenbeisser, S., Sadeghi, A.R.: Cryptography meets watermarking: Detecting watermarks with minimal- or zero-knowledge disclosure. In: XI European Signal Processing Conference. Volume I. (2003) 446–449
121. Damgård, I.: Commitment schemes and zero-knowledge protocols. In: Lectures on data security: modern cryptology in theory and practise. Volume 1561 of Lecture Notes in Computer Science., Springer-Verlag (1998) 63–86
122. Schneier, B.: Applied cryptography. Computer Networking and Distributed Systems. John Wiley & Sons (1994)
123. Fujisaki, E., Okamoto, T.: A practical and provably secure scheme for publicly verifiable secret sharing and its applications. In: Proceedings of EUROCRYPT'98. Volume 1403 of Lecture Notes in Computer Science., Springer (1998) 32–46

124. Damgård, I., Fujisaki, E.: A statistically-hiding integer commitment scheme based on groups with hidden order. In: ASIACRYPT 2002: 8th International Conference on the Theory and Application of Cryptology and Information Security. Volume 2501 of Lecture Notes in Computer Science., Spriger-Verlag (2002) 125–142
125. Piva, A., Corazzi, D., Rosa, A.D., Barni, M.: Zero knowledge st-dm watermarking. In III, E.J.D., Wong, P.W., eds.: Security, Steganography, and Watermarking of Multimedia Contents VIII, SPIE, San José, California, USA (2006)
126. Katzenbeisser, S.: On the integration of watermarks and cryptography. In: International Workshop on Digital Watermarking. (2003) 50–60
127. Adelsbach, A., Rohe, M., Sadeghi, A.R.: Overcoming the obstacles of zero-knowledge watermark detection. In: Proceedings of Multimedia and Security Workshop. (2004) 46–55
128. Menezes, A.J., van Oorschot, P.C., Vanstone, S.A.: Handbook of Applied Cryptography. CRC Press (2001) 5th reprint.
129. Adelsbach, A., Rohe, M., Sadeghi, A.R.: Non-interactive watermark detection for a correlation-based watermarking scheme. In: Communications and Multimedia Security: 9th IFIP TC-6 TC-11International Conference, CMS 2005. Volume 3677 of Lecture Notes in Computer Science., Spriger-Verlag (2005) 129–139
130. Abrardo, A., Barni, M.: Informed watermarking by means of orthogonal and quasi-orthogonal dirty paper coding. IEEE Transactions on Signal Processing **53** (2005) 824–833
131. Katzenbeisser, S., Petitcolas, F.A.P.: Defining security in steganographic systems. In Edward J. Delp III, Wong, P.W., eds.: Security, Steganography, and Watermarking of Multimedia Contents IV. Volume 4675., San Jose, California, USA, SPIE (2002) 50–56
132. Cachin, C.: An information-theoretic model for steganography. Information and Computation **192** (2004) 41–56