Improving Steganographic Security with Source Biasing

Eli Dworetzky, Edgar Kaziakhmedov, Jessica Fridrich

IH&MMSEC 2024

BINGHAMTON UNIVERSITY
STATE UNIVERSITY OF NEW YORK

What is source biasing?

Alice prefers using covers that are more difficult to steganalyze

- textured images
- noisy images (taken at high ISO)
- images in which embedding does not trigger a detector

Alice is **source biasing** if she selects cover images based on this preference

Is it safe?

It depends on

- what the Warden knows about the cover source (Kerckhoffs' principle)
- Warden is commonly given the knowledge of the source

Potentially dangerous for Alice

- if the Warden detects a change in cover source
- can be independent of the message length

Prior art

- Select a subset of a given size from existing datasets on which steganalysis is the least reliable
- Zhang and Wang (IEEE Access 2019) consider impact on source and force cover features to be "typical" in terms of MMD
- Giboulot et al. (TIFS 2023) consider multiple cover sources (of different difficulty) when Warden makes decision based on a single image

Our take

- Algorithm for sampling the cover source should be considered a part of the embedding algorithm
- To detect source change, Warden must pool evidence
  - Cast within batch steganography and pooled steganalysis
- Warden is given the knowledge of the cover source, stego method, and payload

1. Theoretical analysis within a simplified model
   1. biasing morphs Warden's ROC
   2. asymptotic biasing for constant detectability (extension of the square root law)
2. Experimental verification of the phenomena seen in the model

Detector-centric approach

We model the effect of embedding and the source itself through soft outputs of a steganography detector d

- permits formulating steganalysis and source biasing **jointly** through a single hypothesis test
- models are estimable in practice
- we observe a close match between model and experiments on real datasets

Modeling soft output of a detector

Given a bag of n cover images $(X_1, \ldots, X_n)$

$$
\begin{array}{ll}
\text{Cover:} & d(X_i) \sim \mathcal{N}(0,1) \\
\text{Stego:} & d(X_i(\alpha)) \sim \mathcal{N}(b_i\alpha, 1), \quad 0 \le \alpha \le 1
\end{array}
$$

- $X_i(\alpha)$ ... ith cover image embedded with payload α_i
- $b_i \ge 0$ ... **slope** of the detector's response on ith cover image

Source model

Alice's cover source has two types of images: easy and hard.

Hard to steganalyze (complex content / strong noise)

- $b_i = \varepsilon > 0$ (small)
- selected with probability $p \in (0, 1)$

Easy to steganalyze (smooth content / weak noise)

- $b_i = 1$
- selected with probability $1 - p$

Unbiased source $\Longrightarrow$ coin flip with weights p and $1 - p$

Source biasing

Source biasing amounts to biasing the coin flip:

- Alice selects images with $b_i = \varepsilon$ with probability $q \geq p$

$$b_i \sim \mathcal{B}(q)$$

$\mathcal{B} =$ Bernoulli distribution on $\{\varepsilon, 1\}$

$K =$ number of images with slope ε in the bag

- No biasing $\Longrightarrow K \sim \mathrm{Binom}(n, p)$
- Biasing $\Longrightarrow K \sim \mathrm{Binom}(n, q)$

Warden's hypothesis test

Given a bag of images $\mathbf{Y} = (Y_1, \ldots, Y_n)$, $y_i = d(Y_i)$

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

K is the number of “hard” images, α_i payload potentially residing in Y_i

Optimal pooler is the LRT

$$L(\mathbf{b},\mathbf{y}) = \underbrace{\sum_{i=1}^{n} y_i b_i \alpha_i(K) - \frac{1}{2}\sum_{i=1}^{n} b_i^2 \alpha_i^2(K)}_{\substack{\text{LRT of mean-shifted Gaussian} \\ \text{Stego detection}}} + \underbrace{K \log \frac{q}{p} + (n-K)\log\frac{1-q}{1-p}}_{\substack{\text{LRT of binomial r.v.} \\ \text{Source bias detection}}}$$

Warden's pooler's ROC

Given a decision threshold x

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

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

E_0,E_1,V depend on p, q, α, payload spreading strategy, $b_1, \ldots, b_n$
$Q(x)$ Gaussian tail probability function

Bivalued payload spreading

A bag of n covers with slopes

$$(\underbrace{\varepsilon, \dots, \varepsilon}_{k}, \underbrace{1, \dots, 1}_{n-k})$$

Embed α_ε in all ε images and α_1 in images with slope 1

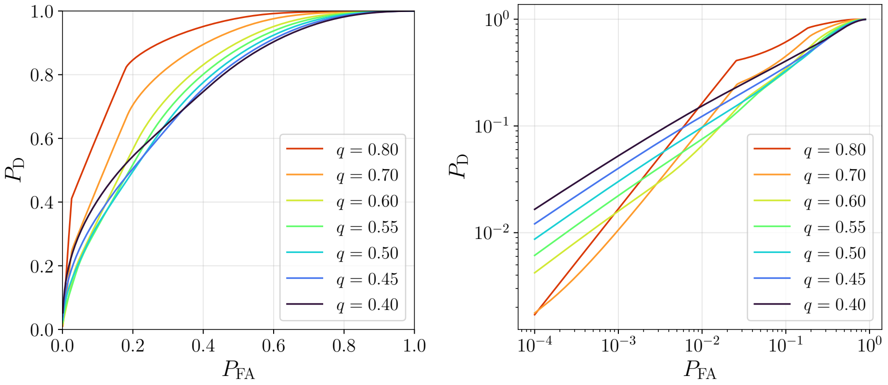
$$(\underbrace{\varepsilon, \dots, \varepsilon}_{\alpha_\varepsilon}, \underbrace{1, \dots, 1}_{\alpha_1})$$

where $(\alpha_\varepsilon, \alpha_1)$ satisfy the payload constraint ($r(n)$ is the rate)

$$r(n)n = k\alpha_\varepsilon + (n-k)\alpha_1$$

Greedy sender embeds in ε images only (if message fits) or embeds them fully and puts the rest in images with $b = 1$

Effect of biasing on ROC

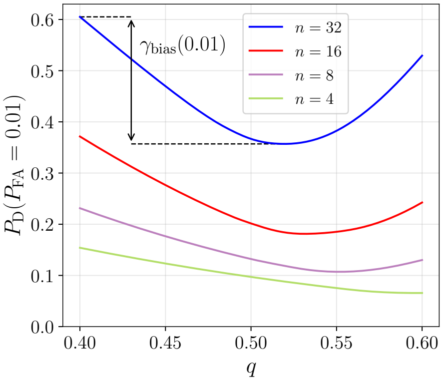


Cover source: $p = 0.4$, $\varepsilon = 0.01$
Bag size $n = 4$
Payload: $\alpha = 1$ (Greedy sender)

Bias Gain

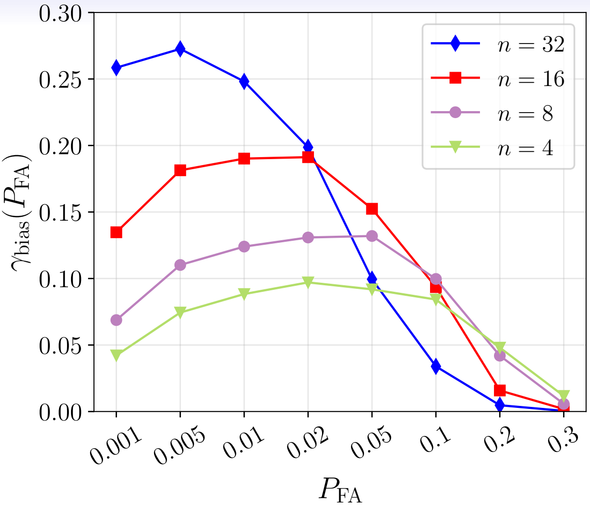
Alice gains in security due to biasing for small P_{FA}:

$$\gamma_{\mathrm{bias}}(P_{\mathrm{FA}}) = P_{\mathrm{D}}(P_{\mathrm{FA}}, p) - \min_{p \leq q} P_{\mathrm{D}}(P_{\mathrm{FA}}, q)$$



Cover source: $p = 0.4$, $\varepsilon = 0.01$ Payload: $\alpha = 1$

Bias gain vs. P_{FA}



Cover source: $p = 0.4$, $\varepsilon = 0.01$ Payload: $\alpha = 1$

Asymptotic Biasing Theorem

Alice adjusts her rate $r(n)$ and bias $q(n)$ as $n \to \infty$

For any bivalued sender

- If both $r(n)$ **and** $q(n) - p$ decay faster than $\frac{1}{\sqrt{n}}$
  - Warden is randomly guessing eventually
- If **at least one of** $r(n)$ **or** $q(n) - p$ decay slower than $\frac{1}{\sqrt{n}}$
  - Alice is caught eventually
- If $r(n)$ **and** $q(n) - p$ decay at critical rate $\frac{1}{\sqrt{n}}$
  - Fixed statistical detectability

Asymptotic Biasing Theorem

Alice adjusts her rate $r(n)$ and bias $q(n)$ so that as $n \to \infty$

$$\underbrace{r^2(n)n \to c_r}_{\text{SRL condition}} \qquad \underbrace{(q(n) - p)\sqrt{n} \to c_p}_{\text{new biasing condition}}$$

For any bivalued sender

- When $c_r = 0$ and $c_p = 0$, asymptotic perfect security
- When $c_r = \infty$ or $c_p = \infty$, asymptotic perfect detectability

Special case for **greedy** and **uniform** senders when $c_r < \infty$ and $c_p < \infty$,

- Warden's **limiting ROC is Gaussian** with deflection

$$d^2_{\text{greedy}} = \frac{c_r \varepsilon^2}{p} + \frac{c_p^2}{p(1-p)} \qquad d^2_{\text{uniform}} = (p\varepsilon^2 + 1 - p)c_r + \frac{c_p^2}{p(1-p)}$$

Experiments on real dataset

ALASKA II (75k grayscale images) divided into four subsets:

1. For training Alice's detector (SRNet) used for spreading payload
2. For training Warden's detector (SRNet)
3. For forming bags to train Warden's pooler (Random forest, $2n+2$ dim. feature extracted from bag $(X_1, \ldots, X_n)$)
4. For forming bags for evaluation

We now have a continuum of slopes

Estimating slopes

Given detector (SRNet) d and cover image X:

- Slope b of X with capacity C bpp was estimated from fully embedded stegos $X(C)$

$$b = \frac{\overline{d(X(C))} - d(X)}{C}$$

- This estimator is reasonable for Greedy sender
- Average taken over 100 simulated embeddings with random stego keys

Greedy sender

Given bag $(X_1, \ldots, X_n)$

- Alice estimates the slopes $(b_1, \ldots, b_n)$ with her detector
- Orders b_i from the smallest to the largest
- Embeds fully one by one till payload is embedded

Continuous biasing

Let F denote CDF of slopes b, and $U \sim \mathrm{Unif}[0,1]$

- Unbiased sampling of images via inverse transform sampling:

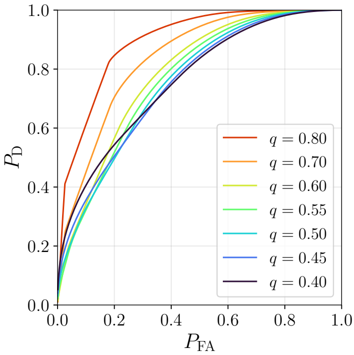
$$F^{-1}(U)$$

- Biased sampling of images via a modification:

$$F^{-1}(G_q^{-1}(U))$$

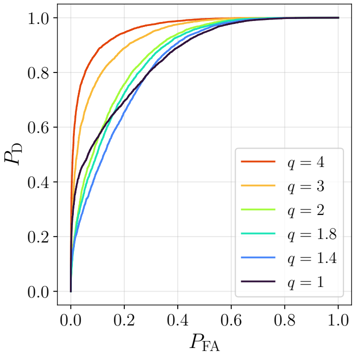
- G_q is the CDF of $\mathrm{Beta}(1/q, 1)$ where $q \geq 1$ is the biasing parameter ($q = 1$ no bias)
- Intuitively, we are sampling quantiles of F non-uniformly with a Beta r.v.

ROCs of Warden's pooler



Bivalued model

$\alpha = 1$, $p = 0.4 \leq q \leq 1$

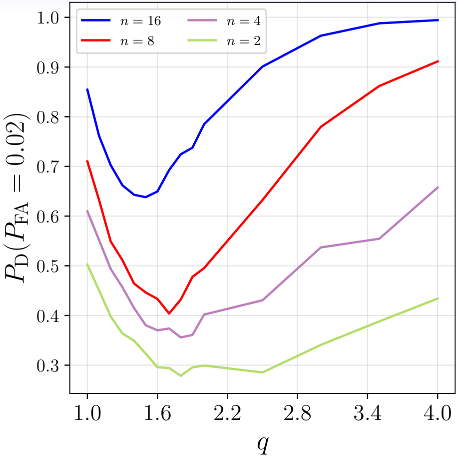


Alaska II (continuous slopes)

$\alpha = 0.5$, $q \geq 1$

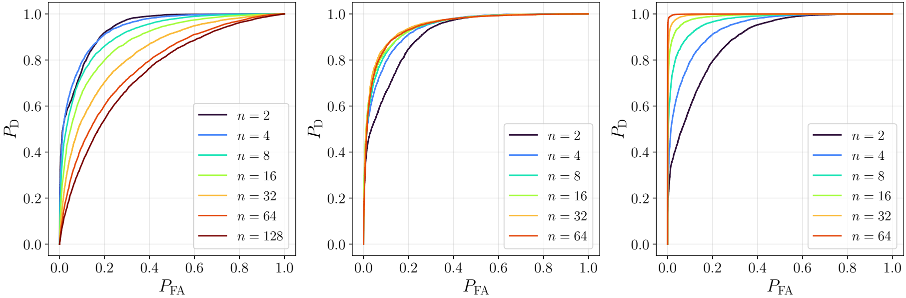
Both: bag size $n = 4$

Bias gain: P_{D} at $P_{\mathrm{FA}} = 0.02$ vs. q



- Bag sizes $n = 2, 4, 8, 16$
- Payload $\alpha = 0.7$

Asymptotic biasing theorem verification



Sub-critical rates

At critical rates

Sub-critical bias
Super-critical rate

Asymptotic trends of the ROC of Warden's pooler on binarized ALASKA II for uniform sender

Conclusions

1. Biasing morphs Warden's pooler's ROC in a complex way
   - for small P_{FA},steganographer gains (smaller P_{D})
   - for large P_{FA}, steganographer loses
2. Bias gain $=$ decrease in P_{D} for small fixed P_{FA}
3. Asymptotic biasing theorem (extension of the SRL)
   - critical scaling for payload and bias for constant asymptotic statistical detectability