PRACTICAL STRATEGIES FOR CONTENT-ADAPTIVE BATCH STEGANOGRAPHY AND POOLED STEGANALYSIS

Rémi Cogranne Member, IEEE, *Vahid Sedighi* Member, IEEE, *Jessica Fridrich* Fellow, IEEE [*]

ICD - LM2S - UMR 6281 CNRS
Troyes University of Technology
Troyes, France
remi.cogranne@utt.fr

Department of ECE
Binghamton University
Binghamton, NY 13902-6000
{vsedigh1,fridrich}@binghamton.edu

ABSTRACT

This paper investigates practical strategies for distributing payload across images with content-adaptive steganography and for pooling outputs of a single-image detector for steganalysis. Adopting a statistical model for the detector's output, the steganographer minimizes the power of the most powerful detector of an omniscient Warden, while the Warden, informed by the payload spreading strategy, detects with the likelihood ratio test in the form of a matched filter. Experimental results with state-of-the-art content-adaptive additive embedding schemes and rich models are included to show the relevance of the results.

Index Terms— Batch steganography, pooled steganalysis, adaptive embedding, optimal detection.

1. INTRODUCTION

Steganography alters innocuously looking cover objects in order to communicate in secrecy. This work focuses on steganography in *digital images*, arguably the most popular and most studied cover objects.

Recent years have seen a remarkable progress in steganography and steganalysis. Syndrome trellis codes [1] gave birth to numerous modern, content-adaptive data hiding algorithms [2, 3, 4]. The science of detection of hidden data called steganalysis has also remarkably improved over the past few years with the introduction of rich media models [5, 6] and new machine learning tools [7, 8].

In batch steganography, the payload is spread over multiple covers while pooled steganalysis jointly analyzes multiple objects for detection. Introduced in [9], these two topics are among the most pressing open problems today [10]. Payload spreading strategies for non-adaptive hiding schemes and targeted detectors were studied in [11, 12, 13] with the conclusion that the payload should either be concentrated in as few covers as possible or spread evenly.

Pooled steganalysis was studied in [12] under the assumption that the Warden knows the payload sizes but not their assignment to individual images. In a different setup, the authors of [14] used a local outlier factor to identify the steganographer in the wild. The topic of learning optimal pooling functions appeared in [15], and the problem of sequential detection was investigated in [16].

In this paper, we work in spatial domain because of the availability of good and tractable models that recently lead to important advances in steganography and steganalysis [4, 17]. By adopting a statistical model of Warden's detector, the problems of batch steganography and pooled steganalysis are formalized within a unified theoretical framework that allows us to design and study realistic payload spreading strategies that can be viewed as approximations of optimal but practically infeasible approaches.

In Section 2, we adopt a model of Warden's detector and formalize optimal pooled steganalysis in Section 3 and optimal content-adaptive batch steganography in Section 4. Several practical alternatives for optimal batch embedding strategies are introduced in Section 4. Experiments with content-adaptive embedding schemes and steganalysis detectors built with rich models appear in Section 5. The paper is concluded in Section 6.

2. PROBLEM STATEMENT

Abstracting away from specific details, we assume that the actors are aware of the state-of-the-art in the corresponding fields, namely that the steganographer uses content-adaptive embedding and the steganalyst a classifier trained on possibly high-dimensional features from $\mathbb{R}^d$. For now, we will disregard the details of exactly how the classifier is trained and simply assume that, when applied to a feature from a single

[*] This work has been done while Rémi Cogranne was a visiting scholar at Binghamton University. The work on this paper was supported by Air Force Office of Scientific Research under the research grants number FA9950-12-1-0124. 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.

image, the classifier θ returns a scalar output: $\theta: \mathbb{R}^d \to \mathbb{R}$ that is subsequently compared with a threshold to reach the decision.

Denoting the ith image with $\mathbf{x}^{(i)} = (x_{kl}^{(i)})$ and its representation in the feature space as $\mathbf{z}^{(i)} \in \mathbb{R}^d$, the steganographers generate a source of I images $\mathbf{x}^{(i)}$, $i = 1, \dots, I$, that are either all cover or all stego embedded with payloads R_i. The number of images is assumed to be arbitrarily large. On the other hand, the Warden inspects a set of B images $\mathbf{x}^{(i)}$, $i = 1, \dots, B$, with a classifier trained with a high-dimensional feature set. Due to the way the features are built and the fact that the test statistic is a projection of high-dimensional features, the Warden's single-image detector output, denoted $\theta^{(i)} = \theta(\mathbf{z}^{(i)})$, is a sample from a Gaussian distribution $\mathcal{N}(0, \sigma^2)$ [16, 18].

Given $B \geq 1$ images $\mathbf{x}^{(i)}$, $i = 1, \dots, B$, in Warden's *pooling bag*, the Warden faces the following hypothesis testing problem:

$$\mathcal{H}_0 : \theta^{(i)} \sim \mathcal{N}(0, \sigma^2),\ \forall i \tag{1}$$

$$\mathcal{H}_1 : \theta^{(i)} \sim \mathcal{N}\left(\mu_i(R_i), \sigma^2\right),\ \forall i, \tag{2}$$

where $\mu_i(R_i)$ is the expected shift of the detection statistic (over messages) when embedding payload size R_i in $\mathbf{x}^{(i)}$.

In (2), we adopted the so-called shift hypothesis [9], meaning that the embedding affects only the mean of the detector output but not its distribution. Note that we allow this shift to be a different function of the payload size for each image, hence the subscript i of μ. In addition, because we did not impose any assumption on the steganographers' payload spreading strategy, R_i can also be different for each image.

3. OPTIMAL POOLED STEGANALYSIS AND PRACTICAL APPROXIMATIONS

From the formulation of the hypotheses in (2), the Warden's problem consists of maximizing the detection accuracy given a set of B inspected images $\mathbf{x}^{(i)}$, $i = 1, \dots, B$.

The case in which the Warden does not have any information about the payload strategy used by the steganographers, $\mathbf{R} = (R_1, \dots, R_B)$, has been addressed in [16]. The conclusion of this prior work is that the steganalyst should simply average all outputs of the single-image detector $\theta^{(i)}$. In this paper, we consider an "omniscient" Warden who knows the spreading strategy $\mathbf{R} = (R_1, \dots, R_B)$ and the expectations $\mu_i(R_i)$. In this case, (2) reduces to a test between simple hypotheses for which the Most Powerful (MP) test that maximizes the detection power for a given false-alarm probability is the Likelihood Ratio (LR) test, the matched filter [20]:

$$\Lambda_B = \frac{1}{\sigma \left\|\boldsymbol{\mu}_B\right\|_2} \sum_{i=1}^{B} \mu_i(R_i)\theta^{(i)}, \tag{3}$$

where $\boldsymbol{\mu}_B = (\mu_1(R_1), \dots, \mu_B(R_B))$ denotes the outputs of Warden's detector on all images embedded with payloads $\mathbf{R}$ and $\|\mathbf{x}\|_2$ is the Euclidean norm of $\mathbf{x}$. The term $\left(\sigma \left\|\boldsymbol{\mu}_B\right\|_2\right)^{-1}$ is just a normalization factor.

From Eq. (3) and from the distribution of the detector's output (2), it is immediate that Λ_B follows

$$\Lambda_B \sim \begin{cases} \mathcal{N}(0,1) & \text{under } \mathcal{H}_0 \\ \mathcal{N}\left(\sqrt{B}\left\|\boldsymbol{\mu}_B\right\|_2, 1\right) & \text{under } \mathcal{H}_1. \end{cases} \tag{4}$$

Note that the power of this LRT only depends on $\left\|\boldsymbol{\mu}_B\right\|_2$.

Since in practice the steganalyst does not know the expectations $\mu_i(R_i)$, they need to be estimated. In this paper, we chose a polynomial regression of second degree in three variables: the payload R_i, the change rate r_i caused by embedding payload R_i, and ϱ_i, the deflection coefficient based on MiPOD's [4] cover model:

$$\varrho_i = \sum_{kl=1}^{N_p} (r_{kl}^{(i)})^2 (\sigma_{kl}^{(i)})^{-4}, \tag{5}$$

where $(\sigma_{kl}^{(i)})^2$ is the variance of pixel $x_{kl}^{(i)}$, estimated from its neighborhood as in [4, Section V], and $r_{kl}^{(i)}$ is the change rate of pixel $x_{kl}^{(i)}$. The regressor coefficients were estimated from the training part of the image database (images available to the Warden) using the least square estimator. Replacing the detector output expectation in the optimal LR under $\mathcal{H}_1$ by the estimates $\hat{\mu}_i$ gives us the "estimated LR":

$$\widehat{\Lambda}_B = \frac{1}{\sigma \left\|\hat{\boldsymbol{\mu}}_B\right\|_2} \sum_{i=1}^{B} \hat{\mu}_i \theta^{(i)}. \tag{6}$$

4. BATCH EMBEDDING BY MINIMIZING STATISTICAL DETECTABILITY (AND EMPIRICAL SOLUTIONS)

Once the steganographers have chosen an embedding scheme and have been granted access to a source of digital images $\mathbf{x}^{(i)}$, $i = 1, \dots, I$, their task is to select a payload spreading strategy $\mathbf{R} = (R_1, \dots, R_B)$ to minimize detectability. An omniscient steganographer with perfect knowledge of the expectation of Warden's detector output $\mu_i(R_i)$ for all $i = 1, \dots, I$ can find optimal combination of payloads, $\mathbf{R}^\star$, by minimizing the statistical detectability against the MP adversary or the expectation of the LR Λ_B (4), $\left\|\boldsymbol{\mu}_B\right\|_2$:

$$\mathbf{R}^\star = \arg\min_{\mathbf{R}} \sum_{i=1}^{B} \mu_i(R_i)^2, \tag{7}$$

$$\text{s.t.} \quad R = \sum_{i=1}^{B} R_i, \tag{8}$$

where Eq. (8) is the steganographer's payload constraint.

Evaluating for each image the expectation of the detector's output $\mu_i(R_i)$ can be very cumbersome. Additionally, the steganographers will likely be ignorant of the detector used by the Warden and especially how it has been trained. In this paper, we thus consider the following alternatives to optimal spreading that are feasible to implement in practice:

1. **Trust the steganography [Image Merging Sender (IMS)]**. The steganographer merges all B images into one and lets the embedding algorithm spread the payload.
2. **Trust the cover model [Detectability Limited Sender (DeLS)]**. The steganographer adopts a cover model and spreads payload over images so that each image from the bag contributes with the same value as the KL divergence (deflection coefficient) ϱ_i (5).
3. **Trust the distortion [Distortion Limited Sender (DiLS)]**. The steganographer spreads payload over images so that each image from the bag contributes with the same value of distortion.

Among the above strategies, the only one that minimizes a distortion (detectability) over all pixels from all images is the IMS. Here, the unknown expectation of the detectors' output is replaced with the distortion function on which the (adaptive) embedding scheme is based. The DeLS strategy is the only one that operates with statistical detectability. Ideally, the steganographer should minimize the total KL divergence $\sum_{i=1}^{B} \varrho_i$. However, this would be rather expensive to implement especially for large B and also computationally infeasible to test (see below). Similar complexity issues arise when minimizing the sum of distortions over all B images. Spreading the payload by finding a fixed value of the deflection coefficient (distortion) for all B images that communicates the required payload can be implemented much more efficiently.

It is important to note that experimental evaluation of each strategy in practice would be very time consuming as the Warden needs to determine the vector of payloads $\mathbf{R}$ and extract features *for each bag*, which makes the training very expensive. We hence adopt two more simplifying assumptions that will allow us to execute the experiments with a significantly lower computational burden. We will assume that over time the steganographers maintain an average communicated payload $\overline{R}$. Having access to a large number of images I, for each spreading strategy the steganographer determines the payload that would be embedded in the ith image if all $I \gg B$ images were in the embedding bag. This payload is essentially a *tag* attached to each image with the tag value depending on the spreading strategy. For large enough bags B, the actual payloads determined for each bag will be approximately the same as the "asymptotic" tags. This will allow us to execute the experiments by pre-computing the feature vectors for all images in the training (and testing) set and then performing the pooled steganalysis by randomly drawing B features from the testing set. Note that under this simplification, the embedding is "bagless" as the tags are determined only by $\overline{R}$ and the spreading strategy.

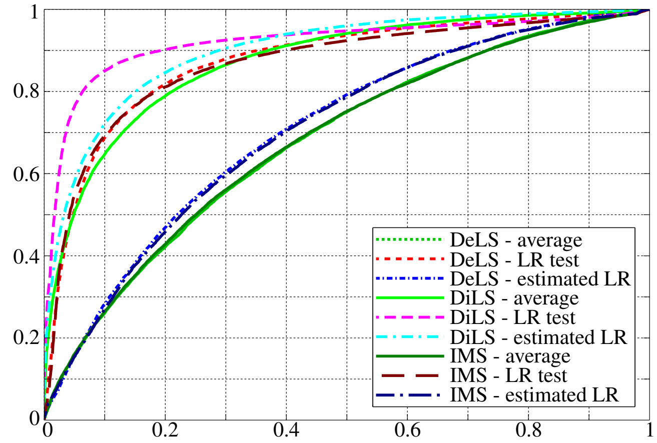

Fig. 1. Batch strategies and pooling methods evaluated using ROC curves for S-UNIWARD at $\overline{R} = 0.2$ bpp.

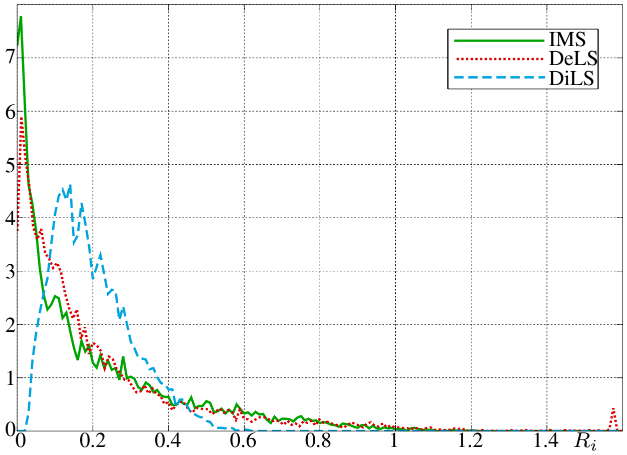

Fig. 2. Empirical distribution of payload size across BOSSbase images for the three proposed batch strategies.

5. NUMERICAL RESULTS

We use three additive state-of-the-art embedding schemes: S-UNIWARD [3], HILL [21], and MiPOD [4]. All three algorithms were modified to avoid making embedding changes in saturated pixels (grayscale with value 255) to prevent weakening their security as documented in [4, 22].

The image source in all experiments is BOSSbase 1.01 [23] split 10 times randomly into 5,000 images for training (Warden's database) and 5,000 for testing (Alice's database), unless mentioned otherwise. All experiments were conducted with a single-image LSMR-based linear classifier described in [8] trained on SRM features [5]. Since the Warden is omniscient, she trains on the same stego source produced by Alice, which depends on the embedding scheme, average secret communication rate $\overline{R}$, and spreading strategy.

Due to space limitations, we only include the most interesting results regarding the proposed batch steganography strategies and pooling methods. Table 1 shows the accuracy of three pooling methods as a function of the bag size B used by

Table 1. Evolution of minimal total probability of error under equal prior P_{E} as a function of pooling bag size B, $\overline{R} = 0.2$ bpp.

Embedding scheme	Batch strategy	Pooling method	Bag size: 1	2	5	10	20	50	100	200
HILL	DiLS	Average	0.4224	0.3930	0.3372	0.2811	0.2137	0.1134	0.0487	0.0175
		Optimal LR (3)	0.4224	0.3778	0.2936	0.2224	0.1513	0.0720	0.0380	0.0112
		Estimated LR (6)	0.4224	0.3852	0.3128	0.2431	0.1685	0.0725	0.0210	0.0120
	DeLS	Average	0.4521	0.4368	0.4077	0.3739	0.3291	0.2525	0.1782	0.1039
		Optimal LR (3)	0.4521	0.4272	0.3782	0.3304	0.2812	0.2043	0.1452	0.0903
		Estimated LR (6)	0.4521	0.4352	0.4030	0.3666	0.3225	0.2446	0.1737	0.1005
	IMS	Average	0.4663	0.4506	0.4277	0.4026	0.3663	0.3046	0.2369	0.1594
		Optimal LR (3)	0.4663	0.4444	0.3988	0.3515	0.2930	0.2150	0.1528	0.1014
		Estimated LR (6)	0.4663	0.4483	0.4208	0.3923	0.3516	0.2838	0.2171	0.1408
MiPOD	DeLS/DiLS	Average	0.4444	0.4260	0.3924	0.3528	0.3004	0.2138	0.1375	0.0665
		Optimal LR (3)	0.4444	0.4178	0.3638	0.3129	0.2581	0.1852	0.1334	0.0586
		Estimated LR (6)	0.4444	0.4258	0.3885	0.3482	0.2948	0.2074	0.1341	0.0670
	IMS	Average	0.4610	0.4477	0.4204	0.3905	0.3527	0.2802	0.2129	0.1312
		Optimal LR (3)	0.4610	0.4368	0.3863	0.3200	0.2522	0.1590	0.1080	0.0691
		Estimated LR (6)	0.4610	0.4419	0.4095	0.3680	0.3206	0.2391	0.1693	0.0967

Eve for steganalysis. Note that the DiLS is by far the worst, though for larger bag size the difference becomes rather small, while the DeLS performs almost as well as the IMS. The ranking of steganographic algorithms, batch strategies, and pooling methods also almost always remain consistent over the bag size, which is in agreement with the proposed statistical model presented in Sections 2–3. Regarding the accuracy of the pooling methods, as one can expect the optimal LR test always has the best performance while the proposed "estimated LR" reaches a comparable performance only for DiLS. For the most secure DeLS and IMS batch strategies, the performance of the estimated LR drops significantly in comparison with the optimal test, indicating thus a possible detection improvement with a more accurate predictor of the detector output. The average test proposed in [16] for an ignorant Warden does not know the batch strategy and performs much worse across all batch strategies.

The P_{E} for bag size $B = 1$ in the table corresponds to the performance of the single-image detector. It should be contrasted with the single-image detector trained to detect the most commonly considered uniform spreading strategy or payload-limited sender (PLS) that embeds the mean payload $\overline{R} = 0.2$ bpp in every image. The detector of the PLS achieves $P_{\mathrm{E}} \approx 0.35$ for both HILL and MiPOD while the DeLS and IMS senders are detected by the single-image detector at $P_{\mathrm{E}} \approx 0.46$, which testifies to the rather large gain in security due to payload spreading. It is also worth pointing out that, while MiPOD has been evaluated as the most secure additive embedding scheme [8, 4, 22], it appears slightly less secure than HILL in the batch mode as HILL seems to spread the payload more efficiently across multiple images.

The ROC curves in Fig. 1 show the probability of correct detection as a function of the false-alarm probability for S-UNIWARD [3] with mean payload $\overline{R} = 0.1$ bpp and pooling bag size $B = 100$. We note that, in agreement with the statistical model described in Sec. 2, the ranking of batch strategies and pooling methods is consistent for almost all false-alarm probabilities. Also note the small differences between the accuracy of pooling methods for DiLS (lightest curves) compared to other batch strategies and how the proposed "estimated LR" fails to match the performance of the optimal LR. Finally, note how close the IMS and DeLS spreading strategies are in terms of security. This, of course, depends on the statistical model of the cover Alice uses and on the distortion function on which the embedding is based.

Last, but not least, Fig. 2 provides a useful insight about the batch strategies. This figure shows the empirical distribution of payloads among all 10,000 BOSSbase images for all batch strategies for S-UNIWARD at mean payload $\overline{R} = 0.2$ bpp. We note that the IMS and DeLS strategies are similar in the sense that both tend to put small payloads in the vast majority of images and allocate most of the hidden data in a limited number of images for which the detection is the most challenging, such as very textured images. On the other hand, the DiLS often embeds a payload that is close to the mean payload $\overline{R}$ and thus adapts to image content to a much smaller degree. Qualitatively similar results have been observed for other embedding schemes and all tested mean payloads.

6. CONCLUSIONS

In this paper, we study the problem of content-adaptive batch steganography and pooled steganalysis for an omniscient Warden aware of the payload-spreading strategy and equipped with a single-image detector trained as a classifier between the cover and stego sources. By adopting a statistical model for the output of the single-image detector, we derive the optimal pooling function as a likelihood ratio in the form of a matched filter and its approximations realizable in practice. We also consider several batch strategies that can be efficiently implemented in practice and test them, together with pooling strategies on state-of-the-art steganography, drawing numerous interesting conclusions for practitioners of steganography as well as the steganalyst.

7. REFERENCES

[1] T. Filler, J. Judas, and J. Fridrich, "Minimizing additive distortion in steganography using syndrome-trellis codes," *IEEE Trans. On Information Forensics and Security*, vol. 6, no. 3, pp. 920–935, September 2011.

[2] T. Penný, T. Filler, and P. Bas, "Using high-dimensional image models to perform highly undetectable steganography," in *Proc. Information hiding*, Calgary, Canada, June 28–30, 2010, vol. 6387 of LNCS, pp. 161–177.

[3] V. Holub, J. Fridrich, and T. Denemark, "Universal distortion design for steganography in an arbitrary domain," *EURASIP Journal on Information Security*, vol. 2014:1, 2014.

[4] V. Sedighi, R. Cogranne, and J. Fridrich, "Content-adaptive steganography by minimizing statistical detectability," *IEEE Trans. On Information Forensics and Security*, vol. 11, no. 2, pp. 221–234, 2016.

[5] J. Fridrich and J. Kodovský, "Rich models for steganalysis of digital images," *IEEE Trans. On Information Forensics and Security*, vol. 7, no. 3, pp. 868–882, June 2011.

[6] T. Denemark, V. Sedighi, V. Holub, R. Cogranne, and J. Fridrich, "Selection-channel-aware rich model for steganalysis of digital images," in *Proc. IEEE Workshop on Information Forensics and Security (WIFS)*, Atlanta, GA, December 3–5, 2014.

[7] J. Kodovský, J. Fridrich, and V. Holub, "Ensemble classifiers for steganalysis of digital media," *IEEE Trans. On Information Forensics and Security*, vol. 7, no. 2, pp. 432–444, 2012.

[8] R. Cogranne, V. Sedighi, T. Pevný, and J. Fridrich, "Is ensemble classifiers needed for steganalysis in high-dimensional feature spaces?," in *IEEE Workshop on Information Forensics and Security (WIFS)*, Rome, Italy, November 16–19 2015.

[9] A. D. Ker, "Batch steganography and pooled steganalysis," in *Proc. Information hiding*, Alexandria, VA, July 10–12, 2006, vol. 4437 of LNCS, pp. 265–281.

[10] A. D. Ker, P. Bas, R. Böhme, R. Cogranne, S. Craver, T. Filler, J. Fridrich, and T. Pevný, "Moving steganography and steganalysis from the laboratory into the real world," in *Proc. ACM IH&MMSec*, Montpellier, France, June 17–19, 2013.

[11] A. D. Ker, "Batch steganography and the threshold game," in *Proc. SPIE, Electronic Imaging, Security, Steganography, and Watermarking of Multimedia Content IX*, E. J. Delp and P. W. Wong, Eds., San Jose, CA, January 29–February 1, 2007, vol. 6505, p. 04 1–13.

[12] A. D. Ker, "Perturbations hiding and the batch steganography problem," in *Proc. Information hiding*, Santa Barbara, CA, June 19–21, 2008, vol. 5284 of LNCS, pp. 45–59.

[13] A. D. Ker and Tomas Penný, "Batch steganography in the real world," in *Proc. ACM MM&Sec*, J. Dittmann, S. Craver, and S. Katzenbeisser, Eds., Coventry, UK, September 6–7, 2012, pp. 1–10.

[14] A. D. Ker and T. Penný, "The steganographer is the outlier: Realistic large-scale steganalysis," *IEEE Trans. On Information Forensics and Security*, vol. 9, no. 9, pp. 1424–1435, September 2014.

[15] T. Penny and I. Nikolaev, "Optimizing pooling function for pooled steganalysis," in *IEEE Workshop on Information Forensics and Security (WIFS)*, Nov 2015, pp. 1–6.

[16] R. Cogranne, "A sequential method for online steganalysis," in *IEEE Workshop on Information Forensics and Security (WIFS)*, November 2015, pp. 1–6.

[17] A. D. Ker, T. Pevný, and P. Bas, "Rethinking optimal embedding," in *Proc. ACM IH&MMSec*, Vigo, Spain, June 20–22, 2016.

[18] R. Cogranne and J. Fridrich, "Modelling and extending the ensemble classifier for steganalysis of digital images using hypothesis testing theory," *IEEE Trans. On Information Forensics and Security*, vol. 10, no. 2, pp. 2627–2642, December 2015.

[19] R. Böhme and A. D. Ker, "A two-factor error model for quantitative steganalysis," in *Proc. SPIE, Electronic Imaging, Security, Steganography, and Watermarking of Multimedia Content VIII*, San Jose, CA, January 16–19, 2006, vol. 6072, pp. 59–74.

[20] S. M. Kay, *Fundamentals of Statistical Signal Processing, Volume II: Detection Theory*, vol. II, Upper Saddle River, NJ: Prentice Hall, 1998.

[21] B. Li, M. Wang, and J. Huang, "A new cost function for spatial image steganography," in *IEEE, International Conference on Image Processing (ICIP)*, Paris, France, October 27–30, 2014.

[22] V. Sedighi and J. Fridrich, "Effect of saturated pixels on security of steganography schemes for digital images," in *IEEE, International Conference on Image Processing (ICIP)*, Phoenix, Arizona, September 25–28, 2016.

[23] P. Bas, T. Filler, and T. Pevný, "Break our steganography system – the ins and outs of organizing BOSS," in *Proc. Information hiding*, Prague, Czech Republic, May 18–20, 2011, vol. 6958 of LNCS, pp. 59–70.