Open access • Journal Article • DOI:10.1109/TIFS.2007.916006

Kerckhoffs-Based Embedding Security Classes for WOA Data Hiding

François Cayre, Patrick Bas

Institutions: École nationale supérieure d'ingénieurs électriciens de Grenoble

Published on: 01 Mar 2008 - IEEE Transactions on Information Forensics and Security (IEEE)

Topics: Provable security, Digital watermarking, Cryptography and Steganography

Kerckhoffs-based embedding security classes for WOA data-hiding

François Cayre, Patrick Bas

Kerckhoffs-based embedding security classes for WOA data-hiding

François Cayre and Patrick Bas

Abstract— **It has recently been discovered that using pseudo-random sequences as carriers in spread-spectrum techniques for data-hiding is not at all a sufficient condition for ensuring data-hiding security. Using proper and realistic a priori hypothesis on the messages distribution, it is possible to accurately estimate the secret carriers by casting this estimation problem into a Blind Source Separation problem. After reviewing relevant works on spread-spectrum security for watermarking, we further develop on this topic to introduce the concept of security *classes* which broaden previous notions in watermarking security and fills the gap with steganography security as defined by Cachin. We define four security classes, namely, by order of creasing security: insecurity, key-security, subspace-security and stego-security. To illustrate these views, we present two new modulations for *truly* secure watermarking in the Watermark-Only-Attack (WOA) framework. The first one is called Natural Watermarking and can be made either stego-secure or subspace-secure. The second is called Circular Watermarking and is key-secure. We show that Circular Watermarking has robustness comparable to that of the insecure classical spread spectrum. We shall also propose information leakage measures to highlight the security level of our new spread-spectrum modulations.**

Index Terms— **Spread spectrum watermarking, security.**

EDICS Category: WAT-SSPM

I. Introduction

SINCE the first attempts of defining steganography and watermarking security, there have been obvious similarities between the two notions. We shall motivate our views with remarks that encompass both steganography and watermarking. The concept of steganography has been first defined with modern terminology in Simmons' founding work on subliminal channels and his prisoners' problem [1]. Following Simmons, Alice and Bob are in jail and they want to, possibly, devise an escape plan by exchanging hidden messages in innocent-looking cover contents. These messages are to be conveyed to one another by a common warden who eavesdrops all contents and can choose to interrupt the communication if they appear to be stego-contents. In this particular case, Eve is called a *passive* warden. This setup was commonly regarded as Simmons' original problem in which the security of the communication process is only partially affected: the transmission channel can be broken indeed, but it cannot be read or modified by the warden.

However, considering the introduction of [1], one reads: "The warden is willing to allow the prisoners to exchange messages in the hope that he can *deceive* at least one of them into accepting as a genuine communication from the other either a fraudulent message created by the warden himself or a modification by him of a genuine message." Actually, since the very beginning, Simmons stated his prisoners' problem with an *active* warden who can affect the security of the communication process. The tasks of Eve can be very different in essence, she may want to:

1) *detect* whether Alice and Bob share hidden messages, and if yes:
2) *estimate* their hidden messages,
3) *tamper* their communications.

One also has to note that both passive/active behaviours of Eve (i.e. permitting only innocuous messages to be transmitted and tampering with Alice and Bob communications – provided estimation is required prior to tampering) are of *equal* importance to her. Strangely enough, modern steganographers usually cast the steganalysis problem into a *detection* problem and restrict it to the passive behaviour of the warden. Such works like Cachin's [2] and more recently Ker's [3] follow this research line. We are however aware of a new research line in data-hiding security relying on complexity [4]. This work does not follow this line and relies on information-theoretic arguments.

It is useful to distinguish between Cachin and Ker setups from now on. In Cachin's setup, Eve is supposed to perform a test for every and each separate content being circulated between Alice and Bob: Eve performs no accumulation of the (possibly stego) contents. Extending Cachin's setup, Ker [3] introduced the concept of batch steganography and pooled steganalysis in which the hidden message is to be disseminated over a set of (possibly non) marked contents. In Ker's views, accumulation of the contents can improve Eve's knowledge.

In other recent works on steganography, an active warden was only supposed to *jam* [5] the communication channel between Alice and Bob, not to *tamper* with the message itself. Although being highly interesting, this point of view does not fit our approach of the game between the warden and the prisoners.

The framework proposed by Simmons should be related to the definitions of watermarking security given two decades after. Definitions proposed by Comesaña *et al.* [6] claim that *"attacks to security are those aimed at gaining knowledge about the secrets of the system (e.g. the embedding and/or the detection keys)."* This definition is coherent with the definition proposed earlier by Kalker [7]: *"watermark security refers to the inability by unauthorised users to have access to the raw watermarking channel"*. It implies that it is not possible to either modify the embedded information or to copy it to

This work was supported (in part) by the European Commission under grant IST-2002-507932 ECRYPT and French national funding under grants ANR-06-SETI-009 Nebbiano, RIAM Estivale and ARA TSAR.

another content if the watermarking scheme is secure: although it is always possible to modify the embedded information, a secure scheme does not allow control on how this information is modified. Performing an attack that estimates the secret key used for embedding and then copy the embedded message to another content using the estimated key is a threat on the security of a watermarking scheme. Obviously, Comesaña and Kalker definitions, along with Simmons' active warden, present useful insights to devise a common approach to data-hiding embedding security, encompassing both watermarking and steganography.

Description of attacker's knowledges and behaviours

Watermarking security was first considered from the point of view of security level assessment. In [8], the Diffie and Hellman methodology is adapted to digital watermarking and yields a classification of the attacks according to the type of information Eve has access to:

- Known-Message Attack (KMA) occurs when an attacker has access to several pairs of watermarked contents and corresponding hidden messages,
- Kown-Original Attack (KOA) occurs when an attacker has access to several pairs of watermarked contents and their corresponding original versions,
- Watermark-Only Attack (WOA) occurs when an attacker has only access to several watermarked contents.

This classification has been further extended with the Constant-Message Attack (CMA) [9] where the attacker observes several watermarked contents and only knows that the unknown hidden message is the same in all contents.

This classification also pertains to data-hiding in general. Obviously, the WOA setup is clearly related to the prisoners' problem: Eve can only be sure to observe stego (or not) contents (this is not a KOA setup since she cannot observe *pairs* of original and stego contents) and, without some social engineering she cannot know about the hidden messages (KMA is a too strong assumption for this problem).

Another issue with information security is the way the well-known Kerckhoffs' principle is applied [13]. The Kerckhoffs' principle states that Alice and Bob shall only rely on some previously shared secret for privacy. It *also* states that Alice and Bob must consider that Eve knows everything on their communication process but their secret. We found little trace of the Kerckhoffs' principle in the data-hiding literature, namely:

- in [2], the principle is said to be respected because of the very existence of a secret key,
- in [8], the authors allow Eve to know about the decoder,
- in [14], it is even assumed that Eve knowing anything on Alice and Bob communication process is a too "strong" assumption.

Summarizing the above considerations, it appears that Simmons' original prisoners' problem was cast into a detection problem that only addresses the passive behaviour of the warden. We believe it is due to the well-known relationship between the Kullback-Leibler divergence and statistical tests [2]. We also believe a more damageable issue is to somewhat neglect the prudent Kerckhoffs' principle when dealing with data-hiding security.

This work aims at looking back at the prisoners' problem (or equivalently to stay in the WOA setup) with the following assumptions in mind:

1) Eve may take a greater advantage on Alice and Bob if she allows them to communicate, even if she detects stego contents (i.e. introduce some sort of conspiracy theory in the game);
2) the warden and the prisoners all performed a detailed Kerckhoffs analysis of the way secret information is *embedded* into host contents.

The first assumption is somewhat related to Ker's pooled steganalysis [3]: Eve implicitly stores the contents circulating between Alice and Bob. This means that in our framework, we assume a twofold strategy for Eve: the first step is devoted to the analysis of the marked contents and the second step is the attack of transmitted contents. In the first step, Eve can analyze stego contents at will to try to estimate the secret key shared by Alice and Bob: her role is merely to act as a passive warden. In the second step, Eve will act as an active warden: if she knows she did accumulate enough information on the secret key, she will either try to jam the hidden channel or to tamper with it.

Like other works, we consider Alice and Bob use only one key. Of course, in real applications, especially in steganography, it is highly desirable to change the key at every communication between Alice and Bob. However, conservative worst-case security analysis (from the prisoners' point of view) can be based on such an assumption.

We shall eventually illustrate our views with spread-spectrum (SS) based data-hiding techniques, of which two are new. This paper is organized as follows: Sec. II defines the so-called embedding security classes, Sec. III develops some views on SS-based data-hiding security, Sec. IV illustrates a simple SS-based scheme designed to provide provably secure undetectability, Sec. V illustrates another SS-based data-hiding scheme designed to improve the robustness of the previous one. Finally section VI presents a theoretical and practical evalutation of the security of the presented schemes.

II. Embedding security classes

We are aware that Simmons' original problem dealt with messages. Without loss of generality, we shall however consider that Alice and Bob exchange *contents* that can either be innocent or stego contents. We depict in Fig. 1 the general setting known as the prisoners' problem.

Like Simmons, we consider Alice and Bob before-hand found a way to share a common secret, called the key. We pay no attention to the key channel anymore. Unlike Simmons, who worked with messages emanating from authentication protocols, we shall distinguish a coding/decoding stage and an embedding/extraction stage in Alice and Bob communication process (see [15] for a more detailed analysis of the hidden channel).

To us, the coding stage relates to the way the binary message $\mathbf{m}$ is transformed into one codeword $\mathbf{c}$. This transformation

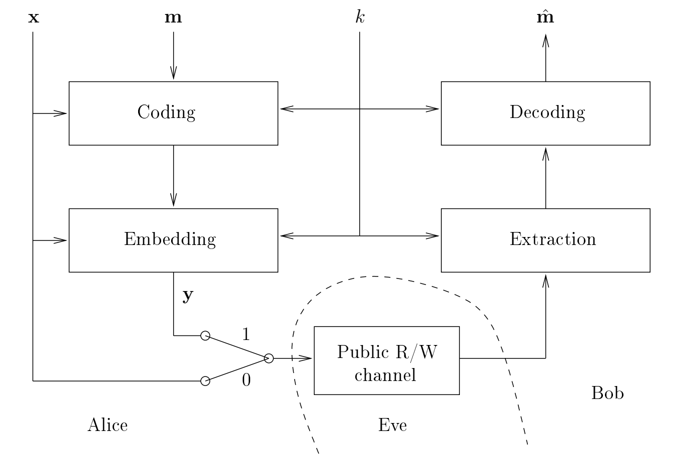

Fig. 1. The prisoners' problem. Eve has read/write access to Alice and Bob communication channel.

can be done using the host content feature vector $\mathbf{x}$ in the case of informed coding [16], [17]. The stego vector $\mathbf{y}$ is afterwards generated during the embedding stage taking into account both $\mathbf{x}$ and $\mathbf{c}$ (in watermarking this is related to using some sort of informed-embedding [19], [25] or not [18]). Implicitly, we consider contents are to be represented by their real-valued feature vector. Note that it is important to make the distinction between the embedding scheme and the coding scheme: the embedding scheme defines the way the codewords are embedded in the host signal, the coding scheme defines the way the codewords are generated according to different requirements (robustness to different categories of noises, tracing, etc.).

Following Cachin's model, Alice might want to fool Eve by (possibly randomly) sending either $\mathbf{x}$ or $\mathbf{y}$. Thus the 0/1 switch before the input of the public channel. This switch is to be discussed later on in Sec. II. Like Ker [3], we also assume several contents are to be sent to Bob through Eve. Note that we implicitly assume that Eve, since she can be an active warden, has full read/write access to Alice and Bob contents. Like Cachin, we assume Eve has some knowledge of what an innocent content should look like. Moreover, contrary to most authors, we assume Eve can run the embbeding and extraction functions at will, with any key. And we assume Alice and Bob are aware of that. Of course, Alice and Bob also can run these functions at will, with any key. This is where we restrict the application of Kerckhoffs' principle to the sole *embedding* and *extraction* functions in the prisoners' problem.

A. Notations for security

In the rest of this paper, we use the following notations:

- N_c is the size of the hidden payload (in bits),
- N_v is the size of the stego or host vector (in samples),
- N_o is the number of observed contents,
- $\mathbf{X}$ is a set of vectors representing a collection of N_o original contents, each element $\mathbf{x}$ of $\mathbf{X}$ is a N_v-dimensional random vector.
- $\mathbf{Y}$ is a set of N_v-long vectors representing a collection of N_o stego contents, each element $\mathbf{y}$ of $\mathbf{Y}$ is a N_v-dimensional random vector.

We also need to define what we consider as a secret key in this work. We use the general formalism proposed by Costa [20] where the embedding process has to consider a set of couples (codewords,messages). The set of codewords can be defined by $\mathbf{C} = \{\mathbf{c}_i \ ; \ \mathbf{c}_i \in \mathbb{R}^{N_v} \ , \ 1 \leq i \leq N_{cwd}\}$ where N_{cwd} represents the number of codewords. Each codeword $\mathbf{c}_i$ is associated with an element of the set $\mathbf{M} = \{\mathbf{m}_1, ..., \mathbf{m}_p\}$ with $p = 2^{N_c}$, by an application $A : \mathbf{C} \rightarrow \mathbf{M}$. However, because our study only consider the WOA setup, we don't have any a-priori knowledge of the embedded messages and consequently it is not possible to estimate the application A. In this context the secret key $\mathbf{K}$ is reduced (and considered to be equal) to the set of codewords $\mathbf{C}$. Note that this definition implies that two keys $\mathbf{K}_1$ and $\mathbf{K}_2$ are different (we use further the symbol $\dot{\neq}$ to denotes the difference between two keys) if $\mathrm{card}(\mathbf{K}_1 \cup \mathbf{K}_2) \neq \mathrm{card}(\mathbf{K}_1)$. Moreover we define the set $\mathcal{K}$ as the set of all possible keys, $\mathcal{K} = \{\mathbf{K}_1, ..., \mathbf{K}_{N_k}\}$, where N_k is the number of possible keys. It should be noted that practically, the secret key is generated using a seed that initializes a pseudo-random number generator with a given output repetition period (PRNG). Therefore, even if one transforms the output of a PRNG to get Gaussian signals, the set of possible Gaussian signals is related to the repetition period of the PRNG and is therefore countable.

B. Definitions of embedding security classes

Applying Kerckhoffs' principle to the embedding function allows to assume that both Alice and Eve can build a perfect estimation of different pdfs (espcially the pdf of the original contents and the pdf of the watermarked contents, see infra). The game of security is then defined taking into consideration the knowledge of:

- $p(\mathbf{X})$, the probabilistic model of N_o host contents. Since we are considering the WOA setup, Alice, Bob and Eve are able to model the joint distribution of $\mathbf{X}$: $p(\mathbf{X}) = p(\mathbf{x}_0, \ldots, \mathbf{x}_{N_o-1})$. This hypothesis stems from some sort of a worst-case consideration (from Alice and Bob point of view), where the attacker was able to model the original contents.
- $p(\mathbf{Y})$, the probabilistic model of N_o watermarked contents. Each content has been watermarked using a different key. This model can be built by the attacker using his knowledge of the embedding function.
- $p(\mathbf{Y}_\mathbf{K})$, the probabilistic model of N_o watermarked contents. Each content has been watermarked using the same unknown key $\mathbf{K}$. This is the model that the attacker can build while observing the collection of watermarked contents without any knowledge on the secret key.
- $p(\mathbf{Y}|\mathbf{K}_i)$, the probabilistic model of N_o watermarked contents. Each content has been watermarked using the same known key $\mathbf{K}_i$. This is the model that the attacker can build while applying the Kerckhoffs' principle, e.g. while embedding random messages into a collection of watermarked contents using his own key $\mathbf{K}_i$.

Since host contents are assumed to be independent, the previous models are the products of marginals, i.e.: $p(\mathbf{X}) =$

$p(\mathbf{x}_0) \times \ldots \times p(\mathbf{x}_{N_o-1})$. The same holds for $p(\mathbf{Y})$, and $p(\mathbf{Y}|\mathbf{K})$. Thus, definitions of embedding security classes in the sequel also holds for the marginals. However, we prefer to use joint probabilities in order to highlight the fact that the pirate can accumulate several contents.

Finally, Eve's ultimate goal is to estimate the constant $\mathbf{K_e}$ which maximizes the likelihood $p(\mathbf{Y_K}|\mathbf{K_e})$. Since Eve's behaviour can be very different (depending whether she acts as a passive or as an active warden), we devise accordingly four security classes for the embedding function.

Definition 1 (INSECURITY): An embedding function is **insecure** iff (if and only if) :

$$\begin{aligned} &\exists\, \mathbf{K}_1 \in \mathcal{K}\,,\; p(\mathbf{Y}|\mathbf{K}_1) = p(\mathbf{Y_K}) \\ &\text{and}\; \forall\, \mathbf{K}_2\;, \mathbf{K}_2 \dot{\neq} \mathbf{K}_1,\; p(\mathbf{Y}|\mathbf{K}_2) \neq p(\mathbf{Y_K})\,. \end{aligned} \tag{1}$$

An embedding function is then called insecure if there exists an unique key $\mathbf{K}_1$ whose associated model of watermarked contents with this key $p(\mathbf{Y}|\mathbf{K}_1)$ matches the model of the observations $p(\mathbf{Y_K})$[1]. It implies that the maximum likelihood estimation of the secret key is possible, the worst method being the exhaustive search considering the N_k different keys. However we will see in the next section that more clever techniques are possible when a embedding function is insecure.

Definition 2 (KEY-SECURITY): An embedding function is **key-secure** iff:

$$\begin{aligned} &\exists\, \mathcal{S}_\mathbf{K} \subset \mathcal{K}\,, \mathrm{card}(\mathcal{S}_\mathbf{K}) > 1\;, \\ &\forall\, \mathbf{K}_1 \in \mathcal{S}_\mathbf{K}\;,\; p(\mathbf{Y}|\mathbf{K}_1) = p(\mathbf{Y_K})\,. \end{aligned} \tag{2}$$

We can define $\mathcal{S}_\mathbf{K}$ as the invariant subset of the key $\mathbf{K}$. Note that we obviously have $\mathbf{K} \in \mathcal{S}_\mathbf{K}$. $\mathcal{S}_\mathbf{K}$ represents the set of keys which does not modify the probabilistic model of the observations. When a watermarking scheme is said insecure we can claim that $\mathcal{S}_\mathbf{K}$ does not exist. If this subset equals $\mathcal{K}$ then the algorithm is called subspace-secure (see next definition): since many authors in the literature coined the term private subspace for something close to the invariant subset, we later on will use the term *invariant subspace* instead of invariant subset.

Note that even if it is impossible to estimate the secret key $\mathbf{K}$ for key-security, it is possible to estimate the secret subspace $\mathcal{S}_\mathbf{K}$ and to reduce the uncertainty of the estimation of the secret key. The security of key-secure embedding schemes relies on the number of possible keys included in $\mathcal{S}_\mathbf{K}$ which is $\mathrm{card}(\mathcal{S}_\mathbf{K})$. As we will see further in the paper, Circular Watermarking enables to achieve key-security and the invariant-subspace associated to the key is included in an hypersphere.

Note that Doërr *et al.* defined the subspace related to a secret key for SS watermarking schemes as the set of all keys belonging to the hyperplane where the collection of watermarked signals $\mathbf{Y_K}$ share the same covariance matrix. We can call such subspace a covariant-subspace. The definition of subspace invariance proposed in this paper is more accurate because the density functions are directly considered and not only their second-order statistics. Nevertheless it is important to have the possibility to estimate either the invariant-subspace or the covariant-subspace for security purposes. If one of these subspaces is known, then it is possible to decrease the robustness of the watermarking scheme regarding AWGN attack for example and to design a random worst case attack where the attacking vector $\mathbf{v}$ belongs to the private subspace [22].

Key-security consequently means that it is impossible for the attacker to estimate the secret key $\mathbf{K}$ even if it is possible to estimate the subspace $\mathcal{S}_\mathbf{K}$.

The concept of key-security points out the existing thin frontier between data-hiding robustness and security. It deals with security because it states that the secret key cannot be disclosed and it deals with robustness because it allows random scrambling of the whole hidden information at low distortion. Consequently, we regard key-security as the minimum required class when one does not want to allow unauthorized read/write access to the secret channel.

Definition 3 (SUBSPACE-SECURITY): An embedding function is **subspace-secure** iff :

$$\forall\, \mathbf{K}_1 \in \mathcal{K}\;,\; p(\mathbf{Y}|\mathbf{K}_1) = p(\mathbf{Y_K})\,. \tag{3}$$

Subspace-security means that even in the case of an exhaustive search, Eve will not be able to distinguish between the right secret key and any wrong key. Consequently, it will be impossible for Eve to estimate the invariant-subspace $\mathcal{S}_\mathbf{K}$ associated with the secret key $\mathbf{K}$. In other words, the conditional-pdf $p(\mathbf{Y}|\mathbf{K})$ does not depend on the key $\mathbf{K}$ which is equivalent to state that $\mathbf{Y}$ and $\mathbf{K}$ are independent.

Note that subspace-security implies key-security: subspace-security allows to choose any two keys $\mathbf{K}_1$ and $\mathbf{K}_2$ for which $p(\mathbf{Y}|\mathbf{K}_1) = p(\mathbf{Y}|\mathbf{K}_2) = p(\mathbf{Y_K})$ holds and to obtain the property of key-security.

It is also important to point out that, by definition, subspace-security implies no information leakage between the watermarked contents and the key as defined by [23]. This is because subspace-security states that the right key is equivalent to any other (wrong) key: Eve can extract no knowledge from her observations. Consequently:

$$\text{Subspace-security} \Leftrightarrow I(\mathbf{Y_K}, \mathbf{K}) = 0\,. \tag{4}$$

Definition 4 (STEGO-SECURITY): An embedding function is **stego-secure** iff:

$$\forall\, \mathbf{K}_1 \in \mathcal{K}\;,\; p(\mathbf{Y}|\mathbf{K}_1) = p(\mathbf{X})\,. \tag{5}$$

Stego-security states that knowledge of $\mathbf{K}$ does not help to make the difference between $p(\mathbf{X})$ and $p(\mathbf{Y})$.

Note that stego-security implies subspace-security. However, subspace-security does not imply stego-security. One

[1] The notion of insecurity defined here is very close to the notion of array processing called *identifiability* [21] where a priori information about the sources is used to perform parameters estimation of the system.

example will be given in Sec. V of this paper. This definition implies that $p(\mathbf{Y}|\mathbf{K}_1) = p(\mathbf{Y}|\mathbf{K}_2) = \cdots = p(\mathbf{Y}|\mathbf{K}_{N_k}) = p(\mathbf{Y}) = p(\mathbf{X})$ which is equivalent to a zero Kullback-Leibler divergence (definition of "perfect secrecy" proposed by Cachin [2]):

$$\text{Stego-security} \Rightarrow D_{KL}(p(\mathbf{Y})||p(\mathbf{X})) = 0\,. \qquad (6)$$

Practically it says that it is impossible for Eve to decide whether a content has been processed through the embedding function or not (the 0/1 switch of Fig. 1).

One can finally summarize the relationships between embedding security classes with the diagram of Fig. 2.

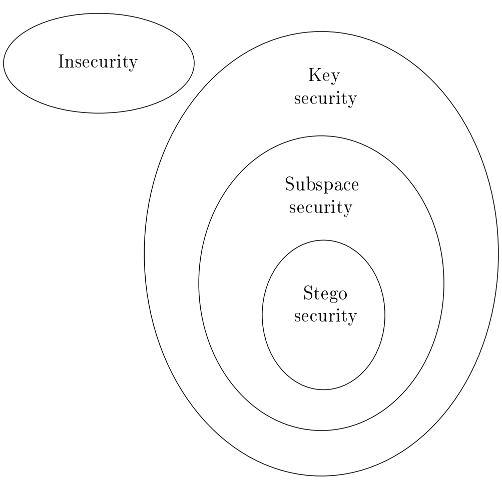

Fig. 2. Diagram for embedding security classes.

C. ϵ-Stego, ϵ-Subspace and ϵ-Key security

As proposed in [2], another way to measure similarities between density functions is to use the Kullback-Leibler divergence $D_{KL}(p(\mathbf{A})||p(\mathbf{B}))$ which equals 0 when $p(\mathbf{A}) = p(\mathbf{B})$. If $D_{KL}(p(\mathbf{A})||p(\mathbf{B})) \leq \epsilon$ it is possible to perform binary hypothesis testing to decide whether the scheme belongs to a specific class of security (either Stego-, Subspace- or Key-security). Consequently, if we call α the probability that Eve does not detect the class of the scheme and β the probability that Eve decides that the scheme belongs to a class when it is wrong, then (see Theorem 2 of [2]):

$$\alpha \log \frac{\alpha}{1-\beta} + (1-\alpha) \log \frac{1-\alpha}{\beta} \leq \epsilon\,. \qquad (7)$$

It is afterwards possible to translate our definitions of Stego-security, Subspace-security and Key-security into respectively ϵ-Stego-security, ϵ-Subspace-security and ϵ-Key-security.

Definition 5 (ϵ-KEY-SECURITY): An embedding function is **ϵ-key-secure** iff:

$$\begin{gathered} \exists\, \mathcal{S}_{\mathbf{K}} \subset \mathcal{K}\,, \mathrm{card}(\mathcal{S}_{\mathbf{K}}) > 1\,, \\ \forall\, \mathbf{K}_1 \in \mathcal{S}_{\mathbf{K}}\,,\; D_{KL}(p(\mathbf{Y}_{\mathbf{K}}) \| p(\mathbf{Y}|\mathbf{K}_1)) \leq \epsilon\,. \end{gathered} \qquad (8)$$

Definition 6 (ϵ-SUBSPACE-SECURITY): An embedding function is **ϵ-subspace-secure** iff:

$$\forall\, \mathbf{K}_1 \in \mathcal{K}\,,\; D_{KL}(p(\mathbf{Y}_{\mathbf{K}}) \| p(\mathbf{Y}|\mathbf{K}_1)) \leq \epsilon\,. \qquad (9)$$

Definition 7 (ϵ-STEGO-SECURITY): An embedding function is **ϵ-stego-secure** iff:

$$\forall\, \mathbf{K}_1 \in \mathcal{K}\,,\; D_{KL}(p(\mathbf{X}) \| p(\mathbf{Y}|\mathbf{K}_1)) \leq \epsilon\,. \qquad (10)$$

Note that because $\mathbb{E}_{\mathbf{K}_1}[D_{KL}(p(\mathbf{X}) \| p(\mathbf{Y}|\mathbf{K}_1))] \geq D_{KL}(p(\mathbf{X}) \| p(\mathbf{Y}))$, see [24] Theorem 4.3.6, and we require ϵ-stego-security be true for every $\mathbf{K}_1$, the proposed definition of ϵ-stego-security encompasses the definition of ϵ-security as proposed by Cachin for steganography.

D. Possible attacks

According to which security class the embedding function belongs, Eve has several options:

1) if the scheme is stego-secure, she cannot get any information from the transmitted contents;
2) if the scheme is subspace-secure but not stego-secure, Eve is not able to estimate $\mathcal{S}_{\mathbf{K}}$ (neither $\mathbf{K}$), but she is able to distinguish stego contents from innocent ones, e.g. she will be able to perform steganalysis. The embedding function does not respect Cachin's perfect secrecy but it is still secure for watermarking in the way defined by [23];
3) if the scheme is key-secure but not subspace-secure, Eve shall be able to estimate, given enough observations[2], the subspace $\mathcal{S}_{\mathbf{K}}$ but not the secret key $\mathbf{K}$. She will be able to concentrate the energy of her attack into the invariant-subspace of the codewords. Practically, this means that it will be possible to jam the message with a smaller distortion than in the previous case.
4) if the scheme is insecure, the estimation of $\mathbf{K}$ is possible and the security of the system is bound to be broken. She will be able to have access to the covert channel. More precisely, in a pure WOA framework, she will be able only to notice differences between hidden messages or flip the bits while minimizing the distortion (knowledge of some messages is needed to gain full read-write access to the hidden channel).

In the sequel, we shall present examples of SS-based schemes that are stego-secure, subspace-secure and key-secure. Their performances are to be assessed against those of I/SS [25], which have already been shown to be insecure [8].

III. On SS-based data-hiding security

This section first presents the principles of two popular SS watermarking schemes. We afterwards present an estimation

[2] Estimation of the number of observations required for the estimation of the subspace or the key is out of the scope of this paper.

technique that uses Independent Component Analysis and enables to estimate the secret key of a SS watermarking scheme when it is possible. Note that this attack is devoted to the class of SS schemes, and that other estimation methods can be used according to the watermarking scheme or the host statistics. For example, sub-gaussian watermark components can be estimated using Blind Source Separation techniques [10] and for other schemes that uses informed coding, clustering [11] or set-membership approaches [9] can be used.

A. SS embedding

We borrow notations from [8]. Let $\mathbf{x} \in \mathbb{R}^{N_v}$ be a host vector in which we want to hide a message $\mathbf{m} \in \{0,1\}^{N_c}$. Let r be the rate of the data-hiding channel:

$$r = \frac{N_c}{N_v}.$$

Further, we need a secret key $\mathbf{K} \in \mathbb{N}$ used to initialize a PRNG (Pseudo-Random Number Generator) in order to get N_c secret carriers $\{\mathbf{u}_i\}$. Using Gram-Schmidt orthogonalization, we can ensure that:

$$\forall\ i \neq j\ \ < \mathbf{u}_i | \mathbf{u}_j >= 0.$$

Further, we normalize each $\mathbf{u}_i$ such that:

$$\forall\ i\ \|\mathbf{u}_i\|^2 = N_v.$$

This means that for N_v large enough, we can assume that $\forall\ i\ \sigma^2_{\mathbf{u}_i} \simeq \sigma^2_{\mathbf{u}} = 1$ since a Gaussian PRNG is expected to produce a zero-mean output ($\mathbb{E}[u] = 0$).

¿From Eve's point of view, $\mathbf{K}$ and $\{\mathbf{u}_i\}$ are equivalent representation of the secret. We implicitly assume that Eve will focus on the $\{\mathbf{u}_i\}$ rather than on $\mathbf{K}$. Note that due to the constraints on the norm of the carriers, the different keys are located on the N_c-hypersphere of radius N_v that belongs to Span $(\{\mathbf{u}_i\})$.

Using a modulation $s : \{0,1\} \to \mathbb{R}$, we are able to construct the watermark signal $\mathbf{w}$:

$$\mathbf{w} = \sum_{i=0}^{N_c-1} \mathbf{u}_i s(\mathbf{m}(i)). \tag{11}$$

Classical SS uses a modulation called the BPSK modulation:

$$s_{BPSK}(\mathbf{m}(i)) = \gamma(-1)^{\mathbf{m}(i)}, \tag{12}$$

where γ allows to achieve a given distorsion. A more efficient modulation (from the robustness point of view) is the linear approximation of ISS (Improved Spread Spectrum [25]):

$$s_{ISS}(\mathbf{m}(i)) = \alpha(-1)^{\mathbf{m}(i)} - \lambda \frac{< \mathbf{x}|\mathbf{u}_i >}{\|\mathbf{u}_i\|^2}, \tag{13}$$

where α and λ are computed to achieve an average distorsion and to minimize the error probability. One also generally wants to achieve a desired Watermark-to-Content power Ratio (WCR) in decibels, or possibly an expectation of it:

$$WCR = 10 \log \left(\frac{\sigma^2_{\mathbf{w}}}{\sigma^2_{\mathbf{x}}} \right). \tag{14}$$

Without loss of generality, we can assume that $\sigma^2_{\mathbf{u}} = \sigma^2_{\mathbf{u}_i} = \sigma^2_{\mathbf{x}} = 1$ since we can consider that the power ratio of the carriers and the host signal is taken into account in the modulation process. Additionally, we also assume additive embedding to construct the watermarked signal $\mathbf{y}$:

$$\mathbf{y} = \mathbf{x} + \mathbf{w}. \tag{15}$$

Using the correlation (normalized by N_v) $z_{\mathbf{v},\mathbf{u}}$ between a vector $\mathbf{v}$ and a carrier $\mathbf{u}$:

$$z_{\mathbf{v},\mathbf{u}} = \frac{1}{N_v} < \mathbf{v}|\mathbf{u} >= \frac{1}{N_v} \sum_{j=0}^{N_v-1} \mathbf{v}(j)\mathbf{u}(j), \tag{16}$$

one can deduce the simple decoding rule to output $\hat{\mathbf{m}}$ the estimated message from $\mathbf{y}'$ a potentially attacked version of $\mathbf{y}$:

$$\begin{array}{ll} \hat{\mathbf{m}}(i) = 1 & \text{if} \quad z_{\mathbf{y}',\mathbf{u}_i} < 0, \\ \hat{\mathbf{m}}(i) = 0 & \text{if} \quad z_{\mathbf{y}',\mathbf{u}_i} > 0. \end{array} \tag{17}$$

A common way in the watermarking community to assess robustness is to add an AWGN to the watermarked vector. Later on, we shall therefore add a noise $\mathbf{n} \sim \mathcal{N}(0, \sigma^2_{\mathbf{n}})$ for BER simulations. The power of the attack will be expressed in terms of Watermarked-Content-to-Noise power Ratio (WCNR):

$$WCNR = 10 \log \left(\frac{\sigma^2_{\mathbf{y}}}{\sigma^2_{\mathbf{n}}} \right). \tag{18}$$

Moreover, let N_o be the number of observations Eve has access to. In the sequel, we generally use matrices as column-wise collections of several realizations of a template vector. For example, $\mathbf{Y} \in \mathcal{M}_{N_v \times N_o}(\mathbb{R})$ is the matrix of watermarked contents Eve has collected, and $\mathbf{S} \in \mathcal{M}_{N_c \times N_o}(\mathbb{R})$ is the matrix of modulated messages. N_o data-hiding operations are described as:

$$\mathbf{Y} = \mathbf{X} + \mathbf{W} = \mathbf{X} + \mathbf{US}. \tag{19}$$

In the sequel, we assume *all host signals* are *i.i.d* (independent and identically-distributed) Gaussian noises with $\mathcal{N}(0,1)$ pdf.

A hypothesis of utmost importance is that the messages are supposed to be independently drawn according to the Bernouilli $\mathcal{B}(\frac{1}{2})$ distribution. Also, we will denote as $p(\mathbf{x})$, $p(\mathbf{y})$, $p(\mathbf{y}|\mathbf{U})$ the pdf of the host vector, the pdf of the watermarked vector and the pdf of the watermarked vector given the knowledge of the carriers, respectively.

B. SS-based embedding security and ICA

To study the (in)security of any SS Watermarking scheme, we have to estimate $p(\mathbf{Y}|\mathbf{U}) = p(\mathbf{X} + \mathbf{US}) = p(\mathbf{X}) * p(\mathbf{US})$. Since $\mathbf{X}$ is modeled by an *i.i.d* Gaussian process and since $\mathbf{U}$ is constant, the security of this scheme relies on the possibility to estimate the density of the modulation $p(\mathbf{S})$.

It is important to point out that Eq. 19 states that BPSK-based spread-spectrum watermarking can be seen as a noisy mixture of carriers. Noise is the original content and the mixture is parameterized by the modulation of the message. In this setup the problem of carriers estimation is just what is commonly known as blind source separation (BSS). Given proper a priori knowledge, one typically wants to recover $\mathbf{S}$ (the *sources* in BSS theory) and possibly $\mathbf{U}$ (the *mixing matrix* in BSS theory). It is very insightful to notice that on one hand one advantage of BSS theory is that it makes no assumption on $\mathbf{U}$ the mixing matrix, but only on $\mathbf{S}$, the sources. On the other hand, other methods may use the fact that the columns of $\mathbf{U}$ are orthogonal to perform its estimation [12].

Given our fundamental hypothesis that the messages are drawn independently and that the carriers are scaled orthonormal, the projection of each carrier gives independent components:

$$p(<\mathbf{y}|\mathbf{u}_1>, ..., <\mathbf{y}|\mathbf{u}_{N_c}>) = \prod_{i=1}^{N_c} p(<\mathbf{y}|\mathbf{u}_i>),$$

and our attacker shall therefore rely on ICA (Independent Component Analysis) to achieve his goal.

To assess the insecurity of a SS-based technique, we have decided to adopt the following methodology which is generally used in BSS benchmarks:

1) We generate N_o observations of watermarked contents and generate the matrix of observations $\mathbf{Y}$.
2) We whiten the observed signals using Principal Component Analysis. A reduction of dimension is therefore performed to reduce the searching time. If we consider that each host signal is generated from an *i.i.d.* process, the subspace containing the secret key will be included into a N_c-dimensional space of different variance [26]. We consequently select the subspace generated by eigenvectors corresponding to the N_c highest eigenvalues.
3) We run the FastICA algorithm [27] on this subspace to estimate the independent components and the independent basis vectors (*e.g.* the secret carriers).
4) We compute the normalized correlation c between each original and estimated carriers. A value of c close to 1 means that the estimation of the component is accurate. An estimation close to 0 means that the estimation is erroneous. For $N_c = 2$, we may evaluate the estimation accuracy by plotting a 2D constellation of points of coordinates (c_1, c_2). A successful estimation will then provide a point close to one of the four cardinal points $(0, 1)$, $(0, -1)$, $(1, 0)$, $(-1, 0)$[3].

We have applied this ICA-based carrier estimation for both SS and ISS embedding. Fig. 3 depicts the normalized correlation between the original and estimated carriers for 100 experiments considering each time 1000 watermarked vectors. We can notice that the estimations are globally more accurate for SS than for ISS with the BSS technique we used. In this case, this is mainly due to the fact that the variance of the watermarked signal after ISS is smaller than that after SS and consequently reduces the accuracy of the subspace estimation. Both SS and ISS were experimented at the same level of distorsion: the Watermark-to-Content Ratio (WCR) was set to $-21dB$.

Please note that ICA algorithms also have the two fundamental limitations (due to seeking independent components) which encompass those defined in the previous section:

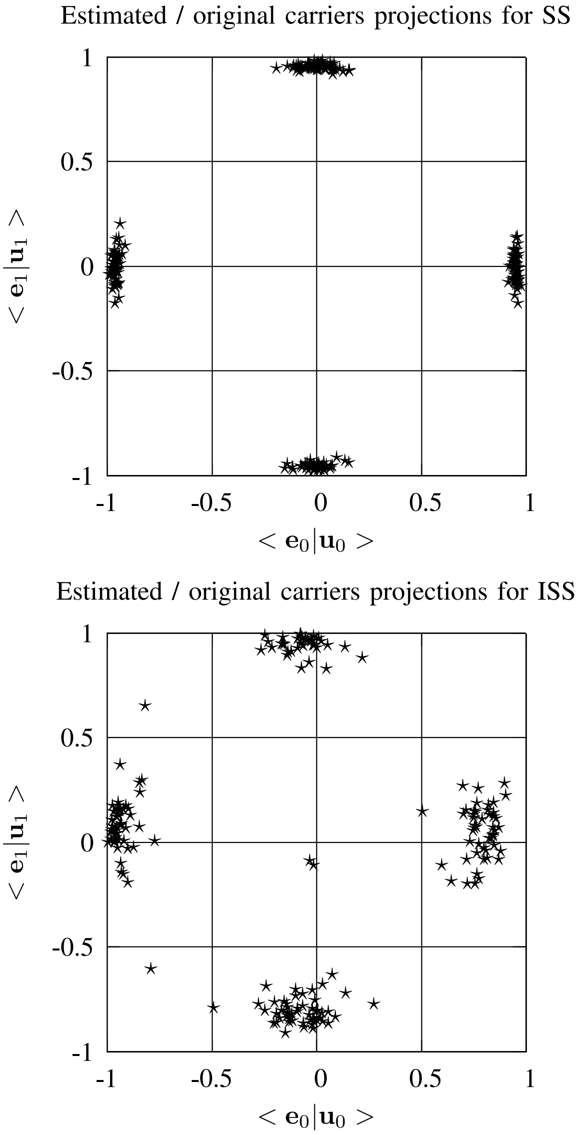

Fig. 3. Normalised correlations between the two estimated carriers and $\mathbf{u}_i \sim \mathcal{N}(0, 1)$ the real ones. For both schemes, $N_o = 1000$, $WCR = -21dB$ and $N_v = 512$.

- it cannot recover the correct ordering of the mixing matrix columns;
- it outputs vectors that are only colinear to the mixing matrix columns.

This natural disinclination means that in the WOA set-up, the set of carriers and their opposites will be considered as representing the very same key. In this context, Key-security will be achieved only if it is impossible to estimate a secret carrier even up to a sign as shown in the following sections.

In Fig. 4, we depict the plot of $z_{\mathbf{y},\mathbf{u}_i}$ for traditional SS and ISS [25]: when the sources are not Gaussian nor dependent, one can observe clusters oriented according to the positions of the secret carriers. Dependency and Gaussianity are two solutions to avoid such clusters to arise. Theses ideas are explained in the next section.

However it is important to point out that any ICA algorithm

[3] We use $N_c = 2$ for illustration purposes, dealing with more bits would require to use the Hungarian method [28] to assign original and estimated carriers prior to the computation of the normalized correlation c.

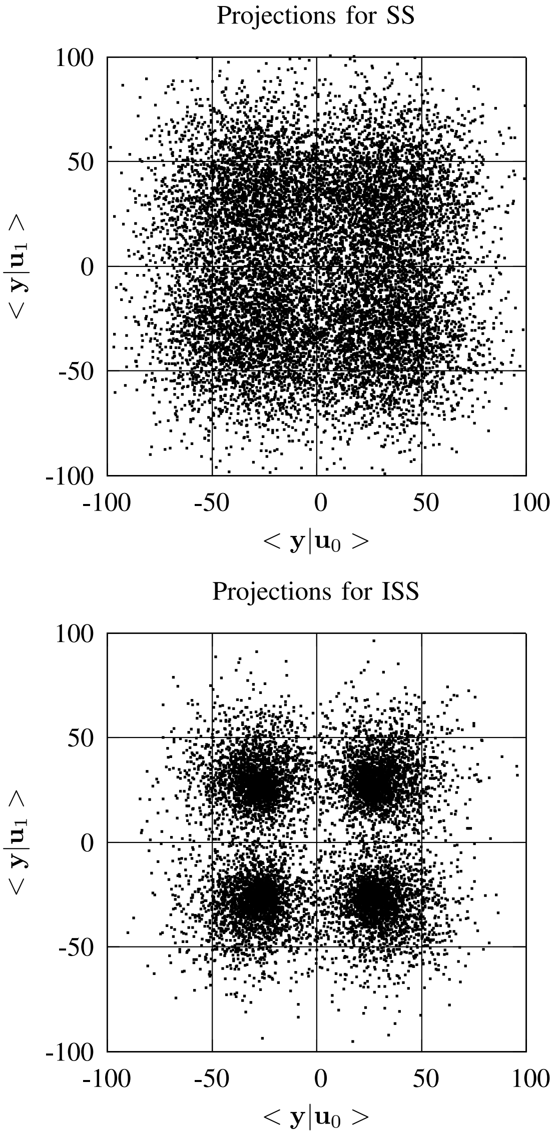

Fig. 4. Joint-distribution of two carriers for SS (top) and ISS (bottom). The formation of clusters allows to estimate the secret carriers by means of ICA. $N_c = 2$, $N_v = 512$, $WCR = -20dB$.

is known to fail (i.e. it outputs random sources and mixing matrix) in the two following cases:

- the sources are not independent,
- the sources are *i.i.d.* Gaussian signals of same variance.

Therefore, a successful approach to forbid accurate estimation of the carriers is to artificially make the sources become Gaussian and *i.i.d.* (Sec. IV: Natural Watermarking) or dependent (Sec. V: Circular Watermarking).

IV. Natural Watermarking

The goal of this section is to devise a secure SS-based watermarking scheme for the WOA framework. Natural Watermarking (NW) was named after its ability to (possibly) preserve the original pdf of the distribution of $z_{\mathbf{x},\mathbf{u}_i}$ during embedding, i.e. $z_{\mathbf{x},\mathbf{u}_i} \sim z_{\mathbf{y},\mathbf{u}_i}$. First, provided $\mathbf{x}$ has symmetrical pdf, one can easily show by Central Limit Theorem (CLT) argument that for N_v large enough:

$$z_{\mathbf{x},\mathbf{u}_i} \sim \mathcal{N}\left(0, \frac{\sigma_{\mathbf{x}}^2 \sigma_{\mathbf{u}_i}^2}{N_v}\right). \tag{20}$$

A. NW as SI model-based watermarking

NW modulation uses side-information (SI) at the encoder to increase security, whereas it has been common during the last years to use it for increasing robustness [19]. Natural watermarking can be seen as the spread-spectrum version of model-based steganography [29].

NW modulation is defined as:

$$s_{NW}(\mathbf{m}(i)) = -\left(1 + \eta(-1)^{\mathbf{m}(i)} \frac{< \mathbf{x}|\mathbf{u}_i >}{| < \mathbf{x}|\mathbf{u}_i > |}\right) \frac{< \mathbf{x}|\mathbf{u}_i >}{\|\mathbf{u}_i\|^2}. \tag{21}$$

This modulation is more easily viewed as a model-based projection on the different vectors $\mathbf{u}_i$ followed by a η-scaling along the direction of $\mathbf{u}_i$. NW basically checks whether $z_{\mathbf{x},\mathbf{u}_i}$ lies on the desired side of the Gaussian curve, see Fig. 5. If not, it simply performs a model-based symmetry before applying a scaling. Also note that the condition for correct decoding is obviously:

$$\eta \geq 1. \tag{22}$$

¿From the security point of view, the original Bernouilli modulations are modified according to the values of the projection $z_{\mathbf{x},\mathbf{u}_i}$ which have Gaussian distribution. Again, by CLT argument, we have for N_v large enough:

$$s_{NW} \sim \mathcal{N}\left(0, \sigma^2_{s_{NW}}\right). \tag{23}$$

The fact that the sources in NW follow a Gaussian distribution ensures that NW is at least key-secure under our assumptions (WOA framework and independent messages), see Sec. VI-B. This clearly relates to the inability of ICA to separate sources in this case.

Since we assume $\mathbf{x} \sim \mathcal{N}(0, \mathbf{I}_{N_v})$ ($\mathbf{I}_{N_v}$ is the identity matrix of size $N_v \times N_v$), one has obviously:

$$\mathbf{y}|\mathbf{K} \sim \mathcal{N}(0, \mathbf{J}_{N_v}),$$

with $\mathbf{J}_{N_v} = \mathbf{I}_{N_v}$ only if $\eta = 1$. This means that NW is stego-secure for $\eta = 1$. Otherwise it is just key-secure. Indeed, having $\eta > 1$ implies that subspace-security cannot be met. A similar result was already found by other means [30], [31]: the authors forced a Gaussian stego-distribution to keep the same as the Gaussian cover-distribution by means of an adequate scaling.

B. NW features

¿From Appendix I, we have the following theoretical expectation of WCR for NW (which is actually a lower bound):

$$WCR_{NW} = 10 \log_{10}\left(\frac{(1+\eta^2)N_c}{N_v}\right). \tag{24}$$

We confirm on Fig. 6 that there is no difference between this last approximation and the practical measurements. This last plot was performed targeting stego-security ($\eta = 1$) but the

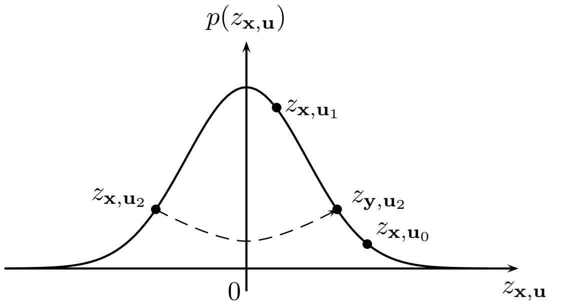
$p(z_{\mathbf{x},\mathbf{u}})$
$z_{\mathbf{x},\mathbf{u}_1}$
$z_{\mathbf{x},\mathbf{u}_2}$
$z_{\mathbf{y},\mathbf{u}_2}$
$z_{\mathbf{x},\mathbf{u}_0}$
0
$z_{\mathbf{x},\mathbf{u}}$

Fig. 5. Natural watermarking for $\mathbf{m} = \{1, 1, 1\}$ ($N_c = 3$). Only the third bit calls for a model-based projection followed by a scaling.

WCR expectation was found to be equally adequate when $\eta > 1$. Note that considering $\eta < 1$ is meaningless, see Eq. 22.

Additionally, the expression of the BER for the AWGN channel is the following (see Appendix II):

$$P_e = \int_0^{+\infty} \mathcal{G}_{\sigma_W^2+\sigma_N^2}(t) \text{ erfc } \left(\frac{\sigma_W t}{\sqrt{2}\sigma_N \sqrt{\sigma_W^2 + \sigma_N^2}} \right) dt, \tag{25}$$

with $\mathcal{G}_{\sigma_W^2+\sigma_N^2}(t) = \frac{1}{\sqrt{2\pi(\sigma_W^2+\sigma_N^2)}} \exp\left(\frac{-t^2}{2(\sigma_W^2+\sigma_N^2)}\right)$.

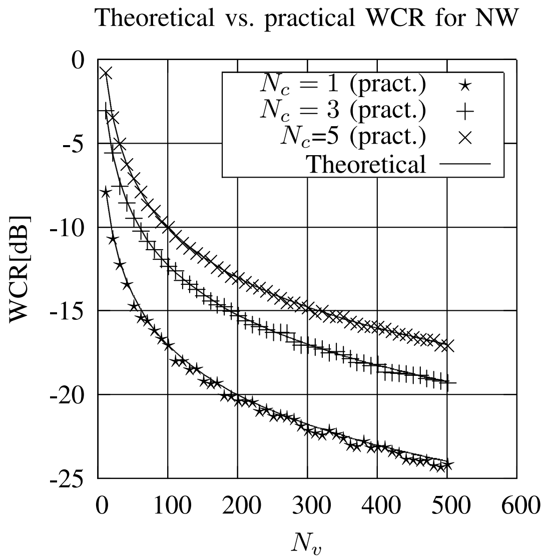

Fig. 6. Comparison between theoretical and practical WCR ($\eta = 1$).

If one wants to specify a target average WCR, the parameter η has the following expression:

$$\eta = \sqrt{\frac{N_v}{N_c} \times 10^{\frac{WCR}{10}} - 1}. \tag{26}$$

Therefore, the maximum number of bits to be securely hidden (i.e. $\eta = 1$) in $\mathbf{x}$ is:

$$N_c^{max} = \frac{N_v}{2} \times 10^{\frac{WCR}{10}}. \tag{27}$$

Fig. 9 shows the performance of NW ($\eta = 1$) compared to other SS-based schemes. It is not surprising that SS always outperforms NW since it has a security constraint to meet that SS does not have to. Another remark is that NW does not achieve stego-security when the cover distribution is not Gaussian: in this case it only achieves subspace-security if $N_c = N_v$ or key-security if $N_c \neq N_v$. This means that for practical applications, NW can only be used to embed some hidden information into noisy components. This conclusion somewhat complies with Fridrich's advice to use noisy images for steganography [32]. All in all, NW is more a theoretical scheme than a practical one.

C. NW for stego-security: $\eta = 1$

When $\eta = 1$, NW simply amounts to the implementation of the well-known Householder reflection. We depict on Fig. 7 the distribution of the projection of two secret carriers on watermarked contents. We can see that neither cluster nor principal directions arise with NW, all other parameters being equal to SS and ISS embedding depicted in Fig. 4. Note also that the theoretical and practical evaluations of the security of Natural Watermarking will be evaluated in Sec. V-D and Sec. VI. It is not possible to estimate the secret keys because the Gaussian joint distribution of the projection of the carriers in the watermarked contents is circular (see below) and consequently any estimation of independent components (the carriers) is hopeless [33]. More importantly, circularity implies the definition of key-security since all the carriers that belong to the hypersphere provide the same density functions. Fig. 7 presents the Gaussian joint distribution of two carriers. As we can see, it is not possible to find the directions that are associated to each carrier.

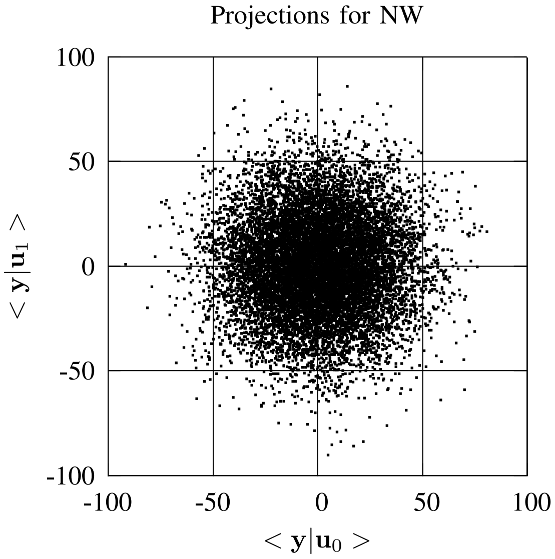

Fig. 7. Joint-distribution of two carriers for NW. NW does not produce any cluster, thus leading to key-security. In this particular setting when $\eta = 1$, NW also achieves stego-security.

V. Circular embedding and watermarking

Looking back at Fig. 4 and Fig. 7, it is clear that NW robustness can get much better by improving the separation of the decoding regions. It is the goal of the coding technique that we describe in this section under the term of Circular Watermarking (CW).

A. CW definition

One can easily check that the joint-distribution of the projections of the secret carriers on the host signal using NW is circular.

Let $p(z_{\mathbf{x},\mathbf{u}_0}, \ldots, z_{\mathbf{x},\mathbf{u}_{N_c-1}})$ be the joint-distribution of the projection of the secret carriers on the host signal. Formally, we call *circular* any watermarking scheme which exhibits the following property:

$$p(z_{\mathbf{x},\mathbf{u}_0}, \ldots, z_{\mathbf{x},\mathbf{u}_{N_c-1}}) = p\left(\rho\right), \tag{28}$$

where

$$\rho = \sqrt{\sum_{i=0}^{N_c-1} z^2_{\mathbf{x},\mathbf{u}_i}}.$$

Incidentally, note that NW is clearly circular because in that case we have $p(\rho) = \frac{1}{\sqrt{2\pi\sigma^2}^{N_c}} \exp\left(\frac{-\rho^2}{2\sigma^2}\right)$.

B. A practical implementation of CW based on ISS

While Eq. 28 leaves many degrees of freedom for devising a circular watermarking scheme, we present here a practical implementation based on the well-known ISS modulation [25]. We could also have based our implementation on classical SS, but certainly at the cost of a lower robustness. For the sake of simplicity, we shall refer to this very implementation as CW in the sequel. The basic idea is to randomly spread the clusters of ISS (which are depicted on Fig. 4) on the whole decoding regions while preserving the circularity.

To this aim, let us construct a normalized [34] vector $\mathbf{d} \in \mathbb{R}^{N_c}$ from another random vector $\mathbf{g} \sim \mathcal{N}(0, \mathbf{I}_{\mathbf{N_c}})$. Each $\mathbf{d}(i)$ coefficient is constructed as follows:

$$\mathbf{d}(i) = \frac{|\mathbf{g}(i)|}{||\mathbf{g}||}. \tag{29}$$

This vector will be independently drawn at each embedding and uniformly distributed on the positive orthant of the hypersphere. Our CW implementation requires exactly the same computations for ISS-parameters α and λ [25], which we intentionally omit here:

$$s_{CW}(\mathbf{m}(i)) = \alpha(-1)^{\mathbf{m}(i)}\mathbf{d}(i) - \lambda\frac{z_{\mathbf{x},\mathbf{u}_i}}{||\mathbf{u}_i||}. \tag{30}$$

C. CW features

Naming of the vector $\mathbf{d}$ in Eq. 29 was chosen on purpose, since CW offers an analogy with the well-known DC-DM (Distorsion Compensated Dither Modulation [35]) watermarking scheme where the dither is used to hide the location of the quantization cells. However, note that by construction CW is invariant to the scaling attack, contrarily to DC-DM schemes. We show on Fig. 8 the analogous of Fig. 4 and Fig. 7 for CW. Since Circular Watermarking renders carrier modulation jointly circular, we obtain dependency among message modulations, thus even powerful practical BSS attacks [36] are hopeless when using CW. There are no independent directions to allow for reliable carrier estimation using ICA.

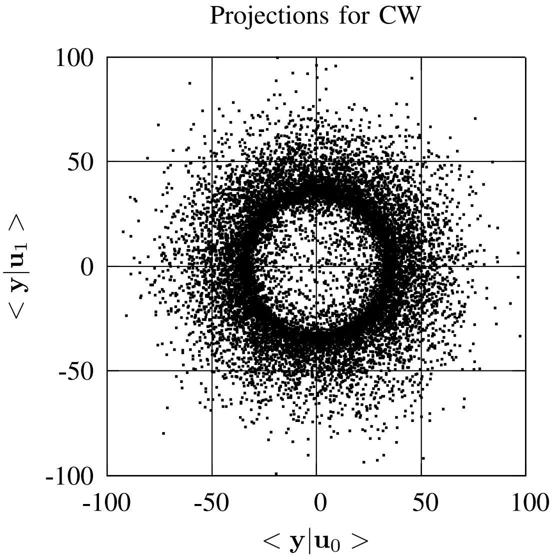

Fig. 8. Joint-distribution of the projection of two carriers for CW. The clusters of ISS have been spread over the entire corresponding decoding region thus leading to key-security.

Additionaly, the expression of the BER for the AWGN channel is given in Appendix III. We depict on Fig. 9 the BER comparison between SS, ISS, NW and CW. We believe this figure points out what is the cost of true security for SS-based watermarking techniques. Interestingly enough, at typical WCR of -21dB, CW performs close to SS but is always outperformed by ISS. At higher WCR however (-15dB), the performance of CW compared to other modulations degrades.

D. Evaluation of NW and CW security using BSS

The aim of this section is to assess the theoretical properties of NW and CW. We have used the estimation setup proposed in section III-B for classical SS and ISS considering the same parameters ($N_c = 2$, $N_o = 1000$). The distortion for NW remains the same ($WCR = -21dB$). Normalized correlations between the two estimated and original carriers are depicted on Fig. 10 for 100 different trials. For NW, the estimation of the secret carriers is unsuccessful because every point is very close to the origin for each trial. The CW plot illustrates the fact that in this case the watermark subspace is estimated (the distance between each point and the origin is close to 1), but that the estimation of the two carriers is not possible because each trial leads to a point which seems to be randomly chosen on the unitary circle.

VI. Theoretical and practical analysis of the security of Natural and Circular Watermarking

A. Information theoretic constraints

As previously, for the sake of simplicity and clarity, we assume in the following notations that $\mathbf{Y}$ is a set of N_o random vectors, each of size N_v: $\mathbf{Y} = (\mathbf{y}_1^\top, \ldots, \mathbf{y}_{N_o}^\top) =$

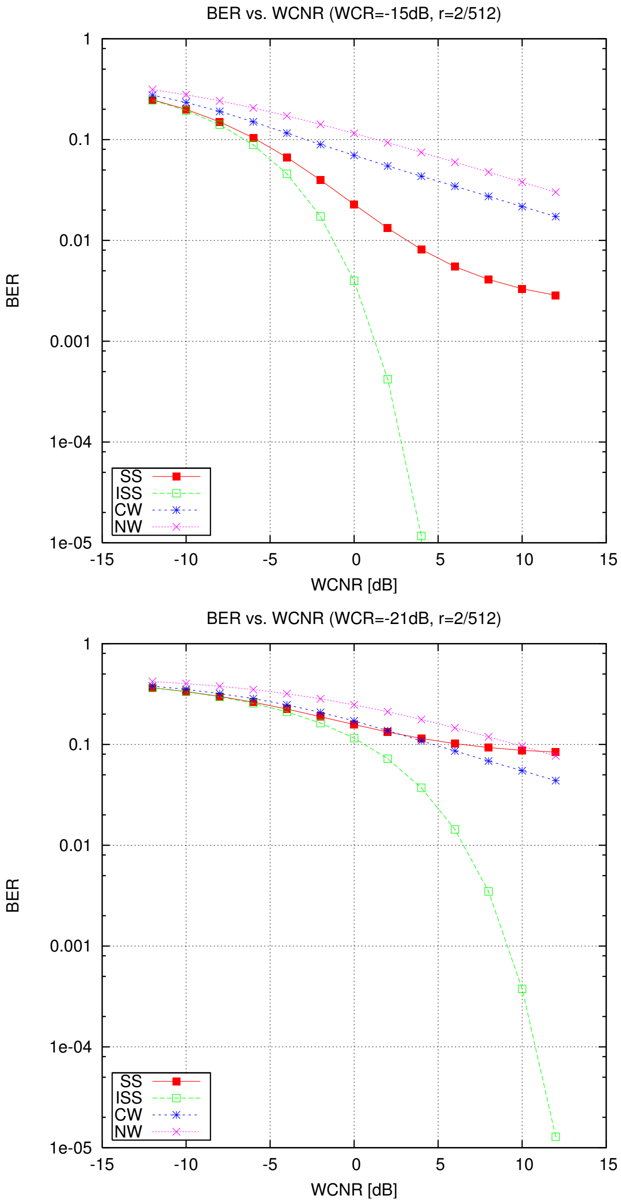

Fig. 9. BER comparison for SS, ISS, NW ($\eta = 1$) and CW. WCR=-21dB.

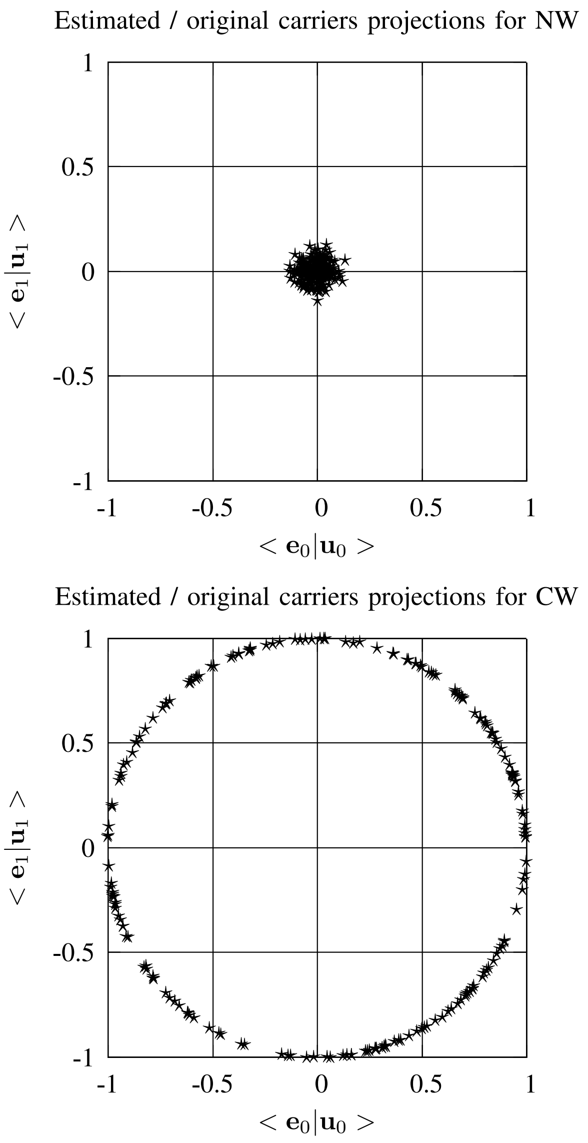

Fig. 10. Normalized correlations between the two estimated carriers and the original ones. For both schemes: $N_o = 1000$, $N_c = 2$ and $N_v = 512$.

$(\mathbf{y}_{1,1}, \ldots, \mathbf{y}_{N_v,1}, \ldots, \mathbf{y}_{1,N_o}, \ldots, \mathbf{y}_{N_v,N_o})$ and that $\mathbf{U}$ is a set of N_c random carriers, each of size N_v: $\mathbf{U} = (\mathbf{u}_1^\top, \ldots, \mathbf{u}_{N_c}^\top) = (\mathbf{u}_{1,1}, \ldots, \mathbf{u}_{N_v,1}, \ldots, \mathbf{u}_{1,N_c}, \ldots, \mathbf{u}_{N_v,N_c})$. Consequently $\mathbf{Y}$ represents a random vector of size $N_v \times N_o$ and $\mathbf{U}$ represents a random vector of size $N_v \times N_c$.

The assessment of the security of a data-hiding scheme has been already proposed by Cachin [2] and extended to the WOA, KOA and KMA scenarios by Comesaña *et al.* [6]. It is defined by the mutual information between the observed watermarked contents and the secret key that has been used to watermark these contents, given the hypothesis that the secret key is constant for each realization of a set of watermarked random vectors $\mathbf{Y}$. In the case of WOA, it can be written as:

$$I(\mathbf{Y}_\mathbf{U}, \mathbf{U}). \tag{31}$$

The definition of mutual information is linked with the differential entropy by the relation:

$$I(\mathbf{Y}_\mathbf{U}, \mathbf{U}) = h(\mathbf{Y}_\mathbf{U}) - h(\mathbf{Y}_\mathbf{U}|\mathbf{U}). \tag{32}$$

It has been shown [6] that perfect secrecy may be achieved iff:

$$I(\mathbf{Y}_\mathbf{U}, \mathbf{U}) = 0, \tag{33}$$

which means that $\mathbf{Y}_\mathbf{U}$ and $\mathbf{U}$ are independent sets of random vectors and that it is not possible to gain any information about $\mathbf{U}$ observing $\mathbf{Y}_\mathbf{U}$.

B. Theoretical evaluation of security for non-robust Natural Watermarking ($\eta = 1$)

In this section we propose to calculate the different pdfs $p(\mathbf{Y}|\mathbf{U})$ and $p(\mathbf{Y}, \mathbf{U})$ and then apply Eq. 32 to compute the information leakage. Considering the embedding formula, for $N_c = N_o = 1$ we have, considering all the different possibilities of embedding :

$$\begin{aligned} p(\mathbf{Y_U}|\mathbf{U}) = p(\mathbf{Y}|\mathbf{U}) &= \frac{1}{4}(\\ p(\mathbf{Y}|\mathbf{U}, b = 0, \mathbf{Y}^t\mathbf{U} \geq 0) &+ p(\mathbf{Y}|\mathbf{U}, b = 1, \mathbf{Y}^t\mathbf{U} \geq 0) + \\ p(\mathbf{Y}|\mathbf{U}, b = 0, \mathbf{Y}^t\mathbf{U} < 0) &+ p(\mathbf{Y}|\mathbf{U}, b = 1, \mathbf{Y}^t\mathbf{U} < 0)). \end{aligned}$$

By grouping the second and third terms together and the first and the fourth term, this is equivalent to

$$p(\mathbf{Y}|\mathbf{U}) = \frac{1}{2}p(\mathbf{X}) + \frac{1}{2}p((\mathbf{I}_{N_v} - 2\frac{\mathbf{U}\mathbf{U}^\top}{N_v})\mathbf{X}), \qquad (34)$$

where $\mathbf{I}_{N_v}$ represents the identity matrix of size $N_v \times N_v$. Under the *i.i.d.* host signal distribution assumption, the last pdf of equation Eq. 34 is equal to $p(\mathbf{X})/|\det(\mathbf{I}_{N_v} - 2\mathbf{U}\mathbf{U}^\top/N_v)|$ [37]. Furthermore, $\mathbf{I}_{N_v} - 2\mathbf{U}\mathbf{U}^\top/N_v$ is an elementary reflection and an orthogonal matrix so $|\det(\mathbf{I}_{N_v} - 2\mathbf{U}\mathbf{U}^\top/N_v)| = 1$ and we obtain:

$$p(\mathbf{Y}|\mathbf{U}) = p(\mathbf{X}). \qquad (35)$$

If $N_c > 1$, $\mathbf{U}$ is composed of a set of N_c orthogonal different carriers and we can perform the analysis presented above for each carrier independently: Eq. 35 is still valid. If $N_o > 1$, it is also possible to decompose the N_o watermarked observations as $\mathbf{Y} = \mathbf{H}\mathbf{X}$ where $\mathbf{H}$ is a $N_oN_v \times N_oN_v$ orthogonal matrix. Finally we have $p(\mathbf{Y}|\mathbf{U}) = p(\mathbf{X})$ for all secret key $\mathbf{U}$ and consequently the Natural Watermarking embedding is shown to be stego-secure in the case of *i.i.d.* Gaussian host signals. Moreover, using Eq. 35 and the Bayes' theorem, we have:

$$\begin{aligned} p(\mathbf{Y_U}) &= \int_{\mathcal{K}} p(\mathbf{Y}, \mathbf{U}) d\mathbf{U} \\ &= \int_{\mathcal{K}} p(\mathbf{U}) p(\mathbf{Y}|\mathbf{U}) d\mathbf{U} \\ &= p(\mathbf{X}) \int_{\mathcal{K}} p(\mathbf{U}) d\mathbf{U} \end{aligned}$$

and we finally have:

$$p(\mathbf{Y_U}) = p(\mathbf{Y}|\mathbf{U}), \qquad (36)$$

which means that there is no information leakage for Natural Watermarking when $\eta = 1$:

$$I_{NW}(\mathbf{Y_U}, \mathbf{U}) = 0. \qquad (37)$$

C. Practical estimation of mutual information for random vectors

In the case of $\eta > 1$ (robust NW) or the case of CW the computation of Eq. 31 is not analytically tractable and consequently we have to use a practical estimation. This is essentially due to the fact that the pdf $p(\mathbf{Y})$ is impossible to explicit literally in those cases.

For low dimensions, practical solutions to compute differential entropy and mutual information are based on histogram and kernel-based pdf estimation [38]. However when the dimension of the random vector is too high (greater than 3), such methods are not accurate enough because they suffer from the curse of dimensionality (the number of samples that are necessary to estimate the pdf grows exponentially with the number of variables). Consequently we have decided in our experiments to use an estimator of differential entropy based on K-Nearest Neighbour (KNN) [39], [40]. The approximation of the differential entropy of a N_v-dimensional vector $\mathbf{V}$ is given by:

$$\hat{h}(\mathbf{V}) = \psi(N_v) - \psi(k) + \log \mathcal{V}_{N_v} + \frac{N_v}{N_r} \sum_{i=1}^{N_r} \log \epsilon(i, k), \qquad (38)$$

where:

- $\psi(x)$ is the digamma function,
- k is the order of the nearest neighbour,
- V_{N_v} is the volume of the unitary N_v-dimensional sphere,
- N_r is the number of vectors considered,
- and $\epsilon(i, k)$ is the distance between the i^{th} vector and its k^{th} nearest neighbour.

This estimation function enables to compute accurate estimation of the differential entropy of a random vector in high dimension. For example, for a vector composed of 10 *i.i.d.* normal components of respective variances 1,2,...,10, the theoretical entropy is equal to $0.5\log((2\pi e)^{N_v}10!) \approx 21.74$ nats and after 100 trials the average approximation obtained using Eq. 38 is 22.01 nats (variance of 0.005) for $k = 3$ and $N_r = 1000$.

D. Set-up and practical results

This section uses Eq. 38 to derive an approximation of Eq. 32 and calculate the information leakage for the different schemes. $\hat{h}(\mathbf{Y})$ is computed by generating a set of $N_r = 1000$ realizations. Logically, one realization is computed using the same key but different keys are used for each realization. The approximation $\hat{h}(\mathbf{Y}|\mathbf{U})$ is computed using also N_r realizations, but in this case, all the realizations are generated using an arbitrary fixed key.

In our set-up we have chosen $N_c = 2$, $N_v = 8$, and the value of the mutual information is an average after 100 trials. Fig. 11 depicts the mutual information $I(\mathbf{Y}, \mathbf{U})$ considering only one observation ($N_o = 1$) and different embedding distortions for different SS-based watermarking schemes. The lowest distortion (-3.0 dB) corresponds to the distortion obtained for non-robust Natural Watermarking and $\sigma^2_{\mathbf{x}} = 1$. From this figure, we can observe two important properties. Firstly, the information leakage for non robust NW is equal to zero and the theoretical results presented above are confirmed. Secondly, considering only one observation, we can order the security of the studied watermarking schemes: the scheme having the most important information leakage is ISS, it is also the one that offers the most important robustness. Then CW and SS produces approximately the same information leakage with these parameters. Finally NW, even with its robust implementation, is the most secure of the proposed schemes.

Fig. 12 represents the approximation of the information leakage for the four schemes and an increasing number of observations. The practical results confirm again the results obtained in Sec. VI-B: there is no information leakage for NW ($\eta = 1$) even considering several observations. The information leakage for CW is lower than the information leakage for ISS or SS. Note however that the measure of mutual information does not enable to distinguish key-secure schemes from non-secure schemes. CW which is key-secure produces a positive information leakage and another measure should be used to define key-security in an information theoretic formulation taking into account $\mathcal{S}_{\mathbf{K}}$. CW information leakage therefore corresponds to the progressive disclosure of the whole private subspace.

Note that one might consider the WCRs of Figs. 11 and 12 not realistic. However, it was our intention to focus on WCRs leading to high information leakage values.

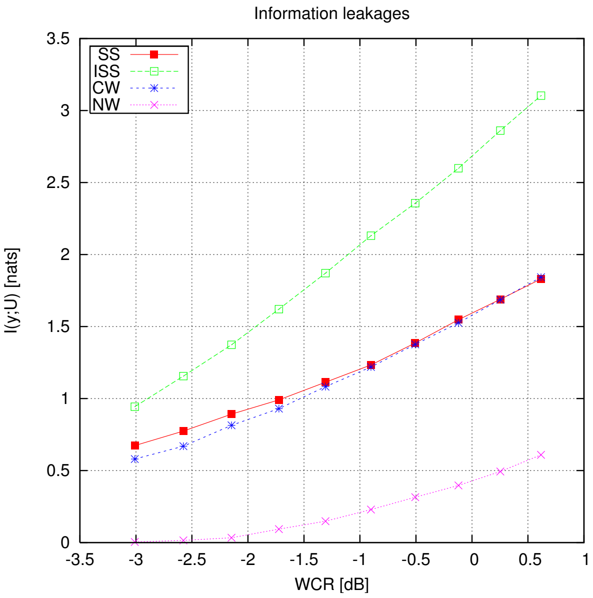

Fig. 11. Mutual information for NW ($\eta = 1$), CW, SS and ISS for one observation ($N_o = 1$) and different Watermark to Content Ratios (average after 100 trials). $N_c = 2$, $N_v = 8$, $N_r = 1000$. For ISS and CW: $NDR = 0dB$ (see [25] for definition of NDR).

VII. Conclusion

This paper aims at making a new entry in the field of framework-oriented papers for data-hiding [41]. It proposed a detailed analysis of the security of the embedding function for data-hiding following Kerckhoffs' principle. It leads to the definition of four security classes, which were illustrated by two new SS-based modulations for improved security. Stego-security is shown to be related to previous works on steganography. Subspace-security establishes the thin line between data-hiding robustness and security: if it is impossible to estimate the private subspace then no low-distorsion watermark removal is achievable. Key-security focuses on the impossibility for an attacker to estimate anything but random estimates of the secret key. Insecurity relates to data-hiding schemes where estimation of the secret key is achievable. Natural watermarking can be made stego-secure if the host signal is *i.i.d.* Gaussian and $\eta = 1$, otherwise it is only subspace-secure if $N_c = N_v$ or key-secure if $N_c \neq N_v$. Results show that Natural Watermarking has a relatively poor robustness. On the other hand, Circular Watermarking was devised to increase robustness at the cost of achieving only key-security which, we believe, offers good security for many appplications. Results are further assessed using information leakage measures that exhibit the superior security of our two new modulations compared to classical or improved spread spectrum. We believe this work broadens the choice for data-hiding application designers when high security is needed. Future works include application of these modulations to real media and subsequent assessment of security in the wild, possibly with the use of independent subspace analysis.

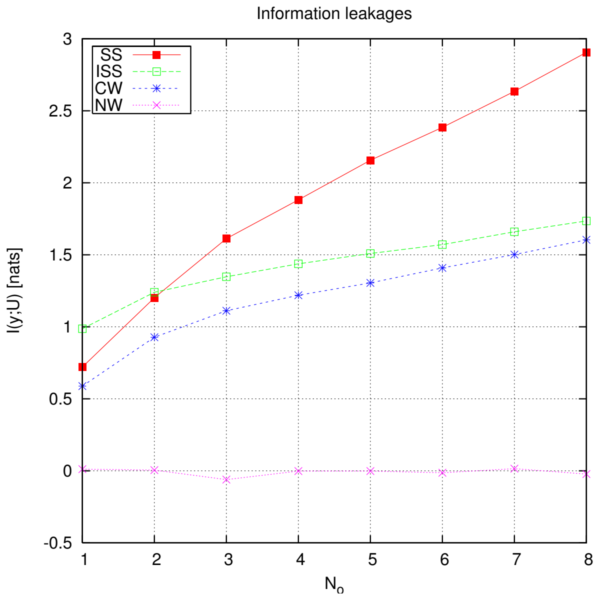

Fig. 12. Mutual information $I(\mathbf{Y_U};\mathbf{U})$ [nats] for NW, CW, SS and ISS for different number of observations (average after 10 trials). ($N_c = 2$, $N_v = 8$, $N_r = 5000$, $WCR = -3dB$). For ISS and CW: $NDR = 0dB$ (see [25] for definition of NDR).

Appendix I
WCR for NW

Like ISS, NW takes into account the very realization of the host signal. Therefore, we have to find the expectation of the WCR. The first point is to get the expectation of $\mathbb{E}[< \mathbf{x}, \mathbf{u}_i >^2$]:

$$< \mathbf{x}, \mathbf{u}_i >^2 = \sum_{k=0}^{N_v-1} (\mathbf{x}(k)\mathbf{u}_i(k))^2 + 2 \sum_{0\leq k<l}^{l\leq N_v-1} \mathbf{x}(k)\mathbf{u}_i(k)\mathbf{x}(l)\mathbf{u}_i(l). \tag{39}$$

Then, since $\mathbf{x}$ and $\mathbf{u}_i$ are considered independent variables, and by linearity of the expectation operator:

$$\begin{aligned} \mathbb{E}[< \mathbf{x}, \mathbf{u}_i >^2] &= \sum_{k=0}^{N_v-1} \mathbb{E}[\mathbf{x}^2(k)]\mathbb{E}[\mathbf{u}_i^2(k)] + \\ & \quad 2 \sum_{0\leq k<l}^{l\leq N_v-1} \underbrace{\mathbb{E}[\mathbf{x}(k)]}_{0} \underbrace{\mathbb{E}[\mathbf{u}_i(k)]}_{0} \underbrace{\mathbb{E}[\mathbf{x}(l)]}_{0} \underbrace{\mathbb{E}[\mathbf{u}_i(l)]}_{0} \\ &= \sum_{k=0}^{N_v-1} \sigma_{\mathbf{x}}^2 \sigma_{\mathbf{u}_i}^2 . \end{aligned}$$

Finally:

$$\mathbb{E}[< \mathbf{x}, \mathbf{u}_i >^2] = N_v \sigma_{\mathbf{x}}^2 \sigma_{\mathbf{u}}^2 . \tag{40}$$

Now we want to compute the watermark signal power. Let us first rearrange its expression by assuming that, averaging on several messages, $\Pr[\mathbf{m}(i) = 1] = \Pr[\mathbf{m}(i) = 0] = \frac{1}{2}$ and taking into account orthogonality of the $\mathbf{u}_i$:

$$\begin{aligned}
\mathbb{E}[\|\mathbf{w}\|^2] &= \sum_{i=0}^{N_c/2} (1+\eta)^2 \frac{\mathbf{E}[< \mathbf{x}, \mathbf{u}_i >^2]}{\|\mathbf{u}\|^2} + \\
&\quad \sum_{i=1+N_c/2}^{N_c-1} (1-\eta)^2 \frac{\mathbb{E}[< \mathbf{x}, \mathbf{u}_i >^2]}{\|\mathbf{u}\|^2} \\
&= \frac{(1+\eta)^2 + (1-\eta)^2}{\|\mathbf{u}\|^2} \sum_{i=0}^{N_c/2} \mathbb{E}[< \mathbf{x}, \mathbf{u}_i >^2] \\
&= \frac{2(1+\eta^2)}{N_v \sigma_{\mathbf{u}}^2} \sum_{i=0}^{N_c/2} \mathbb{E}[< \mathbf{x}, \mathbf{u}_i >^2].
\end{aligned}$$

After summation:

$$\begin{aligned}
\mathbb{E}[\|\mathbf{w}\|^2] &= \frac{N_c(1+\eta^2)}{N_v \sigma_{\mathbf{u}}^2} \mathbb{E}[< \mathbf{x}, \mathbf{u}_i >^2] \\
&= (1+\eta^2) N_c \sigma_{\mathbf{x}}^2.
\end{aligned}$$

Finally, the expectation of the WCR expressed in dB is:

$$\mathbb{E}[WCR] = 10 \log_{10} \left(\frac{(1+\eta^2) N_c}{N_v} \right). \qquad (41)$$

APPENDIX II
BER FOR NW AND $\eta = 1$

The correlation of the watermarked signal with a Gaussian noise of law $\mathcal{N}(0, \sigma_n^2)$ is Gaussian with law $\mathcal{N}(0, \sigma_N^2)$ and corresponding pdf:

$$f(x_1) = \frac{1}{\sqrt{2\pi}\sigma_N} \exp\left(\frac{-x_1^2}{2\sigma_N^2} \right), \qquad (42)$$

where $\sigma_N^2 = \sigma_{\mathbf{u}}^2 \sigma_{\mathbf{n}}^2$. If a 0 is coded, the correlation value z with a watermarked signal has a pdf of a half-Gaussian function:

$$\begin{aligned}
g(x_1) &= \frac{2}{\sqrt{2\pi}\sigma_W} \exp\left(\frac{-x_1^2}{2\sigma_W^2} \right) \quad \text{if} \quad x_1 \leq 0, &(43)\\
g(x_1) &= 0 \quad \text{if} \quad x_1 > 0, &(44)
\end{aligned}$$

where $\sigma_W^2 = \sigma_{\mathbf{u}}^2 \sigma_{\mathbf{x}}^2$.

The pdf of the watermarked signal which undergoes noise is:

$$(g * f)(t) = \int_{-\infty}^{+\infty} g(x_1) f(t - x_1) dx_1, \qquad (45)$$

which can be expressed as:

$$(g * f)(t) = A(t) \int_{-\infty}^{0} \exp - \frac{\left(x_1 - \frac{\sigma_W^2 t}{\sigma_N^2 + \sigma_W^2} \right)^2}{2 \left(\frac{\sigma_W \sigma_N}{\sqrt{\sigma_W^2 + \sigma_N^2}} \right)^2} dx_1, \qquad (46)$$

with $A(t) = \frac{\exp\left(\frac{-t^2}{2\sigma_N^2} + \frac{\sigma_W^2 t^2}{2\sigma_N^2 (\sigma_N^2 + \sigma_W^2)} \right)}{\pi \sigma_W \sigma_N}$.

After a variable substitution we can compute the integral part $I(t)$ as:

$$I(t) = \frac{\sqrt{2}\sigma_W \sigma_N}{\sqrt{\sigma_W^2 + \sigma_N^2}} \frac{\sqrt{\pi}}{2} \text{ erfc } \left(\frac{\sigma_W t}{\sqrt{2}\sigma_N \sqrt{\sigma_W^2 + \sigma_N^2}} \right), \qquad (47)$$

and finally :

$$P_e = \int_0^{+\infty} \mathcal{G}_{\sigma_W^2 + \sigma_N^2}(t) \text{ erfc } \left(\frac{\sigma_W t}{\sqrt{2}\sigma_N \sqrt{\sigma_W^2 + \sigma_N^2}} \right) dt, \qquad (48)$$

with $\mathcal{G}_{\sigma_W^2 + \sigma_N^2}(t) = \frac{1}{\sqrt{2\pi(\sigma_W^2 + \sigma_N^2)}} \exp\left(\frac{-t^2}{2(\sigma_W^2 + \sigma_N^2)} \right)$.

APPENDIX III
BER FOR CW (ISS-BASED IMPLEMENTATION WITH 2 CARRIERS)

In what follows, we assume α_{iss} and λ_{iss} arec computed from [25]. The correlation with a Gaussian noise of law $\mathcal{N}(0, \sigma_n^2)$ is Gaussian with a 2D corresponding pdf:

$$f(x_1, x_2) = \frac{1}{2\pi\sigma_N^2} \exp\left(\frac{-(x_1^2 + x_2^2)}{2\sigma_N^2} \right), \qquad (49)$$

with $\sigma_N^2 = \sigma_u^2 \sigma_n^2$.

The marginal pdf $f(x_2)$ is:

$$f(x_2) = \frac{1}{\sqrt{2\pi}\sigma_N} \exp\left(\frac{-x_2^2}{2\sigma_N^2} \right). \qquad (50)$$

After the Circular implementation of ISS we can assume that the pdf of the correlations with the carriers for watermarked contents if the first coded bit corresponds to 0 is given by:

$$g(x_1, x_2) = \frac{2}{4\pi\sqrt{\pi}\sigma_N \alpha_{iss}} \exp\left(- \frac{\left(\sqrt{x_1^2 + x_2^2} - \sqrt{2}\alpha_{iss} \right)^2}{2\sigma_N^2} \right) \text{ if } x_1 \leq 0, \qquad (51)$$

$$g(x_1, x_2) = 0 \text{ if } x > 0, \qquad (52)$$

with α_{iss} calculated from [25] and:

$$\sigma_{iss}^2 = (1 - \lambda_{iss})^2 \sigma_X^2 / (N_v \sigma_U^2).$$

The marginal pdf $g(x_2)$ is given by:

$$g(x_2) = \int_{-\infty}^{+\infty} g(x_1, x_2) dx_1, \qquad (53)$$

and must be computed numerically.

The pdf of the correlation of the watermarked signal which undergoes noise is:

$$(g * f)(t) = \int_{-\infty}^{+\infty} g(x_2) f(t - x_2) dx_2, \qquad (54)$$

and the probability of error P_e is expressed as :

$$P_e = \int_0^{+\infty} (g * f)(t) dt. \qquad (55)$$

Also note that these last two expressions have to be computed numerically.

ACKNOWLEDGMENT

The authors would like to take this opportunity to thank Teddy Furon for helpful discussions and Amaury Lendasse for insights on differential entropy estimation. The authors want to acknowledge the anonymous reviewers who helped to significantly improve the quality of this paper.

References

[1] G.J. Simmons, *The prisoners' problem and the subliminal channel*, Advances in Cryptology, Proc. CRYPTO'83, pp. 51–67, 1984.

[2] C. Cachin, *An Information-Theoretic Model for Steganography*, Information Hiding Workshop (IH) 2002, Lecture Notes in Computer Science 1525, pp. 306–318, Springer-Verlag, 1998.

[3] A.D. Ker, *Batch Steganography and Pooled Steganalysis*, Proc. 8th Information Hiding workshop, to appear in Springer LNCS, 2006.

[4] S. Katzenbeisser, *Computational Security Models for Digital Watermarks*, in Workshop on Image Analysis for Multimedia Interactive Services (WIAMIS'05), 2005.

[5] R. Chandramouli, *A Mathematical Framework for Active Steganalysis*, in Proc. ACM/Springer Multimedia Systems Special Issue on Multimedia Watermarking, vol. 9, pp. 303–311, Sept. 2003.

[6] P. Comesaña, L. Pérez-Freire and F. Pérez-González, *Fundamentals of Data-Hiding Security and their Application to Spread Spectrum Analysis*, Information Hiding Workshop (IH) 2005, Lectures Notes in Computer Science 3727, pp. 146–160, Springer-Verlag, 2005.

[7] T. Kalker, *Considerations on Watermarking Security*, in Proc. IEEE Conf. Multimedia Signal Processing, MMSP'01, pp. 201–206, Cannes, France, 2001.

[8] F. Cayre, C. Fontaine and T. Furon, *Watermarking Security: Theory and Practice*, IEEE Trans. Signal Processing, vol. 53, No. 10, pp. 3976–3987, oct. 2005.

[9] L. Pérez-Freire, F. Pérez-González and P. Comesaña, *Secret Dither Estimation in Lattice-Quantization Data-Hiding: A Set-Membership Approach*, in Edward J. Delp III and Ping W. Wong, Editors, *Security, Steganography and Watermarking of Multimedia Contents*, VIII, San Jose, California, USA, Jan. 06, SPIE.

[10] P. Bas and J. Hurri, *Vulnerability of DM Watermarking of Non-i.i.d. Host Signals to Attacks Utilizing the Statistics of Independent Components*, IEE Proceedings, Trans. on Information Security, vol. 153, pp. 127–139, 2006.

[11] P. Bas and G. Doërr, *Practical Security Analysis of Dirty Paper Trellis Watermarking*, Proc. 9th Information Hiding Workshop, St-Malo, France, jun. 2007.

[12] L. Pérez-Freire and F. Pérez-Gonzalez, *Disclosing Secrets in Watermarking and Data-Hiding*, Proc. 3rd. WAVILA Challenge, WACHA'07, St-Malo, France, jun. 2007.

[13] A. Kerckhoffs, *La Cryptographie Militaire*, Journal des Sciences Militaires, vol. IX, pp. 5–38, Jan. 1883, pp. 161–191, Feb. 1883.

[14] I.S. Moskowitz, G.E. Longdon and L. Chang, *A New Paradigm Hidden in Steganography*, Proc. Workshop on New Security Paradigms, WNSP'00, ACM Press, 2001.

[15] I.J. Cox, M.L. Miller and A.L. McKellips, *Watermarking as Communication with Side Information*, in Proc. IEEE, vol. 87, No. 7, pp. 1127–1141, 1999.

[16] M.L. Miller, G.J. Doërr and I.J. Cox, *Applying Informed Coding and Embedding to Design a Robust High-Capacity Watermark*, IEEE Trans. Image Processing, vol. 13, No. 6, pp. 792–807, 2004.

[17] A. Abrardo and M. Barni, *Informed Watermarking by Means of Orthogonal and Quasi-Orthogonal Dirty Paper Coding*, IEEE Trans. Signal Processing, vol. 53, No. 2, pp. 824–833, 2005.

[18] I.J. Cox, J. Kilian, F.T. Leighton and T. Shamoon, *Secure Spread Spectrum Watermarking for Multimedia*, IEEE Trans. Image Processing, vol. 6, Issue 12, pp. 1673–1687, dec. 1997.

[19] M.L. Miller, I.J. Cox and J. Bloom, *Informed Embedding Exploiting Image and Detector Information during Watermark Insertion*, in Proc. IEEE Intl. Conference on Image Processing, ICIP'00, vol. III, pp. 1–4, 2000.

[20] M.H. Costa, *Writing on Dirty Paper*, IEEE Trans. Information Theory, 29(3), pp. 439–441, May 1983.

[21] B. Hochwald and A. Nehorai, *Identifiability in Array Processing Models with Vector-sensor Applications*, IEEE Trans. Signal Processing, 44(1), Jan. 1996.

[22] J.E. Vila-Forcén, S. Voloshynovskiy, O. Koval, F. Pérez-González and T. Pun, *Worst Case Additive Attack Against Quantization-based Data-hiding Methods*, In Proc. of SPIE Photonics West, Electronic Imaging 2005, Security, Steganography, and Watermarking of Multimedia Contents VII (EI120), San Jose, USA, January 16-20 2005.

[23] P. Comesaña, L. Pérez-Freire and F. Pérez-González, *An Information-Theoretic Framework for Assessing Security in Practical Watermarking and Data-Hiding Scenarios*, 6th Intl. Workshop on Image Analysis for Multimedia Interactive Services, Montreux, Switzerland, April 2005.

[24] R.E. Blahut, *Principles and Pratice of Information Theory*, Addison-Wesley Publishing Co., ISBN 0-201-10709-0, 1987.

[25] H.S. Malvar and D. Florêncio, *Improved Spread Spectrum: A New Modulation Technique for Robust Watermarking*, IEEE Trans. Signal Processing, vol. 53, pp. 898–905, apr. 2003.

[26] G.J. Doërr, and J.-L. Dugelay, *Danger of Low-Dimensional Watermarking Subspaces*, ICASSP 2004, 29th IEEE International Conference on Acoustics, Speech, and Signal Processing, May 17-21, 2004, Montreal, Canada.

[27] A. Hyvärinen, *Fast and Robust Fixed-Point Algorithms for Independent Component Analysis*, IEEE Trans. on Neural Networks 10(3):626-634, 1999.

[28] H.W. Kuhn, *The Hungarian Method of Solving the Assignment Problem*, Naval Res. Logistics Quart., 2:83–97, 1955.

[29] P. Sallee, *Model-Based Steganography*, Int. Workshop on Digital Watermarking (IWDW) 2003, Lecture Notes in Computer Science 2939, pp. 154–167, Springer-Verlag, 2003.

[30] Y. Wang and P. Moulin, *Steganalysis of Block-Structured Stegotext*, in Proc. Security, Stenography and Watermarking of Multimedia Contents, San Jose, CA, USA, vol. 5306, pp. 477-488, SPIE, 2004.

[31] K. Solanki, K. Sullivan, U. Madhow, B.S. Manjunath and S. Chandrasekaran, *Provably Secure Steganography: Achieving Zero K.-L. Divergence Using Statistical Restoration*, to appear in Proc. IEEE Intl. Conference on Image Processing, ICIP'06, Atlanta, GA, USA, 2006.

[32] J. Fridrich, M. Goljan and R. Du, *Reliable Detection of LSB Steganography in Grayscale and Color Images*, in Proc. of the ACM Workshop on Multimedia and Security, Ottawa, Canada, October 5, 2001, pp. 27-30.

[33] A. Hyvärinen, J. Karhunen, and E. Oja, *Independent Component Analysis*, John Wiley & Sons, 2001.

[34] D. Knuth, *The Art of Computer Programming*, vol. 2 Seminumerical Algorithms, 2nd. Ed., Addison-Wesley, 1988.

[35] J.J. Eggers, R. Bäuml, T. Tzschoppe and B. Girod, *Scalar Costa Scheme for Information Embedding*, IEEE Trans. Signal Processing, vol. 51, No. 4, pp. 1003–1019, apr. 2003.

[36] F. Cayre, C. Fontaine and T. Furon, *Watermarking Attack: Security of WSS Techniques*, Int. Workshop Digital Watermarking (IWDW) 2004, IWDW 2004 Best Paper Award, Lecture Notes in Computer Science 3304, pp. 171–183, Springer-Verlag, 2005.

[37] A. Papoulis and U. Pillai, *Probability, Random Variables and Stochastic Processes*, Mc Graw Hill, 2002.

[38] D.W. Scott, *Multivariate Density Estimation: Theory, Practice, and Visualization*, John Wiley & Sons, 1992.

[39] L.F. Kozachenko and N.N. Leonenko, *A Statistical Estimate for the Entropy of a Random Vector*, Problems of Information Transmission, 23(2):9–16, 1987.

[40] A. Kraskov, H. Stögbauer and P. Grassberger, *Estimating Mutual Information*, Physical review. E, Statistical, Nonlinear, and Soft Matter Physics, 69(2):1–16, 2004.

[41] O. Altun, G. Sharma, M.U. Celik and M. Bocko, *A Set Theoretic Framework for Watermarking and its Application to Semifragile Tamper Detection*, IEEE Trans. Information Forensics and Security, vol. 1, no. 4, pp. 479–492, Dec. 2006.