Improving Steganographic Security with Source Biasing

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

Edgar Kaziakhmedov
Department of ECE
Binghamton University
Binghamton, NY, USA
ekaziak1@binghamton.edu

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

ABSTRACT

By selecting covers in which steganographic embedding is harder to detect, the steganographer can decrease the chances of being caught by the Warden. On the other hand, sampling from the cover source with a bias is detectable on its own. In this paper, we study this trade-off theoretically within a simple source model. Our analysis predicts the existence of "bias security gain" when the sender selects the sampling bias optimally. Sampling with a bias initially morphs the ROC of Warden's detector to be asymmetrical, lowering the true positive rate for small false alarm rates. We provide a theorem, analogous to the square root law, for the joint critical rates of sampling bias and payload that achieve asymptotically constant detectability. Our analysis is verified experimentally.

CCS CONCEPTS

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

KEYWORDS

steganalysis, steganography, source, source biasing, cover selection, bias gain

ACM Reference Format:
Eli Dworetzky, Edgar Kaziakhmedov, and Jessica Fridrich. 2024. Improving Steganographic Security with Source Biasing. In *Proceedings of the 2024 ACM Workshop on Information Hiding and Multimedia Security (IH&MMSec '24), June 24–26, 2024, Baiona, Spain.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3658664.3659646

1 INTRODUCTION

In the classical setup of steganography, there exists a source of cover objects from which Alice and Bob draw to communicate covertly. The Warden tests whether the objects they exchange follow a known distribution [6]. As is the case in most work on digital media steganography by cover modification, Alice's stego method is assumed to be imperfectly secure as she is unable to exactly preserve the cover distribution due to the sheer complexity of natural images and the diversity of development from RAW capture and subsequent post processing [2]. Thus, in practice Alice settles for disturbing the cover source model as little as possible to comply with some detectability or payload requirement.

The majority of research in this field has focused on the steganographic scheme and the allocation of payload across images [13, 14, 16, 19, 22]. One seemingly simple measure for Alice to improve security is to draw covers from her source with a bias and prefer selecting complex or noisy images in which the embedding changes are harder to detect. This should indeed improve the security with respect to stego detectors. On the other hand, sampling from the cover source with a bias will change the distribution as well and Alice will become vulnerable to a source detector. With this in mind, it is natural to ask if there is an optimal bias, neither too big nor too small, that best avoids being caught by either detector.

Most prior art in cover selection for steganography deals with the problem of selecting a subset of images from a given dataset to create a new cover source that is harder to steganalyze [18, 20–23, 25, 26]. Such work ignores the fact that sending only stego-friendly images is suspicious on its own. As argued in the next section, a cover selection algorithm should be considered a part of the embedding scheme, which would allow the Warden to detect whether Alice selects covers from her source with a bias. In [21], the authors consider the impact of cover selection on the source by enforcing the maximum mean discrepancy [11] between the selected subset of covers and a randomly selected subset to be "typical." This work is experimental and heuristic in nature, and it is not clear how far one can bias to enjoy a security benefit, how big this benefit is, and what form it has.

The novelty of our work is that we approach cover selection from a theoretical point of view within a source model. The Warden's hypothesis test considers the effect of embedding and source biasing *jointly* within the context of batch steganography and pooled steganalysis [14]. The most closely related prior art is [10], where the authors study how Alice should choose her cover from multiple sources within a game theoretic setup when the Warden makes a decision based on a single image.

After motivating our approach in the next section, Section 3 introduces the key concepts, including the detector response curve, sender's embedding strategy and biasing, and the modeling assumptions that facilitate our theoretical analysis. In Section 4, we formulate Warden's hypothesis test as a joint detection of steganography and source biasing and derive a closed-form expression for the receiver operating characteristic (ROC) of Warden's optimal detector. This allows us to obtain the main results, which include morphing of Warden's ROC with increased bias, the sender's bias gain, and an asymptotic result when an infinite number of images are communicated. To validate our analysis, we report on experiments with digital images and deep learning detectors in Section 5; the results confirm the existence of the bias gain, the theoretically predicted behavior of the Warden's detector, and its dependence

IH&MMSec '24, June 24–26, 2024, Baiona, Spain

ACM ISBN 979-8-4007-0637-0/24/06.
https://doi.org/10.1145/3658664.3659646

on the parameters of the steganographic channel. The paper is concluded in Section 6.

2 MOTIVATION

To motivate the setup within which we carry out our analysis and experiments, in this section we discuss the concept of cover biasing and its detection. While in this paper we work with covers in the form of digital images represented in the spatial domain, we expect the methodology and our conclusions to apply to JPEG images and also for steganography in other digital media.

Intuitively, creating a slight preference for stego-friendly images might go a long way in improving security without raising suspicion. Depending on how Alice selects her covers, one could argue that detecting a bias in sampling from the cover source is not even possible or at least not possible to do accurately. For example, when Alice uses her images as covers, the type of images taken can vary greatly depending on numerous unpredictable circumstances not related to steganography, such as travel, world events, changes in photographic equipment and editing software, her own evolution as a photographer, etc. Moreover, modeling sources is difficult given the typical dimensionality of images, the wide spectrum of development pipeline parameters and post-processing options, and the number of images required for estimating such models.

Additionally, we need to ask whether observing a source of images that is more friendly to steganography is a sufficient argument for steganography actually being used. The answer depends on the degree to which Kerckhoffs' principle is employed, i.e., what kind of information is available to the Warden. In Natural Steganography (NS) [1], for example, Alice uses her knowledge of the development pipeline to embed into low ISO images, making stego images look like covers taken at higher ISO. Although NS does not involve cover selection, the Warden would tend to see higher ISO images communicated by Alice; but is this suspicious on its own? The security of NS comes from the Warden not knowing the true ISO distribution of Alice's camera—a higher ISO distribution would not be justifiably suspicious. However, if the Warden knew that, e.g., Alice uses automatic exposure settings, the true ISO distribution would be known and NS would be highly detectable. Note that this evidence of stego-friendly images or change in source is indirect as it is often not specific to any given embedding method. However, in steganalysis there are many examples of detectors that rely on such types of evidence, such as compatibility attacks. In the JPEG compatibility attack [9], a block of pixels that could not have arisen from any block of quantized DCTs by decompression is taken as evidence of steganography even though the incompatibility could be due to other kinds of manipulations, such as image retouching, removing dust specks, recoloring one's eyes, etc.

Ultimately, we consider a change in how the cover source is sampled to be a part of the embedding algorithm. Thus, we grant the Warden knowledge of the cover source so that testing for biased sampling is possible,[1] in accordance with the information theoretic definition of steganographic security [6]. In fact, not giving the Warden any information about Alice's original cover source leads to a degenerate situation because quite significant portions of popular image datasets, such as ALASKA II [7], contain images in which steganography is virtually undetectable using state-of-the-art detectors, such as images taken with very high ISO setting.

Regarding the source modeling and its biasing, one could consider models in some steganalysis feature space as was the approach taken in [21]. However, high-dimensional models are difficult to use for an analytic study, which is our main goal. Instead, we model the source through soft (scalar) outputs of a steganography detector as in [8]. While this makes the model dependent on a given steganographic method and detector, as will be seen below this approach facilitates tractable analysis with interpretable closed-form solutions that provide insight into the trade-off between the gain of biased sampling and the added vulnerability to a source detector. It also intuitively makes sense for Alice to prefer selecting covers for which the embedding does not affect the soft output of the detector. In the extreme case when she only embeds images that do not respond to embedding, a warden equipped with the same detector would not be able to detect steganography. Finally and most importantly, using a steganography detector for detecting both the use of steganography and a change in sampling covers allows us to consider the problem of detecting steganography and source *jointly* through a single hypothesis formulation instead of having to consider a stego detector and a separate source detector and then deal with the difficult problem of fusing their outputs.

3 MODELING FRAMEWORK

In this section, we lay out the basic assumptions and statistical models that will facilitate our analysis of the effect of source biasing on security. We assume the Warden has a single-image steganography detector (SID), which is a mapping $d : \mathcal{X} \to \mathbb{R}$ that assigns to each image a scalar referred to as the soft output (or response) of the detector. To motivate and justify our models, we will assume that sampling from $\mathcal{X}$ is a two-stage process. First, Alice selects a "scene" and then acquires it with an imaging sensor. Conceivably, she could take as many acquisitions of the same scene as she wishes. These acquisitions will slightly differ due to, e.g., photonic noise and electronic noise, and will be concentrated around a noise-free version of the scene. We call this distribution conditioned on the scene the acquisition oracle. To avoid the complexity of modeling the oracle, we model the soft outputs of Warden's detector as in [8].

3.1 Modeling soft output of Warden's detector

Suppose that Alice has n cover images X_i, $i = 1, \dots, n$, which are independent samples from n acquisition oracles. Denoting a Gaussian random variable with mean μ and variance σ^2 as $\mathcal{N}(\mu, \sigma^2)$, we assume that

$$d(X_i) \sim \mathcal{N}(\mu_i, \sigma_i^2), \tag{1}$$

where μ_i and σ_i^2 are the expected value and variance of d on cover images X_i generated by the acquisition oracle for the ith scene. Since the acquisitions are concentrated on a small subset of $\mathcal{X}$ and since differentiable non-linear functions[2] are approximately linear on sufficiently 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. See [24] and [8] for further discussion.

[1] And we assume a source change is due to steganography.

[2] Modern SIDs are often neural networks with differentiable structure.

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), $d(X_i(\alpha_i))$, is also Gaussian[3]

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

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 [19] about the *global* distribution of detector response which is not satisfied for modern steganalyzers built with machine learning (see Sec. 3.2 in [24]).

3.2 Detector response curve

For a ternary embedding algorithm and cover acquisition X_i with maximal embedding capacity $C_i \le \log_2 3$, the response curve (RC) of detector d is the function $\varrho_i : [0, C_i] \to \mathbb{R}$ defined by

$$\varrho_i(\alpha) = \mathbb{E}[d(X_i(\alpha))|X_i]. \tag{3}$$

with $\alpha \in [0, C_i]$. In other words, $\varrho_i(\alpha)$ is the expected value of the response $d(X_i)$ when embedding X_i with random messages and stego keys for a given α and a fixed cover X_i.

Since the detector is trained to be sensitive to embedding changes but not acquisition noise, we assume the expected increase in detector response is uniform across all possible acquisitions X_i of the ith acquisition oracle

$$\varrho_i(\alpha) - \varrho_i(0) = s_i(\alpha). \tag{4}$$

This assumption allows us to compute the expected shift $s_i(\alpha)$ from a specific cover image, which simplifies analysis and practical implementations.

3.3 Source model and biasing

As explained above, our source model will be defined through a detector's response curves. First, we adopt the simplifying assumption that response curves are linear

$$\varrho_i(\alpha) - \varrho_i(0) = b_i\alpha, \tag{5}$$

where $b_i \in [0, \infty)$ is the slope of the linear response curve. Even though the response curves of typical detectors built with machine learning are not linear, they are approximately linear when $\varrho_i(\alpha) - \varrho_i(0)$ is small (see, e.g., Figure 3 in [24]).

Within the class of linear RCs (5), our source model boils down to modeling the slopes. In particular, we assume that the slopes b_i follow a two-valued distribution controlled by parameter $p \in (0, 1)$:

$$b_i \sim \mathcal{B}(p) = \begin{cases} \varepsilon & \text{with probability } p \\ 1 & \text{with probability } 1-p. \end{cases} \tag{6}$$

We assume that the cover source contains only two types of images – those with smooth content where steganography is easily detectable (slope $b_i = 1$) and images with complex textures or noisy images with $b_i = \varepsilon \ll 1$. While this may appear as a rather drastic simplification, it allows us to analyze steganography with source biasing via closed-form expressions, make specific predictions on the impact of biasing, and verify these findings in practice.

[3]The random variable $X_i(\alpha_i)$ is generated by sampling X_i from the oracle and embedding a random message with a random stego key.

In the context of our source model (6), source biasing simply involves sampling images so that the distribution of slopes follows the biased distribution $b_i \sim \mathcal{B}(q)$ given the biasing parameter $q \in (0, 1)$. Clearly, Alice should choose $q \ge p$ so that she is more likely to embed difficult-to-steganalyze images.

3.4 Alice's embedding strategy

The proper framework for analyzing the effect of source biasing is the paradigm of batch steganography and pooled steganalysis [14]. Indeed, it is not possible to detect changes in cover source sampling based on a single image. The Warden needs to analyze multiple images. Alice, on the contrary is free to spread her payload across multiple images as well.

Suppose that Alice samples a bag (or collection) of n images $X_1, \dots, X_n$ with response curve slopes $b_1, \dots, b_n$, respectively. Alice wants to communicate some rate $r(n)$ measured in bpp. For simplicity, assume all images have relative embedding capacity $C_i = \log_2 3$ bpp. Alice embeds α_i bpp to each image X_i subject to her payload constraint $\sum_{i=1}^n \alpha_i = r(n) \cdot n$ and $\alpha_i \in [0, \log_2 3]$, $\forall i$.

Let $K = |\{i : b_i = \varepsilon\}|$ be the number of images whose RC slopes are ε, i.e., $K \sim \text{Binom}(q, n)$ is a binomial random variable due to the source model (6). For our theoretical analysis in the next section, we consider a *bivalued spreading strategy* that assigns $\alpha_i = \alpha_\varepsilon$ to all images with $b_i = \varepsilon$ and $\alpha_i = \alpha_1$ to images with $b_i = 1$. The pair $\{\alpha_\varepsilon, \alpha_1\}$ is determined to satisfy the payload constraint $r(n)n = K\alpha_\varepsilon + (n-K)\alpha_1$. A special type of a bivalued strategy is the uniform sender, which assigns $\alpha_i = r(n)$ to all images X_i is with $\alpha_\varepsilon = \alpha_1 = r(n)$. The greedy sender prefers to embed the payload in ε images. Formally, if $r(n)n \le K \log_2 3$, the greedy sender selects $\alpha_\varepsilon = r(n)n/K$ and $\alpha_1 = 0$. When $r(n)n > K \log_2 3$, $\alpha_\varepsilon = \log_2 3$ and $\alpha_1 = (r(n)n - K \log_2 3)/(n-K)$.

4 EFFECT OF SOURCE BIASING ON SECURITY

In this section, we formulate Warden's hypothesis test and derive the most powerful pooled detector. Then, we analyze and discuss how its performance is affected by source biasing.

4.1 Optimal pooler

In practice, Alice and the Warden will use their own SIDs for detector-informed spreading and pooled steganalysis, respectively. To avoid complicating the analysis with a mismatch between their SIDs, we assume in this section that Alice uses Warden's SID d. When verifying the analysis in practice in Section 5, we carry out experiments with both SIDs matched as well as mismatched. We will also assume that the parameter p is known both to Alice and the Warden and the Warden knows her spreading strategy, biasing parameter q, and rate $r(n)$. Moreover, we assume that the embedding does not change the response curve slope b_i, which is a reasonable assumption for small rates $r(n)$.

Referring to our model of Warden's detector soft output on cover (1) and stego (2) images, to avoid modeling the distribution of the variances σ_i^2 across scenes and the oracle itself, we assume that all variances are the same across scenes $\sigma_i^2 = 1$ for all i. Furthermore,

we grant the Warden the knowledge of μ_i, which greatly simplifies the problem as the Warden now faces the following simple hypothesis testing problem:

$$\begin{aligned} \mathcal{H}_0: &\quad b_i \sim \mathcal{B}(p), \quad y_i \sim \mathcal{N}(0,1) &&\text{for all } i \\ \mathcal{H}_1: &\quad b_i \sim \mathcal{B}(q), \quad y_i \sim \mathcal{N}(b_i\alpha_i(K),1) &&\text{for all } i \end{aligned} \tag{7}$$

The most powerful detector for (7) is the likelihood ratio test (LRT)

$$L(\mathbf{b},\mathbf{y}) = \sum_{i=1}^{n} y_i b_i \alpha_i(K) - \frac{1}{2}\sum_{i=1}^{n} b_i^2\alpha_i^2(K) + K\log\frac{q}{p} + (n-K)\log\frac{1-q}{1-p}. \tag{8}$$

We introduce two functions to help condense some of the expressions. The first is the steganographic deflection coefficient of a bag (the quadratic term in the expectation of $L(\mathbf{b},\mathbf{y})$ under $\mathcal{H}_1$) conditioned on the event $K=k$:

$$\Delta^2(k) = \frac{1}{2}\sum_{i=1}^{n} b_i^2\alpha_i^2(k). \tag{9}$$

Next is the likelihood ratio between two binomial random variables as a function of k

$$L_{\text{binom}}(k) = k\log\frac{q}{p} + (n-k)\log\frac{1-q}{1-p}. \tag{10}$$

The false alarm and correct detection probabilities of $L(\mathbf{b},\mathbf{y})$ are

$$P_{\text{FA}}(x) = \sum_{k=0}^{n}\binom{n}{k}p^k(1-p)^{n-k}Q\left(\frac{x-E_0(k)}{\sqrt{V(k)}}\right) \tag{11}$$

$$P_{\text{D}}(x) = \sum_{k=0}^{n}\binom{n}{k}q^k(1-q)^{n-k}Q\left(\frac{x-E_1(k)}{\sqrt{V(k)}}\right), \tag{12}$$

owing to the law of total probability where

$$E_0(k) = L_{\text{binom}}(k) - \Delta^2(k), \tag{13}$$

$$E_1(k) = L_{\text{binom}}(k) + \Delta^2(k), \tag{14}$$

$$V(k) = 2\Delta^2(k). \tag{15}$$

4.2 Morphing of ROC

Figure 1 shows the ROCs of Warden's detector Eqs. (11)–(12) for $p = 0.4$, the greedy sender, and relative payload $r = r(n) = 1$. Each ROC corresponds to a different q. The top figure has $\varepsilon = 0.01$ and $n = 4$ while the bottom figure has $\varepsilon = 0.05$ and $n = 8$. With increasing bias $q - p$, the ROCs for both cases "morph" in a similar manner. The Warden loses on detection power for small false alarms and gains for large false alarms. Since for practical steganalysis it is important to keep false alarms low, one can say that the biasing initially helps the steganographer. For sufficiently large bias, the biasing eventually decreases the security. Intuitively, the negative effect of biasing will be perceived sooner for larger n as the change in the source becomes easier to detect. The jagged character of the ROCs for the largest bias is due to $L_{\text{binom}}(k)$ (10) and it is more pronounced for smaller bags n.

For the two examples shown in the figure, the parameters ε, r, and p were chosen to approximately match the experimental setup with "binarized ALASKA II" in Section 5 so that one can contrast the experiments with the theoretically derived morphing of the ROC.

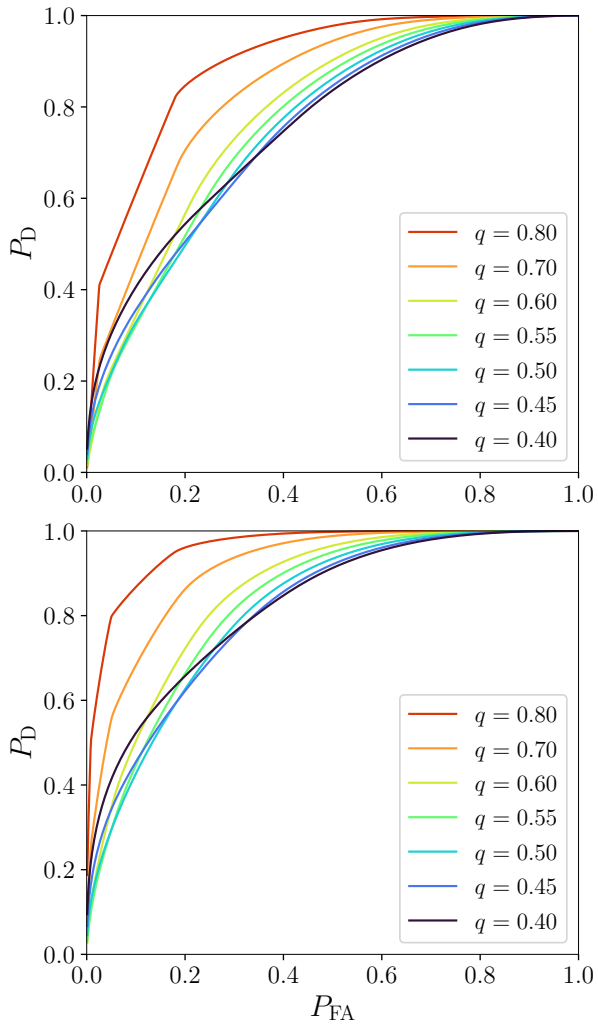

Figure 1: Examples of how ROCs for Warden's LRT pooler (8) morph with bias q. Top: $\varepsilon = 0.01$, $n = 4$. Bottom: $\varepsilon = 0.05$, $n = 8$. Both: greedy sender, $p = 0.4$ and $r = 1$.

We wish to stress that the ROC morphing is a general phenomenon that can be observed for a wide range of the parameters ε, r, p, and various bag sizes n. The morphing also motivates us to define the bias gain at a selected P_{FA} as the largest drop in P_{D} of Warden's optimal pooler the steganographer can achieve by selecting the bias optimally:

$$\gamma_{\text{bias}}(P_{\text{FA}}) = P_{\text{D}}(P_{\text{FA}},p) - \min_{p\le q} P_{\text{D}}(P_{\text{FA}},q) \tag{16}$$

where $P_{\text{D}}(P_{\text{FA}},q)$ is the power of Warden's optimal pooler at false alarm P_{FA} when the steganographer samples from the source with $\mathcal{B}(q)$. Note that the bias gain is also a function of the steganographic method, payload size, the spreading strategy, ε, and Warden's SID. Equipped with this measure, we plot two more figures to better convey the nature of the bias gain.

Figure 2 shows the true positive rate P_{D} for the false alarm fixed to $P_{\text{FA}} = 0.01$ as a continuous function of q across four different bag

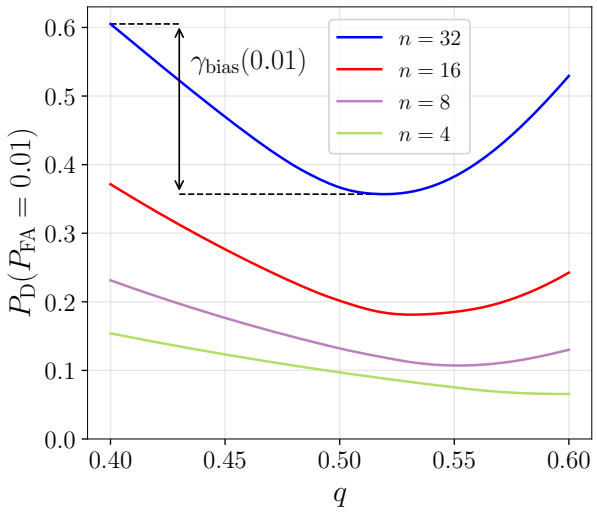

Figure 2: True positive rate P_{D} at $P_{\text{FA}} = 0.01$ versus q for the greedy sender and four bag sizes $n \in \{4, 8, 16, 32\}$, $r = 1$, $\varepsilon = 0.01$, $p = 0.4$, $q \in [0.4, 0.6]$. As n increases, the bias gain $\gamma_{\text{bias}}(0.01)$ increases and the optimal q decreases.

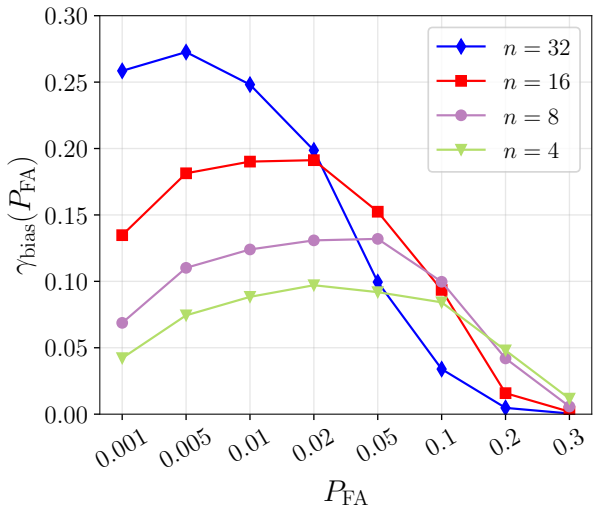

Figure 3: Bias gain $\gamma_{\text{bias}}(P_{\text{FA}})$ for four different bag sizes for the greedy sender, $r = 1$, $\varepsilon = 0.01$, $p = 0.4$. The figure demonstrates the security benefit (i.e. decrease in P_{D}) of biasing for small false alarm rates.

sizes. For maximal bias gain (16), Alice should bias more strongly for smaller bag sizes n than for larger ones, which is intuitive because it is harder to detect the source change from smaller bags. Also, the gain is larger for larger bags ($n = 32$). These observations appear universal across other choices of the parameters ε, r, and p.

Figure 3 is another way to show how the ROCs morph with bias. In this figure, we plot $\gamma_{\text{bias}}(P_{\text{FA}})$ for a range of fixed false alarms P_{FA} for four different bag sizes and the greedy sender with $r = 1$ and $\varepsilon = 0.01$. For each P_{FA} and n, the value of q differs and is chosen to minimize $P_{\text{D}}(P_{\text{FA}}, q)$ as in Eq. (16).

4.3 Biasing for large n

In this section, we study the effect of biasing on security in the asymptotic case when the number of communicated images n approaches infinity. An already established general result is the square root law (SRL), which states that the critical rate for Alice to achieve constant detectability is $r(n) \propto 1/\sqrt{n}$. Larger rates allow the Warden to build an arbitrarily accurate detector while smaller rates will lead to asymptotically perfect security. For a precise formulation of this result, the reader is referred to, e.g. [15]. When Alice additionally samples from the cover source with a bias, she needs to asymptotically adjust her bias as well to prevent becoming asymptotically perfectly detectable. The following Asymptotic Biasing Theorem (ABT), which is proved in the appendix, summarizes these findings for bivalued payload spreading strategies (recall Section 3.4).

Theorem 4.1. *[Asymptotic Biasing Theorem] For a bivalued cover source (6) and a bivalued spreading strategy, when Alice adjusts her rate $r(n)$ and bias $q(n)$ to follow critical rates in the sense that these two limits exist $c_r = \lim_{n\to\infty} r^2(n)n$, $c_p = \lim_{n\to\infty}(q(n)-p)\sqrt{n}$, the following holds regarding the security of the steganographic channel:*

When $c_r = 0$ and $c_p = 0$, the communication is asymptotically perfectly secure.

When $c_r = \infty$ or $c_p = \infty$, the communication is asymptotically perfectly detectable.

When Alice uses the greedy or uniform senders and $c_r < \infty$ and $c_p < \infty$, the performance of Warden's most powerful detector is described with a Gaussian ROC with deflection $d^2_{\text{greedy}} = \frac{c_r\varepsilon^2}{p} + \frac{c_p^2}{p(1-p)}$ and $d^2_{\text{uniform}} = (p\varepsilon^2 + 1 - p)c_r + \frac{c_p^2}{p(1-p)}$.

The theorem essentially states that in order to prevent asymptotically perfect detection due to source biasing, the sender needs to scale the bias $q(n) - p$ by $1/\sqrt{n}$, which is analogous to scaling rate $r(n)$ by $1/\sqrt{n}$ to achieve constant detectability per the SRL. When both the payload and the bias are adjusted at their critical rates, the ROC becomes symmetrical and Gaussian. In the third part of the ABT, observe that the first terms of d^2_{greedy} and d^2_{uniform} only depend on Alice's rate while the second terms only depend on her bias and are strictly monotone in c_p. Thus, when communicating a square root rate $r(n) \propto 1/\sqrt{n}$, Alice should avoid asymptotically biasing altogether in the sense that $c_p = 0$; she will only become more detectable if $c_p > 0$. This theorem is validated experimentally on real images with poolers built with machine learning in Section 5.6.

We note that the third part of the ABT is a special case only for the greedy and uniform senders since our proof directly computes the deflection. We conjecture that this part generalizes to all bivalued senders. This extension would require adopting some strategy on how α_ε and α_1 change with n.

5 EXPERIMENTS

The purpose of the experiments in this section is to find out whether the effect of source biasing on Warden's detector observed in our models transfers to the real world. In particular, we are interested in verifying the morphing of the ROC, the qualitative trends of the bias gain and optimal bias as a function of bag size, and the asymptotic

scaling theorem. We present two sets of experiments. In Section 5.5, we verify the theoretical results from Section 4 on an image dataset that approximately contains only easy and difficult-to-steganalyze images. In Section 5.6, we investigate the security effects of biasing on a realistic dataset which necessitates a more general method of biasing. In all experiments, Alice uses the embedding algorithm HILL [17] simulated to perform on the rate–distortion bound.

5.1 Datasets

All experiments were executed on ALASKA II [7] developed as in [7] without the final JPEG compression step. This dataset contains 75,000 images, which we randomly split into three disjoint parts for our experiments: Split A, Split W, and Split BSPS.[4] Split A consists of 25k images reserved for Alice, and Split W consists of 25k images reserved for the Warden. We describe how Alice and the Warden use Split A and W for training SIDs in Section 5.4. The remaining 25k images in Split BSPS are further divided into Split P (one third of Split BSPS) used to form bags for training the Warden's pooler while the remaining two thirds, named TST, are used for evaluation.

5.2 Estimating response curves

In practice, we estimate image X_i's response curve (3), ϱ_i, and its slope b_i for a SID d as follows. For each $\alpha \in \mathcal{P}$ in the grid of payloads

$$\mathcal{P} = \{0.05, 0.1, 0.2, \dots, 1.4, 1.5\}, \tag{17}$$

we compute $\varrho_i(\alpha)$ by averaging the detector response of 100 embeddings with random stego keys. For values of α between the grid points, $\varrho_i(\alpha)$ is computed by linear interpolation. Given that X_i has embedding capacity $C_i \le \log_2 3$, we estimate the slope as $b_i = (\varrho_i(C_i) - \varrho_i(0))/C_i$.

5.3 Greedy sender

Since the images in our evaluation datasets have a continuous range of slopes, we modified the greedy sender (Section 3.4) in the following way. Given a set of n images, the sender (Alice) uses a SID to estimate their response curves, orders them from the smallest slope to the largest, and then embeds the images one by one at their capacity until the required payload is embedded. The last image may be embedded only partially.

5.4 Single-image detectors and poolers

We wish to point out that both the sender and the Warden use SIDs to achieve their goals. Alice's detector will be denoted as d^{A}, while d^{W} will be used for Warden's SID. For our experiments, we trained two different detectors that will be given to Alice and the Warden: two versions of SRNet [3] trained on Split A and Split W denoted by SRNetA and SRNetW, respectively. Both were pre-trained on ImageNet with the binary task of steganalyzing J-UNIWARD [12] (the so-called JIN pre-training exactly as described in [5]). The refinement of all detectors to detect HILL was done with stego images embedded with relative payloads randomly drawn from $\mathcal{P}$. Each split was randomly partitioned into disjoint subsets of 22k, 1k, and 2k images for training, validation, and testing, respectively. The CNNs logit was used as the detector's soft response d.

[4] BSPS stands for "batch steganography & pooled steganalysis".

For fixed bag size n, the Warden's pooling function was implemented as a random forest (RF) [4] (Python's package `scikit-learn`) on a $2n + 2$-dimensional feature vector[5] extracted from all images in the bag

$$(d^{\mathrm{W}}(X_1), \dots, d^{\mathrm{W}}(X_n), b_1, \dots b_n, n_\varepsilon/n, \pi_{\mathrm{CORR}}), \tag{18}$$

where $d^{\mathrm{W}}(X_i)$ is the soft output of Warden's detector and b_i is the response curve slope of the ith image, n_ε is the number of ε-type images in the bag, and

$$\pi_{\mathrm{CORR}} = \sum_{i=1}^{n} \alpha_i d^{\mathrm{W}}(X_i) \tag{19}$$

is the correlation of Warden's SID soft outputs with payloads that might reside in the images. We note that the slopes as well as the payloads are computed from the cover versions of X_i, which makes Warden's pooler clairvoyant. We do this for the reasons of excessive computational complexity as estimating the response curves from multiple embeddings from each image in the bag and estimating the payloads is very computationally expensive. One can also think of this clairvoyant pooler as the worst case scenario for Alice. Moreover, there is evidence (see Section 7.2 in [24]) that the performance of the Warden's pooler is largely unaffected when the Warden estimates the payloads and the response curve slopes from the images at hand.

We remind the reader that we use Split P to generate 5k cover and 5k stego bags for training the pooler. The features are normalized to zero sample mean and unit variance. A grid search was used to estimate the RF hyper parameters, which include the number of estimators and the maximum depth. Every step of the grid search is Monte-Carlo cross-validated with 5 iterations with fixed train-test split ratio of 2 : 1.

5.5 Experiments on binarized ALASKA II

The experiments in this section are executed on a subset of Split BSPS that we call "binarized ALASKA II" which is obtained by rejection sampling to make the dataset closer to the binomial source model considered in our analysis in Sections 3–4. To stay within the spirit of the model, this dataset consists of two groups of images: easy-to-steganalyze images with steep response curves (M-type images) and hard-to-steganalyze images with almost flat response curves (ε-type images). This grouping was based on the slopes of their response curves. For an ε-type image X, we requested that $10^{-4} \le b_X \le 0.08$, while for M-type images $0.8 \le b_X \le 3.2$. The groups contained $N_\varepsilon = 3204$ and $N_M = 7418$ images with average slopes 0.0213 and 2.001. As mentioned above, binarized ALASKA II was split into a training set for the Warden's pooler (one third), named Split P, and an evaluation set (two thirds), named TST.

The biasing was executed as described in Section 3.3. When Alice samples from the binarized set without a bias, she selects an ε-type image with probability $p = N_\varepsilon/(N_\varepsilon + N_M) \doteq 0.3016$ and an M-type image with probability $1 - p$. When sampling with a bias, Alice first selects the group (ε-type with probability $q > p$) and then randomly draws an image from the group. This corresponds in spirit to the biasing considered within our model. We remind the reader that

[5] We experimented with a variety of feature vectors but only report on the ones that performed the best.

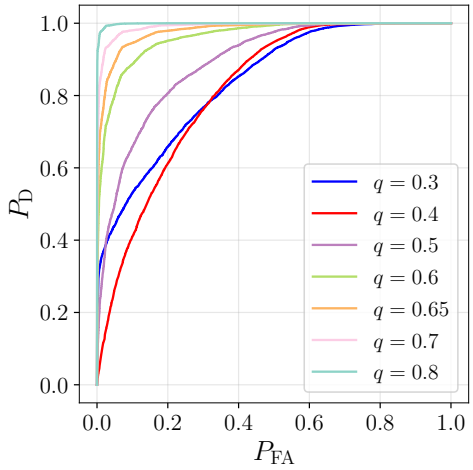

Figure 4: Example of ROC's morphing for Warden's RF pooler with bias q for greedy sender and $n = 16$, $r(n) = 0.5$ on binarized ALASKA II. The figure demonstrates a qualitative match with the theoretical results.

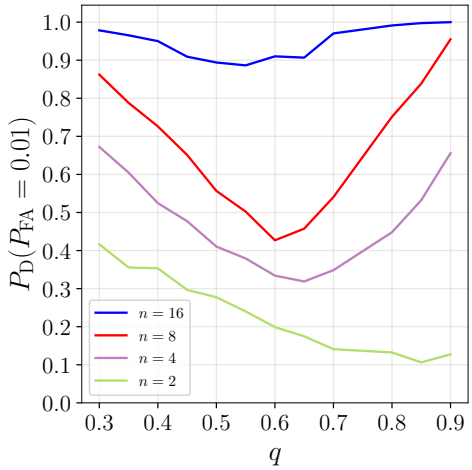

Figure 5: True positive rate P_D at $P_{FA} = 0.01$ versus q for the greedy sender and four bag sizes, $n \in \{2, 4, 8, 16\}$, $r(n) = 1$ on binarized ALASKA II. The optimal bias decreases with increased bag size as observed in our model in Figure 2.

Alice forms bags from TST. Once a bag is formed, the selected sender (greedy or uniform) is used to spread the payload across images in the bag. In case the bag does not have enough capacity to embed the requested payload, the bag is resampled until the capacity requirement is met.[6] Eventually, all n images $X_1, \dots, X_n$ in the bag are passed through Warden's detector d^W to obtain soft outputs $d^W(X_i)$.

As our first batch of experiments, we verified the morphing of the ROC of Warden's pooler with increased bias. We experimented with the following combinations of bag sizes and rates $n, r(n) \in \{(2, 1), (4, 1), (8, 1), (16, 0.5), (16, 1)\}$. In all cases, the ROCs quite closely mimicked the predicted morphing, including the lowered true positive rates for small false alarms, increased true positive rate for larger false alarms, and the "jagged" shape for strong source biasing. Figure 4 shows a typical example for $n = 16$ and $r(n) = 0.5$, which matches the trends predicted by our analysis (Figure 1). In Figure 5, we plot P_D for a fixed P_{FA} as a function of the bias for several bag sizes. Again, we see a qualitatively close match with Figure 2 in terms of the optimal value of the bias decreasing with increased bag size. The bias gain γ_{bias} also initially increases with increased bag size except for the largest bag size $n = 16$ because the P_D at this level of P_{FA} approaches 1.

Next, we turn our focus to verifying the scaling as expressed by the ABT. We use the term sub-SR (sub square root) for the situation when the rate (or bias) decreases faster with n than what would guarantee asymptotic constant statistical detectability: $r^2(n)n \to 0$ (or $(q(n)-p)\sqrt{n} \to 0$). To the contrary, super-SR (super square root) adjustment means that the adjustment is slower, which induces perfect detection: $r^2(n)n \to \infty$ (or $(q(n)-p)\sqrt{n} \to \infty$). To this end, we executed the following three experiments.

(1) Adjusting both the rate and the bias sub-SR to verify that the performance of Warden's pooler approaches random guessing. See Figure 6 (left) and Figure 7.
(2) Adjusting the rate and bias for constant statistical detectability to verify the convergence and symmetrization of the ROCs. See Figure 6 (middle).
(3) Adjusting the rate super-SR and bias sub-SR to verify that asymptotically the overly large rate gives perfect detectability regardless of bias gain. See Figure 6 (right).

For our experiments, we fix the relative payload for bag size $n = 4$ at $r(4) = 0.5$ and the biasing parameter $q(4) = 0.4$, which is the optimal bias for this bag size and rate as determined experimentally. With increasing bag size n, the relative payload and the biasing are scaled as

$$q(n) - p = (q(4) - p)\left(\frac{4}{n}\right)^{1-\lambda_b} \tag{20}$$

$$r(n) = r(4)\left(\frac{4}{n}\right)^{1-\lambda_r}. \tag{21}$$

The scaling for the rate and bias is controlled by the two exponents λ_r and λ_b, respectively. Sub-SR adjustment corresponds to $\lambda_r, \lambda_b < 1/2$ (the rate and bias decay more rapidly with n) while super-SR scaling uses $\lambda_r, \lambda_b > 1/2$. When $\lambda_r = \lambda_b = 1/2$, the third part of the ABT predicts asymptotic constant statistical detectability and a Gaussian ROC. This is indeed confirmed in Figure 6 (middle). Note that the ROCs for the largest bag sizes are visually indistinguishable. In Figure 6 (right), the biasing is reduced more aggressively than the rate (sub-SR for bias $\lambda_b = 1/4$, super-SR for rate $\lambda_r = 3/4$), and we observe the Warden's pooler approach perfect detection. When both the rate and bias are adjusted with $\lambda_r = 1/4$, $\lambda_b = 1/4$ (sub-SR), the ROCs gradually flatten effectively approaching that of a random guesser (Figure 6 left) in agreement with the ABT. Since some detection power remains even for the largest bag size, we include a log-log plot of $0.5 - P_E(n)$ vs. bag size n in Figure 7 to verify that the ROCs are indeed tending towards random guessing. Assuming the ROCs are Gaussian, then using the

[6] This happens infrequently and only when embedding very large payloads.

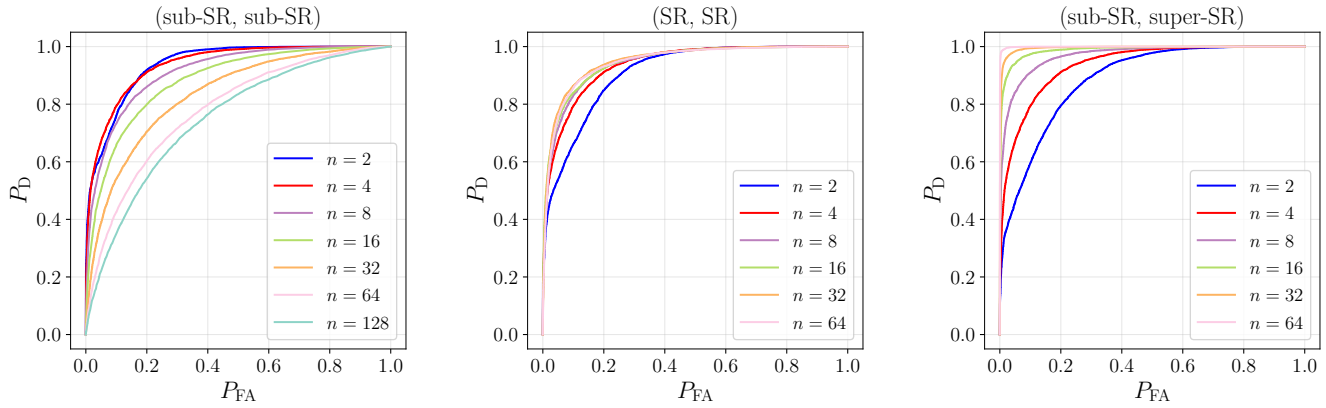

Figure 6: Asymptotic trends of the ROC of Warden's pooler on binarized ALASKA II for uniform sender when adjusting the bias and the rate based on Eqs.(20) and (21). The first parameter in the parentheses is the scaling exponent used for biasing, while the second parameter is exponent for the rate.

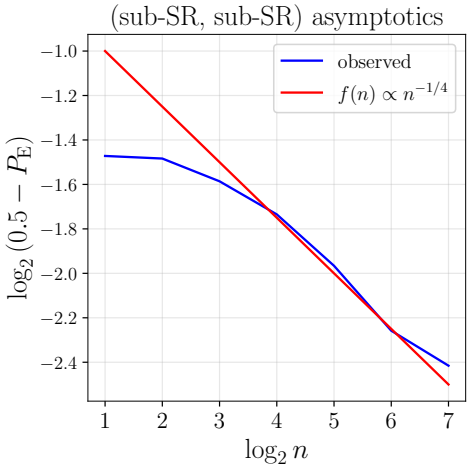

Figure 7: Log-log plot of $0.5 - P_E(n)$ vs. bag size n when adjusting both the rate and bias sub-SR on binarized ALASKA II (see Figure 6). We experimentally confirm a tendency toward perfect security.

fact that the deflection (KL-divergence) for this particular scaling of the rate is proportional to $n^{-1/2}$ we have

$$P_E(n) = Q\left(\frac{1}{2}n^{-1/4}\right) \doteq \frac{1}{2} - \frac{1}{2\sqrt{2\pi}}n^{-1/4} \tag{22}$$

by the first-order Taylor expansion of $Q(x)$ at $x = 0$. Thus, $0.5 - P_E(n) \propto n^{-1/4}$ and we should observe $0.5 - P_E(n)$ tend toward a line of slope $-1/4$ in a log-log plot. Note that to stay closer with our modeling assumptions, Alice's SID used for spreading was the same as Warden's SID but the conclusions drawn and the qualitative nature of the results remain valid when they use different SIDs.

5.6 Biasing in real life conditions

To find out the impact of biasing in a more realistic dataset and under more realistic conditions, we executed the next experiment on the original (i.e., non-binarized) Split BSPS of ALASKA II. We used the same greedy sender as in the previous section but the biasing algorithm needed an adjustment because now we have a continuous distribution of slopes.

Consider a continuous and strictly increasing CDF F. Biasing the sampling of F can be implemented by a modified inverse transform sampling algorithm. Without biasing, a sample from F can be generated by first sampling a uniform r.v. $U \sim \mathcal{U}[0,1]$ and then computing $F^{-1}(U)$ where F^{-1} is the inverse CDF, or quantile function. Now consider a beta random variable $Y \sim \text{Beta}(\alpha^*, \beta^*)$ whose CDF is denoted by G_{α^*,β^*}. If we instead compute $F^{-1}(G^{-1}_{\alpha^*,\beta^*}(U))$, our sample will follow the distribution given by

$$\mathbb{P}(F^{-1}(G^{-1}_{\alpha^*,\beta^*}(U)) \le u) = \mathbb{P}(U \le G_{\alpha^*,\beta^*}(F(u))) = G_{\alpha^*,\beta^*}(F(u)). \tag{23}$$

for $u \in [0,1]$ (0 for $u < 0$, 1 for $u > 1$). Observe that $Y \overset{d}{=} G^{-1}_{\alpha^*,\beta^*}(U)$ (equal in distribution) so we can view our biasing method as sampling quantiles of F using a beta random variable instead of the usual $\mathcal{U}[0,1]$. We chose the beta distribution since $\text{Beta}(1,1) \overset{d}{=} \mathcal{U}[0,1]$, its support is $[0,1]$, and G_{α^*,β^*} is continuous in its parameters $\alpha^*, \beta^* > 0$.

Suppose we obtain a dataset of N i.i.d. samples $X_1, \ldots, X_N \sim F$. Using $\mathbf{1}_A$ to denote the indicator function of an event A, we consider the empirical distribution function (ECDF) of the samples

$$\widetilde{F}_N(u) = \frac{1}{N}\sum_{k=1}^{N} \mathbf{1}_{X_k \le u}. \tag{24}$$

Our goal is to generate n biased samples from $\widetilde{F}_N$ without replacement. We use the inverse transform sampling method, except to implement sampling without replacement, we sample from a sequence of ECDFs. Specifically, to sample the ith variable given we already obtained samples $X_{k_1}, X_{k_2}, \ldots, X_{k_{i-1}}$ (indexed by $k_1, \ldots, k_{i-1}$) we perform inverse transform sampling using the ECDF:

$$\widetilde{F}^{(i)}_N(u) = \frac{1}{N-i}\sum_{\substack{k \in \{1,\ldots,N\} \\ k \notin \{k_1,\ldots,k_{i-1}\}}} \mathbf{1}_{X_k \le u}, \tag{25}$$

Algorithm 1 A procedure for obtaining samples from a finite dataset in a biased (or non-uniformly random) fashion. The procedure returns the indices used to uniquely query a dataset.

```
// Set dataset size N, number of samples n, biasing parameter q ≥ 1
Z_1, ..., Z_n ← obtain n iid realizations of Beta(1/q, 1)
for i = 1 to n:
  m_i ← ceil((N − i + 1)Z_i)
for i = 1 to n:
  k_i ← m_i
  for j = 1 to i − 1:
    if m_{i−j} ≤ k_i:
      k_i ← k_i + 1
return k_1, ..., k_n
```

where the sum is over all indices $K = 1, \ldots, N$ not equal to indices already sampled $k_1, \ldots, k_{i-1}$. Algorithm 1 depicts how the sampling is done without explicitly computing the ECDFs.

To relate this procedure to a biased sampling of bags, we associate the distribution of slopes to the CDF F and the size of evaluation set to N. Alice samples covers with a bias by selecting α^*, β^*. For easier analysis and interpretation, it is advantageous to have a single biasing parameter, which we again denote q. The choice of how the beta distribution should be parameterized by q is non-trivial. We tried two simple parametrizations: $\mathrm{Beta}(1/q, 1)$ and $\mathrm{Beta}(1, q)$ both for $q \geq 1$. We found that the former, $\mathrm{Beta}(1/q, 1)$, permits a larger bias gain likely due to the latter parametrization $\mathrm{Beta}(1, q)$ morphing the right tail of the distribution of slopes too aggressively. Indeed, the bivalued spreading strategies (and spreading strategies in general [24]) prioritize images belonging to the left tail, so morphing the right tail contributes relatively less to the gain in steganographic security. Hence, for all continuous biasing experiments below, we use the parametrization $\mathrm{Beta}(1/q, 1)$ in Algorithm 1.

For the following experiments, the sender uses $d^{\mathrm{A}} = \mathrm{SRNetA}$ for biasing and for spreading while the Warden uses $d^{\mathrm{W}} = \mathrm{SRNetW}$. The feature vector for the machine-learning built pooler is the same as above (18) with the only difference that we removed the quantity n_ε/n because there are no groups of images with distinct slopes in the evaluation dataset—the slopes are continuously distributed.

Figure 8 shows the ROCs of Warden's pooler for bag size $n = 4$ when increasing the biasing parameter q in $\mathrm{Beta}(1/q, 1)$ used for continuous biasing. We can see that many of the ROC morphing characteristics observed in our model and binarized ALASKA II are preserved. For instance, with increased bias $P_{\mathrm{D}}(P_{\mathrm{FA}})$ is monotonically increasing for large enough fixed P_{FA}, and we see the same initial decrease in $P_{\mathrm{D}}(P_{\mathrm{FA}})$ for low fixed P_{FA}. However, the jagged characteristic of the ROCs is gone due to the continuous nature of our biasing.

Figure 9 presents $P_{\mathrm{D}}(P_{\mathrm{FA}} = 0.02)$ vs. the continuous biasing parameter $q \geq 1$ for various bag sizes. We see that the trend of optimal q monotonically decreasing in bag size n is preserved. Additionally, we see that γ_{bias} exhibits similar monotonic trends seen in the experiments on binarized ALASKA II and the model. Similar to Figure 5, for $n = 16$, γ_{bias} decreases as P_{D} approaches 1.

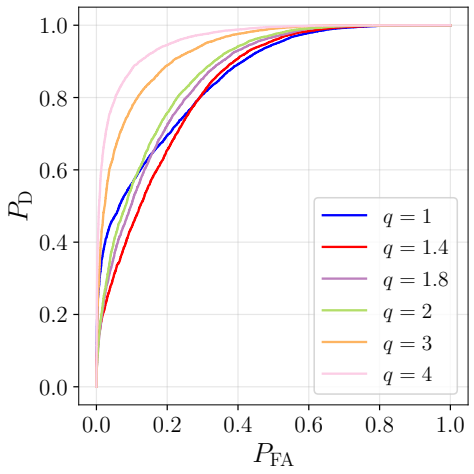

Figure 8: ROCs of Warden's pooler for bag size $n = 4$ and rate $r = 0.5$ for a range of continuous biasing parameter values $q \geq 1$.

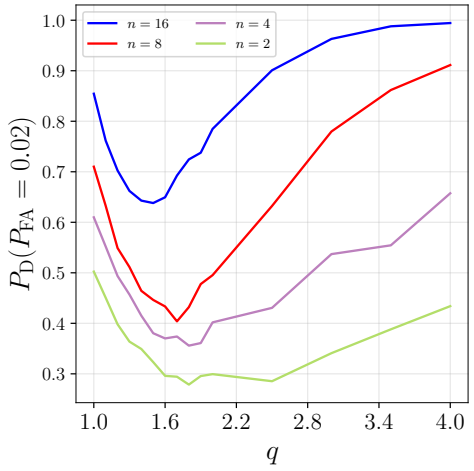

Figure 9: P_{D} at $P_{\mathrm{FA}} = 0.02$ vs. the continuous biasing parameter q for bag sizes $n = 2, 4, 8, 16$ and fixed rate $r = r(n) = 0.7$.

To demonstrate that the bias gain manifests robustly w.r.t. Warden's pooler, besides the results with the clairvoyant pooler, we provide limited evidence[7] for a realistic (non-clairvoyant) version of this pooler when the Warden estimates the response curve slopes and payloads from images at hand using feedback from her own detector d^{W}. To be more precise, b_i and α_i in the feature vector are replaced with Warden's estimates and π_{CORR} (19) is also computed with the estimates. Moreover, we add to our study another simple pooler that is agnostic w.r.t. slopes and payloads. It is trained as RF on a subset of the feature vector 18 with only n soft outputs $d^{\mathrm{W}}(X_i)$. Figure 10 confirms that the bias gain robustly manifests for all three poolers irrespective of Warden's knowledge of payloads and slopes.

[7] Experiments with non-clairvoyant pooler are very computationally demanding.

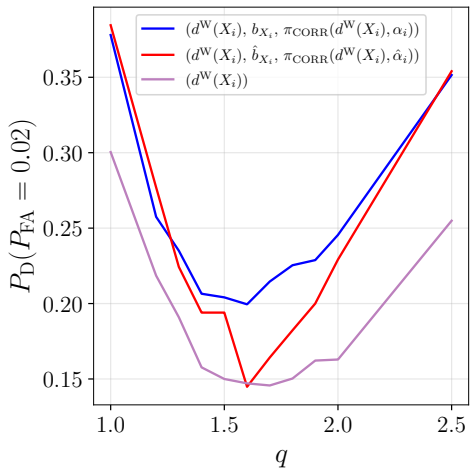

Figure 10: P_{D} at $P_{\mathrm{FA}} = 0.02$ vs. the continuous biasing parameter q for bag size $n = 2$ and rate $r = 0.5$ for three different poolers described in the text.

6 CONCLUSIONS

The steganographer can decrease her chances of being caught by the Warden by selecting covers in which steganographic embedding changes are less detectable. This, however, changes the cover source, which is detectable on its own. In this paper, we study the trade off between the increased security with respect to detectors of embedding changes and the vulnerability to source detectors to find out the impact of biasing the source. We begin with a theoretical study from a simple source model and then confirm the theoretically predicted findings experimentally with a source of real images and modern detectors built with deep learning. Our findings can be summarized as follows:

(1) With increased bias towards selecting images that are harder to steganalyze, the true positives of Warden's optimal pooler start decreasing for small false alarm rates but increase for large false alarms, which indicates a steganographer's gain.
(2) Measuring this "bias gain" as the loss of Warden's detection power at a fixed false alarm rate, with increased number of uses of the stego channel, the optimal value of the bias decreases.
(3) As the number of communicated images approaches infinity, both the payload and the bias need to be adjusted at their critical rates for constant asymptotic statistical detectability.

Our experiments with real datasets and detectors indicate that cover selection can indeed produce a significant gain in security in practice, which is relevant for practitioners. We note our findings qualitatively align with the recent work [10]; choosing difficult sources helps Alice only up to a certain point. Many intriguing open questions remain to be answered, including optimizing the source biasing algorithm and jointly optimizing the biasing and payload allocation among multiple images.

ACKNOWLEDGMENTS

The work on this paper was supported by NSF grant No. 2028119.

APPENDIX

Proof of Theorem

Before starting the proof, we state several useful facts. We write $f(n) \asymp g(n)$ when $f(n) = g(n)(1 + o(1))$ as $n \to \infty$.

For any x,

$$x \log \frac{q}{p} + (1-x) \log \frac{1-q}{1-p} = (x-p)\ell(p,q) - D_{\mathrm{KL}}(p||q) \tag{26}$$

$$\ell(p,q) = \log \frac{q}{p} \frac{(1-p)}{(1-q)}. \tag{27}$$

When $(q(n)-p)\sqrt{n} \to c_p$, by using Taylor expansion of $\log(1+x)$ at $x = 0$, it is straightforward to show that

$$\sqrt{n}\ell(p,q) = \frac{c_p}{p(1-p)} + O(n^{-1/2}) \tag{28}$$

$$nD_{\mathrm{KL}}(p||q) = \frac{c_p^2}{2p(1-p)} + O(n^{-1/2}). \tag{29}$$

——

A version of the De Moivre–Laplace theorem follows from the Stirling's formula: Let $c > 0$ and $1/2 < a < 2/3$. For any k, $|k-np| < cn^a$, as $n \to \infty$

$$\binom{n}{k} p^k (1-p)^{n-k} = \frac{1}{\sqrt{2\pi np(1-p)}} e^{-\frac{(k-pn)^2}{2np(1-p)}} \left(1 + O\left(1/n^{2-3a}\right)\right). \tag{30}$$

For $k < np$ ($k > np$), the left (right) tails of the cumulative distribution function of the binomial distribution are bounded by the Hoeffding's bound

$$\Pr\{\mathrm{Binom}(p,n) \lesseqgtr k\} \leq e^{-2n(p-k/n)^2}, \tag{31}$$

which translates for $k < np - cn^a$ and $k > np + cn^a$ to

$$\Pr\{\mathrm{Binom}(p,n) \lesseqgtr k\} \leq e^{-2c^2 n^{2a-1}}. \tag{32}$$

——

Given $Y \sim \mathcal{N}(\mu, \sigma^2)$ and $Z \sim \mathcal{N}(0,1)$ independent, we have by the definition of the Q function and the law of total probability that

$$\mathbb{E}[Q(Y)] = \mathbb{E}[\mathbb{P}(Z > Y|Y)] = \mathbb{P}(Z > Y). \tag{33}$$

Since $Z - Y \sim \mathcal{N}(-\mu, \sigma^2 + 1)$, we can compute the right hand side via the Q function

$$\mathbb{P}(Z > Y) = Q\left(\frac{\mu}{\sqrt{\sigma^2+1}}\right). \tag{34}$$

Asymptotic perfect security

The first part of the theorem follows from bounds on the KL-divergence between the distributions of the HT (7):

$$\begin{aligned} \mathrm{D}_{\mathrm{KL}}(\mathcal{H}_0||\mathcal{H}_1) &= -\mathbb{E}\left[L(\mathbf{b},\mathbf{y})|\mathcal{H}_0\right] \\ &= \mathbb{E}\left[\Delta^2(K)|\mathcal{H}_0\right] - \mathbb{E}\left[L_{\mathrm{binom}}(K)|\mathcal{H}_0\right]. \end{aligned} \tag{35}$$

For a bivalued spreader with $\alpha_\varepsilon, \alpha_1$, $\Delta^2(k)$ is upper bounded by the deflection for a spreader that puts all payload into images with slope 1. Such images receive $\alpha_1 = nr(n)/(n-k)$ bpp as long as $nr(n) \leq (n-k)\log_2 3$. Since $nr^2(n) \to 0$, $nr^2(n) < \delta$ for any $\delta > 0$ for sufficiently large n, and thus $nr(n) < n^{1/2}\delta^{1/2}$. This sub-linear

payload will thus fit into images with slope 1 for all $k \le ns$, for any $0 \le s < 1$ for sufficiently large n. We write

$$\begin{aligned}\mathbb{E}\left[\Delta^2(K)|\mathcal{H}_0\right] &\overset{(H)}{\asymp} \frac{1}{2}\sum_{k=0}^{ns}\binom{n}{k}p^k(1-p)^{n-k}k\frac{n^2r^2(n)}{(n-k)^2}\\ &\le \frac{1}{2}\sum_{k=0}^{ns}\binom{n}{k}p^k(1-p)^{n-k}ns\frac{n^2r^2(n)}{(n-ns)^2}\\ &= \frac{1}{2}nr^2(n)\frac{s}{(1-s)^2}\sum_{k=0}^{ns}\binom{n}{k}p^k(1-p)^{n-k}\\ &\le \frac{1}{2}nr^2(n)\frac{s}{(1-s)^2} \end{aligned} \tag{36}$$

which approaches 0 as $n \to \infty$. The first asymptotic equality $\overset{(H)}{\asymp}$ follows from Hoeffding inequality (32) and the fact that $\Delta^2(k)$ is $o(n^2)$ independently of k.

Next, we inspect the right-most term in (35), which is the KL-divergence between $\mathrm{Binom}(p,n)$ and $\mathrm{Binom}(q,n)$. Performing a Taylor expansion at $q=p$, we get

$$\begin{aligned}-\mathbb{E}\left[L_{\mathrm{binom}}(K)|\mathcal{H}_0\right] &= pn\log\frac{p}{q} + (1-p)n\log\frac{1-p}{1-q}\\ &= \frac{n(q-p)^2}{2p(1-p)} + nO((q-p)^3). \end{aligned} \tag{37}$$

By supposition, this also approaches 0 and thus completes the proof of the first statement of the theorem.

Asymptotic perfect detectability

For the second statement, we note that the optimal spreader (the least detectable spreader) in a source of images all of the same slope ε is the uniform spreader. This is because $\frac{1}{2}\sum_{i=1}^n \varepsilon^2\alpha_i^2$ is minimal subject to $\sum_{i=1}^n \alpha_i = nr(n)$ when $\alpha_i = nr(n)/n = r(n)$ for all i. By the classical SRL result, this payload-limited sender is asymptotically perfectly detectable when $n^{1/2}r(n) \to \infty$. Since a cover source consisting of only images with slope ε is harder to steganalyze than a source consisting of a mixture of images with slopes ε and 1, we can conclude the following. If the optimal spreading in a strictly more difficult source is asymptotically perfectly detectable, then any spreading in an easier source must also be asymptotically perfectly detectable.

All that remains to show is that when $n(q(n)-p)^2 \to \infty$ there exists an asymptotically perfect source detector. The test that achieves this performance is the count of images with slope ε, which we denote K. Since $K \sim \mathrm{Binom}(p,n)$ under $\mathcal{H}_0$ and $K \sim \mathrm{Binom}(q,n)$ under $\mathcal{H}_1$, the normalized test

$$\frac{K-pn}{\sqrt{n}} \overset{(d)}{\to} \begin{cases}\mathcal{N}(0, p(1-p)) & \text{under } \mathcal{H}_0\\ \mathcal{N}\left(\sqrt{n}(q-p), q(1-q)\right) & \text{under } \mathcal{H}_1\end{cases}$$

by invoking the De Moivre–Laplace theorem, which proves asymptotic perfect detectability.

Case of root-rate and root -biasing

For the proof, we limit ourselves to the greedy sender (Section 3.4) to simplify the arguments.

We begin the proof by finding the steganographic deflection (9) for the greedy sender. Since the absolute payload satisfying $r(n)n = c_r^{1/2}n^{1/2}(1+o(1))$ ensures $r(n)n/\log_2 3 < pn - cn^a$ for any $c > 0$ for sufficiently large n, for $k \ge pn - cn^a$

$$\Delta^2_{\mathrm{greedy}}(k) = \frac{1}{2}\sum_{i=1}^{k}\varepsilon^2\left(\frac{r(n)n}{k}\right)^2 = \frac{Cn}{2k}, \tag{38}$$

where we denoted for brevity $C = \varepsilon^2 c_r(1+o(1))$. Next, we have for P_{FA} of optimal Warden's pooler (11):

$$\begin{aligned}P_{\mathrm{FA}}(x) &\overset{(t)}{\asymp} \sum_{k=pn-cn^a}^{pn+cn^a}\binom{n}{k}p^k(1-p)^{n-k}Q\left(\frac{x-E_0(k)}{\sqrt{V(k)}}\right)\\ &\overset{(a)}{\asymp} \sum_{k=pn-cn^a}^{pn+cn^a}\frac{1}{\sqrt{2\pi np(1-p)}}e^{-\frac{(k-pn)^2}{2np(1-p)}}Q\left(\frac{x-E_0(k)}{\sqrt{V(k)}}\right)\\ &\overset{(b)}{=} \sum_{|l|\le cn^{a-1/2}}\frac{1}{\sqrt{2\pi np(1-p)}}e^{-\frac{l^2}{2p(1-p)}}\times\\ &\quad Q\left(\frac{x+\frac{C}{2(l/\sqrt{n}+p)}-\left[l\sqrt{n}\ell(p,q)-nD_{\mathrm{KL}}(p||q)\right]}{\sqrt{\frac{C}{l/\sqrt{n}+p}}}\right)\end{aligned} \tag{39}$$

The approximation (t) is due to (32) and the fact that $Q(x) \le 1$, (a) is from (30), and in (b) we used (38), (13), and $l = (k-pn)/\sqrt{n}$, which increments by $1/\sqrt{n}$.

Next, we consider (39) as a Riemann sum approximation to an integral. The order of approximation is the length of the interval, which is $1/\sqrt{n}$. We add the rest of the entire real line to the integration (which adds a multiplicative factor of $1+o(1)$ by (32)) and obtain

$$P_{\mathrm{FA}}(x) \asymp \int_{-\infty}^{\infty}\sqrt{\frac{1}{2\pi p(1-p)}}\exp\left(-\frac{u^2}{2p(1-p)}\right)$$

$$\times Q\Bigg(x\sqrt{\frac{u/\sqrt{n}+p}{C}}+\frac{1}{2}\sqrt{\frac{C}{u/\sqrt{n}+p}} \tag{40}$$

$$-\sqrt{\frac{u/\sqrt{n}+p}{C}}\left[\sqrt{n}\ell(p,q)u - nD_{\mathrm{KL}}(p||q)\right]\Bigg)\mathrm{d}u \tag{41}$$

$$\asymp \int_{-\infty}^{\infty}\sqrt{\frac{1}{2\pi p(1-p)}}\exp\left(-\frac{u^2}{2p(1-p)}\right) \tag{42}$$

$$\times Q\left(x\sqrt{\frac{p}{\varepsilon^2 c_r}}+\frac{1}{2}\sqrt{\frac{\varepsilon^2 c_r}{p}}-\sqrt{\frac{p}{\varepsilon^2 c_r}}\left[\frac{c_p}{p(1-p)}u-\frac{c_p^2}{2p(1-p)}\right]\right)\mathrm{d}u. \tag{43}$$

The last approximation is due to (28) and (29), taking the limit $n \to \infty$, and swapping the limit and integration, which is permissible because the integrand converges uniformly on all closed intervals. We further rewrite (43)

$$P_{\mathrm{FA}}(x) \asymp \int_{-\infty}^{\infty}\sqrt{\frac{1}{2\pi p(1-p)}}\exp\left(-\frac{u^2}{2p(1-p)}\right)Q(D-Au)\mathrm{d}u, \tag{44}$$

where $A = \sqrt{\frac{p}{\varepsilon^2 c_r}}\frac{c_p}{p(1-p)}$ and $D = x\sqrt{\frac{p}{\varepsilon^2 c_r}} + \frac{1}{2}\sqrt{\frac{\varepsilon^2 c_r}{p}} + \sqrt{\frac{p}{\varepsilon^2 c_r}}\frac{c_p^2}{2p(1-p)}$. Next, we use substitution $u = (D-z)/A$ and (33)–(34)

$$P_{\text{FA}}(x) \asymp \int_{-\infty}^{\infty} \sqrt{\frac{1}{2\pi A^2 p(1-p)}} \exp\left(-\frac{(z-D)^2}{2A^2 p(1-p)}\right) Q(z)\text{d}z$$
$$= \mathbb{E}[Q(Y)] = Q\left(\frac{\mu}{\sqrt{\sigma^2+1}}\right) \tag{45}$$

with $Y \sim \mathcal{N}(\mu, \sigma^2)$, $\mu = D$ and $\sigma^2 = A^2 p(1-p) = \frac{c_p^2}{\varepsilon^2 c_r (1-p)}$. This allows to finally obtain

$$P_{\text{FA}}(x) \asymp Q\left(\frac{x\sqrt{\frac{p}{\varepsilon^2 c_r}} + \frac{1}{2}\sqrt{\frac{\varepsilon^2 c_r}{p}} + \sqrt{\frac{p}{\varepsilon^2 c_r}}\frac{c_p^2}{2p(1-p)}}{\sqrt{1+\frac{c_p^2}{\varepsilon^2 c_r(1-p)}}}\right) = Q\left(\frac{x+\frac{1}{2}d^2}{d}\right) \tag{46}$$

where $d^2 = \frac{\varepsilon^2 c_r}{p} + \frac{c_p^2}{p(1-p)}$.

For $P_{\text{D}}(x)$ (12), while reminding that $E_1(k) = L_{\text{binom}}(k) + \Delta^2(k)$, a series of quite similar steps and arguments can be applied by simply replacing p with $p + n^{-1/2}c_p(1+o(1))$. The Gaussian approximation to the binomial term will now be shifted by c_p:

$$P_{\text{D}}(x) \asymp \int_{-\infty}^{\infty} \sqrt{\frac{1}{2\pi p(1-p)}} \exp\left(-\frac{(u-c_p)^2}{2p(1-p)}\right)$$
$$\times Q\left(x\sqrt{\frac{p}{\varepsilon^2 c_r}} - \frac{1}{2}\sqrt{\frac{\varepsilon^2 c_r}{p}} - \sqrt{\frac{p}{\varepsilon^2 c_r}}\left[\frac{c_p}{p(1-p)}u - \frac{c_p^2}{2p(1-p)}\right]\right)\text{d}u$$
$$= \int_{-\infty}^{\infty} \sqrt{\frac{1}{2\pi p(1-p)}} \exp\left(-\frac{u^2}{2p(1-p)}\right) Q(D-Au)\text{d}u \tag{47}$$

which can be further rewritten using the same substitution $u = (D-z)/A$ as

$$P_{\text{D}}(x) \asymp \int_{-\infty}^{\infty} \sqrt{\frac{1}{2\pi A^2 p(1-p)}} \exp\left(-\frac{(D-z-c_pA)^2}{2A^2 p(1-p)}\right) Q(z)\text{d}z$$
$$= \mathbb{E}[Q(Y)] = Q\left(\frac{\mu}{\sqrt{\sigma^2+1}}\right) \tag{48}$$

with $Y \sim \mathcal{N}(\mu, \sigma^2)$, $\mu = D - c_pA$ and $\sigma^2 = A^2 p(1-p) = \frac{c_p^2}{\varepsilon^2 c_r(1-p)}$, which gives

$$P_{\text{D}}(x) \asymp Q\left(\frac{x-\frac{1}{2}d^2}{d}\right). \tag{49}$$

In summary, asymptotically the ROC of Warden's optimal pooler is Gaussian $P_{\text{D}}(P_{\text{FA}}) = Q\left(Q^{-1}(P_{\text{FA}}) - d\right)$.

The same steps as above can be followed to obtain the equivalent result for the uniform sender, $d^2_{\text{uniform}} = (p\varepsilon^2 + 1 - p)c_r + \frac{c_p^2}{p(1-p)}$,and potentially for any bivalued sender once adopting a policy for adjusting sender's rates α_ε and α_1 with n.

REFERENCES

[1] P. Bas. Steganography via cover-source switching. In *2016 IEEE International Workshop on Information Forensics and Security (WIFS)*, pages 1–6, December 4-7 2016.

[2] R. Böhme. *Improved Statistical Steganalysis Using Models of Heterogeneous Cover Signals*. PhD thesis, Faculty of Computer Science, Technische Universität Dresden, Germany, 2008.

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

[4] L. Breiman. Random forests. *Machine Learning*, 45:5–32, October 2001.

[5] J. Butora, Y. Yousfi, and J. Fridrich. How to pretrain for steganalysis. In D. Borghys and P. Bas, editors, *The 9th ACM Workshop on Information Hiding and Multimedia Security*, Brussels, Belgium, June 22–25, 2021. ACM Press.

[6] C. Cachin. An information-theoretic model for steganography. *Information and Computation*, 192(1):41–56, July 2004.

[7] R. Cogranne, Q. Giboulot, and P. Bas. ALASKA–2: Challenging academic research on steganalysis with realistic images. In *IEEE International Workshop on Information Forensics and Security*, New York, NY, December 6–11, 2020.

[8] E. Dworetzky and J. Fridrich. Explaining the bag gain in batch steganography. *IEEE Transactions on Information Forensics and Security*, 18:3031–3043, 2023.

[9] E. Dworetzky, E. Kaziakhmedov, and J. Fridrich. Advancing the JPEG compatibility attack: Theory, performance, robustness, and practice. In Y. Yousfi, C. Pasquini, and A. Bharati, editors, *The 11th ACM Workshop on Information Hiding and Multimedia Security*, Chicago, IL, June 28–30, 2023. ACM Press.

[10] Q. Giboulot, T. Pevný, and A. D. Ker. The non-zero-sum game of steganography in heterogeneous environments. *IEEE Transactions on Information Forensics and Security*, 18:4436–4448, 2023.

[11] A. Gretton, K. M. Borgwardt, M. Rasch, B. Schölkopf, and A. J. Smola. A kernel method for the two-sample-problem. In B. Schölkopf, J. Platt, and T. Hoffman, editors, *Advances in Neural Information Processing Systems 19*, pages 513–520. MIT Press, Cambridge, MA, 2007.

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

[13] X. Hu, J. Ni, W. Zhang, and J. Huang. Efficient JPEG batch steganography using intrinsic energy of image contents. *IEEE Transactions on Information Forensics and Security*, 16:4544–4558, 2021.

[14] A. D. Ker. Batch steganography and pooled steganalysis. In J. L. Camenisch, C. S. Collberg, N. F. Johnson, and P. Sallee, editors, *Information Hiding, 8th International Workshop*, volume 4437 of Lecture Notes in Computer Science, pages 265–281, Alexandria, VA, July 10–12, 2006. Springer-Verlag, New York.

[15] A. D. Ker. The square root law of steganography. In M. Stamm, M. Kirchner, and S. Voloshynovskiy, editors, *The 5th ACM Workshop on Information Hiding and Multimedia Security*, Philadelphia, PA, June 20–22, 2017. ACM Press.

[16] A. D. Ker and Tomas Pevný. Batch steganography in the real world. In J. Dittmann, S. Craver, and S. Katzenbeisser, editors, *Proceedings of the 14th ACM Multimedia & Security Workshop*, pages 1–10, Coventry, UK, September 6–7, 2012.

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

[18] H. Sajedi and M. Jamzad. Using contourlet transform and cover selection for secure steganography. *International Journal of Information Security*, 9(5):337–352, October 2010.

[19] V. Sedighi, R. Cogranne, and J. Fridrich. Practical strategies for content-adaptive batch steganography and pooled steganalysis. In *Proceedings IEEE, International Conference on Acoustics, Speech, and Signal Processing*, March 5–9, 2017.

[20] M. S. Subhedar and V. H. Mankar. Curvelet transform and cover selection for secure steganography. *Multimedia Tools Applications*, 77(7):8115–8138, April 2018.

[21] Z. Wang and X. Zhang. Secure cover selection for steganography. *IEEE Access*, 7:57857–57867, May 2019.

[22] Z. Wang, X. Zhang, and Z. Yin. Joint cover-selection and payload-allocation by steganographic distortion optimization. *IEEE Signal Processing Letters*, 25(10):1530–1534, October 2018.

[23] S. Wu, Y. Liu, S. Zhong, and Y. Liu. What makes the stego image undetectable? In *Proc. 7th Int. Conf. Internet Multimedia Comput. Servic (ICIMCS)*, pages 1–6, Hunan, China, 2015.

[24] Y. Yousfi, E. Dworetzky, and J. Fridrich. Detector-informed batch steganography and pooled steganalysis. In J. Butora, C. Veilhauer, and B. Tondi, editors, *The 10th ACM Workshop on Information Hiding and Multimedia Security*, Santa Barbara, CA, 2022. ACM Press.

[25] X. Yu, K. Chen, W. Zhang, Y. Wang, and N. Yu. Improving the embedding strategy for batch adaptive steganography. In *Proc. IWDFW*, pages 248–260, Jeju Island, South Korea, October 2018. Springer.

[26] Z. Zhao, Q. Guan, X. Zhao, H. Yu, and C. Liu. Embedding strategy for batch adaptive steganography. In *Proc. IWDFW*, pages 494–505, Beijing, China, September 2016.