Toward Reliable JPEG Steganalysis (at QF100)

Etienne Levecque, John Klein, Patrick Bas, Jan Butora

Toward Reliable JPEG Steganalysis (at QF100)

Etienne Levecque
Univ. Lille, CNRS, Centrale Lille, UMR 9189 CRIStAL Lille, France
etienne.levecque@cnrs.fr

John Klein
Univ. Lille, CNRS, Centrale Lille, UMR 9189 CRIStAL Lille, France
john.klein@univ-lille.fr

Patrick Bas
Univ. Lille, CNRS, Centrale Lille, UMR 9189 CRIStAL Lille, France
patrick.bas@cnrs.fr

Jan Butora
Univ. Lille, CNRS, Centrale Lille, UMR 9189 CRIStAL Lille, France
jan.butora@cnrs.fr

***Abstract*—Contrary to classical steganalysis methods focused on the detection performance, this paper proposes a reliable steganalysis method targeting two goals: prescribing a given, potentially very small, False-Positive (FP) rate while maximizing the detection accuracy. This is the first step toward operational steganalysis where the forensics agent needs an accurate FP rate in order to make a decision. Because JPEG steganalysis at Quality Factor (QF) 100 is associated with an accurate and robust distribution of Cover images by computing the rounding error after JPEG decompression, we use this domain to derive statistical tests associated with theoretical FP rates. A Kolmogorov-Smirnov test applied on different pixel positions of the 8×8 block, associated with an aggregation strategy, and a block filtering pre-processing are combined to propose an accurate yet reliable detector for FP going below 10^{-4}. Experiments are both assessed on synthetic images and the "wild" ALASKA database using J-UNIWARD.**

***Index Terms*—Steganalysis, reliability, false-positive rate, Kolmogorov-Smirnov test**

I. Introduction

Because the goal of Eve, the warden performing steganalysis, is to *decide* from the output of a given detector if the image under scrutiny should be considered either as Cover (Negative outcome) or Stego (Positive), the operational use of such a detector depends on its performances in terms of False Positive (FP) and False Negative (FN) rates. Most steganalysis detectors from the literature are validated in terms of probability of error which generally penalizes equally FPs from FNs and aggregate these two criteria into a single scalar. From a broader decision-theoretic standpoint, and also in connection with the practical needs of end-users, merging these two criteria may not be desirable. In this article, we investigate ways to obtain a detector with a strong control on FP rate and, given that control, an FN rate as small as possible which is reminiscent of the Neyman-Pearson classification framework. While fixing the FN rate and minimizing FPs is another possibility [1], as will be motivated in the following paragraphs, we stress that the desired control is meant to be valid for a vast class of JPEG Stego sources[1] which calls for a precise statistical model of the Cover class and rules out most (discriminative) machine learning approaches.

[1]The only restriction is on the quality factor (QF) which is restricted to 100 for reasons to be exposed in the sequel. However, the detector should work for several acquisition devices, processing pipelines, or payloads.

A. Toward Reliable Steganalysis

We motivate in this paper the need to design both *accurate* and *reliable* detectors in steganalysis. If on one hand, accuracy is related to the power of the test performed by the classifier, i.e. the True Positive (TP) rate; reliability is related to the possibility of controlling the error rates, and particularly the FP rate.

Practically, FP rates may be estimated using Monte-Carlo simulations (i.e. by testing on a large database of Cover images), but this is only possible if the prescribed FP rate is relatively high, for example prescribing a FP rate of 10^{-9} would require more than 10 billion images to have a reliable estimation, which is expensive from a computational point of view.

Moreover, to control the FP rate, the diversity of the Cover images needs also to be controlled. It has been shown for example in [2], [3] that a detector trained from a source made of a base of images developed with a given software will be efficient on a test base also developed with this same software. However, it can become completely inefficient if the tested images are developed with another software, while the visual content between the two bases is extremely close. This problem, known as the Cover Source Mismatch (CSM) can complicate considerably the task of the warden.

Consequently, Eve needs to have theoretical guarantees on her FP rates. In order to be practical, the relation between the detection threshold used to discriminate Cover from Stego images should be related to the FP rate by an explicit function, and Monte-Carlo simulations can be used to assess, whenever it is practically possible, the validity of the targeted FP rate.

It is also important to mention that the framework of reliable steganalysis is different from standard steganalysis since it now can be seen as a two-faced problem:

1) On one hand, Eve needs to accurately control the FP error rate for potentially different sources of images and arbitrary small rates (targeting 10^{-6}),

2) On the other hand, the power of the test needs also to be maximized. This second constraint of course is dependent of the Stego scheme and the embedding rate.

A naive detector, which is an extreme point in the scope of this new trade-off, consists in deciding that each steganalyzed image is Cover, hence fixing the FP rate to zero, but in this case, the power of the test would be null as well. On the other side, Eve can use a steganalyzer associated with very high TP rates, but if the FP rate is too important, the detector will also be useless for her.

B. Prior works

As we shall see in section II, error rates can be controlled within the framework of statistical hypothesis testing by computing the p-value associated with a given test, which is equivalent to deriving the distribution of the statistic used to perform the test. We briefly highlight previous works following rather similar approaches.

The pioneering work of Cogranne and Retraint [4] (see e.g. Sec. 4 and Fig. 4) proposed to model the distribution of the Generalized Likelihood Ratio Test (GLRT) as a Normal distribution, for LSB matching in the spatial domain and on RAW images. Note however that such a test relies on the knowledge of the payload size, and also that the estimation of noise associated with the Cover distribution is reliable only if the noise is independently distributed, which is not the case for various sources of images.

Previous works leverage the computation of p-values in steganalysis, but this was either to assess the validity of the Cover distribution or for detection purposes. In [5], Ker used the Kolmogorov-Smirnov statistic (see section II for more details) and its associated p-value on RS steganalysis to check if the distribution of the Cover images after JPEG compression and downscaling are similar to the distributions of Cover images without compression.

In [6], Westfeld and Pfitmaan computed the p-value to detect segments of the image that could be considered as carrying the payload after various embeddings operating in the LSB domain, one of the objectives there was to estimate the payload size.

C. Outline of the paper

This paper proposes to design a reliable steganalysis scheme in the JPEG domain. To control the distribution of Cover contents, and to be immune to the CSM, the detection is only performed for a quality factor equal to 100. This specific setting enables obtaining a reliable distribution of the Cover images, by computing the rounding error after the inverse DCT transform [7], which is invariant to the CSM (see section III).

Section II presents different statistical tests and strategies to aggregate several p-values generated from the same test image on potential dependent samples.

Section IV investigates on synthetic images the impact of different hypotheses such as the CLT, or the i.i.d. property of the rounding errors. Section V presents practical results and technical considerations necessary to achieve reliable steganalysis. Both the control of the error rate and the detection performance are benchmarked.

II. Statistical tests and FP rates control

A statistical test is a procedure consisting in checking if a (null) hypothesis $\mathcal{H}_0$ should be rejected based on collected data. This is usually done by computing a statistic from the data and, based on the precisely known distribution of this statistic, by checking if the data is a rare event (having a probability no greater than a prescribed small level α). If so, $\mathcal{H}_0$ is rejected otherwise $\mathcal{H}_0$ cannot be rejected meaning that the test is inconclusive. In our context, $\mathcal{H}_0$ is the fact that an analyzed image is Cover and the processed data is $\mathbf{e}$. The corresponding detector assigns the Stego class to the image when $\mathcal{H}_0$ is rejected and to the Cover class otherwise. Consequently, the level α is the (theoretical) FP rate achieved by the detector.

In [7], the authors envisaged using a likelihood ratio test. This test also relies on the wrapped Gaussian model and because of the limitations evoked in the previous section, it is not possible to derive the exact distribution of the likelihood ratio statistic. It also requires a model for the alternative hypothesis $\mathcal{H}_1$ (Stego class) which is impossible without the knowledge of the embedding rate.

We present the Kolmogorov-Smirnov test in the next subsection, followed by the Bonferroni procedure to control the Family-Wise Error Rate (FWER) of multiple hypothesis testing.

A. Kolmogorov-Smirnov (KS) Test

To test the equality between the sample distribution from a given image and the Cover one, the Kolmogorov-Smirnov (KS) test [8] appears as the most straightforward choice of non-parametric test regarding assumptions. The KS test can compare a sample with a one-dimensional reference probability distribution if the latter is continuous and has a known cumulative distribution function. Moreover, the KS test can also compare two samples. It can be instrumental if the cumulative distribution of the reference distribution is not known but can be sampled.

Let $\mathbf{X}$ and $\mathbf{Y}$ be one dimensional samples of size n and n' respectively, let $F_{\mathbf{X}}$ and $F_{\mathbf{Y}}$ be the empirical cumulative distributions defined for every $x \in \mathcal{R}$ such as $F_{\mathbf{X}}(x) = \frac{1}{n}\sum_{i=0}^{n} \mathbb{1}_{[-\infty;x]}(X_i)$ with $\mathbb{1}_{[-\infty;x]}(X) = 1$ if $X \leq x$, 0 otherwise. Finally, let F be a cumulative density function.

The KS statistic is the supremum vertical distance between the two cumulative distributions:

$$\begin{array}{lll} D_n & = \sup_x |F_{\mathbf{X}}(x) - F(x)| & \text{One sample KS test} \\ D_{n,n'} & = \sup_x |F_{\mathbf{X}}(x) - F_{\mathbf{Y}}(x)| & \text{Two-sample KS test} \end{array}$$

Under the null hypothesis, the distribution of the statistic is known and tends toward a Kolmogorov distribution when n and n' tend toward infinity. In practice, assuming we observed a KS distance d between the two cumulative distributions, accurate approximations of the distribution can be used to

obtain the p-value: $p = P(D_{n,n'} > d)$ or $P(D_n > d)$ depending on the chosen test type (two or one-sample).

As we shall see in section IV, in our case the null hypothesis might be different for each of the 64-pixel positions in a block. This leads to 64 different hypothesis testing and the same number of p-values. Since our global test is to know if the image is Cover or not, we are facing a multiple hypothesis testing problem.

B. Bonferroni procedure and Family-Wise Error Rate (FWER)

Multiple hypothesis testing problems occur when inferring a parameter from a set of tests. The more tests are performed, the more likely we are to observe an outlier statistic and thus an error. Assuming we have a family of m different hypothesis, we denote the number of type I error (false positive outcome) as $V \in [0, m]$. The Family-Wise Error Rate (FWER) is defined by the probability of observing at least one type I error among the m hypothesis: FWER $= P(V > 0)$

Let $\alpha > 0$ be our type I error parameter and let p be the p-value of one of our test. We reject the null hypothesis if $p \leq \alpha$. By definition of a p-value, $P(p \leq \alpha) \leq \alpha$. So by correcting the threshold α to $\frac{\alpha}{m}$ (or multiplying the p-value p by m), without any assumption on the dependence of the test,

$$\text{FWER} = P(V > 0) \leq m \times P\left(p \leq \frac{\alpha}{m}\right) \leq m\frac{\alpha}{m} = \alpha.$$

This correction is called the Bonferroni procedure [8] and will be used to aggregate our 64 tests into a single one.

III. JPEG steganalysis at QF100

We formalize here the steganalysis problem, which consists in analyzing JPEG images compressed with a Quality Factor of 100. Note that QF100 images are quite popular on the internet, since around 15% of uploaded images on Flickr are compressed at QF100 [9].

JPEG images are defined on 8×8 blocks and every following notation and property is block independent so, for simplicity, let $\mathbf{c}$ denote a raw Cover image composed of a single block 8×8. The two-dimensional Discrete Cosine Transform is a function denoted DCT and its inverse is DCT^{-1}. We introduce the following notations:

$\tilde{\mathbf{c}}$	$=$	$\text{DCT}\{\mathbf{c}\}$	DCT domain image
$[\tilde{\mathbf{c}}]$			Compressed JPEG image
$\hat{\mathbf{c}}$	$=$	$\text{DCT}^{-1}\{[\tilde{\mathbf{c}}]\}$	Decompressed image

where $[.]$ is the rounding function. Since we are working at QF100, the quantification step is equal to one and does not appear in those expressions. Each position in the 8×8 block can lead to different statistical models, so everywhere in this paper, $0 \leq i, j \leq 7$ will be used to index pixels, and $0 \leq k, l \leq 7$ will index DCT coefficients.

The stenographer has no other choice than to insert its secret message by modifying the DCT coefficients $[\tilde{c}]_{kl}$, typically by making ± 1 moves on these coefficients. Butora and Fridrich [7] showed that these modifications leave a highly detectable pattern in the spatial rounding error after decompression. We define this rounding errors as follows:

$\mathbf{e}$	$=$	$\hat{\mathbf{c}} - [\hat{\mathbf{c}}]$	rounding error in pixel domain
$\mathbf{u}$	$=$	$\tilde{\mathbf{c}} - [\tilde{\mathbf{c}}]$	rounding error in DCT domain

Moreover, the authors also suggest that these pixel errors $\mathbf{e}$ can be adequately modeled by a wrapped Gaussian distribution, denoted by $\nu\left(0, \frac{1}{12}\right)$, where 0 is the mean and $\frac{1}{12}$ is the variance. This statistical model could lead to the type of control sought in this paper. Unfortunately, this is not the case for small FP rates, as the following paragraphs will illustrate.

1) Associated hypotheses and distributions: The model proposed in [7] is built upon the hypothesis that rounding errors in the DCT domain (during JPEG compression) are independent and identically distributed. The aforementioned hypothesis consists in $u_{kl} \perp\!\!\!\perp u_{k'l'}$ whenever $(k, l) \neq (k', l')$ and $u_{kl} \sim \mathcal{U}_{\left[-\frac{1}{2};\frac{1}{2}\right]}$ (uniform distribution on the $\left[-\frac{1}{2}; \frac{1}{2}\right]$ interval)[2].

A closer look at these random variables (see IV-B) reveals that none of these assumptions hold even if, in first approximation, the wrapped Gaussian model efficiently characterizes $\mathbf{e}$.

IV. Playing in a sandbox

A. Starting in a controlled environment ...

We introduce our approach through a simple framework where most difficulties are controlled. In this sandbox, an image $\mathbf{x}$ is composed of n independent and identically distributed pixels sampled from a uniformly discrete distribution between 0 and 255. This way we can ensure that the host content follows the i.i.d. assumption. This image is split into

[2]the symbol $\perp\!\!\!\perp$ defines two random variables which are independently distributed.

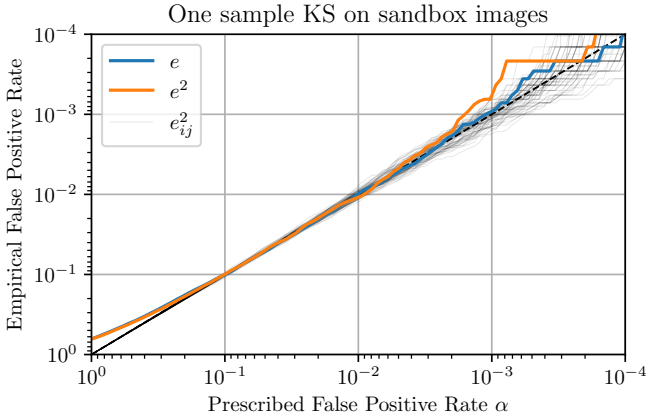


Fig. 1: Empirical false positive rate depending on theoretical false positive rate α. Computed on 1M random images $\mathbf{x} \sim \mathcal{U}_{\{0,255\}}^{N}$. Black lines are the 64 $KS\left(e_{ij}^2, \nu\left(0, \frac{1}{12}\right)^2\right)$ without Bonferroni procedure and aggregation.

8×8 blocks. Thus, we end up with 64 rounding error samples $(\mathbf{e}_{ij})$ of the same sizes.

Our family of null hypothesis is

$$\mathcal{H}_0 = \left\{ \mathcal{H}_0^{i,j} : \mathbf{e}_{ij} \sim \nu\left(0, \frac{1}{12}\right), \quad \forall\, 0 \leq i,j \leq 7 \right\}.$$

$\mathcal{H}_0^2$ respectively denotes the family of null hypothesis concerning the squared rounding error whose, by abuse of notation, is denoted by ν^2.

The wrapped Gaussian cumulative distribution function and the one for the squared error are derived from [7] as follows. For all $x \in \mathbb{R}$,

$$F_{\nu(\mu,s)}(x) = \frac{1}{\sqrt{2\pi}s} \sum_{n\in\mathbb{Z}} \left[\Phi\left(-\tfrac{1}{2} - \mu + [\mu] - n\right) - \Phi\left(-x - \mu + [\mu] - n\right)\right],$$

and,

$$F_{\nu^2(\mu,s)}(x) = 2F_{\nu(\mu,s)}(\sqrt{|x|}) - 1,$$

where Φ is the cumulative distribution function of the normal distribution.

Using the one sample KS test, we obtain a set of 64 p-values $\mathbf{p} = (p_{ij})$ and we can correct them using the Bonferroni procedure. The new p-values are simply $\mathbf{p}^* = (p^*_{ij})$, where $p^*_{ij} = 64 \times p_{ij}$ for all $0 \leq i,j \leq 7$.

Finally, we define the following classifier:

$$\delta_\alpha(\mathbf{x}) = \begin{cases} \text{Stego if } \min_{i,j} p^*_{ij} \leq \alpha, \\ \text{Cover otherwise.} \end{cases}$$

The probability of type I error of this classifier is the probability to classify a Cover image as stego:

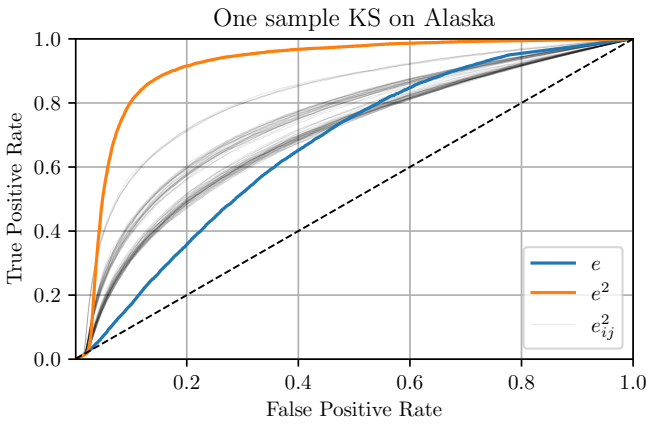
$$P_{\mathbf{x}\in\text{cover}}\left(\delta_\alpha(\mathbf{x}) = \text{stego}\right) = \text{FWER}(\mathcal{H}_0) \leq \alpha.$$

We use the spatial rounding error $\mathbf{e}$ but also its (elementwise) squared value $\mathbf{e}^2$. The motivation behind the squared error is twofold.

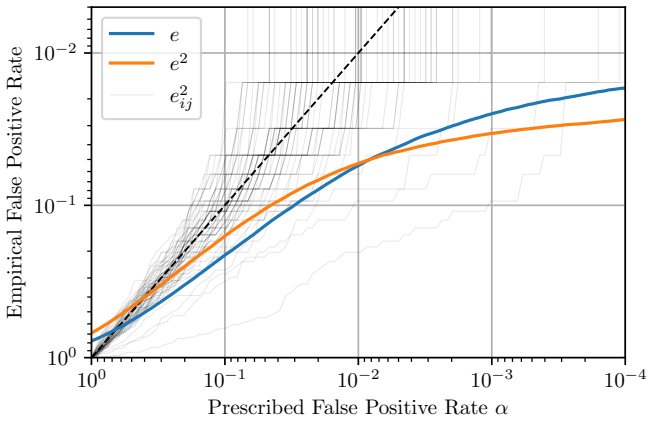
First, depending on the implementation of the rounding error, $x \mapsto x - [x]$ or $x \mapsto [x] - x$ do not have the same support which can lead to a shift in the mean. The rounding function itself can also have different implementations, but by "folding" the distribution on itself, there is no such difference anymore. This can be done using the absolute value function or the square function. Note also that since the distribution of the rounding error is symmetric, "folding" the distribution generates a more accurate empirical CDF for the KS test, it can be seen as a way to artificially double the number of observed samples.

Second, we know that steganalysers using the variance of the rounding error are quite powerful. Since the variance is computed from the squared error, we hope to benefit from this detection power (see sections IV and V).

Figure 1 represents the empirical false positive rate as a function of α. It was obtained using "sandbox images" as explained earlier. Both the error and the squared error give promising results on the guarantees. The slight offset around $\alpha = 1$ is due to the Bonferroni procedure but remains consistent with the upper bound.



(a) ROC Curve



(b) P_{FA} Comparison

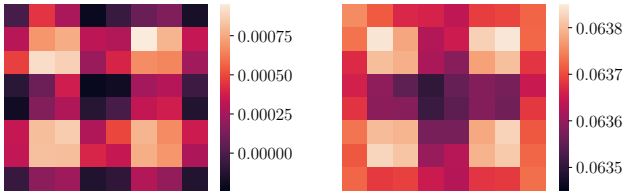
Fig. 2: Steganalysis results on Alaska dataset: 10k Stego images embedded with J-UNIWARD at payload 0.1 bpnzac and 70k Cover images. Legend is shared with figure 1 : in blue $KS\left(e, \nu\left(0, \frac{1}{12}\right)\right)$, in orange $KS\left(e^2, \nu^2\left(0, \frac{1}{12}\right)\right)$. Black lines are the results without Bonferroni procedure and aggregation for the squared error.

B. ... and going into the wild

The Alaska database [9] has been generated to promote an important diversity by using a large number of development pipelines and camera sensors. We now benchmark our steganalysis scheme on JPEG images at QF100 directly downloaded from https://alaska.utt.fr. Here results are very different. Indeed, Figure 2 shows two things. First, the ROC curves show a better power for the "folded" rounding error. Second, the upper bound on the false positive rate is violated for both features. We can think of the following hypotheses to explain this result:

(a) In uniform areas of the image, the rounding error has intra and inter blocks correlations which have an impact on the overall distribution.

(b) The i.i.d. wrapped Gaussian model is not accurate enough. For example, dependencies inside the same



(a) Mean (b) Variance

Fig. 3: Empirical mean and variances of rounding errors for the 64 pixels positions, estimation using 10^4 images.

sample $\mathbf{e}_{ij}$ are not modeled. One can also empirically observe a different variance $\mathrm{Var}(\mathbf{e}_{ij})$ between corners, borders, and inner positions in the block where the wrapped Gaussian has a single variance for each position (see Figure 3).

V. Practical approaches

In this section, we modify the classifier [3] w.r.t. two aspects. A filter on the variance of the pixel values in one block is introduced to favor noisy blocks. The test is also changed from one-sample KS test to a two-sample KS test in order to obtain a more reliable null hypothesis.

A. Filtering the rounding errors

In natural images, the content and the processing pipeline can create correlations between pixels values but section IV shows that independent pixels lead to a correct false positive rate upper bound. So if we filter out the most correlated pixels and keep only the noisiest ones, we tend toward the independent noise setting of the sandbox framework. This is done using the variance of the pixel values in each block. If the variance of a block is less or equal than a threshold s, the rounding errors of these blocks are discarded, otherwise, they are kept to perform the test.

If most images have a sufficient amount of textures to be kept by this filter, some others will lose most of their blocks, if not all. Performing the KS test on a few samples is possible but can lead to outliers very easily. That is why we only classify images with at least 100 blocks remaining after filtering. The value of s is studied in Figure 4 and shows that the higher the variance threshold, the closer to the guarantee. However, the higher this threshold, the more images are not classified because they do not have enough blocks remaining. For the rest of the paper, we choose to fix $s = 20$ such that only noisy blocks are used and only 3% images are not classified in Alaska. This filter should deal with the first hypothesis (a).

B. Increasing reliability with two-sample tests

The possibility that the wrapped Gaussian is not accurate enough can be bypassed by using the two-sample KS test. The idea would be to use some Cover images at the disposition of the steganalyser to build 64 reference samples of rounding errors. A larger reference sample should have better detection power but is associated with a larger complexity to compute the 64 tests. A small reference can lead to a poor representation of the true distribution and increase the chance of a false positive. For the rest of this paper, around 100 images were used to build a sample vector of 75×10^3 rounding errors. Practically, we can notice that false-positive guarantees are already obtained with 15 images to build the training set, but we prefer to take some margin, which also boosts the detection power by learning a more accurate and generalizable model. This change can be seen as a change in the null hypothesis and should address the second point (b).

[3]The code related to this classifier can be found on Gitlab.

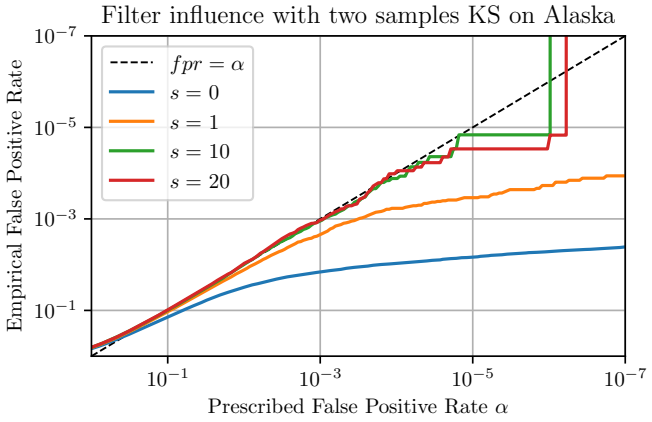


Fig. 4: Impact of the block filtering threshold s on the reliability of the detector for 70k Cover images. Note that 3% of the images are not classified for $s = 20$.

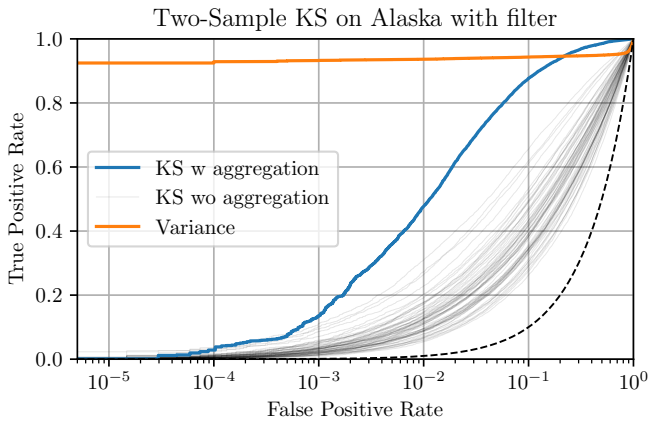
C. Steganalysis performances

Figure 5a shows the results of this detector on the Alaska dataset embedded with J-UNIWARD using a payload of 0.1 bpnzac (bit per non-zero AC coefficient). Both detection and false positive rate guarantee are satisfactory. Note however that the proposed classifier is, for low FP rates, very far to be as powerful as the score based on the variance of the rounding error proposed in [7], [10]. This illustrates the trade-off that we are currently facing between classical and reliable steganalysis.

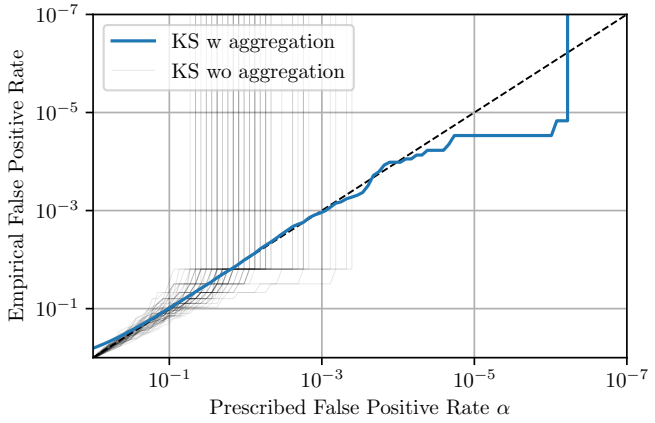
Figure 6 shows that the detection power increases with higher payloads. Since Cover images are not changing when the payload changes, we always have the same amount of false positives. The dataset was voluntarily unbalanced toward Cover images (70k Cover vs 10k stego) to be able to observe low false-positive values.

VI. Conclusions and perspectives

In this paper, the rounding error is used to build a reliable steganalysis detector for JPEG images at QF100, where the practical/actual false positive rate is upper bounded by a prescribed FP rate α that can be chosen by the forensics agent. This detector is composed of 64 two-sample KS tests for



(a) ROC Curve



(b) P_{FA} Comparison

Fig. 5: Steganalysis results on Alaska: 10k Stego images embedded with J-UNIWARD at payload 0.1 bpnzac and 70k Cover images. Note that 3% of 80k images were not classified because of the filter. Black lines are the results without the Bonferroni procedure and aggregation.

every position in the 8×8 blocks where the reference sample is obtained from a training phase with known Cover images. The p-values of every test are corrected by the Bonferroni procedure and used for the decision to increase the power of the test by aggregating different observations.

We have also noticed that the i.i.d. wrapped Gaussian approximation of the rounding error proposed in [7] struggles to represent the complexity of natural images but the two-sample test combined with a filter on the noisiest blocks of the image increases considerably the reliability of the detector.

Future perspectives encompass the study of the robustness of this detector for other JPEG compressors than the one used to generate the Alaska database. Bigger datasets or smart Cover sampling can also be used to confidently observe tiny false positive rates such as 10^{-6} or below. Future analyses will also refine the distribution of rounding errors of each pixel position.

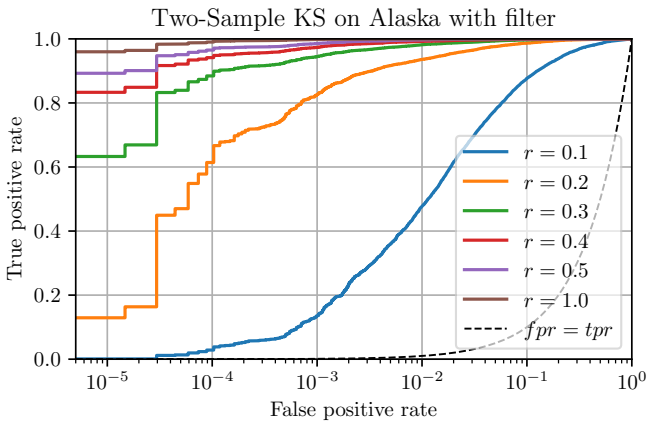


Fig. 6: Steganalysis results on Alaska: 10k Stego images embedded with J-UNIWARD at different embedding rates (bpnzac) and 70k Cover images. Note that 3% of 80k images were not classified because of the filter.

Acknowledgment

The work presented in this paper received funding from the European Union's Horizon 2020 research and innovation program under grant agreement No 101021687 (project "UNCOVER").

References

[1] T. Pevny and A. D. Ker, "Exploring non-additive distortion in steganography," in *Proceedings of the 6th ACM Workshop on Information Hiding and Multimedia Security*, 2018, pp. 109–114.

[2] J. Kodovskỳ, V. Sedighi, and J. Fridrich, "Study of cover source mismatch in steganalysis and ways to mitigate its impact," in *Media watermarking, security, and forensics 2014*, vol. 9028. SPIE, 2014, pp. 204–215.

[3] Q. Giboulot, R. Cogranne, D. Borghys, and P. Bas, "Effects and Solutions of Cover-Source Mismatch in Image Steganalysis," *Signal Processing: Image Communication*, Aug. 2020. [Online]. Available: https://hal-utt.archives-ouvertes.fr/hal-02631559

[4] R. Cogranne and F. Retraint, "An asymptotically uniformly most powerful test for lsb matching detection," *Information Forensics and Security, IEEE Transactions on*, vol. 8, no. 3, pp. 464–476, 2013.

[5] A. D. Ker, "Quantitative evaluation of pairs and rs steganalysis," in *Security, Steganography, and Watermarking of Multimedia Contents VI*, vol. 5306. SPIE, 2004, pp. 83–97.

[6] A. Westfeld and A. Pfitzmann, "Attacks on steganographic systems," in *International workshop on information hiding*. Springer, 1999, pp. 61–76.

[7] J. Butora and J. Fridrich, "Reverse jpeg compatibility attack," *IEEE Transactions on Information Forensics and Security*, vol. 15, pp. 1444–1454, 2019.

[8] E. L. Lehmann, J. P. Romano, and G. Casella, *Testing statistical hypotheses*. Springer, 2005, vol. 3.

[9] R. Cogranne, Q. Giboulot, and P. Bas, "The ALASKA Steganalysis Challenge: A First Step Towards Steganalysis "Into The Wild"," in *ACM IH&MMSec*, Paris, France, Jul. 2019. [Online]. Available: https://hal.archives-ouvertes.fr/hal-02147763

[10] R. Cogranne, "Selection-channel-aware reverse jpeg compatibility for highly reliable steganalysis of jpeg images," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 2772–2776.