Secure Payload Scaling in Detector-Informed Batch Steganography: The Mismatched Detectors Case

Eli Dworetzky
Department of ECE
Binghamton University
Binghamton, NY, USA
edworet1@binghamton.edu

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

Abstract

This paper deals with the problem of batch steganography and pooled steganalysis when the sender uses a steganography detector to spread chunks of the payload across a bag of cover images while the Warden uses a possibly different detector for her pooled steganalysis. We investigate how much information can be communicated with increasing bag size n at a fixed statistical detectability of Warden's detector. Specifically, we are interested in the scaling exponent γ of the secure payload $P(n) = cn^\gamma$. We approach this problem both theoretically from a statistical model of the soft output of a detector and practically using experiments on real datasets when giving both actors different detectors implemented as convolutional neural networks and a classifier with a rich model. While the effect of the detector mismatch depends on the payload allocation algorithm and the type of mismatch, in general the mismatch decreases the constant of proportionality c as well as the exponent γ. This stays true independently of who has the superior detector. Many trends observed in experiments qualitatively match the theoretical predictions derived within our model. Finally, we summarize our most important findings as lessons for the sender and for the Warden.

CCS Concepts

• **Security and privacy**; • **Computing methodologies** → **Neural networks**; **Image manipulation**;

Keywords

Batch steganography, pooled steganalysis, secure payload, scaling, mismatched detectors

ACM Reference Format:
Eli Dworetzky and Jessica Fridrich. 2025. Secure Payload Scaling in Detector-Informed Batch Steganography: The Mismatched Detectors Case. In *ACM Workshop on Information Hiding and Multimedia Security (IH&MMSEC '25), June 18–20, 2025, San Jose, CA, USA.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3733102.3733134

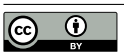
1 Introduction

In batch steganography and pooled steganalysis [13], the sender (Alice) spreads her secret payload across a bag of images to decrease the chances of being caught while the Warden inspects all images in the bag to detect the use of steganography. Recognizing that detectability of steganographic embedding strongly depends on image content, in the past researchers looked at various payload spreading strategies that assign payload chunks to images based on content complexity. For instance, the Image Merging Sender (IMS) [18] considers a bag of cover images as a single image and spreads payload by embedding this larger "image" using a content-adaptive steganographic algorithm. Recently, payload spreading strategies have been proposed that make use of a trained detector. They include the Minimum Deflection Sender (MDS) [20], the Shift Limited Sender (SLS) [20], and the greedy sender [8]. All three in some way assign the largest payload to images that, when embedded, elicit the smallest detector response.

In this paper, we study how much information Alice can send in bags of increasing size n as $n \to \infty$. Since in practice Alice is unlikely to use very large bags, perhaps a more sensible interpretation of our research goal is addressing the question of how Alice should adjust her payload over n uses of the stego channel to stay within a limited risk of being detected. The most closely related prior art is [8], where the authors studied this problem primarily for the case when Alice and the Warden share the same detector, which Alice uses for payload allocation and the Warden for detection. The secure payload size $P(n)$ that guarantees constant statistical detectability was observed to scale as $P(n) \propto n^\gamma$ with $\gamma \approx 0.85$ for bag sizes up to $n = 16,000$. This surprising super-square root law (SRL) [15] scaling was attributed to the fact that in the dataset used by the authors (spatial domain ALASKA II [5]) a non-negligible fraction of images appear to have a vanishing response to embedding in terms of a detector's soft output, an observation the authors conjecture holds for all typical image datasets.

While the main focus of this prior art was the matched detector case, the authors did provide some limited results on scaling when Alice spreads her payload with a different detector than the one used by the Warden for detection (the mismatched detector case). The detectors were implemented as two different architectures of a convolutional neural network (CNN). The experiments revealed that, depending on the payload spreading strategy and Warden's pooler, the secure payload may follow the same super SRL scaling but could also exhibit an exponent strictly smaller than 0.80.

The current paper begins where this prior art ends. Our goal is a more detailed study of the secure payload scaling when both actors use different detectors. We experiment with three types of state-of-the-art CNNs and a classifier based on the spatial rich model (SRM) [10], and study the scaling for all mismatched as well as matched cases. To explain the trends observed in experiments, we formalize and quantify detector mismatch mathematically within



IH&MMSEC '25, San Jose, CA, USA

ACM ISBN 979-8-4007-1887-8/2025/06
https://doi.org/10.1145/3733102.3733134

a statistical model of the soft output of a detector in a way that is relevant to secure payload scaling. A scalar detector-mismatch parameter $\theta \geq 0$ is introduced to quantify the drop in the scaling exponent γ due to the mismatch. We explain how this parameter can be estimated in practice and show that it can predict the impact of detector mismatch on the scaling.

The paper is structured as follows. In the next section, we introduce notational conventions and a list of commonly used abbreviations and acronyms. Section 3 contains the details of two payload spreading strategies used by Alice for batch steganography that will be studied in this paper. In Section 4, we adopt a statistical model of the soft output of Warden's detector and derive the most powerful pooler for pooled steganalysis. The concepts of secure payload and a detectability limited sender (DLS) are formalized in Section 5, including the implementation details of the DLS. The scaling of secure payload with matched and mismatched detectors is theoretically analyzed within our model in Section 6. Based on this analysis, we introduce the detector mismatch parameter and explain how it affects the scaling of secure payload. The setup of our experiments, including the datasets, detector training, and other implementation details are described in Section 7. All experimental results are presented and contrasted with the theoretical analysis in Section 8. After discussing the implications of our findings to Alice and the Warden in Section 9, the paper is concluded in Section 10.

2 Notation

Throughout the paper, cover images are denoted with X, while images intercepted by the Warden (either cover or stego) are represented with the symbol Y. Boldface symbols are used for n-tuples of objects. In particular, a bag of n cover images will be denoted $\mathbf{X} = (X_1, \ldots, X_n)$ while a bag intercepted by the Warden will be denoted $\mathbf{Y} = (Y_1, \ldots, Y_n)$. Both Alice and the Warden use a single-image detector (SID) d to achieve their respective goals. Throughout this paper, the superscripts 'A' and 'W' will be used for quantities and objects that depend on Alice's and Warden's detector, respectively. Alice uses d^{A} to spread her payload while the Warden uses d^{W} to detect steganography. Formally, a SID is a mapping $d : \mathcal{X} \to \mathbb{R}$, where $\mathcal{X}$ is the space of all images. When d is a trained CNN, we use its stego logit as the soft output. For a detector implemented as an LCLC classifier [6] with a rich model, the soft output is the projection of the rich feature on the weight vector.

A real valued random variable Z restricted to interval $[a, b]$ will be denoted $Z[a, b]$. Gaussian random variable with mean μ and variance σ^2 will be denoted $\mathcal{N}(\mu, \sigma^2)$, while $U[a, b]$ stands for the uniform random variable on $[a, b]$.

Common abbreviations and acronyms: SRM = spatial rich model, RC = response curve, SID = single image detector, SLS = shift limited sender, bpp = bits per pixel, DLS = detectability limited sender, PLS = payload limited sender, CNN = convolutional neural network, WLOG = without loss of generality.

3 Detector-informed batch steganography

In this section, we describe how Alice uses her detector to allocate payload chunks to individual images in the bag. She uses detector feedback for this task based on how her detector responds to embedding in a given cover image.

3.1 Response curve

Let C be the maximum embedding capacity of a cover image $X \in \mathcal{X}$. For example, for a ternary embedding scheme in the spatial domain, $C \leq \log_2 3$ bits per pixel (bpp). Given a cover image X, detector d^{A}, and an embedding scheme, the response curve (RC) is the function $\varrho^{\mathrm{A}} : [0, C] \to \mathbb{R}$ defined by

$$\varrho^{\mathrm{A}}(\alpha) = \mathbb{E}[d^{\mathrm{A}}(X(\alpha; K))], \quad 0 \leq \alpha \leq C, \tag{1}$$

where $X(\alpha; K)$ is X embedded with a secret message of relative length α bpp and stego key K. The expectation is taken over random messages and stego keys. Furthermore, we define the (expected) shift of the detector response

$$s^{\mathrm{A}}(\alpha) = \varrho^{\mathrm{A}}(\alpha) - \varrho^{\mathrm{A}}(0). \tag{2}$$

Note that the RC and the shift depend on the image, the detector, and the embedding scheme.

3.2 Payload spreading

In this paper, we consider two payload spreading strategies (senders) that make use of the RCs. The SLS and greedy senders were selected because they are computationally inexpensive, which is important since we run experiments with bags consisting of thousands of images for a wide range of bag sizes and various combinations of detectors. The greedy sender has been included mainly because it is easier to analyze mathematically. As will be seen later in this paper, it is overly aggressive with its payload assignment and is thus very sensitive to detector mismatch.

Let us assume that the sender has a bag of n cover images $X_1, \ldots, X_n$ with embedding capacities $0 \leq C_i \leq \log_2 3$ bpp. WLOG, let us assume that the images are ordered by their detector shifts at capacity $s_1^{\mathrm{A}}(C_1) \leq \cdots \leq s_n^{\mathrm{A}}(C_n)$. Let $P \in \left[0, \sum_{i=1}^n C_i\right]$ bpp be the total payload the sender wants to communicate in the bag and k be the largest integer for which $\sum_{i=1}^k C_i < P$. The greedy sender fully embeds images $X_1, \ldots, X_k$ with $\alpha_i = C_i$ while X_{k+1} holds the last chunk $\alpha_{k+1} = P - \sum_{i=1}^k C_i$ bpp and $X_{k+2}, \ldots, X_n$ are left empty.

The shift limited sender (SLS) [20] finds the smallest $\delta > 0$ that leads to the same expected detector output shift when embedding payload α_i in X_i, satisfying $\sum_{i=1}^n \alpha_i = P$, and $\delta = s_i^{\mathrm{A}}(\alpha_i)$ for all i for which $s_i^{\mathrm{A}}(C_i) \geq \delta$. For images that do not satisfy this condition (e. g., images with "flat" response curves), the SLS sets $\alpha_i = C_i$. In other words, SLS enforces the shift hypothesis [13].

In the rare case when the RC of an image is not monotonically increasing, the greedy sender uses the absolute value $|s_i^{\mathrm{A}}(\alpha)|$ and SLS uses the cumulative max of $|s_i^{\mathrm{A}}(\alpha)|$ in their implementations.

4 Warden's pooler

In this section, we explain in detail the pooled steganalysis executed by the Warden. To this end, we impose a statistical model on the soft output of her detector, which allows us to derive the most powerful pooler within the adopted model.

Having intercepted a bag of n images $\mathbf{Y} = (Y_1, \ldots, Y_n)$, the Warden applies d^{W} to each Y_i and then pools $d^{\mathrm{W}}(Y_i)$, $i = 1, \ldots, n$ with a pooler $\pi : \mathbb{R}^n \to \mathbb{R}$, arriving at $\pi(d^{\mathrm{W}}(Y_1), \ldots, d^{\mathrm{W}}(Y_n))$ as her detection statistic. For brevity, we will slightly abuse notation using $\pi(\mathbf{Y}) = \pi(d^{\mathrm{W}}(Y_1), \ldots, d^{\mathrm{W}}(Y_n))$.

We will assume that sampling covers from $\mathcal{X}$ is a two-stage process. First, Alice selects a "scene" and then acquires it with a digital camera. If she were to take multiple images of the exact same scene, they would slightly differ due to the sensor noise but will generally concentrate around the noise-free version of the scene. To avoid the complexity of modeling the images themselves, we instead model the soft outputs of Warden's detector as in [7].

4.1 Statistical model

Given the scene of the ith cover image, we model the distribution of detector outputs on acquisitions X_i of this scene as

$$d^{\mathrm{W}}(X_i) \sim \mathcal{N}(\mu_i, \sigma^2). \tag{3}$$

Since the acquisitions are concentrated on a small subset of $\mathcal{X}$, d^{W} will be approximately linear on such small neighborhoods, the Gaussianity can be heuristically justified by the central limit theorem at least for the case of RAW captures where the acquisition noise is independent across pixels. While the variance of the response, σ^2 , is generally a function of the image X_i, we assume it is constant to simplify the modeling. The reader is referred to [20] and [7] for a more in-depth discussion of these modeling assumptions.

Since stego schemes strive to preserve statistical properties of X_i, the embedding process will also preserve the concentration. Therefore, by the same argument we assume that the detector output on the stego image embedded with relative payload α_i bits per pixel (bpp), $d^{\mathrm{W}}(X_i(\alpha_i))$, is also Gaussian

$$d^{\mathrm{W}}(X_i(\alpha_i)) \sim \mathcal{N}(\mu_i + s_i^{\mathrm{W}}(\alpha_i), \sigma^2). \tag{4}$$

Note that we assume only the mean is affected by embedding but not the variance. This *local* shift hypothesis is a much weaker assumption than the shift hypothesis [18] about the distribution of detector response across scenes which is not satisfied for modern steganalyzers built with machine learning (see, e. g., Sec. 3.2 in [20]).

4.2 Warden's hypothesis test

Given a bag of intercepted images $\mathbf{Y} = (Y_1, \dots, Y_n)$ the Warden faces the following hypothesis test assuming the $d^{\mathrm{W}}(Y_i)$ are independent random variables:

$$\begin{aligned} \mathcal{H}_0 &: \quad d^{\mathrm{W}}(Y_i) \sim \mathcal{N}(\mu_i, \sigma^2) \quad \text{for all } i \\ \mathcal{H}_1 &: \quad d^{\mathrm{W}}(Y_i) \sim \mathcal{N}(\mu_i + s_i^{\mathrm{W}}(\alpha_i), \sigma^2) \quad \text{for all } i, \end{aligned} \tag{5}$$

where α_i is the relative payload potentially residing in the ith image Y_i and $s_i^{\mathrm{W}}(\alpha_i) = \varrho_i^{\mathrm{W}}(\alpha_i) - \varrho_i^{\mathrm{W}}(0)$. Assuming the parameters of the distributions in the hypothesis test (5) are known to the Warden (this includes the payloads α_i and shifts $s_i^{\mathrm{W}}(\alpha_i)$), the test becomes simple and, due to the independence, the Warden's most powerful pooled detector is the correlator

$$\pi_{\mathrm{corr}}(\mathbf{Y}) = \sum_{i=1}^{n} d^{\mathrm{W}}(Y_i) s_i^{\mathrm{W}}(\alpha_i). \tag{6}$$

The detectability of steganography is then determined by the deflection coefficient

$$\Delta^2 = \frac{1}{\sigma^2} \sum_{i=1}^{n} (s_i^{\mathrm{W}}(\alpha_i))^2. \tag{7}$$

5 Secure payload

In this section, we define the concept of secure payload and describe a detectability-limited sender (DLS) that will be studied in this paper. This material closely follows Section Secure Payload from [8], hence we include only a condensed description and refer the reader to the original publication for more details.

Given a fixed steganographic scheme and spreading strategy, we define the *secure payload of a bag of size n* at detectability $\delta \geq 0$, $P_\delta(n)$, as the largest total payload P that can be communicated in a bag of n images that satisfies

$$\mathbb{E}[\Delta^2] \leq \delta, \tag{8}$$

where the expectation is taken over all bags of size n sampled independently from the cover source.

5.1 Detectability-limited sender

A detectability-limited sender adjusts the payload in the bag in order to satisfy (8). Given a desired statistical detectability $\delta \geq 0$ for the Warden's pooler π and bag size n, a DLS determines the maximal secure payload size $P_\delta(n)$. In experiments, we use an empirical measure of detectability in the form of Warden's minimum total average error probability $P_{\mathrm{E}} = \min \frac{1}{2}(P_{\mathrm{FA}} + P_{\mathrm{MD}})$, where P_{D} and P_{FA} are the detector's power and false alarm rate, respectively.

The DLS used in this paper is specifically an empirical DLS; it fixes detectability across a collection of N bags rather than detectability within a statistical model of a specific bag. To achieve a fixed detectability across N bags, we use a payload-limited sender to solve for the total payload that achieves the desired detectability δ. This is implemented as a binary search for $P_\delta(n)$. Note that a DLS implemented this way embeds the same payload $P_\delta(n)$ in every bag while a DLS that fixes detectability based on the statistical model of each bag would embed variable payloads depending on the bag.

The binary search looks for payload size P bpp that elicits a chosen P_{E} for Warden's pooler on bags of size n. The images in bags are sampled without replacement and then payload P is embedded using a spreading strategy. Each stego image $X_i(\alpha_i)$ was generated using a random key, producing the stego bag $\mathbf{X}(\boldsymbol{\alpha})$. Warden's pooled detector is then applied to both cover and stego bags, yielding $\pi(\mathbf{X})$ and $\pi(\mathbf{X}(\boldsymbol{\alpha}))$ respectively. Repeating these steps for each of the N bags, we compute the empirical detectability $P_{\mathrm{E}}(P, n, N)$ from the collection of $2N$ data points $\left\{\pi(\mathbf{X}^{(m)}), \pi(\mathbf{X}^{(m)}(\boldsymbol{\alpha}^{(m)}))\right\}_{m=1}^{N}$ where the superscript (m) signifies the mth sample cover / stego bag. The search ultimately solves for the payload P such that $P_{\mathrm{E}}(P, n, N) = \delta$. We denote the payload found this way by $P_\delta(n)$, omitting the dependence on N since $P_{\mathrm{E}}(P, n, N)$ converges to a limit for large enough N.

5.2 Simulating the DLS

Implementing the DLS described above is rather expensive as it involves running the embedding simulator and computing forward passes of Warden's detector for $O(N \times n)$ images per iteration of the binary search. Since we wish to study a wide variety of mismatched detectors, this further increases the computational demands. To speed up our experiments, instead of actually embedding the cover images and applying the detector to them, we merely sampled the Gaussian model of the images' RCs (4) in a Monte Carlo fashion.

As shown in Figure 4 in the original publication [8], this simulation gives essentially the same results in terms of scaling of the secure payload as executing the DLS as described in the previous section.

6 Effect of detector mismatch on scaling

Modeling a mismatch between two detectors is generally difficult and even ill posed since the mismatch can have many forms and practical consequences. In this paper, we model the mismatch in a way that reflects our goal, which is to study its impact on secure payload scaling. To explain our reasoning and motivate our approach, we first review a relevant result from [8]. As in this prior art, we restrict our study to the greedy sender because the results are available in a closed form as one only needs to adopt a model for detector shifts at capacity $s_i^{\mathrm{W}}(C_i)$. To carry out a similar analysis for the SLS, it would be necessary to model the entire response curves.

For simplicity and WLOG, we will assume that $\sigma^2 = 1$ and that the embedding capacity of all images is the same, $C_i = C = \log_2 3$ for all i. Finally, we will ignore the contribution of the last partially embedded image to the deflection as this will become negligible as n tends to infinity.

We begin with adopting a model of the square detector shift at capacity, $(s_i^{\mathrm{W}}(C))^2$, when sampling the image space $\mathcal{X}$ according to the cover distribution:

$$(s_i^{\mathrm{W}}(C))^2 \sim F^{\mathrm{W}}, \tag{9}$$

where F^{W} is a cumulative distribution function (CDF) supported on $[0, \infty)$. We will assume that $F^{\mathrm{W}}(x) > 0$ for $x > 0$ and that it is continuous and invertible on some right neighborhood of zero. Note that F^{W} generally depends on the cover source, steganographic method, and Warden's detector, as also highlighted by the superscript 'W'.

6.1 Matched detectors case

The following secure payload scaling theorem has been proved for the case of matched detectors in [8].

Theorem 6.1. *The secure payload of the greedy sender for bag size n and detectability $\delta > 0$ is*

$$P_\delta(n) = \log_2 3 \times k(n) \quad \textit{(bpp)}, \tag{10}$$

where $k(n)$ is the largest integer satisfying $n \times \int_0^{k(n)/n} (F^{\mathrm{W}})^{-1}(x)\mathrm{d}x \le \delta$. Moreover, when $F^{\mathrm{W}}(x) \propto x^\beta$, $\beta > 0$, on some right neighborhood of 0, $P_\delta(n) \propto n^\gamma$,

$$\gamma(\beta) = \frac{1}{1+\beta}. \tag{11}$$

In this paper, we draw inspiration from the proof of this result, specifically from the following Lemma proved in the Appendix [8]. Consider n i.i.d. random variables $Z_i \sim F^{\mathrm{W}}$ as well as the kth order statistic of the Z_i, denoted by $Z_{(i)}$. Suppose that $k = k(n)$ is some function of n satisfying $cn^{1/2} \le k(n)$ (secure payload is at least $\propto n^{1/2}$) for some $c > 0$ and $k(n)/n \to 0$ (secure payload is sublinear) as $n \to \infty$. Then, $F(Z_{(k)})\frac{n+1}{k(n)} \to 1$ in probability as $n \to \infty$ and $F^{\mathrm{W}}(Z_{(i)})$, $1 \le i \le k(n)$, are uniform on $[0, k(n)/n]$:

$$F^{\mathrm{W}}(Z_{(i)}) \sim U[0, k(n)/n]. \tag{12}$$

This result serves as the starting point for modeling the detector mismatch below.

6.2 Mismatched detectors

The greedy sender implemented by Alice embeds the payload in $k(n)$ images with the smallest shifts $(s_i^{\mathrm{A}}(C))^2$ w.r.t. her detector d^{A}. When the detectors are mismatched, however, these images do not need to have the smallest shift w.r.t. d^{W}. The more mismatched the detectors are, the more likely Alice is to embed in images that rank in terms of the Warden detector shifts $(s_i^{\mathrm{W}}(C))^2$ very differently. Let $X_{[i]}$ be the cover images in the bag ordered by $(s_i^{\mathrm{A}}(C))^2$ from the smallest to the largest. Due to the mismatch, $X_{[i]}$ no longer coincide with Warden's order statistics $X_{(i)}$ when ordering the images by $(s_i^{\mathrm{W}}(C))^2$. We model this detector mismatch mathematically by claiming that $F^{\mathrm{W}}\left((s_{[i]}^{\mathrm{W}}(C))^2\right)$ for $i = 1, \dots, k(n)$ are sampled by Alice independently and uniformly from a *larger interval* (c. f., (12))

$$F^{\mathrm{W}}\left((s_{[i]}^{\mathrm{W}}(C))^2\right) \sim U[0, n^\theta k(n)/n], \tag{13}$$

where the scalar $\theta \ge 0$ quantifies the strength of the mismatch. Note that (13) implies that $k(n) \le n^{1-\theta}$ since $\frac{k(n)}{n} n^\theta \le 1$. For matched detectors ($\theta = 0$), this condition means that secure payload is at most linear. To obtain the scaling result for mismatched detectors below, we will assume a slightly stronger condition $\frac{k(n)}{n} n^\theta \to 0$ as $n \to \infty$.

At first glance the above definition of the mismatch may seem a bit contrived. In reality, the upper bound of the support of $F^{\mathrm{W}}\left((s_{[i]}^{\mathrm{W}}(C))^2\right)$ could be any function of n lower bounded by $k(n)/n$, so why scale $k(n)/n$ by the factor n^θ? There are two important advantages of capturing the mismatch in this fashion. First, the parametric model (13) can be verified with inexpensive experiments. Second, we can easily obtain the payload scaling result for mismatched detectors as follows.

Let $Z \sim F^{\mathrm{W}}$ be the random variable modeling the distribution of $(s_i^{\mathrm{W}}(C))^2$. Denoting for compactness $a_{k,n} = (F^{\mathrm{W}})^{-1}\left(k(n)n^{\theta-1}\right)$, from (13) we see that the $(s_{[i]}^{\mathrm{W}}(C))^2$ are equal in distribution to $Z[0, a_{k,n}]$, i.e., Z restricted to $[0, a_{k,n}]$. Since all $k(n)$ images selected by Alice are fully embedded, the expected deflection (7) of Warden's detector is[1]

$$\mathbb{E}[\Delta^2] = k(n)\mathbb{E}[Z[0, a_{k,n}]]. \tag{14}$$

The term $\mathbb{E}[Z[0, a_{k,n}]]$ can be computed using the definition of conditional expectation

$$\mathbb{E}[Z[0, a_{k,n}]] = \frac{1}{F^{\mathrm{W}}(a_{k,n})} \int_0^{a_{k,n}} x \mathrm{d}F^{\mathrm{W}}(x).$$

This allows us to compute the secure payload $k(n) \times \log_2 3$ from the condition

$$\begin{aligned} \delta &= k(n)\mathbb{E}[Z[0, a_{k,n}]] \\ &= n^{1-\theta} \int_0^{k(n)n^{\theta-1}} (F^{\mathrm{W}})^{-1}(x)\mathrm{d}x. \end{aligned} \tag{15}$$

[1] Recall WLOG that $\sigma^2 = 1$ and the partially embedded image is ignored.

Observe that, as $n \to \infty$, the region of integration is a vanishing right-neighborhood of 0. Hence, the asymptotics of the secure payload $k(n) \times \log_2 3$ are determined by the (left-tail) asymptotics of $F^{\mathrm{W}}(x)$ as $x \to 0^+$.

As a corollary, we consider the case when $F^{\mathrm{W}}(x) \propto x^{\beta}$, $\beta > 0$, on a small right-neighborhood of 0. It follows that $(F^{\mathrm{W}})^{-1}(x) \propto x^{1/\beta}$, and we have $\delta = n^{1-\theta}\frac{\beta}{\beta+1}\left(kn^{\theta-1}\right)^{\frac{\beta+1}{\beta}}$ for sufficiently large n. Solving for $k(n)$ yields the following secure payload scaling with detector mismatch

$$k(n) \propto n^{\frac{1-\theta}{1+\beta}}. \tag{16}$$

Such a case is reasonable to consider because many positive random variables have distribution functions with left-tail power scaling, e.g., Gamma and Beta prime.

We wish to point out that θ is not symmetrical w.r.t. the pair of detectors d^{A}, d^{W} in the sense that θ will generally change if Alice and the Warden swap their detectors. This is because the way θ is defined, it inherently measures the drop in the scaling exponent of Alice's secure payload w.r.t. the Warden's matched case when she spreads with d^{W} (case d^{W},d^{W} vs. d^{A},d^{W}). It is thus Warden-centric. In particular, θ does not tell us how much Alice loses if she estimates the secure payload w.r.t. her detector (Alice's matched case) when the Warden has a different detector (case d^{A},d^{A} vs. d^{A},d^{W}). Assessing this within our model is far from straightforward because the deflection changes both due to different ordering of images in the bag and the fact that the shifts are computed w.r.t. a different detector.

7 Setup of experiments

In this section, we first explain the setup of our experiments, the datasets, detector training, the way we estimate θ, and various other implementation details. The setup stems from our assumptions: a) the cover source and the embedding scheme are available to both the sender and the Warden, b) Alice's payload spreading strategy and the payload size embedded in a bag are known to the Warden.

7.1 Datasets and detectors

All experiments are executed on the image dataset ALASKA II [5] developed to the spatial domain as 8-bit grayscale images using the BOSSbase script [1]. We consider this dataset as more realistic than developing the RAW files as in [5] because the randomized development pipeline can produce excessively noisy images with completely flat response curves.

The dataset contains 75,000 images, which we randomly split into three disjoint parts of the same size for our experiments (Splits 1–3). Split 1 and Split 2 are used for training detectors while Split 3 was used for assessing the secure payload scaling. In all experiments, the sender uses the embedding algorithm HILL [16], which is simulated to perform on the rate–distortion bound.

Alice and Bob use one of four possible detectors: SRNet [2], the Efficient Net B4 [17, 19], SE-ResNet18 (Xu2) [12], and the LCLC classifier [6] trained on the SRM [10]. The CNNs were pre-trained on ImageNet with the binary task of steganalyzing J-UNIWARD [11] (the so-called JIN pre-training exactly as described in [3]). The refinement of all three CNNs and training of the LCLC to detect HILL was done with stego images embedded with random payloads in the following manner. Given a cover image with relative capacity $0 \le C \le \log_2 3$ bpp, the relative payload was drawn uniformly randomly from the set $C\mathcal{P} \coloneqq \{Cx : x \in \mathcal{P}\}$, a grid of payloads scaled by C where

$$\mathcal{P} = \{0.03, 0.05, 0.07, 0.1, 0.2, ..., 0.9, 1.0\}. \tag{17}$$

SRNet and SRM were trained on Split 1 while B4 and Xu2 were trained on Split 2. Each split was randomly partitioned into disjoint subsets of 22k, 1k, and 2k images for training, validation, and testing, respectively. The CNNs logit is used as the detector's response, while for the SRM based classifier, we use the projection of the feature vector on the weight vector.

The response curves were estimated using the same scaled grid of payloads above, $C\mathcal{P}$ (again, C depends on the image at hand). We use a grid of payloads scaled by the relative capacity C so that all estimated response curves have the same resolution, i.e., number of grid points. Using a fixed grid of payloads may cause some images with lower relative capacities to have response curves undefined at the largest grid points. We computed the average detector response $\hat{\varrho}(\alpha)$ and the standard deviation of detector outputs $\hat{\sigma}(\alpha)$ using 100 stego images (with different PRNG seeds in the embedding simulator) for each payload $\alpha \in C\mathcal{P}$ bpp. To draw a sample from $\mathcal{N}(\hat{\varrho}(\alpha), \hat{\sigma}^2(\alpha))$ for general α when simulating the DLS (Section 5.2), we linearly interpolate $\hat{\varrho}(\alpha)$, $\hat{\sigma}^2(\alpha)$ from the two closest grid points from $C\mathcal{P}$. The secure payload of the DLS was estimated on images from Split 3. To determine the scaling of the secure payload across a range of bag sizes, the binary search (Section 5.1) is repeated for $n \in \{2^4, 2^5, \ldots, 2^{12}\}$. The number of bags used in the binary search was $N = 500$ for all bag sizes.

In our experiments we mainly use two poolers: The simple average, which is agnostic w.r.t. the sender's payload allocation strategy

$$\pi_{\mathrm{avg}}(\mathbf{Y}) = \frac{1}{n}\sum_{i=1}^{n} d^{\mathrm{W}}(Y_i), \tag{18}$$

and a version of the correlator introduced in [20]

$$\overline{\pi}_{\mathrm{corr}}(\mathbf{Y}) = \sum_{i=1}^{n} d^{\mathrm{W}}(Y_i)\overline{s}(\alpha_i). \tag{19}$$

Here, $\overline{s}(\alpha)$ is a logistic fit to embedding shifts $s_k^{\mathrm{W}}(\alpha)$ across the testing set (2k images) of the split on which d^{W} was trained (see Section 6.3 in [20] for more details). We use this correlator instead of π_{corr} (6) because this theoretically optimal pooler is unrealizable in practice as it needs the shifts $s_i^{\mathrm{W}}(\alpha_i)$ of Warden's detector, which would necessitate Warden's access to cover images. We note that $\overline{\pi}_{\mathrm{corr}}$ is still clairvoyant because it is given Alice's payloads α_i. Using the correlator $\overline{\pi}_{\mathrm{corr}}$ is thus conservative as the secure payload determined with this pooler will likely be smaller than when the Warden needs to guess what kind of detector Alice uses for spreading and use it to estimate α_i from the images at hand.

7.2 Estimating detector mismatch parameter θ

In this section, we describe how the detector mismatch parameter θ is estimated in our experiments. For brevity, from now on we use $s_i^{\mathrm{W}}, s_i^{\mathrm{A}}$ instead of the more bulky $s_i^{\mathrm{W}}(C_i), s_i^{\mathrm{A}}(C_i)$.

Fixing a pair of Alice's and Warden's detectors d^{A},d^{W}, and a sender (greedy), we estimate θ for bag size n in the following manner. Let $\mathbf{X} = (X_1, \dots, X_n)$ be a bag of randomly selected cover images and $k(n)$ be the secure payload determined for that bag size. We first order the images by their shift at capacity w.r.t. Alice's detector from the smallest to the largest. The permutation is denoted using the subscript $[i]$ so that $s^{\mathrm{A}}_{[1]} \le \cdots \le s^{\mathrm{A}}_{[n]}$ where $s^{\mathrm{A}}_{[i]}$ is the shift in response of d^{A} for image $X_{[i]}$ embedded at capacity. Since Alice is using the greedy sender, she fully embeds the first $k(n)$ images $X_{[1]}, \dots, X_{[k(n)]}$. The shifts of Warden's detector d^{W} for $X_{[1]}, \dots, X_{[k(n)]}$ are denoted by $s^{\mathrm{W}}_{[1]}, \dots, s^{\mathrm{W}}_{[k(n)]}$ which are *unordered*. Assuming the uniformity in Eq. (13) holds, computing $\max_{i=1,\dots,k(n)} F^{\mathrm{W}}\left((s^{\mathrm{W}}_{[i]})^2\right)$ would give a biased estimate of the quantity $k(n)n^{\theta}/n$. Since the CDF F^{W} must be estimated empirically and images are sampled without replacement, we opted to instead use the 98th percentile of the list $F^{\mathrm{W}}\left((s^{\mathrm{W}}_{[i]})^2\right)$, $i = 1, \dots, k(n)$ for robustness against right-tail outliers. We further sort the $F^{\mathrm{W}}\left((s^{\mathrm{W}}_{[i]})^2\right)$ in increasing order and select the ℓ_{98}th entry, where $\ell_{98} = \mathrm{ceil}(0.98 \times k(n))$. More precisely, the 98th percentile of $F^{\mathrm{W}}\left((s^{\mathrm{W}}_{[i]})^2\right)$ for the bag $\mathbf{X}$ can be expressed as

$$p^{\neq}_{98}(n) = \min_{i=1,\dots,k(n)} \left\{ F^{\mathrm{W}}\left((s^{\mathrm{W}}_{[i]})^2\right) : |\mathcal{J}(i)| \ge 0.98 \times k(n) \right\}. \quad (20)$$

where $|\mathcal{J}(i)|$ counts the number of indices $j \in \{1, \dots, k(n)\}$ for which $F^{\mathrm{W}}\left((s^{\mathrm{W}}_{[j]})^2\right) \le F^{\mathrm{W}}\left((s^{\mathrm{W}}_{[i]})^2\right)$.To reduce estimation variance, we took the average of $p^{\neq}_{98}(n)$ over $N = 2000$ bags to be estimate of $k(n)n^{\theta}/n$ (see Eq. (13)). The same algorithm was applied to the matched case when Alice spreads with Warden's detector and hence the images are ordered by shifts w.r.t. the Warden's detector, $X_{(i)}$. Since Alice and Warden share the same ordering, we simply have

$$p^{=}_{98}(n) = F^{\mathrm{W}}\left((s^{\mathrm{W}}_{(\ell_{98})})^2\right). \quad (21)$$

Although our theory assumes the support of $F^{\mathrm{W}}\left((s^{\mathrm{W}}_{[i]})^2\right)$ increases by a factor of n^{θ}, we will estimate the power scaling as if θ depended on n, $\theta(n)$, and then average our estimates across n. For each n, we thus obtain one estimate of $\theta(n)$ as

$$\hat{\theta}(n) = \frac{1}{\log n} \log \frac{p^{\neq}_{98}(n)}{p^{=}_{98}(n)}. \quad (22)$$

Experimentally, the estimator $\hat{\theta}(n)$ is approximately constant in n which confirms the validity of our modeling assumption from (13). The final estimate of θ is obtained by averaging $\theta(n)$ over all bag sizes $n \in \{2^4, \dots, 2^{12}\}$

$$\overline{\theta} = \frac{1}{9} \sum_{k=4}^{12} \hat{\theta}(2^k). \quad (23)$$

We heuristically average across n since we are ultimately concerned with obtaining a single value for θ over the entire range of bag sizes.

8 Results

In this section, we present and discuss all the results. We begin with the case of matched detectors and then study the secure payload scaling with mismatched detectors.

Figure 1 contains eight log-log subplots of the secure payload size $P_{\delta}(n)$ as a function of n at detectability $\delta = P_{\mathrm{E}} = 0.2$ for the SLS (first row) and greedy (second row) senders. Each subplot contrasts the scaling of the secure payload when Alice spreads with four different detectors while the Warden detects with a fixed detector (column). We use π_{avg} for SLS as Warden's pooler while $\overline{\pi}_{\mathrm{corr}}$ was used for greedy. The average pooler for SLS is close to the optimal correlator because the soft output of all four detectors on covers is approximately Gaussian and the SLS by definition induces a shift in the soft output. The legend highlights Alice's choice for the detector and includes the slope of a line fit to the secure payload across all bag sizes $n = 2^4$ to 2^{12}. The matched case detector is in boldface. The bottom chart in each column contains a color-coded graphical representation of the value of the detector mismatch parameter θ estimated as explained in Section 7.2.

8.1 Matched detectors

We first comment on the cases with matched detectors. In Figure 2 left, we show the secure payload for all four cases of matched Alice's and Warden's detectors for the SLS sender extracted from the subplots in Figure 1. The scaling exponent varies from $\gamma = 0.79$ for $d^{\mathrm{A}} = d^{\mathrm{W}} = \mathrm{SRM}$ to $\gamma = 0.85$ when both actors share Xu2. It is further worth mentioning that the secure payload when both actors use SRM is about 3 times larger than when they share an SRNet.

The differences in the scaling exponent are related to the CDF of the square shifts at capacity. By the scaling theorem for matched detectors (Section 6.1), the exponent is determined by the distribution of the squared shifts $(s^{\mathrm{W}}(C))^2$ of Warden's detector in the right neighborhood of zero. In Figure 2 right, we show the log-log plots of the empirical CDF of $(s^{\mathrm{W}}(C))^2$ for all four detectors. For each detector, we compute its empirical CDF's slope β (shown in the legend) by fitting[2] a line to all points left of $(s^{\mathrm{W}}(C))^2 = 2^{-3}$. We computed estimates of the secure payload scaling exponent using the plug-in estimator $\gamma(\beta) = 1/(\beta + 1)$. These estimates from the CDF are contrasted with the observed scaling exponents γ (line fits) in Figure 2 (left). While the values of $\gamma(\beta)$ are slightly underestimated, they do correctly predict that the SRM will have the smallest exponent and Xu2 the largest. The ordering $\gamma(\beta)$ predicts between SRNet vs. SRNet (red) and B4 vs. B4 (purple) is swapped compared to γ, however this mis-estimation is likely due to the bending of B4's CDF.

To make this paper self contained, we provide a condensed intuitive explanation for why a super SRL secure payload scaling is observed. It is intuitively due to the fact that the cover source contains a non-negligible fraction of images with diminishing response to embedding as captured by the CDF (Figure 2). We make a parallel with the closest related result in prior art, which is Ker's square root law of secure payload for content-adaptive steganography [14]. Instead of spreading payload across images, the author analyzes secure payload for adaptive steganography while capturing the detectability of embedding with pixel costs (Fisher information). Most importantly, the author bans the existence of a non-negligible

[2]To the right of 2^{-3}, the CDFs are non-linear and for sufficiently large bags greedy in the matched detector case will almost exclusively embed in images for which $(s^{\mathrm{W}}(C))^2 \le 2^{-3}$.

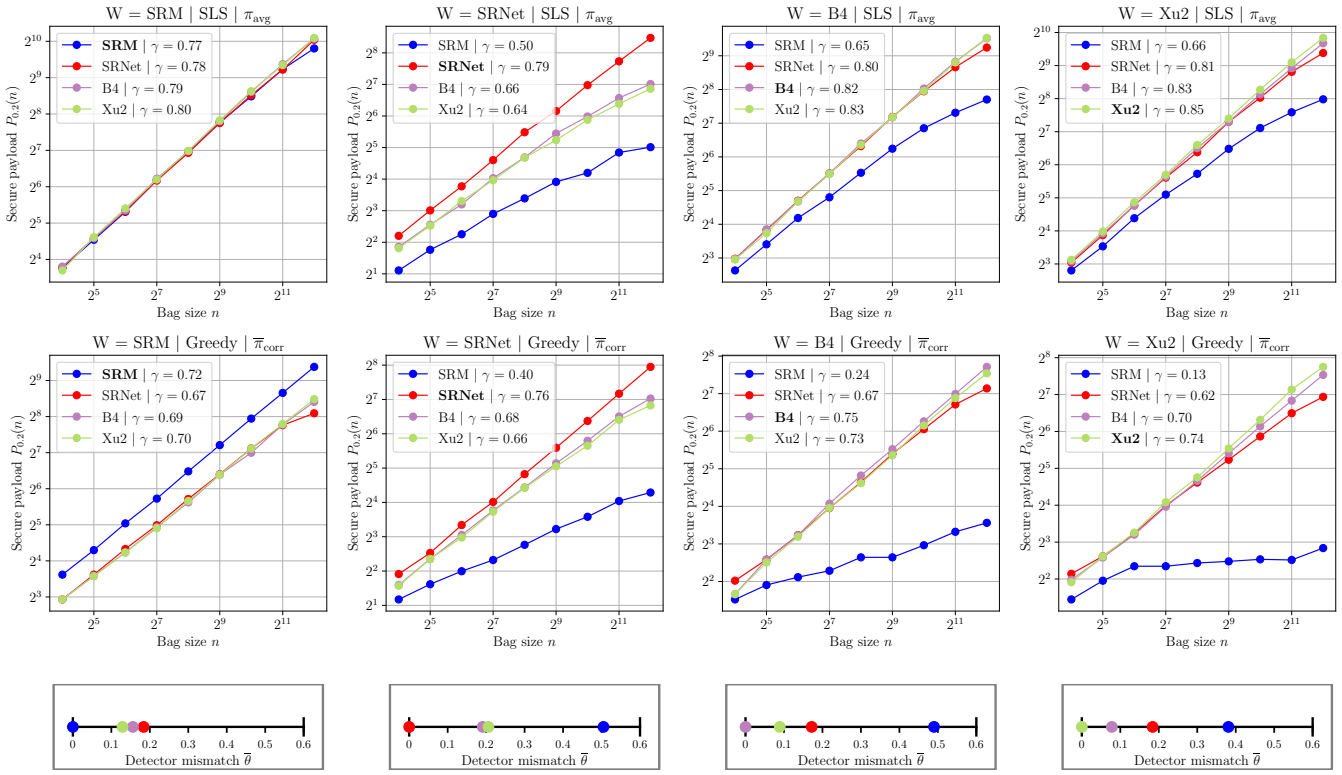


Figure 1: Log-log plot of secure payload vs. bag size n for SLS (first row) and greedy (second row) senders and different combinations of Alice's and Warden's detectors. SLS is steganalyzed with the π_{avg} pooler while the greedy with $\overline{\pi}_{\text{corr}}$. For every bag size n, the Warden's pooler achieves constant detectability $P_{\text{E}} = 0.2$. The choice of Alice's detector appears in the legend with the slope of line fit to the secure payload (matched case in boldface font). The bottom charts (third row) contain color-coded graphical representations of the estimated detector mismatch parameter θ.

source of free bits in the form of pixels' diminishing costs. While this ban is reasonable on the level of pixels or DCT coefficients, soft outputs of a detector exhibit a larger diversity. While it is unclear how to estimate the detectability on the pixel level, it is more feasible to estimate it for entire images based on their response curves or the shifts $s^{\text{W}}(C)$ for the greedy sender. The reader is referred to Section Discussion in [8] for a more detailed discussion on this topic, including the limitations of the modeling approach and what one could expect asymptotically for much larger bag sizes.

8.2 Mismatched detectors

Inspecting Figure 1, when Alice spreads with her own detector rather than Warden's (matched case, boldface), there is a loss in secure payload cn^{γ} both in terms of the multiplicative constant c and the scaling exponent γ. This loss generally depends on the type of the mismatch and the sender. It is notably smaller when the Warden has the inferior detector (SRM).

We wish to point out that while for the matched case the secure payload appears to follow a power law, for mismatched detectors the curves exhibit some "bending" in the log-log plot, indicating a more complex dependence. Except when Alice uses SRM, the secure payload size initially exhibits super SRL power scaling close to the matched case and only starts bending for the largest tested bags. This has practical consequences for Alice when she sets up her communication channel. We elaborate on this aspect in more detail in Section 9. For now, we interpret the experimental results shown in Figure 1 separately for the three cases when the inferior detector is given to the Warden, then to Alice, and when both actors use a CNN.

CNN vs. SRM: As our first type of detector mismatch, we review the case when the inferior detector (SRM) is given to the Warden (first column of subplots in Figure 1) while Alice spreads with a CNN, the superior detector. The secure payload for the SLS sender stays the same regardless of which CNN Alice uses for spreading. For the greedy sender, however, when Alice spreads with a different detector, the secure payload is smaller and so is the scaling exponent (the difference is however very small) despite the fact that she uses a better detector. This is in agreement with our analysis from Section 6.2: as long as the Warden uses a different detector (not necessarily better), the scaling exponent will be smaller by $\theta/(\beta + 1) \geq 0$. In other words, the Warden does not need to have the better detector, she only needs a *different* detector to observe a decrease in the scaling exponent.

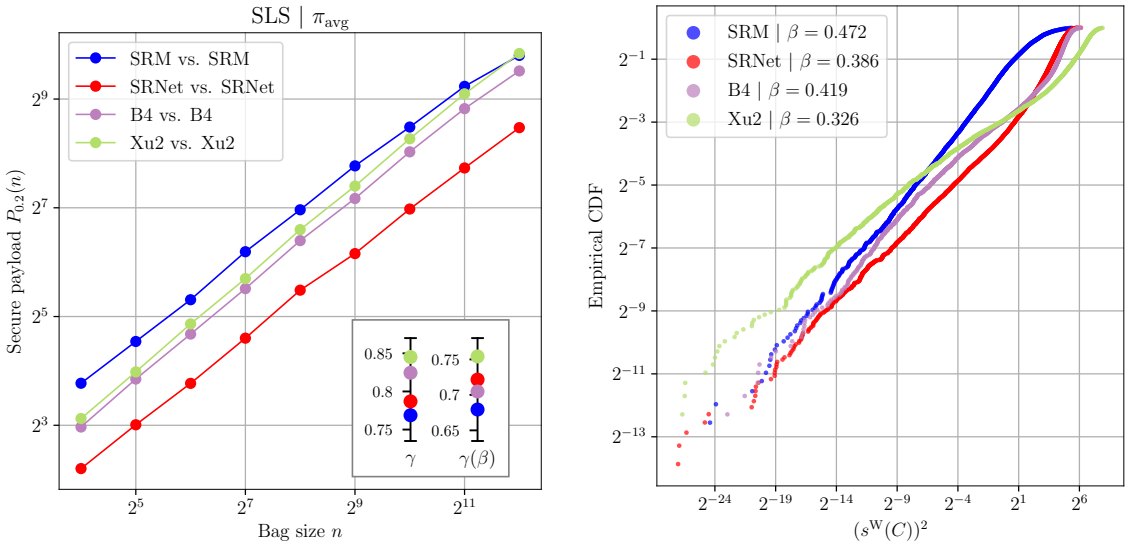


Figure 2: Left: Scaling of secure payload for the SLS sender for matched detectors. While the secure payload for the SRM classifier scales with a smaller exponent than for the CNNs, the secure payload is still larger for SRM as it is an inferior detector w.r.t. network detectors. Right: Log-log plot of the empirical CDF F^{W} of $(s^{\text{W}}(C))^2$ and the line fit with an estimated slope β for all four Warden's detectors.

SRM vs. CNN: We now inspect the case when Alice spreads with the inferior detector (SRM). To this end, we extract the corresponding curves for each sender from Figure 1 and show them in two separate plots in Figure 3. When Alice spreads with the inferior detector (SRM) the scaling exponent as well as the multiplicative constant clearly decrease w.r.t. Alice's matched case, SRM vs. SRM. For the SLS, we still observe super SRL scaling when the Warden uses B4 or Xu2 ($\gamma = 0.65$ and 0.66, respectively) but when she uses SRNet, the scaling follows the SRL with $\gamma = 0.50$. The greedy sender suffers a significantly larger loss as we observe sub-SRL scaling with exponents ranging from $\gamma = 0.13$ when the Warden is given Xu2 to $\gamma = 0.40$ when she uses SRNet. This can be intuitively explained by the aggressiveness of the greedy sender – Alice trusts a poor detector and in the end pays dearly for her choice as she would be better off spreading uniformly, in which case her secure payload would follow the SRL scaling.

CNN vs. CNN: Finally, we take a closer look at the impact of detector mismatch when both Alice and the Warden use a CNN detector (red, purple and green curves) for the SLS. Following the last three columns of subplots in Figure 1, we first comment on the SLS sender. Here, the mismatch also decreases the secure payload and the scaling exponent but the decrease is very small when the Warden uses B4 and Xu2 ($\gamma \geq 0.80$). When the Warden uses SRNet, the scaling exponent decreases from $\gamma = 0.79$ for the matched case to approximately 0.65 when Alice spreads with B4 and Xu2. If we were to advise Alice on her choice of a detector, from among the three tested CNNs, she should use an SRNet. In summary, the secure payload for a CNN vs. CNN mismatch still follows a super SRL scaling up to the largest bag size of 4096. Similar conclusions can be reached for the greedy sender.

Finally, we comment on the predictive power of the detector mismatch parameter $\overline{\theta}$ shown in the bottom row of figures in Figure 1. The color coding helps convey the fact that in all cases $\overline{\theta}$ correctly predicts which mismatch type exhibits the largest and the smallest deviation from the matched scaling.

9 Discussion

We now summarize the lessons learned from our experiments and from the theoretical study. For clarity, we begin with lessons for Alice and then for the Warden.

To Alice: Alice needs to keep in mind that a mismatch between her and Warden's detector generally decreases her secure payload cn^{γ} both in terms of the constant of proportionality c and the scaling exponent γ. And this is true even when she has the superior detector. The SLS sender implemented with feedback from a trained CNN detector seems particularly robust to detector mismatch and offers super SRL payload scaling for bag sizes up to 1000.

Alice needs to avoid using payload spreaders that are overly aggressive (e. g., greedy) especially when implemented with a poor detector (SRM). We observed that the greedy sender can exhibit significantly sub SRL payload scaling under these circumstances. Here, we reiterate that the greedy sender was included in our study because it lends itself to a tractable analysis and provides insight that would be difficult to obtain for the SLS sender.

Our analysis gives some feedback to Alice on how to setup her covert communication channel and how to use it in practice. Our analysis starts with Alice deciding on an acceptable detectability δ assuming the Warden is allowed to analyze all images she will ever send. While we use P_{E} as a measure of detectability in our

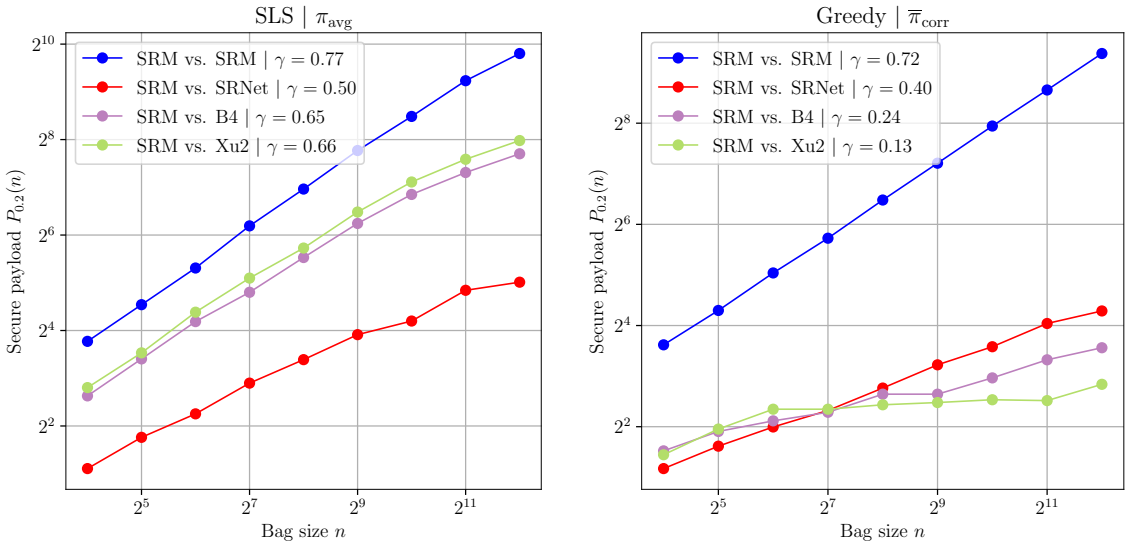


Figure 3: Secure payload scaling when Alice uses SRM and spreads with SLS (left) and greedy (right) while Warden detects with four different detectors.

experiments, she can certainly use other measures, such as weighted AUC [4] or P_{D} at a fixed P_{FA}. Then, for a given steganographic method she trains her SID d^{A} and determines the secure payload $P_\delta(n)$ as a function of n with respect to d^{A}. Considering that for mismatched detectors the secure payload slightly deviates from a power law for the largest bags and taking into account that the under mismatched detectors her secure payload will be smaller, she selects the maximal number of images, n_0, she will ever send using her steganographic channel (e. g., a social network account). Then, she selects n_0 images at random from her cover source and computes the payload chunks α_i, $i = 1, \dots, n_0$, that would be assigned to her cover images if she was to embed the entire secure payload $P_\delta(n)$ among all n_0 covers. In practice, to embed a specific message of m bits she selects the images one by one (e.g., of fixed size n_{pixels}) until the total accumulated payload becomes larger or equal to the size of the desired message—she finds the smallest N such that $\sum_{i=1}^{N} \alpha_i n_{\text{pixels}} \geq m$ and sends her message in a bag of N images. For the next message, she continues with images $N+1, \dots, n_0$. When she uses up all her covers, she will have exhausted her stego channel at detectability δ and needs to stop using the account for steganography. We wish to stress that Alice cannot guarantee detectability w.r.t. an unknown detector. When Warden's detector $d^{\text{W}}$is unknown, Alice can only estimate $P_\delta(n)$ w.r.t. d^{W} using her own simulated setup and / or be conservative with her payload size. Our analysis provides some guidance in terms of the effect of detector mismatch on the secure payload.

To the Warden: The Warden needs to keep her detector as *secret* as possible and needs to make an effort to have a *different* detector than Alice. Should Alice get a hold of Warden's detector, she can enjoy super SRL secure payload scaling with a very large exponent $\gamma \geq 0.8$ (Section 8.1). Any information about the detector leaked to Alice empowers her in terms of secure payload. Say, Alice knows that the Warden uses a CNN for detection. As discussed in Section 8.2, even when she trains a different CNN architecture for payload spreading than what the Warden used, while the scaling exponent decreases, it can still be significantly larger than 0.5 (super SRL).

Under the most realistic conditions, the Warden will not know how Alice allocates her payloads and Alice will be ignorant about Warden's pooled detector. The proper framework to study this adversarial setup is via Game Theory. In [9], the authors studied a zero-sum game with payoff function in the form of the deflection of a shared (matched) detector with payload allocation as Alice's strategy and the coefficients in a linear pooler as Warden's strategy. They showed that this Payload Allocation Game has a unique weak Nash equilibrium in pure strategies with the minimum deflection sender (MDS) as the equilibrium strategy.[3] A possible future direction is to expand on this result for the case of mismatched detectors.

10 Conclusions

With repetitive use of the stego channel, the Warden is allowed to take into consideration the entire collection of images ever communicated by Alice to conclude whether she is using steganography. We say that the Warden is executing pooled steganalysis. If Alice wishes to control her risk of being detected, she needs to know how to adjust the payload with increased number of communicated images n. In this paper, we phrase the problem of finding the secure payload that guarantees a prescribed statistical detectability by Warden's detector within the framework of batch steganography and pooled steganalysis. In particular, Alice spreads her message across images based on feedback from a detector trained to detect her steganographic scheme—she performs detector-informed batch

[3]The MDS is also Alice's optimal strategy when the Warden is omniscient [7, 20].

steganography. The main goal of our paper is to determine the scaling of the secure payload with n when the detector used by Alice for spreading and the detector used by the Warden for detection are different. Many interesting questions come to mind, such as what happens to the secure payload scaling when Alice has a detector that is superior to Warden's and when it is the other way around.

We carry out experiments with four different types of detectors—three deep convolutional neural networks and one classifier with SRM features. We also model and quantify the mismatch mathematically within a statistical model of the soft output of Warden's detector. In summary, we observed that detector mismatch generally *decreases the super square root scaling exponent* observed when Alice spreads with Warden's detector (the "matched case"). This decrease generally depends on the spreading strategy and the type of mismatch. Surprisingly, this decrease is present even when Eve has the *superior detector*, as long as it is different from Alice's. Finally, the mismatch as quantified within the adopted statistical model seems to qualitatively match the trends observed in experiments.

We have the following messages for Alice and the Warden:

> *Alice should be conservative with her spreading strategy (avoid overly aggressive spreaders) and keep in mind that the payload scaling exponent will be smaller even when the Warden has the inferior detector.*
> *Asymptotically, it is important for the Warden to have a different detector than Alice. The Warden also needs to make every effort to not reveal how her detector is built else she risks super square root law payloads being passed without detection.*

In the future, we plan to extend our study to the JPEG domain. We also plan to inspect a wider range of datasets to see if the super square root secure payload scaling is a universal phenomenon that occurs in typical image sources.

Acknowledgments

The work on this paper was supported by NSF grant No. 2324991.

References

[1] P. Bas, T. Filler, and T. Pevný. 2011. Break Our Steganographic System – the Ins and Outs of Organizing BOSS. In *Information Hiding, 13th International Conference* (Lecture Notes in Computer Science, Vol. 6958), T. Filler, T. Pevný, A. Ker, and S. Craver (Eds.). Prague, Czech Republic, 59–70.

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

[3] J. Butora, Y. Yousfi, and J. Fridrich. 2021. How to Pretrain for Steganalysis. In *The 9th ACM Workshop on Information Hiding and Multimedia Security*, D. Borghys and P. Bas (Eds.). ACM Press, Brussels, Belgium.

[4] R. Cogranne, Q. Giboulot, and P. Bas. 2019. The ALASKA Steganalysis Challenge: A first step towards Steganalysis "Into The Wild". In *The 7th ACM Workshop on Information Hiding and Multimedia Security*, R. Cogranne and L. Verdoliva (Eds.). ACM Press, Paris, France.

[5] R. Cogranne, Q. Giboulot, and P. Bas. 2020. ALASKA–2: Challenging Academic Research on Steganalysis with Realistic Images. In *IEEE International Workshop on Information Forensics and Security*. New York, NY.

[6] R. Cogranne, V. Sedighi, T. Pevný, and J. Fridrich. 2015. Is Ensemble Classifier Needed for Steganalysis in High-Dimensional Feature Spaces?. In *IEEE International Workshop on Information Forensics and Security*. Rome, Italy.

[7] E. Dworetzky and J. Fridrich. 2023. Explaining the Bag Gain in Batch Steganography. *IEEE Transactions on Information Forensics and Security* 18 (2023), 3031–3043.

[8] E. Dworetzky, E. Kaziakhmedov, and J. Fridrich. 2024. Secure Payload Scaling for Source Adaptive Payload Allocation, A. Alattar, N. D. Memon, and G. Sharma (Eds.). San Francisco, CA.

[9] E. Dworetzky, B. Tondi, M. Barni, and J. Fridrich. 2025. Payload Allocation in Batch Steganography: A Game-Theoretic Perspective. *IEEE Signal Processing Letters* (2025). In preparation.

[10] J. Fridrich and J. Kodovský. 2011. Rich Models for Steganalysis of Digital Images. *IEEE Transactions on Information Forensics and Security* 7, 3 (June 2011), 868–882.

[11] V. Holub, J. Fridrich, and T. Denemark. 2014. 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).

[12] J. Hu, L. Shen, and G. Sun. 2018. Squeeze-and-excitation networks. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 7132–7141.

[13] A. D. Ker. 2006. Batch Steganography and Pooled Steganalysis. In *Information Hiding, 8th International Workshop* (Lecture Notes in Computer Science, Vol. 4437), J. L. Camenisch, C. S. Collberg, N. F. Johnson, and P. Sallee (Eds.). Springer-Verlag, New York, Alexandria, VA, 265–281.

[14] A. D. Ker. 2017. On the Relationship Between Embedding Costs and Steganographic Capacity. In *The 5th ACM Workshop on Information Hiding and Multimedia Security*, M. Stamm, M. Kirchner, and S. Voloshynovskiy (Eds.). ACM Press, Philadelphia, PA.

[15] A. D. Ker. 2017. The Square Root Law of Steganography. In *The 5th ACM Workshop on Information Hiding and Multimedia Security*, M. Stamm, M. Kirchner, and S. Voloshynovskiy (Eds.). ACM Press, Philadelphia, PA.

[16] B. Li, M. Wang, and J. Huang. 2014. A new cost function for spatial image steganography. In *Proceedings IEEE, International Conference on Image Processing, ICIP*. Paris, France.

[17] T. Mingxing and V. L. Quoc. 2019. EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. In *Proceedings of the 36th International Conference on Machine Learning, ICML*, Vol. 97. 6105–6114.

[18] V. Sedighi, R. Cogranne, and J. Fridrich. 2017. Practical Strategies for Content-Adaptive Batch Steganography and Pooled Steganalysis. In *Proceedings IEEE, International Conference on Acoustics, Speech, and Signal Processing*.

[19] Y. Yousfi, J. Butora, E. Khvedchenya, and J. Fridrich. 2020. ImageNet Pre-trained CNNs for JPEG Steganalysis. In *IEEE International Workshop on Information Forensics and Security*. New York, NY.

[20] Y. Yousfi, E. Dworetzky, and J. Fridrich. 2022. Detector-informed Batch Steganography and Pooled Steganalysis. In *The 10th ACM Workshop on Information Hiding and Multimedia Security*, J. Butora, C. Veilhauer, and B. Tondi (Eds.). ACM Press, Santa Barbara, CA.