# Information-Theoretic Analysis of Security in Side-Informed Data Hiding

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

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

**Abstract.** In this paper a novel theoretical security analysis will be presented for data hiding methods with side-information, based on Costa's dirty paper scheme. We quantify the information about the secret key that leaks from the observation of watermarked signals, using the mutual information as analytic tool for providing a fair comparison between the original Costa's scheme, Distortion Compensated - Dither Modulation and Spread Spectrum.

## 1 Introduction

In this paper a novel theoretical security analysis of data hiding methods based on Costa's capacity-achieving scheme [1] is presented. Security in this schemes is introduced by parameterizing the codebook by means of a secret key $\mathbf{\Theta}$, in such a way that an unauthorized agent, who does not know that key, will not be able, for instance, to decode the embedded message, because he does not know what particular codebook has been used in the embedding stage. The motivation for the analysis we will present here is the fact that information about the secret key may leak from the observation of watermarked signals, and this information leakage can be exploited by an attacker to refine his knowledge about the key $\mathbf{\Theta}$, in order to gain access to the decoding of secret embedded messages, removal of the watermark with low attacking distortion, or even generation of forged watermarked signals. In this sense, our analysis is inspired by Shannon's work [2] in the field of cryptography. A previous security analysis in the field of data hiding following this rationale has been accomplished in [3], but only for spread spectrum methods, and using the Fisher Information Matrix as analytic tool. Instead, our approach relies on the framework presented in [4], using the mutual information to measure the information leakage about the secret key, which in

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

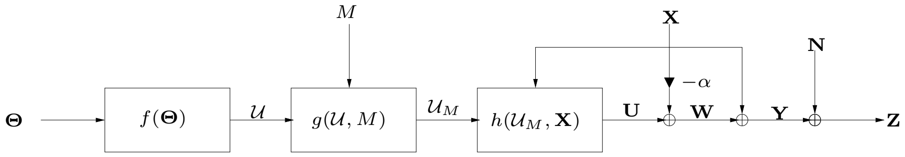


**Fig. 1.** Block diagram of Costa's schemes.

turn can be used to calculate the *residual entropy* (*equivocation* in Shannon's nomenclature) or ignorance about the key for the attacker after a certain number of observations.

We will distinguish two different scenarios where the security will be assessed: in the first scenario, the attacker only has access to $N_o$ signals watermarked with the same key, so the measure of information leakage will be given by the mutual information $I(\mathbf{Y}^1, \mathbf{Y}^2, \dots, \mathbf{Y}^{N_o}; \mathbf{\Theta})$, where $\mathbf{Y}^i$ denotes the $i$-th watermarked vector observed (the superscript will be omitted when only one observation is considered, for simplicity of notation). In the second scenario, the attacker has also access to the embedded messages, hence the information leakage in this case is measured by $I(\mathbf{Y}^1, \mathbf{Y}^2, \dots, \mathbf{Y}^{N_o}; \mathbf{\Theta}|\mathbf{M}^1, \mathbf{M}^2, \dots, \mathbf{M}^{N_o})$. In the rest of this paper, capital letters will denote random variables, whereas its specific values will be denoted by italicized lowercase letters, and boldface letters will indicate vectors of length $N_v$. Both scenarios have already been considered in the analysis in [3] for spread spectrum, under the names of *Watermarked Only Attack* (WOA), and *Known Message Attack* (KMA), so we will adopt here the same terminology.

In the following section, the security analysis of Costa's scheme will be accomplished, whereas Section 3 will be devoted to the analysis of Distortion Compensated - Dither Modulation (DC-DM) [5], which is a practical (suboptimal) implementation of Costa's scheme using structured codebooks. In Section 4, a comparison between the two analyzed schemes and spread spectrum will be given, and the main conclusions will be summarized.

## 2 Random codebooks (Costa's construction)

In Fig. 1 the considered framework is represented. In Costa's construction, the codebook is random by definition; however, this randomness can be parameterized by a secret key $\mathbf{\Theta}$, resulting in a codebook $\mathcal{U} = f(\mathbf{\Theta})$. Depending on the sent message $m$, one coset in the codebook will be chosen, namely $\mathcal{U}_m = g(\mathcal{U}, m)$. Taking into account the host signal $\mathbf{X}$ and the distortion compensation parameter $\alpha$ (which belongs to the interval [0,1]) the encoder will look for a sequence $\mathbf{U} = h(\mathcal{U}_m, \mathbf{X})$ belonging to $\mathcal{U}_m$ such that $|(\mathbf{U}-\alpha\mathbf{X})^t\mathbf{X}| \leq \delta$, for some arbitrarily small $\delta$. The watermark signal will be $\mathbf{W} = \mathbf{U}-\alpha\mathbf{X}$, and the watermarked signal $\mathbf{Y} = \mathbf{X}+\mathbf{W}$. Finally, the decoder will observe $\mathbf{Z} = \mathbf{X}+\mathbf{W}+\mathbf{N}$, where $\mathbf{N}$ is the channel noise, independent of both $\mathbf{X}$ and $\mathbf{W}$.

For the sake of simplicity, in this section we will focus on the analysis of this system when a single observation is available. We will also assume $\mathbf{X}$, $\mathbf{W}$ and

$\mathbf{N}$ to be i.i.d. random vectors with distributions $\mathcal{N}(\mathbf{0}, \sigma_X^2 \mathbf{I}_{N_v})$, $\mathcal{N}(\mathbf{0}, P\mathbf{I}_{N_v})$ and $\mathcal{N}(\mathbf{0}, \sigma_N^2 \mathbf{I}_{N_v})$, respectively, where $\mathbf{I}_{N_v}$ denotes the $N_v$-th order identity matrix. The embedding distortion is parameterized by the *Document to Watermark Ratio*, defined as $\text{DWR} = 10\log_{10}(\sigma_X^2/P)$, and the distortion introduced by the attacking channel is parameterized by the *Watermark to Noise Ratio*, defined as $\text{WNR} = 10\log_{10}(P/\sigma_N^2)$.

### 2.1 Known Message Attack

Since knowledge of the secret key and the sent symbol implies knowledge of the coset in the codebook (i.e., $\mathcal{U}_m$), we can write

$$I(\mathbf{Y}; \mathbf{\Theta}|M) = h(\mathbf{Y}) - I(\mathbf{Y}; M) - h(\mathbf{Y}|\mathcal{U}_M).$$

In App. A.1, we show that if $\alpha > 0.2$, then

$$I(\mathbf{Y}; \mathbf{\Theta}|M) = \frac{N_v}{2} \log\left[\frac{P + \sigma_X^2}{(1-\alpha)^2\sigma_X^2}\right],$$

so

$$h(\mathbf{\Theta}|\mathbf{Y}, M) = h(\mathbf{\Theta}) - \frac{N_v}{2} \log\left[\frac{P + \sigma_X^2}{(1-\alpha)^2\sigma_X^2}\right]. \qquad (1)$$

Since each component of each sequence $\mathbf{U}$ follows a Gaussian distribution with power $P + \alpha^2\sigma_X^2$, and all of them are mutually independent, it follows that

$$h(\mathbf{\Theta}) = \frac{|\mathcal{U}|N_v}{2} \log\left[2\pi e(P + \alpha^2\sigma_X^2)\right],$$

where $|\mathcal{U}| = e^{I(\mathbf{U};\mathbf{Z})} = \left(\frac{[P+\sigma_X^2+\sigma_N^2][P+\alpha^2\sigma_X^2]}{P\sigma_X^2(1-\alpha)^2+\sigma_N^2(P+\alpha^2\sigma_X^2)}\right)^{N_v/2}$.

Eq. (1) shows that the higher the DWR is, the higher the residual entropy becomes, because the host signal is making difficult the estimation of the secret key. On the other hand, the larger $\alpha$, the smaller the residual entropy will be, since the self-noise is reduced and the estimation becomes easier. In Fig. 2, theoretical results are plotted for different values of the DWR.

### 2.2 Watermarked Only Attack

Again, knowledge of the secret key and the sent symbol implies knowledge of the coset in the codebook (i.e., $\mathcal{U}_m$). Therefore, we can write

$$I(\mathbf{Y}; \mathbf{\Theta}) = h(\mathbf{Y}) - I(\mathbf{Y}; M|\mathbf{\Theta}) - h(\mathbf{Y}|\mathcal{U}_M). \qquad (2)$$

In App. A.2, it is shown that if $\alpha > 0.2$

$$I(\mathbf{Y}; \mathbf{\Theta}) = \frac{N_v}{2} \log\left[\frac{(P + \sigma_X^2)\left(P\sigma_X^2(1-\alpha)^2 + \sigma_N^2(P + \alpha^2\sigma_X^2)\right)}{P(P + \sigma_X^2 + \sigma_N^2)(1-\alpha)^2\sigma_X^2}\right]. \qquad (3)$$

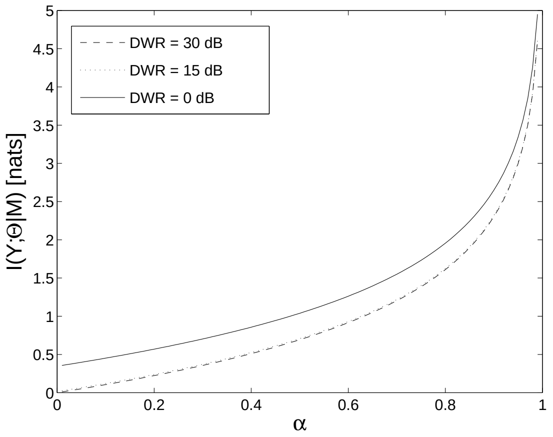


**Fig. 2.** $I(\mathbf{Y};\mathbf{\Theta}|M)$ for Costa in nats vs. $\alpha$ , for different values of DWR and $N_v = 1$.

Be aware that we are assuming that the watermarker transmits at the maximum reliable rate allowed, thus the power of the channel noise will affect the information leakage (this is further explained in App. A.2). For instance, when $\sigma_N^2 = 0$, the supremum of the maximum reliable rate is achieved, so the uncertainty about the sent symbol is also maximum, which complicates the attacker's work, yielding in this case $I(\mathbf{Y};\mathbf{\Theta}) = 0$ (*perfect secrecy* in the Shannon's sense [2]). In any case, using (3) we can write

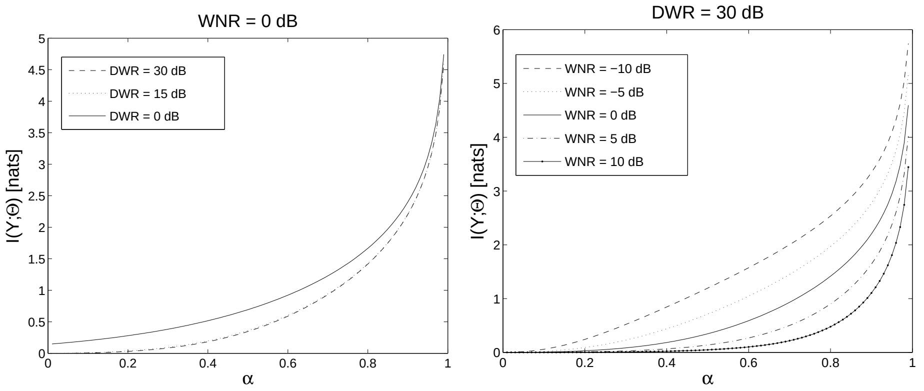
$$h(\mathbf{\Theta}|\mathbf{Y}) = h(\mathbf{\Theta}) - \frac{N_v}{2}\log\left[\frac{(P+\sigma_X^2)\left(P\sigma_X^2(1-\alpha)^2+\sigma_N^2(P+\alpha^2\sigma_X^2)\right)}{P(P+\sigma_X^2+\sigma_N^2)(1-\alpha)^2\sigma_X^2}\right]. \quad (4)$$

Theoretical results are plotted in Fig. 3, showing their dependence on the DWR, the WNR and $\alpha$. Since $I(\mathbf{Y};\mathbf{\Theta})$ depends on the transmission rate and this depends in turn on the WNR, the WNR has been fixed in order to plot the results. Under the light of these plots, several conclusions can be drawn:

- The information leakage increases with $\alpha$, because a smaller self-noise power is introduced.
- Conversely, the information leakage decreases for growing DWR's, because the uncertainty about the watermarked signal given the chosen $\mathbf{U}$ sequence is increased.
- The larger the WNR, the smaller the mutual information, because the embedder can achieve a higher reliable rate, thus increasing the uncertainty of the attacker about the sent symbol, which makes more difficult his job.

## 3 Distortion Compensated - Dither Modulation

We will focus on the scalar version of DC-DM [5] (also known as Scalar Costa Scheme, SCS [6]), for two reasons: first, for simplicity of the analysis, and second,



(a) (b)

**Fig. 3.** $I(\mathbf{Y};\mathbf{\Theta})$ vs. $\alpha$ in Costa, for different values of DWR and WNR = 0 dB (a), and for different values of WNR, setting DWR = 30 dB (b). In both plots, $N_v = 1$.

because it provides the fundamental insights into structured quantization-based methods. In DC-DM, embedding is made component-wise, hence, following the notation of Fig. 1, the codebook for the $k$-th component in DC-DM is given by the lattice

$$\mathcal{U}_{t_k} = \bigcup_{l=0}^{|\mathcal{M}|-1} \left( \alpha \Delta \mathbb{Z} + \alpha l \frac{\Delta}{|\mathcal{M}|} + \alpha t_k \right), \tag{5}$$

being $|\mathcal{M}|$ the number of different symbols, $k = 0, \ldots, N_v - 1$, and $t_k$ is the pseudo-random *dither signal* introduced to achieve randomization of the codebook. Each coset is chosen as a sublattice of (5), resulting in

$$\mathcal{U}_{m,t_k} = g(\mathcal{U}_{t_k}, m) = \alpha \Delta \mathbb{Z} + \alpha m \frac{\Delta}{|\mathcal{M}|} + \alpha t_k.$$

In DC-DM, $\alpha x_k$ is quantized to the nearest $u_k \in \mathcal{U}_{m,t_k}$, where $x_k$ is the $k$-th component of $\mathbf{X}$, which is assumed to be independent and identically distributed (i.i.d.). Thus, the expression of the $k$-th component of the watermarked signal is $y_k = x_k + u_k - \alpha x_k$, which can be rewritten as $y_k = x_k + \alpha \left( Q_{\Lambda_{m,t_k}}(x_k) - x_k \right)$, where $Q_{\Lambda_{m,t_k}}(\cdot)$ is an Euclidean quantizer with uniform step size $\Delta$ with its centroids defined by the shifted lattice $\Lambda_{m,t_k}$, according to the to-be-transmitted message symbol $m$:

$$\Lambda_{m,t_k} = \Delta \mathbb{Z} + m \frac{\Delta}{|\mathcal{M}|} + t_k. \tag{6}$$

The dither signal $\mathbf{t}$ may be any deterministic function of the secret key $\boldsymbol{\theta}$, i.e. $\mathbf{t} = f(\boldsymbol{\theta})$. If function $f$ is unknown, the only observation of watermarked vectors will not provide any information about $\boldsymbol{\theta}$, thus the target of the attacker is to disclose the dither signal used for embedding, or equivalently the location of the centroids. As it is usual in the analysis of quantization-based methods for data hiding [5], [6], we will assume a low-embedding-distortion regime, thus we can consider that the host pdf is uniform inside each quantization bin and all centroids occur with similar probabilities. This assumption (which we will refer to in the sequel as the *flat-host assumption*) implies that we can restrict our attention to the modulo-$\Delta$ version of $Y_k$ without any loss of information, considerably simplifying the theoretical analysis. The security level of the system will depend, obviously, on the statistical distribution of the dither. We show in App. B that the entropy of the watermarked signal $\mathbf{Y}$ only depends on the modulo-$\Delta$ version of the dither, and furthermore the distribution which maximizes the residual entropy is the uniform over the quantization bin; thus, hereafter we will assume that $T_k \sim U(-\Delta/2, \Delta/2)$ with i.i.d samples.

### 3.1 Known Message Attack

This is the simplest case to analyze. When only one watermarked vector is observed ($N_o = 1$), the following equalities hold

$$\begin{aligned} I(\mathbf{Y};\mathbf{T}|\mathbf{M}) &= \sum_{i=1}^{N_v}\sum_{j=1}^{N_v} I(Y_i;T_j|Y_{i-1},\dots,Y_1,T_{j-1},\dots,T_1,\mathbf{M}) & (7)\\ &= \sum_{i=1}^{N_v} I(Y_i;T_i|M_i) = N_v I(Y_i;T_i|M_i), & (8)\end{aligned}$$

where $Y_i$ denotes the $i$-th component of vector $\mathbf{Y}$, (7) follows from the chain rule for mutual informations [7], and (8) follows from the fact that the pairs $Y_i, T_j$ and $Y_i, M_j$ are independent $\forall\ i \neq j$, and furthermore $\{Y_i\}, \{T_i\}, \{M_i\}$ are i.i.d. processes. From the definition of mutual information we have

$$\begin{aligned} I(Y_i;T_i|M_i) &= h(Y_i|M_i) - h(Y_i|T_i, M_i) & (9)\\ &= h(Y_i|M_i = 0) - h(Y_i|T_i = 0, M_i = 0), & (10)\end{aligned}$$

where (10) follows from the flat-host assumption introduced above. Furthermore, due to this assumption, the entropies of (10) can be easily calculated by considering one period of the watermarked signal. Finally, (8) results in

$$\begin{aligned} I(\mathbf{Y};\mathbf{T}|\mathbf{M}) &= N_v I(Y_i;T_i|M_i) = N_v \left(\log(\Delta) - \log((1-\alpha)\Delta)\right)\\ &= -N_v \log(1-\alpha) \text{ nats }, & (11)\end{aligned}$$

so the residual entropy is

$$h(\mathbf{T}|\mathbf{Y},\mathbf{M}) = h(\mathbf{T}|\mathbf{M}) - I(\mathbf{Y};\mathbf{T}|\mathbf{M}) = N_v \log((1-\alpha)\Delta) \text{ nats }. \qquad (12)$$

Fig. 4 shows the result for the mutual information when $N_v = 1$. For the general case of $N_o$ observations one may be tempted to upper bound the mutual information by assuming that all observations will provide the same amount of information, but this bound will be too loose in most cases. For example, we can

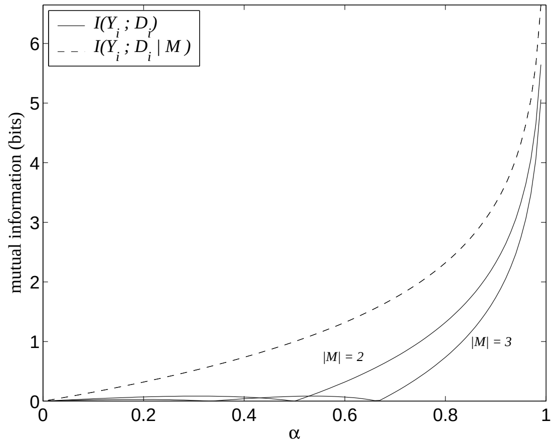


**Fig. 4.** Mutual informations for scalar DC-DM, in KMA and WOA cases with $N_v = 1$

calculate the exact mutual information when $\alpha \geq 0.5$, yielding (see Appendix C.1 for details)

$$I(\mathbf{Y}^1, \ldots, \mathbf{Y}^{N_o}; \mathbf{T}|\mathbf{M}^1, \ldots, \mathbf{M}^{N_o}) = N_v \left( -\log(1-\alpha) + \sum_{i=2}^{N_o} \frac{1}{i} \right) \text{ nats .} \qquad (13)$$

It can be seen in Fig. 5-(a) that the first observations provide most of the information about the secret dither signal. In Fig. 5-(b), numerical results are shown for $\alpha < 0.5$ up to 10 observations, showing that the linear upper bound gets tighter (at least for a small number of observations) when $\alpha$ is decreased.

### 3.2 Watermarked Only Attack

In this case, the only information at hand for the attacker is the watermarked vector; hence, we must calculate the mutual information $I(\mathbf{Y}; \mathbf{T})$. By reasoning as in the KMA case we can write

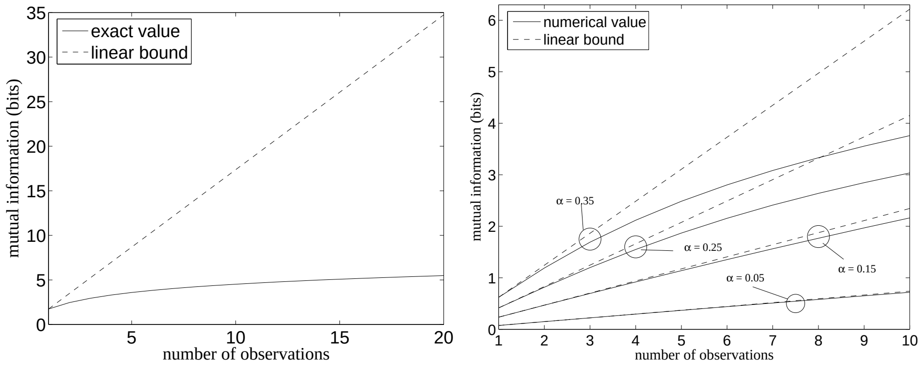
$$I(\mathbf{Y}; \mathbf{T}) = N_v I(Y_i; T_i) = N_v \left( h(Y_i) - h(Y_i|T_i) \right). \qquad (14)$$

Although it is always possible to obtain a theoretical expression for (14), we will calculate it here only for the case of binary signaling ($M_i = \{0, 1\}$), for the sake of simplicity. We have that $h(Y_i) = \log(\Delta)$, as in the KMA case, whereas for the term $h(Y_i|T_i)$ we have $h(Y_i|T_i) = h(Y_i|T_i = 0)$. Thus, we can write

$$I(Y_i; T_i) = \log(\Delta) - h(Y_i|T_i = 0),$$

For calculating $h(Y_i|T_i = 0)$ we must take into account that, for $\alpha < 0.5$, the pdf's associated to adjacent centroids overlap. It is easy to show that

$$h(Y_i|T_i = 0) = \begin{cases} \log(2(1-\alpha)\Delta) & \text{, for } \alpha \geq \frac{1}{2} \\ \log((1-\alpha)\Delta)\frac{(1-2\alpha)}{1-\alpha} + \log(2(1-\alpha)\Delta)\frac{\alpha}{(1-\alpha)} & \text{, for } \alpha < \frac{1}{2} \end{cases}$$



(a) (b)

**Fig. 5.** Mutual information as a function of the number of observations for scalar DC-DM in the KMA case, for $N_v = 1$, and $\alpha = 0.7$ (a), and $\alpha < 0.5$ (b).

With the above expressions, derivation of the equivocation is straightforward. In Fig. 4, results for 2 and 3 transmitted symbols are shown. It can be seen that when $\alpha = 0.5$ (for the binary case, $|\mathcal{M}| = 2$) the information leakage is null [1], thus achieving *perfect secrecy*, in the sense that the attacker can not gain knowledge about the dither, regardless the number of observations; this is because the pdf of the host and that of the watermarked signal are the same. When $\alpha < 0.5$ the information leakage is very small due to overlaps between adjacent centroids. For the case of multiple observations and $\alpha \geq 0.75$ we have (see App. C.2 for details)

$$I(\mathbf{Y}^1, \ldots, \mathbf{Y}^{N_o}; \mathbf{T}) = N_v \left( -\log(1-\alpha) - \log(2) + \sum_{i=2}^{N_o} \frac{1}{i} \right) \text{ nats} .$$

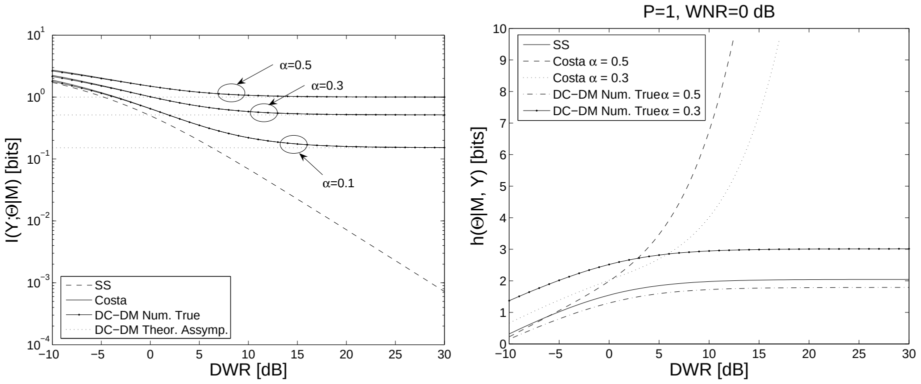
Then, the loss with respect to the KMA case is exactly log(2) nats (i.e. one bit), which is in accordance with the term $I(\mathbf{Y}; M|\mathbf{\Theta})$ of Eq. (2) obtained for Costa's scheme. For $\alpha < 0.75$, only numerical results have been obtained.

## 4 Comparison and conclusions

Fig. 6-(a) shows a comparison between the information leakage in Costa and DC-DM, for several values of $\alpha$. Notice that two different plots are shown for DC-DM: one with the theoretical results obtained in Section 3 under the flat-host assumption (DWR = $\infty$), and another plot with results obtained via numerical integration by considering finite DWR's [2]; it can be seen that both plots coincide

[1] It can be shown that for the general case of $D$-ary signalling, the values of $\alpha = k/D$, with $k = 0, \ldots, D-1$ yield null information leakage.

[2] The exact pdf of the watermarked signal has been numerically computed for finite DWR's by following the guidelines given in [8]



(a) (b)

**Fig. 6.** Comparison of information leakage (a) and residual entropy (b) for different data hiding schemes, with $N_v = 1$ and $N_o = 1$.

for any DWR of practical interest, thus supporting the validity of the flat-host assumption. The other remarkable result is the large resemblance between between Costa and DC-DM with finite DWR's.

An analysis similar to that carried out in Sections 2 and 3 is accomplished for spread spectrum data hiding in [4]; in this case, the secret parameter is the spreading sequence $\mathbf{s}$ used in the embedding stage, which will be assumed to be Gaussian with variance $\sigma_S^2$. The analysis for a single observation $(N_o = 1)$ yields the mutual information $I(\mathbf{Y};\mathbf{S}|\mathbf{M}) = \frac{N_v}{2}\log\left(1 + \sigma_S^2/\sigma_X^2\right)$ and the residual entropy $h(\mathbf{S}|\mathbf{Y}^1, \cdots, \mathbf{Y}^{N_o}, \mathbf{M}) = \frac{N_v}{2}\log\left(2\pi e\sigma_S^2\sigma_X^2/(\sigma_X^2 + \sigma_S^2)\right)$.

Fig. 6-(a) shows that, when compared under the same conditions, the information leakage for the informed embedding methods is larger than those of spread spectrum. However, note that the security level is not only given by the information leakage but depends also on the entropy of the secret key, yielding values for the residual entropy which are compared in Fig. 6-(b), where we can see that the theoretical Costa's scheme provides much larger residual entropies for practical values of the DWR. Similar comparisons could be made for the WOA attack.

Summarizing, we have shown in this paper that an attacker can take advantage of the observation of watermarked signals to gain knowledge about the secret key (when all observations were generated with the same key), and that knowledge of the embedded messages can simplify the attacker's work. Theoretically, the attacker needs, in general, an infinite number of observations to achieve perfect knowledge of the secret key, but in practice this is not necessary, since he only may be interested in obtaining an estimate of that key; in this sense, one can exploit the relationship between the residual entropy and the estimation error in order to find appropriate thresholds which define a given security level, but this path will not be explored here due to lack of space.

Appendix

# A Mutual information for a single observation in Costa's scheme

## A.1 Known Message Attack (KMA)

The mutual information between the received signal and the secret key when the sent message is known by the attacker can be written as

$$I(\mathbf{Y};\mathbf{\Theta}|M) = h(\mathbf{Y}|M) - h(\mathbf{Y}|\mathbf{\Theta}, M) = h(\mathbf{Y}) - I(\mathbf{Y};M) - h(\mathbf{Y}|\mathcal{U}_M). \quad (15)$$

Studying the second term, $I(\mathbf{Y};M) = h(\mathbf{Y}) - h(\mathbf{Y}|M)$, it can be seen to be 0 whenever $f_{\mathbf{Y}}(\mathbf{y}) = f_{\mathbf{Y}|M}(\mathbf{y}|M = m)$ for all the possible values of $m$. Taking into account that $\mathbf{Y} = f_1(\mathbf{\Theta}, M, \mathbf{X})$, this will be true in several cases. For example, if $\mathcal{U}_M$ is a lattice shifted by a random variable uniform over its Voronoi region (as in [9]); since the value of that random variable is not known by the attacker, the former equality is verified and $I(\mathbf{Y};M) = 0$. This will be also the case when $\mathcal{U}_M$ is a random codebook [1]; the attacker could know exactly all the $\mathbf{u}$'s in $\mathcal{U}$, but if he does not know the value of $M$ corresponding to each of them, the best he can do is to apply his a priori knowledge about $P(M = m)$, implying $I(\mathbf{Y};M) = 0$ again; this is the scenario studied here. Nevertheless, in the general case $0 \leq I(\mathbf{Y};M) \leq I(\mathbf{Y};M|\mathbf{\Theta})$.

To compute $h(\mathbf{Y}|\mathcal{U}_M)$ we will focus on the implementations using random codebooks. In those schemes every $\mathbf{u}$ in $\mathcal{U}_M$ has the same probability of being chosen. In order to facilitate the analysis, we will see $\mathbf{y}$ as the combination of a scaled version of $\mathbf{u}$ and a component orthogonal to $\mathbf{u}$, $\mathbf{y} = c\mathbf{u} + \mathbf{u}^{\perp}$; recalling that $\mathbf{u} = \mathbf{w} + \alpha\mathbf{x}$, we can write $\mathbf{u}^{\perp} = \mathbf{x} + \mathbf{w} - c\mathbf{w} - c\alpha\mathbf{x}$. Therefore, the value of $c$ can be computed taking into account that $\sigma_X^2 + P = c^2(P + \alpha^2\sigma_X^2) + \sigma_X^2(1 - c\alpha)^2 + P(1 - c)^2$; after some trivial algebraic operations, one obtains

$$c = \frac{P + \alpha\sigma_X^2}{P + \alpha^2\sigma_X^2}.$$

Since all the variables are Gaussian, if $N_v$ is large enough the samples of $\mathbf{y}$ will be very close to a sphere with radius $\sqrt{N_v \text{Var}\{\mathbf{U}^{\perp}\}}$ centered at some $c\mathbf{u}_o$; these spheres will be disjoint if[3] $\frac{\text{Var}\{\mathbf{U}^{\perp}\}}{c^2} < P$, which is true for any DWR if $\alpha > 0.2$. If this is the case, then we can write $h(\mathbf{Y}|\mathcal{U}_M) = h(\mathbf{Y}|\mathbf{U}) + \log(|\mathcal{U}_M|)$. Concerning $\log(|\mathcal{U}_M|)$, it is easy to see that

$$|\mathcal{U}_M| \approx e^{I(\mathbf{U};\mathbf{X})} = \left(\frac{P + \alpha^2\sigma_X^2}{P}\right)^{N_v/2}. \quad (16)$$

On the other hand,

$$h(\mathbf{Y}|\mathbf{U}) = h(\mathbf{U}^{\perp}) = \frac{N_v}{2}\log\left[2\pi e\frac{(1-\alpha)^2 P\sigma_X^2}{P + \alpha^2\sigma_X^2}\right], \quad (17)$$

[3] It can be shown that this is a sufficient, but not necessary, condition.

so,

$$\begin{aligned} h(\mathbf{Y}|\mathcal{U}_M) &= \frac{N_v}{2}\log\left[2\pi e\frac{(1-\alpha)^2P\sigma_X^2}{P+\alpha^2\sigma_X^2}\right]+\frac{N_v}{2}\log\left[\frac{P+\alpha^2\sigma_X^2}{P}\right] \\ &= \frac{N_v}{2}\log\left[2\pi e(1-\alpha)^2\sigma_X^2\right]. \end{aligned} \tag{18}$$

Note that this value is just an upper bound when the spheres described above are not disjoint.
Finally, the information leakage is given by

$$\begin{aligned} I(\mathbf{Y};\mathbf{\Theta}|M) &= \frac{N_v}{2}\log\left[2\pi e(P+\sigma_X^2)\right]-\frac{N_v}{2}\log\left[2\pi e(1-\alpha)^2\sigma_X^2\right] \\ &= \frac{N_v}{2}\log\left[\frac{P+\sigma_X^2}{(1-\alpha)^2\sigma_X^2}\right]. \end{aligned} \tag{19}$$

### A.2 Watermarked Only Attack (WOA)

In this case, the mutual information between the observations and the secret key is

$$\begin{aligned} I(\mathbf{Y};\mathbf{\Theta}) &= h(\mathbf{Y})-h(\mathbf{Y}|\mathbf{\Theta}) = h(\mathbf{Y})-I(\mathbf{Y};M|\mathbf{\Theta})-h(\mathbf{Y}|\mathbf{\Theta},M) \\ &= h(\mathbf{Y})-I(\mathbf{Y};M|\mathbf{\Theta})-h(\mathbf{Y}|\mathcal{U}_M) = I(\mathbf{Y};\mathbf{\Theta}|M)-I(\mathbf{Y};M|\mathbf{\Theta}). \end{aligned}$$

The only term that has not been analyzed yet is $I(\mathbf{Y};M|\mathbf{\Theta})$, which is the reliable rate that can be reached when the codebook is known. Note that the fact of not knowing the transmitted message produces a decrease in $I(\mathbf{Y};\mathbf{\Theta})$ equal to the transmission rate $I(\mathbf{Y};M|\mathbf{\Theta})$, since the increase in the uncertainty of the sent symbol complicates the attacker's work. In [1] it is shown that

$$I(\mathbf{Y};M|\mathbf{\Theta}) = \frac{N_v}{2}\log\left[\frac{P(P+\sigma_X^2+\sigma_N^2)}{P\sigma_X^2(1-\alpha)^2+\sigma_N^2(P+\alpha^2\sigma_X^2)}\right]. \tag{20}$$

So in this case, assuming again $\alpha > 0.2$, we can write

$$\begin{aligned} I(\mathbf{Y};\mathbf{\Theta}) &= \frac{N_v}{2}\log\left[2\pi e(P+\sigma_X^2)\right]-\frac{N_v}{2}\log\left[\frac{P(P+\sigma_X^2+\sigma_N^2)}{P\sigma_X^2(1-\alpha)^2+\sigma_N^2(P+\alpha^2\sigma_X^2)}\right] \\ &\quad -\frac{N_v}{2}\log\left[2\pi e(1-\alpha)^2\sigma_X^2\right] \\ &= \frac{N_v}{2}\log\left[\frac{(P+\sigma_X^2)\left\{P\sigma_X^2(1-\alpha)^2+\sigma_N^2(P+\alpha^2\sigma_X^2)\right\}}{P(P+\sigma_X^2+\sigma_N^2)(1-\alpha)^2\sigma_X^2}\right]. \end{aligned} \tag{21}$$

## B Optimal distribution for the dither in DC-DM

First, we show that for scalar DC-DM

$$I(\mathbf{Y};\mathbf{T}) = I(\mathbf{Y};\mathbf{T} \bmod \Delta). \tag{22}$$

We will assume without loss of generality that $\mathbf{Y}$ and $\mathbf{T}$ are scalars. Let $f_T(t)$ and $f_Y(y|T=t)$ denote the pdf of the secret key and the pdf of the watermarked signal conditioned on the dither, respectively. Taking into account that, due to the periodicity of the embedding lattices (6), $f_Y(y|T=t) = f_Y(y|T=t+i\Delta)\ \forall\ i$, then it is possible to write

$$\begin{aligned} f_Y(y) &= \int_t f_Y(y|T=t) f_T(t) dt = \int_0^{\Delta} f_Y(y|T=t) \sum_{i=-\infty}^{\infty} f_T(t+i\Delta) dt \\ &= \int_0^{\Delta} f_Y(y|T=t) f_{Tmod\Delta}(t) dt, \end{aligned} \tag{23}$$

where $f_{Tmod\Delta}(t)$ is the pdf of the modulo-$\Delta$-reduced version of $T$, hence equality (22) inmediately follows, whatever the distribution of the host and the dither.

Now, we consider what is the best choice for the dither from a security point of view. For simplicity of notation we define the random variable $Z \triangleq T \bmod \Delta$. It is a known fact that the uniform distribution maximizes the entropy in an interval [7], but the watermarker is interested in maximizing the residual entropy

$$h(Z|Y) = h(Z) - h(Y) + h(Y|Z). \tag{24}$$

In the following discussion we will make use of the flat-host assumption introduced in Section 3, thus we will consider that $-\Delta/2 \leq Y < \Delta/2$. We have that $h(Y|Z) = h(Y|Z=z)\ ,\forall\ z$, thus the rightmost term of (24) does not depend on the distribution of $Z$. Then, we must find $f_Z(z)$ such that $\{h(Z) - h(Y)\}$ is maximum. Let us define a random variable $V$ such that $f_V(v) \triangleq f_Y(y|Z=0)$. Under the flat-host assumption we have that $f_Y(y) = f_V(v) \circledast f_Z(z)$, where $\circledast$ denotes cyclic convolution over $[-\Delta/2, \Delta/2)$. Hence, the maximization problem can be written as

$$\max_{f_Z(z)} \left\{ h(Z) - h(V \oplus Z) \right\},$$

where $\oplus$ denotes modulo-$\Delta$ sum. We have the following lemma:

**Lemma 1.** *$h(Z) \leq h(V \oplus Z)$, with equality if $Z \sim U(-\Delta/2, \Delta/2)$.*

*Proof.* Consider that $f_{\tilde{V}}(\tilde{v})$ is the periodic extension of $f_V(v)$ over $n$ bins, properly scaled to ensure that $f_{\tilde{V}}(\tilde{v})$ is still a valid pdf, i.e.

$$f_{\tilde{V}}(\tilde{v}) = \frac{1}{n} \sum_{i=-n/2}^{n/2-1} f_V(v + i\Delta),$$

and that the same applies for $f_{\tilde{Z}}(\tilde{z})$. Now, define $\tilde{Q} \triangleq \tilde{V} + \tilde{Z}$. This way, $f_{\tilde{Q}}(\tilde{q})$ will be also periodic with period $\Delta$ in $n-2$ bins. Notice that

$$h(\tilde{Z}) = h(Z) + \log(n). \tag{25}$$

Assuming that for sufficiently large $n$ we can neglect the border effects, if we denote by $Q$ the modulo-$\Delta$ version of $\tilde{Q}$, we have that

$$h(\tilde{Q}) = h(Q) + \log(n). \tag{26}$$

We know that $h(\tilde{Z}) \leq h(\tilde{Q})$ [7], hence by (25) and (26) we have $h(Z) \leq h(V \oplus Z)$. To achieve equality it is sufficient to choose $Z$ such that $Z \sim U(-\Delta/2, \Delta/2)$. □

The proof of the lemma shows that the uniform over the quantization bin maximizes the residual entropy.

# C Mutual information for multiple observations in DC-DM

## C.1 Known Message Attack (KMA)

Assuming that the flat-host assumption introduced in Section 3 is valid, we will use the modulo-$\Delta$ version of the pdf of $Y_i$, hence $-\Delta/2 \leq Y_i < \Delta/2$. Without loss of generality, we consider that the transmitted symbol is the same ($M_i = 0$) in the $N_o$ observations. In the following, $\mathbf{Y}_k^{N_o}$ will denote a vector of $N_o$ observations of the $k$-th component of $\mathbf{Y}$, and $Y_{k,i}$ will be the $i$-th observation of that component. The mutual information after $N_o$ observations is given by

$$\begin{aligned} I(\mathbf{Y}_k^{N_o}; T_k | \mathbf{M}_k^{N_o}) &= \sum_{i=1}^{N_o} I(Y_{k,i}; T_k | \mathbf{M}_k^{N_o}, Y_{k,i-1}, \ldots, Y_{k,1}) & (27) \\ &= \sum_{i=1}^{N_o} \Big( h(Y_{k,i} | \mathbf{M}_k^{N_o}, Y_{k,i-1}, \ldots, Y_{k,1}) - \log((1-\alpha)\Delta) \Big). & (28) \end{aligned}$$

The problem here is the calculation of the leftmost conditional entropy term in (28), since the pdf of the $i$-th observation depends on the previous ones. However, the observations are independent when the dither is known, so we can write [4]

$$\begin{aligned} f(y_1, y_2, \ldots, y_{N_o}) &= \int_{t_{int}} \prod_{i=1}^{N_o} f_{Y_i}(y_i | T = t) f_T(t) dt \\ &= \frac{1}{(1-\alpha)^{N_o - 1} \Delta^{N_o}} \int_{t_{int}} f_{Y_{N_o}}(y_{N_o} | T = t) dt, \end{aligned} \tag{29}$$

with $f_T(t) = U(-\Delta/2, \Delta/2)$, and $t_{int}$ is the region of integration, given by

$$t_{int} = t \in (-\Delta/2, \Delta/2] \text{ such that } f(y_i | T = t) \neq 0 \; \forall \, i \leq N_o - 1, \tag{30}$$

which may be composed of disjoint intervals, in general. The obtention of the conditional pdf's by relying on (29) is straightforward, and the conditional entropy of (28) can be calculated as

$$h(Y_i | \mathbf{M}^{N_o}, Y_1, \ldots, Y_{i-1}) = \int h(Y_i | \mathbf{M}^{N_o}, Y_1 = y_1, \ldots, Y_{i-1} = y_{i-1}) dy_1 \ldots dy_{i-1}. \tag{31}$$

[4] We will obviate subindex $k$ in the following discussion, for simplicity of notation.

The integration limits in (30) can be specialized for $\alpha \geq 1/2$, resulting in

$$t_{int} = \begin{cases} \left[\max_i\{y_i - \mu\}, \min_i\{y_i + \mu\}\right) & \text{, if } |y_i - y_j| < 2\mu \; \forall \; i, j < N_o \\ 0 & \text{, otherwise} \end{cases} \tag{32}$$

where $\mu = (1-\alpha)\mathit{\Delta}/2$. For $\alpha > 1/2$ there is no overlapping between adjacent centroids, and the pdf of $Y_i$ conditioned on the previous observations can be analytically calculated, and the following conditional entropy is obtained

$$h(Y_i|\mathbf{M}^{N_o}, y_{i-1}, \ldots, y_1) = \frac{a}{(1-\alpha)\mathit{\Delta}} + \log\left((1-\alpha)\mathit{\Delta}\right) \text{ nats }, \tag{33}$$

where $a$ is half the volume of $t_{int}$. Substituting (33) into (31), we obtain

$$h(Y_i|\mathbf{M}^{N_o}, Y_{i-1}, \ldots, Y_1) = \log((1-\alpha)\mathit{\Delta}) + \frac{1}{(1-\alpha)\mathit{\Delta}} E_{f(y_1,\ldots,y_{i-1})}[a], \tag{34}$$

with

$$a = \frac{1}{2}((1-\alpha)\mathit{\Delta} + \min\{y_1, \ldots, y_{i-1}\} - \max\{y_1, \ldots, y_{i-1}\}). \tag{35}$$

Hence, the conditional entropy depends on the mean value of the integration volume.

Analytical evaluation of (34) can be simplified by assuming that the received samples $y_i$ are all independent but uniformly distributed around an unknown centroid [5] $t$: $Y_i \sim U(t - (1-\alpha)\mathit{\Delta}/2, t + (1-\alpha)\mathit{\Delta}/2)$. Under this assumption, let us define the random variable

$$X \triangleq min(Y_1, Y_2, \ldots, Y_{N_o}) - max(Y_1, Y_2, \ldots, Y_{N_o}).$$

The pdf of $X$ for $N$ observations can readily be shown to be

$$f_X(x) = N(N-1)\frac{(-x)^{N-2}}{((1-\alpha)\mathit{\Delta})^N}[(1-\alpha)\mathit{\Delta} + x],$$

with $x \in (-(1-\alpha)\mathit{\Delta}, 0]$. Hence, the mean value of $a$ results in

$$E_{f_X(x)}[a] = \frac{(1-\alpha)\mathit{\Delta}}{2}\left(1 - \frac{N-1}{N+1}\right),$$

and substituting it in (34), after some algebra, we finally obtain the following expression for the conditional entropy

$$h(Y_i|\mathbf{M}^{N_o}, Y_{i-1}, \ldots, Y_1) = \log((1-\alpha)\mathit{\Delta}) + \frac{1}{i} \text{ nats }, \text{ for } i > 1. \tag{36}$$

Substituting (36) in Eq. (28), we have the final expression for the mutual information

$$I(\mathbf{Y}_k^{N_o}; T|\mathbf{M}^{No}) = -\log(1-\alpha) + \sum_{i=2}^{N_o}\frac{1}{i} \text{ nats }. \tag{37}$$

[5] This simplification is possible since the chosen random variable yields the same mean value as the true distribution.

### C.2 Watermarked Only Attack (WOA)

For the WOA case, we must take into account that the observations may be associated to any of the possible cosets; however, with binary signaling ($M = \{0,1\}$) and $\alpha > 0.75$ there is no overlapping between the adjacent cosets, so the conditional pdf's can be calculated similarly to the case of KMA. Under these assumptions, it is possible to show that

$$h(Y_i|M_i, Y_{i-1}, \ldots, Y_1) = \log((1-\alpha)\mathit{\Delta}) + \frac{1}{i} + \log(2) \text{ nats },$$

hence the mutual information in the WOA case for $\alpha > 0.75$ is given by

$$I(\mathbf{Y}_k^{N_o}; T) = -\log(1-\alpha) - \log(2) + \sum_{i=2}^{N_o} \frac{1}{i} = I(\mathbf{Y}^{N_o}; T|\mathbf{M}^{No}) - \log(2) \text{ nats }.$$

## References

1. Costa, M.H.M.: Writing on dirty paper. IEEE Transactions on Information Theory **29** (1983) 439–441
2. Shannon, C.E.: Communication theory of secrecy systems. Bell system technical journal **28** (1949) 656–715
3. Cayre, F., Fontaine, C., Furon, T.: Watermarking security: application to a WSS technique for still images. In Cox, I.J., Kalker, T., Lee, H., eds.: Third International Workshop on Digital Watermarking. Volume 3304., Seoul, Korea, Springer (2004)
4. Comesaña, P., Pérez-Freire, L., Pérez-González, F.: Fundamentals of data hiding security and their application to Spread-Spectrum analysis. In: 7th Information Hiding Workshop, IH05. Lecture Notes in Computer Science, Barcelona, Spain, Springer Verlag (2005)
5. Chen, B., Wornell, G.: Quantization Index Modulation: a class of provably good methods for digital watermarking and information embedding. IEEE Transactions on Information Theory **47** (2001) 1423–1443
6. Eggers, J.J., Bäuml, R., Tzschoppe, R., Girod, B.: Scalar Costa Scheme for information embedding. IEEE Transactions on Signal Processing **51** (2003) 1003–1019
7. Cover, T.M., Thomas, J.A.: Elements of Information Theory. Wiley series in Telecommunications (1991)
8. Pérez-Freire, L., Pérez-Gonzalez, F., Voloshinovskiy, S.: Revealing the true achievable rates of Scalar Costa Scheme. In: IEEE Int. Workshop on Multimedia Signal Processing (MMSP), Siena, Italy (2004) 203–206
9. Erez, U., Zamir, R.: Achieving $\frac{1}{2}$log(1+SNR) over the Additive White Gaussian Noise Channel with Lattice Encoding and Decoding. IEEE Transactions on Information Theory **50** (2004) 2293–2314