EXTENDING THE REVERSE JPEG COMPATIBILITY ATTACK TO DOUBLE COMPRESSED IMAGES

Jan Butora and Jessica Fridrich, Fellow, IEEE

Binghamton University
Department of ECE
Binghamton, NY
{jbutora1,fridrich}@binghamton.edu

ABSTRACT

The reverse JPEG compatibility attack has recently been introduced as a very accurate and universal steganalysis algorithm for JPEG images with quality 99 or 100. The limitation to these two largest qualities appears fundamental as the prior work on this topic suggests. In this paper, we provide mathematical analysis and demonstrate experimentally that this attack can be extended to double compressed images when the first compression quality is 93 or larger and the second quality equal or larger than the first quality. Comparisons with state-of-the-art deep convolutional neural networks as well as detectors built in the JPEG domain show the merit of this work.

Index Terms— Steganography, steganalysis, reverse JPEG compatibility attack, double compression, rounding errors

1. INTRODUCTION

Recently, a qualitatively new type of attack on JPEG steganography has been introduced, the Reverse JPEG Compatibility Attack (RJCA) [2], which forms the detection statistic from the rounding errors of a decompressed JPEG image. Unlike other detectors, the RJCA is universal (can detect any steganography) and can reliably detect even very short messages. It can thus provide a very high certainty about usage of steganography even from a single intercepted image. The disadvantage of this attack is its limited applicability to only JPEG quality 99 or 100. As the original paper on this topic shows, this limitation is quite fundamental and cannot be overcome due to the nature of the JPEG compression itself.

The main reason why the RJCA works is because, during compression, the discrete cosine transform (DCT) is applied to an integer-valued signal. This allows modeling the rounding errors after decompression as a zero-mean Gaussian with variance $\approx 1/12$ folded into the interval $[-0.5, 0.5)$. The embedding increases the variance of the Gaussian, which begins to fold into a uniform distribution. The attack can be realized by training a classifier on rounding errors [2] or using a simplified likelihood ratio test when the selection channel is known [3]. The attack does not work for lower qualities because the variance of the folded Gaussian increases rapidly with increasing quantization steps, making the distribution of the rounding errors essentially uniform even for cover images.

The main novel idea presented in this paper is the realization that the above-mentioned limitation relates to *single compressed images*, and does not necessarily apply to images that were compressed more than once. We show using mathematical analysis as well as experimentally that the RJCA can be extended to images doubly compressed with qualities $93 \leq Q_1 \leq Q_2$, broadening thus the applicability of this attack in practice. In particular, the attack is extremely accurate when $Q_1 = Q_2$, when the detectors that do not utilize rounding errors perform poorly. Images doubly compressed with the same quality factor naturally arise due to minor retouching, such as removing wrinkles or sensor dust, and adding a visible watermark when the editing tool is set to preserve the compression parameters. Moreover, double compressed JPEG covers can be introduced either by a conscious action of the sender or inadvertently due to the processing pipeline that precedes the actual embedding.

In the next section, we introduce the notation and preliminary concepts. Section 3 analyzes the distribution of rounding errors in the spatial domain after double compression for both cover and stego images. We analyt-

The work on this paper was partially 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 AFOSR or the U.S. Government.

ically show that the rounding errors after decompression can again be modeled as a folded Gaussian distribution if either the quality settings during both compression steps are the same or if $Q_2 = 99$ or 100 and $Q_1 \geq 93$. The theoretical insight is put to test in Section 4, where we report the detection accuracy of the RJCA for J-UNIWARD and benchmark it against SRNet [1] and JRM [4]. The paper is concluded in Section 5.

2. PRELIMINARIES AND NOTATION

Boldface symbols are reserved for matrices and vectors with elementwise multiplication and division denoted $\odot$ and $\oslash$. The uniform distribution on the interval $[a, b]$ will be denoted $\mathcal{U}[a, b]$ while $\mathcal{N}(\mu, \sigma^2)$ is used for the Gaussian distribution with mean μ and variance σ^2. Rounding x to an integer is denoted $[x]$. The set of all integers will be denoted $\mathbb{Z}$. For $X \sim \mathcal{N}(\mu, \sigma^2)$ with $\mu \in \mathbb{Z}$, the rounding error $X - [X] \sim \mathcal{N}_F$, the Gaussian distribution folded on $-1/2 \leq x < 1/2$, with pdf

$$\nu(x; \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \sum_{n \in \mathbb{Z}} \exp\left(-\frac{(x+n)^2}{2\sigma^2}\right). \quad (1)$$

For better readability, we strictly use i, j to index pixels and k, l DCT coefficients. Denoting by x_{ij}, $0 \leq i, j \leq 7$, an 8×8 block of pixels, they are transformed during JPEG compression to DCT coefficients $d_{kl} = \mathrm{DCT}_{kl}(\mathbf{x}) \triangleq \sum_{i,j=0}^{7} f_{kl}^{ij} x_{ij}$, $0 \leq k, l \leq 7$, and then quantized $c_{kl} = [d_{kl}/q_{kl}]$, $c_{kl} \in \{-1024, \dots, 1023\}$, where q_{kl} are quantization steps in a luminance quantization matrix, and $f_{kl}^{ij} = w_k w_l / 4 \cos \pi k(2i+1)/16 \cos \pi l(2j+1)/16$, $w_0 = 1/\sqrt{2}$, $w_k = 1$, $0 < k \leq 7$, are the discrete cosines.

During decompression, the above steps are reversed. For a block of quantized DCTs c_{kl}, the corresponding block of non-rounded pixels after decompression is $y_{ij} = \mathrm{DCT}_{ij}^{-1}(\mathbf{c} \odot \mathbf{q}) \triangleq \sum_{k,l=0}^{7} f_{kl}^{ij} q_{kl} c_{kl}$, $y_{ij} \in \mathbb{R}$. To obtain the final decompressed image, y_{ij} are rounded to integers and clipped to $[0, 255]$.

For color images, the *RGB* representation is typically changed to YC_bC_r (luminance, and two chrominance signals), the luminance Y is processed as above, while the chrominance signals are optionally subsampled, then transformed using DCT, and finally quantized with chrominance quantization matrices [7].

3. ROUNDING ERRORS AND DOUBLE COMPRESSION

In this section, we derive a model for the statistical distribution of the rounding errors in the spatial domain when decompressing a doubly compressed cover image and its stego version. The quantization matrices and quality factors used for the first and second compression will be denoted as $\mathbf{q}^{(1)}$, $\mathbf{q}^{(2)}$ and Q_1, Q_2, respectively.

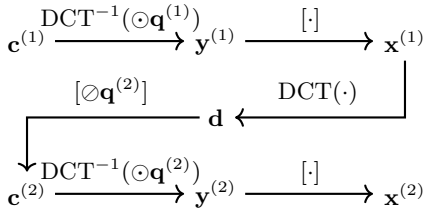

Fig. 1. Double compression pipeline.

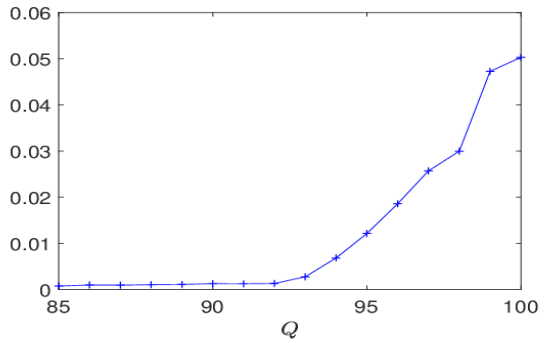

Fig. 2. Relative number of different quantized DCTs when recompressing an image with quality Q with the same quality. Results averaged over 1000 images from BOSSbase 1.01.

3.1. Cover images

Starting with a single-compressed JPEG file represented with quantized DCT coefficients $\mathbf{c}^{(1)}$, we consider the pipeline shown in Figure 1, which consists of decompressing $\mathbf{c}^{(1)}$ to $\mathbf{y}^{(1)}$, rounding to integers $\mathbf{x}^{(1)}$, compressing the second time with quantization matrix $\mathbf{q}^{(2)}$ to obtain DCT coefficients before quantization $\mathbf{d}$ and after quantization $\mathbf{c}^{(2)}$, decompressing to non-rounded pixels $\mathbf{y}^{(2)}$ and rounding to $\mathbf{x}^{(2)}$. Assuming the rounding errors in the spatial domain $u_{ij}^{(1)} = y_{ij}^{(1)} - x_{ij}^{(1)} \sim \mathcal{U}[-1/2, 1/2)$, we have $\mathbb{E}[u_{ij}^{(1)}] = 0$, $Var[u_{ij}^{(1)}] = 1/12$. Since

$$y_{ij}^{(1)} = \mathrm{DCT}^{-1}(\mathbf{c}^{(1)} \odot \mathbf{q}^{(1)}) = \sum_{k,l=0}^{7} f_{kl}^{ij} c_{kl}^{(1)} q_{kl}^{(1)} \quad (2)$$

$$x_{ij}^{(1)} = y_{ij}^{(1)} - u_{ij}^{(1)} = \sum_{k,l=0}^{7} f_{kl}^{ij} c_{kl}^{(1)} q_{kl}^{(1)} - u_{ij}^{(1)}, \quad (3)$$

we can write

$$\begin{aligned} d_{kl} &= \mathrm{DCT}_{kl}(\mathbf{x}^{(1)}) \\ &= \mathrm{DCT}_{kl}(\mathbf{y}^{(1)} - \mathbf{u}^{(1)}) \\ &= c_{kl}^{(1)} \cdot q_{kl}^{(1)} - \sum_{i,j=0}^{7} f_{kl}^{ij} u_{ij}^{(1)}. \end{aligned} \quad (4)$$

Assuming that $u_{ij}^{(1)}$ are mutually independent, from the CLT and orthonormality of the DCT :

$$d_{kl} \sim \mathcal{N}\left(c_{kl}^{(1)} q_{kl}^{(1)}, \frac{1}{12}\right). \quad (5)$$

Denoting the rounding error in the DCT domain during the second compression as $e_{kl} = d_{kl}/q_{kl}^{(2)} - c_{kl}^{(2)} = d_{kl}/q_{kl}^{(2)} - [d_{kl}/q_{kl}^{(2)}]$, from (5), e_{kl} follows a folded Gaussian distribution on $[-1/2, 1/2)$

$$e_{kl} \sim \mathcal{N}_F\left(c_{kl}^{(1)} \frac{q_{kl}^{(1)}}{q_{kl}^{(2)}}, \frac{1}{12(q_{kl}^{(2)})^2}\right) \quad (6)$$

with expectation

$$\mathbb{E}[e_{kl}] = c_{kl}^{(1)} \frac{q_{kl}^{(1)}}{q_{kl}^{(2)}} - \left[c_{kl}^{(1)} \frac{q_{kl}^{(1)}}{q_{kl}^{(2)}}\right]. \quad (7)$$

Continuing our analysis,

$$\begin{aligned} y_{ij}^{(2)} &= \mathrm{DCT}_{ij}^{-1}(\mathbf{c}^{(2)} \odot \mathbf{q}^{(2)}) \\ &= \mathrm{DCT}_{ij}^{-1}(\mathbf{d} - \mathbf{e} \odot \mathbf{q}^{(2)}) \\ &= \mathrm{DCT}_{ij}^{-1}\left(\mathrm{DCT}(\mathbf{y}^{(1)} - \mathbf{u}) - \mathbf{e} \odot \mathbf{q}^{(2)}\right) \\ &= y_{ij}^{(1)} - u_{ij}^{(1)} - \mathrm{DCT}_{ij}^{-1}(\mathbf{e} \odot \mathbf{q}^{(2)}) \\ &= x_{ij}^{(1)} - \eta_{ij}, \end{aligned} \quad (8)$$

where $\eta_{ij} = \sum_{k,l=0}^{7} f_{kl}^{ij} e_{kl} q_{kl}^{(2)}$. Assuming the independence of the rounding errors e_{kl}, the CLT implies

$$\eta_{ij} \sim \mathcal{N}\Big(\sum_{k,l=0}^{7} f_{kl}^{ij} q_{kl}^{(2)} \mathbb{E}[e_{kl}], \sum_{k,l=0}^{7} (f_{kl}^{ij})^2 (q_{kl}^{(2)})^2 Var[e_{kl}]\Big). \quad (9)$$

Thus, $y_{ij}^{(2)}$ follows a Gaussian distribution with mean

$$\mathbb{E}[y_{ij}^{(2)}] = x_{ij}^{(1)} - \sum_{k,l=0}^{7} f_{kl}^{ij} q_{kl}^{(2)} \mathbb{E}[e_{kl}]. \quad (10)$$

Note that for $q_{kl}^{(2)} > 1$, the variance of the folded Gaussian distribution (6) is approximately the same as the variance of the Gaussian that e_{kl} follows, $Var[e_{kl}] \approx 1/(12(q_{kl}^{(2)})^2)$, and thus $Var[y_{ij}^{(2)}] \approx 1/12$.

With this approximation, the rounding error after the second decompression $u_{ij}^{(2)} = y_{ij}^{(2)} - x_{ij}^{(2)}$ follows a Gaussian distribution, which is folded into $[-1/2, 1/2)$, with mean and variance

$$\mathbb{E}[u_{ij}^{(2)}] = -\sum_{k,l=0}^{7} f_{kl}^{ij} q_{kl}^{(2)} \mathbb{E}[e_{kl}] + \left[\sum_{k,l=0}^{7} f_{kl}^{ij} q_{kl}^{(2)} \mathbb{E}[e_{kl}]\right] \quad (11)$$

$$Var[u_{ij}^{(2)}] = 1/12. \quad (12)$$

For the RJCA to work, the distribution of the rounding error cannot be uniform as in this case, the embedding would not change it. In particular, if the expectations (11) are not zero and vary across pixels ij, the resulting mixture becomes practically uniform. On the other hand, when $\mathbb{E}[e_{kl}] = 0$ for most DCT modes kl and blocks, $\mathbb{E}[u_{ij}^{(2)}] \approx 0$ and the RJCA works again. Note that from (7) $\mathbb{E}[e_{kl}] = 0$ when $q_{kl}^{(2)}$ divides $c_{kl}^{(1)} q_{kl}^{(1)}$. Since we need this to be satisfied for the majority of the blocks and irrespectively of the content, we arrive at our first condition:

[C1] $q_{kl}^{(2)}$ divides $q_{kl}^{(1)}$ for most modes kl.

Note that this means that $Q_1 \leq Q_2$. Unless both qualities are equal, however, the double-compressed image will exhibit strong signs of double-compression with gaps and peaks in the DCT histogram, which will make steganography highly detectable using standard steganalysis features, such as the JRM [4]. Thus, from now on, we mainly focus on cases when $Q_1 = Q_2$ while noting that the RJCA remains extremely accurate when $Q_2 = 99$ or $Q_2 = 100$.

Moreover, notice that when $\mathbf{c}^{(1)} = \mathbf{c}^{(2)}$, the double-compressed image is the same as the single-compressed image, and, as already established in [2], the RJCA for single-compressed images works only for qualities 99 and 100. Thus, the second condition for the RJCA to work in doubly-compressed images with $Q_1 = Q_2$ is

[C2] $\mathbf{c}^{(1)} \neq \mathbf{c}^{(2)}$,

which is mainly fulfilled if there are ones in the quantization table or equivalently $Q_2 \geq 93$. This is confirmed in Figure 2 showing the average number of DCT coefficients that changed during recompression with the same quality factor across 1000 images selected from BOSSbase 1.01 at random. This result is not sensitive to the specific implementation of the JPEG compressor.

3.2. Stego images

Given a JPEG cover image represented by DCT coefficients $\mathbf{c}^{(1)}$, the steganographer embeds the secret message into the image after recompression $\mathbf{c}^{(2)}$. We model the steganography by adding steganographic noise $\xi_{kl} \in \{-1, 0, 1\}$, $\Pr\{\xi_{kl} = 1\} = \beta^+$, $\Pr\{\xi_{kl} = -1\} = \beta^-$ to the cover: $s_{kl} = c_{kl}^{(2)} + \xi_{kl}$. Note that $\mathbb{E}(\xi_{kl}) = \beta_{kl}^+ - \beta_{kl}^-$ and $Var[\xi_{kl}] = \beta_{kl}^+ + \beta_{kl}^-$.

Decompressing the stego image block gives

$$\begin{aligned} z_{ij} &= \mathrm{DCT}_{ij}^{-1}(\mathbf{s} \odot \mathbf{q}^{(2)}) \\ &= \mathrm{DCT}_{ij}^{-1}(\mathbf{c}^{(2)} \odot \mathbf{q}^{(2)} + \xi \odot \mathbf{q}^{(2)}) \\ &= x_{ij}^{(1)} - \eta_{ij} + \zeta_{ij}, \end{aligned} \quad (13)$$

Q_1	detector	Q_2							
		93	94	95	96	97	98	99	100
93	e-SRNet	0.0438	0.3678	0.4104	0.3545	0.2845	0.0317	0.0002	0.0002
	eOH-SRNet	0.0485	0.0059	0.0019	0.0024	0.0035	0.0051	0.0001	0.0001
	JRM	0.4360	0.0029	0.0028	0.0016	0.0010	0.0031	0.0064	0.0053
94	e-SRNet		0.0028	0.3356	0.4205	0.1725	0.0994	0.0001	0.0000
	eOH-SRNet		0.0027	0.0076	0.0030	0.0033	0.0060	0.0002	0.0001
	JRM		0.4304	0.0023	0.0022	0.0019	0.0022	0.0050	0.0068
95	e-SRNet			0.0009	0.3449	0.2870	0.0463	0	0.0001
	eOH-SRNet			0.0008	0.0008	0.0038	0.0038	0.0002	0.0001
	JRM			0.4232	0.0067	0.0024	0.0039	0.0052	0.0067
96	e-SRNet				0.0006	0.3251	0.0412	0.0001	0.0001
	eOH-SRNet				0.0004	0.0118	0.0062	0.0001	0.0002
	JRM				0.4196	0.0079	0.0058	0.0068	0.0086
97	e-SRNet					0.0005	0.2055	0.0001	0.0003
	eOH-SRNet					0.0003	0.0482	0.0002	0.0001
	JRM					0.4159	0.0207	0.0070	0.0061
98	e-SRNet						0.0003	0.0001	0.0001
	eOH-SRNet						0.0001	0.0002	0.0001
	JRM						0.4194	0.0031	0.0041
99	e-SRNet							0	0.0001
	eOH-SRNet							0.0001	0
	JRM							0.4127	0.0026
100	e-SRNet							0.0002	0.0001
	eOH-SRNet							0.0001	0.0001
	JRM							0.4126	0.3965

Table 1. Detection error P_{E} with different detectors, J-UNIWARD at 0.4 bpnzac.

where $\zeta_{ij} = \sum_{kl} f^{ij}_{kl} \xi_{kl} q^{(2)}_{kl}$

$$\zeta_{ij} \sim \mathcal{N}\Bigg(\sum_{k,l=0}^{7} f^{ij}_{kl} q^{(2)}_{kl} (\beta^{+}_{kl} - \beta^{-}_{kl}), \sum_{k,l=0}^{7} (f^{ij}_{kl})^2 (q^{(2)}_{kl})^2 (\beta^{+}_{kl} + \beta^{-}_{kl}) \Bigg). \quad (14)$$

For steganography without side information $\beta^{+}_{kl} = \beta^{-}_{kl}$, thus

$$z_{ij} \sim \mathcal{N}\Bigg(x^{(1)}_{ij} - \sum_{k,l=0}^{7} f^{ij}_{kl} q^{(2)}_{kl} \mathbb{E}[e_{kl}], \sum_{k,l=0}^{7} (f^{ij}_{kl})^2 (q^{(2)}_{kl})^2 (\beta^{+}_{kl} + \beta^{-}_{kl} + Var[e_{kl}]) \Bigg). \quad (15)$$

Notice that the rounding error of z_{ij} is a folded Gaussian whose variance is increased due to embedding (c.f. Eq. (9) with Eq. (15)) and whose mean is now *non-zero*, dependent on the rounding errors in DCT domain. Both contribute to the fact that in stego images, these Gaussians will start folding into a uniform distribution with increased payload (change rates).

4. RESULTS

All experiments in this paper are executed on the union of the popular datasets BOSSbase 1.01 and BOWS2, each containing 10,000 grayscale images downsampled to 256×256 using 'imresize' with default parameters in Matlab. The detectors were trained on all BOWS2 images and a randomly selected 4,000 BOSSbase images, with 1,000 BOSSbase images used for validation and 5,000 for testing.

Table 1 shows the detection error under equal priors on the testing set for J-UNIWARD at 0.4 bpnzac. The cover JPEG images were doubly compressed with the first quality factor being represented by rows and the second quality factor by columns. We only show the cases when $93 \leq Q_1 \leq Q_2$ and also when $Q_1, Q_2 \in \{99, 100\}$, since these cases satisfy condition [C1]. Three detectors are tested: SRNet [1] trained on the rounding errors after decompressing the JPEG image (e-SRNet), JRM with the ensemble classifier [5], and OneHot network [9] combined with e-SRNet (eOH-SRNet), which is implemented as OneHot-SRNet in the original paper with clipping threshold $T = 5$. The SRNet, however, takes the rounding errors on the input instead of the spatial representation of the image. We want to point out that both network based detectors converge to their optimum extremely quickly, within 20k iterations. Even though e-SRNet fails for some combinations of the compression qualities, such as (96, 97), double compression with such combinations of quality factors leads to peaks and valleys in cover DCT histograms, which allows very accurate detection with JRM and other prior art [6, 8, 10, 9]. Note that these detectors perform rather poorly whenever $Q_1 = Q_2$. The eOH-SRNet provides overall reliable detection.

The condition [C2] dictates that the RJCA will work whenever the (equal) quality factors are at least 93 and that can be confirmed in Table 1. Results for lower qualities are not included because RJCA stops working there, in agreement with the analysis from Section 3.

5. CONCLUSIONS

The reverse JPEG compatibility attack is an extremely accurate, universal, and quite simple steganalysis technique that was originally shown to be limited to high quality factors (99 or 100). In this paper, we extend this attack to cover images that are doubly compressed with quality factors $93 \leq Q_1 \leq Q_2$. By analyzing the distribution of the rounding errors in the spatial domain, we arrived at two conditions that need to be satisfied for the attack to work. The conclusions reached from the theoretical considerations match our experimental results. In combination with the OneHot-SRNet, the detector provides the most reliable detection across all above combinations of quality factors. In particular, the compatibility attack works extremely reliably also when $Q_1 = Q_2$, which is the case when all other tested detectors (SRNet and JRM) perform rather poorly.

6. 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] J. Butora and J. Fridrich. Reverse JPEG compatibility attack. *IEEE Transactions on Information Forensics and Security*, 15:1444–1454, 2020.

[3] R. Cogranne. Selection-channel-aware reverse JPEG compatibility for highly reliable steganalysis of JPEG images. In *Proceedings IEEE, International Conference on Acoustics, Speech, and Signal Processing*, pages 2772–2776, Barcelona, Spain, May 4–8, 2020.

[4] J. Kodovský and J. Fridrich. Steganalysis of JPEG images using rich models. In A. Alattar, N. D. Memon, and E. J. Delp, editors, *Proceedings SPIE, Electronic Imaging, Media Watermarking, Security, and Forensics 2012*, volume 8303, pages 0A 1–13, San Francisco, CA, January 23–26, 2012.

[5] 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.

[6] J. L. Davidson P. Parajape. Double-compressed JPEG detection in a steganalysis system. In *Annual ADFSL Conference on Digital Forensics, Security, and Law*, May 30, 2012.

[7] W. Pennebaker and J. Mitchell. *JPEG: Still Image Data Compression Standard*. Van Nostrand Reinhold, New York, 1993.

[8] Y. Yang, X. Kong, and C. Feng. Double-compressed JPEG images steganalysis with transferring feature. *Multimedia Tools and Applications*, 77, February 2018.

[9] Y. Yousfi and J. Fridrich. An intriguing struggle of cnns in jpeg steganalysis and the onehot solution. *IEEE Signal Processing Letters*, 27:830–834, 2020.

[10] Y. Zhou, W. W. Y. Ng, and Z. He. Effects of double jpeg compression on steganalysis. In *International Conference on Wavelet Analysis and Pattern Recognition*, pages 106–112, 2012.