Effect of JPEG Quality on Steganographic Security

Jan Butora
Binghamton University
Department of ECE
Binghamton, NY
jbutora1@binghamton.edu

Jessica Fridrich
Binghamton University
Department of ECE
Binghamton, NY
fridrich@binghamton.edu

ABSTRACT

This work investigates both theoretically and experimentally the security of JPEG steganography as a function of the quality factor. For a fixed relative payload, modern embedding schemes, such as J-UNIWARD and UED-JC, exhibit surprising non-monotone trends due to rounding and clipping of quantization steps. Their security generally increases with increasing quality factor but starts decreasing for qualities above 95. In contrast, old-fashion steganography, such as Jsteg, OutGuess, and model-based steganography, exhibit complementary trends. The results of empirical detectors closely match the trends exhibited by the KL divergence computed between models of cover and stego DCT modes. In particular, our analysis shows that the main reason for the complementary trends is the way modern schemes attenuate embedding change rates with increasing spatial frequency. Our model also provides guidance on how to adjust the embedding algorithm J-UNIWARD to improve its security for JPEG quality factor 100.

KEYWORDS

Steganography, Steganalysis, JPEG, quality factor, generalized Gaussian

ACM Reference Format:
Jan Butora and Jessica Fridrich. 2019. Effect of JPEG Quality on Steganographic Security. In *ACM Information Hiding and Multimedia Security Workshop (IHMMSec '19), July 3–5, 2019, TROYES, France.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3335203.3335714

1 INTRODUCTION

The JPEG format is the most ubiquitous image format in use today due to its ability to efficiently compress visual data without introducing perceivable artifacts and the fact that it is supported across all platforms by all applications capable of displaying imagery. It is also a quite complex format because the compression algorithm is controlled by numerous parameters and settings, such as the selection of the color representation, quantization matrices, chrominance subsampling, and the specific implementation of the Discrete Cosine Transform (DCT). Surprisingly little research is available on the effect of the above choices on detectability of steganography.

IHMMSec '19, July 3–5, 2019, TROYES, France

ACM ISBN 978-1-4503-6821-6/19/06…$15.00
https://doi.org/10.1145/3335203.3335714

Arguably, the most influential settings in JPEG compression are the quantization matrices, which control the trade-off between the file size and image quality. As this paper shows using both empirical detectors and theoretical arguments, the impact of quantization on security is quite complex and depends on the specific embedding algorithm. Most notably, for relative payload fixed in terms of bits per non-zero AC DCT coefficient (bpnzac) the security of "old" embedding methods, such as Jsteg [13] (or any generic LSB flipper), OutGuess [10], and Model-Based Steganography (MBS) [11], decreases with increasing JPEG quality factor (QF) but starts increasing for qualities close to 100, while the trend is just the *opposite* for modern embedding schemes, such as J-UNIWARD [7] and UED-JC [4]. Hints of this can be observed, but are not explicitly commented upon, in previous work with steganalyzers implemented using the JPEG Rich Model (JRM) and the JPEG Projection Spatial Rich Model (JPSRM) (Table 1 in [5]), detectors using the JPEG-phase-aware features (Fig. 5 and 6 in [6]), as well as detectors implemented as Convolutional Neural Networks (CNNs) [2].

In Section 2, we introduce the notation and the model of DCT coefficients used in Section 3 to quantify the impact of embedding using the KL divergence between cover and stego models of individual DCT modes. The datasets used in this paper as well as the estimation of the model from images are described in Section 4. Theoretical predictions derived from the model are validated experimentally using machine-learning based steganalyzers in Section 5. In Section 6, we provide a more intuitive explanation of the observed non-monotone security trends and identify the modulation of change rates across spatial frequencies as the key element responsible for the observed complementary trends. In the same section, we also use our model to find an adjustment of embedding change rates of J-UNIWARD to improve its security for quality factor 100. A summary and future directions appear in Section 7.

2 JPEG IMAGE MODEL

In this section, we first introduce the notation followed by a model of JPEG DCT coefficients that will later be used in Section 3 to assess the impact of steganographic embedding changes on security.

2.1 Notation

For simplicity, we only consider $n_1 \times n_2$ 8-bit grayscale images x_{ij}, $1 \le i \le n_1$, $1 \le j \le n_2$, with n_1 and n_2 multiples of 8. The (a,b)th 8×8 block of pixels, $1 \le a \le n_1/8$, $1 \le b \le n_2/8$, formed by pixels with indices $8(a-1)+i+1, 8(b-1)+j+1$, $0 \le i,j \le 7$, will be denoted $\mathbf{x}^{(a,b)} = (x_{ij}^{(a,b)})$. Similarly, the (a,b)th 8×8 block of unquantized and quantized DCT coefficients will be denoted $\mathbf{c}^{(a,b)} = (c_{ij}^{(a,b)})$ and $\mathbf{d}^{(a,b)} = (d_{ij}^{(a,b)})$, respectively, where $d_{kl}^{(a,b)} = \left[c_{kl}^{(a,b)}/q_{kl}\right]$

with q_{kl} denoting the luminance quantization steps and $[x]$ the operation of rounding to integers.

Denoting the 8×8 matrix of ones with boldface $\mathbf{1}$, the standard quantization matrix for quality factor $Q \in \{1, 2, \ldots, 100\}$ is

$$\mathbf{q}(Q) = \begin{cases} \max\left\{\mathbf{1}, \left[2\mathbf{q}(50)\left(1 - \frac{Q}{100}\right)\right]\right\}, & Q > 50 \\ \min\left\{255 \times \mathbf{1}, \left[\mathbf{q}(50)\frac{50}{Q}\right]\right\}, & Q \leq 50, \end{cases} \tag{1}$$

where the luminance quantization matrix for quality factor 50 is

$$\mathbf{q}(50) = \begin{pmatrix} 16 & 11 & 10 & 16 & 24 & 40 & 51 & 61 \\ 12 & 12 & 14 & 19 & 26 & 58 & 60 & 55 \\ 14 & 13 & 16 & 24 & 40 & 57 & 69 & 56 \\ 14 & 17 & 22 & 29 & 51 & 87 & 80 & 62 \\ 18 & 22 & 37 & 56 & 68 & 109 & 103 & 77 \\ 24 & 35 & 55 & 64 & 81 & 104 & 113 & 92 \\ 49 & 64 & 78 & 87 & 103 & 121 & 120 & 101 \\ 72 & 92 & 95 & 98 & 112 & 100 & 103 & 99 \end{pmatrix}. \tag{2}$$

We use the symbol $\mathbb{Z}$ for the set of all integers, $\Gamma(x)$ for the gamma function, and $H_2(x) = -x\log_2 x - (1-x)\log_2(1-x)$, $H_3(x) = H_2(x) + x$ for the binary and ternary entropy expressed in bits.

2.2 Model

Unquantized DCT coefficients $c_{kl}^{(a,b)}$ are modeled as 64 independent channels (modes (k, l)). The coefficients in each mode (k, l) across all blocks in the image (a, b) are assumed to be independent realizations of a random variable with the generalized Gaussian (GG) distribution

$$c_{kl}^{(a,b)} \sim g(x; \gamma_{kl}, w_{kl}), \tag{3}$$

with zero mean, shape parameter $\gamma_{kl} > 0$, and width parameter $w_{kl} > 0$:

$$g(x; \gamma, w) = \frac{\gamma}{2w\Gamma\left(\frac{1}{\gamma}\right)} \exp\left(-\left|\frac{x}{w}\right|^{\gamma}\right). \tag{4}$$

We note that the variance of the GG distribution is $v = \frac{w^2\Gamma(3/\gamma)}{\Gamma(1/\gamma)}$.

Quantized DCTs from a cover image, $d_{kl}^{(a,b)}$, across all blocks (a, b), follow the quantized GG probability mass function $P_{kl}^{(c)}(m) \triangleq \Pr\{d_{kl}^{(a,b)} = m\}$, $m \in \mathbb{Z}$:

$$P_{kl}^{(c)}(m) = \int_{q_{kl}(m-\frac{1}{2})}^{q_{kl}(m+\frac{1}{2})} g(x; \gamma_{kl}, w_{kl})\mathrm{d}x = \omega(m; q_{kl}, \gamma_{kl}, w_{kl}) \tag{5}$$

$$\omega(m; q, \gamma, w) = \begin{cases} \frac{1}{2}\left[\underline{\Gamma}\left(\frac{1}{\gamma}, \left(\frac{q(|m|+\frac{1}{2})}{w}\right)^{\gamma}\right)\right. \\ \left.-\underline{\Gamma}\left(\frac{1}{\gamma}, \left(\frac{q(|m|-\frac{1}{2})}{w}\right)^{\gamma}\right)\right] & \text{for } m \neq 0 \\ \underline{\Gamma}\left(\frac{1}{\gamma}, \left(\frac{q}{2w}\right)^{\gamma}\right) & \text{for } m = 0 \end{cases} \tag{6}$$

where

$$\underline{\Gamma}(x, z) = \frac{1}{\Gamma(x)} \int_0^z t^{x-1}e^{-t}\mathrm{d}t, \tag{7}$$

is the normalized lower incomplete gamma function.

3 EMBEDDING MODELS

For old steganographic systems, it is easier to obtain the impact of embedding on the distribution of quantized DCT coefficients because the schemes are not adaptive to content. Instead, a fixed embedding operation is typically applied to a selected subset of coefficients with a fixed change rate β determined by the size of the secret payload to be embedded.

Using the GG model of cover DCT coefficients, we can express the total expected number of non-zero quantized DCT coefficients N_0, the number of DCT coefficients different from 0 and 1, N_{01}, and the number of non-zero AC DCT coefficients, $N_{0\mathrm{AC}}$, as

$$N_0 = n_1 n_2 \left(1 - \frac{1}{64}\sum_{k,l=0}^{7} P_{kl}^{(c)}(0)\right) \tag{8}$$

$$N_{01} = n_1 n_2 \left(1 - \frac{1}{64}\sum_{k,l=0}^{7} \left[P_{kl}^{(c)}(0) + P_{kl}^{(c)}(1)\right]\right) \tag{9}$$

$$N_{0\mathrm{AC}} = n_1 n_2 \left(1 - \frac{1}{64} - \frac{1}{64}\sum_{(k,l)\neq(0,0)} P_{kl}^{(c)}(0)\right). \tag{10}$$

3.1 Generic LSB flipper

By a generic LSB flipper (LSBF), we understand an algorithm that embeds messages by replacing the Least Significant Bits (LSBs) of pseudo-randomly selected quantized DCT coefficients that are not equal to 0 or 1 with message bits. For example, the embedding algorithm Jsteg falls into this category. LSB replacement is the most popular type of steganography because it is simple and can be applied to virtually any sampled signal. As of October 2017, out of 2863 tools available on the Internet capable of hiding data in digital images, 1024 (36%) of them embed secrets by manipulating LSBs.[1]

Assuming an absolute payload of M bits to be embedded, the probability of changing a quantized DCT coefficient not equal to zero or one is thus $\beta = M/(2N_{01})$, where N_{01} is the number of all DCT coefficients in the cover image not equal to zero or one, the maximum number of bits that can be embedded. In terms of the relative payload α in bits per non-zero AC DCT coefficient (bpnzac) and in terms of bits per pixel (bpp), $M = \alpha N_{0\mathrm{AC}}$ and $M = \alpha n_1 n_2$, respectively. Thus, using (9) and (10), the change rates w.r.t. N_{01} are

$$\beta = \frac{\alpha N_{0\mathrm{AC}}}{2N_{01}} \quad \alpha \text{ in bpnzac} \tag{11}$$

$$\beta = \frac{\alpha n_1 n_2}{2N_{01}} \quad \alpha \text{ in bpp}. \tag{12}$$

[1] N. Johnson, "IoT Forensic Considerations and Steganography Beyond Images." Invited talk presented at the Network and Cloud Forensics Workshop, IEEE Conference on Communications and Network Security, October 9–11, 2017, Las Vegas, Nevada, USA.

Quantized DCT coefficients in the stego image follow the p.m.f. $P^{(s)}_{kl}$, $0 \le k,l \le 7$:

$$\begin{aligned} P^{(s)}_{kl}(2m) &= (1-\beta)P^{(c)}_{kl}(2m) + \beta P^{(c)}_{kl}(2m+1) \quad m \neq 0 \\ P^{(s)}_{kl}(2m+1) &= \beta P^{(c)}_{kl}(2m) + (1-\beta)P^{(c)}_{kl}(2m+1) \quad m \neq 0 \\ P^{(s)}_{kl}(m) &= P^{(c)}_{kl}(m), \quad m \in \{0,1\}. \end{aligned} \tag{13}$$

3.2 OutGuess

OutGuess embedding proceeds in two stages – embedding and correction. First, the secret message is embedded using LSBR as in the generic LSBF. Then, more changes are introduced in unused DCT coefficients to preserve the global histogram of DCT coefficients. This introduces the following impact on quantized DCT coefficients in the stego image:

$$P^{(s)}_{kl}(2m) = \begin{cases} (1-\beta)P^{(c)}_{kl}(2m) + \beta \frac{P^{(c)}(2m)}{P^{(c)}(2m+1)} P^{(c)}_{kl}(2m+1) & m > 0 \\ (1-\beta)P^{(c)}_{kl}(2m) + \beta \frac{P^{(c)}(2m+1)}{P^{(c)}(2m)} P^{(c)}_{kl}(2m+1) & m < 0 \end{cases}$$

$$P^{(s)}_{kl}(2m+1) = \begin{cases} \beta P^{(c)}_{kl}(2m) + (1-\beta) \frac{P^{(c)}(2m)}{P^{(c)}(2m+1)} P^{(c)}_{kl}(2m+1) & m > 0 \\ \beta P^{(c)}_{kl}(2m) + (1-\beta) \frac{P^{(c)}(2m+1)}{P^{(c)}(2m)} P^{(c)}_{kl}(2m+1) & m < 0 \end{cases}$$

$$P^{(s)}_{kl}(m) = P^{(c)}_{kl}(m), \quad m \in \{0,1\}$$

where $P^{(c)}$stands for the global p.m.f. of DCT coefficients in the cover image.

3.3 nsF5

For nsF5, the maximum number of bits that can be embedded is equal to the number of non-zero AC DCT coefficients in the cover image, $N_{0\mathrm{AC}}$. Assuming optimal source coding, nsF5 modifies the fraction $\beta = H_2^{-1}(M/N_{0\mathrm{AC}})$ of all non-zero AC DCT coefficients, where H_2^{-1} is the inverse binary entropy function. For relative payload α,

$$\beta = H_2^{-1}\left(\frac{\alpha N_{0\mathrm{AC}}}{N_{0\mathrm{AC}}}\right) = H_2^{-1}(\alpha) \quad \alpha \text{ in bpnzac} \tag{14}$$

$$\beta = H_2^{-1}\left(\frac{\alpha n_1 n_2}{N_{0\mathrm{AC}}}\right) \quad \alpha \text{ in bpp.} \tag{15}$$

Quantized DCT coefficients in the stego image follow

$$\text{For } (k,l) \neq (0,0): \tag{16}$$

$$P^{(s)}_{kl}(m) = \begin{cases} (1-\beta)P^{(c)}_{kl}(m) + \beta P^{(c)}_{kl}(m+1) & m > 0 \\ p_{kl}(0) + \beta P^{(c)}_{kl}(1) + \beta P^{(c)}_{kl}(-1) & m = 0 \\ (1-\beta)P^{(c)}_{kl}(m) + \beta P^{(c)}_{kl}(m-1) & m < 0 \end{cases} \tag{17}$$

$$P^{(s)}_{00}(m) = P^{(c)}_{00}(m).$$

3.4 LSBM

We also work out the impact for a generic embedding scheme that uses LSB matching (LSBM) applied to all non-zero DCT coefficients. Even though such an embedding scheme has not been proposed before, it does make sense to include this case in our study for completeness. Denoting the number of all non-zero DCT coefficients with N_0, under optimal source coding the total change rate applied to each non-zero DCT is $\beta = H_3^{-1}(M/N_0)$, where H_3^{-1} is the inverse ternary entropy. For relative payload α,

$k\backslash l$	0	1	2	3	4	5	6	7
0	2.24	0.43	0.40	0.38	0.37	0.37	0.36	0.35
1	0.48	0.46	0.43	0.43	0.42	0.42	0.41	0.40
2	0.45	0.45	0.44	0.42	0.42	0.42	0.41	0.41
3	0.45	0.45	0.43	0.43	0.42	0.42	0.42	0.41
4	0.44	0.45	0.44	0.42	0.43	0.42	0.42	0.41
5	0.43	0.45	0.44	0.43	0.43	0.42	0.42	0.41
6	0.41	0.44	0.43	0.43	0.42	0.43	0.42	0.42
7	0.40	0.42	0.42	0.42	0.42	0.42	0.42	0.41

$k\backslash l$	0	1	2	3	4	5	6	7
0	709	2.89	1.06	0.52	0.32	0.21	0.14	0.08
1	5.87	2.24	1.08	0.68	0.49	0.34	0.25	0.15
2	2.27	1.47	0.89	0.53	0.39	0.29	0.21	0.15
3	1.46	1.05	0.67	0.49	0.36	0.27	0.19	0.14
4	0.91	0.76	0.57	0.39	0.31	0.24	0.18	0.12
5	0.61	0.57	0.45	0.33	0.27	0.20	0.16	0.11
6	0.36	0.42	0.32	0.27	0.21	0.17	0.13	0.10
7	0.22	0.25	0.22	0.19	0.16	0.13	0.10	0.08

Table 1: Top/bottom: Shape/width parameter of GG models of unquantized DCT coefficients in each DCT mode (k,l) estimated from 2000 randomly selected BOSSbase images.

$k\backslash l$	0	1	2	3	4	5	6	7
0	.16807	.26092	.23824	.07496	.05008	.00831	.00485	.00395
1	.26807	.22638	.20590	.05708	.01411	.00121	.00159	.00252
2	.24810	.20875	.07469	.04893	.00455	.00088	.00063	.00196
3	.20128	.06112	.05147	.01152	.00102	.00006	.00018	.00119
4	.05848	.04286	.00513	.00048	.00011	.00001	.00004	.00025
5	.05434	.00646	.00105	.00051	.00006	.00002	.00003	.00021
6	.00630	.00202	.00044	.00012	.00004	.00002	.00005	.00030
7	.00311	.00067	.00030	.00015	.00006	.00014	.00032	.00069

Table 2: Average change rates $\overline{\beta}_{kl}$ across DCT modes (k,l) for J-UNIWARD at 0.4 bpnzac for JPEG QF 95 in BOSSbase.

$$\beta = H_3^{-1}\left(\frac{\alpha N_{0\mathrm{AC}}}{N_0}\right) \quad \alpha \text{ in bpnzac} \tag{18}$$

$$\beta = H_3^{-1}\left(\frac{\alpha n_1 n_2}{N_0}\right) \quad \alpha \text{ in bpp.} \tag{19}$$

The stego p.m.f. of quantized DCT coefficients is for $|m| > 1$, $|m| = 1$, and $m = 0$, respectively, and for all k,l :

$$P^{(s)}_{kl}(m) = \begin{cases} (1-\beta)P^{(c)}_{kl}(m) + \frac{\beta}{2}P^{(c)}_{kl}(m+1) + \frac{\beta}{2}P^{(c)}_{kl}(m-1) \\ (1-\beta)P^{(c)}_{kl}(m) + \frac{\beta}{2}P^{(c)}_{kl}\left(m + \frac{m}{|m|}\right) \\ P^{(c)}_{kl}(0) + \frac{\beta}{2}P^{(c)}_{kl}(1) + \frac{\beta}{2}P^{(c)}_{kl}(-1) \end{cases} \tag{20}$$

3.5 J-UNIWARD

The steganographic scheme J-UNIWARD modifies quantized DCT coefficients with probabilities determined by the local content of the cover image. This non-stationarity significantly complicates

modeling the impact of embedding. For simplicity, we will assume that J-UNIWARD applies a certain change rate $\overline{\beta}_{kl}$ to all coefficients (including zeros and the DC term) from mode (k, l) in all blocks. These change rates will be determined by averaging the change rates in each DCT mode across a number of images for each JPEG quality factor Q and payload α separately (Section 4.2). The impact on the p.m.f. of each DCT mode will thus be for all k, l, m :

$$P^{(s)}_{kl}(m) = (1 - \overline{\beta}_{kl})P^{(c)}_{kl}(m) + \frac{\overline{\beta}_{kl}}{2}P^{(c)}_{kl}(m+1) \tag{21}$$

$$+ \frac{\overline{\beta}_{kl}}{2}P^{(c)}_{kl}(m-1). \tag{22}$$

Allowing the change rate to be different across the modes captures the fact that the cost of an embedding change in J-UNIWARD depends on the quantization step q_{kl} and thus on the DCT mode. This model is limited, however, because it does not capture the content adaptivity of J-UNIWARD.

3.6 UED-JC

In UED steganography (Uniform Embedding Distortion), the cost of changing a DCT coefficient is proportional to its reciprocal value (UED-SC algorithm as originally introduced in [3]). The more advanced version called UED-JC [4] considers four intra and inter-block neighbors of the coefficient to determine the cost (see Section III-C in [4]). This makes the embedding adaptive to content.

To model the impact of embedding, we adopt the same simplification as for J-UNIWARD – the change rates are assumed to depend on the spatial frequency k, l but not on the physical location within the image as in Eq. (21), and are estimated from a set of images for each quality factor as explained in the next section.

3.7 Security

Security will be measured with the KL divergence between the cover and stego p.m.f.s :

$$D_{\mathrm{KL}}(P^{(c)}||P^{(s)}) \triangleq \sum_{k,l=0}^{7} D_{\mathrm{KL}}(P^{(c)}_{kl}||P^{(s)}_{kl}) \tag{23}$$

$$= \sum_{k,l=0}^{7}\sum_{m=-L}^{L} P^{(c)}_{kl}(m)\log\frac{P^{(c)}_{kl}(m)}{P^{(s)}_{kl}(m)}, \tag{24}$$

where, for numerical evaluation, L was selected to obtain $P^{(c)}_{kl}(m) < 10^{-15}$ for $|m| > L$.

4 DATASETS AND MODEL ESTIMATION

All experiments in this paper were carried out on the union of BOSSbase 1.01 and BOWS2 datasets, each with 10,000 grayscale images, resized from their original size 512×512 to 256×256 using `imresize` with default setting in Matlab. Cover JPEG images were obtained in Matlab using the command `imwrite`. The decompression to the spatial domain for experiments with empirical detectors was obtained by multiplying the DCT coefficients by quantization steps and applying the block inverse DCT without rounding or clipping, `idct2`, in Matlab.

For training empirical detectors, we randomly selected 4,000 images from BOSSbase and the entire BOWS2 dataset with 1,000 BOSSbase images set aside for validation. The remaining 5,000 BOSSbase images were used for testing. In summary, $2 \times 14,000$ cover and stego images were used for training, $2 \times 1,000$ for validation, and $2 \times 5,000$ for testing. This dataset and the split into training and testing has been used for design of many modern deep learning architectures for steganalysis, including the YeNet [15], the Yedroudj-Net [16], and the SRNet [1].

4.1 Estimating GG models of DCT modes

A total of $N = 2000$ grayscale uncompressed images were selected from BOSSbase at random and subjected to block-wise DCT without quantization or rounding. The GG parameters shown in Table 1 were estimated from all N images using the method of moments [9] for each DCT mode (k, l) separately. Note that the DC term was approximated with a rather wide distribution similar to a Gaussian ($\gamma = 2.24$) while all AC modes exhibit spiky distributions with a similar value of the shape parameter, $0.35 \le \gamma \le 0.48$, with the vast majority around $\gamma \approx 0.42$ but a widely varying width $0.08 \le w \le 5.87$.

4.2 Estimating change rates for J-UNIWARD and UED-JC

Different N randomly chosen images were used for computing the average change rates $\overline{\beta}_{kl}(\alpha, Q)$ for each DCT mode (k, l), payload α, and quality factor Q. Let $\beta^{(a,b)}_{kl}(\mathbf{x}, \alpha, Q)$ denote the change rates returned by the embedding simulator for (a, b)th block in image $\mathbf{x}$. The values $\overline{\beta}_{kl}$ were obtained as averages over all blocks (a, b) and all N images $\mathbf{x}$:

$$\overline{\beta}_{kl}(\alpha, Q) = \frac{64}{Nn_1n_2}\sum_{a=1}^{n_1/8}\sum_{b=1}^{n_2/8}\sum_{\mathbf{x}}\beta^{(a,b)}_{kl}(\mathbf{x}, \alpha, Q). \tag{25}$$

For compactness, in the rest of this paper we will often drop the explicit dependence of $\overline{\beta}_{kl}$ on α and Q.

Table 2 shows an example of the average change rates $\overline{\beta}_{kl}$ for J-UNIWARD for quality factor 95 and relative payload 0.4 bpnzac. Note that the change rate is the largest for low spatial frequencies and much smaller for high frequencies. This is because the embedding costs of J-UNIWARD are larger for larger quantization steps q_{kl}, which roughly correspond to higher spatial frequencies.

5 EXPERIMENTS

In this section, we report the results of all experiments, which include the accuracy of empirical detectors as a function of the JPEG quality factor for several algorithms and payloads contrasted with the KL divergence computed from the model of JPEG coefficients introduced in Section 2. The investigation focuses on the case when the relative payload is fixed in terms of bpnzac because it is far more interesting than for bpp, which we briefly comment upon in Section 5.3.

5.1 Modern steganography

The initial investigation deals with J-UNIWARD [7]. Two types of empirical detectors were studied: the ensemble classifier [8] with Gabor Filter Residual (GFR) features [12], as a representative of the paradigm of rich models, and the Steganalysis Residual Network

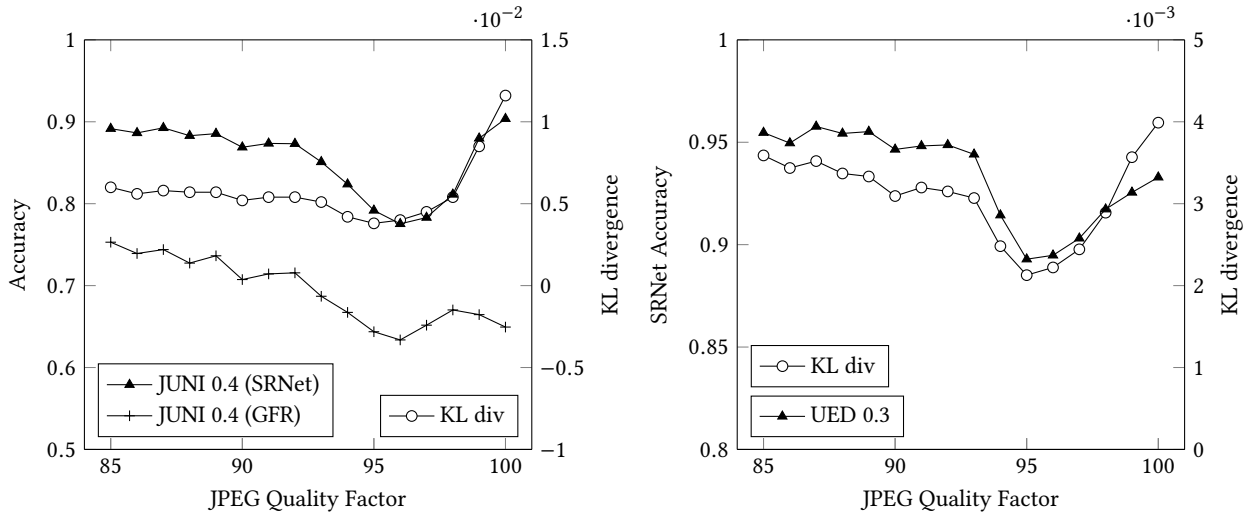

Figure 1: Left: Detection accuracy of J-UNIWARD as a function of quality factor for payload 0.4 bpnzac using SRNet and GFR with ensemble (left axis) and the KL divergence between cover and stego models for the same payload (right axis). Right: UED-JC for 0.3 bpnzac.

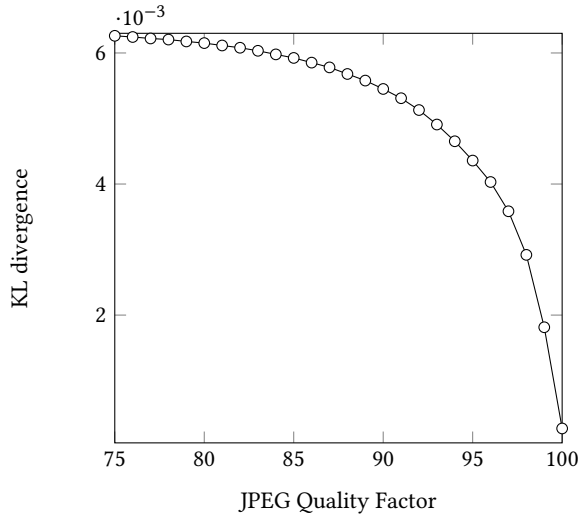

Figure 2: KL divergence as a function of the QF for J-UNIWARD at 0.4 bpnzac when using non-rounded and non-maximized quantization matrices.

(SRNet) [1] as a representative of detectors built using deep learning. Based on the experiments reported in [1], the SRNet currently provides the most accurate detection of modern JPEG steganography over other competing architectures designed for the JPEG domain [14, 17].

Figure 1 left shows the performance of both detectors for payload 0.4 bpnzac across JPEG qualities 85–100 in terms of the correct classification accuracy $1 - P_{\text{E}}$, where

$$P_{\text{E}} = \min \frac{1}{2}(P_{\text{MD}} + P_{\text{FA}}) \tag{26}$$

is the often used minimum average detection error under equal priors, and P_{MD} and P_{FA} the missed-detection and false-alarm rates. The right y axis shows the scale of the KL divergence (23) computed between the cover model (5) and the stego model of J-UNIWARD (21). With the exception of GFR for quality 99 and 100, both empirical detectors closely mimic the variations of the KL divergence across all quality factors, including the small "ripples" at 86, 88, 90, and 93, due to rounding and clipping of the quantization steps (1) as well as the minimum around quality 95–96. To confirm the origin of the ripples, in Figure 2 we show the KL divergence for J-UNIWARD at 0.4 bpnzac, when the quantization steps q_{kl} are not rounded to integers and not clipped to 1 (when removing "max" and rounding "[.]" in (2)) with $\mathbf{q}(100) \triangleq \mathbf{q}(99)/10$ as (1) would produce a matrix of zeros for quality 100. The KL divergence for J-UNIWARD in this case monotonically and smoothly decreases with increased quality factor Q.

Furthermore, still inspecting Figure 1, the SRNet provides markedly better detection than GFR. In particular, GFR appears to significantly under-perform w.r.t. the SRNet for quality factors above 98. For the two largest quality factors 99 and 100, the KL divergence predicts that the detection should be much more accurate than what the SRNet exhibits, perhaps indicating a possible space for improvement. Since the SRNet generally offers much better detection than GFR, all remaining experiments, unless otherwise mentioned, are executed with the SRNet as the empirical detector.

In Figure 1 right, we show the detectability of UED-JC across quality factors for a fixed payload 0.3 bpnzac. The trends of the empirical detector, including the small variations between QF 85 and 91 due to quantization step rounding again closely match the KL divergence computed between the models. As with J-UNIWARD, the KL divergence values seem to suggest that the empirical detector under-performs for qualities near 100.

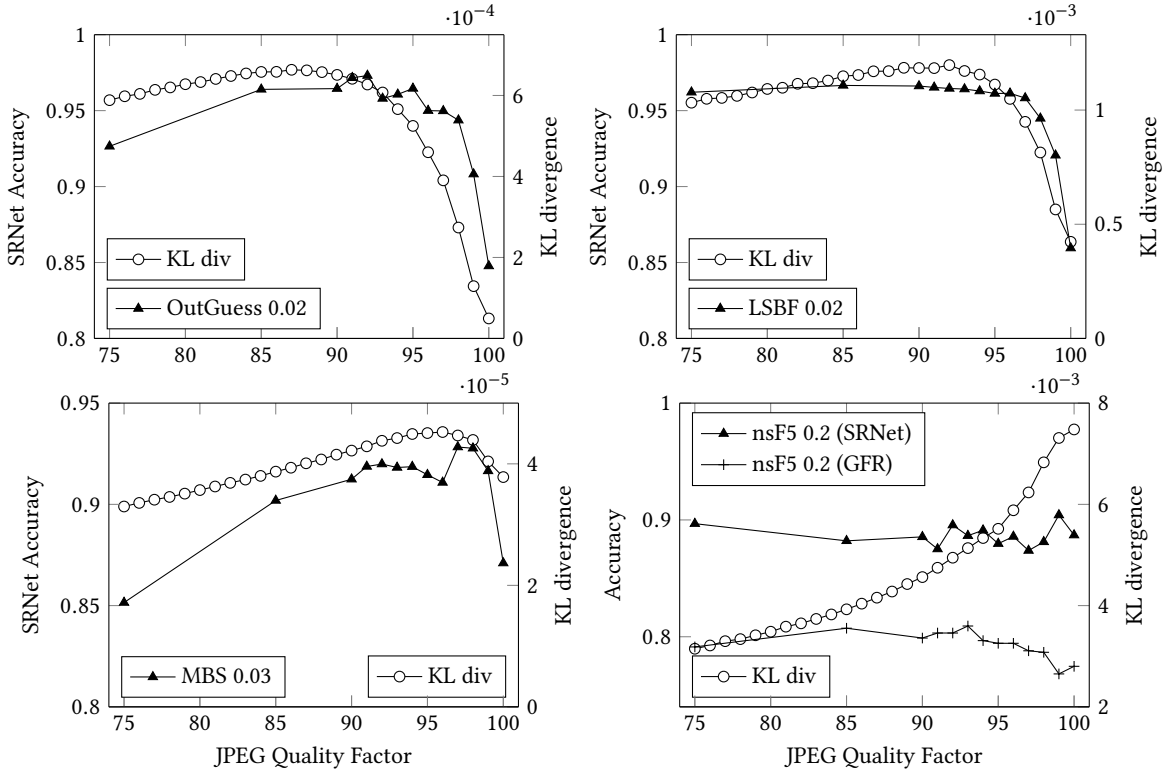

Figure 3: By rows: Detection accuracy of SRNet for OutGuess at 0.02 bpnzac, LSBF at 0.02 bpnzac, MBS at 0.03 bpnzac, and nsF5 at 0.2 bpnzac (left axis). The right axis is for the KL divergence (23) between cover and the corresponding stego models.

Before we move to older steganographic paradigms in the next section, we note that for the experiments reported above, the SRNet was initially trained as described in the original publication [1] from randomly initialized filters for quality factor 85 as this is when both J-UNIWARD and UED-JC are the most detectable. Curriculum training via the quality factor was used to train for 86, 87, . . ., 100 and was always run for 100k iterations with LR 10^{-3} after which the LR was lowered to 10^{-4} for an additional 50k iterations.

5.2 Old steganography

In this section, the relationship between the empirical detection accuracy and the KL divergence between the cover and stego models has been investigated for a generic LSB flipper, OutGuess, model-based steganography (MBS), generic ternary embedding in non-zero DCT coefficients (LSBM), and nsF5. The results are summarized in graphical form in Figure 3.

In contrast to J-UNIWARD and UED-JC, except for nsF5, all embedding methods exhibit the same qualitative trend – their empirical security decreases with increasing quality factor but this trend eventually reverses for larger quality factors. Since the details of how MBS handles embedding a payload smaller than the maximal payload have not been available to the authors, the KL divergence displayed in the graph showing MBS is for LSBM.

The corresponding KL divergence between the models relatively well matches the empirical results. The nsF5 was the only embedding algorithm for which the KL divergence exhibited a different trend than the empirical detectors (Figure 3 bottom right). While both the SRNet and the ensemble with GFR exhibit approximately constant detectability, the model predicts an increasing KL divergence. This could mean that either our model fails to capture the impact of embedding correctly for this algorithm or that the empirical detectors increasingly under-perform for larger quality factors. We hypothesize that the latter explanation is more likely for the following reason. The increase of the KL divergence for nsF5 is primarily due to the increased number of zeros in stego images since the embedding operation always decreases the absolute value of DCT coefficients. Detecting this increase or, equivalently, estimating the number of zeros in the cover from the stego image, however, seems to be a difficult task in practice. We intend to investigate this issue as part of our future effort.

For all embedding methods, the SRNet was first trained from scratch for QF 95 because this is the range with the easiest detection. The detectors for the remaining QFs were trained using curriculum training via the quality factor in quality factor steps of one.

5.3 Fixed bpp

For completeness, we briefly report the results obtained when the payload is fixed in terms of bits per pixel (bpp) rather than bpnzac.

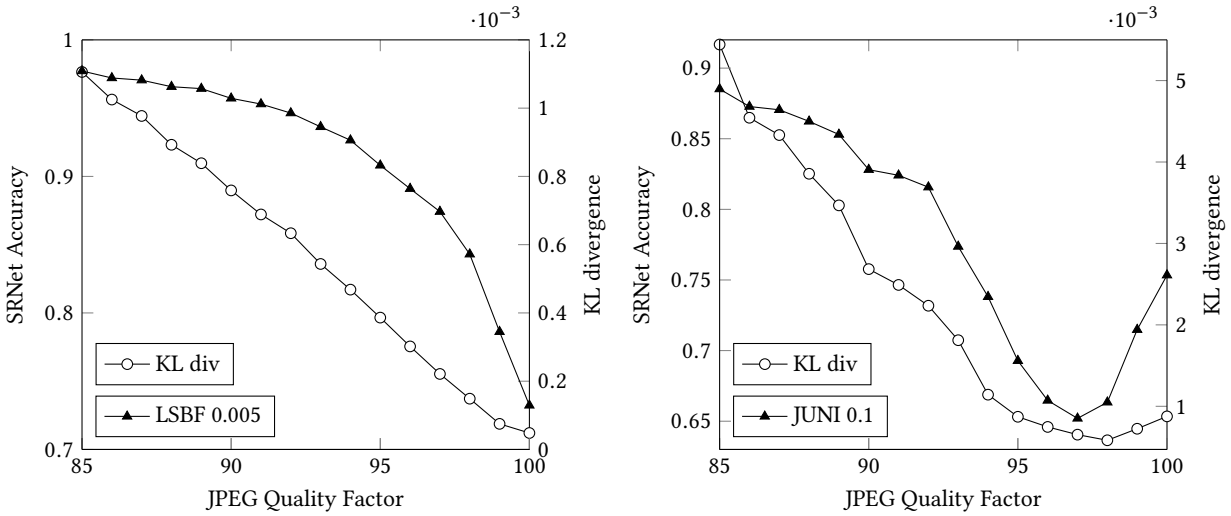

Figure 4: Accuracy of SRNet and the KL divergence between cover and stego models for LSBF at 0.005 bpp (left) and J-UNIWARD (right) at 0.1 bpp.

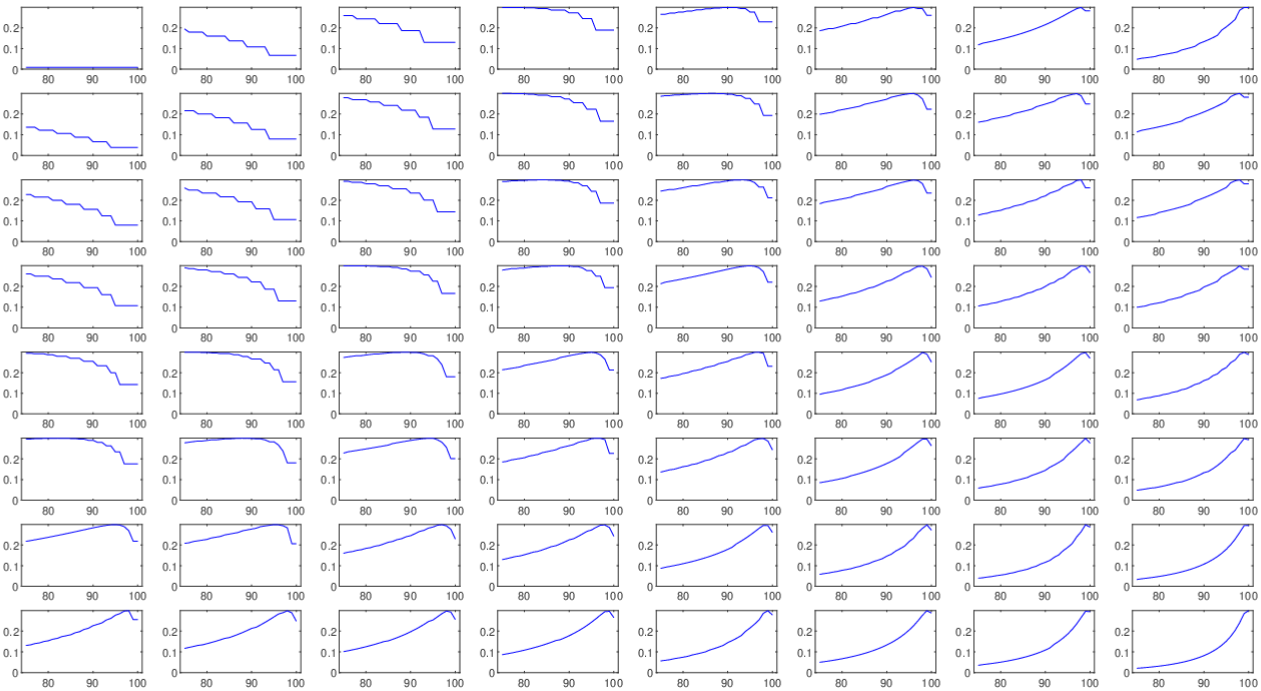

Figure 5: Fisher information $I_{kl}(Q)$ for generic LSBM as a function of the quality factor $75 \leq Q \leq 100$ for all 64 DCT modes with k and l corresponding to rows and columns, respectively.

A relative payload fixed in terms of bpp means that the same number of bits is embedded for all quality factors and all steganographic algorithms. Since the number of non-zero DCT coefficients strictly increases with increased quality factor, the "effective size" of the cover for old steganography paradigms increases. Our model predicts a strictly decreasing KL divergence for all old stego methods. As an example, in Figure 4 left we show the SRNet accuracy and the KL divergence for LSBF at payload 0.1 bpp.

In contrast, for modern steganography, the detectability decreases but starts increasing for qualities close to 100. Figure 4 right shows the detection accuracy of the SRNet and the KL divergence between cover and stego models for J-UNIWARD when

fixing the relative payload at 0.1 bpp. The model correctly predicts the lowest detectability around 97–98 as well as the small "ripples" between 85 and 93.

6 ANALYSIS

In this section, we present a more intuitive explanation of the complementary security trends observed for old and modern steganography. This requires inspecting in more detail how the KL divergence of individual DCT modes changes with increasing quality factor. We first study old steganography paradigms and then modern schemes.

6.1 Old steganography

We work with the generic LSBM (20) with global change rate β w.r.t. all non-zero DCT coefficients as this will simplify our arguments. The leading term of the Taylor expansion of the KL divergence (23) with respect to β is :

$$D_{\mathrm{KL}}(P^{(c)}||P^{(s)}) \doteq \frac{\beta^2}{2}\sum_{k,l=0}^{7} I_{kl}, \tag{27}$$

where

$$I_{kl} = \sum_{m} \frac{1}{P^{(c)}_{kl}(m)}\left(\frac{\partial P^{(s)}_{kl}(m)}{\partial \beta_{kl}}\bigg|_{\beta_{kl}=0}\right)^2. \tag{28}$$

is the steganographic Fisher information for mode (k,l). Thus, to understand the trends w.r.t. the quality factor Q, we need to inspect I_{kl} as a function of Q. First, we take a look at the range $Q \leq 95$.

Figure 5 shows $I_{kl}(Q)$ for $75 \leq Q \leq 100$ with the y-axis scale unified across all modes. Note that the Fisher information for low frequency modes decreases, it exhibits a non-monotone trend for medium frequencies, and sharply increases for high frequencies. With increasing Q, the increase in $I_{kl}(Q)$ for high spatial frequencies is larger than the decrease of $I_{kl}(Q)$ for low spatial frequencies, which clarifies the security trend of old embedding methods observed in the previous section. Note that this trend can be reversed by letting the change rates decrease with increasing $k+l$ as is the case for modern steganography.

The seemingly complex behavior of $I_{kl}(Q)$ w.r.t. Q is caused by the fact that old steganography does not embed into zeros. To see why, we point out that the cover p.m.f. $P^{(c)}_{kl}$ (5) depends only on the ratio $w_{kl}/q_{kl}(Q)$ (see Eq. (6)), the effective width of the GG model after quantizing the (k,l)th mode with quantization step $q_{kl}(Q)$. Figure 6 (solid line, right y-axis) shows the Fisher information I as a function of the ratio w/q for $\gamma = 0.4$.[2] Note that I exhibits a maximum at $w/q \approx 0.04$. In contrast, when allowing embedding into zeros, I becomes strictly decreasing w.r.t. Q (the dashed line, left y-axis shows $\log_{10} I$ in Figure 6). For DCT modes (k,l) for which $w_{kl}/q_{kl}(Q) \leq 0.04$, increasing the quality factor leads to increased Fisher information $I_{kl}(Q)$. This occurs for high frequency modes because the width of their GG fit is smaller (Table 1). For modes with $w_{kl}/q_{kl}(Q) \geq 0.04$, $I_{kl}(Q)$ decreases with increased Q. The non-monotone behavior of $I_{kl}(Q)$ for medium frequencies is due to the ratio $w_{kl}/q_{kl}(Q)$ moving past 0.04 as Q increases.

[2] From Table 1, $\gamma \approx 0.4$ across all AC modes.

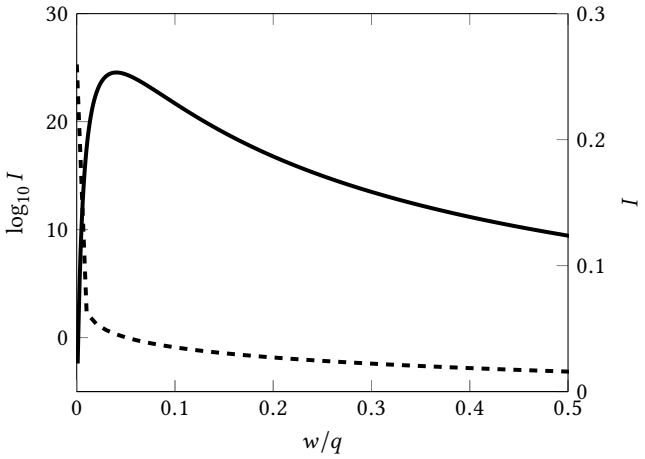

Figure 6: Solid line and right y-axis: Fisher information I of LSBM as a function of the ratio w/q for $\gamma = 0.4$. Dashed line and left y-axis: Logarithm of Fisher information of LSBM when embedding into zeros.

Once $Q > 95$, for low spatial frequencies the quantization steps $q_{kl}(Q)$ start "flattening out" at 1, which means that the ratio $w_{kl}/q_{kl}(Q)$ stops increasing and thus $I_{kl}(Q)$ no longer decreases with Q. The Fisher information of high and medium-frequency modes start decreasing as the ratios $w_{kl}/q_{kl}(Q)$ grow larger than 0.04, eventually reversing the detectability trend for old steganography.

6.2 Modern steganography

Explaining the trend reversal for modern steganography is more complicated due to the modulation of change rates across spatial frequencies and their dependence on the quality factor. We can no longer factor out the global change as in (27) and need to consider $d_{kl}(Q) = \overline{\beta}^2_{kl}(Q) I_{kl}(Q)$ as functions of Q.

Generally speaking, for low-frequency modes, $d_{kl}(Q)$ decrease with increasing Q. For medium and high frequencies, however, d_{kl} starts to sharply increase for $Q > 95$ (see Figure 7). This rapid increase is responsible for the reversal of the detectability trend observed for modern embedding schemes for high quality factors, which holds for fixed relative payload in both bpnzac and bpp :

> Attenuation of change rates across modes is not optimal.

This observation gives a clue on a possible improvement of J-UNIWARD, which we briefly delve into in the next section.

6.3 Improving J-UNIWARD

In the previous section, we concluded from our model that the increase of detectability (KL divergence) of J-UNIWARD for high quality factors is due to improper modulation of embedding change rates. In particular, the change rates for high spatial frequencies should be attenuated more aggressively than what the embedding distortion of J-UNIWARD dictates. This shows a possible way to improve its security.

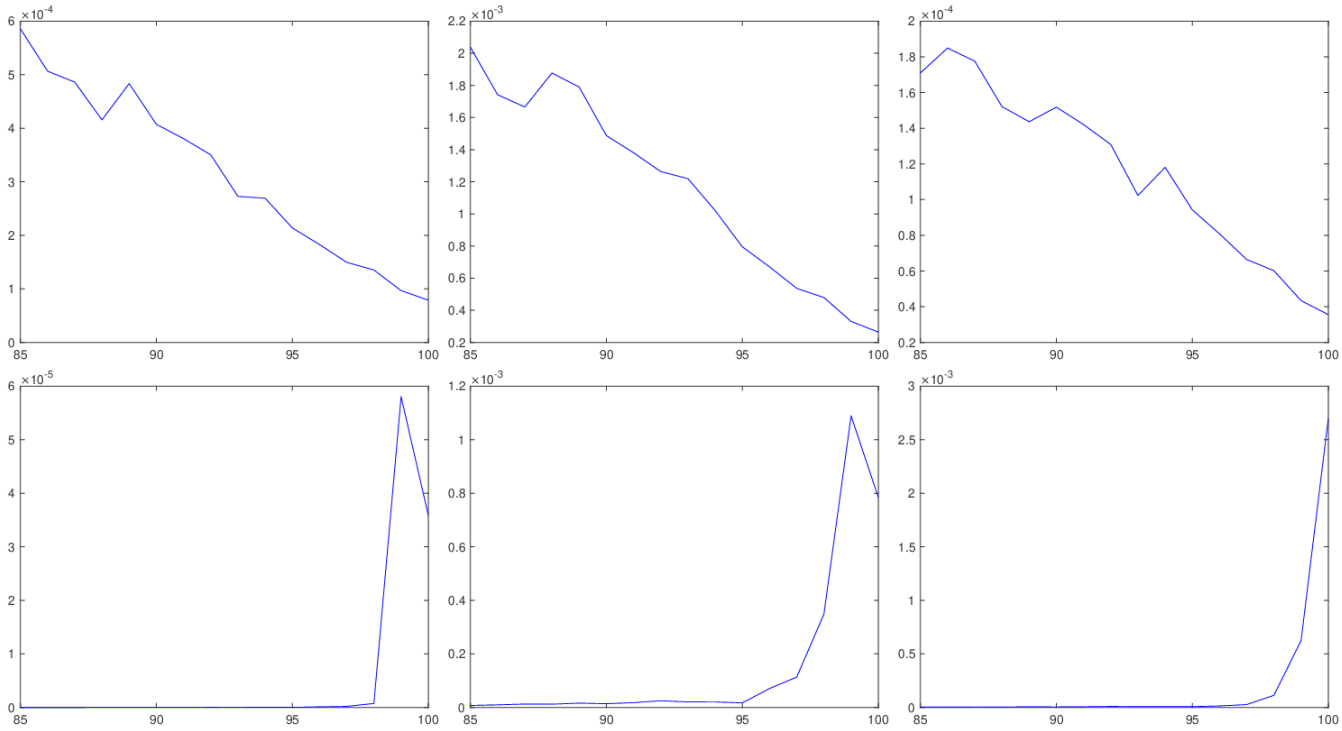

Figure 7: By rows: the leading term of the KL divergence $d_{kl}(Q)$ in modes $(0,1),(0,2),(1,0),(4,4),(1,7),(7,7)$ as a function of the quality factor Q. J-UNIWARD, 0.4 bpnzac.

Recalling that $\overline{\beta}_{kl}(\alpha, Q)$ is the average change rate applied by J-UNIWARD to mode (k,l) for a given payload α and quality factor Q, we find $\tilde{\beta}_{kl}(\alpha, Q)$, minimizing the leading term of the KL divergence while communicating on average the same entropy:

$$\min_{\tilde{\beta}_{kl}} \sum_{k,l=0}^{7} \tilde{\beta}_{kl}^2(\alpha, Q) I_{kl}(Q) \tag{29}$$

$$\sum_{k,l=0}^{7} H_3(\tilde{\beta}_{kl}) = \sum_{k,l=0}^{7} H_3(\overline{\beta}_{kl}). \tag{30}$$

Since the DC term is difficult to model, we avoid optimizing it and instead set $\tilde{\beta}_{00} = \overline{\beta}_{00}$. The change rates $\tilde{\beta}_{kl}$, $k + l > 0$, found in this manner are indeed smaller for high frequencies ($k > 5$ or $l > 5$) and larger for low and medium frequencies. Figure 8 shows $\overline{\beta}_{kl}$ and $\tilde{\beta}_{kl}$ for $Q = 100$ and relative payload $\alpha = 0.1$ bpp. Note that $\tilde{\beta}_{kl} > \overline{\beta}_{kl}$ for low frequencies and $\tilde{\beta}_{kl} < \overline{\beta}_{kl}$ for high spatial frequencies. Also, while $\tilde{\beta}_{kl}$ decrease with increased frequency, the smallest values of $\overline{\beta}_{kl}$ roughly correspond to the largest entries in $\mathbf{q}(50)$ (see Eq. (2)).

To incorporate this adjustment into the embedding algorithm, we first convert both $\overline{\beta}_{kl}$ and $\tilde{\beta}_{kl}$ to embedding costs

$$\tilde{\varrho}_{kl} = \ln\left(\frac{1}{\tilde{\beta}_{kl}} - 2\right) \tag{31}$$

$$\overline{\varrho}_{kl} = \ln\left(\frac{1}{\overline{\beta}_{kl}} - 2\right). \tag{32}$$

Given the matrix of J-UNIWARD's embedding costs in (a,b)th 8×8 block of image $\mathbf{x}$ as $\rho_{kl}^{(a,b)}(\mathbf{x})$, we modulate them

$$\rho_{kl}^{(a,b)}(\mathbf{x}) \to \rho_{kl}^{(a,b)}(\mathbf{x}) \frac{\tilde{\varrho}_{kl}}{\overline{\varrho}_{kl}}. \tag{33}$$

These modulated costs would then be used in an embedding simulator or STCs for practical embedding in image $\mathbf{x}$. Note that the modulation (33) depends on payload α as well as the quality factor Q.

This heuristic adjustment of the embedding change rates did indeed improve J-UNIWARD's security. For quality factor 100, the accuracy of SRNet decreased by 2.14%. The network detector was trained by seeding with detector trained on J-UNIWARD for $\alpha = 0.1$ bpp and the corresponding quality factor. The LR was 10^{-3} for the first 100k iterations, lowered to 10^{-4} for an additional 50k iterations.

To further validate this approach, we carried out the same experiment for quality factors 100 for J-UNIWARD at $\alpha = 0.4$ bpnzac. In this setting, the security was improved by 1.12% in terms of SRNet accuracy.

Due to limited space and time, we postpone a more detailed investigation to our future work. In particular, more detailed study needs to be executed regarding the change rate adjustment across payloads and quality factors as well as for other embedding schemes. The limited experiment in this section should thus be thought

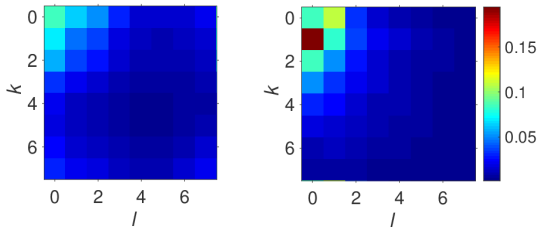

Figure 8: Left: $\overline{\beta}_{kl}$, right: $\tilde{\beta}_{kl}$ for quality factor $Q = 100$ and relative payload $\alpha = 0.1$ bpp.

of more as a promising direction and additional evidence for the predictive power of our theoretical approach.

7 CONCLUSIONS

This paper investigates how the detectability of JPEG steganography changes with the quality factor when fixing the relative payload. While older embedding paradigms become progressively more detectable up until quality 90–95 after which their detectability decreases, modern steganography exhibits complementary trends. This behavior is explained by modeling a JPEG file as 64 independent channels with a quantized generalized Gaussian distribution. The KL divergence between cover and stego distributions closely matches the detectability obtained with empirical detectors. The only tested algorithm for which our theoretical analysis failed to match the results of empirical detectors is nsF5. We hypothesize that this is due to the inability of empirical detectors to assess the number of zeros in a JPEG file, indicating a possible improvement of detection of nsF5 for larger quality factors.

By analyzing the Fisher information as a function of the width of the GG model, we offer a more intuitive explanation of the observed trends. For old embedding paradigms, the contribution of high-frequency modes to detectability increases faster with increased quality than the decrease in detectability in low-frequency modes. This trend can be reversed by decreasing the change rates with increased spatial frequency. For modern steganography, the loss of security of J-UNIWARD for high quality factors has been linked to slightly improper modulation of change rates across spatial frequencies. A heuristic adjustment of the change rates based on the insight obtained from the model indeed lead to an improved security of J-UNIWARD for quality factor 100.

A by-product of our analysis is a better understanding of why older embedding paradigms are much less secure than modern schemes: the comparatively large change rates for high-frequency modes in older schemes substantially increase the KL divergence but contribute little to the total payload because they contain fewer nonzero coefficients. Modern steganography addresses this problem by decreasing the change rate with increasing spatial frequency.

Numerous imaging devices and image editing software use nonstandard quantization matrices, which were not investigated in this work. However, the authors are fairly confident that the findings of this paper qualitatively generalize to custom quantization matrices with respect to a generalized concept of JPEG quality defined by a suitably chosen distance (metric) between quantization matrices.

Despite the fact that our model cannot properly capture content adaptivity of modern steganography, its predictive power allowed us to explain the security trends w.r.t. JPEG quality factor and improve the security of J-UNIWARD for the largest quality factor. The use of the model for steganography is a topic that deserves a more extensive study and is thus left for future research.

ACKNOWLEDGMENTS

The work on this paper was supported by NSF grant No. 1561446 and by DARPA under agreement number FA8750-16-2-0173. The U.S. Government is authorized 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 DARPA or the U.S. Government.

REFERENCES

[1] M. Boroumand, M. Chen, and J. Fridrich. Deep residual network for steganalysis of digital images. *IEEE Transactions on Information Forensics and Security*, 14(5):1181–1193, May 2019.

[2] M. Chen, M. Boroumand, and J. Fridrich. Reference channels for steganalysis of images with convolutional neural networks. In R. Cogranne and L. Verdoliva, editors, *The 7th ACM Workshop on Information Hiding and Multimedia Security*, Lecture Notes in Computer Science, Paris, France, 2019.

[3] L. Guo, J. Ni, and Y.-Q. Shi. An efficient JPEG steganographic scheme using uniform embedding. In *Fourth IEEE International Workshop on Information Forensics and Security*, Tenerife, Spain, December 2–5, 2012.

[4] L. Guo, J. Ni, and Y. Q. Shi. Uniform embedding for efficient JPEG steganography. *IEEE Transactions on Information Forensics and Security*, 9(5):814–825, May 2014.

[5] V. Holub and J. Fridrich. Challenging the doctrines of JPEG steganography. In A. Alattar, N. D. Memon, and C. Heitzenrater, editors, *Proceedings SPIE, Electronic Imaging, Media Watermarking, Security, and Forensics 2014*, volume 9028, pages 02–1–02–7, San Francisco, CA, February 3–5, 2014.

[6] V. Holub and J. Fridrich. Low-complexity features for JPEG steganalysis using undecimated DCT. *IEEE Transactions on Information Forensics and Security*, 10(2):219–228, February 2015.

[7] V. Holub, J. Fridrich, and T. Denemark. Universal distortion design for steganography in an arbitrary domain. *EURASIP Journal on Information Security, Special Issue on Revised Selected Papers of the 1st ACM IH and MMS Workshop*, 2014:1, 2014.

[8] J. Kodovský, J. Fridrich, and V. Holub. Ensemble classifiers for steganalysis of digital media. *IEEE Transactions on Information Forensics and Security*, 7(2):432–444, April 2012.

[9] S. Meignen and H. Meignen. On the modeling of DCT and subband image data for compression. *IEEE Transactions on Image Processing*, 4(2):186–193, February 1995.

[10] N. Provos. Defending against statistical steganalysis. In *10th USENIX Security Symposium*, pages 323–335, Washington, DC, August 13–17, 2001.

[11] P. Sallee. Model-based methods for steganography and steganalysis. *International Journal of Image Graphics*, 5(1):167–190, 2005.

[12] X. Song, F. Liu, C. Yang, X. Luo, and Y. Zhang. Steganalysis of adaptive JPEG steganography using 2D Gabor filters. In P. Comesana, J. Fridrich, and A. Alattar, editors, *The 3rd ACM IH&MMSec. Workshop*, Portland, Oregon, June 17–19, 2015.

[13] D. Upham. Steganographic algorithm JSteg. Software available at http://zooid.org/ paul/crypto/jsteg.

[14] G. Xu. Deep convolutional neural network to detect J-UNIWARD. In M. Stamm, M. Kirchner, and S. Voloshynovskiy, editors, *The 5th ACM Workshop on Information Hiding and Multimedia Security*, Philadelphia, PA, July 3–5, 2017.

[15] J. Ye, J. Ni, and Y. Yi. Deep learning hierarchical representations for image steganalysis. *IEEE Transactions on Information Forensics and Security*, 12(11):2545–2557, November 2017.

[16] M. Yedroudj, F. Comby, and M. Chaumont. Yedroudj-net: An efficient CNN for spatial steganalysis. In *IEEE ICASSP*, pages 2092–2096, Alberta, Canada, April 15–20, 2018.

[17] J. Zeng, S. Tan, B. Li, and J. Huang. Large-scale JPEG image steganalysis using hybrid deep-learning framework. *IEEE Transactions on Information Forensics and Security*, 13(5):1200–1214, 2018.