Breaking Cauchy Model-Based JPEG Steganography with First Order Statistics

Rainer Böhme and Andreas Westfeld

Technische Universität Dresden
Institute for System Architecture
01069 Dresden, Germany
{rainer.boehme,westfeld}@mail.inf.tu-dresden.de

Abstract. The recent approach of a model-based framework for steganography fruitfully contributes to the discussion on the security of steganography. In addition, the first proposal for an embedding algorithm constructed under the model-based paradigm reached remarkable performance in terms of capacity and security. In this paper, we review the emerging of model-based steganography in the context of decent steganalysis as well as from theoretical considerations, before we present a method to attack the above-mentioned scheme on the basis of first order statistics. Experimental results show a good detection ratio for a large test set of typical JPEG images. The attack is successful because of weaknesses in the model and does not put into question the generalised theoretical framework of model-based steganography. So we discuss possible implications for improved embedding functions.

1 Introduction

Steganography is the art and science of hiding information such that its presence cannot be detected. Unlike cryptography, where anybody on the transmission channel notices the flow of information but cannot read its content, steganography aims to embed a confidential message in unsuspicious data, such as image or audio files [18]. Like in cryptography, the Kerckhoffs principle [16] also applies to steganography: Security relies on publicly known algorithms that are parameterised with secret keys.

Steganalysis is the task to attack steganographic systems. For a successful attack, it is sufficient for an adversary to prove the existence of a hidden message in a carrier even if she cannot decrypt the content. Whereas in most cases the existence of steganography can only be expressed in probabilities, the literature suggests a somewhat weaker notion for successful attacks. A steganographic algorithm is considered as broken if there exists a method that can determine whether or not a medium contains hidden information with a success rate better than random guessing.

1.1 Related Embedding Schemes and Successful Attacks

Judging from the set of available steganographic tools, digital images are the most popular carrier for steganographic data, likely because of being both op-

P. Samarati et al. (Eds.): ESORICS 2004, LNCS 3193, pp. 125–140, 2004.

erable and plausible. The plausibility of steganographic target formats increases with the amount of data transmitted in the respective format. Regarding the WWW and E-mail, JPEG images are widely used, and therefore they are an ideal target format.

Jsteg [22], released in 1993, is probably the first steganographic tool to embed into JPEG images. Embedding is accomplished by replacing the least significant bits of quantised coefficients that describe the image data in the frequency domain. Even though, this simple method can be reliably detected with the Chi-square attack (χ^2) [26]. This attack exploits the pair wise dependencies of adjacent bins in the histogram, which occur after embedding of uniformly distributed message bits.

To prevent this attack, the algorithm F5 [24] uses a different embedding function, adapting the least significant bits to the message by decreasing the coefficients' absolute values. In addition, F5 implements two steganographic techniques to lower the risk of detection for messages below the full capacity. *Matrix encoding* [4] minimises the amount of modifications per message bit by carefully selecting the modified coefficients. A *permutative straddling* function spreads the message equally over the whole image.

OutGuess [19], another algorithm, also replaces the least significant bits, but additionally introduces correction bits to preserve the first order statistics. Thus, it is not vulnerable to the Chi-square attack. OutGuess voluntarily limits the maximum steganographic content to 6 % of the file size (about half as much as the before mentioned algorithms support) in order to realise *plausible deniability*: At first, a secret message is embedded together with error correction codes. Then, a second harmless message can be embedded, which acts as alibi for the case that the concealed communication is actually discovered.

Both OutGuess and F5 can be detected by computing *calibrated statistics*. Uncompressing a JPEG image and re-compressing it after a slight transformation in the spatial domain accomplishes this. A comparison of both marginal statistics, of the examined image and of the re-compressed image, reveals the existence of a hidden message [10, 11].

Apart from targeted attacks, which are constructed for particular embedding functions, *blind attacks* [17, 7] do not assume knowledge about the functionality of particular algorithms. Blind methods extract a broad set of statistical features, which might be subject to changes due to embedding. Then, a classifier is trained with a large number of typical images, both pristine carriers and stegotexts. Although suffering from lower prediction reliability than targeted attacks, blind attacks have the advantage of easy adaptability to new embedding functions. While in this case targeted attacks have to be altered or redesigned, blind attacks just require a new training.

1.2 Towards Model-Based Steganography

There have been several attempts to formalise the security of steganographic systems from an information theoretic point of view. Based on Anderson and Petitcolas' [1] initial idea to argue with entropy measures of carrier, stegotexts, and hidden messages, Zöllner et al. [29] show that information theoretical secure steganography is not feasible in general. As a result, they introduce the notion of an *in-deterministic steganographic function*. This concept implies that the steganographic carrier, represented as random variable $X = (X_{\mathrm{det}}, X_{\mathrm{indet}})$, can be split up into a deterministic part X_{det} and an in-deterministic part X_{indet}. Zöllner et al. assume that an adversary has knowledge about deterministic parts of a carrier, ranging from general assumptions about marginal distributions – for example, the typical macro structures of natural images – to specific information about an actual carrier, such as the possibility to verify the accuracy of a digitised photograph by comparing it to the depicted scene. Hence, the deterministic part must not be altered to carry steganographic data. The in-deterministic part, however, is assumed to be uniformly distributed random noise, which has been introduced, for example, by quantising the signal with an analogue-digital converter. Apart from meta-information, such as proportion and marginal distribution, the adversary has no knowledge about the actual shape of X_{indet}. Under this assumption, X_{indet} can be replaced with a similar distributed payload message X^*_{indet} (i. e., compressed data can be considered as uniformly distributed) to compose a stegotext $X^* = (X_{\mathrm{det}}, X^*_{\mathrm{indet}})$.

Though this approach sounds simple in theory, its practical application suffers from the problem to separate X_{indet} from X_{det}. This separation is not only complicated by the varying qualitative assumptions about which information an adversary can gain about the carrier – in more general terms, this is a question of the adversary model –, but also by the difficulty to consider all possible dependencies between

1. the “noise” and the structure of a carrier, and
2. the intra-dependencies within the “noise” part[1].

So, most attempts to separate X_{det} from X_{indet} are rather naive. The most widely used one is *least significant bit* (LSB) embedding, which implicitly assumes the k LSBs as X_{indet}, and the remaining bits as X_{det}. A couple of successful attacks against this scheme [5, 9, 12, 26, 28] proves the inadequacy of this approach.

Also arguing with information theory, Cachin [3] describes the relation of the relative entropy between the probability distributions of carrier data and stegotexts to the error probabilities in a hypothesis test of a passive adversary. He introduces the concept of ε-security, denoting an upper bound for the binary relative entropy $d(\alpha, \beta) \leq \varepsilon$. In steganographic hypothesis tests, α is the probability that the adversary falsely suspects a hidden message in a pristine carrier (also *false positives*, or *type I error*), and β is probability that the adversary does not detect a hidden message (*misses*, or *type II error*).

These information theoretic considerations, however, seemed to have only marginal influence on the design of specific steganographic algorithms. Eventually, Sallee's work [21] contributes to remedy this unsatisfactory situation. His proposal of a *model-based* approach to steganography can be interpreted as an

[1] We put the term *noise* in inverted commas, because if it were *real* (i.e., uncorrelated) noise, we would not face the described problems.

evolutionary combination of the above mentioned concepts coupled with strong implications for the design of steganographic algorithms.

Model-based steganography adapts the division of the carrier into a deterministic random variable X_{det} and an in-deterministic one X_{indet}[2]. In contrast to the previous approaches, model-based steganography does not assume X_{indet} to be independently and uniformly distributed. Therefore the developers propose to find suitable models for the distribution of X_{indet}, which reflect the dependencies with X_{det}. The general model is parameterised with the actual values of X_{det} of a concrete cover medium, which leads to a cover specific model. The purpose of this model is to determine the conditional distributions $P(X_{\mathrm{indet}}|X_{\mathrm{det}} = x_{\mathrm{det}})$. Then, an arithmetic decompression function[3] is used to fit uniformly distributed message bits to the required distribution of X_{indet}, thus replacing X_{indet} by X^*_{indet}, which has similar statistic properties and contains the confidential message. Figure 1 shows a block diagram of the general model-based embedding process.

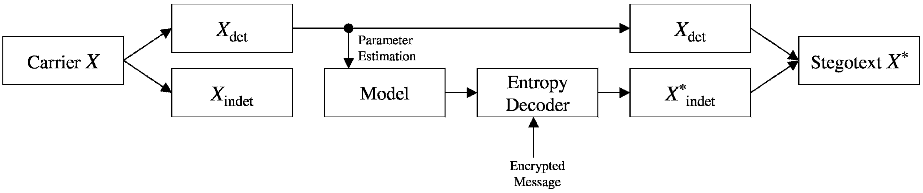

Fig. 1. Block diagram of the principles of model-based steganography

In addition to these general considerations, the initial work on model-based steganography contains a proposal for a concrete embedding function for JPEG images in the frequency domain. The purpose of this paper is to point to weaknesses of this concrete model, which allow an adversary to separate stegotexts from innocuous images.

The remainder of this paper is structured as follows: In the next section we explain the functionality of the actual embedding scheme, before we discuss its weaknesses in Section 3 in order to construct an operable attack. In Section 4, we report experimental results evaluating the performance of the presented detection method. In the final Section 5, we discuss the insights in a more general context and derive implications towards ever more secure steganography.

2 Model-Based Steganography for JPEG Images

In this section, we briefly explain the steganographic algorithm for JPEG images proposed in [21]. As we acknowledge the theoretical framework of the model-based approach, we expect further development under the new framework. So,

[2] Sallee [21] denotes X_{indet} as X_α and X_{det} as X_β. We do not follow this convention because the symbols α and β usually stand for error probabilities and might lead to confusion in other contexts.

[3] The idea of employing a decompression functions to generate arbitrary target distributions has been described in the literature as *mimic function* [23].

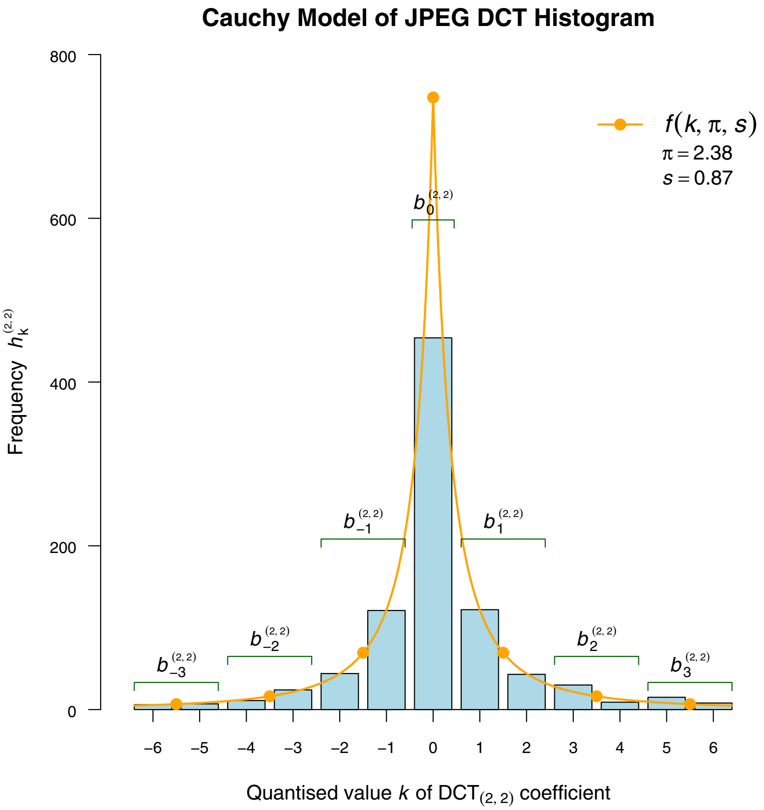

Fig. 2. Typical histogram of JPEG coefficients and approximated Cauchy distribution with data points. The frequency of the low precision bins $b_k^{(i,j)}$ is not modified during embedding

we will refer to the examined method as MB1, because it was the first one derived from the general considerations of model-based steganography.

Standardised JPEG compression cuts a greyscale image into blocks of 8 × 8 pixels, which are separately transformed into the frequency domain by a two dimensional *discrete cosine transformation* (DCT). The resulting 64 DCT coefficients, $(i, j) : i, j = 1, \ldots, 8$, are quantised with a quality dependent quantisation step size and further compressed by a lossless Huffman entropy encoder. The MB1 algorithm, as most steganographic schemes for JPEG files, embeds the steganographic semantic by modifying the quantised values. This ensures a lossless transmission of the hidden message bits.

Since individual modifications are not detectable without knowledge of the original values, an attacker reverts to the marginal statistics over all blocks of an image. Hence, the MB1 algorithm has been designed to preserve these distributions. Figure 2 depicts an example histogram of a selected $\mathrm{DCT}_{(2,2)}$ coefficient. The histogram shape is typical for all JPEG DCT coefficients except $\mathrm{DCT}_{(1,1)}$, which are therefore excluded from embedding (i. e., the $\mathrm{DCT}_{(1,1)}$ coefficients belong to X_{det}).

Let us denote $h_k^{(i,j)}$ as the number of quantised $\mathrm{DCT}_{(i,j)}$ coefficients equal to k in a given image. We will further refer to this quantity as the k-th *high precision bin* of the histogram $h^{(i,j)}$. By contrast, the *low precision bins* comprise several high precision bins. Without restricting generality, we focus on the case when a low precision bin $b_k^{(i,j)}$ $(k \neq 0)$ contains exactly two high precision bins, so that

$$b_k^{(i,j)} = \begin{cases} h_{2k+1}^{(i,j)} + h_{2k}^{(i,j)} & \text{for } k < 0 \\ h_0^{(i,j)} & \text{for } k = 0 \\ h_{2k-1}^{(i,j)} + h_{2k}^{(i,j)} & \text{for } k > 0 \ . \end{cases}$$

To avoid the case differentiation and simplify the notation, we will write further equations only for $k > 0$. Furthermore, $q \in [0,1]$ denotes the quality parameter of a JPEG compression that is used to compute the quantisation tables.

The MB1 algorithm defines the size of the low precision bins $b_k^{(i,j)}$ as part of X_{det}, while the distribution within the low precision bins (i. e., the corresponding high precision bins $h_{2k-1}^{(i,j)}$ and $h_{2k}^{(i,j)}$) is considered as part of X_{indet}. The embedding function alters the quantised DCT coefficients, so that

1. the new values belong to the same low precision bin, and
2. the conditional distribution of $h_{2k-1}^{(i,j)}$ and $h_{2k}^{(i,j)}$ from a given $b_k^{(i,j)}$ keeps coherent according to a model.

This is accomplished by altering coefficient values of $2k-1$ to $2k$ and vice versa. In contrast to simple LSB overwriting, the conditional probabilities of occurrence $P(X_{\mathrm{indet}}|X_{\mathrm{det}} = x_{\mathrm{det}})$, actually $P(h_{2k-1}^{(i,j)}|b_k^{(i,j)})$, are derived from the model in dependency of the low precision bin $b_k^{(i,j)}$. As it is obvious that the probabilities for all high precision bins sum up to 1 in each low precision bin,

$$P(h_{2k-1}^{(i,j)}|b_k^{(i,j)}) + P(h_{2k}^{(i,j)}|b_k^{(i,j)}) = 1, \qquad \forall\ i, j, k,$$

we further refer only to the $P(h_{2k-1}^{(i,j)}|b_k^{(i,j)})$ as $p_k^{(i,j)}$. The required $p_k^{(i,j)}$ is adjusted to the shape of a Modified Generalised Cauchy (MGC) distribution $f(k, \pi, s)$:

$$p_k^{(i,j)} \quad = \quad \frac{f(2k-1, \pi, s)}{f(2k-1, \pi, s) + f(2k, \pi, s)}$$

The density function of the MGC distribution applied is defined as follows:

$$f(k, \pi, s) = \frac{p-1}{2s}(|k/s| + 1)^{-\pi}$$

The scale parameter s and the location parameter π are computed separately for all DCT modes by a maximum likelihood estimation over the low precision bins $b^{(i,j)}$. Then, $p_k^{(i,j)}$ is determined for all low precision bins $b_k^{(i,j)}, k \neq 0$ of each DCT mode, but $\mathrm{DCT}_{(1,1)}$ coefficients and zero value coefficients $b_0^{(i,j)}$ are excluded from embedding. An arithmetic entropy decoder [27, cited from [21]] is used to fit the compressed and encrypted – thus uniformly distributed – message bits $m \sim U$ to a discrete vector with defined symbol probabilities $p_k^{(i,j)}$ and $1 - p_k^{(i,j)}$ [4]. As $b^{(i,j)}$ is not modified due to embedding, the receiver can recompute the model parameters and thus extract the message.

One way to evaluate the performance of an embedding algorithm is the embedding efficiency. According to [24], the embedding efficiency in JPEG files can be defined as the average message bits encoded per change of a coefficient. The application of an arithmetic decoder is an elegant way to achieve an exceptionally high embedding efficiency. Sallee reports embedding efficiencies between 2.06 and 2.16 bits per change for test images with $q = 80\,\%$ [21]. Other decent algorithms achieve values between 1.0 and 2.0 (OutGuess), or just under 2.0 (F5)[5]. Also in terms of capacity, MB1 performs on the upper end of the range. The capacity is defined as ratio of message bits per transmitted bits. MB1 reaches values of just under 14 %, which is slightly better than F5 and Jsteg (about 13 % and 12 %, respectively), and clearly above OutGuess (below 7 %).

Being explicitly designed as proof of concept, the developers of MB1 concede that the simple model does not include higher order statistics. However, they claim it to be "resistant to first order statistical attacks" [21, p. 166]. First order statistics are all measures describing data regardless of the inter-dependencies between observations, such as mean, variance, and histograms. Higher order statistics consider the relationship between observations and their position in the dataset; for example correlations between adjacent pixels in an image. As a rule of thumb, if the results of a statistical measure are invariant to any permutation of the data, then it is first order statistics.

Until today, none of the existing attacks against other algorithms also works on MB1, and no targeted attack has been published. Though, it is not surprising that a blind attack with special second order features, such as blockiness measures and co-occurrence tables, can discriminate between plain carriers and MB1 stegotexts [7]. But according to the outlook in the initial paper, we soon expect improved methods also taking into account some second order statistics. Whereas research clearly goes into this direction, it is somewhat important and also surprising that MB1 steganography is also vulnerable from the believed safe side: In the following section, we present a detection method which is completely based on first order statistics.

[4] The use of a *de*-coder might sound surprising, however, entropy considerations suggest that the length of the symbol stream increases with the skewness of the target distribution. For all $p_k^{(i,j)} \neq 0.5$ the amount of symbols and of consumed coefficients dominates the length of the message bit stream.

[5] Matrix encoding in F5 leads to higher efficiencies if the capacity is not fully used.

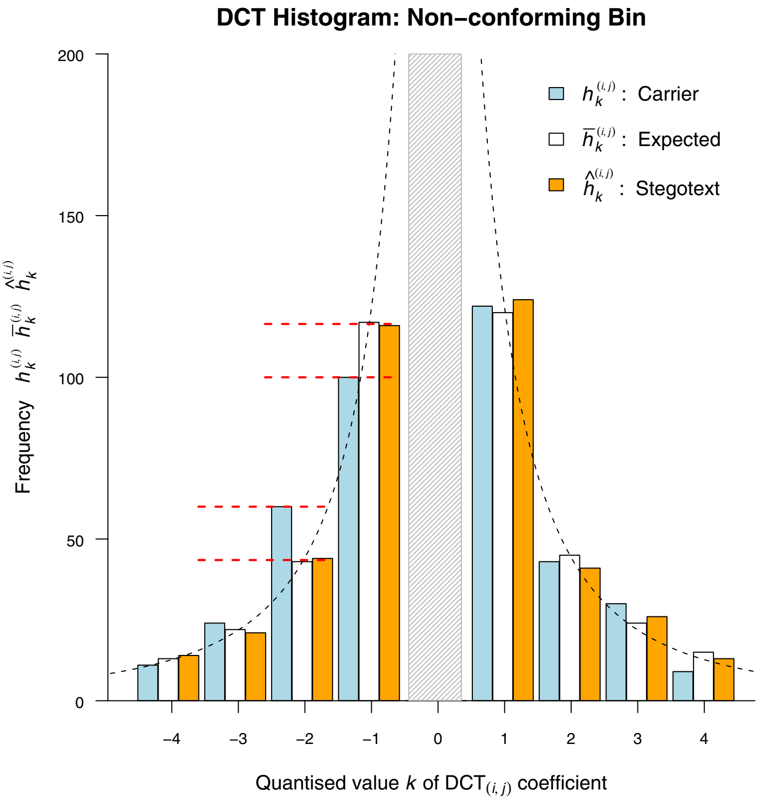

Fig. 3. Example DCT histogram with non-conforming bin $b_{-1}^{(i,j)}$. The divergences in the carrier between actual frequency and Cauchy-based expected frequencies disappear after MB1 embedding

3 Detection Method

The main idea of the proposed attack can be summarised as follows: Although the Cauchy distribution generally fits well to the DCT histograms, there are outlier bins in natural images. After embedding, these non-conforming bins are adjusted to the density function of the model distribution.

The construction of an attack can be structured into two steps: First, a test discriminates non-conforming bins from conforming ones. The test is run on all independent low precision bins of a particular JPEG image. Second, the count of positive test results is compared to an empirical threshold value for natural carrier images. If a questionable image contains less non-conforming bins than plain carrier images usually have, it is likely that the histograms are smoothed by the MB1 embedding function and thus, the image is classified as steganogram.

Figure 3 depicts a DCT histogram with a typical outlier in the low precision bin b_{-1}. The bins $b_0^{(i,j)}$ are excluded from embedding, so the respective bars are blanked out in the histogram. It is clearly visible that the original frequencies h_{-1} and h_{-2} (left bars of the triples) differ from the expected frequencies (middle bars). The expected frequencies and the frequencies in the stegotext (right bars) are fitted to the same level.

The differences can be measured by a contingency test between the observed frequencies and expected frequencies of both high precision bins represented in one low precision bin. To calculate the expected frequencies, we model the symbol output of the arithmetic decoder as a Bernoulli distributed random variable $Y(p)$ with $p = p_k^{(i,j)}$ [6]. So the high precision histogram bins of stegotexts $\hat{h}^{(i,j)}$ follow a Binomial distribution

$$\hat{h}_{2k-1}^{(i,j)} \sim B(b_k^{(i,j)}, p_k^{(i,j)}) \qquad \text{, and}$$

$$\hat{h}_{2k}^{(i,j)} \sim B(b_k^{(i,j)}, 1 - p_k^{(i,j)}).$$

The expected frequencies $\bar{h}^{(i,j)}$ are given by the expected values of B:

$$\bar{h}_{2k-1}^{(i,j)} = E(B(b_k^{(i,j)}, p_k^{(i,j)})) = b_k^{(i,j)} \cdot p_k^{(i,j)}.$$

An adversary can compute these values by refitting the model for $p_k^{(i,j)}$ because the low precision bins are not altered. Then a contingency table is used to perform Pearsons's χ^2-test, whether or not individual low precision bins are conform to the model (see Table 1). The distribution function $Q(\chi^2, \mathrm{df})$ of the χ^2 distribution gives an error probability for the null hypothesis that the contrasted frequencies are independent. The test will reject the null for non-conforming bins if $p < p_{\mathrm{lim}}$.

Table 1. Contingency test for non-conforming low precision bins

	High precision bin		
	left	right	$\sum$
Observed frequencies	$h_{2\cdot k-1}^{(i,j)}$	$h_{2\cdot k}^{(i,j)}$	$b_k^{(i,j)}$
Expected frequencies	$\bar{h}_{2\cdot k-1}^{(i,j)}$	$\bar{h}_{2\cdot k}^{(i,j)}$	$b_k^{(i,j)}$
$p = Q(\chi^2, \mathrm{df} = 1)$			

To explore the average count of non-conforming bins in typical JPEG images, contingency tests are run on the low precision bins $b_1^{(i,j)}$ and $b_{-1}^{(i,j)}$ for 63 DCT modes of a set of 100 JPEG images (altogether 126 tests per image). These images were randomly drawn from a large number of digital photographs with the

[6] The assumption that the symbol output is drawn from a Bernoulli distribution is a worst case assumption. Any "better" arithmetic decoding algorithm would – apart from reducing the entropy – on average fit the stegotext bin sizes closer to the expected sizes and thus lead to more arbitrable contingency tests.

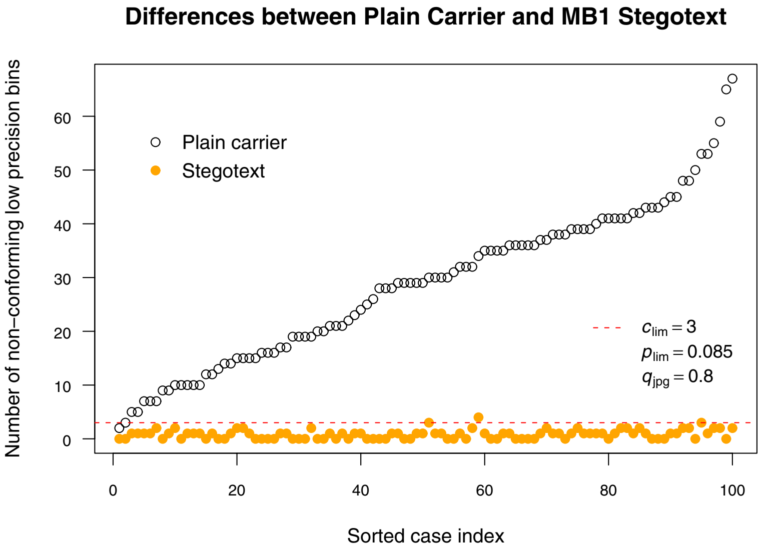

Fig. 4. Non-conformities with the assumed Cauchy distribution are typical for JPEG images. MB1 is detectable because it erases these particularities

resolution 800×600 and a JPEG quality parameter of $q = 0.8$. Figure 4 contrasts the results to 100 full capacity stegotexts created from the same carriers. It is obvious that a threshold of, say, $c_{\text{lim}} = 3$ can quite reliably discriminate the two sets.

At last, two more details of the contingency test are worth to mention: First, as the test is unreliable for low frequency numbers in any of the cells, tables with a minimal cell value below 3 are excluded from the evaluation. Second, the reliability of the test depends on the number of DCT coefficients unequal to zero. Since this number varies both with the size of the test image and with the quantisation step size derived from q, the critical probability p_{lim} has to be adjusted to the above mentioned parameters. This method allows an optimal differentiation in terms of low error probabilities α and β of the stegotext detection.

4 Experimental Results

The reliability of the proposed detection method was assessed using a test database of about 300 images from a digital camera[7]. To reduce unwanted influences or atypical artefacts due to previous JPEG compression [8], all images were scaled down to a resolution of 800×600 pixels and stored as JPEG with six different quality settings, $q = 0.4, 0.5, \ldots, 0.9$. In all experiments, only the luminance component of colour images has been regarded. All analyses were accomplished with the *R Project for Statistical Computing* [20, 14].

[7] Sony Cybershot DSC-F55E, 2.1 mega-pixel.

To generate comparable histogram sets of plain carrier and steganograms, 63 DCT histograms were extracted from all images. The plain carrier histograms were transformed to equivalent stegotext histograms by replacing the high precision bins with random numbers drawn from a Binomial distribution, using before determined parameters from the model:

$$\hat{h}_{2k-1}^{(i,j)} = R_{\mathrm{binom}}(b_k^{(i,j)}, p_k^{(i,j)})$$

$$\hat{h}_{2k}^{(i,j)} = b_k^{(i,j)} - \hat{h}_{2k-1}^{(i,j)}$$

Furthermore it is obvious that limiting capacity leads to smaller changes in the histograms and thus shorter messages are less detectable. To estimate this effect we also varied the capacity usage in 10 levels from full capacity down to 10 % for all test images and quality factors. This leads to a set of 1.2 M stegotext DCT histograms (equivalent to 18,120 stegotext images), which were compared to the same amount of plain carrier histograms. Explorative analyses of suitable bins for the contingency test revealed that the bins $b_{-1}^{(i,j)}$ and $b_1^{(i,j)}$ yielded to the best results for all DCT modes. So, all other bins were excluded from the evaluation.

In a first experiment, the proposed attack was run on a subset of this database with 100 % capacity usage and $q = 0.8$. Here, all images could be correctly classified with $p_{\lim} = 0.014$ (corresp. $\chi^2 = 6$, $df = 1$). The threshold $c_{\lim} = 2$ was fixed in all experiments. Attacks on steganograms with lower capacity usage or lower q cause misclassifications. The number of false positives (α) and misses (β) depends on the choice of the threshold parameters.

To further explore the relationship between α and β, the attack was repeated multiple times with different $p_{\lim} \in [0.0001, 0.2]$ [8], and the resulting error rates were plotted in a *receiver operating characteristics* (ROC) diagram shown in Figure 5. Here again, we reached good discriminatory power for capacity usages higher than 80 %, and still acceptable detection rates for capacities above 50 %. Hence, we can state that the MB1 algorithm is broken with first order statistics.

The qualitative interpretation of the shape of ROC curves can be quantified in an aggregated measure of the reliability of a detection method. Unfortunately, different quality measures in the literature complicate comparisons between different studies. Some authors argue with the probability $(1 - \beta)$ for a fixed proportion of false positives, say $\alpha = 1\,\%$ [17]. Others compare α values for a fixed detection rate of $\beta = 1 - \beta = 50\,\%$ [15]. In this paper, we follow the third approach from [7], which reflects both α and β: The detection reliability ρ is defined as $\rho = 2A - 1$, where A is the area under the ROC curve. It is normalised, so that $\rho = 0$ indicates no discriminatory power at all (i. e., random guessing) and $\rho = 1$ stands for a perfect detection.

Table 2 reports the empirical ρ values derived from the test images for different allocations of capacity, and different quantisation factor q. The minimum

[8] In fact, varying the underlying χ^2 threshold leads to equivalent results since the (computing intensive) transformation function Q is strictly monotonic decreasing.

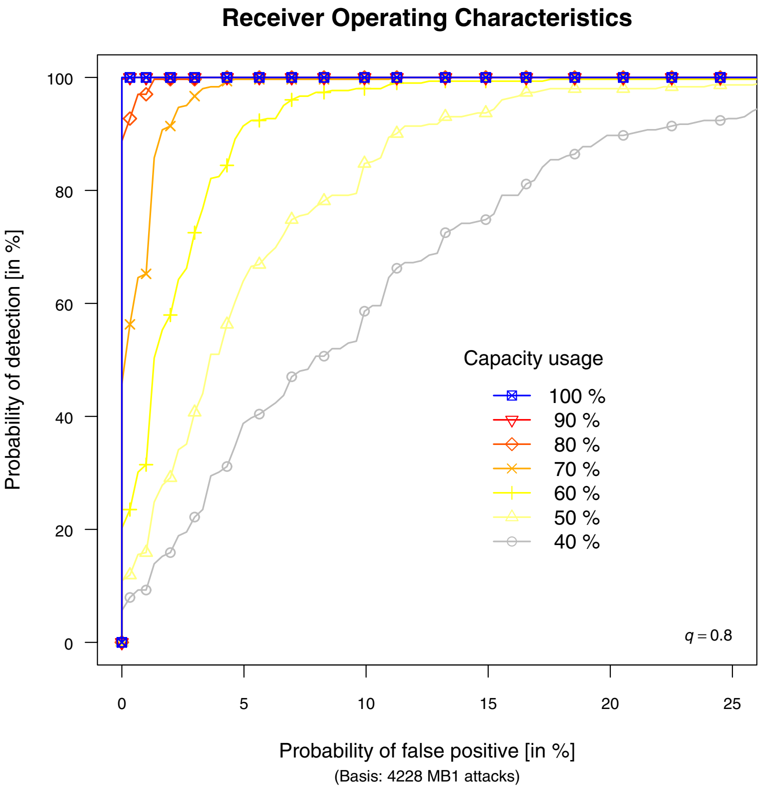

Fig. 5. Dicriminatory power of attacks on MB1 for different capacity usages

embedding rate that is detectable under the arbitrary definition of 'reliablility' stated in [15], namely $\alpha = 5\,\%$ and $\beta = 50\,\%$, is about $45\,\%$ of the capacity of JPEG images with $q = 0.8$. Note that these figures reflect average estimations. The actual detectability is likely to vary for certain carrier images.

5 Discussion and Conclusion

It is important to emphasise that this vulnerability of the MB1 scheme is rather a problem of the specific model used in the scheme, than a weakness of the general approach of model-based steganography. Having said this, the successful attack is still somewhat surprising, because the theoretical considerations given in the original paper [21, p. 166] suggest that possible vulnerabilities come from analyses of higher order statistics, which are not reflected in the model. However, the proposed detection method only uses characteristics of first order statistics, which were considered as safe.

The remainder of this section addresses three open subjects. First, we point to limits of the proposed attack and propose future improvements. Then, we discuss

Table 2. Experimental attacks: Detection reliablility ρ

Capacity	Avg. message size	JPEG quality q					
usage	per file size	0.9	0.8	0.7	0.6	0.5	0.4
100 %	13.1 %	**1.0000**	**1.0000**	**0.9999**	**1.0000**	**1.0000**	**0.9979**
90 %	11.8 %	**1.0000**	**1.0000**	**0.9997**	**1.0000**	**0.9996**	**0.9963**
80 %	10.5 %	**0.9989**	**0.9982**	**0.9949**	**0.9970**	**0.9940**	**0.9826**
70 %	9.2 %	**0.9890**	**0.9850**	**0.9797**	**0.9777**	**0.9740**	**0.9596**
60 %	7.9 %	**0.9593**	**0.9527**	**0.9440**	**0.9322**	**0.9292**	**0.9202**
50 %	6.6 %	**0.9012**	**0.8898**	**0.8782**	0.8615	0.8552	0.8519
40 %	5.2 %	0.8057	0.7906	0.7796	0.7624	0.7509	0.7516
30 %	3.9 %	0.6576	0.6476	0.6457	0.6185	0.6063	0.6180
20 %	2.6 %	0.4568	0.4549	0.4583	0.4295	0.4214	0.4362
10 %	1.3 %	0.2289	0.2294	0.2357	0.2160	0.2133	0.2252

The ROC curves for values printed bold-face meet a reliability criterium of more than 50 % detection rate with less than 5 % false positives.

possible countermeasures to prevent this attack, before we finally conclude with more general implications for the design of new and better embedding functions.

This attack is considered as proof of concept and as first step towards more precise attacks. It has been mainly driven by the feeling that an embedding algorithm offering a payload capacity of about 13 % of the transferred data is very likely to be detectable with a targeted attack. Similar to the evolution of attacks against LSB steganography after the publication of [26], we expect that this existential break will lead to far better attacks, which shall be able to estimate the hidden message length and thus will even detect tiny messages.

A possible extension to this attack can be the anticipation of common image processing steps. As the experiments were run on a limited set of test images directly loaded from a digital camera, we cannot generalise our results to all kind of carrier data. At first, computer generated images may have different characteristics than natural images. This is a reason why the literature suggests that this type of carrier should be avoided for steganography [18]. Even though, natural images are often subject to image manipulation. It is likely that some of these algorithms, e. g. blurring, also result in smoother DCT histograms. The challenge to detect these manipulations in advance and thus reduce the number of false positives, or even to distinguish between "white collar" image processing and "black hat" Cauchy-based steganography, is subject to further research.

Thinking about possible countermeasures, the ad hoc solution is as old as digital steganography, namely a reduction in capacity usage. Nevertheless it is an interesting research question, how this limitation can be optimally accomplished for model-based steganography. Whereas the conditional distributions for individual bins depend on the deterministic part X_{det}, a careful selection of the skipped bins or coefficients may lead to a far better ratio between security and capacity, than random selection. A similar approach is described in [6]. This method exactly preserves the low precision bins without a model – albeit in the spatial domain of losslessly compressed images, and in a less optimal manner. Therefore it is not vulnerable to attacks on first order statistics, but still detectable due to other flaws [2].

Refining the model could be another promising countermeasure. As a rather fussy preservation of the properties of the actual carrier is often superfluous and also complicates the embedding function, we could imagine to model a certain amount of non-conforming bins, and to randomly intersperse the stegotext with outliers.

Despite the specific obstacles of MB1, the model-based approach offers a promising framework for the design of adaptive steganographic algorithms. The clear link between information theoretic considerations and the design of actual algorithms contributes to structure the research area. A generalisation from the concrete vulnerabilities suggests two implications for the design of more secure embedding functions.

First, it is dangerous to give up the information superiority of the colluding communication partners. The described attack on MB1 is successful, because an adversary can re-compute the model parameters. If the adversary had no access to the Cauchy distribution, she would not be able to compute the expected frequencies. Hence, future algorithms should either consider to make the parameter retrieval key dependant, or perform an embedding operation which does not require the receiver to know the exact model. The recently developed *wet paper codes* [13] seem to be a promising technique to tackle this problem.

Second, the reliability of statistical attacks increases with the amount of observations. Although MB1 already computes distinct models for each of the 63 usable DCT modes, an even more detailed segmentation of individually modelled – and maybe even locally correlated – statistics breaks the steganalyst's advantage of large numbers. Apart from including second order dependencies into the models, the challenge to harden future algorithms against the here discussed weaknesses can be accomplished by modelling the carrier medium with a multiple of key dependent models.

To conclude, as it is common sense that the ultimate and provable secure model cannot exist [1, 21, 29], the core contribution of this paper is pointing out that future models should reflect the particularities that made this attack successful.

Acknowledgement

The work on this paper was supported by the Air Force Office of Scientific Research under the research grant number FA8655-03-1-3A46. The U. S. Government is authorised to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright notation there on. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of the Air Force Office of Scientific Research, or the U. S. Government.

References

1. Anderson, R., Petitcolas, F. A. P.: On the Limits of Steganography. *IEEE Journal of Selected Areas in Communications* **16** (1998) 474–481
2. Böhme, R., Westfeld, A.: Exploiting Preserved Statistics for Steganalysis. Paper presented at the Sixth Workshop on Information Hiding, Toronto, Canada (2004, May)
3. Cachin, C.: An Information-Theoretic Model for Steganography. In: Aucsmith, D. (ed.): Information Hiding. Second International Workshop, LNCS 1525, Springer-Verlag, Berlin Heidelberg (1998) 306–318
4. Crandall, R.: Some Notes on Steganography. Posting to a mailing list on steganography (1998) `http://os.inf.tu-dresden.de/~westfeld/crandall.pdf`
5. Dumitrescu, S., Wu, X., Wang, Z.: Detection of LSB Steganography Via Sample Pair Analysis. In: Petitcolas, F. A. P. (ed.): Information Hiding. Fifth International Workshop, LNCS 2578, Springer-Verlag, Berlin Heidelberg (2003) 355–372
6. Franz, E.: Steganography Preserving Statistical Properties. In: Petitcolas, F. A. P. (ed.): Information Hiding. Fifth International Workshop, LNCS 2578, Springer-Verlag, Berlin Heidelberg (2003) 278–294
7. Fridrich, J.: Feature-based Steganalysis for JPEG Images and its Implications for Future Design of Steganographic Schemes. Paper presented at the Sixth Workshop on Information Hiding, Toronto, Canada (2004, May)
8. Fridrich, J., Goljan, M., Du, R.: Steganalysis Based on JPEG Compatibility. In: Tescher, A. G., Vasudev, B., Bove,V. M., Jr. (eds.): Proceedings of SPIE, Multimedia Systems and Applications IV, Denver, CO (2001) 275–280
9. Fridrich, J., Goljan, M., Du, R.: Reliable Detection of LSB Based Image Steganography. Proceedings of the ACM Workshop on Multimedia and Security (2001) 27–30
10. Fridrich, J., Goljan, M., Hogea, D.: Attacking the OutGuess. Proceedings of the ACM Workshop on Multimedia and Security (2002)
11. Fridrich, J., Goljan, M., Hogea, D.: Steganalysis of JPEG Images: Breaking the F5 Algorithm. In: Petitcolas, F. A. P. (ed.): Information Hiding. Fifth International Workshop, LNCS 2578, Springer-Verlag, Berlin Heidelberg (2003) 310–323
12. Fridrich, J., Goljan, M., Soukal, D.: Higher-order Statistical Steganalysis of Palette Images. In: Delp, E. J., Wong, P. W. (eds.): Proceedings of SPIE, Security and Watermarking of Multimedia Contents V (2003) 178–190
13. Fridrich, J., Goljan, M., Soukal, D.: Perturbed Quantization Steganography Using Wet Paper Codes. Paper to be presented at the ACM Workshop on Multimedia and Security, Magdeburg, Germany (2004, September 20–21)
14. Ihaka, R., Gentlemen, R.: R – A Language for Data Analysis and Graphics. *Journal of Computational Graphics and Statistics* **5** (1996) 299–314
15. Ker, A.: Improved Detection of LSB Steganography in Grayscale Images. Paper presented at the Sixth Workshop on Information Hiding, Toronto, Canada (2004, May)
16. Kerckhoffs, A.: La cryptographie militaire. *Journal des sciences militaires* **XI** (1883) 5–38, 161–191, `http://www.cl.cam.ac.uk/~fapp2/`
17. Lyu, S., Farid, H.: Detecting Hidden Messages Using Higher-Order Statistics and Support Vector Machines. In: Petitcolas, F. A. P. (ed.): Information Hiding. Fifth International Workshop, LNCS 2578, Springer-Verlag, Berlin Heidelberg (2003) 340–354

18. Petitcolas, F. A. P., Anderson, R. J., Kuhn, M. G.: Information Hiding – A Survey. *Proceedings of the IEEE* **87** (1999) 1062–1078
19. Provos, N.: OutGuess – Universal Steganography (2001) `http://www.outguess.org/`
20. The R Project for Statistical Computing, `http://www.r-project.org/`.
21. Sallee, P.: Model-Based Steganography. In: Kalker, T., et al. (eds.): International Workshop on Digital Watermarking, LNCS 2939, Springer-Verlag, Berlin Heidelberg (2004) 154–167
22. Upham, D: Jsteg (1993) `http://ftp.funet.fi/pub/crypt/cypherpunks/applications/jsteg/`
23. Wayner, P.: Mimic Functions. *Cryptologia* **16** (1992) 193–214
24. Westfeld, A.: F5 – A Steganographic Algorithm. High Capacity Despite Better Steganalysis. In: Moskowitz, I. S. (ed.): Information Hiding. Fourth International Workshop, LNCS 2137, Springer-Verlag, Berlin Heidelberg (2001) 289–302
25. Westfeld, A.: Detecting Low Embedding Rates. In: Petitcolas, F. A. P. (ed.): Information Hiding. Fifth International Workshop, LNCS 2578, Springer-Verlag, Berlin Heidelberg (2003) 324–339
26. Westfeld, A., Pfitzmann, A.: Attacks on Steganographic Systems. In: Pfitzmann, A. (ed.): Information Hiding. Third International Workshop, LNCS 1768, Springer-Verlag, Berlin Heidelberg (2000) 61–76
27. Witten, I. H., Neal, R., M., Cleary, J. G.: Arithmetic Coding for Data Compression. *Communications of the ACM* **20** (1987) 520–540
28. Zhang, X., Wang, S., Zhang, K.: Steganography with Least Histogram Abnormality. In: Gorodetsky et al. (eds.): MMM-ACNS 2003, LNCS 2776, Springer-Verlag, Berlin Heidelberg (2003) 395–406
29. Zöllner, J., Federrath, H., Klimant, H., Pfitzmann, A., Piotraschke, R., Westfeld, A., Wicke, G., Wolf, G.: Modelling the Security of Steganographic Systems. In: Aucsmith, D. (ed.): Information Hiding. Second International Workshop, LNCS 1525, Springer-Verlag, Berlin Heidelberg (1998) 334–354