On Reversibility of Random Binning Based Data-hiding Techniques: Security Perspectives

Sviatoslav Voloshynovskiy*
CVML,CUI - University of Geneva
24, rue General Dufour
1211, Geneva, Switzerland
svolos@cui.unige.ch

Oleksiy Koval
CVML,CUI - University of Geneva
24, rue General Dufour
1211, Geneva, Switzerland
koval@cui.unige.ch

Emre Topak
CVML,CUI - University of Geneva
24, rue General Dufour
1211, Geneva, Switzerland
emre.topak@cui.unige.ch

José Emilio Vila-Forcén
CVML,CUI - University of Geneva
24, rue General Dufour
1211, Geneva, Switzerland
jose.vila@cui.unige.ch

Thierry Pun
CVML,CUI - University of Geneva
24, rue General Dufour
1211, Geneva, Switzerland
pun@cui.unige.ch

ABSTRACT

Reversibility of data-hiding refers to the reconstruction of original host data at the decoder from the stego data. Previous works on the subject are concentrated on the reversibility of data-hiding techniques from multimedia perspectives. However, from the security point of view, that at our knowledge was not exploited in existing studies, reversibility could be used by an attacker to remove the complete trace of watermark data from the stego data in the sense of designing the worst case attack. Thus, the aim of this paper is to analyze the reversibility of data-hiding techniques based on random binning from the security perspectives.

Categories and Subject Descriptors

H.4 [**Information Systems Applications**]: Miscellaneous.

General Terms

performance, theory.

Keywords

reversibility, data-hiding, random binning, exhaustive search, security leakage analysis.

1. INTRODUCTION

Data-hiding refers to the ways of reliable communication of information embedded into a host data for the purposes of copyright protection, content authentication, broadcast monitoring, database indexing, etc.

Initially proposed data-hiding methods consider the host data as a source of interference to the data-hiding communication. Within this class of methods, the information to be embedded, i.e., watermark data, is generated without considering the host data according to *random coding-based techniques* [4]. Later, *random binning-based techniques* were introduced that are inspired by Gel'fand-Pinsker approach [5] for communications in channels where the channel state is non-causally available at the encoder. These methods take the host data into account in the generation of watermark that lead to a substantial increase in the achievable embedding rate in comparison to the random coding based techniques.

Depending on the particular application area, there are specific requirements to the data-hiding method to be applied. In medical and military applications, content authenticity and availability of original host data at the decoder should be provided simultaneously. Moreover, in some cases, the original host data might need to be reproduced from the stego data over the time due to its unavailability or access restrictions. These needs bring discussions about the recovery of the original host data from the stego data, i.e., the reversibility of watermark embedding.

The reversibility of random binning-based data-hiding techniques from multimedia perspectives is considered in [12] for a multiuser communications scenario. It is provided for the scenarios of authorized, i.e. with the knowledge of key, and unauthorized, i.e. without the knowledge of key, users. It is shown that in the noisy communications assumption the unauthorized user is able to estimate the host data using optimal Minimum Mean Square Error (MMSE) estimator with the same estimation accuracy as for the authorized one under optimal selection of a compensation parameter. However, in the noiseless communications assumption, the authorized user can completely recover the host data that is

*(For further information: contact S. Voloshynovskiy, e-mail: svolos@cui.unige.ch, http://sip.unige.ch)

MM&Sec'06, September 26–27, 2006, Geneva, Switzerland.

never the case for the unauthorized user.

Having considered the reversibility from multimedia perspectives and taking into account the fact that it could be used in the development of more involved attacks, this paper handles the same problem from security perspectives.

Notations We use capital letters to denote scalar random variables X, small letters x to denote their realizations. The superscript N is used to designate length-N vectors $\mathbf{x} = x^N = [x[1], x[2], ..., x[N]]^T$ with k^{th} element $x[k]$. We use $X \sim p_X(x)$ or simply $X \sim p(x)$ to indicate that a random variable X is distributed according to $p_X(x)$. The mathematical expectation of a random variable $X \sim p_X(x)$ is denoted by $E_{p_X}[X]$ or simply by $E[X]$ and $Var[X]$ denotes the variance of X. Calligraphic fonts $\mathcal{X}$ denote sets $X \in \mathcal{X}$ and $|\mathcal{X}|$ denotes the cardinality of set $\mathcal{X}$. We use $H(X)$,$h(X)$ and $I(X;Y)$ to denote the entropy of X, differential entropy of X and the mutual information between X and Y respectively. $\mathbf{0}$ denotes a vector of zeros and $\mathbf{I}_N$ denotes the $N \times N$ identity matrix. We also define the watermark-to-image ratio (WIR) as WIR=$10\log_{10}\frac{\sigma_W^2}{\sigma_X^2}$ and the watermark-to-noise ratio (WNR) as WNR $= 10\log_{10}\frac{\sigma_W^2}{\sigma_Z^2}$, where σ_X^2, σ_W^2, σ_Z^2 represent the variances of host data, watermark and noise, respectively.

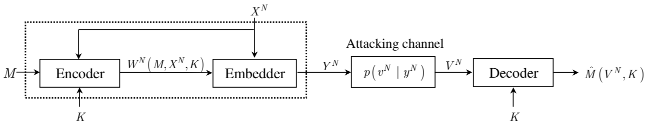


Figure 1: Communication setup for random binning based data-hiding methods

2. INFORMATION THEORETIC MODELING OF DATA-HIDING SECURITY

Let's consider the communication scenario given in Fig.1, where a particular user defined by a $K \in \{1, 2, \ldots, |\mathcal{K}|\}$ is trying to reliably communicate a message $M \in \{1, 2, \ldots, |\mathcal{M}|\}$, $|\mathcal{M}| = 2^{NR}$, within the host data X^N through the attacking channel $p(v^N|y^N)$. This is a generalization of the Gel'fand-Pinsker problem [5], where the channel state X^N is non-causally available at the encoder and used in the generation of watermark data W^N. As illustrated in Fig. 2, the encoder and decoder share the knowledge of keys and corresponding codebooks, which are constituted from unique, typical [4] codewords U^N.

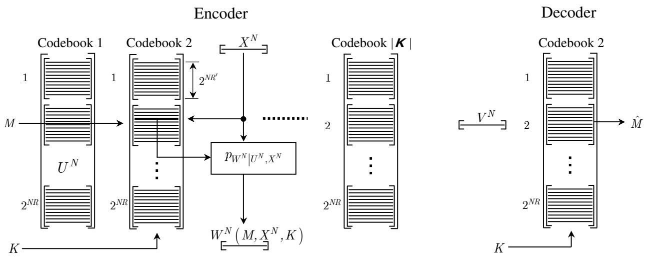


Figure 2: Codebook construction for a generalized Gel'fand-Pinsker based data-hiding scenario

Based on the message M, key K and host data X^N the encoder generates the watermark W^N. It looks in the M^{th} bin of K^{th} codebook for a codeword U^N that is jointly typical [5] with X^N. After finding the jointly-typical (U^N, X^N) pair, the encoder generates the watermark data W^N according to the probabilistic mapping $p(w^N|u^N, x^N)$. Following, the watermark W^N is embedded into the host data X^N according to $p(y^N|w^N, x^N)$, resulting in the stego data Y^N. This operation is subject to a distortion constraint D^E between X^N and Y^N such that:

$$\begin{aligned} d^N(X^N, Y^N) &: \mathcal{X}^N \times \mathcal{Y}^N \to R^+, \\ d^N(X^N, Y^N) &\leq D^E. \end{aligned} \tag{1}$$

Afterwards, the stego data Y^N is sent to the discrete memoryless attack channel (DMC) $p(v^N|y^N) = \prod_{i=1}^N p_{V|Y}(v_i|y_i)$, which is also subject to a distortion constraint D^A between Y^N and V^N as:

$$\begin{aligned} d^N(Y^N, V^N) &: \mathcal{Y}^N \times \mathcal{V}^N \to R^+, \\ d^N(Y^N, V^N) &\leq D^A. \end{aligned} \tag{2}$$

Finally, the decoder makes a decision about what was the sent message $\hat{M}$ based on the received data V^N and key K. It looks in the K^{th} codebook for a codeword U^N that is jointly typical with the received data V^N and declares the index $\hat{M}$ of the bin as the decoded message, where it finds this codeword U^N.

Capacity of this communication setup is given by [5]:

$$R \leq C = \frac{1}{N} \max_{p(u^N, w^N|x^N)} [I(U^N; V^N) - I(U^N; X^N)], \tag{3}$$

where $U^N \in \mathcal{U}^N(K = k)$, with $|\mathcal{U}(K = k)| \leq |\mathcal{X}||\mathcal{W}| + 1$ for all $K \in \{1, 2, \ldots, |\mathcal{K}|\}$. To reflect the technicality behind the codebook and watermark generations as well as channel degradations, we assume that $p(k, x^N, u^N, w^N, y^N, v^N) = p(k)p(x^N)p(u^N|x^N)\underline{\mathbf{1}}\{u^N \in \mathcal{U}^N(K = k)\}p(w^N|u^N, x^N)$ $p(y^N|x^N, w^N)p(v^N|y^N)$. Here, $\underline{\mathbf{1}}\{.\}$ denotes the indicator function.

Costa considered the Gel'fand-Pinsker problem for the Ga-ussian case and mean-square error distance for the distortion constraints [3]. In his formulation, the host data X^N

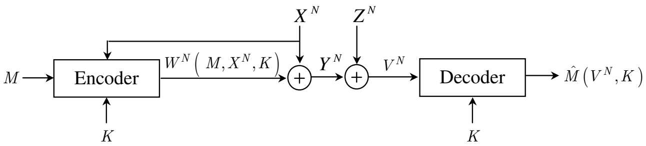


Figure 3: Costa setup for random binning based data-hiding methods

is Gaussian $X^N \sim \mathcal{N}(\mathbf{0}, \sigma_X^2\mathbf{I}_N)$ and the attacking DMC becomes an additive noise Z^N, $Z^N \sim \mathcal{N}(\mathbf{0}, \sigma_Z^2\mathbf{I}_N)$. He defined the following relationship between the codeword U^N, the host data X^N and the watermark data W^N:

$$U^N = W^N + \alpha X^N, \tag{4}$$

with α as the optimization parameter maximizing the achievable rate given in (3):

$$R(\alpha) = \frac{1}{2}\log_2 \frac{\sigma_W^2(\sigma_W^2 + \sigma_X^2 + \sigma_Z^2)}{\sigma_W^2\sigma_X^2(1-\alpha)^2 + \sigma_Z^2(\sigma_W^2 + \alpha^2\sigma_X^2)}. \tag{5}$$

Costa has shown that (5) is maximized when:

$$\alpha = \alpha_{opt} = \frac{\sigma_W^2}{\sigma_W^2 + \sigma_Z^2}, \quad (6)$$

and in this case it is equal to the capacity of an additive white Gaussian channel (AWGN), where there is no host interference at all:

$$R(\alpha_{opt}) = C^{AWGN} = \frac{1}{2}\log_2\left(1 + \frac{\sigma_W^2}{\sigma_Z^2}\right). \quad (7)$$

However, selection of α_{opt} requires a priori knowledge of noise variance σ_Z^2 at the encoder.

Let's replace the fixed DMC $p(v^N|y^N)$, or additive noise Z^N in Costa setup, with an "active attacker", who will be operating before the decoder on the stego data Y^N as indicated in Fig. 4. The overall objective of such an attacker

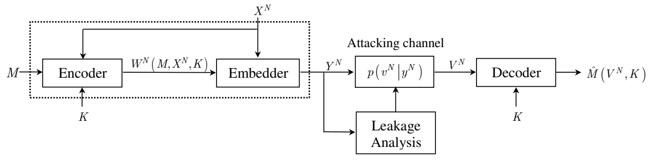


Figure 4: Insertion of an active attacker into the scenario for the security analysis of random binning based data-hiding methods

would be to destroy the reliable communications of the message M. There are different approaches to achieve this objective and they could be classified as follows:

- *Signal processing based attacks:* The idea of this group of attacks is to try to produce the most accurate estimation of the host data X^N from the stego data Y^N using the knowledge of statistical properties of the watermark W^N and host data X^N [11]. They have the disadvantage that this estimate leads to poor results as $\sigma_X^2 \to \infty$.

- *Desynchronization attacks:* This group of attacks targets to destroy the synchronization between the encoder and decoder by applying some geometrical transformation to Y^N from a set of geometrical transformations with large cardinality. A detailed analysis of these attacks and possible countermeasures from the point of designing geometrically robust data-hiding schemes are provided in [10].

- *Sensitivity attacks:* With an assumption of unlimited access to the decoder for the attacker, this group of attacks is based on finding the decision/detection regions of the decoder by observing its response to incremental changes in the amount of attack introduced to Y^N.

Design of a secure data-hiding system has to comply with the Kerckhoffs security principle [6], where it is assumed that the attacker has access to the codebooks and the knowledge of encoding, embedding and decoding algorithms. What remains secret for the attacker is the knowledge of the message M and key K. Moreover, particular codebook designs might lead to security leakages about the codebook construction. Although none of the attacks given above considers these leakages in their design, more involved attacks based on them are provided in [10]. Therefore, the goal of an active attacker would be to develop the worst case attacking strategy that consists in the removal of the watermark data W^N from the stego data Y^N by using all available public information about codebook construction, encoding, embedding and decoding algorithms as well as statistics about the watermark and host data. For this purpose, it is sufficient for the attacker to have the knowledge of the codeword U^N. This argument can be justified in Costa setup, where under such a circumstance he will simply subtract it from the stego data Y^N and will get a scaled version of the host data X^N (when $\alpha \neq 1$):

$$\begin{aligned} U^N &= W^N + \alpha X^N, & (8)\\ Y^N &= W^N + X^N, & (9)\\ Y^N - U^N &= (1-\alpha)X^N. & (10) \end{aligned}$$

Moreover, having access to the codebooks, from the knowledge about the communicated codeword U^N, the attacker can find which particular user K was communicating and what was the message M communicated by checking in which codebook and in which particular message bin this codeword is found. In fact, this is the fundamental difference of the proposed approach from previously considered security perspectives in assessing the security of a scheme in terms of trial efforts to reveal the codeword U^N instead of the message M or key K.

Attacker in the current scenario can follow two different approaches in order to find the communicated codeword U^N: *exhaustive search* (ES) and *security leakage analysis* (SLA) . In the coming sections, these approaches will be investigated in details.

3. EXHAUSTIVE SEARCH APPROACH

In the case of ES, the attacker will look in each codebook for a codeword U^N that is jointly typical with the stego data Y^N. Assuming that in total there exists $|\mathcal{K}|$ codebooks each with 2^{NR} bins and $2^{NR'}$ codewords in each bin, the complexity of this search for the attacker will be upper bounded as:

$$\mathcal{O}_{ES}^A \leq |\mathcal{K}| 2^{N(R+R')}. \quad (11)$$

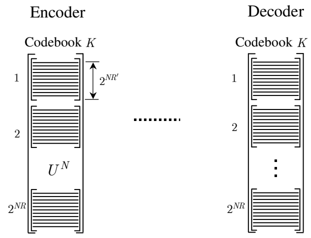
A graphical presentation of ES is given in Fig. 5. The data-hider performs the decoding with the knowledge of key based on V^N with the same codebook used at both encoder and decoder. Contrarily, the attacker observing Y^N performs the search among all possible U^N due to the lack of key knowledge.

In Gel'fand-Pinsker setup, if one assumes $N(R+R') = I(U^N;V^N)$ for the errorless performance of decoder, the upper bound on the complexity in (11) becomes:

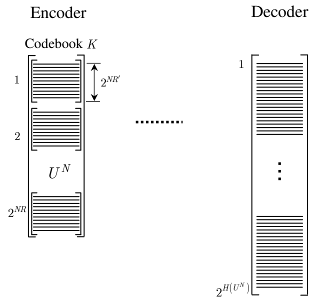
$$\mathcal{O}_{ES}^A \leq |\mathcal{K}| 2^{I(U^N;V^N)}. \quad (12)$$

Next stage in finding the upper bound on $\mathcal{O}_{ES}^A$ would be to calculate the maximum value that $|\mathcal{K}|$ might take. If we consider the generation of all possible typical codewords $U^N \in \mathcal{U}^N$ and distribute them into the codebooks in a non-repetitive manner[1] with $2^{N(R+R')}$ codewords in each, the

[1]Here, we are interested in not in finding a particular pair of (m,k) used, but in the reversibility attack.



(a)



(b)

Figure 5: Decoding structure for: (a) the data-hider and (b) the attacker.

maximum value of $|\mathcal{K}|$ will be:

$$|\mathcal{K}_{max}| \leq \frac{2^{H(U^N)}}{2^{N(R+R')}} = \left\lfloor \frac{2^{H(U^N)}}{2^{I(U^N;V^N)}} \right\rfloor, \tag{13}$$

where $\lfloor\ \rfloor$ denotes a rounding to the nearest smaller integer. In this case, the complexity of the search for the attacker will be:

$$\mathcal{O}^A_{ES} \leq |\mathcal{K}_{max}| 2^{N(R+R')} \leq \left\lfloor 2^{H(U^N)} \right\rfloor. \tag{14}$$

As an example, for the Costa setup taking into account continuous character of random variables, the upper bound on $|\mathcal{K}_{max}|$ is:

$$|\mathcal{K}_{max}| \leq \left\lfloor 2^{h(U^N|V^N)} \right\rfloor$$

$$= \left\lfloor \left(2\pi e \frac{\sigma_W^2 \sigma_X^2 (1-\alpha)^2 + \sigma_Z^2 \left(\sigma_W^2 + \alpha^2 \sigma_X^2\right)}{\sigma_W^2 + \sigma_X^2 + \sigma_Z^2} \right)^{\frac{N}{2}} \right\rfloor, \tag{15}$$

and the upper bound on $\mathcal{O}^A_{ES}$ becomes:

$$\mathcal{O}^A_{ES} \leq \left\lfloor \left(2\pi e(\sigma_W^2 + \alpha^2\sigma_X^2)\right)^{\frac{N}{2}} \right\rfloor. \tag{16}$$

This bound should be interpreted as the volume of ambiguity with respect to the searched codeword U^N. If, for example, $h(X) \to -\infty$, this volume reduces to the unique codeword U^N. Further, it should be understood from the context for which variables this quantity is applied.

In Figs. 6 and 7, the behaviour of $|\mathcal{K}_{max}|$ given by (15) vs. WNR is plotted when $N = 1$ for several values of α and WIRs of -6 dB and -16 dB respectively. Similarly, in Fig. 8, a graph of $\mathcal{O}^A_{ES}$ given in (16) vs. WIR is presented when $N = 1$ for several values of α. The behaviour of $\mathcal{O}^A_{ES}$ for

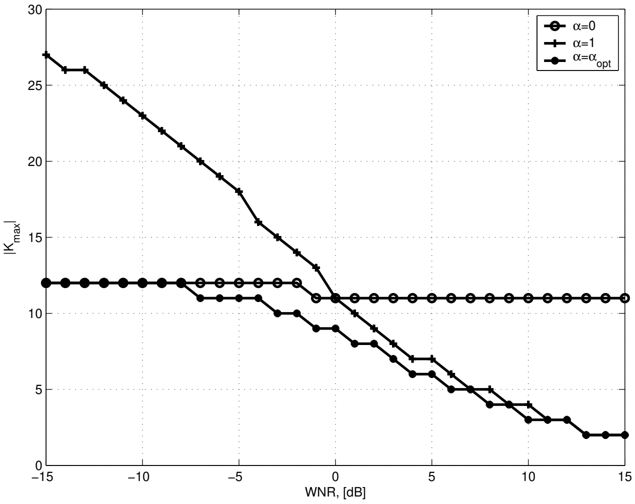


Figure 6: Maximum number of users vs. WNR for WIR=-6 dB.

$\alpha = 0$, as expected from (16), is independent from WIR. As WIR$\to \infty$, $\mathcal{O}^A_{ES}$ for all α values converges to the one for $\alpha = 0$.

From the security point of view for the data-hider, a natural question would be if the attacker can really find the jointly typical $(U^N; Y^N)$ pair after the above number of checks. In this search, the attacker might encounter with some codewords U^N, which were not the one selected at the encoder, that are also jointly typical with the stego data Y^N. The probability of finding such a codeword U^N is $2^{-I(U^N;Y^N)}$[2]. Therefore, the total probability of decoding error for the attacker will be bounded as:

$$\begin{aligned} P_e^{ES} &\leq \left(|\mathcal{K}| 2^{N(R+R')} - 1\right) 2^{-I(U^N;Y^N)}, \\ &< |\mathcal{K}| 2^{N(R+R')} 2^{-I(U^N;Y^N)}, \end{aligned} \tag{17}$$

where $U^N \in \mathcal{U}^N$. Injecting the value of $|\mathcal{K}_{max}|$ from (13) into (17), one obtains for the Costa setup:

$$\begin{aligned} P_e^{ES} &\leq \frac{2^{h(U^N)}}{2^{N(R+R')}} 2^{N(R+R')} 2^{-I(U^N;Y^N)} = 2^{h(U^N|Y^N)}, \\ &= \left(2\pi e \frac{\sigma_W^2 \sigma_X^2 (1-\alpha)^2}{\sigma_W^2 + \sigma_X^2}\right)^{\frac{N}{2}}. \end{aligned} \tag{18}$$

For particular values of α parameter of Costa, (18) can be rewritten as follows:

- $\underline{\alpha = 0}$

$$P_e^{ES} \leq \left(2\pi e \frac{\sigma_W^2 \sigma_X^2}{\sigma_W^2 + \sigma_X^2}\right)^{\frac{N}{2}}, \tag{19}$$

[2] We do not consider here a case when the found $\hat{U}^N$ is close to the true U^N in some sense that can be applied for the reversibility similar to remodulation attack ([13], [8]) and sensitivity attack ([7], [1], [2]). We are interested in to find a probability to have the exact U^N for the complete removal of watermark from Y^N.

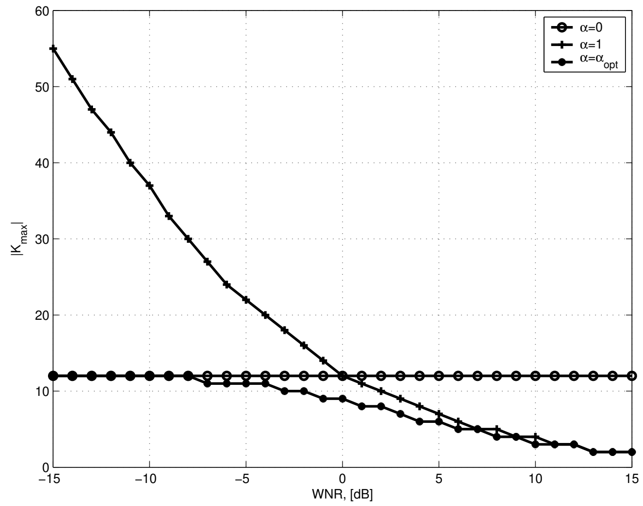


Figure 7: Maximum number of users vs. WNR for WIR=-16 dB.

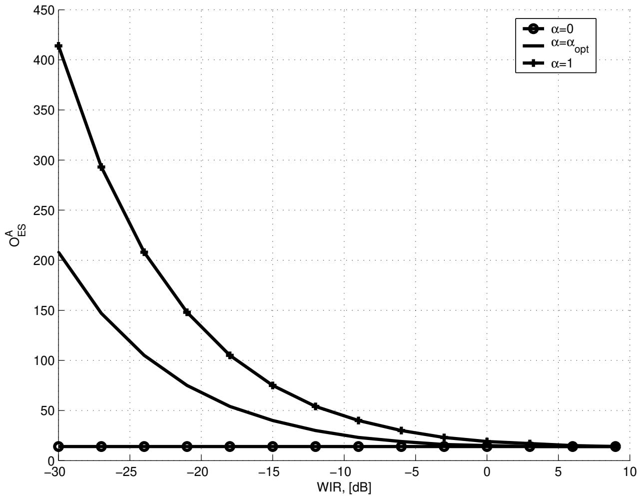


Figure 8: The complexity of exhaustive search for the attacker vs. WIR.

- $\underline{\alpha = \alpha_{opt}}$

$$P_e^{ES} \leq \left(2\pi e \frac{\sigma_W^2 \sigma_X^2 (\sigma_Z^2)^2}{(\sigma_W^2+\sigma_X^2)(\sigma_W^2+\sigma_Z^2)^2}\right)^{\frac{N}{2}}, \tag{20}$$

- $\underline{\alpha = 1}$

$$P_e^{ES} \leq \left(2\pi e \frac{\sigma_W^2 \sigma_X^2 (1-1)^2}{\sigma_W^2+\sigma_X^2}\right)^{\frac{N}{2}} = 0. \tag{21}$$

Therefore, when $\alpha = 1$, the probability of decoding error is 0 for the attacker who is performing an exhaustive search. However, this result is not surprising due to the fact that $Y^N = U^N$ and $h(U^N|Y^N) = -\infty$ when $\alpha = 1$, i.e., U^N is directly given by observing Y^N.

A possible defense strategy of the data-hider would be to make the upper bound on the probability of decoding error for the attacker non-zero as $N \to \infty$:

$$\begin{aligned}
&|\mathcal{K}| 2^{N(R+R')} 2^{-I(U^N;Y^N)} > 0,\\
&2^{\log|\mathcal{K}|} 2^{N(R+R')} 2^{-I(U^N;Y^N)} > 0,\\
&2^{N(\frac{1}{N}\log|\mathcal{K}| + R + R' - \frac{1}{N} I(U^N;Y^N))} > 0,\\
&\frac{1}{N}\log|\mathcal{K}| + R + R' - \frac{1}{N} I(U^N;Y^N) \geq 0.
\end{aligned} \tag{22}$$

Putting the assumption $N(R+R') = I(U^N;V^N)$ from Gel'fand Pinsker setup into (22), the inequality becomes:

$$\begin{aligned}
\frac{1}{N}\log|\mathcal{K}| &\geq \frac{1}{N}\left(I(U^N;Y^N) - I(U^N;V^N)\right),\\
\log|\mathcal{K}| &\geq I(U^N;Y^N) - I(U^N;V^N),\\
|\mathcal{K}| &\geq 2^{\left(I(U^N;Y^N) - I(U^N;V^N)\right)}.
\end{aligned} \tag{23}$$

The result in (23) is the minimum number of codebooks that the data-hider should construct so that the attacker following an exhaustive search scenario will have a non-zero probability of decoding error. It can be seen that this value is always smaller than the maximum value that $|\mathcal{K}|$ can take, i.e., $|\mathcal{K}_{max}|$, given by (13).

In the Costa Setup, when the values of $I(U^N;Y^N)$ and $I(U^N;V^N)$ for the Gaussian data are substituted into (23), one has:

$$\begin{aligned}
|\mathcal{K}| &\geq 2^{\left(I(U^N;Y^N) - I(U^N;V^N)\right)} = \\
&= \left(\frac{(\sigma_W^2+\alpha^2\sigma_X^2)(\sigma_W^2+\sigma_X^2)}{\sigma_W^2\sigma_X^2(1-\alpha)^2}\right)^{\frac{N}{2}} \times \\
&\times \left(\frac{(\sigma_W^2+\sigma_X^2+\sigma_Z^2)(\sigma_W^2+\alpha^2\sigma_X^2)}{\sigma_W^2\sigma_X^2(1-\alpha)^2+\sigma_Z^2(\sigma_W^2+\alpha^2\sigma_X^2)}\right)^{-\frac{N}{2}}, \\
&= \left(\frac{(\sigma_W^2+\sigma_X^2)}{\sigma_W^2\sigma_X^2(1-\alpha)^2}\right)^{\frac{N}{2}} \times \\
&\times \left(\frac{\sigma_W^2\sigma_X^2(1-\alpha)^2+\sigma_Z^2(\sigma_W^2+\alpha^2\sigma_X^2)}{(\sigma_W^2+\sigma_X^2+\sigma_Z^2)}\right)^{\frac{N}{2}}.
\end{aligned} \tag{24}$$

For particular values of α, (24) takes the following values:

- $\underline{\text{For } \alpha = 0}$

$$\begin{aligned}
|\mathcal{K}| &\geq \left(\frac{(1+\frac{\sigma_X^2}{\sigma_W^2})}{\frac{\sigma_X^2}{\sigma_W^2}} \frac{\frac{\sigma_X^2}{\sigma_W^2}+\frac{\sigma_Z^2}{\sigma_W^2}}{(1+\frac{\sigma_X^2}{\sigma_W^2}+\frac{\sigma_Z^2}{\sigma_W^2})}\right)^{\frac{N}{2}}, \\
&= \left\lceil \left(\frac{10^{0.2WIR}}{10^{0.1WIR}10^{0.1WNR}+10^{0.1WIR}+10^{0.1WNR}} + \right.\right. \\
&\qquad \left.\left. +1\right)^{\frac{N}{2}} \right\rceil.
\end{aligned} \tag{25}$$

In Fig. 9, a plot of lower bound on $|\mathcal{K}|$ in (25) vs. WIR is given when $N = 1$ for several values of WNR.

- $\underline{\text{For } \alpha = \alpha_{opt}}$

$$\begin{aligned}
|\mathcal{K}| &\geq \left(\frac{(\sigma_W^2+\sigma_X^2)(\sigma_Z^2+\sigma_W^2)}{\sigma_X^2\sigma_Z^2}\right)^{\frac{N}{2}}, \\
&= \left\lceil (1+10^{0.1WIR})(1+10^{0.1WNR}) \right\rceil^{\frac{N}{2}}.
\end{aligned} \tag{26}$$

In Fig. 10, a plot of lower bound on $|\mathcal{K}|$ in (26) vs. WIR is given when $N = 1$ for several values of WNR.

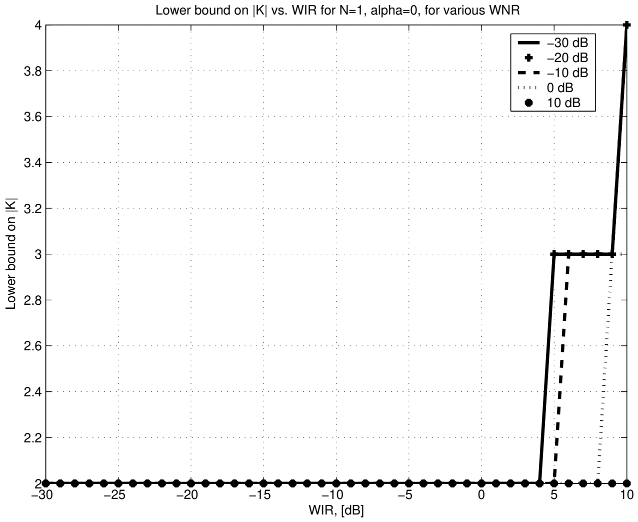


Figure 9: The lower bound on $|\mathcal{K}|$ vs. WIR for $\alpha = 0$.

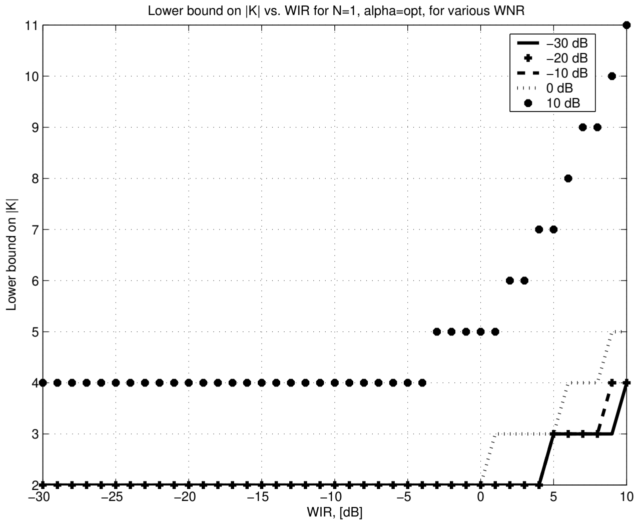


Figure 10: The lower bound on $|\mathcal{K}|$ vs. WIR for $\alpha = \alpha_{opt}$.

- For $\alpha = 1$

$$|\mathcal{K}| \geq \left(\frac{(\sigma_W^2 + \sigma_X^2)}{0} \frac{\sigma_Z^2(\sigma_W^2 + \sigma_X^2)}{(\sigma_W^2 + \sigma_X^2 + \sigma_Z^2)} \right)^{\frac{N}{2}} = \infty. \quad (27)$$

Thus, the data-hider has to generate infinitely many codebooks in order to make the probability of decoding error non-zero for the attacker, which is not possible at all. This result could be justified by the fact that for the Costa Setup, when $\alpha = 1$, $W^N = U^N - X^N$ and $Y^N = X^N + W^N = U^N$, i.e., all information about the codeword U^N is immediately available to the attacker after the observation of stego data Y^N. However, according to (10), when $\alpha = 1$, the knowledge of U^N does not give the attacker any possibility of reversibility.

Consequently, a proper selection of total number of codebooks from the data-hider side will make the attacker's exhaustive search efforts useless by leading to a non-zero probability of decoding error for the attacker. In the case when $\alpha = 1$, reversibility is avoided by definition according to (10).

4. SECURITY LEAKAGE ANALYSIS

In the SLA, the attacker tries to estimate U^N based on his observation of the stego data Y^N. The uncertainty about U^N after observing Y^N will be reduced from $h(U^N)$ to $h(U^N|Y^N)$, where $h(U^N|Y^N)$ depends on the variance of the estimation accuracy $\sigma^2_{U|Y}$. Therefore, the exhaustive search in the previous section will convert to a search for a jointly-typical codeword U^N with the stego data Y^N among a set of codewords collected from all codebooks. For the Gaussian data and large N, these codewords would constitute a sphere around $\hat{U}^N = E\left[U^N|Y^N\right]$ with radius $\sqrt{N\sigma^2_{U|Y}}$ as shown in Fig. 11 and the volume of this set would be at most $2^{h(U^N|Y^N)}$. A practical implementation of the SLA is given in Fig. 12.

Therefore, disregarding the complexity of collecting these codewords from all codebooks, the complexity of security leakage analysis for the data-hider will be upper bounded

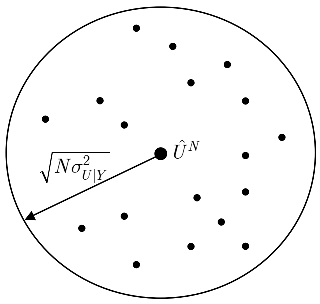


Figure 11: Set of codewords considered for SLA.

by:

$$\mathcal{O}^A_{SLA} \leq \left\lfloor 2^{h(U^N|Y^N)} \right\rfloor \leq \mathcal{O}^A_{ES} \leq \left\lfloor 2^{h(U^N)} \right\rfloor. \quad (28)$$

The upper bounds on the complexity for the attacker in the case of exhaustive search and in the case of security leakage analysis become equal only if $I(U^N; Y^N) = 0$, which corresponds to the Shannon's condition of perfect secrecy [9].

In Fig. 13, the upper bound on the complexity of security leakage analysis for the attacker vs. WIR is plotted for Costa setup by calculating the corresponding value of $h(U^N|Y^N)$:

$$\begin{aligned}\mathcal{O}^A_{SLA} &\leq \left\lfloor 2^{h(U^N|Y^N)} \right\rfloor \\ &= \left\lfloor \left(2\pi e \frac{\sigma_W^2 \sigma_X^2 (1-\alpha)^2}{\sigma_W^2 + \sigma_X^2} \right)^{\frac{N}{2}} \right\rfloor, \quad (29)\end{aligned}$$

for particular α values when $N = 1$. Compared to Fig. 8, one can observe the inverse behavior of the upper bound on $\mathcal{O}^A_{SLA}$ with respect to that on $\mathcal{O}^A_{ES}$. The largest value for the upper bound on $\mathcal{O}^A_{SLA}$ is obtained when $\alpha = 0$. Actually, in this case $U^N = W^N$, so, the estimation of codeword U^N corresponds to the estimation of watermark data W^N. There is an interesting situation when $\alpha = 1$ since $h(U^N|Y^N) = -\infty$ due to the fact that $W^N = U^N - X^N$ and $Y^N = X^N + W^N = U^N$. Therefore, the codeword U^N is im-

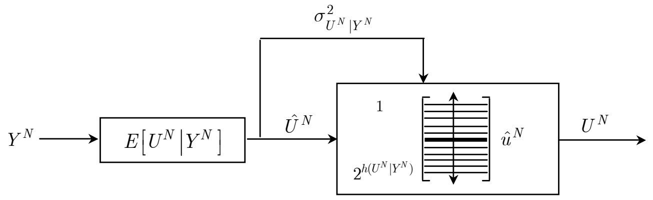


Figure 12: A practical SLA system: MMSE Estimator with the search in the volume of the closest codewords.

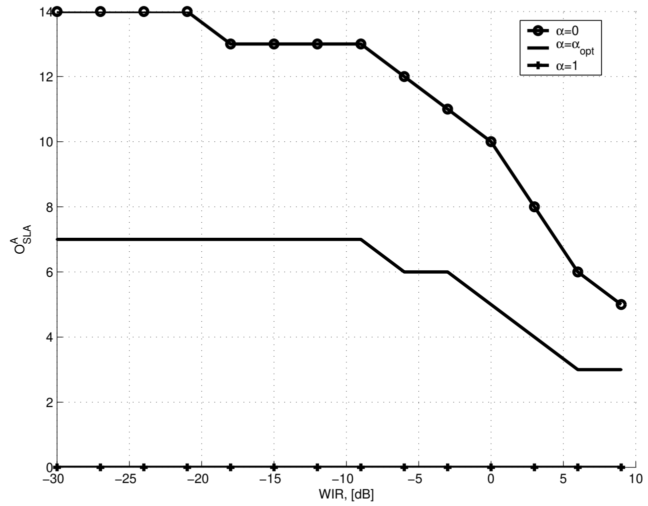


Figure 13: Upper bound on the attacker complexity vs. WIR in the case of SLA.

mediately available to the attacker upon his observation of the stego data Y^N. However, this knowledge does not grant reversibility for the attacker when $\alpha = 1$ according to (10). Finally, for $\alpha = \alpha_{opt}$, the upper bound on $\mathcal{O}^A_{SLA}$ demonstrates a behaviour between the upper bounds on $\mathcal{O}^A_{SLA}$ for the cases $\alpha = 0$ and $\alpha = 1$. Another observation after comparing the plots of upper bounds on $\mathcal{O}^A_{SLA}$ and $\mathcal{O}^A_{ES}$ is that $\mathcal{O}^A_{SLA} \le \mathcal{O}^A_{ES}$. The equality $\mathcal{O}^A_{SLA} = \mathcal{O}^A_{ES}$ is realizable only for $\alpha = 0$ as WIR$\rightarrow -\infty$ since $I(U^N;Y^N) \rightarrow 0$ in this case. In such a situation, the attacker cannot benefit from the security leakages in reducing the complexity of search for the codeword U^N that will constitute a jointly typical (U^N, Y^N) pair with the stego data Y^N.

Having analyzed the complexity of security leakage analysis for the attacker, next question would be if the attacker can succeed with this search in finding the codeword U^N, which was the one sent by the encoder. Since a random codeword U^N and the stego data Y^N is jointly-typical with a probability of $2^{-I(U^N;Y^N)}$, one can bound the probability of decoding error for the attacker by:

$$\begin{aligned} P_e^{SLA} &\le \left(2^{h(U^N|Y^N)} - 1\right) 2^{-I(U^N;Y^N)}, \\ &\le 2^{h(U^N|Y^N)} 2^{-I(U^N;Y^N)}. \end{aligned} \tag{30}$$

Comparing (30) with the P_e^{ES} defined by (18), one can conclude that $P_e^{SLA} \le P_e^{ES}$ since $h(U^N|Y^N) \le h(U^N)$.

For Costa setup, putting the values of $h(U^N|Y^N)$ and $I(U^N;Y^N)$ into (30), the upper bound on the probability of decoding error for the attacker becomes:

$$P_e^{SLA} \le \left(2\pi e \frac{(\sigma_W^2 \sigma_X^2 (1-\alpha)^2)^2}{(\sigma_W^2 + \alpha^2 \sigma_X^2)(\sigma_W^2 + \sigma_X^2)^2}\right)^{\frac{N}{2}}. \tag{31}$$

The upper bound on P_e^{SLA} takes the following particular values for:

- $\underline{\alpha = 0}$

$$P_e^{SLA} \le \left(2\pi e \frac{\sigma_W^2 (\sigma_X^2)^2}{(\sigma_W^2 + \sigma_X^2)^2}\right)^{\frac{N}{2}}, \tag{32}$$

- $\underline{\alpha = \alpha_{opt}}$

$$\begin{aligned} P_e^{SLA} \le &\left(2\pi e \frac{\sigma_W^2 (\sigma_X^2 \sigma_Z^4)^2}{((\sigma_W^2 + \sigma_Z^2)^2 + \sigma_W^2 \sigma_X^2)}\right)^{\frac{N}{2}} \times \\ &\times \left(\frac{1}{(\sigma_W^2 + \sigma_X^2)^2 (\sigma_W^2 + \sigma_Z^2)^2}\right)^{\frac{N}{2}}, \end{aligned} \tag{33}$$

- $\underline{\alpha = 1}$

$$P_e^{SLA} \le \left(2\pi e \frac{(\sigma_W^2 \sigma_X^2 (1-1)^2)^2}{(\sigma_W^2 + 1^2 \sigma_X^2)(\sigma_W^2 + \sigma_X^2)^2}\right)^{\frac{N}{2}} = 0. \tag{34}$$

Therefore, when $\alpha = 1$, the probability of decoding error is 0 for the attacker who is following a security leakage analysis strategy. As in the case of exhaustive search, this result is obvious due to the fact that $Y^N = U^N$ for $\alpha = 1$.

5. REVERSIBILITY ATTACK

Attacker, who could get the knowledge of U^N after applying ES or SLA attack could further perform reversibility. For this purpose, one would make an estimate of X^N based on the information about U^N and observation of Y^N as [12]:

$$\begin{aligned} \hat{X}^N &= E\left[X^N | Y^N, U^N\right], \\ &= \frac{1}{1-\alpha}\left(Y^N - U^N\right), \\ &= \frac{1}{1-\alpha}\left(X^N + W^N - \alpha X^N - W^N\right), \\ &= X^N, \end{aligned} \tag{35}$$

with the distortion [12]:

$$\begin{aligned} D^r(\alpha) &= \frac{\sigma_W^2 \sigma_X^2 \sigma_Z^2}{\alpha \sigma_X^2 \sigma_Z^2 + \sigma_W^2 (\sigma_X^2 (1-\alpha)^2 + \sigma_Z^2)}, \\ &= \frac{\sigma_W^2 \sigma_X^2 0}{\alpha \sigma_X^2 0 + \sigma_W^2 (\sigma_X^2 (1-\alpha)^2 + 0)}, \\ &= 0, \end{aligned} \tag{36}$$

since the attacker is working on the noise-free stego data Y^N, i.e., $\sigma_Z^2 = 0$. Therefore, depending on the availability of knowledge about α, perfect reversibility of the Costa setup is possible for the attacker. However, when $\alpha = 1$, although $U^N = Y^N$ without any requirement to perform an ES or SLA attack, reversibility is not possible at all according to (35).

As a countermeasure to the reversibility from the attacker side, the data-hider could consider the following defense strategies:

- To increase the complexity in finding U^N: In the case of ES (11), it is possible for the data-hider to increase the complexity of attacker search for U^N by generating as many codebooks as possible for the communications protocol, where $\mathcal{K}_{max}$ is the maximum (13).

- To increase the probability of error in finding U^N: In the case of ES, it is possible to make the upper bound on the probability of error non-zero for the attacker in finding U^N provided that the number of codebooks satisfies the condition in (23). A generalization of this condition for Costa setup is given in (24), where one can observe that for $\alpha = 1$, it is not possible to make the upper bound non-zero since the number of codebooks should be infinite. In the case of SLA, the upper bound would be non-zero according to (30), unless $\alpha = 1$ (34).

- To avoid reversibility: In the case of $\alpha = 1$, one cannot achieve perfect reversibility according to (35) and (36). Therefore, relaxing two previous strategies and even given directly U^N as $Y^N = U^N$, one can prevent the reversibility attack. However, in this case, some other attacks might be used.

6. CONCLUSIONS

In this paper, reversibility of data-hiding techniques based on random binning is investigated from the security point of view. It is shown that an attacker with privileges given by the Kerckhoffs security principle can introduce more involved attacks in the sense of designing the worst case attack. These attacks can be based on either exhaustive search or security leakage analysis approaches. Each approach is analyzed in terms of the complexity and the probability of decoding error for the attacker. Moreover, from the side of data-hider, it is shown that with a proper selection of total number of codebooks, the exhaustive search efforts of the attacker turn out to be hopeless. Similarly, security leakage analysis should be considered further both from the same perspective and from the design of a more involved attack. Furthermore, when $\alpha = 1$, reversibility is not possible for the attacker in both cases due to system design particularities given by (10).

7. ACKNOWLEDGMENTS

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, FP6-507609-SIMILAR and Swiss IM2 projects.

8. REFERENCES

[1] M. E. Choubassi and P. Moulin. New sensitivity analysis attack. In E. J. Delp and P. W. Wong, editors, *Security, Steganography, and Watermarking of Multimedia Contents*, volume 5681 of *Proceedings of SPIE*, pages 734–745. SPIE, 2005.

[2] P. Comesaña, L. Pérez-Freire, and F. Pérez-González. The return of the sensitivity attack. In M. Barni, I. J. Cox, T. Kalker, and H. J. Kim, editors, *IWDW*, volume 3710 of *Lecture Notes in Computer Science*, pages 260–274. Springer, 2005.

[3] M. Costa. Writing on dirty paper. *IEEE Trans. on Information Theory*, 29(3):439–441, May 1983.

[4] T. Cover and J. Thomas. *Elements of Information Theory*. Wiley and Sons, New York, 1991.

[5] S. Gel'fand and M. Pinsker. Coding for channel with random parameters. *Probl. Control and Inf. Theory*, 9(1):19–31, 1980.

[6] A. Kerckhoffs. La cryptographie militaire. *Journal des sciences militaires*, 9:5–38, 1883.

[7] J.-P. M. G. Linnartz and M. van Dijk. Analysis of the sensitivity attack against electronic watermarks in images. In D. Aucsmith, editor, *Information Hiding*, volume 1525 of *Lecture Notes in Computer Science*, pages 258–272. Springer, 1998.

[8] M. K. Mihcak, R. Venkatesan, and M. Kesal. Cryptanalysis of discrete-sequence spread spectrum watermarks. In *Proceedings of the 5th International Information Hiding Workshop (IH 2002)*, Noordwijkerhout, The Netherlands, October 2002.

[9] C. E. Shannon. Communication theory of secrecy systems. *Bell System Technical Journal*, 28:656–715, October 1949.

[10] E. Topak, S. Voloshynovskiy, O. Koval, M. Mihcak, and T. Pun. Towards geometrically robust data-hiding with structured codebooks. *ACM Multimedia Systems Journal, Special Issue on Multimedia and Security*, 11(2), 2005.

[11] S. Voloshynovskiy, A. Herrigel, N. Baumgaertner, and T. Pun. A stochastic approach to content adaptive digital image watermarking. In *Third International Workshop on Information Hiding*, volume 1768, pages 212–236, Dreyden, Germany, September 29 - Ocaober 1st 1999.

[12] S. Voloshynovskiy, O. Koval, E. Topak, J. Vila-Forcén, P. Comesana, and T. Pun. On reversibility of random binning techniques: multimedia perspectives. In *9th IFIP TC-6 TC-11 Conference on Communications and Multimedia Security (CMS 2005)*, Salzburg, Austria, September 2005.

[13] S. Voloshynovskiy, S. Pereira, A. Herrigel, N. Baumgärtner, and T. Pun. A generalized watermark attack based on stochastic watermark estimation and perceptual remodulation. In *IS&T/SPIE's 12th Annual Symposium, Electronic Imaging 2000: Security and Watermarking of Multimedia Conteht II*, volume 3951, pages 358–370, San Jose, California USA, 23–28 January 0000.