# On Joint Coding for Watermarking and Encryption

Neri Merhav

Department of Electrical Engineering
Technion - Israel Institute of Technology
Haifa 32000, ISRAEL
merhav@ee.technion.ac.il

**Abstract**

In continuation to earlier works where the problem of joint information embedding and lossless compression (of the composite signal) was studied in the absence [9] and in the presence [10] of attacks, here we consider the additional ingredient of protecting the secrecy of the watermark against an unauthorized party, which has no access to a secret key shared by the legitimate parties. In other words, we study the problem of joint coding for three objectives: information embedding, compression, and encryption. Our main result is a coding theorem that provides a single–letter characterization of the best achievable tradeoffs among the following parameters: the distortion between the composite signal and the covertext, the distortion in reconstructing the watermark by the legitimate receiver, the compressibility of the composite signal (with and without the key), and the equivocation of the watermark, as well as its reconstructed version, given the composite signal. In the attack–free case, if the key is independent of the covertext, this coding theorem gives rise to a *threefold* separation principle that tells that asymptotically, for long block codes, no optimality is lost by first applying a rate–distortion code to the watermark source, then encrypting the compressed codeword, and finally, embedding it into the covertext using the embedding scheme of [9]. In the more general case, however, this separation principle is no longer valid, as the key plays an additional role of side information used by the embedding unit.

**Index Terms:** Information hiding, watermarking, encryption, data compression, separation principle, side information, equivocation, rate–distortion.

# 1 Introduction

It is common to say that encryption and watermarking (or information hiding) are related but they are substantially different in the sense that in the former, the goal is to protect the secrecy of the *contents* of information, whereas in the latter, it is the very *existence* of this information that is to be kept secret.

In the last few years, however, we are witnessing increasing efforts around the *combination* of encryption and watermarking, which is motivated by the desire to further enhance the security of sensitive information that is being hidden in the host signal. This is to guarantee that even if the watermark is somehow detected by a hostile party, its contents still remain secure due to the encryption. This combination of watermarking and encryption can be seen both in recently reported research work (see, e.g., [1],[2],[6],[8],[14],[16] and references therein) and in actual technologies used in commercial products with a copyright protection framework, such as the CD and the DVD. Also, some commercial companies that provide Internet documents, have in their websites links to copyright warning messages, saying that their data are protected by digitally encrypted watermarks (see, e.g., `http://genealogy.lv/1864Lancaster/copyright.htm`).

This paper is devoted to the information–theoretic aspects of joint watermarking and encryption together with lossless compression of the composite signal that contains the encrypted watermark. Specifically, we extend the framework studied in [9] and [10] of joint watermarking and compression, so as to include encryption using a secret key. Before we describe the setting of this paper concretely, we pause then to give some more detailed background on the work reported in [9] and [10].

In [9], the following problem was studied: Given a covertext source vector $X^n = (X_1, \ldots, X_n)$, generated by a discrete memoryless source (DMS), and a message $m$, uniformly distributed in $\{1, 2, \ldots, 2^{nR_e}\}$, independently of $X^n$, with $R_e$ designating the embedding rate, we wish to generate a composite (stegotext) vector $Y^n = (Y_1, \ldots, Y_n)$ that satisfies the following requirements: (i) Similarity to the covertext (for reasons of maintaining quality), in the sense that a distortion constraint, $Ed(X^n, Y^n) = \sum_{t=1}^{n} Ed(X_t, Y_t) \leq nD$, holds, (ii) compressibility (for reasons of saving storage space and bandwidth), in the sense that the normalized entropy, $H(Y^n)/n$, does not exceed some threshold $R_c$, and (iii) reliability in decoding the message $m$ from $Y^n$, in the sense that the decoding error probability is ar-

bitrarily small for large $n$. A single–letter characterization of the best achievable tradeoffs among $R_c$, $R_e$, and $D$ was given in [9], and was shown to be achievable by an extension of the ordinary lossy source coding theorem, giving rise to the existence of $2^{nR_e}$ *disjoint* rate–distortion codebooks (one per each possible watermark message) as long as $R_e$ does not exceed a certain fundamental limit. In [10], this setup was extended to include a given memoryless attack channel, $P(Z^n|Y^n)$, where item (iii) above was redefined such that the decoding was based on $Z^n$ rather than on $Y^n$, and where, in view of requirement (ii), it is understood that the attacker has access to the compressed version of $Y^n$, and so, the attacker decompresses $Y^n$ before the attack and re–compresses it after. This extension from [8] to [9] involved an different approach, which was in the spirit of the Gel'fand–Pinsker coding theorem for a channel with non–causal side information (SI) at the transmitter [5]. The role of SI, in this case, was played by the covertext.

In this paper, we extend the settings of [9] and [10] to include encryption. For the sake of clarity of the exposition, we do that in several steps.

In the first step, we extend the attack–free setting of [9]: In addition to including encryption, we also extend the model of the watermark message source to be an arbitrary DMS, $U_1, U_2, \ldots$, independent of the covertext, and not necessarily a binary symmetric source (BSS) as in [9] and [10]. Specifically, we now assume that the encoder has three inputs (see Fig. 1): The covertext source vector, $X^n$, an independent (watermark) message source vector $U^N = (U_1, \ldots, U_N)$, where $N$ may differ from $n$ if the two sources operate in different rates, and a secret key (shared also with the legitimate decoder) $K^n = (K_1, \ldots, K_n)$, which, for mathematical convenience, is assumed to operate at the same rate as the covertext. It is assumed, at this stage, that $K^n$ is independent of $U^N$ and $X^n$. Now, in addition to requirements (i)-(iii), we impose a requirement on the equivocation of the message source relative to an eavesdropper that has access to $Y^n$, but not to $K^n$. Specifically, we would like the normalized conditional entropy, $H(U^N|Y^n)/N$, to exceed a prescribed threshold, $h$ (e.g., $h = H(U)$ for perfect secrecy). Our first result is a coding theorem that gives a set of necessary and sufficient conditions, in terms of single–letter inequalities, such that a triple $(D, R_c, h)$ is achievable, while maintaining reliable reconstruction of $U^N$ at the legitimate receiver.

A few words are now in order about the secrecy metric $H(U^N|Y^n)/N$, whose evident weakness (even when $h = H(U)$) is that it does not rule out sublinear learning rates at

the eavesdropper's side. Notwithstanding this weakness, this secrecy metric has been used in many other information–theoretic works on cryptography (see, e.g., [18],[19] and many others). Perhaps a more natural criterion for security could be the distortion associated with the best estimate that an eavesdropper can get from the cryptogram. Yamamoto [19] has made an attempt to analyze such a criterion, but at the price of a gap between the upper and lower bounds on achievable performance. One obvious fact is that an equivocation level $h$ guarantees that this distortion will be lower bounded by $D(H(U)-h)$, where $D(\cdot)$ is the distortion–rate function of the source $\{U_i\}$. So, equivocation and distortion are related in the sense that a certain level of $h$ guarantees a desirable distortion level. Another alternative is the stronger notion of secrecy due to Maurer. However, it is considerably more difficult to work with.

Returning to the present work, in the second step, we relax the requirement of perfect (or almost perfect) watermark reconstruction, and assume that we are willing to tolerate a certain distortion between the watermark message $U^N$ and its reconstructed version $\hat{U}^N$, that is, $Ed'(U^N,\hat{U}^N)=\sum_{i=1}^N Ed'(U_i,\hat{U}_i)\le ND'$. For example, if $d'$ is the Hamming distortion measure then $D'$, of course, designates the maximum allowable bit error probability (as opposed to the block error probability requirement of [9] and [10]). Also, in this case, it makes sense to impose a requirement regarding the equivocation of the *reconstructed* message, $\hat{U}^N$, namely, $H(\hat{U}^N|Y^n)/N\ge h'$, for some prescribed constant $h'$. The rationale is that it is $\hat{U}^N$, not $U^N$, that is actually conveyed to the legitimate receiver, and hence there is an incentive to protect the secrecy of $\hat{U}^N$. We will take into account both equivocation requirements, with the understanding that if one of them is superfluous, then the corresponding threshold ($h$ or $h'$ accordingly) can always be set to zero. Our second result then extends the above–mentioned coding theorem to a single–letter characterization of achievable quintuples $(D,D',R_c,h,h')$. As will be seen, this coding theorem gives rise to a threefold separation theorem, that separates, without asymptotic loss of optimality, between three stages: rate–distortion coding of $U^N$, encryption of the compressed bitstream, and finally, embedding the resulting encrypted version using the embedding scheme of [9]. The necessary and sufficient conditions related to the encryption are completely decoupled from those of the embedding and the stegotext compression.

In the third and last step, we drop the assumption of an attack–free system and we assume a given memoryless attack channel, in analogy to [10]. Again, referring to Fig. 1, it

should be understood that the stegotext $Y^n$ is stored (or transmitted) in compressed form, and that the attacker decompresses $Y^n$ before the attack and re–compresses after (the compression and decompression units are omitted from the figure). As it will turn out, in the case of a memoryless attack, there is an interaction between the encryption and the embedding, even if the key is still assumed independent of the covertext. In particular, it will be interesting to see that the key, in addition to its original role in encryption, serves also as SI that is available to both encoder and decoder (see Fig. 2).[1] Also, because of the dependence between the key and the composite signal, and the fact that the content provider (at the encoder side) may wish to store the compressed composite signal at its own end, it is reasonable to let the compressibility constraint correspond also to the conditional entropy of $Y^n$ given $K^n$, that is, *private* compression as opposed to the previously considered *public* compression, without the key, which enables decompression but not decryption (when these two operations are carried out by different, remote units). Accordingly, we will consider both the conditional and the unconditional entropies of $Y^n$, i.e., $H(Y^n)/n \leq R_c$ and $H(Y^n|K^n)/n \leq R_c'$. Our final result then is a coding theorem that provides a single–letter characterization of the region of achievable six–tuples $(D, D', R_c, R_c', h, h')$.

Interestingly, this characterization remains essentially unaltered even if there dependency between $K^n$ and $X^n$ is introduced.[2] In this context, the system designer confronts an interesting dilemma regarding the desirable degree of statistical dependence between $K^n$ and $X^n$, which affects the dependence between $K^n$ and $Y^n$. On the one hand, strong dependence can reduce the entropy of $Y^n$ given $K^n$ (and thereby reduce $R_c'$), and can also help in the embedding process: For example, the extreme case of $K^n = X^n$ (which corresponds to *private* watermarking since the decoder actually has access to the covertext) is particularly interesting because in this case, for the $K^n$, there is no need for any external resources of randomness, in addition to the randomness of $X^n$ that is already available. On the other hand, when there is strong dependence between $K^n$ and $Y^n$, the secrecy of the watermark might be sacrificed since $H(K^n|Y^n)$ decreases as well. An interesting point, in this context, is that the Slepian–Wolf encoder [15] (see Fig. 2) is used to generate, from $K^n$, random bits that are essentially independent of $Y^n$ (as $Y^n$ is generated only after the encryption). All

[1] This idea of the double role of the secret key has been explored also in [7] in the context of compression of encrypted data.

[2] In fact, the choice of the conditional distribution $P(K^n|X^n)$ is a degree of freedom that can be optimized subject to the given randomness resources.

these aspects will be seen in detail in Section 4, and even more so, in Section 6.

The remaining parts of this paper are organized as follows: In Section 2, we set some notation conventions. Section 3 will be devoted to a formal problem description and to the presentation of the main result for the attack–free case with distortion–free watermark reconstruction (first step described above). In Section 4, the setup and the results will be extended along the lines of the second and the third steps, detailed above, i.e., a given distortion level in the watermark reconstruction and the incorporation of an attack channel. Finally, Sections 5 and 6 will be devoted to the proof of the last (and most general) version of the coding theorem, with Section 5 focusing on the converse part, and Section 6 – on the direct part.

## 2 Notation Conventions

We begin by establishing some notation conventions. Throughout this paper, scalar random variables (RV's) will be denoted by capital letters, their sample values will be denoted by the respective lower case letters, and their alphabets will be denoted by the respective calligraphic letters. A similar convention will apply to random vectors and their sample values, which will be denoted with same symbols superscripted by the dimension. Thus, for example, $A^\ell$ ($\ell$ – positive integer) will denote a random $\ell$-vector $(A_1, ..., A_\ell)$, and $a^\ell = (a_1, ..., a_\ell)$ is a specific vector value in $\mathcal{A}^\ell$, the $\ell$-th Cartesian power of $\mathcal{A}$. The notations $a_i^j$ and $A_i^j$, where $i$ and $j$ are integers and $i \leq j$, will designate segments $(a_i, \ldots, a_j)$ and $(A_i, \ldots, A_j)$, respectively, where for $i = 1$, the subscript will be omitted (as above). For $i > j$, $a_i^j$ (or $A_i^j$) will be understood as the null string. Sequences without specifying indices are denoted by $\{\cdot\}$.

Sources and channels will be denoted generically by the letter $P$, or $Q$, subscripted by the name of the RV and its conditioning, if applicable, e.g., $P_U(u)$ is the probability function of $U$ at the point $U = u$, $P_{K|X}(k|x)$ is the conditional probability of $K = k$ given $X = x$, and so on. Whenever clear from the context, these subscripts will be omitted. Information theoretic quantities like entropies and mutual informations will be denoted following the usual conventions of the information theory literature, e.g., $H(U^N)$, $I(X^n; Y^n)$, and so on. For single–letter information quantities (i.e., when $n = 1$ or $N = 1$), subscripts will be omitted, e.g., $H(U^1) = H(U_1)$ will be denoted by $H(U)$, similarly, $I(X^1; Y^1) = I(X_1; Y_1)$ will be denoted by $I(X; Y)$, and so on.

## 3 Problem Definition and Main Result for Step 1

We now turn to the formal description of the model and the problem setting for step 1, as described in the Introduction. A source $P_X$, henceforth referred to as the *covertext source* or the *host source*, generates a sequence of independent copies, $\{X_t\}_{t=-\infty}^{\infty}$, of a finite–alphabet RV, $X \in \mathcal{X}$. At the same time and independently, another source $P_U$, henceforth referred to as the *message source*, or the *watermark source*, generates a sequence of independent copies, $\{U_i\}_{i=-\infty}^{\infty}$, of a finite–alphabet RV, $U \in \mathcal{U}$. The relative rate between the message source and the covertext source is $\lambda$ message symbols per covertext symbol. This means that while the covertext source generates a block of $n$ symbols, say, $X^n = (X_1, \ldots, X_n)$, the message source generates a block of $N = \lambda n$ symbols, $U^N = (U_1, \ldots, U_N)$ (assuming, without essential loss of generality, that $\lambda n$ is a positive integer). In addition to the covertext source and the message source, yet another source, $P_K$, henceforth referred to as the *key source*, generates a sequence of independent copies, $\{K_t\}_{t=-\infty}^{\infty}$, of a finite–alphabet RV, $K \in \mathcal{K}$, independently[3] of both $\{X_t\}$ and $\{U_i\}$. The key source is assumed to operate at the same rate as the covertext source, that is, while the covertext source generates the block $X^n$ of length $n$, the key source generates a block of $n$ symbols as well, $K^n = (K_1, \ldots, K_n)$. As the probability distribution $P_K$ of the key source is given, its entropy $H(K)$ is dictated. The entropy $H(K)$ has a dual meaning: It refers both to the available amount of randomness resources, and to the rate at which $K^n$ should be conveyed to the legitimate decoder (i.e., the capacity of the secure channel in between [11]).

Given $n$ and $\lambda$, a block code for *joint watermarking, encryption, and compression* is a mapping $f_n : \mathcal{U}^N \times \mathcal{X}^n \times \mathcal{K}^n \to \mathcal{Y}^n$, $N = \lambda n$, whose output $y^n = (y_1, \ldots, y_n) = f_n(u^N, x^n, k^n) \in \mathcal{Y}^n$ is referred to as the *stegotext* or the *composite signal*, and accordingly, the finite alphabet $\mathcal{Y}$ is referred to as the *stegotext alphabet*. Let $d : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}^+$ denote a single–letter distortion measure between covertext symbols and stegotext symbols, and let the distortion between the vectors, $x^n \in \mathcal{X}^n$ and $y^n \in \mathcal{Y}^n$, be defined additively across the corresponding components, as usual.

An $(n, \lambda, D, R_c, h, \delta)$ code is a block code for joint watermarking, encryption, and compression, with parameters $n$ and $\lambda$, that satisfies the following requirements:

[3]The assumption of independence between $\{K_t\}$ and $\{X_t\}$ is temporary and made now primarily for the sake of simplicity of the exposition. It will be dropped later on.

1. The expected distortion between the covertext and the stegotext satisfies

$$\sum_{t=1}^{n} Ed(X_t, Y_t) \leq nD. \tag{1}$$

2. The entropy of the stegotext satisfies

$$H(Y^n) \leq nR_c. \tag{2}$$

3. The equivocation of the message source satisfies

$$H(U^N|Y^n) \geq Nh. \tag{3}$$

4. There exists a decoder $g_n : \mathcal{Y}^n \times \mathcal{K}^n \to \mathcal{U}^N$ such that

$$P_e \stackrel{\Delta}{=} \Pr\{g_n(Y^n, K^n) \neq U^N\} \leq \delta. \tag{4}$$

For a given $\lambda$, a triple $(D, R_c, h)$ is said to be *achievable* if for every $\epsilon > 0$, there is a sufficiently large $n$ for which $(n, \lambda, D + \epsilon, R_c + \epsilon, h - \epsilon, \epsilon)$ codes exist. The *achievable region* of triples $(D, R_c, h)$ is the set of all achievable triples $(D, R_c, h)$. For simplicity, it is assumed[4] that $H(K) \leq \lambda H(U)$ as this upper limit on $H(K)$ suffices to achieve perfect secrecy.

Our first coding theorem is the following:

**Theorem 1** *A triple $(D, R_c, h)$ is achievable if and only if the following conditions are both satisfied:*

*(a)* $h \leq H(K)/\lambda$.

*(b) There exists a channel $\{P_{Y|X}(y|x),\ x \in \mathcal{X},\ y \in \mathcal{Y}\}$ such that: (i) $H(Y|X) \geq \lambda H(U)$, (ii) $R_c \geq \lambda H(U) + I(X;Y)$, and (iii) $D \geq Ed(X,Y)$.*

As can be seen, the encryption, on the one hand, and the embedding and the compression, on the other hand, do not interact at all in this theorem. There is a complete decoupling between them: While condition (a) refers solely to the key and the secrecy of the watermark, condition (b) is only about the embedding–compression part, and it is a replica of the conditions of the coding theorem in [9], where the role of the embedding rate, $R_e$ (see Introduction above), is played by the product $\lambda H(U)$. This suggests a very simple

[4] At the end of Section 4 (after Theorem 4), we discuss the case where this limitation (or its analogue in lossy reconstruction of $U^N$) is dropped.

separation principle, telling that in order to attain a given achievable triple $(D, R_c, h)$, first compress the watermark $U^N$ to its entropy, then encrypt $Nh$ bits (out of the $NH(U)$) of the compressed bit–string (by bit–by–bit XORing with the same number of compressed key bits), and finally, embed this partially encrypted compressed bit–string into the covertext, using the coding theorem of [9] (again, see the Introduction above for a brief description of this).

## 4 Extensions to Steps 2 and 3

Moving on to Step 2, we now relax requirement no. 4 in the above definition of an $(n, \lambda, D, R_c, h, \delta)$ code, and allow a certain distortion between $U^N$ and its reconstruction $\hat{U}^N$ at the legitimate decoder. More precisely, let $\hat{\mathcal{U}}$ denote a finite alphabet, henceforth referred to as the *message reconstruction alphabet*. Let $d' : \mathcal{U} \times \hat{\mathcal{U}} \to \mathbb{R}^+$ denote a single–letter distortion measure between message symbols and message reconstruction symbols, and let the distortion between vectors $u^N \in \mathcal{U}^N$ and $\hat{u}^N \in \hat{\mathcal{U}}^N$ be again, defined additively across the corresponding components. Finally, let $R_U(D')$ denote the rate–distortion function of the source $P_U$ w.r.t. $d'$, i.e.,

$$R_U(D') = \min\{I(U;\hat{U}) : \ Ed'(U,\hat{U}) \leq D'\}. \tag{5}$$

It will now be assumed that $H(K) \leq \lambda R_U(D')$, for the same reasoning as before.

Requirement no. 4 is now replaced by the following requirement: There exists a decoder $g_n : \mathcal{Y}^n \times \mathcal{K}^n \to \hat{\mathcal{U}}^N$ such that $\hat{U}^N = (\hat{U}_1, \ldots, \hat{U}_N) = g_n(Y^n, K^n)$ satisfies:

$$\sum_{i=1}^{N} Ed'(U_i, \hat{U}_i) \leq ND'. \tag{6}$$

In addition to this modification of requirement no. 4, we add, to requirement no. 3, a specification regarding the minimum allowed equivocation w.r.t. the reconstructed message:

$$H(\hat{U}^N|Y^n) \geq Nh', \tag{7}$$

in order to guarantee that the secrecy of the reconstructed message is also secure enough. Accordingly, we modify the above definition of a block code as follows: An $(n, \lambda, D, D', R_c, h, h')$ code is a block code for joint watermarking, encryption, and compression with parameters $n$ and $\lambda$ that satisfies requirements 1–4, with the above modifications of requirements 3 and

4. For a given $\lambda$, a quintuple $(D, D', R_c, h, h')$ is said to be *achievable* if for every $\epsilon > 0$, there is a sufficiently large $n$ for which $(n, \lambda, D + \epsilon, D' + \epsilon, R_c + \epsilon, h - \epsilon, h' - \epsilon)$ codes exist.

Our second theorem extends Theorem 1 to this setting:

**Theorem 2** *A quintuple $(D, D', R_c, h, h')$ is achievable if and only if the following conditions are all satisfied:*

*(a)* $h \leq H(K)/\lambda + H(U) - R_U(D')$.

*(b)* $h' \leq H(K)/\lambda$.

*(c) There exists a channel $\{P_{Y|X}(y|x),\ x \in \mathcal{X},\ y \in \mathcal{Y}\}$ such that: (i) $\lambda R_U(D') \leq H(Y|X)$, (ii) $R_c \geq \lambda R_U(D') + I(X;Y)$, and (iii) $D \geq Ed(X,Y)$.*

As can be seen, the passage from Theorem 1 to Theorem 2 includes the following modifications: In condition (c), $H(U)$ is simply replaced by $R_U(D')$ as expected. This means that the lossless compression code of $U^N$, in the achievability of Theorem 1, is now replaced by a rate–distortion code for distortion level $D'$. Conditions (a) and (b) now tell us that the key rate (in terms of entropy) should be sufficiently large to satisfy both equivocation requirements. Note that the condition regarding the equivocation w.r.t. the clean message source is softer than in Theorem 1 as $H(U) - R_U(D') \geq 0$. This is because the rate–distortion code for $U^N$ already introduces an uncertainty of $H(U) - R_U(D')$ bits per symbol, and so, the encryption should only complete it to the desired level of $h$ bits per symbol. This point is discussed in depth in [19]. Of course, by setting $D' = 0$ (and hence also $h' = h$), we are back to Theorem 1.

We also observe that the encryption and the embedding are still decoupled in Theorem 2, and that an achievable quintuple can still be attained by separation: First, apply a rate–distortion code to $U^N$, as mentioned earlier, then encrypt $N \cdot \max\{h + R_U(D') - H(U), h'\}$ bits of the compressed codeword (to satisfy both equivocation requirements), and finally, embed the (partially) encrypted codeword into $X^n$, again, by using the scheme of [9]. Note that without the encryption and without requirement no. 2 of the compressibility of $Y^n$, this separation principle is a special case of the one in [12], where a separation theorem was established for the Wyner–Ziv source (with SI correlated to the source at the decoder) and the Gel'fand–Pinsker channel (with channel SI at the encoder). Here, there is no SI correlated to the source and the role of channel SI is fulfilled by the covertext. Thus, the

new observation here is that the separation theorem continues to hold in the presence of encryption and requirement no. 2.

Finally, we turn to step 3, of including an attack channel (see Fig. 1). Let $\mathcal{Z}$ be a finite alphabet, henceforth referred to as the *forgery alphabet*, and let $\{P_{Z|Y}(z|y),\ y \in \mathcal{Y},\ z \in \mathcal{Z}\}$ denote a set of conditional PMF's from the stegotext alphabet to the forgery alphabet. We now assume that the stegotext vector is subjected to an attack modelled by the memoryless channel,

$$P_{Z^n|Y^n}(z^n|y^n) = \prod_{t=1}^{n} P_{Z|Y}(z_t|y_t). \tag{8}$$

The output $Z^n$ of the attack channel will henceforth be referred to as the *forgery*.

It is now assumed and that the legitimate decoder has access to $Z^n$, rather than $Y^n$ (in addition, of course, to $K^n$). Thus, in requirement no. 4, the decoder is redefined again, this time, as a mapping $g_n : \mathcal{Z}^n \times \mathcal{K}^n \to \hat{\mathcal{U}}^N$ such that $\hat{U}^N = g_n(Z^n, K^n)$ satisfies the distortion constraint (6). As for the equivocation requirements, the conditioning will now be on both $Y^n$ and $Z^n$, i.e.,

$$H(U^N|Y^n, Z^n) \geq Nh \quad \text{and} \quad H(\hat{U}^N|Y^n, Z^n) \geq Nh', \tag{9}$$

as if the attacker and the eavesdropper are the same party (or if they cooperate), then s/he may access both. In fact, for the equivocation of $U^N$, the conditioning on $Z^n$ is immaterial since $U^N \to Y^n \to Z^n$ is always a Markov chain, but it is not clear that $Z^n$ is superfluous for the equivocation w.r.t. $\hat{U}^N$ since $Z^n$ is one of the inputs to the decoder whose output is $\hat{U}^N$. Nonetheless, for the sake of uniformity and convenience (in the proof), we keep the conditioning on $Z^n$ in both equivocation criteria.

Redefining block codes and achievable quintuples $(D, D', R_C, h, h')$ according to the modified requirements in the same spirit, we now have the following coding theorem, which is substantially different from Theorems 1 and 2:

**Theorem 3** *A quintuple $(D, D', R_c, h, h')$ is achievable if and only if there exist RV's $V$ and $Y$ such that $P_{KXVYZ}(k, x, v, y, z) = P_X(x)P_K(k)P_{VY|KX}(v, y|k, x)P_{Z|Y}(z|y)$, where the alphabet size of $V$ is bounded by $|\mathcal{V}| \leq |\mathcal{K}|\cdot|\mathcal{X}|\cdot|\mathcal{Y}|+1$, and such that the following conditions are all satisfied:*

*(a)* $h \leq H(K|Y)/\lambda + H(U) - R_U(D')$.

*(b)* $h' \leq H(K|Y)/\lambda$.

*(c)* $\lambda R_U(D') \leq I(V;Z|K) - I(V;X|K)$.

*(d)* $R_c \geq \lambda R_U(D') + I(X;Y,V|K) + I(K;Y)$.

*(e)* $D \geq Ed(X,Y)$.

First, observe that here, unlike in Theorems 1 and 2, it is no longer true that the encryption and the embedding (along with stegotext compression) are decoupled, yet the rate–distortion compression of $U^N$ is still separate and decoupled from both. In other words, the separation principle applies here in a partial manner only. Note that now, although $K$ is still assumed independent of $X$, it may, in general, depend on $Y$. On the negative side, this dependence causes a reduction in the equivocation of both the message source and its reconstruction, and therefore $H(K|Y)$ replaces $H(K)$ in conditions (a) and (b). On the positive side, on the other hand, this dependence introduces new degrees of freedom in enhancing the tradeoffs between the embedding performance (condition (c)) and the compressibility (condition (d)).

The achievability of Theorem 3 involves essentially the same stages as before (rate–distortion coding of $U^N$, followed by encryption, followed in turn by embedding), but this time, the embedding scheme is a conditional version of the one proposed in [10], where all codebooks depend on $K^n$, the SI given at both ends (see Fig. 2). An interesting point regarding the encryption is that one needs to generate, from $K^n$, essentially $nH(K|Y)$ random bits that are *independent* of $Y^n$ (and $Z^n$), in order to protect the secrecy against an eavesdropper that observes $Y^n$ and $Z^n$. Clearly, if $Y^n$ was given in advance to the encrypting unit, then the compressed bitstring of an optimal lossless source code that compresses $K^n$, given $Y^n$ as SI, would have this property (as if there was any dependence, then this bitstring could have been further compressed, which is a contradiction). However, such a source code cannot be implemented since $Y^n$ itself is generated from the encrypted message, i.e., *after* the encryption. In other words, this would have required a circular mechanism, which may not be feasible. A simple remedy is then to use a *Slepian–Wolf encoder* [15], that generates $nH(K|Y)$ bits that are essentially independent of $Y^n$ (due to the same consideration), without the need to access the vector $Y^n$ to be generated. For more details, the reader is referred to the proof of the direct part (Section 6).

Observe that in the absence of attack (i.e., $Z = Y$), Theorem 2 is obtained as a special case of Theorem 3 by choosing $V = Y$ and letting both be independent of $K$, a choice

which is simultaneously the best for conditions (a)–(d) of Theorem 3. To see this, note the following simple inequalities: In conditions (a) and (b), $H(K|Y) \leq H(K)$. In condition (c), by setting $Z = Y$, we have

$$\begin{aligned} I(V;Y|K) - I(V;X|K) &\leq I(V;X,Y|K) - I(V;X|K) \\ &= I(V;Y|X,K) \\ &\leq H(Y|X,K) \\ &\leq H(Y|X). \end{aligned} \tag{10}$$

Finally in condition (d), clearly, $I(K;Y) \geq 0$ and since $X$ is independent of $K$, then $I(X;Y,V|K) = I(X;Y,V,K) \geq I(X;Y)$. Thus, for $Z = Y$, the achievable region of Theorem 3 is a subset of the one given in Theorem 2. However, since all these inequalities become equalities at the same time by choosing $V = Y$ and letting both be independent of $K$, the two regions are identical in the attack–free case.

Returning now to Theorem 3, as we observed, $K^n$ is now involved not only in the role of a cipher key, but also as SI available at both encoder and decoder. Two important points are now in order, in view of this fact.

First, one may argue that, actually, there is no real reason to assume that $K^n$ is necessarily independent of $X^n$. The idea of dropping this independence assumption was suggested also earlier in [13]. In this situation, it is more plausible to think of $K^n$ in the spirit of its new role, namely, as (synthetic) *side information*, rather than in its original role, i.e., strictly as a cryptographic key (which is normally assumed to be an independent source of randomness at a certain rate), although $K^n$ is still used in order to protect the secrecy of $U^N$ and $\hat{U}^N$. The idea then is as follows: If the user has control of the mechanism of generating $K^n$, then s/he might implement, in general, a channel $P_{K^n|X^n}(k^n|x^n)$ by using the available independent randomness resources,[5] taking advantage of the randomness of the covertext. This can be done by using the notion of channel simulation (see, e.g., [17]). Let us assume that this channel is stationary and memoryless, i.e.,

$$P_{K^n|X^n}(k^n|x^n) = \prod_{t=1}^{n} P_{K|X}(k_t|x_t) \tag{11}$$

with the single–letter transition probabilities $\{P_{K|X}(k|x)\ x \in \mathcal{X},\ k \in \mathcal{K}\}$ left as a degree of freedom for design. Given the covertext $X^n$, one generates $K^n$ using this channel,

[5]These randomness resources are, in fact, purely random, independent bits, which can now be redefined as our secret "key", in the original meaning of this term.

and then $K^n$ is shared with the legitimate decoder.[6] While so far, we assumed that $K^n$ was independent of $X^n$, the other extreme is, of course, $K^n = X^n$ (corresponding to private watermarking). Note, however, that in the attack–free case, in the absence of the compressibility requirement no. 2 (say, $R_c = \infty$), no optimality is lost by assuming that $K^n$ is independent of $X^n$, since the only inequality where we have used the independence assumption, in the previous paragraph, corresponds to condition (d).

The second point is that in Theorems 1–3, so far, we have defined the compressibility of the stegotext in terms of $H(Y^n)$, which is suitable when the decompression of $Y^n$ is *public*, i.e., without access to $K^n$. However, since the content provider may wish to store the stegotext $Y^n$ (in the presence of $K^n$) for possible future use, it may make sense to measure the compressibility of the stegotext also in a *private* regime, i.e., in terms of the *conditional* entropy, $H(Y^n|K^n)$.

Our last (and most general) version of the coding theorem below takes these two points in to account. Specifically, let us impose, in requirement no. 2, an additional inequality,

$$H(Y^n|K^n) \leq nR_c', \tag{12}$$

where $R_c'$ is a prescribed constant, and let us redefine accordingly the block codes and the achievable region in terms of six–tuples $(D, D', R_c, R_c', h, h')$. We now have the following result:

**Theorem 4** *A six–tuple $(D, D', R_c, R_c', h, h')$ is achievable if and only if there exist RV's $V$ and $Y$ such that $P_{KXVYZ}(k, x, v, y, z) = P_{XK}(x, k)P_{VY|KX}(v, y|k, x)P_{Z|Y}(z|y)$, where the alphabet size of $V$ is bounded by $|\mathcal{V}| \leq |\mathcal{K}|\cdot|\mathcal{X}|\cdot|\mathcal{Y}|+1$, and such that the following conditions are all satisfied:*

*(a)* $h \leq H(K|Y)/\lambda + H(U) - R_U(D')$.

*(b)* $h' \leq H(K|Y)/\lambda$.

*(c)* $\lambda R_U(D') \leq I(V;Z|K) - I(V;X|K)$.

*(d)* $R_c \geq \lambda R_U(D') + I(X;Y,V|K) + I(K;Y)$.

[6]Note that now there is a distinction between the required available randomness rate, which is $H(K^n|X^n)/n = H(K|X)$ [17], and the rate at which the key must be conveyed to the legitimate decoder, which remains $H(K)$ (as the decoder has no access to $X^n$). This is in contrast to the case of independence between $K^n$ and $X^n$, where these two parameters coincide.

*(e)* $R'_c \geq \lambda R_U(D') + I(X;Y,V|K)$.

*(f)* $D \geq Ed(X,Y)$.

Note that the additional condition, (e), is similar to condition (d) except for the term $I(K;Y)$. Also, in the joint PMF of $(K,X,V,Y,Z)$ we are no longer assuming that $K$ and $X$ are independent. It should be pointed out that in the presence of the new requirement regarding $H(Y^n|K^n)$, it is more clear now that introducing dependence of $(V,Y)$ upon $K$ is reasonable, in general. In the case $K = X$, that was mentioned earlier, the term $I(V;X|K)$, in condition (c), and the term $I(X;Y,V|K)$, in conditions (d) and (e), both vanish. Thus, both embedding performance and compression performance improve, like in private watermarking.

Finally, a comment is in order regarding the assumption $H(K) \leq \lambda R_U(D')$, which implies that $H(K|Y)$ cannot exceed $\lambda R_U(D')$ either. If this assumption is removed, and even $H(K|Y)$ is allowed to exceed $\lambda R_U(D')$, then Theorem 4 can be somewhat further extended. While $h$ cannot be further improved if $H(K|Y)$ is allowed to exceed $\lambda R_U(D')$ (as it already reaches the maximum possible value, $h = H(U)$, for $H(K|Y) = \lambda R_U(D')$), it turns out that there is still room for improvement in $h'$. Suppose that instead of one rate–distortion codebook for $U^N$, we have many *disjoint* codebooks. In fact, it has been shown in [9] that there are exponentially $2^{NH(\hat{U}|U)}$ disjoint codebooks, each covering the set of typical source sequences by jointly typical codewords. Now, if $H(K|Y) > \lambda R_U(D')$, we can use the $T = nH(K|Y) - NR_U(D')$ excess bits of the compressed key (beyond the $NR_U(D')$ bits that are used to encrypt the binary of representation of $\hat{U}^N$), so as to select one of $2^T$ codebooks (as long as $T < NH(\hat{U}|U)$), and thus reach a total equivocation of $nH(K|Y)$ as long as $nH(K|Y) \leq NH(\hat{U})$, or equivalently, $H(K|Y) \leq \lambda H(\hat{U})$. The equivocation level $h' = H(\hat{U})$ is now the "saturation value" that cannot be further improved (in analogy to $h = H(U)$ for the original source). This means that condition (b) of Theorem 4 would now be replaced by the condition

$$h' \leq \min\{H(\hat{U}), H(K|Y)/\lambda\}. \tag{13}$$

But with this condition, it is no longer clear that the best test channel for lossy compression of $U^N$ is the one that achieves $R_U(D')$, because for the above modified version of condition (b), it would be best to have $H(\hat{U})$ as large as possible (as long as it is below $H(K|Y)/\lambda$),

which is in partial conflict with the minimization of $I(U;\hat{U})$ that leads to $R_U(D')$. Therefore, a restatement of Theorem 4 would require the existence of a channel $\{P_{\hat{U}|U}(\hat{u}|u),\ u\in\mathcal{U},\ \hat{u}\in\hat{\mathcal{U}}\}$ (in addition to the existing requirement of a channel $P_{VY|KX}$), such that the random variable $\hat{U}$ takes now part in the compromise among *all* criteria of the problem. This means that in conditions (a),(c),(d), and (e) of Theorem 4, $R_U(D')$ should be replaced by $I(U;\hat{U})$, and there would be an additional condition (g): $Ed'(U,\hat{U})\le D'$. Condition (a), in view of the earlier discussion above, would now be of the form:

$$h\le\min\{H(U),H(K|Y)/\lambda+H(U)-I(U;\hat{U})\}\equiv H(U)-[I(U;\hat{U})-H(K|Y)/\lambda]_+, \quad (14)$$

where $[z]_+\stackrel{\Delta}{=}\max\{0,z\}$. Of course, under the assumption $H(K)\le\lambda R_U(D')$, that we have used thus far,

$$H(\hat{U})\ge I(U;\hat{U})\ge R_U(D')\ge H(K)/\lambda\ge H(K|Y)/\lambda, \quad (15)$$

in other words, $\min\{H(\hat{U}),H(K|Y)/\lambda\}$ is always attained by $H(K|Y)/\lambda$, and so, the dependence on $H(\hat{U})$ disappears, which means that the best choice of $\hat{U}$ (for all other conditions) is back to be the one that minimizes $I(U;\hat{U})$, which gives us Theorem 4 as is.

It is interesting to point out that this additional extension gives rise to yet another step in the direction of invalidating the separation principle: While in Theorem 4 only the encryption and the embedding interacted, yet the rate–distortion coding of $U^N$ was still independent of all other ingredients of the system, here even this is no longer true, as the choice of the test channel $P_{\hat{U}|U}$ takes into account also compromises that are associated with the encryption and the embedding.

Note that this discussion applies also to the *classical* joint source–channel coding, where there is no embedding at all: In this case, $X$ is a degenerate RV (say, $X\equiv 0$, if $0\in\mathcal{X}$), and so, the mutual information terms depending on $X$ in conditions (c), (d) and (e), all vanish, the best choice of $V$ is $V=Y$ (thus, the r.h.s in condition (c) becomes the capacity of the channel $P_{Z|Y}$ with $K$ as SI at both ends), and condition (f) may be interpreted as a (generalized) power constraint (with power function $\phi(y)=d(0,y)$). Nonetheless, the new versions of conditions (a) and (b) remain the same as in eqs. (13) and (14). This is to say that the violation of the separation principle occurs even in the classical model of a communication system, once security becomes an issue and one is interested in the security of the reconstructed source.

## 5 Proof of the Converse Part of Theorem 4

Let an $(n, \lambda, D+\epsilon, D'+\epsilon, R_c+\epsilon, R'_c+\epsilon, h-\epsilon, h'-\epsilon)$ code be given. First, from the requirement $H(Y^n|K^n) \leq n(R'_c+\epsilon)$, we have:

$$\begin{aligned} n(R'_c+\epsilon) &\geq H(Y^n|K^n) && (16)\\ &= H(Y^n|U^N,K^n) + I(U^N;Y^n|K^n) \\ &\geq H(Y^n|U^N,K^n) + I(U^N;Z^n|K^n) \\ &= H(Y^n|U^N,K^n) + I(U^N;Z^n,K^n) && (17)\end{aligned}$$

where the second inequality comes from the data processing theorem ($U^N \to Y^n \to Z^n$ is a Markov chain given $K^n$) and the last equality comes from the chain rule and the fact that $U^N$ and $K^n$ are independent. Define $\tilde{V}_t = (X_{t+1}^n, U^N, K^{t-1}, Z^{t-1})$, $J$ – as a uniform RV over $\{1,\ldots,n\}$, $X = X_J$, $K = K_J$, $Y = Y_J$, $V' = \tilde{V}_J$, and $V = (\tilde{V}_J, J) = (V', J)$. Now, the first term on the right–most side of eq. (17) is further lower bounded in the following manner.

$$\begin{aligned} H(Y^n|U^N,K^n) &\geq I(X^n;Y^n|U^N,K^n) \\ &= I(X^n;Y^n,U^N,K^n) - I(X^n;U^N,K^n) \\ &= \sum_{t=1}^n I(X_t;Y^n,U^N,K^n|X_{t+1}^n) - I(X^n;K^n) && (18)\\ &= \sum_{t=1}^n I(X_t;Y^n,U^N,K^n,X_{t+1}^n) - nI(X;K) && (19)\\ &\geq \sum_{t=1}^n I(X_t;K_t,Y_t,U^N,K^{t-1},Z^{t-1},X_{t+1}^n) - nI(X;K) && (20)\\ &= \sum_{t=1}^n I(X_t;K_t,Y_t,\tilde{V}_t) - nI(X;K) \\ &= n[I(X;K,Y,V'|J) - I(X;K)] \\ &= n[I(X;K,Y,V',J) - I(X;K)] && (21)\\ &= nI(X;Y,V|K) && (22)\end{aligned}$$

where (18) is due to the chain rule and fact that $(X^n, K^n)$ is independent of $U^N$ (hence $U^N \to K^n \to X^n$ is trivially a Markov chain), (19) is due to the memorylessness of $\{(X_t, K_t)\}$, (20) is due to the data processing theorem, and (21) follows from the fact that $\{X_t\}$ is stationary and so, $X = X_J$ is independent of $J$. The second term on the

right–most side of eq. (17) is in turn lower bounded following essentially the same ideas as in the proof of the converse to the rate–distortion coding theorem (see, e.g., [3]):

$$\begin{aligned} I(U^N; Z^n, K^n) &= H(U^N) - H(U^N|Z^n, K^n) \\ &= \sum_{i=1}^{N} [H(U_i) - H(U_i|U^{i-1}, Z^n, K^n)] \\ &= \sum_{i=1}^{N} I(U_i; U^{i-1}, Z^n, K^n) \\ &\geq \sum_{i=1}^{N} I(U_i; [g_n(Z^n, K^n)]_i) \\ &\geq \sum_{i=1}^{N} R_U(Ed'(U_i, [g_n(Z^n, K^n)]_i)) \\ &\geq NR_U\left(\frac{1}{N}\sum_{i=1}^{N} Ed'(U_i, [g_n(Z^n, K^n)]_i)\right) \\ &\geq NR_U(D' + \epsilon), \end{aligned} \tag{23}$$

where $[g_n(Z^n, K^n)]_i$ denotes the $i$-th component projection of $g_n(Z^n, K^n)$, i.e., $\hat{U}_i$ as a function of $(Z^n, K^n)$. Combining eqs. (17), (22), and (23), we get

$$n(R_c' + \epsilon) \geq NR_U(D' + \epsilon) + nI(X; Y, V|K). \tag{24}$$

Dividing by $n$, we get

$$R_c' + \epsilon \geq \lambda R_U(D' + \epsilon) + I(X; Y, V|K). \tag{25}$$

Using the arbitrariness of $\epsilon$ together with the continuity of $R_U(\cdot)$, we get condition (e) of Theorem 4.

Condition (d) is derived in the very same manner except that the starting point is the inequality $n(R_c + \epsilon) \geq H(Y^n)$, and when $H(Y^n)$ is further bounded from below, in analogy to the chain of inequalities (17), there is an additional term, $I(K^n; Y^n)$, that is in turn lower bounded in the following manner:

$$\begin{aligned} I(K^n; Y^n) &\geq \sum_{t=1}^{n} I(K_t; Y_t) \\ &= nI(K; Y|J) \\ &= n[H(K|J) - H(K|J, Y)] \\ &\geq n[H(K) - H(K|Y)] \\ &= nI(K; Y), \end{aligned} \tag{26}$$

where the first inequality is because of the memorylessness of $\{K_t\}$, and the second inequality comes from the facts that conditioning reduces entropy (in the second term) and that $K$ is independent of $J$ (again, due to the stationarity of $\{K_t\}$). This gives the additional term, $I(K;Y)$, in condition (d).

Condition (c) is obtained as follows:

$$\begin{aligned}
NR_U(D'+\epsilon) &\leq I(U^N;K^n,Z^n) \\
&= I(U^N;K^n,Z^n) - I(U^N;K^n,X^n) \\
&\leq \sum_{t=1}^{n}[I(\tilde{V}_t;K_t,Z_t) - I(\tilde{V}_t;K_t,X_t)] \qquad (27) \\
&= n[I(V';K,Z|J) - I(V';K,X|J)] \\
&\leq n[I(V',J;K,Z) - I(V',J;K,X)] \qquad (28) \\
&= n[I(V;K,Z) - I(V;K,X)] \\
&= n[I(V;Z|K) - I(V;X|K)], \qquad (29)
\end{aligned}$$

where the first inequality is (23), the first equality is due to the independence between $U^N$ and $(K^n,X^n)$, the second inequality is an application of [5, Lemma 4], the third inequality is due to the fact that $I(K,Z;J) \geq 0$ and $I(K,X;J) = 0$ (due to the stationarity of $\{(K_t,X_t)\}$), and the last equality is obtained by adding and subtracting $I(V;K)$. Again, since this is true for every $\epsilon > 0$, it holds also for $\epsilon = 0$, due to continuity.

As for condition (f), we have:

$$D + \epsilon \geq \frac{1}{n}\sum_{t=1}^{n} Ed(X_t,Y_t) = Ed(X,Y), \qquad (30)$$

and we use once again the arbitrariness of $\epsilon$. Regarding condition (b), we have:

$$\begin{aligned} nH(K|Y) &\geq nH(K|Y,J) \\ &= \sum_{t=1}^{n} H(K_t|Y_t) \\ &\geq \sum_{t=1}^{n} H(K_t|K^{t-1},Y^n) \\ &= H(K^n|Y^n) \\ &= H(K^n|Y^n,Z^n) \\ &\geq I(K^n;\hat{U}^N|Y^n,Z^n) \\ &= H(\hat{U}^N|Y^n,Z^n) - H(\hat{U}^N|Y^n,Z^n,K^n) \\ &= H(\hat{U}^N|Y^n,Z^n) \\ &\geq N(h'-\epsilon), \end{aligned} \tag{31}$$

where the last equality is due to the fact that $\hat{U}^N$ is, by definition, a function of $(Z^n, K^n)$, and the last inequality is by the hypothesis that the code achieves an equivocation of at least $N(h'-\epsilon)$. Dividing by $N$ and taking the limit $\epsilon \to 0$, leads to $h' \leq H(K|Y)/\lambda$, which is condition (b). Finally, to prove condition (a), consider the inequality $nH(K|Y) \geq H(\hat{U}^N|Y^n,Z^n)$, that we have just proved, and proceed as follows (see also [19]):

$$\begin{aligned} nH(K|Y) &\geq H(\hat{U}^N|Y^n,Z^n) \\ &\geq H(\hat{U}^N|Y^n,Z^n) + N(h-\epsilon) - H(U^N|Y^n,Z^n) \\ &= N(h-\epsilon) - H(U^N) + I(U^N;Y^n,Z^n) - \\ &\quad I(\hat{U}^N;Y^n,Z^n) + I(\hat{U}^N;U^N) + H(\hat{U}^N|U^N) \\ &\geq N[h-\epsilon-H(U)+R_U(D'+\epsilon)] + \\ &\quad [I(U^N;Y^n,Z^n) - I(\hat{U}^N;Y^n,Z^n) + H(\hat{U}^N|U^N)], \end{aligned} \tag{32}$$

where the second inequality follows from the hypothesis that the code satisfies $H(U^N|Y^n,Z^n) \geq N(h-\epsilon)$, and the third inequality is due to the memorylessness of $\{U_i\}$, the hypothesis that $\sum_{i=1}^{N} Ed'(U_i,\hat{U}_i) \leq N(D'+\epsilon)$, and the converse to the rate–distortion coding theorem. Now, to see that the second bracketed term is non–negative, we have the following chain of

inequalities:

$$\begin{aligned} & I(U^N;Y^n,Z^n) - I(\hat{U}^N;Y^n,Z^n) + H(\hat{U}^N|U^N) \\ = \;& I(U^N;Y^n,Z^n) - H(Y^n,Z^n) + H(Y^n,Z^n|\hat{U}^N) + H(\hat{U}^N|U^N) \\ \geq \;& I(U^N;Y^n,Z^n) - H(Y^n,Z^n) + H(Y^n,Z^n|U^N,\hat{U}^N) + H(\hat{U}^N|U^N) \\ = \;& I(U^N;Y^n,Z^n) - H(Y^n,Z^n) + H(Y^n,Z^n,\hat{U}^N|U^N) \\ \geq \;& I(U^N;Y^n,Z^n) - H(Y^n,Z^n) + H(Y^n,Z^n|U^N) \\ = \;& 0. \end{aligned} \tag{33}$$

Combining this with eq. (32), we have

$$nH(K|Y) \geq N[h - \epsilon - H(U) + R_U(D' + \epsilon)]. \tag{34}$$

Dividing again by $N$, and letting $\epsilon$ vanish, we obtain $h \leq H(K|Y)/\lambda + H(U) - R_U(D')$, which completes the proof of condition (a).

To complete the proof of the converse part, it remains to show that the alphabet size of $V$ can be reduced to $|\mathcal{K}| \cdot |\mathcal{X}| \cdot |\mathcal{Y}| + 1$. To this end, we extend the proof of the parallel argument in [10] by using the support lemma (cf. [4]), which is based on Carathéodory's theorem. According to this lemma, given $J$ real valued continuous functionals $f_j$, $j = 1, ..., J$ on the set $\mathcal{P}(\mathcal{X})$ of probability distributions over the alphabets $\mathcal{X}$, and given any probability measure $\mu$ on the Borel $\sigma$-algebra of $\mathcal{P}(\mathcal{X})$, there exist $J$ elements $Q_1, ..., Q_J$ of $\mathcal{P}(\mathcal{X})$ and $J$ non-negative reals, $\alpha_1, ..., \alpha_J$, such that $\sum_{j=1}^{J} \alpha_j = 1$ and for every $j = 1, ..., J$

$$\int_{\mathcal{P}(\mathcal{X})} f_j(Q)\mu(dQ) = \sum_{i=1}^{J} \alpha_i f_j(Q_i). \tag{35}$$

Before we actually apply the support lemma, we first rewrite the relevant mutual informations of Theorem 4 in a more convenient form for the use of this lemma. First, observe that

$$\begin{aligned} I(V;Z|K) - I(V;X|K) &= H(Z|K) - H(Z|V,K) - H(X|K) + H(X|V,K) \\ &= H(Z|K) - H(X|K) + H(K,X|V) - H(K,Z|V). \end{aligned} \tag{36}$$

and

$$
\begin{aligned}
I(X;Y,V|K) &= I(X;V|K) + I(X;Y|V,K) && (37)\\
&= H(X|K) - H(X|V,K) + H(X|V,K) - H(X|V,Y,K)\\
&= H(X|K) - H(X|V,Y,K)\\
&= H(X|K) - H(K,X,Y|V) + H(K,Y|V). && (38)
\end{aligned}
$$

For a given joint distribution of $(K,X,Y)$, and given $P_{Z|Y}$, $H(Z|K)$ and $H(X|K)$ are both given and unaffected by $V$. Therefore, in order to preserve prescribed values of $I(V;Z|K) - I(V;X|K)$ and $I(X;V,Y|K)$, it is sufficient to preserve the associated values $H(K,X|V) - H(K,Z|V)$ and $H(K,X,Y|V) - H(K,Y|V)$. Let us define then the following functionals of a generic distribution $Q$ over $\mathcal{K}\times\mathcal{X}\times\mathcal{Y}$, where $\mathcal{K}\times\mathcal{X}\times\mathcal{Y}$ is assumed, without loss of generality, to be $\{1,2,...,m\}$, $m = |\mathcal{K}|\cdot|\mathcal{X}|\cdot|\mathcal{Y}|$:

$$f_i(Q) = Q(k,x,y), \quad i \stackrel{\Delta}{=} (k,x,y) = 1,...,m-1 \tag{39}$$

$$f_m(Q) = \sum_{k,x,y} Q(k,x,y) \sum_z P_{Z|Y}(z|y) \log \frac{\sum_{x,y} Q(k,x,y) P_{Z|Y}(z|y)}{Q(k,x)}. \tag{40}$$

Next define

$$f_{m+1}(Q) = \sum_{k,x,y} Q(k,x,y) \log \frac{Q(k,y)}{Q(k,x,y)}. \tag{41}$$

Applying now the support lemma, we find that there exists a random variable $V$ (jointly distributed with $(K,X,Y)$), whose alphabet size is $|\mathcal{V}| = m+1 = |\mathcal{K}|\cdot|\mathcal{X}|\cdot|\mathcal{Y}|+1$ and it satisfies simultaneously:

$$\sum_v \Pr\{V=v\} f_i(P(\cdot|v)) = P_{KXY}(k,x,y), \ i = 1,...,m-1, \tag{42}$$

$$\sum_v \Pr\{V=v\} f_m(P(\cdot|v)) = H(K,X|V) - H(K,Z|V), \tag{43}$$

and

$$\sum_u \Pr\{V=v\} f_{m+1}(P(\cdot|v)) = H(K,X,Y|V) - H(K,Y|V). \tag{44}$$

It should be pointed out that this random variable maintains the prescribed distortion level $Ed(X,Y)$ since $P_{XY}$ is preserved. By the same token, $H(K|Y)$ and $I(K;Y)$, which depend only on $P_{KY}$, are preserved as well. This completes the proof of the converse part of Theorem 4.

## 6 Proof of the Direct Part of Theorem 4

In this section, we show that if there exist RV's $(V, Y)$ that satisfy the conditions of Theorem 4, then for every $\epsilon > 0$, there is a sufficiently large $n$ for which $(n, \lambda, D+\epsilon, D'+\epsilon, R_c+\epsilon, R'_c+\epsilon, h-\epsilon, h'-\epsilon)$ codes exist. One part of the proof is strongly based on a straightforward extension of the proof of the direct part of [10] to the case of additional SI present at both encoder and decoder. Nonetheless, for the sake of completeness, the full details are provided here. It should be pointed out that for the attack–free case, an analogous extension can easily be offered to the direct part of [9].

We first digress to establish some additional notation conventions associated with the method of types [4]. For a given generic finite–alphabet random variable (RV) $A \in \mathcal{A}$ (or a vector of RV's taking on values in $\mathcal{A}$), and a vector $a^\ell \in \mathcal{A}^\ell$ ($\ell$ – positive integer), the empirical probability mass function (EPMF) is a vector $P_{a^\ell} = \{P_{a^\ell}(a'),\ a' \in \mathcal{A}\}$, where $P_{a^\ell}(a')$ is the relative frequency of the letter $a' \in \mathcal{A}$ in the vector $a^\ell$. Given $\delta > 0$, let us denote the set of all $\delta$-typical sequences of length $\ell$ by $T^\delta_{P_A}$, or by $T^\delta_A$ (if there is no ambiguity regarding the PMF that governs $A$), i.e., $T^\delta_A$ is the set of the sequences $a^\ell \in \mathcal{A}^\ell$ such that

$$(1-\delta)P_A(a') \le P_{a^\ell}(a') \le (1+\delta)P_A(a') \tag{45}$$

for every $a' \in \mathcal{A}$. For sufficiently large $\ell$, the size of $T^\delta_A$ is well–known [4] to be bounded by

$$2^{\ell[(1-\delta)H(A)-\delta]} \le |T^\delta_A| \le 2^{\ell(1+\delta)H(A)}. \tag{46}$$

It is also well–known (by the weak law of large numbers) that:

$$\Pr\left\{A^\ell \notin T^\delta_A\right\} \le \delta \tag{47}$$

for all $\ell$ sufficiently large. For a given generic channel $P_{B|A}(b|a)$ and for each $a^\ell \in T^\delta_A$, the set of all sequences $b^l$ that are jointly $\delta$-typical with $a^\ell$, will be denoted by $T^\delta_{P_{B|A}}(a^\ell)$, or by $T^\delta_{B|A}(a^\ell)$ if there is no ambiguity, i.e., $T^\delta_{B|A}(a^\ell)$ is the set of all $b^\ell$ such that:

$$(1-\delta)P_{a^\ell}(a')P_{B|A}(b'|a') \le P_{a^\ell b^\ell}(a', b') \le (1+\delta)P_{a^\ell}(a')P_{B|A}(b'|a'), \tag{48}$$

for all $a' \in \mathcal{A}, b' \in \mathcal{B}$, where $P_{a^\ell b^\ell}(a', b')$ denotes the fraction of occurrences of the pair $(a', b')$ in $(a^\ell, b^\ell)$. Similarly as in eq. (46), for all sufficiently large $\ell$ and $a^\ell \in T^\delta_A$, the size of $T^\delta_{B|A}(a^\ell)$ is bounded as follows:

$$2^{\ell[(1-\delta)H(B|A)-\delta]} \le |T^\delta_{B|A}(a^\ell)| \le 2^{\ell(1+\delta)H(B|A)}. \tag{49}$$

Finally, observe that for all $a^\ell \in T_A^\delta$ and $b^\ell \in T_{B|A}^\delta(a^\ell)$, the distortion $d(a^\ell, b^\ell) = \sum_{j=1}^{\ell} d(a_j, b_j)$ is upper bounded by:

$$d(a^\ell, b^\ell) \leq \ell(1+\delta)^2 \sum_{a',b'} P_A(a')P_{B|A}(b'|a')d(a',b') \stackrel{\Delta}{=} \ell(1+\delta)^2 Ed(A,B). \tag{50}$$

Let $(K, X, V, Y, Z)$ be a given random vector that satisfies the conditions of Theorem 4. We now describe the mechanisms of random code selection and the encoding and decoding operations. For a given $\epsilon > 0$, fix $\delta$ such that $2\delta + \max\{2 \cdot \exp\{-2^{n\delta}\} + 2^{-n\delta}, \delta^2\} \leq \epsilon$. Define also

$$\epsilon_1 \stackrel{\Delta}{=} \delta[1 + H(V|K) + H(V|K,X)], \tag{51}$$

$$\epsilon_2 \stackrel{\Delta}{=} \delta[1 + H(Y|K,V) + H(Y|K,X,V)], \tag{52}$$

and

$$\epsilon_3 \stackrel{\Delta}{=} \delta[1 + H(V|K) + H(V|Z,K)]. \tag{53}$$

*Generation of a rate–distortion code*:

Apply the type–covering lemma [4] and construct a rate–distortion codebook that covers $T_U^\delta$ within distortion $N(D'+\epsilon)$ w.r.t. $d'$, using $2^{NR_U(D')}$ codewords.

*Generation of the encrypting bitstream*:

For every $k^n \in T_K^\delta$, randomly select an index in the set $\{0, 1, \ldots, 2^{n[H(K|Y)+\delta]} - 1\}$ with a uniform distribution. Denote by $s^J(k^n) = (s_1(k^n), \ldots, s_J(k^n))$, $s_j(k^n) \in \{0,1\}$, $j = 1, \ldots, J$, the binary string of length $J = n[H(K|Y)+\delta]$ that represents this index. (Note that $s^J(k^n)$ can be interpreted as the output of the Slepian–Wolf encoder for $K^n$, where $Y^n$ plays the role of SI at the decoder [15].)

*Generation of an auxiliary embedding code*:

We first construct an auxiliary code capable of embedding $2^{NR_U(D')}$ watermarks by a random selection technique. First, $M_1 = 2^{nR_1}$, $R_1 = I(V;Z|K) - \epsilon_3 - \delta$, sequences $\{V^n(i,k^n)\}$, $i \in \{1, \ldots, M_1\}$, are drawn independently from $T_{V|K}^\delta(k^n)$ for every $k^n \in T_K^\delta$. For every such $k^n$, let us denote the set of these sequences by $\mathcal{C}(k^n)$. The elements of $\mathcal{C}(k^n)$ are evenly distributed among $M_U \stackrel{\Delta}{=} 2^{NR_U(D')}$ bins, each of size $M_2 = 2^{nR_2}$, $R_2 = I(X;V|K) + \epsilon_1 + \delta$

(this is possible thanks to condition (c) of Theorem 4, provided that the inequality therein is strict). A different (encrypted) message of length $L = NR_U(D') = n\lambda R_U(D')$ bits is attached to each bin, identifying a sub-code that represents this message. We denote the codewords in bin number $m$ ($m \in \{1, 2, \ldots, M_U\}$), by $\{V^n(m, j, k^n)\}$, $j \in \{1, 2, \ldots, M_2\}$.

*Stegotext sequence generation*:

For each auxiliary sequence (in the above auxiliary codebook of each $\delta$–typical $k^n$), $V^n(m, j, k^n) = v^n$, a set of $M_3 \stackrel{\Delta}{=} 2^{nR_3}$, $R_3 = I(X; Y|V, K) + \epsilon_2 + \delta$, stegotext sequences $\{Y^n(j', v^n, k^n)\}$, $j' \in \{1, \ldots, M_3\}$, are independently drawn from $T^{\delta}_{Y|VK}(v^n, k^n)$. We denote this set by $\mathcal{C}(v^n, k^n)$.

*Encoding*:

Upon receiving a triple $(u^N, x^n, k^n)$, the encoder acts as follows:

1. If $u^N \in T^{\delta}_U$, let $w^L = (w_1, \ldots, w_L)$, $w_i \in \{0, 1\}$, $i = 1, \ldots, L$ be the binary representation of the index of the rate–distortion codeword for the message source. For $k^n \in T^{\delta}_K$, let $s^J(k^n) = (s_1(k^n), \ldots, s_J(k^n))$ denote binary representation string of the index of $k^n$. Let $\tilde{w}^L = (\tilde{w}_1, \ldots, \tilde{w}_L)$, where $\tilde{w}_j = w_j \oplus s_j(k^n)$, $j = 1, \ldots, J$, and $\tilde{w}_j = w_j$, $j = J+1, \ldots, L$, and where $\oplus$ denotes modulo 2 addition i.e., the XOR operation.[7] The binary vector $\tilde{w}^L$ is the (partially) encrypted message to be embedded. Let $m = \sum_{l=1}^{L} \tilde{w}_l 2^{l-1} + 1$ denote the index of this message. If $u^N \notin T^{\delta}_U$ or $k^n \notin T^{\delta}_K$, an arbitrary (error) message $\tilde{w}^L$ is generated (say, the all–zero message).

2. If $(k^n, x^n) \in T^{\delta}_{KX}$ find, in bin number $m$, the first $j$ such that $V^n(m, j, k^n) = v^n$ is jointly typical, i.e., $(k^n, x^n, v^n) \in T^{\delta}_{KXV}$, and then find the first $j'$ such that $Y^n(j', v^n, k^n) = y^n \in \mathcal{C}(v^n, k^n)$ is jointly typical, i.e., $(k^n, x^n, v^n, y^n) \in T^{\delta}_{KXVY}$. This vector $y^n$ is chosen for transmission. If $(k^n, x^n) \notin T^{\delta}_{KX}$, or if there is no $V^n(m, j, k^n) = v^n$ and $Y^n(j', v^n, k^n) = y^n$ such that $(k^n, x^n, v^n, y^n) \in T^{\delta}_{KXVY}$, an arbitrary vector $y^n \in \mathcal{Y}^n$ is transmitted.

*Decoding*:

Upon receiving $Z^n = z^n$ and $K^n = k^n$, the decoder finds all sequences $\{v^n\}$ in $\mathcal{C}(k^n)$ such that $(k^n, v^n, z^n) \in T^{\delta}_{KVZ}$. If all $\{v^n\}$ found belong to the same bin, say, $\hat{m}$, then $\hat{m}$ is decoded

[7] Note that since $H(K)$ is assumed smaller than $\lambda R_U(D')$, then so is $H(K|Y)$, and therefore $J \leq L$.

as the embedded message, and then the binary representation vector $\hat{w}^L = (\hat{w}_1, \ldots, \hat{w}_L)$ corresponding to $\hat{m}$ is decrypted, again, by modulo 2 addition of its first $J$ bits with $s^J(k^n)$. This decrypted binary $L$–vector is then mapped to the corresponding reproduction vector $\tilde{u}^N$ of the rate–distortion codebook for the message source. If there is no $v^n \in \mathcal{C}(k^n)$ such that $(k^n, v^n, z^n) \in T^\delta_{KVZ}$ or if there exist two or more bins that contain such a sequence, an error is declared.

We now turn to the performance analysis of this code in all relevant aspects. For each triple $(k^n, x^n, u^N)$ and particular choices of the codes, the possible causes for incorrect watermark decoding are the following:

1. $(k^n, x^n, u^N) \notin T^\delta_{KX} \times T^\delta_U$. Let the probability of this event be defined as $P_{e_1}$.

2. $(k^n, x^n, u^N) \in T^\delta_{KX} \times T^\delta_U$, but in bin no. $m$ there is no $v^n$ s.t. $(k^n, x^n, v^n) \in T^\delta_{KXV}$. Let the probability of this event be defined as $P_{e_2}$.

3. $(k^n, x^n, u^N) \in T^\delta_{KX} \times T^\delta_U$ and in bin no. $m$ there is $v^n$ s.t. $(k^n, x^n, v^n) \in T^\delta_{KXV}$, but there is no $y^n \in \mathcal{C}(v^n, k^n)$ s.t. $(k^n, x^n, v^n, y^n) \in T^\delta_{KXVY}$. Let the probability of this event be defined as $P_{e_3}$.

4. $(k^n, x^n, u^N) \in T^\delta_{KX} \times T^\delta_U$ and in bin no. $m$ there is $v^n$ and $y^n \in \mathcal{C}(v^n, k^n)$ such that $(k^n, x^n, v^n, y^n) \in T^\delta_{KXVY}$, but $(k^n, v^n, z^n) \notin T^\delta_{KVZ}$. Let the probability of this event be defined as $P_{e_4}$.

5. $(k^n, x^n, u^N) \in T^\delta_{KX} \times T^\delta_U$ and in bin no. $m$ there is $v^n$ and $y^n \in \mathcal{C}(v^n, k^n)$ such that $(k^n, x^n, v^n, y^n) \in T^\delta_{KXVY}$, and $(k^n, v^n, z^n) \in T^\delta_{KVZ}$, but there exists another bin, say, no. $\tilde{m}$, that contains $\tilde{v}^n$ s.t. $(k^n, \tilde{v}^n, z^n) \in T^\delta_{KVZ}$. Let the probability of this event be defined as $P_{e_5}$.

If none of these events occur, the message $\tilde{w}^L$ (or, equivalently, $m$) is decoded correctly from $z^n$, the distortion constraint between $x^n$ and $y^n$ is within $n(D + \epsilon)$ (as follows from (50)), and the distortion between $u^N$ and its rate–distortion codeword, $\tilde{u}^N = \hat{u}^N$, does not exceed $N(D' + \epsilon)$. Thus, requirements 1 and 4 (modified according to eq. (6), with $D' + \epsilon$ replacing $D'$) are both satisfied. Therefore, we first prove that the probability for none of the events 1–5 to occur, tends to unity as $n \to \infty$.

The average probability of error $P_e$ in decoding $m$ is bounded by

$$P_e \leq \sum_{i=1}^{5} P_{e_i}. \tag{54}$$

The fact that $P_{e_1} \to 0$ follows immediately from (47). As for $P_{e_2}$, we have:

$$P_{e_2} \stackrel{\Delta}{=} \prod_{j=1}^{M_2} \Pr\{(k^n, x^n, V^n(m,j,k^n)) \notin T^{\delta}_{KXV}\}. \tag{55}$$

Now, by (46), for every $j$ and every $(k^n, x^n) \in T^{\delta}_{KX}$:

$$\begin{aligned}
\Pr\{V^n(m,j,k^n) \notin T^{\delta}_{V|KX}(k^n,x^n)\} &= 1 - \Pr\{V^n(m,j,k^n) \in T^{\delta}_{V|KX}(k^n,x^n)\} \\
&= 1 - \frac{|T^{\delta}_{V|KX}(k^n,x^n)|}{|T^{\delta}_{V|K}(k^n)|} \\
&\leq 1 - \frac{2^{n[(1-\delta)H(V|K,X)-\delta]}}{2^{n(1+\delta)H(V|K)}} \\
&= 1 - 2^{-n[I(X;V|K)+\epsilon_1]}.
\end{aligned} \tag{56}$$

Substitution of (56) into (55) provides us with the following upper bound:

$$P_{e_2} \leq \left[1 - 2^{-n[I(X;V|K)+\epsilon_1]}\right]^{M_2} \leq \exp\left\{ - 2^{nR_2} \cdot 2^{-n[I(X;V|K)+\epsilon_1]} \right\} \to 0, \tag{57}$$

double–exponentially rapidly since $R_2 = I(X;V|K) + \epsilon_1 + \delta$. To estimate $P_{e_3}$, we repeat the same technique:

$$P_{e_3} \stackrel{\Delta}{=} \prod_{j'=1}^{M_3} \Pr\{(k^n, x^n, v^n, Y^n(j', v^n, k^n)) \notin T^{\delta}_{KXVY}\}. \tag{58}$$

Again, by the property of the typical sequences, for every $j'$ and $(k^n, x^n, v^n) \in T^{\delta}_{KXV}$:

$$\Pr\{Y^n(j', v^n, k^n) \notin T^{\delta}_{Y|KXV}(k^n, x^n, v^n)\} \leq 1 - 2^{-n[I(X;Y|V,K)+\epsilon_2]}, \tag{59}$$

and therefore, substitution of (59) into (58) gives

$$P_{e_3} \leq \left[1 - 2^{-n[I(X;Y|V,K)+\epsilon_2]}\right]^{M_3} \leq \exp\left\{ - 2^{nR_3} \cdot 2^{-n[I(X;Y|V,K)+\epsilon_2]} \right\} \to 0, \tag{60}$$

double–exponentially rapidly since $R_3 = I(X;Y|V,K)+\epsilon_2+\delta$. The estimation of $P_{e_4}$ is again based on properties of typical sequences. Since $Z^n$ is the output of a memoryless channel $P_{Z|Y}$ with input $y^n = Y^n(j', v^n, k^n)$ and by the assumption of this step $(k^n, x^n, v^n, y^n) \in T^{\delta}_{KXVY}$, from (47) and the Markov lemma [3, Lemma 14.8.1], we obtain

$$P_{e_4} = \Pr\{(k^n, x^n, v^n, y^n, Z^n) \notin T^{\delta}_{KXVYZ}\} \leq \delta, \tag{61}$$

and similarly to $P_{e_1}$, $P_{e_4}$ can be made as small as desired by an appropriate choice of $\delta$.

Finally, we estimate $P_{e_5}$ as follows:

$$\begin{aligned} P_{e_5} &= \Pr\{\exists \tilde{m} \neq m : (k^n, V^n(\tilde{m}, j, k^n), z^n) \in T^{\delta}_{KVZ}\} \quad (62) \\ &\leq \sum_{\tilde{m}\neq m,\ j\in\{1,2,\ldots,M_2\}} \Pr\{(k^n, V^n(\tilde{m}, j, k^n), z^n) \in T^{\delta}_{KVZ}\} \\ &= (2^{NR_U(D')} - 1)2^{nR_2} \Pr\{(k^n, V^n(\tilde{m}, j, k^n), z^n) \in T^{\delta}_{KVZ}\} \\ &\leq 2^{nR_1} 2^{-n[I(V;Z|K)-\epsilon_3]}. \quad (63) \end{aligned}$$

Now, since $R_1 = I(V;Z|K) - \epsilon_3 - \delta$, $P_{e_5} \to 0$. Since $P_{e_i} \to 0$ for $i = 1, \ldots, 5$, their sum tends to zero as well, implying that there exist at least one choice of an auxiliary code and related stegotext codes that give rise to the reliable decoding of $\tilde{W}^L$.

Now, let us denote by $N_c$ the total number of composite sequences in a codebook that corresponds to a $\delta$–typical $k^n$. Then,

$$\begin{aligned} N_c &= M_U \cdot M_2 \cdot M_3 \\ &= 2^{n[\lambda R_U(D') + I(X;V|K) + I(X;Y|V,K) + \epsilon_1 + \epsilon_2 + 2\delta]} \\ &= 2^{n[\lambda R_U(D') + I(X;Y,V|K) + \epsilon_1 + \epsilon_2 + 2\delta]}. \quad (64) \end{aligned}$$

Thus,

$$\begin{aligned} H(Y^n|K^n) &\leq \log N_c \\ &= n[\lambda R_U(D') + I(X;Y,V|K) + \epsilon_1 + \epsilon_2 + 2\delta] \\ &\leq n(R'_c + \epsilon_1 + \epsilon_2 + 2\delta), \quad (65) \end{aligned}$$

where in the last inequality we have used condition (e). For sufficiently small values of $\delta$ (and hence of $\epsilon_1$ and $\epsilon_2$) $\epsilon_1 + \epsilon_2 + 2\delta \leq \epsilon$ and so, the compressibility requirement in the presence of $K^n$ is satisfied.

We next prove the achievability of $R_c$. Let us consider the set of $\delta$–typical key sequences $T^{\delta}_K$, and view it as the union of 0–typical sets (i.e., $\delta$–typical sets with $\delta = 0$), $\{T^0_{Q_K}\}$, where $Q_K$ exhausts the set of all rational PMF's with denominator $n$, and with the property

$$(1 - \delta)P_K(k) \leq Q_K(k) \leq (1 + \delta)P_K(k), \quad \forall k \in \mathcal{K}. \quad (66)$$

Suppose that we have already randomly selected a codebook for one *representative* member $\hat{k}^n$ of each type class $T^0_{Q_K} \subset T^{\delta}_K$ using the mechanism described above. Now, consider the

set of all permutations from $\hat{k}^n$ to every other member of $T^0_{Q_K}$. The auxiliary codebook and the stegotext codebooks for every other key sequence, $k^n \in T^0_{Q_K}$ will be obtained by permuting all (auxiliary and stegotext) codewords of those corresponding to $\hat{k}^n$ according to the same permutation that leads from $\hat{k}^n$ to $k^n$ (thus preserving all the necessary joint typicality properties). Now, in the *union* of all stegotext codebooks, corresponding to all typical key sequences, each codeword will appear at least $(n+1)^{-|\mathcal{K}|\cdot|\mathcal{Y}|} \cdot 2^{n[(1-\delta)H(K|Y)-\delta]}$ times, which is a lower bound to the number of permutations of $\hat{k}^n$ which leave a given stegotext codeword $y^n$ unaltered. The total number of stegotext codewords, $N_Y$, in all codebooks of all $\delta$–typical key sequences (including repetitions) is upper bounded by

$$\begin{aligned} N_Y &= |T^\delta_K| \cdot N_c \\ &\leq 2^{n[(1+\delta)H(K)+\delta]} \cdot 2^{n[\lambda R_U(D')+I(X;Y,V|K)+\epsilon_1+\epsilon_2+2\delta]} \\ &= 2^{n[H(K)+\lambda R_U(D')+I(X;Y,V|K)+\epsilon_1+\epsilon_2+\delta(H(K)+3)]}. \end{aligned} \tag{67}$$

Let $\mathcal{C}$ denote the union of all stegotext codebooks, namely, the set of all *distinct* stegotext vectors across all codebooks corresponding to all $k^n \in T^\delta_K$, and let $N(y^n)$ denote the number of occurrences of a given vector $y^n \in \mathcal{Y}^n$ in all stegotext codebooks. Then, in view of the above combinatorial consideration, we have

$$N_Y = \sum_{y^n \in \mathcal{C}} N(y^n) \geq |\mathcal{C}| \cdot (n+1)^{-|\mathcal{K}|\cdot|\mathcal{Y}|} \cdot 2^{n[(1-\delta)H(K|Y)-\delta]}. \tag{68}$$

Combining eqs. (67) and (68), we have

$$\log|\mathcal{C}| \leq n[\lambda R_U(D') + I(X;Y,V|K) + I(K;Y) + \delta'], \tag{69}$$

where

$$\delta' = \epsilon_1 + \epsilon_2 + \delta(H(K) + H(K|Y) + 4) + |\mathcal{K}| \cdot |\mathcal{Y}| \cdot \frac{\log(n+1)}{n}, \tag{70}$$

which is arbitrarily small provided that $\delta$ is sufficiently small and $n$ is sufficiently large. Thus, the rate required for public compression of $Y^n$ (without the key), which is $(\log|\mathcal{C}|)/n$, is arbitrarily close to $[\lambda R_U(D_1) + I(X;Y,V|K) + I(K;Y)]$, which in turn is upper bounded by $R_c$, by condition (d) of Theorem 4.

Before we proceed to evaluate the equivocation levels, an important comment is in order in the context of public compression (and a similar comment will apply to private compression): Note that a straightforward (and not necessary optimal) method for public

compression of $Y^n$ is simply according to its index within $T_Y^\delta$, which requires about $nH(Y)$ bits. On the other hand, the converse theorem tells us that the compressed representation of $Y^n$ cannot be much shorter than $n[\lambda R_U(D') + I(X;Y,V|K) + I(K;Y)]$ bits (cf. the necessity of condition (d) of Theorem 4). Thus, contradiction between these two facts is avoided only if

$$\lambda R_U(D') + I(X;Y,V|K) + I(K;Y) \leq H(Y), \tag{71}$$

or, equivalently,

$$\lambda R_U(D') + I(X;Y,V|K) \leq H(Y|K). \tag{72}$$

This means that any achievable point $(D, D', R_c, R_c', h, h')$ corresponds to a choice of random variables $(K, X, Y, V)$ that must inherently satisfy eq. (72). This observation will now help us also in estimating the equivocation levels.

Consider first the equivocation w.r.t. the reproduction, for which we have the following chain of inequalities:

$$\begin{aligned}
Nh' &\leq nH(K|Y) && (73)\\
&= nH(K) - nI(K;Y)\\
&= H(K^n) - nI(K;Y) && (74)\\
&= H(K^n|Y^n,Z^n) + I(K^n;Y^n,Z^n) - nI(K;Y)\\
&= H(K^n|Y^n,Z^n) + I(K^n;Y^n) - nI(K;Y) && (75)\\
&= H(K^n|Y^n,Z^n) + H(Y^n) - H(Y^n|K^n) - nI(K;Y)\\
&\leq H(K^n|Y^n,Z^n) + n[\lambda R_U(D') + I(X;Y,V|K) + I(K;Y) + \epsilon] -\\
&\quad -n[\lambda R_U(D'+\epsilon) + I(X;Y,V|K) - \epsilon] - nI(K;Y) && (76)\\
&= H(K^n|Y^n,Z^n) + n\lambda[R_U(D') - R_U(D'+\epsilon)] + n\epsilon\\
&\stackrel{\Delta}{=} H(K^n|Y^n,Z^n) + n\epsilon'\\
&= I(K^n;\hat{U}^N|Y^n,Z^n) + H(K^n|Y^n,Z^n,\hat{U}^N) + n\epsilon'\\
&\leq H(\hat{U}^N|Y^n,Z^n) + H(K^n|Y^n,Z^n,\hat{U}^N) + n\epsilon' && (77)
\end{aligned}$$

where (73) is based on condition (b), (74) is due to the memorylessness of $K^n$, (75) follows from the fact that $K^n \to Y^n \to Z^n$ is a Markov chain, (76) is due to the sufficiency of condition (d) (that we have just proved) and the necessity of condition (e), and $\epsilon'$ vanishes as $\epsilon \to 0$ due to the continuity of $R_U(\cdot)$. Comparing the left–most side and the right–most

side of the above chain of inequalities, we see that to prove that $H(\hat{U}^N|Y^n, Z^n)$ is essentially at least as large as $Nh'$, it remains to show that $H(K^n|Y^n, Z^n, \hat{U}^N)$ is small, say,

$$H(K^n|Y^n, Z^n, \hat{U}^N) \leq n\epsilon' \tag{78}$$

for large $n$. We next focus then on the proof of eq. (78).

First, consider the following chain of inequalities:

$$\begin{aligned} H(K^n|Y^n, Z^n, \hat{U}^N) &\leq H(K^n, S^J(K^n)|Y^n, Z^n, \hat{U}^N) \\ &= H(S^J(K^n)|Y^n, Z^n, \hat{U}^N) + H(K^n|S^J(K^n), Y^n, Z^n, \hat{U}^N) \\ &\leq H(S^J(K^n)|Y^n, \hat{U}^N, W^L) + H(K^n|S^J(K^n), Y^n), \end{aligned} \tag{79}$$

where the second inequality follows from the fact that $W^L$ is function of $\hat{U}^N$ and the fact that conditioning reduces entropy. As for the second term of the right–most side, we have by Fano's inequality

$$H(K^n|S^J(K^n), Y^n) \leq 1 + P_{\rm err} \cdot n \log |\mathcal{K}| \leq n\epsilon'/2 \quad \text{for large enough } n, \tag{80}$$

as $P_{\rm err} \to 0$ is the probability of error associated with the Slepian–Wolf decoder that estimates $K^n$ from its compressed version, $S^J(K^n)$, and the "side information," $Y^n$. As for the first term of the right–most side of (79), we have

$$\begin{aligned} H(S^J(K^n)|Y^n, \hat{U}^N, W^L) &= H(W^L \oplus \tilde{W}^L|Y^n, \hat{U}^N, W^L) \\ &\leq H(\tilde{W}^L|Y^n). \end{aligned} \tag{81}$$

It remains to show that $H(\tilde{W}^L|Y^n) \leq n\epsilon'/2$ as well. In order to show this, we have to demonstrate that for a good code, once $Y^n$ is given, there is very little uncertainty with regard to $\tilde{W}^L$, which is the index of the bin.

To this end, let us suppose that the inequality in (72) is strict (otherwise, we can slightly increase the allowable distortion level $D'$ and thus reduce $R_U(D')$). As we prove in the Appendix, for any given (arbitrarily small) $\gamma > 0$,

$$\Pr\{\exists\, y^n \text{ in the code of } \hat{k}^n \text{ that appears in more than } 2^{n\gamma} \text{ bins}\} \leq |\mathcal{Y}|^n 2^{-(n\gamma - \log e)2^{n\gamma}}, \tag{82}$$

that is, a double–exponential decay. The probability of the union of these events across all representatives $\{\hat{k}^n\}$ of all $T^0_{Q_K} \subset T^\delta_K$ will just be multiplied by the number of $\{T^0_{Q_K}\}$ in

$T_K^\delta$, which is polynomial, and hence will continue to decay double–exponentially. Let us define then the event

$$\{\exists\ y^n \text{ in the stego–codebook of some } \hat{k}^n \text{ that appears in more than } 2^{n\gamma} \text{ bins}\}$$

as yet another error event (like the error events 1–5) that occurs with very small probability. Assume then, that the randomly selected codebook is "good" in the sense that no stegovector appears in more than $2^{n\gamma}$ bins, for any of the representatives $\{\hat{k}^n\}$. Now, given $y^n$, how many candidate bins (corresponding to encrypted messages $\{\tilde{w}^L\}$) can be expected at most? For a given $y^n$, let us confine attention to the $\delta$–conditional type class $T^\delta_{K|Y}(y^n)$ (key sequences outside this set cannot have $y^n$ in their codebooks, as they are not jointly $\delta$–typical with $y^n$). The conditional $\delta$–type class $T^\delta_{K|Y}(y^n)$ can be partitioned into conditional 0–type classes $\{T^0_{Q_{K|Y}}(y^n)\}$, where $Q_{K|Y}$ exhausts the allowed $\delta$–tolerance in the conditional distribution around $P_{K|Y}$, in the same spirit as before. Now, take an arbitrary representative $\tilde{k}^n$ from a given $T^0_{Q_{K|Y}}(y^n)$, and consider the set of all permutations that lead from $\tilde{k}^n$ to all other members $\{k^n\}$ of $T^0_{Q_{K|Y}}(y^n)$. Obviously, the stego–codebooks of all those $\{k^n\}$ have exactly the same configuration of occurrences of $y^n$ as that of $\tilde{k}^n$ (since these permutations leave $y^n$ unaltered), therefore they belong to exactly the same bins as in the codebook of $\tilde{k}^n$, the number of which is at most $2^{\gamma n}$, by the hypothesis that we are using a good code. In other words, as $k^n$ scans $T^0_{Q_{K|Y}}(y^n)$, there will be no new bins that contain $y^n$ relative to those that are already in the codebook of $\tilde{k}^n$. New bins that contain $y^n$ can be seen then only by scanning the other conditional 0–types $\{T^0_{Q_{K|Y}}(y^n)\}$ within $T^\delta_{K|Y}(y^n)$, but the number such conditional 0–types does not exceed the total number of conditional 0–types, which is upper bounded, in turn, by $(n+1)^{|\mathcal{K}|\cdot|\mathcal{Y}|}$ [4]. Thus, the totality of stego–codebooks, for all relevant $\{k^n\}$ cannot give more than $(n+1)^{|\mathcal{K}|\cdot|\mathcal{Y}|}\cdot 2^{n\gamma}$ distinct bins altogether. In other words, for a good codebook:

$$H(\tilde{W}^L|Y^n) \leq \log[(n+1)^{|\mathcal{K}|\cdot|\mathcal{Y}|}\cdot 2^{n\gamma}] = n\left[\gamma + |\mathcal{K}|\cdot|\mathcal{Y}|\cdot\frac{\log(n+1)}{n}\right] \tag{83}$$

which is less than $n\epsilon'/2$ for an appropriate choice of $\gamma$ and for large enough $n$.

Finally, for the equivocation w.r.t. the original message source, we have the following:

$$\begin{aligned} H(U^N|Y^n,Z^n) &= H(\hat{U}^N|Y^n,Z^n) + H(U^N|Y^n,Z^n) - H(\hat{U}^N|Y^n,Z^n) \\ &\geq nH(K|Y) - 2n\epsilon' + H(U^N|Y^n,Z^n) - H(\hat{U}^N|Y^n,Z^n) \\ &= nH(K|Y) + H(U^N) - I(U^N;\hat{U}^N) - I(U^N;Y^n,Z^n) - \\ &\quad H(\hat{U}^N|U^N) + I(\hat{U}^N;Y^n,Z^n) - 2n\epsilon' \\ &\geq nH(K|Y) + H(U^N) - H(\hat{U}^N) - I(U^N;Y^n,Z^n) - \\ &\quad H(\hat{U}^N|U^N) + I(\hat{U}^N;Y^n,Z^n) - 2n\epsilon' \\ &\geq nH(K|Y) + NH(U) - NR_U(D') - 2\epsilon'] - \\ &\quad [I(U^N;Y^n,Z^n) + H(\hat{U}^N|U^N) - I(\hat{U}^N;Y^n,Z^n)], \end{aligned} \tag{84}$$

where first inequality is due to the fact that $H(\hat{U}^N|Y^n,Z^n) \geq n[H(K|Y) - 2\epsilon']$, that we have just shown, and the third is due to the memorylessness of $\{U_i\}$ and the fact that the rate–distortion codebook size is $2^{NR_U(D')}$ and so, $H(\hat{U}^N) \leq NR_U(D')$. Now, the second bracketed expression on the right–most side is the same as in eq. (33), where in the case of this specific scheme, both inequalities in (33) become equalities, i.e., this expression vanishes. This is because in our scheme, $U^N \to \hat{U}^N \to (Y^n,Z^n)$ is a Markov chain (and so, the first inequality of (33) is tight) and because $H(\hat{U}^N|U^N,Y^n,Z^n) \leq H(\hat{U}^N|U^N) = 0$ (as $\hat{U}^N$ is a deterministic function of $U^N$), which makes the second inequality of (33) tight. As a result, we have

$$\begin{aligned} H(U^N|Y^n,Z^n) &\geq N[H(K|Y)/\lambda + H(U) - R_U(D') - 2\epsilon'/\lambda] \\ &\geq N[h + R_U(D') - H(U) + H(U) - R_U(D') - 2\epsilon/\lambda] \\ &= N(h - 2\epsilon'/\lambda), \end{aligned} \tag{85}$$

where we have used condition (a). This completes the proof of the direct part.

# 7 Conclusion and Future Research

We have analyzed optimum tradeoffs between several figures of merit pertaining to the performance of a system that combines watermarking, compression and encryption. We have also characterized the (high–level) structure of codes that asymptotically achieve the performance limits under various degrees of generality of the underlying assumptions.

To summarize, the main ideas that were developed in this work, both in the general level and in the technical level, are the following:

1. The separation priniciple falls apart once an attack channel is introduced. In particular, the fact that the cryptographic key plays the additional role of side information is intesresting phenomenom. Moreover, as more generalilty is added into the model, the separation principle 'collapses' in steps: First, the encryption and the embedding become coupled, but the rate–distortion compression is still separate, and then, in another step of enhancing the generality, the rate–distortion code becomes coupled as well with the other parts of the encoder.

2. In many problems where the separation theorem fails (e.g., in network situations, or when there is dependence between the source and the channel), there are no closed–form single–letter expressions for the achievable region, and optimal coding schemes are not known. The situation in this paper is different: even when separation fails, still, single–letter expressions are available (cf. Theorems 3, 4) and asymptotically optimum coding schemes are offerred (at least in the random coding sense).

3. There are interesting tradeoffs with regard to the desired degree of statistical dependency between the key and the stegotext.

4. A Slepian–Wolf encoder is harnessed in order to extract purely random bits for encryption, which are independent of $Y^n$, in order to circumvent the problem that $Y^n$ is not yet available in the encryption stage.

5. The security of $\hat{U}^N$ is taken into account as an additional criterion.

6. The security of $\hat{U}^N$ is enhanced by using extra key bits to control the choice of the rate–distortion code, by using the fact that there are about $2^{NH(\hat{U}|U)}$ *distinct* codebooks.

7. The random selection of a codebook is carried out for only one "representative" $\hat{k}^n$ in each type class, and then the codebook for every other $k^n$ in the same type class is constructed by permuting the codevectors according to the permutations that lead from $\hat{k}^n$ to $k^n$. This idea proves useful both in estabilishing the achievability of the public compression rate and in proving the achievability of the desired security.

A few leftover problems, to be considered in future work, are the following:

1. In view of the findings of this work, it would be desirable to conduct a more thorough investigation and to gain understanding with regard to conditions under which the separation principle holds here, and in more general frameworks. In particular, it would be interesting to identify all the factors in the system that affect the validity of the separation principle. In this paper, we identified only one such factor – the presence of a non–trivial (memoryless) attack channel.

2. Replacing the present secrecy metric by a stronger one (referring to the discussion in the Introduction).

3. Relaxing the assumption that the channel from $X^n$ to $K^n$ is memoryless.

4. Taking into account requirements on the secrecy of the covertext (in addition or instead of the secrecy of the watermark and its reconstruction).

## Acknowledgements

The author would like to thank Dr. Yossi Steinberg for interesting discussions. Useful comments made by the anonymous referees are acknowledged with thanks.

## Appendix

*Proof of eq. (82).* The probability of obtaining $y^n$ in a single random selection within the codebook of $\hat{k}^n$ is given by

$$\begin{aligned} \Pr\{Y^n(j', V^n(m, j, \hat{k}^n), \hat{k}^n) = y^n\} &= \frac{|T^{\delta}_{V|KY}(k^n, y^n)|}{|T^{\delta}_{V|K}(k^n)|} \cdot \frac{1}{|T^{\delta}_{Y|KV}(k^n, v^n)|} \qquad \text{(A.1)} \\ &\leq \frac{2^{n(1+\delta)H(V|K,Y)}}{2^{n[(1-\delta)H(V|K)-\delta]}} \cdot \frac{1}{2^{n[(1-\delta)H(Y|K,V)-\delta]}} \\ &= 2^{-n[H(Y|K)-\delta'']}, \qquad \text{(A.2)} \end{aligned}$$

where the first factor in the right–hand side of (A.1) is the probability of having a $V^n(m, j, \hat{k}^n) = v^n$ that is typical with $y^n$ and $\hat{k}^n$ (a necessary condition for this $v^n$ to generate the given $y^n$), the second factor is the probability of selecting a given $y^n$ in the random selection of the steogtext code, and where

$$\delta'' = \delta[H(V|K, Y) + H(V|K) + H(Y|K, V) + 2]. \qquad \text{(A.3)}$$

It now follows that the probability $q$ for at least one occurrence of $y^n$ among the stegowords corresponding to a certain bin, in the codebook of $\hat{k}^n$, is upper bounded (using the union bound) by

$$\begin{aligned} q &\leq M_2 \cdot M_3 \cdot 2^{-n[H(Y|K)-\delta'']} \\ &= 2^{-n[H(Y|K)-I(X;V|K)-I(X;Y|V,K)-\delta''-2\delta-\epsilon_1-\epsilon_2]} \\ &= 2^{-n[H(Y|K)-I(X;V,Y|K)-\delta''-2\delta-\epsilon_1-\epsilon_2]} \\ &\stackrel{\Delta}{=} 2^{-n[H(Y|K)-I(X;Y,V|K)-\delta_1]}. \end{aligned} \tag{A.4}$$

We are interested to upper bound the probability that a given $y^n$ appears as a stegoword in more than $2^{n\gamma}$ bins in the codebook of $\hat{k}^n$, for a given $\gamma > 0$. For $i = 1, \ldots, M_U$, let $A_i \in \{0, 1\}$ be the indicator function of the event

$$\{y^n \text{appears as a stegoword in bin no. } i \text{ at least once}\}.$$

Then, clearly $\{A_i\}$ are i.i.d. with $\Pr\{A_i = 1\} = q$. Therefore,

$$\begin{aligned} \Pr\left\{\sum_{i=1}^{M_U} A_i \geq 2^{n\gamma}\right\} &\leq \exp_2\left\{-M_U D\left(\frac{2^{n\gamma}}{M_U}\|q\right)\right\} \\ &= \exp_2\left\{-M_U D\left(2^{-n[\lambda R_U(D')-\gamma]}\|q\right)\right\}, \end{aligned} \tag{A.5}$$

where for $\alpha, \beta \in [0, 1]$, the function $D(\alpha\|\beta)$ designates the binary divergence

$$D(\alpha\|\beta) = \alpha \log \frac{\alpha}{\beta} + (1-\alpha) \log \frac{1-\alpha}{1-\beta}. \tag{A.6}$$

Now, referring to eq. (72), suppose that

$$H(Y|K) \geq \lambda R_U(D') + I(X; V, Y|K) + \delta_1 + 2\gamma. \tag{A.7}$$

Then, clearly,

$$2^{-n[\lambda R_U(D')-\gamma]} > 2^{-n[H(Y|K)-I(X;Y,V|K)-\delta_1]} \geq q \tag{A.8}$$

and so, $\Pr\{\sum_{i=1}^{M_U} A_i \geq 2^{n\gamma}\}$ is further upper bounded by

$$\Pr\left\{\sum_{i=1}^{M_U} A_i \geq 2^{n\gamma}\right\} \leq \exp_2\left\{-M_U D\left(2^{-n[\lambda R_U(D')-\gamma]}\|2^{-n[H(Y|K)-I(X;Y,V|K)-\delta_1]}\right)\right\}. \tag{A.9}$$

To further bound this expression from above, we have to get a lower bound to an expression of the form $D(e^{-na}\|e^{-nb})$ for $0 < a < b$. Applying the inequality $\log(1+x) = -\log(1-$

$\frac{x}{1+x}) \geq \frac{x \log e}{1+x}$, for $x > -1$, we have:

$$\begin{aligned} D(2^{-na} \| 2^{-nb}) &= 2^{-na} \log \frac{2^{-na}}{2^{-nb}} + (1 - 2^{-na}) \log \frac{1 - 2^{-na}}{1 - 2^{-nb}} \\ &= n(b-a)2^{-na} + (1 - 2^{-na}) \log \left( 1 + \frac{2^{-nb} - 2^{-na}}{1 - 2^{-nb}} \right) \\ &\geq n(b-a)2^{-na} + (2^{-nb} - 2^{-na}) \log e \\ &\geq [n(b-a) - \log e] 2^{-na}. \end{aligned} \tag{A.10}$$

Applying this inequality with $a = \lambda R_U(D') - \gamma$ and $b = H(Y|K) - I(X;Y,V|K) - \delta_1$, we get

$$D\left(2^{-n[\lambda R_U(D') - \gamma]} \| 2^{-n[H(Y|K) - I(X;Y,V|K) - \delta_1]}\right) \geq (n\gamma - \log e) 2^{-n[\lambda R_U(D') - \gamma]} \tag{A.11}$$

and so,

$$\Pr\left\{ \sum_{i=1}^{M_U} A_i \geq 2^{n\gamma} \right\} \leq 2^{-(n\gamma - \log e) 2^{n\gamma}}, \tag{A.12}$$

which decays double–exponentially rapidly with $n$. While, this inequality holds for a *given* $y^n$, the probability that $\sum_{i=1}^{M_U} A_i \geq 2^{n\gamma}$ for *some* $y^n \in \mathcal{Y}^n$ would be upper bounded, using the union bound, by $|\mathcal{Y}|^n \cdot 2^{-(n\gamma - \log e) 2^{n\gamma}}$, which still decays double–exponentially. Thus, with very high probability the random selection of stegovectors, for $\hat{k}^n$, is such that no stego codevector $y^n$ appears in more than $2^{n\gamma}$ bins.

# References


[1] A. Adelsbach, S. Katzenbeisser, and A.-R. Sadeghi, "Cryptography meets watermarking: detecting watermarks with minimal or zero knowledge disclosure," preprint 2002. Available on–line at [www-krypt.cs.uni-sb.de/download/papers]

[2] S. C. Cheung and D. K. W. Chiu, "A watermark infrastructure for enterprise document management," *Proc. 36th Hawaii International Conference on System Sciences (HICSS'03)*, Hawaii, 2003.

[3] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, Wiley, New York, 1991.

[4] I. Csiszár and J. Körner, *Information Theory: Coding Theorems for Discrete Memoryless Systems*, Academic Press, 1981.

[5] S. I. Gel'fand and M. S. Pinsker, "Coding for channel with random parameters," *Problems of Information and Control*, vol. 9, no. 1, pp. 19-31, 1980.

[6] A. Jayawardena, B. Murison, and P. Lenders, "Embedding multiresolution binary images into multiresolution watermark channels in wavelet domain," preprint 2000. Available on–line at [www.tsi.enst.fr/∼maitre/tatouage/icassp00/articles].

[7] M. Johnson, P. Ishwar, V. Prabhakan, D. Schonberg, and K. Ramchandran, "On compressing encrypted data," *IEEE Trans. Signal Processing*, vol. 52, no. 10, pp. 2992–3006, October 2004.

[8] K. Kuroda, M. Nishigaki, M. Soga, A. Takubo, and I. Nakamura, "A digital watermark using public–key cryptography for open algorithm," *Proc. ICITA 2002*. Also, available on–line at [http://charybdis.mit.csu.edu.au/∼mantolov/CD/ICITA2002/papers/131-21.pdf].

[9] A. Maor and N. Merhav, "On joint information embedding and lossy compression," *IEEE Trans. Inform. Theory*, vol. 51, no. 8, pp. 2998–3008, August 2005.

[10] A. Maor and N. Merhav, "On joint information embedding and lossy compression in the presence of a stationary memoryless attack channel," *IEEE Trans. Inform. Theory*, vol. 51, no. 9, pp. 3166–3175, September 2005.

[11] N. Merhav, "On the Shannon cipher system with a capacity–limited key–distribution channel," submitted to *IEEE Trans. Inform. Theory*, May 2005.

[12] N. Merhav and S. Shamai (Shitz), "On joint source–channel coding for the Wyner–Ziv source and the Gel'fand–Pinsker channel," *IEEE Trans. Inform. Theory*, vol. 49, no. 11, pp. 2844–2855, November 2003.

[13] P. Moulin and J. A. O'Sullivan, "Information–theoretic analysis of information hiding," *IEEE Trans. Inform. Theory*, vol. 49, no. 3, pp. 563–593, March 2003.

[14] P. Moulin and Y. Wang, "New results on steganographic capacity," *Proc. CISS 2004*, pp. 813–818, Princeton University, March 2004.

[15] D. Slepian and J. K. Wolf, "Noiseless coding of correlated information sources," *IEEE Trans. Inform. Theory*, vol. IT–19, pp. 471–480, 1973.

[16] M. Steinder, S. Iren, and P. D. Amer, "Progressively authenticated image transmission," preprint 1999. Available on–line at [www.cis.udel.edu /amer/PEL/poc/pdf/milcom99-steiner.pdf].

[17] Y. Steinberg and S. Verdú, "Channel simulation and coding with side information," *IEEE Trans. Inform. Theory*, vol. IT–40, no. 3, pp. 634–646, May 1994.

[18] A. D. Wyner, "The wire-tap channel," *Bell System Technical Journal*, vol. 54, no. 8, pp. 1355–1387, Oct. 1975.

[19] H. Yamamoto, "Rate–distortion theory for the Shannon cipher system," *IEEE Trans. Inform. Theory*, vol. 43, no. 3, pp. 827–835, May 1997.

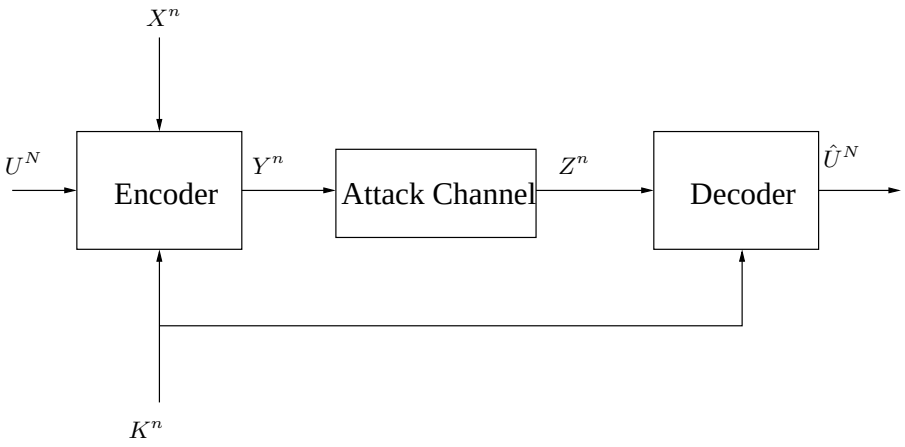


Figure 1: A generic watermarking/encryption system.

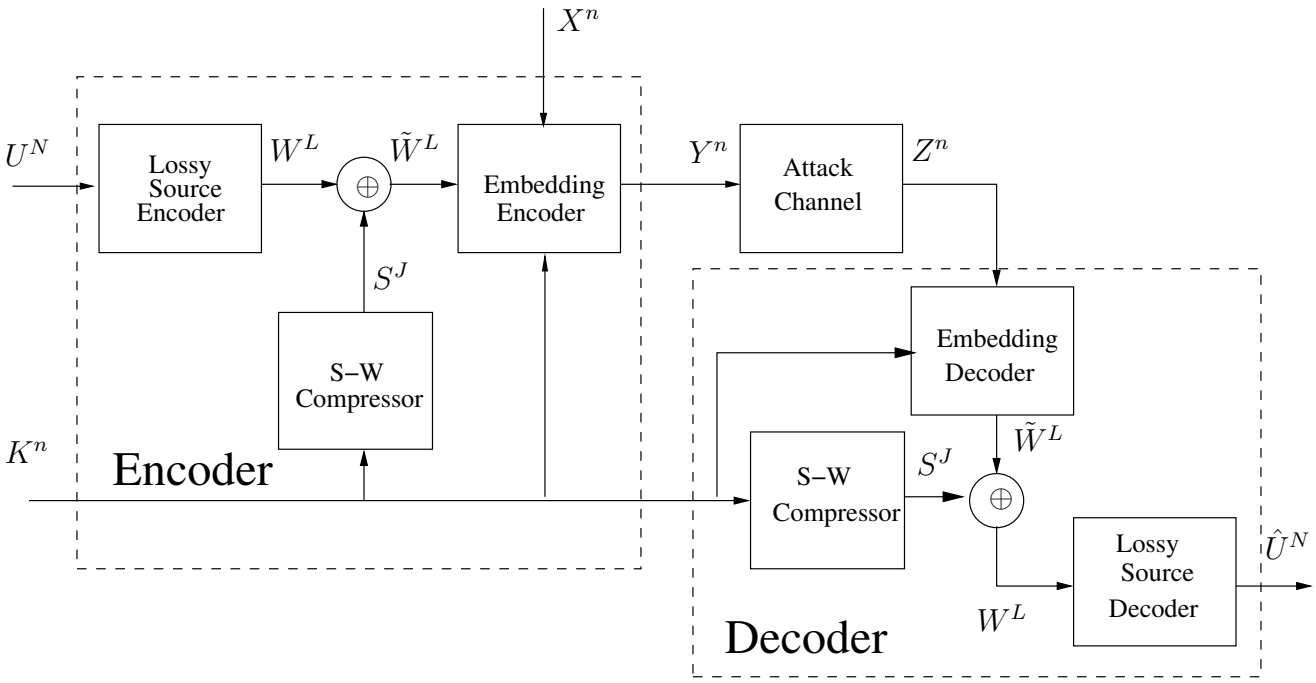


Figure 2: The proposed watermarking/encryption scheme (general case).

Neri Merhav - Biography

**Neri Merhav** (S'86–M'87–SM'93–F'99) was born in Haifa, Israel, on March 16, 1957. He received the B.Sc., M.Sc., and D.Sc. degrees from the Technion, Israel Institute of Technology, in 1982, 1985, and 1988, respectively, all in electrical engineering.

From 1988 to 1990 he was with AT&T Bell Laboratories, Murray Hill, NJ, USA. Since 1990 he has been with the Electrical Engineering Department of the Technion, where he is

now the Irving Shepard Professor. During 1994–2000 he was also serving as a consultant to the Hewlett–Packard Laboratories – Israel (HPL-I). His research interests include information theory, statistical communications, and statistical signal processing. He is especially interested in the areas of lossless/lossy source coding and prediction/filtering, relationships between information theory and statistics, detection, estimation, and Shannon Theory, including topics in joint source–channel coding, source/channel simulation, and coding with side information with applications to information hiding and watermarking systems.

Dr. Merhav was a co-recipient of the 1993 Paper Award of the IEEE Information Theory Society and he is a Fellow of the IEEE since 1999. He also received the 1994 American Technion Society Award for Academic Excellence and the 2002 Technion Henry Taub Prize for Excellence in Research. From 1996 until 1999 he served as an Associate Editor for Source Coding of the IEEE TRANSACTIONS ON INFORMATION THEORY. He also served as a co–chairman of the Program Committee of the 2001 IEEE International Symposium on Information Theory. He is currently on the Editorial Board of FOUNDATIONS AND TRENDS IN COMMUNICATIONS AND INFORMATION THEORY.

## List of Figures

1. Fig.1: A generic watermarking/encryption system.
2. Fig 2: The proposed watermarking/encryption scheme (general case).

## List of Footnotes

1. This idea of the double role of the secret key has been explored also in [7] in the context of compression of encrypted data.
2. In fact, the choice of the conditional distribution $P(K^n|X^n)$ is a degree of freedom that can be optimized subject to the given randomness resources.
3. The assumption of independence between $\{K_t\}$ and $\{X_t\}$ is temporary and made now primarily for the sake of simplicity of the exposition. It will be dropped later on.
4. At the end of Section 4 (after Theorem 4), we discuss the case where this limitation (or its analogue in lossy reconstruction of $U^N$) is dropped.

5. These randomness resources are, in fact, purely random, independent bits, which can now be redefined as our secret "key", in the original meaning of this term.

6. Note that now there is a distinction between the required available randomness rate, which is $H(K^n|X^n)/n = H(K|X)$ [17], and the rate at which the key must be conveyed to the legitimate decoder, which remains $H(K)$ (as the decoder has no access to $X^n$). This is in contrast to the case of independence between $K^n$ and $X^n$, where these two parameters coincide.

7. Note that since $H(K)$ is assumed smaller than $\lambda R_U(D')$, then so is $H(K|Y)$, and therefore $J \leq L$.