Steganographic Capacity Estimation for the Statistical Restoration Framework

Anindya Sarkar[†], Kenneth Sullivan[††] and B. S. Manjunath[†]

†Department of Electrical and Computer Engineering,
University of California,
Santa Barbara, CA 93106
††Mayachitra Inc.,
5266 Hollister Avenue,
Santa Barbara, CA 93111

ABSTRACT

In this paper we attempt to quantify the "active" steganographic capacity - the maximum rate at which data can be hidden, and correctly decoded, in a multimedia cover subject to noise/attack (hence - active), perceptual distortion criteria, and statistical steganalysis. Though work has been done in studying the capacity of data hiding as well as the rate of perfectly secure data hiding in noiseless channels, only very recently have all the constraints been considered together. In this work, we seek to provide practical estimates of steganographic capacity in natural images, undergoing realistic attacks, and using data hiding methods available today.

We focus here on the capacity of an image data hiding channel characterized by the use of statistical restoration to satisfy the constraint of perfect security (under an i.i.d. assumption), as well as JPEG and JPEG-2000 attacks. Specifically we provide experimental results of the statistically secure hiding capacity on a set of several hundred images for hiding in a pre-selected band of frequencies, using the discrete cosine and wavelet transforms, where a perturbation of the quantized transform domain terms by ±1 using the quantization index modulation scheme, is considered to be perceptually transparent. Statistical security is with respect to the matching of marginal statistics of the quantized transform domain terms.

Keywords: channel capacity, earth mover's distance, statistical restoration, optimal hiding fraction, optimal redundancy

1. INTRODUCTION

Steganography is the art of secure communication where the very existence of the communication cannot be detected while steganalysis is the art of detecting the presence of the secret communication. The steganographer has two conflicting requirements - he has to imperceptibly embed a certain amount of data in an innocuous looking host signal (the *cover*), and also ensure that there is minimal statistical difference between the cover and the *stego* (signal containing hidden data). For a practical steganographic system, the hiding has to satisfy perceptual, statistical and attack constraints.

Of these various constraints, the statistical constraint is the most difficult to satisfy for a practical system - there are excellent blind steganalysis methods[1–4] that are able to detect most of the modern-day steganographic schemes. To completely characterize the statistics of images is as yet an unsolved problem. To simplify our task, we focus here on the statistical security under the commonly studied criteria of matching marginal distributions (i.e. first-order histograms). In practice, a steganographic system should be able to withstand a variety of attacks and our goal is to develop a general approach that allows for arbitrary attack channels. To demonstrate the approach we provide here results for a specific attack. In this work, we provide *an end-to-end framework where given an image, an attack channel, and common measures on statistical security, its steganographic capacity can be computed.*

Further author information: (Send correspondence to Anindya Sarkar)
Anindya Sarkar: E-mail: anindya@ece.ucsb.edu, Telephone: 1-805-893-5682

The capacity of data hiding channels with respect to attack has been a well-studied theoretical problem;[5] the rate of perfectly secure data hiding in noiseless channels[6,7] has also been investigated. However, putting all the constraints together as required by a practical steganographic scheme, has only been studied recently (e.g. Wang et al[8]).

In our past work on statistical restoration,[9,10] we presented a steganographic scheme where the first order probability mass function (PMF) of the block-based quantized discrete cosine transform (DCT) coefficients, lying in a certain frequency band and with magnitudes less than a certain threshold, was statistically restored. In,[7] we had obtained a secure hiding rate for the quantization index modulation (QIM) scheme,[11] with the cover signals being generated from arbitrary distributions. Recently, we developed an expression for the maximum allowable hiding fraction[12] for the statistical restoration of first and higher order of co-occurrence statistics. In this paper, the statistical constraint is satisfied by using this hiding fraction in the statistical restoration framework.

2. PROBLEM FORMULATION

Fridrich et al[6] have defined "steganographic capacity" as follows - for a host signal, it is the maximal message length that can be embedded without producing perceptually or statistically detectable distortions. We consider the problem of *active* steganography, in which the adversary can also attack the stego signal (JPEG and JPEG-2000 based compression attacks for our case) - thus introducing an attack constraint. Let X denote the cover signal, S the stego signal, and Y the received signal (all in the same transform domain) at the decoder after attack. The problem of finding the capacity of such active warden stegosystems for an i.i.d. cover signal reduces to maximizing the embedding rate with the following constraints:

Perceptual constraint. The perceptual distortion between the original and stego images in the transform domain should not exceed a certain maximum amount, D_1, for some perceptual distance measure $d(\cdot,\cdot)$. Thus, we must have $d(X,S) \leq D_1$. Distortion constraints for limiting the perceptual distortion have long been used in the information-theoretic analysis of the data hiding problem ([5,11,13]).

Statistical constraint. The embedding process should not modify the statistics of the host signal more than a very small number, ϵ, for some statistical distance measure. Cachin[14] proposed the use of Kullback-Leibler (K-L) divergence for defining the statistical security of steganography. Thus, denoting the cover and stego distributions by P_X and P_S, respectively, the statistical constraint can be given as, $\mathcal{D}(P_X||P_S) \leq \epsilon$, where $\mathcal{D}(\cdot,\cdot)$ is K-L distance measure. For perfect security, we aim at obtaining $P_X = P_S$. For the blind steganalysis schemes[1–4] , there are various other statistical features that are used for detection - e.g. second order features, individual band histograms. In,[12] analysis has been provided for finding the hiding fraction for second-order co-occurrence statistics and for individual frequency bands. With the addition of more constraints, the system will become statistically more secure and the capacity will decrease. However, in the context of this paper, only the first order histogram for the entire hiding band is considered for statistical security.

Attack constraint. The embedded data must be recoverable after the stego signal has undergone an attack distortion of at most D_2, i.e. if $d(S,Y) \leq D_2$.

Let us now present our interpretation of these constraints:

1) The **perceptual distance constraint** can be interpreted as the mean square error between the transform domain feature matrices, before and after hiding. For DCT and discrete wavelet transform (DWT) domains, it is assumed that a distortion of ±1 for the coefficients in the chosen frequency band ensures perceptual transparency of the stego signal - with a proper choice of frequency domain terms for hiding, perceptual transparency for the DCT domain is shown in our prior work.[15,16] Instead of specifying a fixed value for the distortion constraint D_1, our aim is to minimize the perceptual distortion during hiding while ensuring that the other constraints are maintained.

2) **Statistical constraint** can be interpreted as the K-L distance between the first order DCT histograms, before and after hiding. Our aim is to ensure $P_X = P_S$, which ensures that the distance between the PMFs is zero, irrespective of the distance metric (K-L or otherwise) used.

3) For varying the severity of the compression attack, the experiment is repeated for various combinations of quality factors (for data embedding and for generating the compressed image) and compression rates, for JPEG and JPEG-2000 based attacks, respectively. For the **attack constraint**, we do not fix D_2 but our aim is to maximize the hiding rate, by using the minimum redundancy during error correction coding, which ensures zero bit error rate (BER) at the decoder, even after compression attacks. We assume the attack/distortion channel is fixed, and thus the sender can simulate the exact attack and obtain the required redundancy.

2.1 Transform Domains used for Hiding

DCT domain: Hiding occurs in a select band of low and mid-frequency DCT coefficients, for every 8×8 block. After selecting a design quality factor QF_h for hiding, the DCT coefficients for the 8×8 block are divided element-wise by the 8×8 quantization matrix, corresponding to QF_h. The first 19 AC DCT terms, that occur during a zigzag scan, are used for hiding.

DWT domain: For the wavelet domain based scheme, we use the Haar wavelet as the basis function and use the periodic DWT mode, with 3 level decomposition, to obtain a 8×8 matrix of DWT coefficients, given a 8×8 pixel block. In[17] and,[18] a method is described to generate a 8×8 wavelet domain quantization matrix which corresponds to a certain design quality factor. It ensures that the embedding bitrate for DWT based hiding using the wavelet domain quantization matrix is similar to the embedding bitrate obtained using the corresponding quality factor for DCT based hiding. For choosing the frequency band for hiding, a modified scanning procedure is used, as described in[17] . The first 19 DWT terms (leaving the top leftmost LLL coefficient) that occur during the modified scanning procedure are used for embedding.

3. PROBLEM INTRODUCTION: STATISTICAL AND PERCEPTUAL CONSTRAINTS

Let the input feature set available for hiding be X. As per the statistical restoration framework, X is decomposed into two disjoint sets - H for hiding and C for compensation, as in (1). After hiding and compensation, the feature set thus obtained is Y (1). We divide the feature set into bins of unity width and find their respective bin-counts (number of terms per bin). Let $B_X(i)$ denote the number of elements in the i^{th} bin of X. The normalized bin-count is regarded as the probability mass function (PMF). H is changed to $\hat{H}$ after hiding and the objective is to modify C to $\hat{C}$ to ensure that the 1-D PMFs of X and Y, denoted by P_X and P_Y, respectively, are equal, as in (3). Perfect restoration is possible only if the required number of terms in every bin of $\hat{C}$ exceeds zero.

$$X = H \cup C,\ Y = \hat{H} \cup \hat{C},\ H \cap C = \phi,\ \hat{H} \cap \hat{C} = \phi \tag{1}$$

$$\hat{H} \cap \hat{C} = \phi \Rightarrow B_Y(i) = B_{\hat{H}}(i) + B_{\hat{C}}(i),\ \forall\ i \tag{2}$$

$$\text{To obtain } P_Y = P_X,\ \text{ we need } B_Y(i) = B_X(i)\ \Rightarrow\ B_{\hat{C}}(i) = \{B_X(i) - B_{\hat{H}}(i)\} \geq 0,\ \forall\ i \tag{3}$$

Due to QIM based hiding, the perturbation of the quantized transform domain terms in H is limited to [-1,1]. Our compensation framework also ensures that the perturbation to the compensation terms in C while modifying C to $\hat{C}$ is limited to [-1,1], as discussed later in Sec. 3.1. Hence, we do not explicitly include the $d(X,S) \leq D_1$ constraint in the formulation. For maximizing the embedded bitrate, under the perceptual and statistical constraints, the two objectives are:

- find the optimal hiding fraction $\lambda = \frac{|H|}{|X|}$, where $|X|$ denotes the cardinality of the set X
- find the optimal way to modify C to $\hat{C}$, once the optimal hiding fraction is determined.

In[7,12] we present an analysis for computing the optimum hiding fraction λ, which is briefly discussed in Sec. 3.2. The optimal λ indicates how much embedding can occur so that the PMF based statistical restoration can be performed. Once λ is known, we need an optimum way of obtaining Y ($Y = \hat{H} \cup \hat{C}$) from X ($X = H \cup C$) such that the perceptual distortion between X and Y, $d(X, Y)$, is minimized. The connection between the Earth Mover's Distance (EMD) and the statistical restoration framework where we wish to minimize $d(X, Y)$ is explained in.[19] Here, we present a brief overview of the use of EMD for statistical restoration.

3.1 Statistical Restoration and the Earth Mover's Distance

The EMD[20] between two PMFs is defined as the minimum "work" done in converting one PMF to the other. Here, *work* refers to the redistribution of weights among the various bins in the discrete distribution. EMD returns the optimal transportation flows among the bins. For statistical restoration, we have to convert the histogram B_C to $B_{\hat{C}}$, C and $\hat{C}$ being defined in (1), where the normalized histogram is the PMF.

Let S and T denote two signatures, each having M clusters. The weight of each cluster is the fraction of points it contains. Let the center for the k^{th} cluster of S be s_k while the ℓ^{th} cluster center of T is denoted by t_ℓ and the square Euclidean distance between them is called $d_{k\ell}$.

$$d_{k\ell} = (s_k - t_\ell)^2 \tag{4}$$

The EMD problem is "optimally" changing S (considered as the source) to make it as similar as possible to T (the target). For our problem, the source S is the PMF P_C of the compensation coefficients while the target T is $P_{\hat{C}}$, the PMF of $\hat{C}$. The weight of each bin is the PMF value for that bin. Our aim is to find a flow matrix $F = [f_{k\ell}]$, where $f_{k\ell}$ is the flow from the k^{th} bin of S to the ℓ^{th} bin of T that minimizes the total *work* done:

$$WORK(S, T, F) = \sum_{k=1}^{M} \sum_{\ell=1}^{M} d_{k\ell} f_{k\ell} \tag{5}$$

Thus, EMD gives precisely the optimum flows from the bins of C to $\hat{C}$ so as to match P_X to P_Y, where X and Y are defined in (1), under the minimum mean-squared error (MMSE) criterion. It is assumed that the $d(\cdot, \cdot)$ distance function, introduced in Sec. 2, under the perceptual distortion constraint, is the squared Euclidean distance. It has been shown in[21] that while matching of two 1-D signatures in the MMSE sense, the flows are always between two consecutive bins (flow $f_{k\ell} > 0$ iff $|k - \ell| \leq 1$) - hence, while modifying C to $\hat{C}$, as in (1), the absolute distortion for a quantized element in C does not exceed one.

3.2 Computing the Optimal Hiding Fraction for 1-D Histogram based Compensation

While computing the 1-D histograms for the transform domain (DCT/DWT) coefficients, we only consider those with magnitude less than a certain threshold T - the distribution of these coefficients is very peaky near zero and is very low for larger values. For a given T, there are $(2T + 1)$ bins from $[-T, T]$, and we optimally hide in all the bins, except the two extreme ones. For an input message having equal 0's and 1's and using dithering based scalar QIM for hiding, where the dither values are evenly spread in $[-0.5, 0.5]$, the number of terms in the i^{th} bin of $\hat{H}$, $B_{\hat{H}}(i)$ can be approximated as follows,[12] assuming knowledge of the B_X values for the different bins:

$$B_{\hat{H}}(i) \approx \frac{\lambda B_X(i)}{2} + \frac{\lambda B_X(i-1)}{4} + \frac{\lambda B_X(i+1)}{4} \tag{6}$$

for a hiding fraction of λ for all the bins. From (3) and (6), considering the i^{th} bin, λ needs to satisfy:

$$B_{\hat{H}}(i) \leq B_X(i) \Rightarrow \lambda \leq \left\{ \frac{B_X(i)}{\frac{B_X(i-1)}{4} + \frac{B_X(i)}{2} + \frac{B_X(i+1)}{4}} \right\}, \quad -T < i < T \tag{7}$$

$$\text{For ease of notation, we define } \lambda_i = \left\{ \frac{B_X(i)}{\frac{B_X(i-1)}{4} + \frac{B_X(i)}{2} + \frac{B_X(i+1)}{4}} \right\}, \quad -T < i < T \tag{8}$$

For the i^{th} bin, λ_i can be viewed as $\frac{B_X(i)}{B_{\hat{H}}(i)}$ where $B_{\hat{H}}(i)$ is computed using λ=1 in (6). The effective hiding fraction $\lambda^\star(T)$, for a given T, is the minimum of all these λ_i terms (since the hiding fraction $\lambda \leq \lambda_i, \forall i$, using (7) and (8)).

$$\lambda^\star(T) \quad = \quad \min_{-T<i<T} \{\lambda_i : \; \lambda_i > 0\}. \tag{9}$$

The condition ($\lambda_i > 0$) in (9) ensures that the hiding fraction will not be reduced to zero for bins with no elements. An extension of the 1-D case, for restoring higher order co-occurrence statistics in the transform domain, is also explained in.[12]

4. ESTIMATING THE OPTIMAL EMBEDDING RATE - ACCOUNTING FOR CHANNEL DISTORTIONS

In the proposed scheme, hiding is performed by modifying some image coefficients in a certain transform domain. Once the optimum hiding fraction and optimum redundancy factor to be used in the error correction framework are known, the maximum possible databits that can be embedded, while maintaining the perceptual, statistical and attack constraints, can be determined - the sequence of steps involved is shown in Fig. 1.

The overall system flow for DCT based hiding is briefly outlined in Fig. 2. For a generalized framework, the DCT coefficients can be replaced by any other transform domain suitable for hiding and the JPEG attack block, denoted by the $Z \rightarrow Z'$ mapping, can be substituted by any other distortion channel. The maximum allowable size of the message to be embedded, M, depends on $\lambda^\star$ and q_{opt}, the computation of which is shown in Fig. 1. The message, M, is first encoded to R using the turbo-like repeat-accumulate code[22] with redundancy q (RA-q), the data being hidden in a band of low-frequency DCT coefficients, having n elements per 8×8 block. Therefore, the effective number of embedded bits per block=$\frac{n}{q}$ (we use n=19 in our experiments). At the time of embedding, the code symbols corresponding to the DCT coefficients beyond a predetermined threshold (T=30 in our case), are *erased* at the encoder - erasures are denoted by e in Fig. 2 and in (10). It should be noted that the erasure rate is high if DCT elements, which equal zero after rounding, are erased.[15] Since DCT elements $> T$ are erased here, and DCT elements have a peaky PMF, peaking near zero, the erasure rate is small. The DCT elements in the range [-0.5,0.5] are not erased for that would not allow the cover and stego image PMFs to be exactly matched under the statistical restoration framework. Further errors are introduced due to the JPEG compression attack. The embedded data can still be recovered because of the added redundancy using RA codes.

We experimentally obtain the transition probability matrices from R to Z (*depends on the distribution of the image transform domain coefficients and the hiding method*) and from Z to Z' (*depends on the attack channel characteristics*). The mutual information $I(R, Z')$ between the input (R) and output (Z') terms in the probabilistic part of the channel is maximized to compute the capacity $\mathcal{C}_{channel}$ (10) for a given image and a given attack channel, which is then used to compute the minimum redundancy factor needed for perfect data recovery *if an ideal channel code were used* - it equals $\lceil 1/\mathcal{C}_{channel} \rceil$. This minimum redundancy factor is called q_{min} (11).

$$\mathcal{C}_{channel} = \max_{p(r)} \; I(R, Z') = \max_{p(r)} \sum_{r \in \{0,1\}, \; z' \in \{0,1,e\}} p(r, z') \log \left\{ \frac{p(r|z')}{p(r)} \right\} \tag{10}$$

$$q_{min} = 1/\mathcal{C}_{channel} \tag{11}$$

The capacity estimation framework is general enough to be applied for hiding in any other transform domain (affects mapping $R \rightarrow Z$) and for any other attacks (affects mapping $Z \rightarrow Z'$).

There are two parameters to be optimally estimated to maximize the embedding rate for zero BER - the hiding fraction λ and the code redundancy factor q. We separately optimize for λ and q - the estimation of the optimal λ is explained in[12] and in Sec. 3.2. Let the maximum embedding rate (of hidden data) obtained using an ideal channel code be $\mathcal{R}_{max}$ (12) while that practically obtained using the RA code, with optimal redundancy, is $\mathcal{R}_{prac}$ (14), both computed assuming knowledge of the optimal hiding fraction $\lambda^\star$ - the threshold T in (9) is

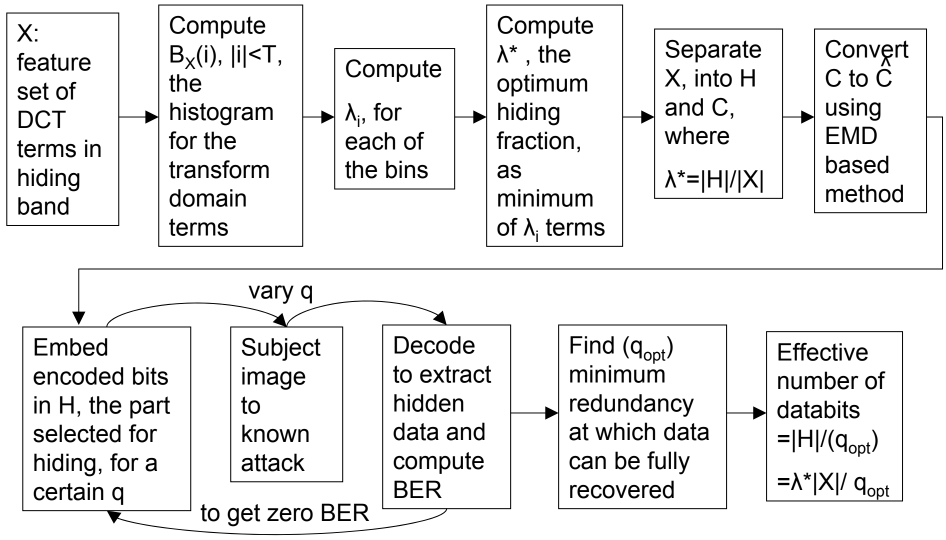

X: feature set of DCT terms in hiding band
Compute B_X(i), |i|<T, the histogram for the transform domain terms
Compute λ_i, for each of the bins
Compute λ* , the optimum hiding fraction, as minimum of λ_i terms
Separate X, into H and C, where
λ*=|H|/|X|
Convert C to Ĉ using EMD based method
vary q
Embed encoded bits in H, the part selected for hiding, for a certain q
Subject image to known attack
Decode to extract hidden data and compute BER
to get zero BER
Find (q_opt) minimum redundancy at which data can be fully recovered
Effective number of databits =|H|/(q_opt)
=λ*|X|/ q_opt

Figure 1. The framework to hide the maximum number of databits, maintaining the statistical (use $\lambda^\star$ as the hiding fraction) and attack constraints (use minimum redundancy that ensures zero BER). The redundancy factor used in the error correction code is denoted by q. The perceptual constraint is implicit in the QIM based hiding scheme. At first, $\lambda^\star$ is computed and then, after separating X into H and C based on $\lambda^\star$, C is converted to $\hat{C}$ using the EMD based formulation to maintain the statistical constraint. Also, the EMD based method ensures minimum perturbations to maintain the perceptual constraint. Then, data is embedded in H under varying q to obtain the optimum redundancy factor q_{opt}.

fixed at 30. The minimum redundancy factor, using RA codes, that produces zero BER for a given image and a given attack channel, called the optimum redundancy factor q_{opt}, is experimentally obtained and $\frac{1}{q_{opt}}$ is regarded as $\mathcal{R}_{undetectable}$ (13) - the rate achievable using a practical code without the statistical constraint. Using $\lambda^\star$% of the available terms for hiding, $\mathcal{R}_{prac}$ can be obtained from $\mathcal{R}_{undetectable}$ (14).

$$\mathcal{R}_{max} = \lambda^\star \cdot \mathcal{C}_{channel} \tag{12}$$

$$\mathcal{R}_{undetectable} = \frac{1}{q_{opt}} \tag{13}$$

$$\mathcal{R}_{prac} = \lambda^\star \cdot \mathcal{R}_{undetectable} \tag{14}$$

5. EXPERIMENTS AND RESULTS

We run the experiments for a variety of design quality factors and attack quality factors. The results are averaged over 500 images for each case. The hiding parameters used for the DCT and DWT based domains are discussed in Sec. 2.1.

The average value of the maximum embedding rate with an ideal channel code, $\mathcal{R}_{max}$ (12), is compared with the rate obtained using the RA code, $\mathcal{R}_{prac}$ (14), for the following transform domains and attack scenarios:

- DCT domain hiding, with JPEG attack
- DWT domain hiding, with JPEG attack
- DWT domain hiding, with JPEG-2000 attack

The hiding rates for both the DCT and DWT domains depend on the quality factor QF_h for hiding. As mentioned in Sec. 2.1, the quantization matrix of the DWT terms can be generated depending on QF_h. It has been shown[15] that when hiding occurs at a certain quality factor, the embedded data can be recovered after JPEG-based compression only if the attack quality factor, QF_a is the same or higher (less severe quantization)

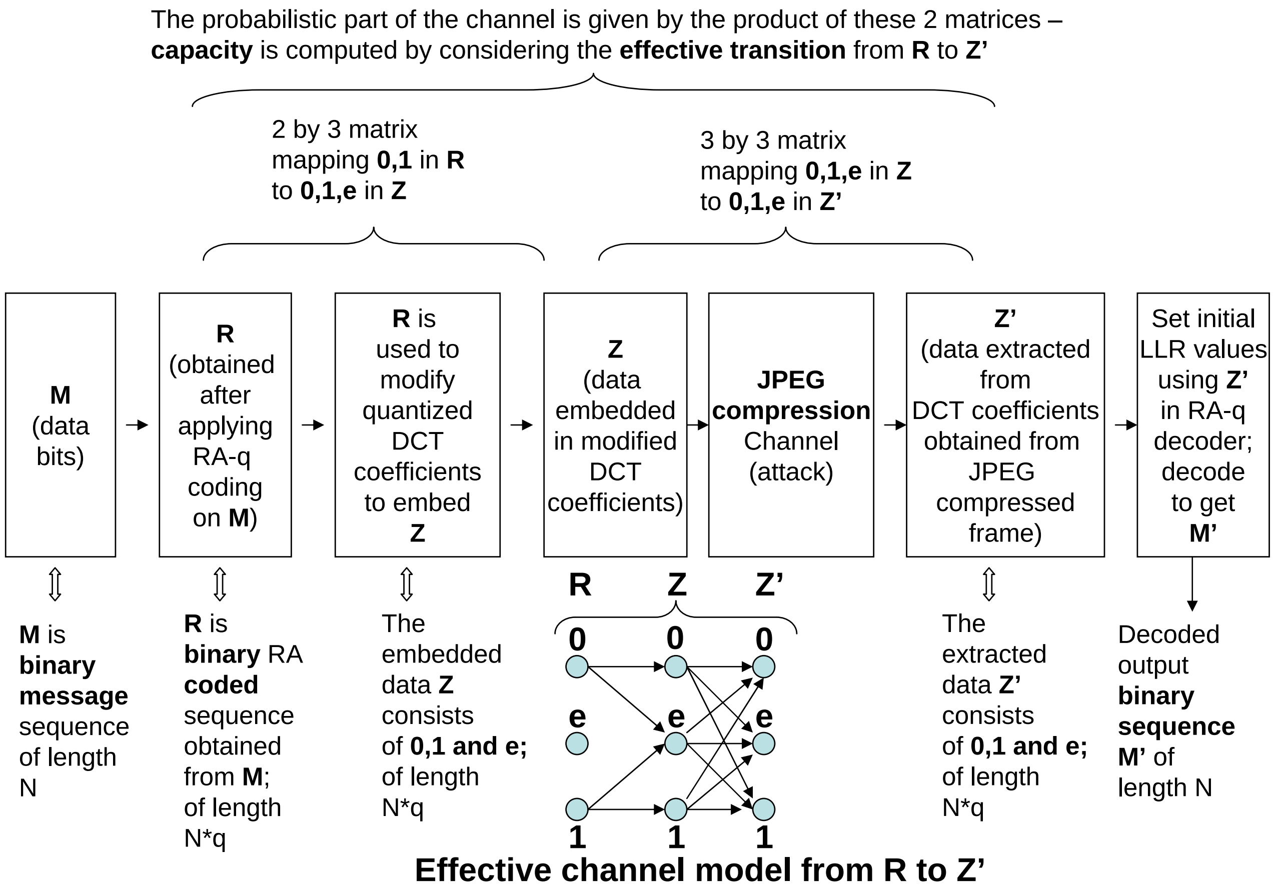
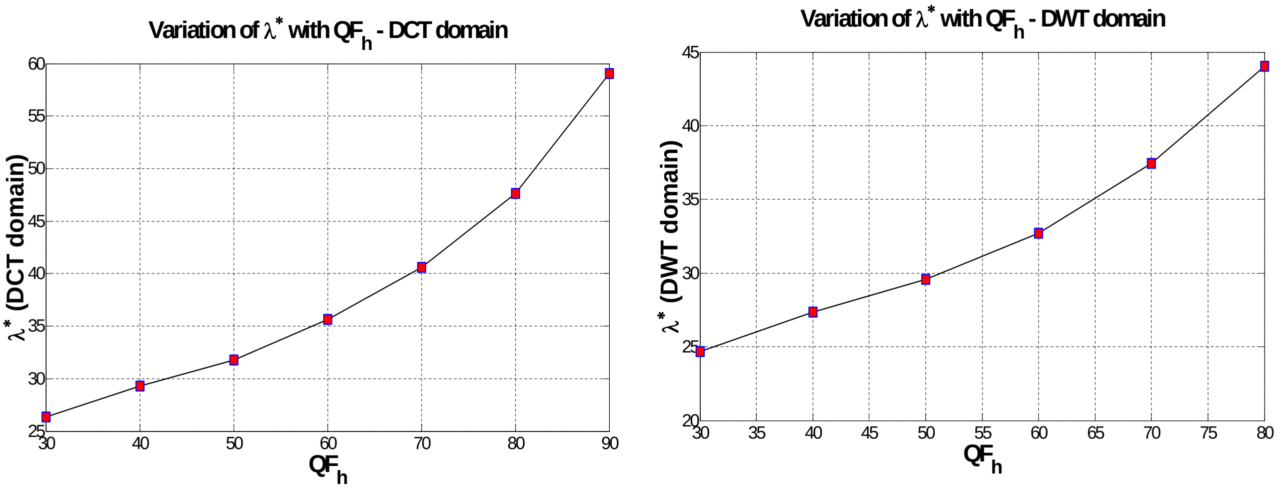

The probabilistic part of the channel is given by the product of these 2 matrices – **capacity** is computed by considering the **effective transition** from **R** to **Z’**
2 by 3 matrix mapping **0,1** in **R** to **0,1,e** in **Z**
3 by 3 matrix mapping **0,1,e** in **Z** to **0,1,e** in **Z’**
M (data bits)
R (obtained after applying RA-q coding on **M**)
R is used to modify quantized DCT coefficients to embed **Z**
Z (data embedded in modified DCT coefficients)
JPEG compression Channel (attack)
Z’ (data extracted from DCT coefficients obtained from JPEG compressed frame)
Set initial LLR values using **Z’** in RA-q decoder; decode to get **M’**
M is **binary message** sequence of length N
R is **binary** RA **coded** sequence obtained from **M**; of length N*q
The embedded data **Z** consists of **0,1 and e;** of length N*q
R
Z
Z’
0
0
0
e
e
e
1
1
1
The extracted data **Z’** consists of **0,1 and e;** of length N*q
Decoded output **binary sequence M’** of length N
Effective channel model from R to Z’

Figure 2. Computation of the data hiding capacity depends on the 2×3 transition probability matrix mapping R to Z' - shown here for DCT domain hiding and for JPEG compression attack (LLR = Log Likelihood Ratio)

than the design quality factor QF_h. When the quantization at the attack stage is more severe than that used while data embedding, the redundancy factor needed for successful data recovery is so high that it makes the effective hiding rate very small. For JPEG-2000 based compression, the compression ratio (CR) is the ratio between the number of bits needed to represent the compressed image and the number of bits that represent the original image - higher CR denotes less severe compression. The compression based attack is repeated for different values of CR.

Our numerical findings are summarized in the figures below. In Fig. 3, it is seen that as the design quality factor for hiding, QF_h, increases, the optimum hiding fraction, $\lambda^\star$ (9) increases. As the design quality factor QF_h decreases, the JPEG quantization matrix consists of larger valued terms (coarser quantization) - hence, the number of zero-valued DCT coefficients increases with a lower quality factor and coarser quantization. With a lower quality factor, the DCT PMF becomes more peaky near 0 - e.g. $B_X(0)$ becomes much greater than $B_X(1)$ and $B_X(-1)$. Using (8), the hiding fraction λ_i for $i = \{-1, 1\}$ becomes smaller with lower QF_h due to the dominance of $B_X(0)$ over $B_X(1)$ and $B_X(-1)$. Hence, $\lambda^\star$ (9), the minimum of the λ_i terms, decreases. $\lambda^\star$ depends on the PMF of the quantized DCT elements, which is determined by the JPEG quantization matrix corresponding to QF_h and not the severity of the attack (QF_a).

The rate $\mathcal{R}_{max}$ (12) is the product of $\lambda^\star$ and $\mathcal{C}_{channel}$. For a fixed QF_h, the variation of $\mathcal{R}_{max}$ with different attacks depends on $\mathcal{C}_{channel}$. In Fig. 4, it is seen that for DCT domain hiding, when the JPEG attack quality factor QF_a is varied, the rate is initially high at $QF_a = QF_h$, and then drops with increased QF_a before rising again. With increased QF_a, the JPEG quantization becomes finer and hence, due to less severe attacks, the channel capacity is expected to increase. This trend holds in general except when QF_a equals QF_h - i.e. the



(a)

(b)

Figure 3. Variation of the optimum hiding fraction, $\lambda^\star$ (%), with the design quality factor QF_h, for DCT and DWT domains.

attack is matched to the design quality factor for hiding. Hence, there is a valley in the $\mathcal{R}_{max}$ vs QF_a plots. The plot of $\mathcal{R}_{prac}$ follows the same trend as $\mathcal{R}_{max}$.

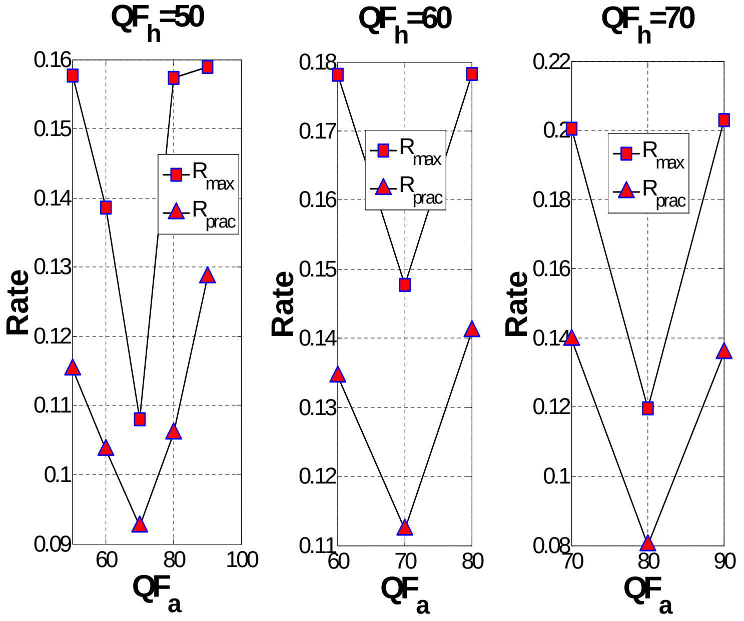


Figure 4. Variation of mean($\mathcal{R}_{max}$) and mean($\mathcal{R}_{prac}$) with the attack quality factor (QF_a) for JPEG attack, for different design quality factor QF_h, for DCT domain hiding.

However, when the hiding is in the DWT domain, the rate does not peak, even when QF_a is matched with QF_h. As QF_a is gradually increased from QF_h, the rate also increases as shown in Fig. 5.

When the hiding is in the DWT domain and the attack is JPEG-2000 based compression, the rate increases with less severe compression, as shown in Fig. 6. With increased CR, the JPEG-2000 based attack is less severe and hence, the channel capacity and the rate increases.

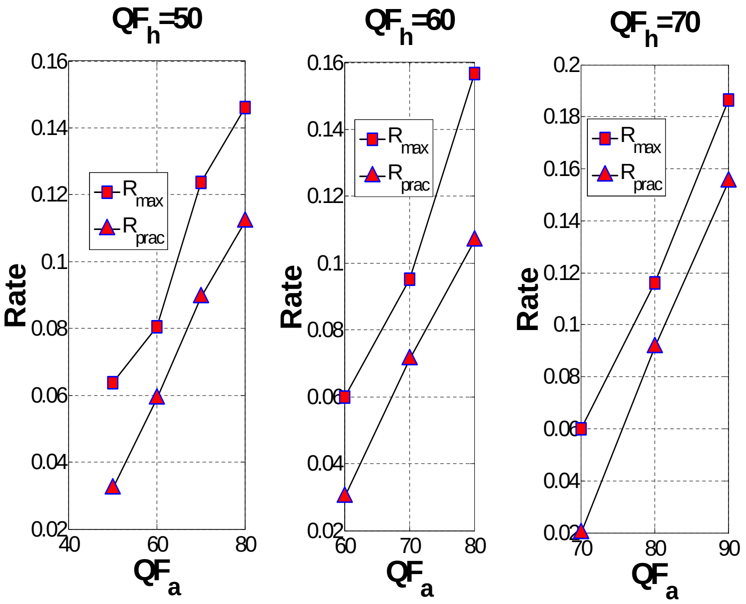


Figure 5. Variation of mean($\mathcal{R}_{max}$) and mean($\mathcal{R}_{prac}$) with the attack quality factor (QF_a) for JPEG attack, for different design quality factor QF_h, for DWT domain hiding.

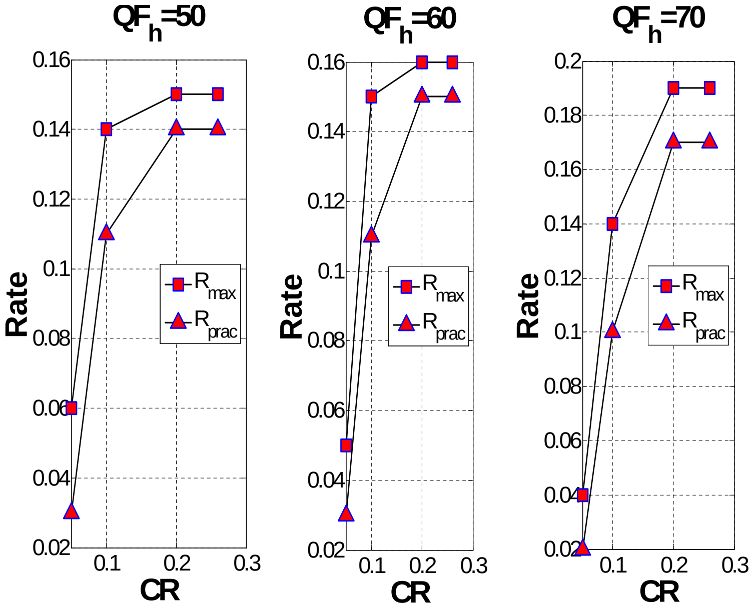


Figure 6. Variation of mean($\mathcal{R}_{max}$) and mean($\mathcal{R}_{prac}$) with the compression ratio (CR) for JPEG-2000 based attack, for different design quality factor QF_h, for DWT domain hiding.

6. CONCLUSION

We have examined here the maximum rate in which a steganographer can embed data subject to three constraints: visual inspection, statistical steganalysis, and error-free transmission despite distortions or attacks between sender and receiver. We have taken an approach that practically satisfies these criteria for a given distortion channel, statistical steganalysis constraint, and perceptual criteria.

Here we have focused on a pair of specific channels (JPEG and JPEG 2000 compression), Cachin's ϵ-divergence statistical steganalysis criteria,[14] and a basic distortion measure, however our approach is not limited to these. From experiments on a diverse set of natural images, we have calculated the maximum rate of zero-divergence hiding subject to compression. Both the maximum theoretical rate and that achievable by modern error correcting codes are found. In many practical cases, the rate is high enough to provide acceptable communication. From these experiments on the compression channels we are able to determine the factors that effect the maximum rate.

In our future work we seek to apply this approach to a broader set of attack channels, statistical criteria, and perceptual distortion measures. Additionally we are interested in tradeoff between allowing an acceptable risk of detection with the increase in embedding rate.

ACKNOWLEDGMENTS

This research is supported in part by a grant from ONR # N00014-05-1-0816.

REFERENCES

1. "http://www.cs.dartmouth.edu/farid/research/steg.m." Code for generating wavelet-based feature vectors for steganalysis, using both coefficient and error statistics.
2. T. Pevny and J. Fridrich, "Multi-class blind steganalysis for JPEG images," in *Proc. of SPIE*, (San Jose, CA), 2006.
3. T. Pevny and J. Fridrich, "Merging Markov and DCT features for multi-class JPEG steganalysis," in *Proc. of SPIE*, (San Jose, CA), 2007.
4. C. Chen, Y. Q. Shi, W. Chen, and G. Xuan, "Statistical moments based universal steganalysis using JPEG-2D array and 2-D characteristic function," in *Proc. ICIP*, pp. 105–108, (Atlanta, GA, USA), Oct. 2006.
5. P. Moulin and J. A. O'Sullivan, "Information-theoretic analysis of information hiding," *IEEE Trans. on Info. Theory* **49**, pp. 563–593, Mar. 2003.
6. J. Fridrich, M. Goljan, D. Hogea, and D. Soukal, "Quantitative steganalysis of digital images: Estimating the secret message length," *ACM Multimedia Systems Journal, Special issue on Multimedia Security* **9**(3), pp. 288–302, 2003.
7. K. Sullivan, K. Solanki, U. Madhow, B. S. Manjunath, and S. Chandrasekaran, "Determing achievable rates for secure, zero-divergence, steganography," in *Proc. ICIP*, pp. 121–124, 2006.
8. Y. Wang and P. Moulin, "Perfectly secure steganography: Capacity, error exponents, and code constructions." submitted to IEEE Trans. Information Theory special issue on Security, Also arxiv:cs.IT/0702161, Feb. 2007.
9. K. Solanki, K. Sullivan, U. Madhow, B. S. Manjunath, and S. Chandrasekaran, "Statistical restoration for robust and secure steganography," in *Proc. ICIP*, pp. II–1118–21, 2005.
10. K. Solanki, K. Sullivan, U. Madhow, B. S. Manjunath, and S. Chandrasekaran, "Provably secure steganography: Achieving zero K-L divergence using statistical restoration," in *Proc. ICIP*, pp. 125–128, 2006.
11. B. Chen and G. W. Wornell, "Quantization Index Modulation: A class of provably good methods for digital watermarking and information embedding," *IEEE Trans. on Info. Theory* **47**, pp. 1423–1443, May 2001.
12. A. Sarkar and B. S. Manjunath, "Estimating steganographic capacity for odd-even based embedding and its use in individual compensation," in *IEEE International Conference on Image Processing (ICIP)*, pp. 409–412, (San Antonio, TX), Sep 2007.
13. A. S. Cohen and A. Lapidoth, "The Gaussian watermarking game," *IEEE Trans. on Info. Theory* **48**, pp. 1639–1667, June 2002.
14. C. Cachin, "An information theoretic model for steganography," *LNCS: 2nd Int'l Workshop on Info. Hiding* **1525**, pp. 306–318, 1998.
15. K. Solanki, N. Jacobsen, U. Madhow, B. S. Manjunath, and S. Chandrasekaran, "Robust image-adaptive data hiding based on erasure and error correction," *IEEE Trans. on Image Processing* **13**, pp. 1627 –1639, Dec 2004.
16. K. Solanki, N. Jacobsen, S. Chandrasekaran, U. Madhow, and B. S. Manjunath, "High-volume data hiding in images: Introducing perceptual criteria into quantization based embedding," in *Proc. ICASSP*, (Orlando, FL, USA), May 2002.

17. R. de Queiroz, C. K. Choi, Y. Huh, and K. R. Rao, “Wavelet transforms in a JPEG-like image coder,” *IEEE Transactions on Circuits and Systems for Video Technology* **7**, pp. 419–424, Apr 1997.

18. “http://scien.stanford.edu/class/psych221/projects/00/shuoyen/.” Overview of JPEG 2000 and Wavelet Compression - Stanford University.

19. A. Sarkar, K. Solanki, U. Madhow, S. Chandrasekaran, and B. S. Manjunath, “Secure steganography: Statistical restoration of the second order dependencies for improved security,” in *Proc. ICASSP*, **2**, pp. II–277–II–280, 2007.

20. Y. Rubner, C. Tomassi, and L. J. Guibas, “The Earth Mover’s Distance as a metric for image retrieval,” *International Journal of Computer Vision* **40**(2), pp. 99–121, 2000.

21. R. Tzschoppe, R. Bauml, and J. Eggers, “Histogram modifications with minimum MSE distortion.” Tech. Rep., Telecom. Lab., Univ. of Erlangen-Nuremberg, Dec 2001.

22. D. Divsalar, H. Jin, and R. J. McEliece, “Coding theorems for turbo-like codes,” in *36th Allerton Conf. on Communications, Control, and Computing*, pp. 201–210, Sept. 1998.