Explaining the Bag Gain in Batch Steganography

Eli Dworetzky and Jessica Fridrich
Binghamton University, Department of ECE

***Abstract*—In batch steganography, the sender distributes the secret payload among multiple images from a "bag" to decrease the chance of being caught. Recent work on this topic described an experimentally discovered phenomenon, which we call the "bag gain": for fixed communication rate, pooled detectors experience a decrease in statistical detectability for initially increasing bag sizes, providing an opportunity for the sender to gain in security. The bag gain phenomenon is universal in the sense of manifesting under a wide spectrum of conditions. In this paper, we explain this experimental observation by adopting a statistical model of detector response. Despite the simplicity of the model, it does capture observed trends in detectability as a function of the bag size, the rate, and cover source properties. Additionally, and surprisingly, the model predicts that in certain cover sources the sender should avoid bag sizes that are too small as this can lead to a bag loss.**

I. Introduction

Batch steganography [11], [12], [14], [17], [9], [15], [19], [21], [22], [20], [25], [24] deals with the situation when the sender spreads her payload among multiple covers to decrease the Warden's chances of detecting the use of this stealth communication channel. Intuitively, images that are harder to steganalyze should receive a larger payload and vice versa. If the sender intends to communicate a fixed payload, she can make her bag size arbitrarily large to achieve her desired security—an infinitely large bag would achieve perfect steganography. To avoid such a degenerate (and uninteresting) case, we will assume that the sender maintains a fixed communication rate instead. For a fixed rate r expressed in terms of bits per pixel (bpp), the sender will eventually be caught due to the square root law (SRL) [13], [16].

This paper builds upon [23] where the authors reported on an interesting phenomenon for batch senders maintaining a positive rate. When pooling evidence from a bag of B images the statistical detectability as a function of B initially decreases with increasing B, then levels off, and eventually increases as the SRL inevitably engages (see Figure 1). The maximal drop in detectability, which we call the *bag gain*, has been observed for all batch senders studied in [23] and for all types of pooled detectors built upon various single-image detectors in the form of rich models as well as convolutional neural networks. It thus appears as a robust phenomenon. The effect of bag size on security was also previously studied in [22] within the context of Gaussian embedding extended to batch senders. While the authors briefly note what appears to be the bag gain in their experiments, it is not clear how and whether their observation, which was obtained with a single-image source detector, extends to a pooled detector. Indeed, as argued below in this paper and as acknowledged by the authors of [22], to properly assess the performance of batch steganography with pooled detectors, one needs to consider the variability of images within bags, which necessitates adopting a model of cover source diversity, an element missing from [22]. Finally, we note that the bag gain did not manifest in previous art [21] because all senders studied in this work embed a variable payload per bag (the rate is maintained only in expectation) based on tags assigned to all images from the cover source computed from an infinitely large bag.

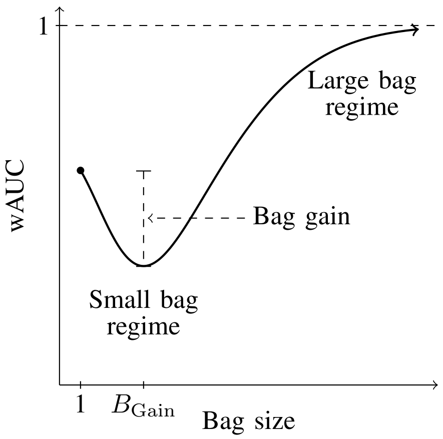

Figure 1. The universal trend of a pooled detector's performance (wAUC) as a function of bag size B with fixed positive rate r. In the small bag regime, it is possible for the steganographer to gain security by selecting a non-trivial bag size. In the large bag regime, the detectability monotonically increases with bag size.

In this paper, we provide an explanation of the bag gain by adopting a model for the soft output of a steganography detector. By suitably simplifying the problem, we become able to analytically study how the bag gain is affected by the detector response, batch sender, cover source, bag size, and communication rate. We argue that the bag gain is due to the differences in how the square root law engages in small and large bags. The bag gain phenomenon is important for practitioners because the security gain can be significant and it occurs for bag sizes that are likely to be used in practice.

In the next section, we describe the general setup for batch steganography and pooled steganalysis as considered in this paper. The purpose of Section III is to adopt suitable modeling assumptions that allow us to derive a closed form expression for the performance of Warden's optimal pooled detector. We also describe the batch sender analyzed in this paper. To capture the diversity of images across bags, in Section IV we adopt a model for the response of a single-image steganography detector on stego images. This model is the key element of our approach as it permits analytic study of the bag gain in Section V, which holds the main bulk of our theoretical results. In particular, we derive a closed-form expression for statistical detectability as a function of the bag size and other parameters describing the cover source and detector response. The derived formulas are contrasted with the performance of a machine learning based pooled detector on real images in Section VI. The model correctly predicts the initial drop in detectability with increasing B. It also captures experimentally obtained trends in detectability vs. the communication rate (Section V-B3). The model additionally predicts a possible bag loss for bag sizes that are too small, which is experimentally confirmed in real datasets. In Section VII, we extend our analysis to a parametrized family of batch senders to study how the bag gain depends on how strongly the senders adapt the payloads to images from the bag. In Section VIII, we contrast our work with relevant prior art on adaptive bag size. The paper is concluded in Section IX.

Throughout this paper, we use $\mathcal{N}(\mu,\sigma^2)$ to denote a normal (Gaussian) distribution with mean μ and variance σ^2. The standard normal tail probability function is denoted $Q(x) = \int_x^\infty (2\pi)^{-1/2} e^{-z^2/2}\mathrm{d}z$. Symbols $\mathbb{P}$ and $\mathbb{E}$ are used for probability and expectation. For a logical statement P, the indicator function, denoted $\mathbf{1}_P$, is equal to 1 when P is true and 0 when P is false. The operation of flooring (rounding to the nearest integer $k \leq x$) is denoted $\lfloor x \rfloor$.

II. Batch steganography formulation

Let $\mathcal{X}$ denote the set of all possible cover images of some fixed size. A cover bag of size B, $\mathbf{X} = (X_0^{(1)}, \dots, X_0^{(B)})$, is formed by independently selecting B cover images $X_0^{(1)}, \dots, X_0^{(B)} \in \mathcal{X}$ according to some probability distribution over $\mathcal{X}$.

To simplify our analysis and without loss on generality of our conclusions, we will assume that each image from $\mathcal{X}$ can be embedded at full capacity of $\log_2 3$ bits per pixel (bpp) with a ternary steganographic scheme. In other words, we assume that images do not contain "wet" pixels [7].

We assume that the steganographer maintains a fixed communication rate $r \in [0, \log_2 3]$ bpp. A batch spreading strategy S is a mapping $\alpha_{r,S} : \mathcal{X}^B \to [0, \log_2 3]^B$ that determines the relative payloads (in bpp) embedded in the B images.[1] When r, S, and $\mathbf{X}$ are clear from context, we simply write $\alpha_i \in [0, \log_2 3]$ to denote the ith component of $\alpha_{r,S}(\mathbf{X})$, i.e., the relative payload embedded in the ith image. The map $\alpha_{r,S}$ must satisfy the payload constraint $\sum_{i=1}^B \alpha_i = rB$. The steganographer produces the ith stego image $X_{\alpha_i}^{(i)}$ by embedding cover $X_0^{(i)}$ with payload of size α_i bpp using a ternary steganographic scheme.

Next, we provide a general formulation of pooled steganalysis. Given an intercepted bag of B images $\mathbf{Y} = (Y^{(1)}, \dots, Y^{(B)})$, the Warden infers whether steganography is being used by performing the following composite hypothesis test:

$$\begin{aligned} \mathcal{H}_0 : &\quad r = 0 \\ \mathcal{H}_1 : &\quad r > 0. \end{aligned} \tag{1}$$

The Warden "pools" the evidence $\mathbf{Y}$ together by using a pooled detector (or "pooler"). We assume the Warden's decision is solely informed by the collection of outputs of a single-image detector, 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. Formally, the Warden's pooler is of the form $\pi : \mathbb{R}^B \to \mathbb{R}$, and she infers whether the sender uses steganography by computing $d(Y^{(i)})$ for all $i = 1, \dots, B$ and comparing $\pi(d(Y^{(1)}), \dots, d(Y^{(B)}))$ against a threshold determined by some application-dependent requirements, such as controlling the false alarm.

In the next two sections, we simplify the formulation above in order to study the bag gain phenomenon analytically. Our approach is *detector-centric* in the sense that we

1) impose statistical models on the response of the detector d on cover and stego images and let all actors share information (next section)
2) model the diversity of bags with a suitably simplified statistical model of the so-called detector response curves that express the dependence of the detector output on message length (Section IV).

III. Modeling assumptions

This paper's goal is to analytically capture and intuitively explain the experimentally observed bag gain phenomenon. This necessitates a rather significant simplification of the setup described in the previous section in terms of what knowledge is available to all actors and in terms of modeling assumptions to facilitate an analytically tractable analysis. To this end, we introduce the concept of acquisition oracle and make specific assumptions about statistical properties of a single-image detector when applied to cover and stego images. We also introduce the batch sender studied in this paper.

Given a collection of cover images indexed by $i = 1, \dots, B$, we consider the specific cover image $X_0^{(i)}$ used by the sender as a sample from an acquisition oracle taking images of the ith cover scene with the same acquisition device. Acquisition noise and possibly small spatial shifts and rotations due to camera shake contribute to the randomness. This oracle will provide us with the means to narrow down the distribution of $d(Y^{(i)})$ under both hypotheses.

A. Gaussianity and local shift hypothesis

First, we take advantage of the fact that, for each i, the distribution of the ith cover acquisition $X_0^{(i)}$ is concentrated on a small subset of $\mathcal{X}$. Since differentiable non-linear functions are approximately linear on sufficiently small neighborhoods, we can employ the central limit theorem (CLT) so that

$$d(X_0^{(i)}) \sim \mathcal{N}(\mu_i, \sigma_i^2), \tag{2}$$

where μ_i and σ_i^2 are the expected value and variance of d on cover images generated by the acquisition oracle for the ith scene. Since stego schemes try to preserve statistical properties of $X_0^{(i)}$, the embedding process will also preserve the concentration. Therefore, by the same argument we assume that $d(X_{\alpha_i}^{(i)})$ is also Gaussian[2]

$$d(X_{\alpha_i}^{(i)}) \sim \mathcal{N}(\mu_i + s_i(\alpha_i), \sigma_i^2) \tag{3}$$

[1] Notice that the mapping is deterministic as we are not considering randomized spreading strategies in this paper.

[2] The random variable $X_{\alpha_i}^{(i)}$ is generated by 1) sampling $X_0^{(i)}$ from the oracle and 2) embedding a random message with a random stego key.

with an additional assumption that only the mean is affected by embedding but not the variance. This *local* shift hypothesis is a much weaker assumption than the shift hypothesis [21] about the *global* distribution of detector response which is not satisfied for modern steganalyzers in the form of rich models and CNNs (see Sec. 3.2 in [23]).

Technically, the variance of $d(X^{(i)}_{\alpha_i})$ also depends on α_i because of the added randomness in the form of the stego key selection and the message itself. We do not consider this dependence in order to further simplify the modeling and also because the acquisition noise dominates the statistical spread because it is stronger than the stego noise.

Finally, to avoid modeling the distribution of the variances σ_i^2 across images from $\mathcal{X}$ and the oracle itself, we assume all variances are the same across scenes $\sigma_i^2 = \sigma^2$.

B. Uniformity of response increase

The response curve (RC) for image $X^{(i)}_0$ and detector d is the function $\varrho_i : [0, \log_2 3] \to \mathbb{R}$ defined by

$$\varrho_i(\alpha) = \mathbb{E}[d(X^{(i)}_{\alpha})|X^{(i)}_0]. \tag{4}$$

Given the payload size α and a fixed cover $X^{(i)}_0$, $\varrho_i(\alpha)$ is the expected value of the response $d(X^{(i)}_{\alpha})$ when embedding $X^{(i)}_0$ with random messages and stego keys.

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

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

for all realizations of $X^{(i)}_0$. This assumption allows us to compute the expected shift $s_i(\alpha)$ from a specific cover image, which simplifies analysis and practical implementations.

C. Warden's test

Equipped with a single-image detector d that adheres to the assumptions above, the Warden's hypothesis test (1) becomes:

$$\begin{aligned} \mathcal{H}_0 :\quad & d(Y^{(i)}) \sim \mathcal{N}(\mu_i, \sigma^2) \quad \text{for all } i \\ \mathcal{H}_1 :\quad & d(Y^{(i)}) \sim \mathcal{N}(\mu_i + s_i(\alpha_i), \sigma^2) \quad \text{for all } i, \end{aligned} \tag{6}$$

where $Y^{(i)}$ are the images from a bag under inspection by the Warden and α_i is the payload residing in the ith image.

Assuming the parameters of the distributions in the hypothesis test (6) are known to the Warden, the test becomes simple and the Warden's most powerful pooled detector is the likelihood ratio test. The detectability of steganography in a single bag is determined by the deflection coefficient

$$\Delta^2(\mathbf{X}) = \sum_{i=1}^{B} \frac{s_i^2(\alpha_i)}{\sigma^2} = \sum_{i=1}^{B} \frac{(\varrho_i(\alpha_i) - \varrho_i(0))^2}{\sigma^2}, \tag{7}$$

where $s_i(\alpha_i)$ can be computed via $\varrho_i(\alpha_i) - \varrho_i(0)$ given any oracle realization $X^{(i)}_0$.

D. Minimum deflection sender

As a batch sender for our study, we selected the detector-informed Minimum Deflection Sender (MDS) introduced in [23] because it is the most amenable to analysis within the context of a statistical model of the detector. As will be argued in Section VII, the bag gain generally manifests for batch senders that minimize the risk of being detected by assigning larger payloads to images that are difficult to steganalyze and smaller payloads to images in which the embedding is more detectable. In particular, the bag gain has also been observed for the detector-agnostic Image Merging Sender (IMS) [21] and detector-aware Shift Limited Sender (SLS) [23].

The MDS makes use of a single-image detector, which we will assume is the same as the one used by the Warden. Given a bag of images $\mathbf{X}$, the MDS selects payloads α_i that minimize the deflection (7). Formally, α_i are found by solving the following optimization problem

$$\begin{aligned} &\text{minimize } \Delta^2(\mathbf{X}), \\ &\text{s.t. } \sum_{i=1}^{B} \alpha_i = rB,\ \alpha_i \in [0, \log_2 3]\ \forall i, \end{aligned} \tag{8}$$

where $r \in [0, \log_2 3]$ is a chosen embedding rate in bpp. A general solution is given in Appendix B.

Granting the Warden and the MDS access to the same detector d makes the MDS the optimal batch sender—it minimizes the power of the Warden's most powerful detector.

E. Discussion

Our setup assumes the actors are omniscient. Among other things, the Warden knows the steganographic method used by the sender, the payloads α_i possibly embedded in each image, the communication rate r, and the bag size B. Moreover, the sender and the Warden share the same single-image detector. While it is certainly of interest to study more relaxed setups and perhaps even probabilistic strategies within game theory, such scenarios would require adopting and justifying additional models on how accurately the Warden can estimate the payloads α_i, on the nature of the mismatch between the detectors, etc. The fact that our conclusions regarding the bag gain based on the simplified setup do capture trends observed in real-life situations testify to their relevance.

Having said this, we wish to point out to the reader that the bag gain has been observed in experiments under much more relaxed conditions, including different pooling strategies, mismatched and qualitatively different detectors built using various machine-learning paradigms, and when the Warden needs to estimate the embedded payloads from the images at hand. The reader is advised to inspect Section 7 in [23] for more details.

IV. Response curve model

In order to analyze the trends of detectability w.r.t. the bag size B and possibly the communication rate r, we must somehow obtain a model of $\Delta^2(\mathbf{X})$ over bags since $\mathbf{X}$ has an underlying distribution. We must be careful with our modeling assumptions to preserve the essential complexities of Eq. (1)

so that the bag gain can properly manifest. Due to the form of the deflection $\Delta^2(\mathbf{X})$ (7), it is sufficient to model the response curves across images, which is easier than modeling natural images and also keeps a tighter connection between the model and practice. In particular, we make the following two assumptions about response curves.

A. Linear response curves

We first assume the response curves are linear in payload

$$\varrho_i(\alpha_i) - \varrho_i(0) = b_i\alpha_i, \tag{9}$$

where $\alpha_i \in [0, \log_2 3]$. This significantly simplifies the problem, permitting a closed-form expression for the payloads α_i embedded by the MDS and its extension in Section VII. Even though the response curves of typical detectors built with machine learning are not linear (see, e. g., Fig. 3 in [23]), they are approximately linear when $\varrho_i(\alpha_i) - \varrho_i(0)$ is small.

B. Binomial model for slopes

Arguably, if all images from the cover source had similar response curves, the MDS would spread payload nearly uniformly, at which point the detectability would need to increase with B from the beginning due to the SRL. The reason for the bag gain is source diversity and the fact that the counts of images that contain very small payloads and those that are embedded nearly fully fluctuate across bags. Thus, in order to simplify the modeling but preserve the essence we adopted a two-valued range for the response curve slopes b_i: $\mathbb{P}(b_i = \varepsilon) = p$ and $\mathbb{P}(b_i = 1) = 1 - p$ where $0 < \varepsilon \ll 1$ and $p \in [0, 1]$. Let C_ε denote the number of response curves with slope ε in a bag of size B. Assuming the images are drawn randomly from the cover source, C_ε follows a binomial distribution on $\{0, 1, \ldots, B\}$.

It is easy to show that if all B images have uniform slope b, the deflection $\sum b^2\alpha_i^2$ is minimal when all images receive uniform payload $\alpha_i = \alpha$. In a bag of two images with different slopes, they both start receiving non-zero payload when embedding a message of any length. More generally via a water filling algorithm (see Appendix B), with increasing rate r all images in the bag start receiving payload until the ones with slope ε saturate at $\log_2 3$. From there, the images with slope 1 absorb the remaining payload.

C. Pooled detector performance measure

The deflection coefficient $\Delta^2(\mathbf{X})$ (7), which depends on ε, r, B, and C_ε, informs us about the the performance of the likelihood ratio detector in a specific bag of images. For fixed ε, r, B, the Receiver Operating Characteristic curve (ROC) of the pooled detector expressing the probability of correct stego bag detection P_D as a function of the probability of false alarm P_FA is the expectation over bags

$$\begin{aligned} P_\mathrm{D}(P_\mathrm{FA}) &= \mathbb{E}[Q(Q^{-1}(P_\mathrm{FA}) - \Delta(\mathbf{X}))] \qquad (10)\\ &= \sum_{k=0}^{\infty} \frac{(-1)^k c_k}{k!} Q^{(k)}(Q^{-1}(P_\mathrm{FA}) - \mathbb{E}[\Delta(\mathbf{X})]), \end{aligned}$$

where c_k is the kth central moment of $\Delta(\mathbf{X}) \triangleq \sqrt{\Delta^2(\mathbf{X})}$ as shown in Appendix A. Keeping only the terms up to $k = 2$ in the sum provides a rather accurate approximation for typical values of our modeling parameters (note that $c_1 = 0$).

In this paper, our reasoning is based on the expectation of the deflection coefficient because it is significantly easier to analyze than the ROC (10). While the expected deflection informs us about the ROC over bags *indirectly* (as seen from (10)), many qualitative properties observed for the expected deflection do propagate to common scalar ROC measures, such as the weighted Area Under the Curve (wAUC) [4].

V. Explaining the bag gain

In this section, we explain the performance trends using the binomial linear model for response curves. We begin by simply assuming that images can hold an arbitrarily large amount of payload. As we progress through this section, we will incorporate more realistic constraints in order to capture which pieces of the model are responsible for certain phenomena we observe in practice.

A. Unbounded embedding capacity

First, we analyze the case of unbounded embedding capacity for all images from the bag. We believe it is useful to start with this case as it 1) clearly captures important trends in the small bag regime, 2) is analytically tractable, and 3) serves to build the reader's intuition as to why a bag gain should occur in the first place. Studying the unbounded case will also help underscore the impact of finite embedding capacity on the observed trends later seen in Section V-B.

Based on Eq. (43) in Appendix B, the MDS payloads for the unbounded case are given, for all i, by

$$\alpha_i = \frac{rB}{b_i^2 \sum_{k=1}^{B} \frac{1}{b_k^2}} = \frac{rB\varepsilon^2}{b_i^2(C_\varepsilon + (B - C_\varepsilon)\varepsilon^2)}, \tag{11}$$

since

$$\sum_{k=1}^{B} \frac{1}{b_k^2} = C_\varepsilon\varepsilon^{-2} + (B - C_\varepsilon). \tag{12}$$

Utilizing (11) and (12), the deflection simplifies to

$$\Delta^2(\mathbf{X}) = \frac{1}{\sigma^2}\sum_{i=1}^{B} b_i^2\alpha_i^2 = \frac{r^2B^2\varepsilon^2}{\sigma^2(C_\varepsilon + (B - C_\varepsilon)\varepsilon^2)}. \tag{13}$$

In this case, the expected deflection becomes

$$\mathbb{E}[\Delta^2(\mathbf{X})] = \frac{r^2B^2\varepsilon^2}{\sigma^2}\sum_{k=0}^{B} \frac{\binom{B}{k}p^k(1-p)^{B-k}}{k + (B-k)\varepsilon^2}, \tag{14}$$

which can be further simplified using Stirling's formula (see, e. g., page 147 in [6]) as $B \to \infty$

$$\begin{aligned} &\binom{B}{pB} \sim 2^{BH_2(p)} \\ \Rightarrow &\binom{B}{pB} p^{pB}(1-p)^{(1-p)B} \sim 2^{BH_2(p)} \times 2^{-BH_2(p)} = 1 \\ \Rightarrow &\qquad \mathbb{E}[\Delta^2(\mathbf{X})] \sim \frac{r^2\varepsilon^2 B}{\sigma^2(p + \varepsilon^2(1-p))}. \qquad (15) \end{aligned}$$

Here, H_2 is the binary entropy function, and $\sim$ means the ratio of both sides tends to 1 as $B \to \infty$.

Figure 2 shows the expected deflection $\mathbb{E}[\Delta^2(\mathbf{X})]$ as a function of the bag size B with the dashed lines drawn to show asymptotic trends. The figure also shows wAUC of Eq (10) as a function of B. For small p, the detectability initially grows due to the SRL because the bags are small and most do not contain any images with slope ε. The growth is steep because it is driven primarily due to payload embedded in images with slope 1. As the bag size increases, however, the detectability starts dropping since the bags are more likely to contain images with small slopes which absorb most of the payload with only a slight contribution to the deflection. The deflection eventually levels off and then linearly increases. This time, the growth is less steep because of the presence of images with slope ε. Thus, the existence of the local maximum and global minimum of expected deflection is fundamentally a consequence of the *SRL switching its growth rate*.

As depicted in Figure 2, the unbounded capacity model predicts two critical bag sizes that depend primarily on p and ε. One is associated with a local maximum, $B_{\max}$, while the other, $B_{\min}$, corresponds to minimal expected deflection. We do not talk about these critical bag sizes as corresponding to bag loss and bag gain *yet* because we define these concepts for the more realistic bounded capacity case using an easily interpretable performance measure (wAUC) in the next section. The closed form for the expected deflection as a function of bag size allows us to study the critical points and obtain insight into the conditions under which the local maximum and the minimum can occur and how they depend on ε and p. Figure 2 tells us that we can then implicitly (but indirectly) draw conclusions about wAUC since the relationships closely transfer as visually portrayed.

Since our model is only defined for positive integers $B \geq 1$ (actual bag sizes), we begin by simplifying the expression in Eq. (14) by using Eq. (15) (the dominant term in the large bag regime) along with the $k = 0$ term (the dominant term in the small bag regime when ε is small) :

$$\mathbb{E}[\Delta^2(\mathbf{X})] \doteq \frac{r^2 B}{\sigma^2}\left((1-p)^B + \frac{\varepsilon^2}{p+\varepsilon^2(1-p)}\right). \tag{16}$$

Notice that Eq. (16) can be defined on the real numbers $B \in \mathbb{R}$. Using mild simplifying assumptions, we can derive closed form approximations for both critical bag sizes. Specifically, using Eq. (16) and the fact that $(1-p)^B = e^{B\ln(1-p)}$, we can approximate the optima by finding solutions to

$$\frac{\partial}{\partial B}\mathbb{E}[\Delta^2(\mathbf{X})] \doteq \left(e^{B\ln(1-p)} + \frac{\varepsilon^2}{p+\varepsilon^2(1-p)}\right) \tag{17}$$

$$+ B\ln(1-p)e^{B\ln(1-p)} = 0. \tag{18}$$

Since $\ln(1-p) < 0$, for small B the term proportional to ε^2 is small compared to the other two terms. Setting $\frac{\varepsilon^2}{(p+\varepsilon^2(1-p))} \approx 0$, we obtain an approximate formula for the first critical bag size corresponding to the local maximum[3]

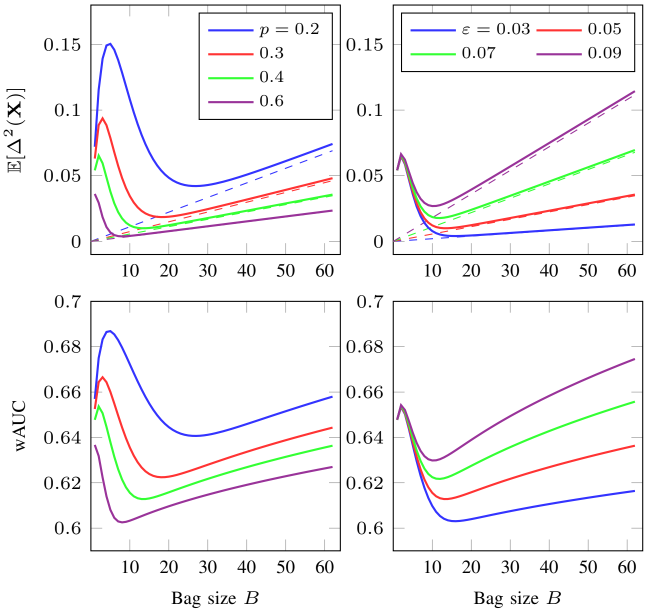

Figure 2. Unbounded capacity model of pooled detector performance. Top row is $\mathbb{E}[\Delta^2(\mathbf{X})]$ as a function of B (dots) with the line (solid) drawn to show asymptotic trends. Bottom row is wAUC of Eq. (10) as a function of B. Left column shows trends w.r.t. p ($\varepsilon = 0.05$ fixed) and right column shows trends w.r.t. ε ($p = 0.4$ fixed). We have $r = 0.3$ and $\sigma^2 = 1$ fixed.

$$0 = e^{B\ln(1-p)}\left(1 + B\ln(1-p)\right)$$

$$\Leftrightarrow B_{\max} \doteq \frac{-1}{\ln(1-p)}. \tag{19}$$

For larger bag sizes, the term proportional to ε^2 cannot be ignored. We rearrange the terms and take log of both sides (keep in mind that $\ln(1-p) < 0$)

$$\frac{\varepsilon^2}{(p+\varepsilon^2(1-p))} = -e^{B\ln(1-p)}\left(1 + B\ln(1-p)\right)$$

$$\Leftrightarrow \quad B\ln(1-p) = \ln\left(\frac{\varepsilon^2}{p+\varepsilon^2(1-p)}\right) \tag{20}$$

$$- \ln\left(-1 - B\ln(1-p)\right). \tag{21}$$

Since the second term on the r.h.s. of this equation is small with respect to the l.h.s., we obtain a first order approximation for the second critical bag size[4]

$$B_{\min} \doteq \frac{\ln\left(\frac{\varepsilon^2}{p+\varepsilon^2(1-p)}\right)}{\ln(1-p)}. \tag{22}$$

From (19), we can deduce that the initial growth associated with the local maximum ceases to manifest with sufficiently large prior probability p of images with small slopes. In particular, $B_{\max} < 1$ for $p \gtrsim 0.63$ in approximate agreement with Figure 2 when working with the exact expected deflection. Additionally, Eq. (22) encapsulates how $B_{\min}$ depends on p and ε (it increases as ε or p decrease). This makes intuitive

[3] The fact that $B_{\max}$ corresponds to a local maximum can be verified by computing the second derivative.

[4] A more precise argument can be made here based on iterative root finding for the equation $B = f(B)$ by showing that $|f'(B)| < 1$ for convergence.

sense as smaller ε means the images can hold larger payload, making the SRL take longer to finish switching its growth rate. Similarly, with a smaller fraction p of such images, it takes larger bags to see their effect on detectability.

B. Bounded embedding capacity

We now show the effect of bounding the embedding capacity to $A = \log_2 3$ bpp and also formally define the bag loss and bag gain. Images with $b_i = \varepsilon$ achieve embedding capacity $\alpha_i = A$ when (c.f. Eq. (11))

$$\frac{rB}{C_\varepsilon + (B - C_\varepsilon)\varepsilon^2} \geq A, \tag{23}$$

which holds iff

$$T := \frac{r/A - \varepsilon^2}{1 - \varepsilon^2} B \geq C_\varepsilon. \tag{24}$$

If $T < C_\varepsilon$, then $\Delta^2(\mathbf{X})$ is given by Eq. (13). However, if $T \geq C_\varepsilon$, we have

$$\alpha_i = \begin{cases} \frac{rB - AC_\varepsilon}{B - C_\varepsilon} & b_i = 1 \\ A & b_i = \varepsilon \end{cases} \tag{25}$$

and so

$$\Delta^2(\mathbf{X}) = \frac{C_\varepsilon \varepsilon^2 A^2}{\sigma^2} + \frac{(rB - C_\varepsilon A)^2}{\sigma^2 (B - C_\varepsilon)}. \tag{26}$$

Thus, we have in expectation

$$\begin{aligned} \mathbb{E}[\Delta^2(\mathbf{X})] &= \mathbb{E}[\Delta^2(\mathbf{X})\mathbf{1}_{T<C_\varepsilon}] + \mathbb{E}[\Delta^2(\mathbf{X})\mathbf{1}_{T\geq C_\varepsilon}] \\ &= \frac{1}{\sigma^2}\left[r^2 B^2 \varepsilon^2 \sum_{k=\lfloor T \rfloor + 1}^{B} \frac{\binom{B}{k} p^k (1-p)^{B-k}}{k + (B-k)\varepsilon^2} \right. \\ &\quad + \sum_{k=0}^{\lfloor T \rfloor} \binom{B}{k} p^k (1-p)^{B-k} \\ &\quad \left. \times \left(k\varepsilon^2 A^2 + \frac{(rB - kA)^2}{B - k} \right) \right]. \end{aligned} \tag{27}$$

In Figure 3, we show the wAUC of Eq. (10) (instead of expected deflection) for various combinations of ε, r, p since we intend to contrast the performance of the model with real life detectors.

1) Bag gain and bag loss: While the exact trend of wAUC w.r.t. B depends on ε, r, and p, one can roughly say that (ignoring for now the small oscillations commented upon in the next section): 1) wAUC can either grow right from $B = 1$, or 2) grow, reach a local maximum, decrease, reach a global minimum (bag gain), and then increase, or 3) exhibit a global minimum without the initial increase. Fundamentally, the local maximum and the global minimum of wAUC are due to the varying statistical makup of small bags as already commented for the unbounded capacity case. Eventually, for large enough B wAUC will approach 1. How fast this happens depends on whether large enough bags contain enough images with small slopes to avoid embedding substantial payload in images with a large slope. This occurs approximately when $p \log_2 3 > r$, at which point wAUC approaches 1 only very slowly, depending on the value of ε. This is why the global minimum appears quite shallow for some combinations of the parameters.

Formally, we define the bag gain γ as the maximum decrease in a chosen detectability measure the batch sender can enjoy by bagging. Since we use wAUC,

$$\gamma = \max_{B \geq 1} \left[\text{wAUC}(1) - \text{wAUC}(B) \right], \tag{28}$$

where wAUC(B) is the wAUC of the pooled detector on bags of size B. Notice that the bag gain can be observed for most combinations of the parameters in Figure 3 but disappears for large enough rates and for larger ε. This is intuitively correct as larger rates force the detectability to grow faster as do larger values of ε.

Besides the global minimum corresponding to the bag gain, wAUC as a function of B may exhibit a local maximum for small bags (for $p \leq 0.3$ in the figure). When the bag gain is positive ($\gamma > 0$), we define bag loss as

$$\nu = \max_{B_{\text{Gain}} > B \geq 1} \left[\text{wAUC}(B) - \text{wAUC}(1) \right], \tag{29}$$

where B_{Gain} is the bag size corresponding to the bag gain.[5] In words, bag loss is the increase in detectability when the sender selects the worst bag size instead of the optimal B_{Gain}. Based on our definition, bag loss is not defined if there is no positive bag gain. Similar to the bag gain, bag loss may not manifest for certain combinations of the parameters.

2) Local oscillations: As shown in Figure 3, the wAUC experiences a transient oscillating / periodic behavior for smaller bag sizes, which can be explained by analyzing expected deflection. The oscillations appear when considering images with bounded capacity and are ultimately due to the quantization of T when computing the bounds for the sums in Eq. (27). In particular, since ε^2 is small, $T \approx rB/A$ which implies $\lfloor T \rfloor$ increments whenever $B \approx Ak/r$ for some positive integer k. In other words, $\lfloor T \rfloor$ is fixed for intervals of length A/r. For example, for $r = 0.3$ we have $A/r \approx 5$ which is approximately the period shown in the corresponding plot in Figure 3. Within each interval, $\mathbb{E}[\Delta^2(\mathbf{X})]$ (and wAUC) changes in a continuous manner and may contain local optima due to the upper sum $\mathbb{E}[\Delta^2(\mathbf{X})\mathbf{1}_{T<C_\varepsilon}]$.

3) Trends w.r.t rate: In the unbounded case, we see that the expected deflection is linearly proportional to r^2 (15). However, in the bounded capacity case, the rate has a non-trivial affect on the performance curves (in terms of wAUC) as seen in Figure 3 and, in particular, the location of B_{Loss} and B_{Gain}. For example, as r increases we see that B_{Gain} decreases for $\varepsilon = 0.06$ and $p = 0.2$, but B_{Gain} increases for $\varepsilon = 0.02$ and $p = 0.4$. Note that if $T < 1$, then approximately $rB < A$ which makes Eq. (27) degenerate to the unbounded model Eq. (14).

VI. Observing trends in real images

In this section, we contrast the trends in detectability w.r.t. bag size from experiments with real images and detectors with those obtained from the model. We measure the performance with wAUC. First, we describe our experimental setup, including the dataset and a single-image detector used by some batch senders and for pooled steganalysis. As mentioned in [23],

[5] B_{Loss} will denote the bag size corresponding to the bag loss.

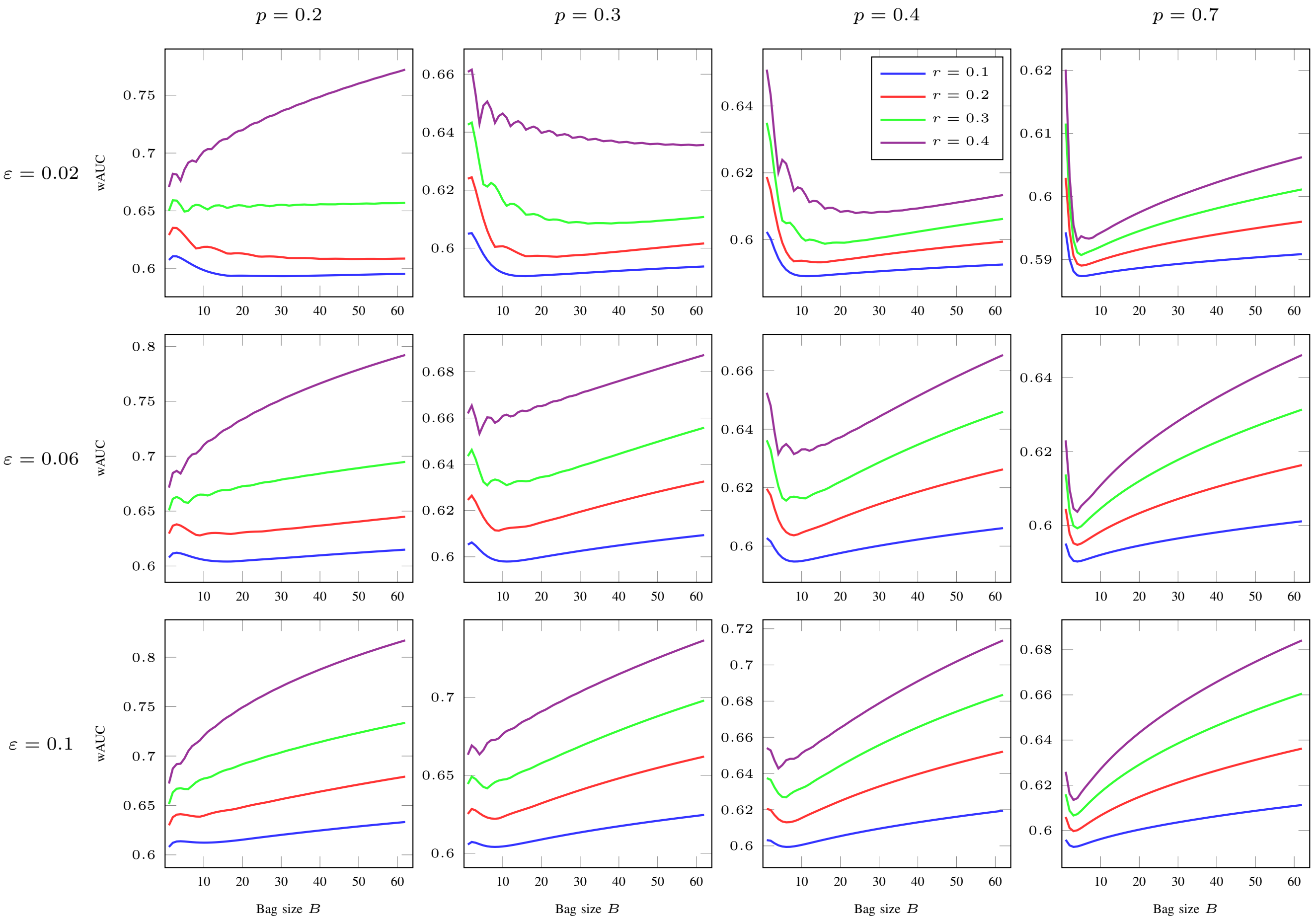

Figure 3. Bounded capacity model of optimal pooled detector performance (wAUC) as a function of B for various combinations of ε, r, p. Rows correspond to fixed ε, columns correspond to fixed p, and colors correspond to fixed r.

the embedding algorithm (whether cost-based or model-based) does not have a significant effect on the bag gain manifesting, so we limit our experiments to the cost-based HILL [18]. All experiments were done on the image dataset ALASKA II [4] developed as in [4] without the final JPEG compression step.[6] We consider two disjoint subsets of ALASKA II images denoted split1 and split2, containing 25,000 images each. Split1 is used to train the shared single-image detector and Warden's pooled detector while split2 is used to assess the performance of batch senders.

The detector-aware senders use a single-image detector d in the form of an SRNet [2] pre-trained on ImageNet with the binary task of steganalyzing J-UNIWARD [8] (the so-called JIN pre-training exactly as described in [3]). The refinement to detect HILL was done on a diverse stego source created using split1 with relative payloads randomly drawn from the uniform distribution on the set of relative payloads

$$\mathcal{P} = \{0.05, 0.1, 0.2, \ldots, 1.4, 1.5\}. \tag{30}$$

In particular, split1 was partitioned into further subsets of 22k, 1k, and 2k images for training, validation, and testing, respectively. The detector-aware senders use the logit as the detector's response.

The Warden is given the sender's detector d for steganalysis. She is also assumed clairvoyant and given the knowledge of the payloads α_i. The reader is referred to [23] for a comprehensive analysis of the situation when the Warden estimates α_i from the images at hand and when she trains her own single-image detector that is possibly different as well as trained on a different dataset from the same source. In particular, as shown in this prior art, the trends of detectability vs. bag size appear to be robust and unaffected by Warden's choices.

Three batch senders are tested: the Image Merging Sender (IMS) and the detector-aware Shift Limited Sender (SLS) and MDS. The IMS treats each bag as one big image and lets the given stego algorithm decide what payload chunk each image will hold. The SLS finds the payloads by requiring that the embedding induces the same shift in the detector response. The MDS, which is described in Section III-D, was implemented using a projected gradient descent method to find optimal payloads since response curves for real images are non-linear. We refer the reader to the original publication for more details [23]. We did not include the batch sender proposed in [22] because it is equivalent to the IMS with an embedding scheme adjusted as in Gaussian embedding.

[6] The authors note that the bag gain was observed on other datasets, such as BOSSbase [5] and BOWS2 [1] (not shown in this paper).

The optimal pooled detector described in Section III-C was used to analytically study and explain the bag gain trends; however, such a pooler is infeasible in practice due to the difficulty of estimating the parameters of the distributions in (6). Thus, all experiments on real images use the LRT pooler, π_{LRT}, as thoroughly studied in [23]. The Warden tests whether the detector output for the ith image of the bag is consistent with the distribution of the detector f_{α_i} on stego images all embedded with the same relative payload α_i:

$$\begin{aligned} \mathcal{H}_0: &\quad d(Y^{(i)}) \sim f_0 \quad \text{for all } i \\ \mathcal{H}_1: &\quad d(Y^{(i)}) \sim f_{\alpha_i} \quad \text{for all } i \end{aligned} \tag{31}$$

with the optimal detector being the log-likelihood ratio

$$\pi_{\mathrm{LRT}}(\mathbf{Y}) = \sum_{i=1}^{B} \log \frac{f_{\alpha_i}\left(d(Y^{(i)})\right)}{f_0\left(d(Y^{(i)})\right)}. \tag{32}$$

The distributions f_{α_i} are estimated empirically using the test set of split1.[7] Both spreading and pooling is done on split2.

We note that [23] investigated three other pooled detectors, including situations when the Warden trained the detector on a different dataset and/or used a different neural architecture or even a qualitatively different detector, such as a rich model. The bag gain was generally observed under all circumstances. For a comprehensive look at bag gain trends across poolers in general, we refer the reader to [23].

A. Trends seen in ALASKA II

Our focus is on trends of detectability w.r.t. bag size B, rate r and for multiple batch senders. Figure 4 shows the detection performance of the LRT pooler π_{LRT}. For each fixed B, r, and sender, we independently form 2000 bags sampled without replacement from split2. The wAUC is computed from the ROC formed by the 2000 samples of bags.

First, notice that all senders exhibit a bag gain, including the detector-agnostic IMS. The bag gain can manifest up to a ~0.15 decrease in pooled detector performance, which can significantly benefit the steganographer in practice. Second, the initial decrease in performance engages quickly so even using bags of size 5, e.g., as opposed to using a single-image is signficiantly advantageous for the steganographer.

Despite the differences between response curves under the binomial model and real image response curves, the trends predicted by our model and shown in Figure 3 provide valuable insight. In particular, the model correctly predicts that for large enough payloads the bag gain disappears. Furthermore, the optimal bag size B_{Gain} increases with decreased rate r except for the smallest value of ε (cover source with images with basically flat response curves). Our model additionally predicts that this increase is smaller in cover sources with fewer hard-to-steganalyze images (smaller p).

One of the clearest differences between IMS and the two detector-aware senders that can be seen in Figure 4 is that the SRL engages a lot sooner for IMS. The main contributing factor is that the two detector-aware senders are more aggressive in utilizing difficult images by embedding them with larger payloads because they are aware of the impact on detectability. Batch senders that are even less aggressive than IMS will eventually not exhibit the bag gain. In the extreme case of a batch sender that assigns the same payloads to all images, the detectability will monotonically increase as per the large bag regime's SRL. In Section VII, we will explain this behavior from a model by introducing a family of batch senders parametrized by a scalar parameter (the Hölder sender) that encompass the uniform sender, the SLS, and MDS.

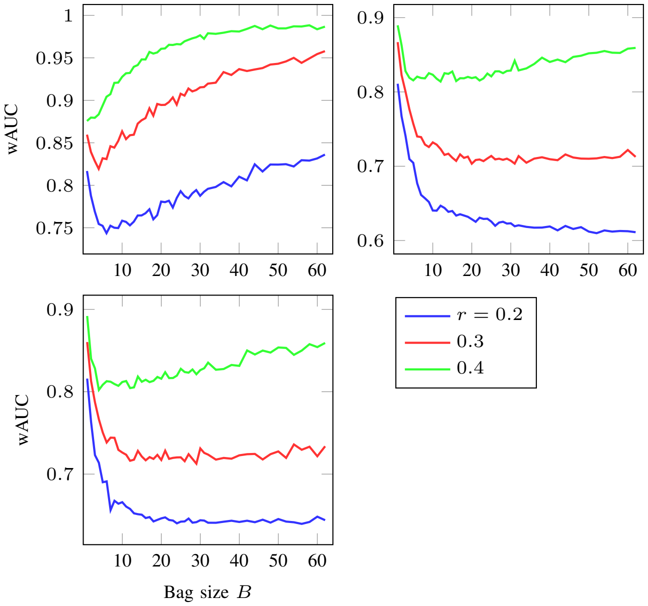

Figure 4. Trends in the performance of π_{LRT} across batch senders for ALASKA II (top left IMS, top right SLS, bottom left MDS). For the lower payloads of SLS and MDS, the SRL requires a much larger bag size to take effect.

Finally, as seen in Figure 3 for some combinations of ε, p, and r our model predicts oscillations in wAUC for small bag sizes and an initial bag loss (local maximum in wAUC) for very small bag sizes. While these higher-order effects were not observed in our experiments on ALASKA II, in the next section we demonstrate that they are real phenomena that can manifest in other datasets with the right diversity of images.

B. Bimodal ALASKA II

As commented on in the previous section, our binomial model of slopes predicts that, for small bag sizes and certain combinations of ε, p, and r, wAUC should exhibit a local maximum, the bag loss, and oscillations that decay with larger bag sizes. Such higher-order effects are not seen in our experiments because the real distribution of response curves in images from ALASKA II is not close enough to the binomial model of slopes.

In order to investigate whether these phenomena can manifest for real images, we construct multiple versions of "bimodal" ALASKA II consisting of two groups of images: 1) easy-to-steganalyze images with steep response curves and 2) hard-to-steganalyze images with almost flat response curves. An approximately bimodal distribution can realistically occur, for example, in a landscape photographer's portfolio when the

[7] Using scipy's gaussian_kde function

Table I
PARAMETERS FOR NARROW AND WIDE ε-M BINNING ON SPLIT2. THE $\#\varepsilon$ AND $\#M$ ARE THE NUMBER OF RCS FROM SPLIT2 THAT QUALIFY AS ε/M-TYPE. THE AVG ε AND AVG M ARE THE SAMPLE AVERAGES OF THE 'SLOPES AT $\alpha = 0$' FOR ε/M-TYPE RCS, RESPECTIVELY.

	ℓ_ε	u_ε	ℓ_M	u_M	$\#\varepsilon$	$\#M$	avg ε	avg M
Narrow	0	0.08	0.8	3.2	873	1767	0.016	1.564
Wide	0	0.15	0.5	9.5	1358	9149	0.027	3.167

majority of the source is low ISO images, which would be the case of images taken during daylight, while the remainder is high ISO images taken during the night (astrophotography).

We propose the following stochastic procedure based on rejection sampling to enforce a distribution of slopes on ALASKA II that more closely matches our model. This will also allow us to parameterize the dataset by p, a source diversity parameter, so we can feasibly observe trends across sources with a varying proportion of easy-to-steganalyze and hard-to-steganalyze images.

First, we perform what we call "ε-M binning" on ALASKA II. Given four non-negative constants $\ell_\varepsilon \le u_\varepsilon \le \ell_M \le u_M$, we say image X has an ε-type RC ϱ_X if for all $\alpha \in \mathcal{P}$, $\ell_\varepsilon \alpha \le \varrho_X(\alpha) - \varrho_X(0) \le u_\varepsilon \alpha$. Similarly, we say image X has an M-type RC if for all $\alpha \in \mathcal{P}$, $\ell_M \alpha \le \varrho_X(\alpha) - \varrho_X(0) \le u_M \alpha$. These response curves can be thought of as having a kind of "Lipschitz" condition on their derivatives since the ε-type, e.g., are contained within the cone formed by $\ell_\varepsilon \alpha$ and $u_\varepsilon \alpha$. Next, when Alice is forming her bag from this artificial ALASKA II source, she samples (uniformly) an image with ε-type RC with probability p and samples an image with M-type RC with probability $1 - p$. In the previous sections, our binomial model had $M = 1$ fixed for notational simplicity in the derivations; note that the equations in Section V can be easily generalized to consider arbitrary $M > \varepsilon$.

As seen in Figure 6, if we take ε (and M) as the sample average of the 'RC slopes at $\alpha = 0$' of the ε/M-type RCs where the slope is estimated using the first three points

$$\hat{b}_X = \frac{1}{2}\left(\frac{\varrho_X(0.05) - \varrho_X(0)}{0.05 - 0} + \frac{\varrho_X(0.1) - \varrho_X(0)}{0.1 - 0}\right), \quad (33)$$

we observe similar behaviors in the size of the bag gain (the maximal drop in detectability) and even the frequency of local oscillations, and the value of B where the SRL regime roughly begins (seen by the decay of the amplitude oscillations and increase in detectability for increasing B). See the values of 'avg ε/M' in Table I for these sample averages of slopes.

In Figure 7, observe that there is still a bag loss even when the rejection sampling uses much wider ε/M bins. This confirms the robustness of a bag loss occurring even in a source that contains a diverse spectrum of real image response curves (which is very different from binomial linear response curves). If easy-to-steganalyze images are common and hard-to-steganalyze images are rare in an image source, it is important to be aware that a bag loss will likely manifest.

VII. GENERALITY OF THE BAG GAIN

In order for the bag gain to occur, the batch sender must prefer embedding more payload in hard-to-steganalyze images and less payload in easy-to-steganalyze images. In the case of the binomial model, this is equivalent to the batch sender putting more payload in images with near flat response curves. This property holds true for the detector-agnostic IMS, Distortion-Limited Sender (DiLS), and Detectability-Limited Sender (DeLS) studied in [21], as well as the detector-aware SLS and MDS. The IMS / DiLS / DeLS are not as extreme as the detector-aware senders since they are not designed to explicitely make use of response curves. However, their spreading still correlates with this preference since content-adaptive steganographic schemes put more payload in regions of complex content which give difficulty to detectors. In situations where the steganographers and Warden are knowledge limited as in [23], even a weak preference to embed more in hard-to-steganalyze images (w.r.t. the Warden's detector) can cause the bag gain to manifest.

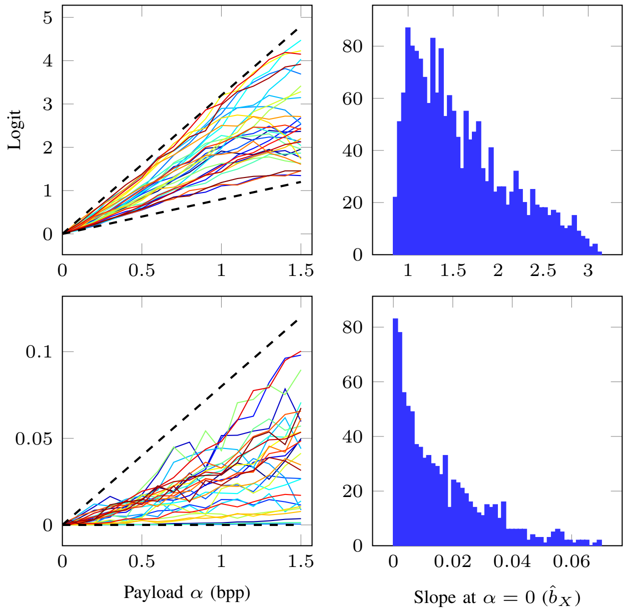

Figure 5. Example of narrow ε-M binning (30 randomly sampled example response curves) and the histograms of initial slopes at $\alpha = 0$ in bimodal ALASKA II. Top row is M-type and bottom row is ε-type.

In this section, we introduce a parametrized family of senders with the parameter controlling how aggressively the sender assigns the payload based on the response curves, including the case when the payload is spread uniformly across all images. By varying this parameter, we can show that the bag gain eventually disappears for sufficiently weak preferences for embedding more payload in harder images.

The Hölder sender can be thought of as a generalization of the MDS (11) as it assigns the following payloads to images:

$$\alpha_i = \frac{rB}{b_i^q \sum_{k=1}^{B} \frac{1}{b_k^q}}, \quad (34)$$

where $q \in \mathbb{R}$ is a parameter. For $q = 2$ and $q = 1$, this sender corresponds to the MDS and SLS, respectively. When $q = 0$, the payload is spread uniformly across all images.

Following the same steps as in Section V-B, the deflection

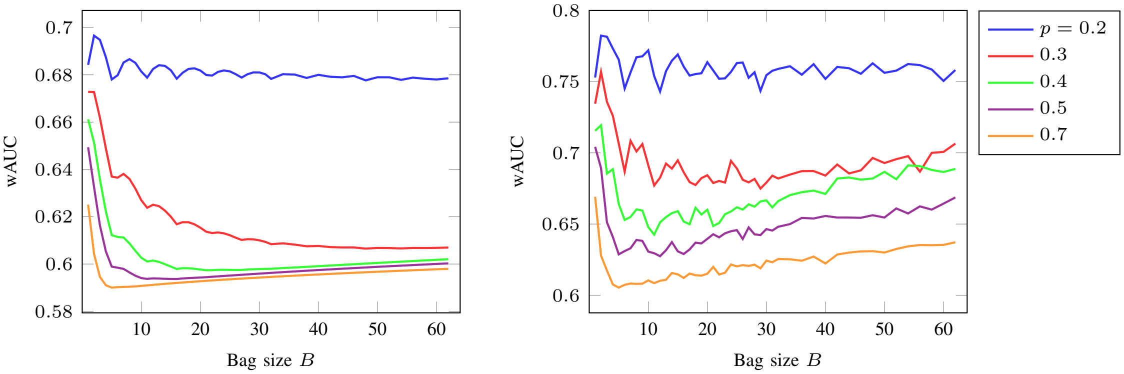

Figure 6. Optimal pooled detector performance for binomial model (left) vs π_{LRT} for bimodal Alaska II (right) using the narrow range parameters given in Table I. The binomial model uses $\varepsilon =$'avg ε' and $M =$'avg M' as given in the table and explained in the text. Rate $r = 0.3$ bpp is fixed.

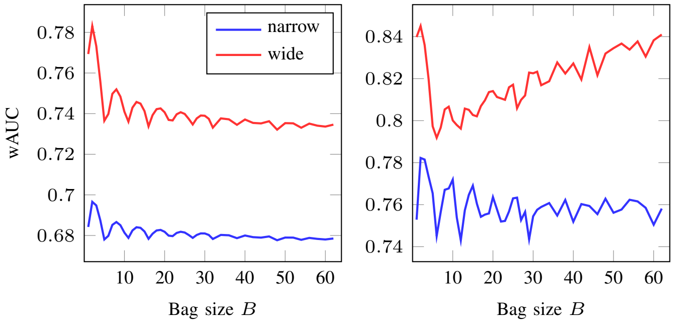

Figure 7. Optimal pooled detector performance for binomial model (left) vs π_{LRT} for bimodal ALASKA II (right) comparing narrow and wider ε-M binning. A bag loss still occurs even for wider bins pointing to the robustness of the phenomenon. Parameters $p = 0.2$ and $r = 0.3$ are fixed.

coefficient for the Hölder sender is

$$\Delta^2(\mathbf{X}) = \begin{cases} \frac{r^2B^2\varepsilon^2}{\sigma^2}\left(\frac{C_\varepsilon+\varepsilon^{2q-2}(B-C_\varepsilon)}{(C_\varepsilon+\varepsilon^q(B-C_\varepsilon))^2}\right) & T_q < C_\varepsilon \\ \frac{C_\varepsilon\varepsilon^2(\log_2 3)^2}{\sigma^2} + \frac{(rB-C_\varepsilon\log_2 3)^2}{\sigma^2(B-C_\varepsilon)} & T_q \geq C_\varepsilon \end{cases} \tag{35}$$

where $T_q = \frac{rB}{(1-\varepsilon^q)\log_2 3} - \frac{\varepsilon^q B}{1-\varepsilon^q}$. Substituting (35) into Eq. (10), we can compute the bag gain γ as given by Eq (28). Figure 8 shows γ as a function of the exponent q for a range of the parameters p (left) and ε (right). As q decreases from $q = 2$ (MDS), the payload assignment is less polarized and the bag gain starts decreasing. It eventually becomes zero and is always zero for uniform spreading ($q = 0$).

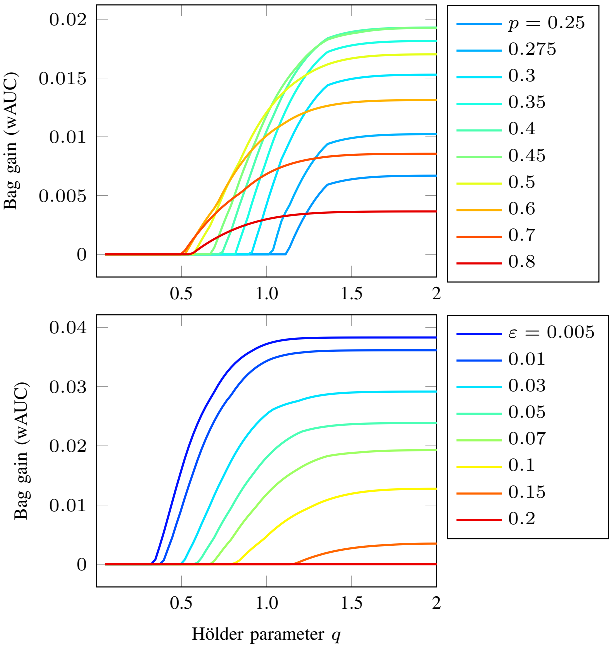

Figure 8. Bag gain measured in wAUC of the bounded capacity model across the family of Hölder spreaders for $q \in [0, 2]$ and various p and ε. The left figure has $\varepsilon = 0.07$ fixed and the right figure has $p = 0.5$ fixed. Both have $r = 0.3$ bpp.

VIII. Relationship to prior work

In this section, we contrast our contribution with previous work [22] that studies optimal bag size in batch steganography. We do so in order to highlight the differences and also to briefly discuss possible future directions by combining both approaches. The authors of [22] extended Gaussian Embedding (GE) to batch steganography. Granting the Warden the knowledge of the underlying distributions, a closed-form expression has been derived for the performance of Warden's likelihood ratio test in a specific collection of bags of images. This was used to implement a batch sender with an adaptive batch size called adaBIM.

The first and the main difference between their work and this paper is the lack of pooled steganalysis. As formulated in the original work of Ker [11], if the steganographer is allowed to spread payload to multiple images, the steganalyst is free to pool evidence from the same multitude of images to reach the conclusion about whether steganographic communication is taking place. In other words, batch steganography needs pooled detectors for proper security assessment. The authors use a performance measure, which is the minimal total detection error P_{E} under equal priors of a single-image detector that distinguishes between the cover source and a stego source whose images contain variable payload "tags" determined by partitioning the dataset into batches and applying GE version of an existing embedding algorithm to the union of all images from the bag to obtain the tags. A pooled detector needs to consider the variability of images in bags, which would necessitate adopting a meta-model on the source. In the case of the GE, it would likely have to be a distribution on the product of cover pixel variances, which opens the possibility to use a similar binomial model within the context of GE. We plan to investigate this direction in the future.

Furthermore, the effect of bag size in [22] is only studied in asymptotic limits of zero or infinite payloads (Theorem 2). For small payloads, the optimal single-image source detector has highest detection error P_{E} when the bag size is equal to the entire image dataset. For large payloads, the highest P_{E} occurs when payloads are assigned using bag size 1. This theorem thus only hints at the existence of optimal bag size w.r.t. P_{E} and a fixed set of bags. The optimal bag size w.r.t. a single-image detector observed in experiments is merely discussed in words without quantitative results.

In contrast, the approach taken in this paper allowed us to relate all essential aspects of a steganographic channel—the cover source diversity, detector response, payload, and bag size—to security under pooled steganalysis. We also believe that working with detector output models leads to a tighter correspondence between the detectability derived from the model and the one obtained experimentally. After all, the model correctly predicts completely new phenomena, such as the bag loss and local oscillations in the small bag regime.

IX. Conclusions

In batch steganography, the secret payload is spread among multiple cover images forming a bag. Within the context of content-adaptive steganography, many batch senders were proposed and studied in the past, such as the image merging sender [21], [22] and the deflection/distortion limited senders [21], as well as two detector-aware senders, the shift limited sender and the minimum deflection sender [23]. When a fixed relative payload is communicated in each bag, batch senders that embed larger payloads in difficult-to-steganalyze images and smaller payloads in easy images exhibit similar trends in terms of detectability vs. the bag size. In this paper, we analyze these trends from the simplest model that captures their essence by considering only two types of images that are "easy" and "difficult" to steganalyze. While the trends depend on the cover source diversity, detector response characteristics, batch sender, and the communication rate, our work offers a simple intuitive explanation.

Assuming that difficult images that can hold large payloads are rare, as the bag size increases, initially the detectability as measured with pooled detectors *increases* due to square root law because only a small fraction of bags contains the difficult images that can carry large payloads without triggering a detector – the square root law thus engages based on embedding primarily in easy images. Once the bag size becomes large enough to contain difficult images with high probability, they hold most of the payload and the detectability begins to *decrease*. Due to the square root law, the detectability eventually levels off, reaching a global minimum, and once more increases but at a speed *slower* than the initial rise depending on the ratio of easy and difficult images in the cover source and the communication rate. The maximum initial rise in detectability is called the *bag loss* while the global minimum corresponds to a *bag gain*. Both phenomena essentially manifest because the average statistical make up of bags differs between small and large bags, which affects how the square root law engages.

While the bag gain was observed experimentally in previous art [23], it was a mere experimental fact that was left unexplained. Our work provides theoretical insight into the manifestation of the bag gain and quantifies how it depends on cover source diversity, detector response, batch sender, and communication rate. The predicted trends closely match experiments with real images. The predicted bag loss, together with some higher-order oscillations, are experimentally confirmed in datasets with suitable diversity. Furthermore, we provide evidence that these phenomena manifest for batch senders that generally assign payloads based on detectability of embedding in individual images sufficiently strongly as bag loss and gain are not observed for uniform batch senders.

X. Acknowledgements

The authors would like to thank Yassine Yousfi for useful discussions and for helping start this research direction. The work on this paper was supported by NSF grant No. 2028119.

Appendix

A. ROC for pooled detector

A pooled detector makes a decision on bags—either it contains cover or stego images. Since the images from each bag are randomly selected from a cover source, some bags will be easier to detect than others, depending on the value of the deflection coefficient Δ^2. In this section, we derive an expression for the ROC of the pooled detector over bags based on the distribution of the deflection coefficient.

For a fixed false-alarm P_{FA}, the probability of correct stego bag detection is

$$P_{\mathrm{D}}(P_{\mathrm{FA}}) = \mathbb{E}[Q(Q^{-1}(P_{\mathrm{FA}}) - \Delta)], \tag{36}$$

the expectation taken over bags. In this paper, Δ is discrete, attaining values from a finite set $\mathcal{D}$. The derivation below, however, is also valid for a continuous-valued Δ. Let $p_\Delta(x)$, $x \in \mathcal{D}$, be the probability mass function of Δ and let $\mu = \mathbb{E}[\Delta]$. Then, using Taylor expansion of $Q(x)$ at $Q^{-1}(P_{\mathrm{FA}}) - \mu$ with the Lagrange form for the remainder, the expected ROC (36) can be written as

$$\begin{aligned}
P_{\mathrm{D}}(P_{\mathrm{FA}}) &= \sum_{\mathcal{D}} Q(Q^{-1}(P_{\mathrm{FA}}) - x) p_\Delta(x) \\
&= \sum_{\mathcal{D}} \Big[\sum_{k=0}^{n-1} \frac{(\mu - x)^k}{k!} Q^{(k)}(Q^{-1}(P_{\mathrm{FA}}) - \mu) \\
&\quad + \frac{(\mu - x)^n}{n!} Q^{(n)}(Q^{-1}(P_{\mathrm{FA}}) - x^*) \Big] p_\Delta(x) \\
&= \sum_{k=0}^{n-1} \frac{(-1)^k c_k}{k!} Q^{(k)}(Q^{-1}(P_{\mathrm{FA}}) - \mu) + R_n \qquad (37)
\end{aligned}$$

where c_k is the kth central moment of Δ, $x^* \in (\mu, x)$, and

$$R_n = \frac{(-1)^n c_n}{n!} Q^{(n)}(Q^{-1}(P_{\mathrm{FA}}) - x^*) \tag{38}$$

is the Lagrange remainder. Note that $Q^{(n)}(x) = \frac{1}{\sqrt{2\pi}}p_n(x)e^{-x^2/2}$ with $p_n(x)$ being the statistician's Hermite polynomial (the physicist's Hermite polynomial is $H_n(x) = 2^{n/2}p_n(\sqrt{2}x)$). Using Cramér inequality [10] for Hermite function defined using physicist's Hermite polynomials $\Psi_n(x) = (2^n n!\sqrt{\pi})^{-1/2}e^{-x^2/2}H_n(x) \leq \pi^{-1/4}$ for all x and all n, it is straightforward to show that

$$|R_n| \leq \frac{c_n}{n!}\sqrt{\frac{n!}{2\pi}} = \frac{c_n}{\sqrt{2\pi n!}}. \tag{39}$$

B. General form of the MDS

Let $r \in [0, \log_2 3]$ be a chosen embedding rate in bpp. Optimal payloads for the MDS are found by minimizing $\Delta^2(\mathbf{X})$ s.t. $\sum_{i=1}^{B}\alpha_i = rB$ and $\alpha_i \in [0, A_i]$ $\forall i$ where $A_i \leq \log_2 3$ is the embedding capacity of the ith image (accounting for wet pixels [7]). The Lagrangian has the form

$$\begin{aligned}\mathscr{L} = {} & \sum_{i=1}^{B} b_i^2\alpha_i^2 - \lambda\left(\sum_{i=1}^{B}\alpha_i - rB\right)\\ & - \sum_{i=1}^{B}\ell_i\alpha_i - \sum_{i=1}^{B}u_i(\alpha_i - A_i),\end{aligned} \tag{40}$$

where ℓ_i and u_i are KKT multipliers that satisfy the lower and upper inequality constraints on α_i, respectively. To be a stationary point, the tuple $(\alpha_1, \ldots, \alpha_B)$ must satisfy

$$\alpha_i = 0,\ \frac{\lambda}{2b_i^2},\ \text{or } A_i\ ;\ \forall i. \tag{41}$$

Let $\mathcal{L}$ and $\mathcal{U}$ denote the sets of indices for which $\alpha_i = 0$ or A_i, respectively. Let $\mathcal{I} = (\mathcal{L} \cup \mathcal{U})^{\mathrm{c}}$ be the set of remaining indices where $0 < \alpha_i < A_i$. From the payload constraint

$$\begin{aligned} rB &= \sum_{k\in\mathcal{L}} 0 + \sum_{k\in\mathcal{I}}\frac{\lambda}{2b_k^2} + \sum_{k\in\mathcal{U}} A_i\\ \Rightarrow\ \lambda &= \frac{rB - \sum_{k\in\mathcal{U}} A_i}{\frac{1}{2}\sum_{k\in\mathcal{I}}\frac{1}{b_k^2}}\\ \Rightarrow \alpha_i &= \frac{rB - \sum_{k\in\mathcal{U}} A_i}{b_i^2\sum_{k\in\mathcal{I}}\frac{1}{b_k^2}}, \end{aligned} \tag{42}$$

for all $i \in \mathcal{I}$. The optimal payload is found numerically by searching over the combinations of $\mathcal{L}$, $\mathcal{I}$, and $\mathcal{U}$.

Note that when $A_i = \infty$ for all i (unbounded embedding capacity), we have $\mathcal{U} = \mathcal{L} = \emptyset$ and (42) simplifies to

$$\alpha_i = \frac{rB}{b_i^2\sum_{k=1}^{B}\frac{1}{b_k^2}}. \tag{43}$$

References

[1] P. Bas and T. Furon. BOWS-2. http://bows2.ec-lille.fr, July 2007.

[2] 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.

[3] 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, 2021. ACM Press.

[4] 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.

[5] T. Filler, T. Pevný, and P. Bas. BOSS (Break Our Steganography System). http://www.agents.cz/boss, July 2010.

[6] J. Fridrich. *Steganography in Digital Media: Principles, Algorithms, and Applications*. Cambridge University Press, 2009.

[7] J. Fridrich, M. Goljan, D. Soukal, and P. Lisoněk. Writing on wet paper. In T. Kalker and P. Moulin, editors, *IEEE Transactions on Signal Processing, Special Issue on Media Security*, volume 53, pages 3923–3935, October 2005. (journal version).

[8] 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.

[9] 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.

[10] J. Indritz. An inequality for Hermite polynomials. 12(6):981–983, 1961.

[11] 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.

[12] A. D. Ker. Batch steganography and the threshold game. In E. J. Delp and P. W. Wong, editors, *Proceedings SPIE, Electronic Imaging, Security, Steganography, and Watermarking of Multimedia Contents IX*, volume 6505, pages 04 1–13, San Jose, CA, January 29–February 1, 2007.

[13] A. D. Ker. A capacity result for batch steganography. *IEEE Signal Processing Letters*, 14(8):525–528, 2007.

[14] A. D. Ker. Perturbation hiding and the batch steganography problem. In K. Solanki, K. Sullivan, and U. Madhow, editors, *Information Hiding, 10th International Workshop*, volume 5284 of Lecture Notes in Computer Science, pages 45–59, Santa Barbara, CA, June 19–21, 2008. Springer-Verlag, New York.

[15] A. D. Ker. Locally square distortion and batch steganographic capacity. *International Journal of Digital Crime and Forensics*, 1(1):29–44, 2009.

[16] 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.

[17] 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.

[18] 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.

[19] L. Li, W. Zhang, C. Qin, K. Chen, W. Zhou, and N. Yu. Adversarial batch image steganography against CNN-based pooled steganalysis. *Signal Processing*, 181:107920–107920, 2021.

[20] T. Pevný and I. Nikolaev. Optimizing pooling function for pooled steganalysis. In *IEEE International Workshop on Information Forensics and Security*, pages 1–6, Rome, Italy, November 16–19, 2015.

[21] 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.

[22] M. Sharifzadeh, M. Aloraini, and D. Schonfeld. Adaptive batch size image merging steganography and quantized Gaussian image steganography. *IEEE Transactions on Information Forensics and Security*, 15:867–879, 2020.

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

[24] A. Zakaria, M. Chaumont, and G. Subsol. Pooled steganalysis in JPEG: how to deal with the spreading strategy? In *IEEE International Workshop on Information Forensics and Security (WIFS)*, pages 1–6, 2019.

[25] N. Zhong, Z. Qian, Z. Wang, X. Zhang, and X. Li. Batch steganography via generative network. *IEEE Transactions on Circuits and Systems for Video Technology*, 31(1):88–97, January 2021.