Security analysis of robust data-hiding with geometrically structured codebooks

E.Topak[a], S.Voloshynovskiy[a], O. Koval[a], M. K. Mihcak[b], and T. Pun[a]

[a]University of Geneva - CUI, 24 rue du General-Dufour, CH 1211, Geneva 4, Switzerland;
[b]Microsoft Research, One Microsoft Way, Redmond, WA, 98052, USA

ABSTRACT

In digital media transfer, geometrical transformations desynchronize the communications between the encoder and the decoder. Therefore, an attempt to decode the message based on the direct output of the channel with random geometrical state fails. The main goal of this paper is to analyze the conditions of reliable communications based on structured codebooks in channels with geometrical transformations. Structured codebooks include codewords that have some features or statistics designed for synchronization purposes. In the design of capacity approaching data-hiding codes, host interference problem should be resolved. The solution to this problem is to perform the message coding based on random binning dependent on host-state. On the other hand, to achieve robustness to geometrical transformations, the codewords should have host independent statistics and encoding should be performed using random coding. To satisfy these conflicting requirements we propose Multiple Access Channel (MAC) framework where the message is split between two encoders designed based on the random binning and random coding principles. The message encoded according to random coding additionally serves for synchronization purposes. Sequentially, all existing methods that are proposed for reliable communications in channels with geometrical transformations are analyzed within the proposed MAC set-up. Depending on the particular codebook design, we classify these methods into two main groups: template-based codebooks and redundant codebooks. Finally, we perform the analysis of security leaks of each codebook structure in terms of complexity of the worst case attack.

Keywords: security analysis, robust data-hiding, structured codebooks, geometrical synchronization, achievable rate, Multiple Access Channel, random coding, random binning, Shannon's equivocation.

1. INTRODUCTION

In the context of digital media distribution via various channels, geometrical transformations, either applied intentionally by an attacker or happen unintentionally as the result of several operations or conversions, desynchronize the communications between the encoder and the decoder.

In terms of implementation, geometrical attacks are very simple and their computational complexity is very low. However, a direct attempt to decode the message from the geometrically distorted data at the decoder will fail due to the mismatch between the distorted watermark codeword in the attacked stego data and the corresponding watermark codeword in the codebook used by the decoder. One of the possible ways to perform a successful decoding is to estimate and to compensate the introduced geometrical transformation.

At the decoder in a classical communications set-up where no framework for the estimation of introduced geometrical transformation is provided, an exhaustive search in the space of all possible geometrical attacks is inevitable. Moreover, as the cardinality of the search space enlarges, the probability of error of the message decoding, or probability of false alarm in the detection, increases. Hence, the problem of geometrical transformations is a fundamental challenge in the design of robust data-hiding systems.

State-of-the-art methods robust to geometrical attacks are based on the approach inspired by the classical communications in channels with random or varying parameters that consists in estimation of the applied

Further author information: (Send correspondence to S. Voloshynovskiy): E-mail: svolos@cui.unige.ch, http://sip.unige.ch

geometrical transformation from the attacked data, known as *channel state estimation* (CSE), and successive compensation of the distortion, i.e. *channel state compensation* (CSC).

In contrast to the data-hiding based on random coding where the watermark codebooks are generated randomly, the watermark codebooks for the channels with geometrical transformations should have a special structure to handle the CSE. In the following, we will refer to these codebooks as *geometrically structured codebooks.* Depending on the particular codebook design, they are classified into two main groups:

- *template-based structured codebooks* in which a specially designed template or a pilot data is used to perform CSE and CSC[1,2];
- *redundant-based structured codebooks* in which codewords have special construction or statistics to aid CSE and CSC[3,4,5,].[6]

Although the practical usefulness of CSE and CSC was demonstrated in the papers referred above, a thorough theoretical analysis of this geometrical synchronization framework still remains an open and little studied problem. Furthermore, the security leakages of structured codebooks should be investigated from the position of designing the worst case attacks to destroy the reliable communications. Therefore, the goal of this paper is to put more light on the security and information-theoretical analysis of geometrically robust data-hiding.

The rest of the paper is organized as follows. In Section 2, the impact of geometrical attacks on the communications performance is considered. Afterwards, in Section 3, the information-theoretic framework to data-hiding synchronization is provided. Section 4 and Section 5 contain the analysis of the template-based structured codebooks and the redundant-based structured codebooks, respectively. In Section 6, the security leaks and attacking strategies for each structured codebook group are investigated. Finally, Section 7 concludes the paper.

Notations: We use capital letters to denote scalar random variables X, bold capital letters to denote vector random variables $\mathbf{X}$, corresponding small letters x and $\mathbf{x}$ to designate the realization of scalar and vector random variables, respectively. The superscript N is used to denote length-N vectors $\mathbf{x} = x^N = \{x[1], x[2], \ldots, x[N]\}$ with i^{th} element $x[i]$. We use $X \sim p_X(x)$ or simply $X \sim p(x)$ to indicate that a random variable X is distributed according to $p_X(x)$. Calligraphic fonts $\mathcal{X}$ designate sets $X \in \mathcal{X}$ and $|\mathcal{X}|$ denotes the cardinality of the set $\mathcal{X}$. $\mathbb{Z}$ and $\mathbb{R}$ stand for the set of integers and the set of real numbers, respectively.

2. INFLUENCE OF GEOMETRICAL ATTACKS ON THE COMMUNICATIONS PERFORMANCE

Consider the following generic additive data-hiding system:

$$\mathbf{Y} = \mathbf{W} + \mathbf{X}, \tag{1}$$

where a stego data $\mathbf{y} \in \mathcal{Y}^N$ of length N is obtained by adding a watermark codeword $\mathbf{w} \in \mathcal{W}^N$ containing an encoded message $m \in \mathcal{M}$, $\mathcal{M} = \{1, 2, \ldots, |\mathcal{M}|\}$ with $|\mathcal{M}| = 2^{NR}$, to a cover data $\mathbf{x} \in \mathcal{X}^N$. $R = \frac{1}{N}\log_2|\mathcal{M}|$ is the rate of communication.

When a certain geometrical transformation $T_A(.)$ is applied to the stego data $\mathbf{Y}$, the attacked data $\mathbf{V}$ is produced as:

$$\mathbf{V} = T_A(\mathbf{Y}), \tag{2}$$

where subscript A represents the class of potentially applied geometrical transformations. Affine, bilinear and projective transformations are among those that A can include. A geometrical transformation consists in change of pixel coordinates of $\mathbf{Y}$ and possibly an accompanying interpolation to fit into the discrete grid of digital data.

A can be parameterized by a set of J parameters $\mathbf{a} = (a_1, a_2, \ldots, a_J)$ such that $\mathbf{a} \in \mathbb{Z}^{J*}$. For example, when A takes the form of affine transformation subclass of general geometrical transformations, a pixel at the

*In general case one can assume $\mathbf{a} \in \mathbb{R}^J$.

coordinates (n_1, n_2) in $\mathbf{Y}$, i.e. $y[n_1, n_2]$, will be transferred to the new coordinates (n_1', n_2') in $\mathbf{V}$, i.e. $v[n_1', n_2']$, according to:

$$\begin{bmatrix} n_1' \\ n_2' \end{bmatrix} = \begin{bmatrix} a_1 & a_2 \\ a_3 & a_4 \end{bmatrix} \begin{bmatrix} n_1 \\ n_2 \end{bmatrix} + \begin{bmatrix} a_5 \\ a_6 \end{bmatrix}. \tag{3}$$

In this case, $\mathbf{a} = (a_1, a_2, a_3, a_4, a_5, a_6)$.

Since the data-hider is not informed about the applied geometrical transformation, it can be considered as random. Then, the total number of geometrical transformations, i.e. cardinality of the space of geometrical transformations, will be $|\mathcal{A}|$. However, in practical data-hiding applications, a geometrical attack space would not include all elements of $|\mathcal{A}|$ due to the visual acceptability constraint. For example, it is not expected that the attacker would rotate a stego data more than 10 degree. Nevertheless, for the sake of generality, we will consider the set of $\epsilon-$typical geometrical transformations[7] $\mathcal{A}_\epsilon^{(J)}(A)$, such that the sample entropy is $\epsilon-$close to the true entropy and $|\mathcal{A}_\epsilon^{(J)}| < |\mathcal{A}|$, as the space of possibly applied geometrical transformations. In the case when $\mathbf{a} \in \mathbb{R}^J$, we refer to the volume of the set instead of cardinality.[7]

Assume that the probability of decoding error for a particular realization of $\mathbf{A} = \mathbf{a}$ is $P_e^{(N)}(\mathbf{a})$ for watermark codewords of length N. Then, the average probability of decoding error $P_e^{G(N)}$ over all possible attacks can be computed by averaging $P_e^{(N)}(\mathbf{a})$ as:

$$P_e^{G(N)} = \sum_{\mathbf{a} \in \mathcal{A}_\epsilon^{(J)}} P_e^{(N)}(\mathbf{a}) p_{\mathbf{A}}(\mathbf{a}). \tag{4}$$

In a theoretical set-up, where the length N of data sequences approaches ∞, the average probability of decoding error $P_e^{G(N)}$ for the *random coding*[7] is upper bounded by $P_e^{G(N)} \leq 2^{NR} |\mathcal{A}_\epsilon^{(J)}| 2^{-N(I(W;V|K)-\delta)}$,[8] where the particular value of K specifies the codebook from the set $\{1, 2, \ldots, |\mathcal{K}|\}$ of all codebooks and δ is an arbitrary small positive number. Similarly, in a theoretical set-up based on the *random binning*,[8] $P_e^{G(N)}$ is upper bounded by $P_e^{G(N)} \leq 2^{N[R+R']} |\mathcal{A}_\epsilon^{(J)}| 2^{-N(I(U;V|K)-\delta)}$, where $\mathbf{u} \in \mathcal{U}^N$ is an auxiliary random variable and R' is the total number of sequences $\mathbf{U}$ that are generated for each message $M = m \in \{1, 2, \ldots, |\mathcal{M}|\}$.[9] In the random coding scenario, if the data-hider communicates with the rate R that satisfies the condition $R \leq I(W; V|K)$, then $P_e^{G(N)} \to 0$ as $N \to \infty$ and $\delta \to 0$. The complexity of decoding for the data-hider is proportional to $2^{NR} |\mathcal{A}_\epsilon^{(J)}|$. Similarly, in the random binning scenario, if the data-hider communicates with the rate R such that $R \leq I(U; V|K) - I(U; X|K)$, then $P_e^{G(N)} \to 0$ as $N \to \infty$ and $\delta \to 0$. In this case, the complexity of decoding is proportional to $2^{N[R+R']} |\mathcal{A}_\epsilon^{(J)}|$. Thus, beside the increase in the complexity of decoding, geometrical attacks do not have any impact on the performance of these theoretical set-ups.

However, in practical situations with a finite N, the encoding is based on random binning or random coding with expurgating bad codewords depending on whether the host state is taken into account or not in the encoding. The decoding is based on a maximum likelihood (ML) technique[10] and $P_e^{G(N)}$ is upper bounded by $P_e^{G(N)} \leq |\mathcal{A}_\epsilon^{(J)}| 2^{-NE_r(R|K)}$, where $E_r(R|K) = \max_{\rho \in [0,1]} \max_{p_{W|K}(w|k)} \left[E_0(\rho, p_{W|K}(w|k)) - \rho R \right]$ and $E_0(\rho, p_{W|K}(w|k)) = -\log_2 \sum_y \left[\sum_x p_{W|K}(w|k) p(y|w)^{\frac{1}{1+\rho}} \right]^{1+\rho}$. Furthermore, as $|\mathcal{A}_\epsilon^{(J)}|$ gets larger, the upper bound for $P_e^{G(N)}$ increases. Hence, in the case of practical set-ups, geometrical transformations completely disable the reliable communications.

In many applications, it is necessary to decrease the cardinality of the search space of the decoder for possible geometrical transformations to reduce the complexity of decoding both in theoretical and practical set-ups and to decrease the average probability of decoding error in practical cases. A way to accomplish this requirement is to introduce a synchronization framework into the scheme in the expense of dedicating some portion of the rate R, originally used for the message transmission, to the communication of synchronization data.

As an illustrative example, in Fig. 1(a), a dot represents a particular geometrical transformation $\mathbf{A} = \mathbf{a}$ in the space $\mathcal{A}_\epsilon^{(J)}$ of typical geometrical transformations. A decoder without a synchronization framework will consider all elements of this space as possibly applied geometrical transformation, i.e. it will perform decoding at each point of this space. However, the use of a geometrical synchronization framework reduces the search space from $\mathcal{A}_{\epsilon}^{(J)}$ to $\mathcal{A}'$ (Fig. 1(b)).

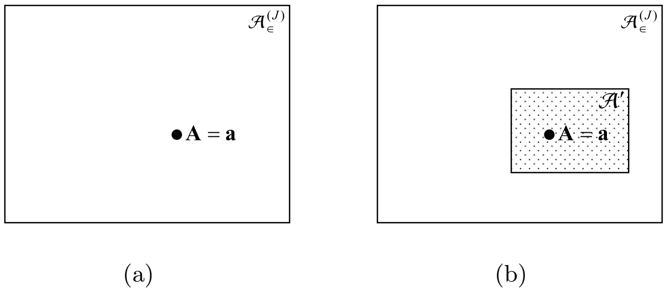

Figure 1. The original geometrical search space $\mathcal{A}_{\epsilon}^{(J)}$ (a) and the constrained search space $\mathcal{A}'$ after application of CSE/CSC (b).

The cardinality of $\mathcal{A}'$, i.e. $|\mathcal{A}'|$, is determined by the accuracy of CSE and CSC and depends on a particular design of structured codebook. As the variance of the estimation error goes to zero, constrained search space $\mathcal{A}'$ reduces to $|\mathcal{A}'| = 1$ ($\mathbf{A} = \mathbf{a}$, Fig. 1).

3. INFORMATION-THEORETIC FRAMEWORK TO DATA-HIDING SYNCHRONIZATION

The host interference to the message communication is an essential problem in the design of a practical capacity achieving robust data-hiding. The message encoding based on the random binning dependent on host state provides the solution to this problem. In contrast, robustness to geometrical attacks with an acceptable complexity requires the codewords of the synchronization part to have special features that are independent from the statistics of the host data.

To resolve these conflicting requirements, we propose the information-theoretic set-up presented in Fig. 2 that is based on a memoryless MAC with side information (SI) about the host state $\mathbf{X}$ non-causally available at one of the encoders. It consists of four alphabets $\mathcal{W}_1, \mathcal{W}_2, \mathcal{X}$ and $\mathcal{V}$, and is denoted by $\{\mathcal{W}_1 \times \mathcal{W}_2, \mathcal{X}, p(v|y), \mathcal{V}\}$. We also assume that the keys K_1 and K_2 are available at corresponding encoders and decoders.

Inputs to the channel, $\mathbf{W_1}$ and $\mathbf{W_2}$, are parts of the watermark $\mathbf{W}$ where $\mathbf{W_1}$ is dedicated to pure message communication and $\mathbf{W_2}$ is additionally used for geometrical synchronization purposes. Message M to be communicated is split into two parts, M_1 and M_2, depending on the rate pair (R_1, R_2) and they are encoded into $\mathbf{W_1}$ and $\mathbf{W_2}$ using the corresponding encoders.

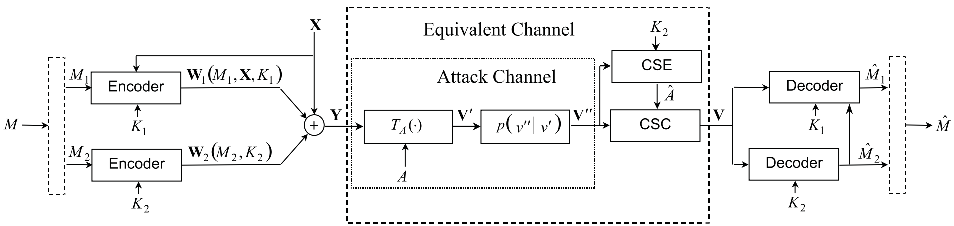

Figure 2. MAC framework to geometrically robust data-hiding.

A $(2^{NR_1}, 2^{NR_2}, N)$ code for the MAC with SI consists of two message sets $\mathcal{M}_1 = \{1, 2, \ldots, 2^{NR_1}\}$ and $\mathcal{M}_2 = \{1, 2, \ldots, 2^{NR_2}\}$, two encoding functions:

$$f_1 \; : \; \{1, 2, \ldots, 2^{NR_1}\} \times \{1, 2, \ldots, |\mathcal{K}_1|\} \times \mathcal{X}^N \rightarrow \mathcal{W}_1^N, \tag{5}$$

$$f_2 \; : \; \{1, 2, \ldots, 2^{NR_2}\} \times \{1, 2, \ldots, |\mathcal{K}_2|\} \rightarrow \mathcal{W}_2^N, \tag{6}$$

and a decoding function:

$$g : \mathcal{V}^N \times \{1, 2, \ldots, |\mathcal{K}_1|\} \times \{1, 2, \ldots, |\mathcal{K}_2|\} \rightarrow \{1, 2, \ldots, 2^{NR_1}\} \times \{1, 2, \ldots, 2^{NR_2}\}. \tag{7}$$

M_1 and M_2 are chosen randomly from the sets $\{1, 2, \ldots, 2^{NR_1}\}$ and $\{1, 2, \ldots, 2^{NR_2}\}$, respectively. The keys K_1 and K_2 determine the particular codebooks that are to be used by the corresponding encoders and decoders. Assuming that joint distribution of messages over the product set $\mathcal{M}_1 \times \mathcal{M}_2$ is uniform, the average probability of error for this code is defined as:

$$P_e^{(N)} = \frac{1}{2^{N(R_1+R_2)}} \sum_{(m_1,m_2) \in \mathcal{M}_1 \times \mathcal{M}_2} Pr[g(\mathbf{V}, K_1, K_2) \neq (m_1, m_2) | (M_1 = m_1, M_2 = m_2)]. \tag{8}$$

A rate pair (R_1, R_2) is said to be achievable, if there exists a $(2^{NR_1}, 2^{NR_2}, N)$ code with $P_e^{(N)} \rightarrow 0$ as $N \rightarrow \infty$. The capacity region of the MAC is the closure of the set of all achievable (R_1, R_2) rate pairs.

Codebook construction: Codebooks for $\mathbf{W}_1$ and $\mathbf{W}_2$ are generated randomly according to the random binning[9] and the random coding[7] principles, respectively, and revealed to corresponding encoders and decoders.

Encoding: A particular message M is partitioned into (M_1, M_2) depending on the rate pair (R_1, R_2). Then, the encoder for M_1 generates $\mathbf{W}_1(M_1, \mathbf{X}, K_1)$ using random binning by taking into account M_1, non-causal host state information $\mathbf{X}$ and the user-specified key K_1. Similarly, encoder for M_2 produces the codeword $\mathbf{W}_2(M_2, K_2)$ using random coding by considering M_2 and the particular key K_2. Afterwards, these two codewords are combined with the host state $\mathbf{X}$ and sent to the equivalent channel. Information transfer via this channel passes the following stages.

Geometrical Transformation: Attacker applies a geometrical transformation $T_A(.)$ from the set of $\epsilon-$typical geometrical transformations $\mathcal{A}_\epsilon^{(J)}(A)$ to the stego data $\mathbf{Y}$.

Probabilistic Channel: In order to prevent complete inversion of the applied geometrical transformation, the attacker might introduce additional noise to the attacked data $\mathbf{V}'$[†]. Assuming that the noise acts as a discrete memoryless channel (DMC), it converts the input $\mathbf{V}'$ to the output $\mathbf{V}''$ in a probabilistic manner according to the channel transition probability $p(\mathbf{v}''|\mathbf{v}') = \prod_{i=1}^{N} p(v_i''|v_i')$.

Synchronization: The output $\mathbf{V}''$ of the probabilistic channel is provided to CSE and CSC blocks for the synchronization. In fact, this is the part where the cardinality of the search space of the decoder for possibly applied geometrical transformations is reduced from $|\mathcal{A}_\epsilon^{(J)}|$ to $|\mathcal{A}'|$. The output $\mathbf{V}$ of this part is sent to decoders. Geometrical transformation, probabilistic channel and synchronization part form the equivalent channel with the input alphabets $\mathcal{W}_1$, $\mathcal{W}_2$, $\mathcal{X}$ and the output alphabet $\mathcal{V}$. Leaving the problem of intersymbol interference (ISI) outside of the scope of this paper, we assume that the channel output is produced according to the probabilistic mapping $p(\mathbf{v}|\mathbf{y}) = \prod_{i=1}^{N} p(v_i|y_i)$.

Decoding: At the lower decoder (Fig. 2) with the knowledge of the key K_2, $\widehat{M_2}$ is decoded first from $\mathbf{V}$ considering $\mathbf{W}_1$ as interference. Then, the output of this decoder (in assumption of errorless decoding of M_2), $\mathbf{W}_2$, is provided to upper decoder and $\widehat{M_1}$ is decoded from $\mathbf{V}$, with the knowledge of the key K_1, after subtracting $\mathbf{W}_2$ (genie-aided decoding[11]). In this way, the interference of $\mathbf{W}_2$ with respect to $\mathbf{W}_1$ is avoided.

[†]The noise in general can be signal dependent that takes into account interpolation effects.

The corresponding achievable rates for the given set-up have been investigated independently for non-watermarking applications[12]:

$$R_1 \leq \frac{1}{N}\left[I(\mathbf{U};\mathbf{V}|\mathbf{W}_2,K_1) - I(\mathbf{U};\mathbf{X}|K_1)\right], \quad (9)$$

$$R_2 \leq \frac{1}{N}\left[I(\mathbf{W}_2;\mathbf{V}|\mathbf{U},K_2)\right], \quad (10)$$

$$R_1 + R_2 \leq \frac{1}{N}\left[I(\mathbf{U},\mathbf{W}_2;\mathbf{V}|K_1,K_2) - I(\mathbf{U};\mathbf{X}|K_1)\right], \quad (11)$$

In (9), the knowledge of $\mathbf{W}_2$ implies the knowledge of K_2 and likewise in (10), the knowledge of $\mathbf{U}$ implies the knowledge of K_1.

The capacity region for the proposed set-up is presented in Fig. 3. The point of interest emphasized in this plot is determined by taking into account the fact that selecting the maximal R_1 value for the message communications has a higher priority than assigning the highest R_2 for the transmission of synchronization data. Therefore, the indicated point is the optimum one from the capacity region which satisfies the above mentioned concern.

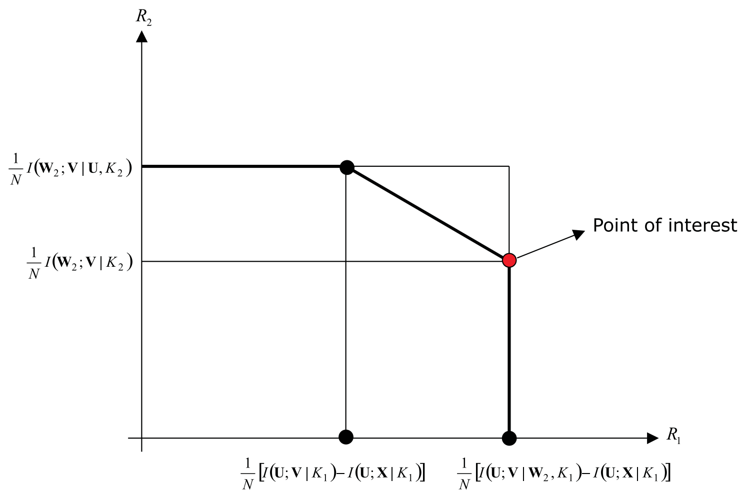

Figure 3. Capacity region of the proposed set-up.

According to the particular way of the synchronization part ($\mathbf{W}_2$) design, structured codebooks can be divided into two main groups: template-based structured codebooks and redundant-based structured codebooks. In the following sections, the properties of these two groups will be investigated in more details.

4. TEMPLATE-BASED STRUCTURED CODEBOOKS

The main idea of template-based synchronization is to use a specially designed pilot to estimate possible geometrical transformations applied to the stego data. Template data itself does not contain any information about the ongoing message transfer, i.e. $R_2 = 0$. It is key-dependent, $\mathbf{W}_2(K_2)$, unique for a given key $K_2 = k$ which is shared by the encoder and the decoder. Once the geometrical transformation is estimated and inverted based on the template $\mathbf{W}_2$, $\mathbf{W}_1$ is decoded from $\mathbf{V}$ after interference of $\mathbf{W}_2$ is canceled by subtraction.

The codebook construction with a template can be considered using Code Division Multiple Access (CDMA) and Space Division Multiple Access (SDMA) signaling approaches.

In case of the CDMA, $\mathbf{W}_1$ and $\mathbf{W}_2$ are transmitted simultaneously using power sharing since there is a constraint on the power of the total input signal $\mathbf{W} = \mathbf{W}_1 + \mathbf{W}_2$ to the channel defined by the distortion $d(\mathbf{X},\mathbf{Y}) = \frac{1}{N}\sum_{i=1}^{N} d(x_i,y_i) \leq \sigma_W^2$. It means that, if the total watermark power is σ_W^2, then $\lambda\sigma_W^2$ portion will be

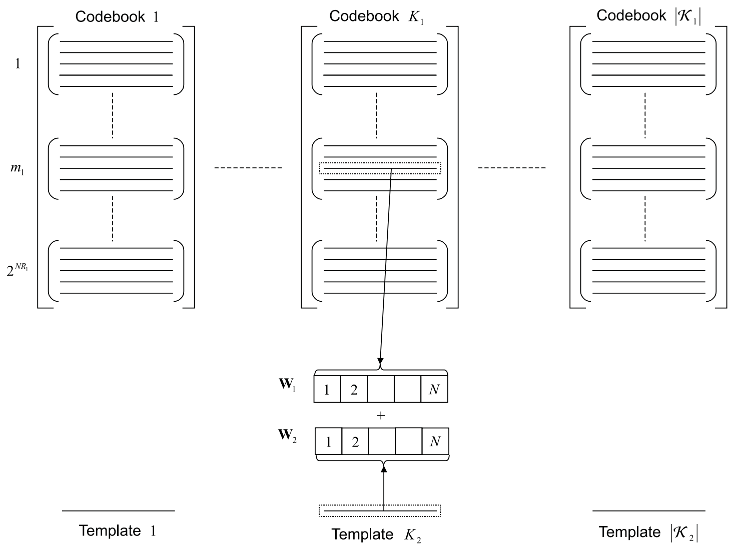

Figure 4. CDMA template-based structured codebooks.

assigned to the communication of $\mathbf{W}_1$ and the rest, $(1-\lambda)\sigma_W^2$, will be used to communicate $\mathbf{W}_2$. An example of template-based structured codebook based on the CDMA is given in Fig. 4.

In case of the SDMA, transmission of $\mathbf{W}_1$ and $\mathbf{W}_2$ is performed in orthogonal space intervals. Thus, interference between these two data is avoided. An example of template-based structured codebook based on the SDMA is given in Fig. 5.

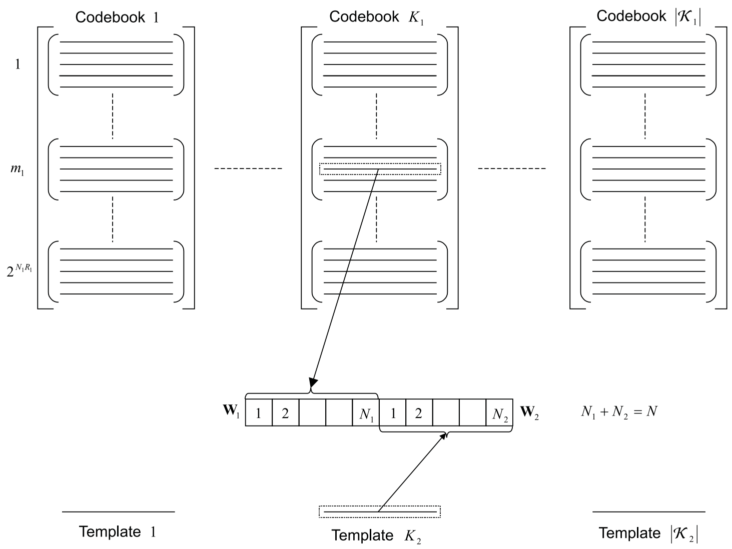

Figure 5. SDMA template-based structured codebooks.

5. REDUNDANT-BASED STRUCTURED CODEBOOKS

In redundant-based structured codebooks, $\mathbf{W}_2$ conveys M_2 part of the message M, i.e. $R_2 \neq 0$, using codewords that have a special construction or statistics designed to aid the synchronization. Once the geometrical transformation is inverted using the special structure of $\mathbf{W}_2$ and M_2 is decoded without error from $\mathbf{V}$, then $\mathbf{W}_1$ is decoded after $\mathbf{W}_2(M_2)$ is subtracted from $\mathbf{V}$.

As in the case of template-based structured codebooks, there are CDMA and SDMA approaches for the construction of redundant-based structured codebooks. An example of redundant-based structured codebook using the CDMA is given in Fig. 6 and another example based on the SDMA is presented in Fig. 7.

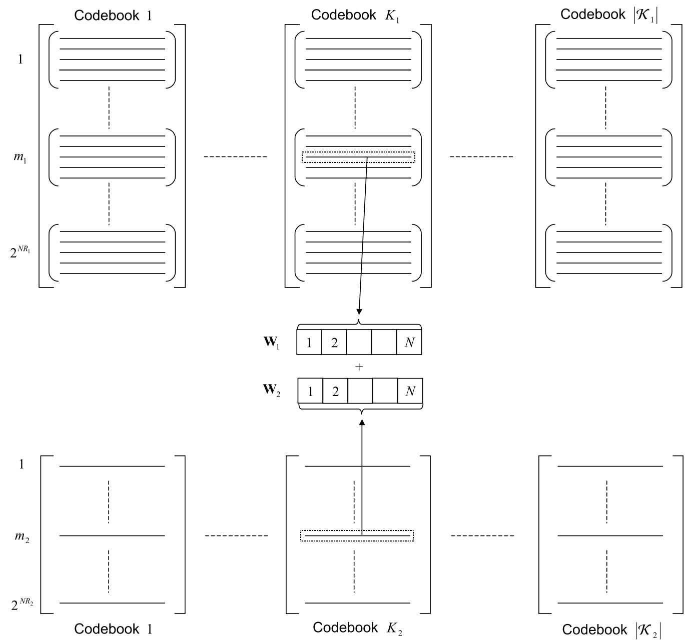

Figure 6. CDMA redundant-based structured codebooks.

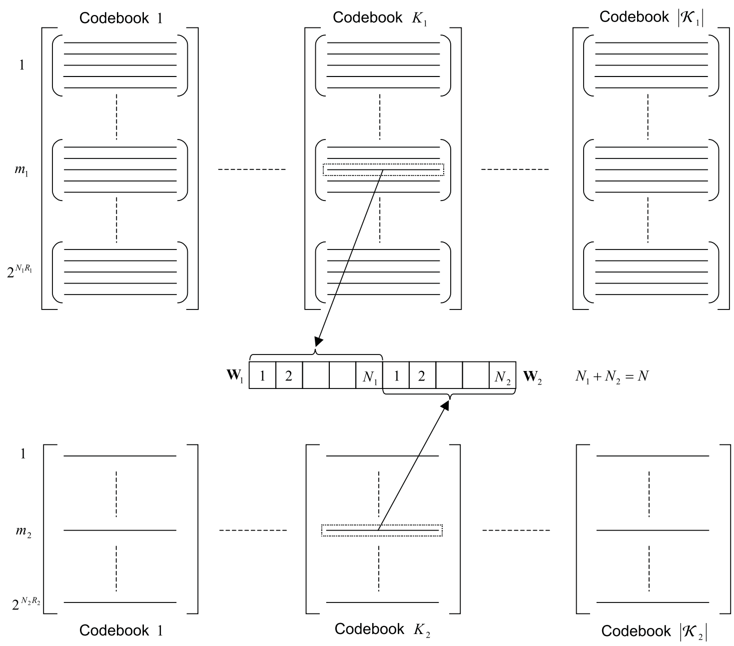

Figure 7. SDMA redundant-based structured codebooks.

6. ANALYSIS OF SECURITY LEAKS AND ATTACKING STRATEGIES

The objective of the attacker that operates between the encoder and the decoder would be to exploit all available prior information about the data-hiding scheme and all security leakages from the observed stego data $\mathbf{Y}$ to destroy reliable communications. In order to comply with *Kerckhoff principle*[13] in the design of a *secure data-hiding scheme*, it is assumed that the attacker has access to encoding and decoding algorithms and has the knowledge of codebooks used at both encoders and decoders as the prior information. Furthermore, it is supposed that the attacker does not know:

- secret keys K_1 and K_2, or particular codebooks that are exploited by encoders and decoders for ongoing communications,
- indexes M_1 and M_2 that are sent by corresponding encoders,
- the original host image $\mathbf{X}$ that carries communicated watermark codewords $\mathbf{W}_1$ and $\mathbf{W}_2$.

Under given conditions, the attacker may apply one of the following *attacking strategies*:

- Statistical signal processing attacks: the attacker exploiting the knowledge of statistics of the watermark and of the host data may estimate the watermark, subtract the estimate from the stego data and add noise, thus avoiding inverse mapping, to decrease the rate of reliable communications;
- Geometrical attacks: the attacker may find signal processing attacks inefficient since in some cases they are even invertible[14] and may decide to increase the complexity of decoding for the data-hider applying a geometrical attack to the stego data for desynchronization, which is simple in terms of implementation;
- Key space search attacks: the attacker with access to the decoder and with the knowledge of codebooks may prefer to perform "cryptographic like" attack by decoding through all possible codebooks, i.e. *exhaustive search*, and to subtract the decoded codeword from the stego data to destroy the communications. Due to the equivocation, every codebook has some security leaks that could simplify the search of the attacker.[16] Moreover, for robustness to geometrical attacks, we further introduce redundancy into the codebook structure. Thus, the attacker may try to benefit from the particular codebook design in reducing the search space.

In the following sections, attacking scenarios that are inspired by the given strategies for each group of structured codebooks based on the proposed MAC framework in Section 3 will be investigated in details for theoretical set-ups, i.e. for $N \to \infty$.

6.1. Attacks against Template-Based Structured Codebooks

Attacks against template-based structured codebooks benefit from the fact that template $\mathbf{W}_2$ is only key-dependent and unique for a particular key $K_2 = k$. Thus, the attacker with the access to codebooks given in Fig. 4 would look for a jointly-typical pair $(\widehat{\mathbf{W}}_2, \mathbf{Y})$. This search has the complexity of $|\mathcal{K}_2|$, where $|\mathcal{K}_2|$ represents the total number of codebooks for $\mathbf{W}_2$. If $\widehat{\mathbf{W}}_2$ is found, the attacker would subtract it from $\mathbf{Y}$ and apply a geometrical transform $\mathbf{A}$ to $\left(\mathbf{Y} - \widehat{\mathbf{W}}_2\right)$ to increase the complexity of decoding for the data-hider.

In this case, the data-hider, who had lost the synchronization framework based on $\mathbf{W}_2$ after this attack, would have to perform decoding at all possible geometrical transformations $\mathcal{A}_\epsilon^{(J)}(A)$. The complexity of decoding for the data-hider will be $|\mathcal{A}_\epsilon^{(J)}|2^{N(R_1+R')}$. Actually, the same result is presented also in Section 2, in the case of a random binning decoder without application of a geometrical synchronization framework.

After $\widehat{\mathbf{W}}_2$ is successfully decoded, instead of applying some geometrical transformation to $\left(\mathbf{Y} - \widehat{\mathbf{W}}_2\right)$, the attacker may further develop following attacks based on security leaks, depending on the key management protocol for K_1 and K_2:

- *The data-hider uses the same key at both encoders, i.e.* $K_1 = K_2 = K$*, and there is a one-to-one correspondence between the codebooks of* $\mathbf{W}_1$ *and* $\mathbf{W}_2$ *for a given key* K: the knowledge of template $\mathbf{W}_2$ implies the knowledge of corresponding codebook for $\mathbf{W}_1$ in such a design. After revealing $\mathbf{W}_2$ with complexity $|\mathcal{K}_2|$, the attacker would search in that particular codebook for a $\mathbf{U}$ that is jointly-typical with $\left(\mathbf{Y} - \widehat{\mathbf{W}}_2\right)$. This search has an additional complexity of $2^{N(R_1+R')}$ trials. After finding $\mathbf{U}$, the attacker can also obtain $\mathbf{X}$. For example, in the Costa set-up,[15] which is proposed for the Gaussian formulation of the Gel'fand-Pinsker problem,[9] $\mathbf{U} = \mathbf{W}_1 + \alpha\mathbf{X}$. Since $\mathbf{Y} - \widehat{\mathbf{W}}_2 = \mathbf{X} + \mathbf{W}_1$, $\mathbf{X}$ can be calculated if the jointly-typical $\left(\mathbf{U}, \left(\mathbf{Y} - \widehat{\mathbf{W}}_2\right)\right)$ pair is found.[17] The possibility for the attacker to obtain $\mathbf{X}$ means the total failure of the communications. Thus, the total complexity of the attacker is bounded by $|\mathcal{K}_2| + 2^{N(R_1+R')}$ trials.

- *The data-hider has different keys for each encoder, i.e.* $K_1 \neq K_2$*, and there is no relationship between the codebooks of* $\mathbf{W}_1$ *and* $\mathbf{W}_2$[2]: this time the knowledge of template $\mathbf{W_2}$ does not provide any information about the codebook from which current $\mathbf{W}_1$ in the stego data is coming. The attacker may perform an exhaustive search in all $|\mathcal{K}_1|$ codebooks for the $\mathbf{U}$ that is jointly-typical with $\left(\mathbf{Y} - \widehat{\mathbf{W}}_2\right)$. The complexity of this search is $|\mathcal{K}_1|2^{N(R_1+R')}$ that results in total complexity of $|\mathcal{K}_2| + |\mathcal{K}_1|2^{N(R_1+R')}$ trials.

- *The data-hider has different keys for each encoder, i.e.* $K_1 \neq K_2$*, but* K_2 *is fixed and is the same for all users*[1]: using the same template for all users, i.e. $|\mathcal{K}_2| = 1$, makes the scheme more susceptible to the attacks. Thus, when compared to the previous cases, it is relatively easy for the attacker to find $\mathbf{W}_2$. However, in terms of destroying the reliable communications, the same exhaustive search should be performed in all $|\mathcal{K}_1|$ codebooks for a jointly-typical $\left(\mathbf{U}, \left(\mathbf{Y} - \widehat{\mathbf{W}}_2\right)\right)$ pair with the complexity of $|\mathcal{K}_1|2^{N(R_1+R')}$ trials. Thus, the overall complexity is $1 + |\mathcal{K}_1|2^{N(R_1+R')}$ trials.

6.2. Attacks against Redundant-Based Structured Codebooks

In the case of redundant-based structured codebooks, codewords are constructed having special features or statistics to facilitate the geometrical synchronization at the decoder[3,4,5]. Therefore, one would expect the attacker to also try to benefit from these statistics in the search of $\mathbf{W}_2$ part. By observing the stego data $\mathbf{Y}$, the attacker could learn the statistics of $\mathbf{W}_2$ even when the key K_2 is not available. Furthermore, the knowledge of statistics for $\mathbf{W}_2$ reduces the ambiguity in finding $\mathbf{W}_2$. For the attacker with access to the codebooks given in Fig. 6, the upper bound for the complexity of finding $\mathbf{W}_2$ is $|\mathcal{K}_2|2^{NR_2}$. Once the attacker obtains $\mathbf{W}_2$, it is subtracted from $\mathbf{Y}$ and a geometrical transformation is applied to $\left(\mathbf{Y} - \widehat{\mathbf{W}}_2\right)$. This causes the data-hider to loose the synchronization framework and the complexity of decoding for the data-hider increases from $2^{N(R_1+R')}$ to $|\mathcal{A}_\epsilon^{(J)}|2^{N(R_1+R')}$.

However, instead of applying a geometrical transformation to $\left(\mathbf{Y} - \widehat{\mathbf{W}}_2\right)$, the attacker may develop the following attacks in order to destroy the reliable communications, depending on the statistical codebook design strategy for $\mathbf{W}_2$:

- *The statistics of* $\mathbf{W}_2$ *are the same for all codebooks*[3,4]: in this case the knowledge of $\mathbf{W}_2$ does not have any significance in reaching $\mathbf{W}_1$. Thus, the attacker has to perform an exhaustive search through all $|\mathcal{K}_1|$ codebooks for the jointly-typical $\left(\mathbf{U}, \left(\mathbf{Y} - \widehat{\mathbf{W}}_2\right)\right)$ pair. The complexity of this search is $|\mathcal{K}_1|2^{N(R_1+R')}$. If the jointly-typical pair is found, it is possible to obtain the realization of $\mathbf{X}$. The total complexity is bounded by $|\mathcal{K}_2|2^{NR_2} + |\mathcal{K}_1|2^{N(R_1+R')}$ trials.

- *The statistics of* $\mathbf{W}_2$ *are different for all user codebooks and there is a one-to-one relationship between the codebooks of* $\mathbf{W}_1$ *and* $\mathbf{W}_2$: in such a codebook design scenario, the knowledge of $\mathbf{W}_2$ restricts the search of the attacker for $\mathbf{W}_1$ in a particular codebook. Thus, the complexity of the attacker search for the jointly-typical $\left(\mathbf{U}, \left(\mathbf{Y} - \widehat{\mathbf{W}}_2\right)\right)$ pair reduces from $|\mathcal{K}_1|2^{N(R_1+R')}$ to $2^{N(R_1+R')}$ when compared to the previous scenario. Thus, the total complexity is reduced to $|\mathcal{K}_2|2^{NR_2} + 2^{N(R_1+R')}$ trials.

6.3. The effect of security leakages on the complexity of the search

In the random coding scenario, where the decoder looks for a jointly typical $(\mathbf{W}(M,K),\mathbf{Y})$ pair, the attacker, who has the knowledge of the decoding rule (or decoder) and targets at destroying reliable communications, has to find $\mathbf{W}(M,K)$. Once $\mathbf{W}(M,K)$ is found, it can be subtracted from $\mathbf{Y}$. Thus, without knowledge of the key K, one will perform an exhaustive search through all codebooks $\{1,2,\ldots,|\mathcal{K}|\}$ and all messages $M=m\in\mathcal{M}$ for the jointly typical $(\mathbf{W}(M,K),\mathbf{Y})$ pair. The complexity of this search will be $|\mathcal{K}|2^{NR}$, where $|\mathcal{K}|$ is the total number of codebooks and 2^{NR} is the number of codewords per codebook[‡].

When the codebooks $\{1,2,\ldots,|\mathcal{K}|\}$ are generated in the way that each one contains unique codewords and every possible $\mathbf{W}$ is included in only one codebook, the exhaustive search for $\mathbf{W}$ is related to the ambiguity $H(\mathbf{W})$ by $2^{H(\mathbf{W})}$. In this limit case, the complexities $|\mathcal{K}|2^{NR}$ and $2^{H(\mathbf{W})}$ will be equal.

However, as proposed by Shannon,[16] observing $\mathbf{Y}$[§] reduces the ambiguity about $\mathbf{W}$ from $H(\mathbf{W})$ to $H(\mathbf{W}|\mathbf{Y})$ as:

$$H(\mathbf{W}|\mathbf{Y}) = H(\mathbf{W}) - I(\mathbf{W};\mathbf{Y}), \tag{12}$$

where $I(\mathbf{W};\mathbf{Y})$ is the amount of information that can be learned about $\mathbf{W}$ by observing $\mathbf{Y}$. If $H(\mathbf{W}|\mathbf{Y})=0$, the knowledge of the current $\mathbf{Y}$ gives the exact value for $\mathbf{W}$, i.e. the complexity of the attacker search is $2^{H(\mathbf{W}|\mathbf{Y})}=1$. Therefore, in a communication scenario with $I(\mathbf{W};\mathbf{Y})\neq 0$, it is possible for the attacker to reduce the complexity $|\mathcal{K}|2^{NR}$ of the exhaustive search for the jointly typical $(\mathbf{W},\mathbf{Y})$ pair.

In the case of random binning, where the decoder looks for a $\mathbf{U}(M,\mathbf{X},K)$ that is jointly typical with the stego data $\mathbf{Y}$, the attacker will try to find the jointly typical $(\mathbf{U}(M,\mathbf{X},K),\mathbf{Y})$ pair through all $\{1,2,\ldots,|\mathcal{K}|\}$ codebooks and all message bins $M=m\in\mathcal{M}$ to be able to destroy reliable communications. The complexity of this search is given by $|\mathcal{K}|2^{N(R+R')}$ where $2^{NR'}$ is the total number of sequences $\mathbf{U}$ in each message bin $M=m$. The knowledge of $\mathbf{U}$ enables the attacker to get the host state $\mathbf{X}$, the message M and the key K.

Similarly to the random coding scenario, when the codebooks are generated by distributing all possible $\mathbf{U}$ sequences to the codebooks uniquely, the complexity of the attacker search depends on the ambiguity $2^{H(\mathbf{U})}$ about $\mathbf{U}$. Therefore, one would expect the complexities $|\mathcal{K}|2^{N(R+R')}$ and $2^{H(\mathbf{U})}$ to be equal in the limit case.

However, attacker's knowledge about the stego data $\mathbf{Y}$ reduces this ambiguity to $H(\mathbf{U}|\mathbf{Y})$ as:

$$\begin{aligned} H(\mathbf{U}|\mathbf{Y}) &= H(\mathbf{U}|\mathbf{X}) - [I(\mathbf{U};\mathbf{Y}) - I(\mathbf{U};\mathbf{X})], && (13)\\ &\leq H(\mathbf{U}) - [I(\mathbf{U};\mathbf{Y}) - I(\mathbf{U};\mathbf{X})], && (14) \end{aligned}$$

where the inequality follows since conditioning reduces the entropy.[7] Thus, as in the random coding case, if $I(\mathbf{U};\mathbf{Y})-I(\mathbf{U};\mathbf{X})\neq 0$, then the attacker search complexity can be decreased from $|\mathcal{K}|2^{N(R+R')}$ based on the observed $\mathbf{Y}$.

7. CONCLUSIONS

In this paper, the conditions of reliable communications based on structured codebooks in channels with geometrical transformations are analyzed from an information-theoretic point of view. Structured codebooks include codewords that have some features or statistics designed for synchronization purposes.

The MAC framework is developed to design the capacity achieving data-hiding codes that are robust to geometrical transformations. The corresponding methods based on the CSE/CSC framework that are proposed for reliable communications in channels with geometrical transformations are classified into two main groups depending on the particular codebook design: template-based codebooks and redundant codebooks. The analysis of security leaks of each codebook structure is performed in terms of complexity of the worst case attack design.

As a continuation of our research, we will consider collusion attacks, when there are several stego data copies produced from different hosts, keys or messages, and will emphasize the role of the host data statistics on

[‡]We do not consider here efficient search strategies similar to Viterbi algorithm.[18]

[§]It should be noticed that the attacker operates directly on the stego data $\mathbf{Y}$ contrarily to the data-hider who has access only to the attacked data $\mathbf{V}$.

the security. We will also extend the proposed set-up to real scenarios, when the data lengths N are finite, the decoding is performed using the ML technique and the probability of error is bounded in terms of error exponents. The particular search algorithms reducing the complexity of the attacker search based on the security leakages $I(\mathbf{W};\mathbf{Y})$ and $I(\mathbf{U};\mathbf{Y}) - I(\mathbf{U};\mathbf{X})$ are also a subject of our ongoing study.

8. ACKNOWLEDGMENT

This paper was partially supported by SNF Professeur Boursier grant PP002–68653, by the European Commission through the IST Programme under contract IST-2002-507932-ECRYPT and Swiss IM2 projects. The authors are also thankful to the members of SIP group for many helpful and interesting discussions during group seminars. The authors also acknowledge interesting discussions during several ECRYPT project meetings with Prof. F. Pérez-González and Prof. M. Barni.

REFERENCES

1. G. B. Rhoads, *Steganography systems*, International Patent WO 96/36163 PCT/US96/06618, November 1996.
2. S. Pereira and T. Pun, *Fast robust template matching for affine resistant image watermarking*, International Workshop on Information Hiding, September 29-October 1, 1999.
3. M. Kutter and F. A. P. Petitcolas, *A fair benchmark for image watermarking systems*, SPIE, 1999.
4. S. Voloshynovskiy, F. Deguillaume and T. Pun, *Content Adaptive Watermarking Based on a Stochastic Multiresolution Image Modeling*, EUSIPCO 2000, Tampere, Finland, September 5-8, 2000.
5. S. Voloshynovskiy, F. Deguillaume and T. Pun, *Multibit Digital Watermarking Robust Against Local Nonlinear Geometrical Distortions*, ICIP 2001, pp. 999-1002, Thessaloniki, Greece, 2001.
6. F. Deguillaume, S. Voloshynovskiy and T. Pun, *Method for the Estimation and Recovering from General Affine Transforms in Digital Watermarking Applications*, SPIE Photonics West, Electronic Imaging 2002, Security and Watermarking of Multimedia Contents IV, San Jose, CA, USA, January 20-24, 2002.
7. T. M. Cover and J. A. Thomas, *Elements of Information Theory*, 1991, John Wiley & Sons, Inc.
8. E. Topak, S. Voloshynovskiy, O. Koval and T. Pun, *Complexity Analysis of Geometrically Robust Data-Hiding Codes in Asymptotic Set-ups*, submitted to EUSIPCO 2005, Antalya, Turkey.
9. S. I. Gel'fand and M. S. Pinsker, *Coding for Channel with Random Parameters*, Problems of Control and Information Theory, Vol. 9 (1), pp. 19-31 (1980).
10. R. G. Gallager, *Information Theory and Reliable Communication*, John Wiley and Sons, New York, 1968.
11. A. Lapidoth and G. Kramer, *Topics in Multi-Terminal Information Theory-Course Notes*, Signal and Information Processing Laboratory, ETHZ, Switzerland, May 27, 2004.
12. S. Kotagiri and J. N. Laneman, *Achievable Rates for Multiple Access Channels with State Information Known at One Encoder*, in Proc. Allerton Conf. Communications, Control, and Computing, Monticello, IL, Oct. 2004.
13. A. Kerckhoff, *La cryptographie militaire*, Journal des sciences militaires 9 (1883), 5-38.
14. M. K. Mihcak, R. Venkatesan, and M. Kesal, *Cryptanalysis of Discrete-Sequence Spread Spectrum Watermarks*, Proceedings of the 5th International Information Hiding Workshop (IH 2002), Noordwijkerhout, The Netherlands, Oct. 2002.
15. M. H. M. Costa, *Writing on Dirty Paper*, IEEE Trans. on Information Theory, Vol. IT-29, No. 3, May 1983.
16. C. E. Shannon, *Communication Theory of Secrecy Systems*, Bell System Technical Journal, 28:656-715, October 1949.
17. S. Voloshynovskiy, O. Koval, E. Topak, J. Vila and T. Pun, *On Reversibility of Random Binning Techniques: Security and Multimedia Perspectives*, to be submitted to the International Workshop on Information Hiding, Barcelona, Catalonia (Spain), June 6-8, 2004.
18. J. Proakis, *Digital Communication*, 3rd ed., New York: McGraw-Hill, 1995.