

# Capacity Laws for Steganography in a Crowd

Andrew D. Ker

adk@cs.ox.ac.uk

Department of Computer Science, University of Oxford  
Oxford, United Kingdom

## ABSTRACT

A steganographer is not only hiding a payload inside their cover, they are also hiding themselves amongst the non-steganographers. In this paper we study asymptotic rates of growth for steganographic data – analogous to the classical Square-Root Law – in the context of a ‘crowd’ of  $K$  actors, one of whom is a steganographer. This converts steganalysis from a binary to a  $K$ -class classification problem, and requires some new information-theoretic tools. Intuition suggests that larger  $K$  should enable the steganographer to hide a larger payload, since their stego signal is mixed in with larger amounts of cover noise from the other actors. We show that this is indeed the case, in a simple independent-pixel model, with payload growing at  $O(\sqrt{\log K})$  times the classical Square-Root capacity in the case of homogeneous actors. Further, examining the effects of heterogeneity reveals a subtle dependence on the detector’s knowledge about the payload size, and the need for them to use negative as well as positive information to identify the steganographer.

## CCS CONCEPTS

- **Security and privacy** → **Information-theoretic techniques**;
- **Mathematics of computing** → **Coding theory**.

## KEYWORDS

steganography, capacity, multiclass detection, square-root law

## 1 INTRODUCTION

Capacity laws quantify our intuition about rates of growth. In steganography, there was a long-held intuition [1] that a steganographer could not act with constant *rate*: for, unless their steganography were perfect in preserving exactly all the statistics of their cover, then embedding at constant rate must leak ever-growing evidence, and eventually-certain detection. This intuition is quantified by the *Square-Root Laws* [5, 9, 11], which are asymptotic results giving a *critical rate* for the size of the steganographer’s payload  $M$  in terms of the size of the cover medium  $N$ : usually  $M = O(\sqrt{N})$ , or  $M = O(\sqrt{N} \log N)$  when source coding is used. The typical form of these theorems is that (i) when  $M$  grows strictly faster than the critical rate then a detector exists whose performance tends to perfect as  $N \rightarrow \infty$ , and (ii) when  $M$  grows strictly slower than the critical rate then any detector tends to asymptotically random behaviour as  $N \rightarrow \infty$ . The theorems are proved for abstract probabilistic models of covers, starting from identical independent binary elements (that we tend to call ‘pixels’) and then showing that richer models – with arbitrary alphabets, heterogeneity, and bounded dependence – satisfy the same results.

Such theorems are rooted in, and use mathematical results from, binary hypothesis testing. The two hypotheses are

$H_0$  : The actor is not a steganographer,

$H_1$  : The actor is a steganographer,

where the ‘actor’ represents the individual whose pixels are being examined.

But steganography is not a solo activity. The reason why steganographic communication, particularly inside digital media covers, is a plausible covert channel is because there are *many* actors transmitting large amounts of information. Thus not only does the steganographer smuggle their payload inside the pixels (or other elements) of their cover, they also hide themselves amongst a crowd of innocent, non-steganographic, actors.

We should have intuition about this situation: the larger the crowd, the more difficult the task of the detector. This is not only in terms of sheer quantities of data that would have to be processed in, for example, scanning social networks for steganographic activity (which is not our topic here). Fundamentally, it is because more innocent actors give the detector more opportunities to make mistakes, and give the steganographer more cover noise in which to hide: their own covers, but also the noise of the other actors. Thus, intuition suggests, when the crowd is larger the steganographer should be able to hide a larger payload for an equivalent risk of detection.

This intuition cannot be captured by the binary hypotheses model, so in this paper we consider a scenario with  $K$  actors of whom exactly one is a steganographer: this is  $K$ -way hypothesis testing with hypotheses

$H_1$  : Actor 1 is the steganographer,

$H_2$  : Actor 2 is the steganographer,

$\vdots$

$H_K$  : Actor  $K$  is the steganographer.

This paper provides capacity laws, with respect to simple abstract cover and embedding models, to quantify how the payload can depend on the size of the crowd. In the simplest case (Sect. 3) the critical rate is  $M = O(\sqrt{N \log K})$ , so that the crowd provides an asymptotic factor of  $\sqrt{\log K}$  extra capacity: a slowly-growing function in  $K$ , but still significant for a steganographer hiding in, for example, a large social media network.

We might wonder whether the amount of information sent by the *innocent* actors, rather than simply the number  $K - 1$  of such actors, makes a difference to the payload that the steganographer can send. Does more information from the others make their innocence more clear (and by implication the steganographer’s guilt more evident) or does it provide additional noise in which to hide? This is captured by a heterogeneous model and we shall see (Sect. 4) that it depends

on the detector knowing the steganographer's rate of scaling  $M$  with  $N$ .

We sketch some extensions of the results in Sect. 5, including relaxing the assumption of exactly one steganographer. This will highlight directions where further research is required.

## 1.1 Notation

Throughout the paper we will use  $K$  for the size of the crowd,  $M$  for the size of the payload, and  $N$  for the size of the cover medium, potentially indexed by  $k$ . All logs will be to natural base.  $f(n) \sim g(n)$  means that  $\frac{f(n)}{g(n)} \rightarrow 1$ , and  $f(n) \lesssim g(n)$  that  $\limsup \frac{f(n)}{g(n)} \leq 1$ , as  $n \rightarrow \infty$ .

$\text{Ber}(p)$  denotes a Bernoulli random variable, taking value 1 with probability  $p$  and 0 with probability  $1 - p$ , and  $\text{Cat}(\mathbf{p})$  a categorical distribution taking discrete values specified by the probability vector  $\mathbf{p}$ . Because all statistical hypotheses in this paper are *simple* – they have no unknown parameters – it will be notationally convenient to identify a hypothesis with the probability measure that it defines. It will be helpful to clarify which random variables are observed, by writing  $D_{\text{KL}}(\mathbf{X} | H_1, H_2)$  for the KL-divergence of the (distribution of the) random variable  $\mathbf{X}$  under hypothesis  $H_1$  and  $H_2$ . We also overload the notation, writing

$$D_{\text{KL}}(\mathbf{p}, \mathbf{q}) = D_{\text{KL}}(\text{Cat}(\mathbf{r}) | \mathbf{r}=\mathbf{p}, \mathbf{r}=\mathbf{q}) = \sum_i p_i \log\left(\frac{p_i}{q_i}\right),$$

(when  $\mathbf{p}$  and  $\mathbf{q}$  each sum to one) and

$$D_{\text{KL}}(p, q) = D_{\text{KL}}(\text{Ber}(r) | r=p, r=q) = p \log\left(\frac{p}{q}\right) + (1-p) \log\left(\frac{1-p}{1-q}\right).$$

In order to maintain the flow of the paper, some technical lemmas have been put into Appendix A.

## 2 THE WARDEN'S ADVANTAGE

In binary classification of a single actor, a detector has a binary output and is characterized by its false positive and false negative rates. In the  $K$ -class case studied in this paper, the detector outputs an estimate of which actor is the steganographer. To formalize the notation, let us write  $\mathcal{X}$  for the space of all possible observations made by the Warden (detector),  $\mathbf{X} \in \mathcal{X}$  for some realization, and  $\mathcal{D} : \mathcal{X} \rightarrow \{1, \dots, K\}$  for a possible decision function. The correct decision has been reached from observations  $\mathbf{X}$  if  $\mathcal{D}(\mathbf{X}) = k$  when  $H_k$  is the true hypothesis.

How should we generalize the measure of performance to the  $K$ -class problem? Inspired by the concept of an *attacker's advantage* in the theory of cryptography and hash functions [14], we propose a *Warden's advantage* to measure the excess probability that the steganographer is correctly identified, above the chance of random guessing:

$$\text{Adv}_+ = \max_{\mathcal{D}} \sum_k \frac{1}{K} \mathbb{P}[\mathcal{D}(\mathbf{X}) = k | H_k] - \frac{1}{K}. \quad (1)$$

If the steganographer sends no payload, their chance of being accused (given that the detector will accuse exactly one actor) is  $\frac{1}{K}$ , and if they can be sure that a certain size of payload ensures  $\text{Adv}_+ < \Delta$  then they can bound the additional risk that the act of steganography has visited upon them. Thus  $\text{Adv}_+$  is an *additive* measure of advantage; we will also briefly consider an alternative

measure of multiplicative advantage:

$$\text{Adv}_\times = \max_{\mathcal{D}} \sum_k \mathbb{P}[\mathcal{D}(\mathbf{X}) = k | H_k], \quad (2)$$

for which a score of  $\Delta$  implies that the steganographer's risk of detection does not exceed  $\Delta$  times their zero-payload risk. We shall see that, with respect to this score, the size of the crowd hinders rather than helps the steganographer.

Note that these definitions imply a uniform prior on the hypothesis  $H_1, \dots, H_K$ . We consider this prior quite reasonable as the numbering of the actors is arbitrary. However, it denies the Warden the possibility of a) having prior suspicions about some of the actors, or b) making use of meta-information such as the number of objects transmitted by each actor. We will briefly consider a non-uniform prior in Sect. 5, but postpone its study to future work.

This paper will prove theorems about  $\text{Adv}_+$  under various assumptions about the distribution of  $\mathbf{X}$  under each  $H_k$ . We will always assume that the actors are independent from each other (save that precisely one is a steganographer): this seems a very reasonable model of behaviour and allows us to state a simple result about  $K$ -way classification between simple hypotheses, analogous to the Neyman-Pearson Lemma, that connects the  $K$ -class optimal detector with likelihood ratios that are commonly used in the theory of binary steganalysis [3].

**CLAIM 1.** *Let  $\mathcal{D} : \mathcal{X} \rightarrow \{1, \dots, K\}$  be the Warden's decision function. The maximum Warden's advantage is achieved when*

$$\mathcal{D}(\mathbf{X}) = \arg \max_k \mathbb{P}[\mathbf{X} | H_k] = \arg \max_k \frac{P_S[\mathbf{X}_k]}{P_C[\mathbf{X}_k]}, \quad (3)$$

where  $\mathbf{X}_k$  represents the part of  $\mathbf{X}$  transmitted by actor  $k$ , and  $P_C$  (resp.  $P_S$ ) its probability distribution in the case that  $k$  is not (resp. is) the steganographer.

**PROOF.** The first equality is simply optimality of the Bayes classifier, and the second follows from independence of the actors given any  $H_k$ , and

$$\mathbb{P}[\mathbf{X} | H_k] = P_S[\mathbf{X}_k] \prod_{l \neq k} P_C[\mathbf{X}_l].$$

□

## 3 HOMOGENEOUS ACTORS

Let us start with the simplest model, in which all  $K$  actors (identical except for the one who is the steganographer) transmit the same number  $N$  of binary cover 'pixels', which all have the same distribution. Furthermore, the payload size  $M$  does not depend on the actor either. This allows us to examine the asymptotics of the Warden's advantage in terms of  $K$ ,  $N$ , and  $M$ , and to demonstrate the mathematical tools that we will use in the rest of the paper, in the context of relatively light notation.

Adopting the simplest Square-Root Law model from [9], we assume binary, independent, cover pixels which are Bernoulli random variables with parameter  $p$ . The embedding process is modelled by each pixel being *used* with probability  $\frac{M}{N}$ , and if used then replaced by a pixel with Bernoulli distribution with parameter  $q \neq p$ . This model avoids any kind of dependency between the cover or stego pixels. The reader may alternatively imagine that the  $N$  items transmitted by each actor are complete objects (e.g. images), each one

used to convey a fixed-length payload by the steganographer with probability  $\frac{M}{N}$ , which have been processed by a binary classifier outputting 0/1 which has false positive rate  $p$  and false negative rate  $1 - q$ .

This model specifies the distribution of  $\mathbf{X}$ , in which pixel  $i$  of actor  $k$  is denoted  $X_k^i$ , under each hypothesis as

$$\begin{aligned} H_k : X_k^i &\sim \text{Ber}(p + r\frac{M}{N}), \quad i = 1, \dots, N, \\ X_l^i &\sim \text{Ber}(p), \quad l \neq k, \quad i = 1, \dots, N, \end{aligned} \quad (4)$$

where  $r = q - p$ : an abbreviation that we will continue throughout the paper. The cases  $p = 0, 1$  must be excluded (a deterministic cover has zero steganographic capacity) and also  $r = 0$  (detection is impossible if the stego distribution is identical to the cover). Without loss of generality we may assume  $p < q$ . Note that, as in most classical Square-Root Laws, we make the conservative assumption that the parameters  $p$  and  $q$  are known to the Warden.

We will demonstrate that the steganographer's risk is asymptotically determined by

$$\frac{M^2}{N \log K},$$

so that (for bounded risk) the embedder can hide a payload of  $O(\sqrt{N \log K})$ , rather than the classical  $O(\sqrt{N})$  in the absence of a crowd. Our results parallel the classical Square-Root Law in having two parts: a proof of asymptotically perfect detection when  $M^2/N \log K \rightarrow \infty$ , and a bound on the Warden's advantage in terms of  $M^2/N \log K$  that forces asymptotically perfect security when  $M^2/N \log K \rightarrow 0$ .

### 3.1 Asymptotically Perfect Detection

It will be convenient to write  $Y_k = \sum_i X_k^i$  for the counting statistics of each actor. Equations (3) and (4) tell us the optimal detector for the Warden:

$$\begin{aligned} \mathcal{D}(\mathbf{X}) &= \arg \max_k \frac{\prod_{i=1}^N (p + r\frac{M}{N})^{X_k^i} (1 - p - r\frac{M}{N})^{1-X_k^i}}{\prod_{i=1}^N p^{X_k^i} (1-p)^{1-X_k^i}} \\ &= \arg \max_k \left( \frac{(p + r\frac{M}{N})(1-p)}{(1-p - r\frac{M}{N})p} \right)^{Y_k} \\ &= \arg \max_k Y_k, \end{aligned}$$

since  $r > 0$ . This is the unsurprising conclusion that, since embedding increases the chance of each pixel taking value 1 and the non-steganographer actors are identical, the optimal detector selects the actor transmitting the largest number of 1s.

A delicate analysis of the maximum of  $K$  random variables is difficult, but a crude bound on the error will be sufficient for our

asymptotic results:

$$\begin{aligned} &\mathbb{P}[\arg \max_i Y_i = k \mid H_k] \\ &\geq \mathbb{P}[Y_k > Np + \frac{1}{2}Mr \wedge \forall l \neq k, Y_l < Np + \frac{1}{2}Mr \mid H_k] \\ &\stackrel{(i)}{\geq} \left[ 1 - \exp\left(-ND_{\text{KL}}\left(p + \frac{1}{2}r\frac{M}{N}, p + r\frac{M}{N}\right)\right) \right] \cdot \\ &\quad \left[ 1 - \exp\left(-ND_{\text{KL}}\left(p + \frac{1}{2}r\frac{M}{N}, p\right)\right) \right]^{K-1} \\ &\stackrel{(ii)}{\geq} \left[ 1 - \exp\left(-\frac{1}{2}r^2\frac{M^2}{N}\right) \right] \cdot \left[ 1 - \exp\left(-\frac{1}{2}r^2\frac{M^2}{N}\right) \right]^{K-1} \\ &\stackrel{(iii)}{\geq} 1 - K \exp\left(-\frac{1}{2}r^2\frac{M^2}{N}\right), \end{aligned} \quad (5)$$

where (i) is the Chernoff bound for sums of independent Bernoulli variables, Lem.A.3(i), (ii) is a uniform lower bound on KL-divergence, Lem. A.2(i), and (iii) is the Weierstrass product inequality. This is enough to bound the Warden's advantage  $\text{Adv}_+(K, N, M)$  below, hence:

CLAIM 2.

$$\text{Adv}_+(K, N, M) \rightarrow 1 - \frac{1}{K} \quad \text{if} \quad \frac{M^2}{N \log K} \rightarrow \infty.$$

PROOF. By (3) and (5),

$$\text{Adv}_+(K, N, M) \geq 1 - \frac{1}{K} - \exp\left(\log K - \frac{1}{2}r^2\frac{M^2}{N}\right).$$

The hypothesis ensures that, in the exponent, the negative term in  $\frac{M^2}{N}$  dominates the positive term in  $\log K$ . Note that  $\text{Adv}_+(K, N, M)$  can never exceed  $1 - \frac{1}{K}$ , which is perfect detection.  $\square$

### 3.2 Asymptotically Zero Advantage

In the traditional Square-Root Law, every discriminator is proved to have asymptotically random performance using tools of information theory, typically by bounding the KL divergence (or some such measure) between the distributions of cover and stego objects. The information-processing theorem then bounds the KL divergence of the output of any detector. For the multiclass classification problem studied here we need a measure of divergence between  $K$  distributions: such measures are sometimes called *multi-divergences*, and have been studied (e.g. [6]) where it is shown that they satisfy information-processing inequalities. A convenient multi-divergence for our purposes is the *Jensen-Shannon divergence* [13], which measures the pairwise differences of each hypothesis (or measure) from their *average* measure:

$$D_{\text{JS}}(\mathbf{X} \mid H_1, H_2, \dots, H_K) = \frac{1}{K} \sum_{k=1}^K D_{\text{KL}}(\mathbf{X} \mid H_k, \bar{H}),$$

where  $\bar{H}$  represents the result of sampling  $\mathbf{X}$  from one of the  $H_k$  with  $k$  chosen uniformly at random, i.e.

$$\mathbb{P}[\mathbf{X} = \mathbf{x} \mid \bar{H}] = \frac{1}{K} \sum_{k=1}^K \mathbb{P}[\mathbf{X} = \mathbf{x} \mid H_k].$$

In our case,

$$\bar{H} : X_k^i \sim \text{Ber}(p + r\frac{M}{NK}), \quad i = 1, \dots, N, \quad k = 1, \dots, K.$$

So in order to bound the Jensen-Shannon divergence, we first bound the KL divergence from  $\bar{H}$ : for sufficiently small  $M/N$ ,

$$\begin{aligned} D_{\text{KL}}(\mathbf{X}_k | H_k, \bar{H}) &\stackrel{(i)}{=} ND_{\text{KL}}(p + r \frac{M}{N}, p + r \frac{M}{KN}) \\ &\stackrel{(ii)}{\leq} \frac{2r^2}{p(1-p)} \frac{(K-1)^2 M^2}{K^2 N}, \end{aligned} \quad (6)$$

and

$$\begin{aligned} D_{\text{KL}}(\mathbf{X}_I | H_k, \bar{H}) &\stackrel{(i)}{=} ND_{\text{KL}}(p, p + r \frac{M}{KN}) \\ &\stackrel{(ii)}{\leq} \frac{2r^2}{p(1-p)} \frac{M^2}{K^2 N}, \end{aligned} \quad (7)$$

where each (i) uses the independence of the pixels emitted by each actor, and each (ii) is by Lemma A.2(i). Then

$$\begin{aligned} D_{\text{JS}}(\mathbf{X} | H_1, H_2, \dots, H_K) &= \frac{1}{K} \sum_{k=1}^K D_{\text{KL}}(\mathbf{X} | H_k, \bar{H}) \\ &= \frac{1}{K} \sum_{k=1}^K \left( D_{\text{KL}}(\mathbf{X}_k | H_k, \bar{H}) + \sum_{l \neq k} D_{\text{KL}}(\mathbf{X}_l | H_k, \bar{H}) \right) \\ &\stackrel{(6,7)}{\leq} \frac{1}{K} \frac{2r^2}{p(1-p)} \left( K \frac{(K-1)^2 M^2}{K^2 N} + K(K-1) \frac{M^2}{K^2 N} \right) \\ &= \frac{2r^2}{p(1-p)} \frac{(K-1)M^2}{KN} \leq 2D \frac{M^2}{N} \end{aligned} \quad (8)$$

for  $D = \frac{r^2}{p(1-p)}$ . Slightly more strongly, in the case  $M/N \rightarrow 0$ , we can use Lemma A.2(iii) instead of Lemma A.2(i) to obtain

$$D_{\text{JS}}(\mathbf{X} | H_1, H_2, \dots, H_K) \sim D \frac{(K-1)M^2}{KN}. \quad (9)$$

We also need to bound the attacker's advantage by the Jensen-Shannon divergence. This is more complex than in binary hypothesis case. Write  $c_{ij} = P[\mathcal{D}(\mathbf{X}) = j | H_i]$ , so that  $(c_{ij})$  is the confusion matrix of the detector  $\mathcal{D}$ ,  $C_j = \sum_i c_{ij}$  for its column sums, and  $\bar{C}$  for its trace  $\sum_i c_{ii}$ . Then we can bound  $\alpha = \text{Adv}_+(K, N, M)$  as follows:

$$\begin{aligned} D_{\text{JS}}(\mathbf{X} | H_1, H_2, \dots, H_K) &\stackrel{(i)}{\geq} D_{\text{JS}}(\mathcal{D}(\mathbf{X}) | H_1, H_2, \dots, H_K) \\ &= \frac{1}{K} \sum_i D_{\text{KL}}(\mathcal{D}(\mathbf{X}) | H_i, \bar{H}) \\ &= \frac{1}{K} \sum_i \sum_j c_{ij} \log \left( \frac{c_{ij}}{\frac{1}{K} C_j} \right) \\ &= \frac{1}{K} \sum_i c_{ii} \log \left( \frac{c_{ii}}{\frac{1}{K} C_i} \right) + \frac{1}{K} \sum_{i \neq j} c_{ij} \log \left( \frac{c_{ij}}{\frac{1}{K} C_j} \right) \\ &\stackrel{(ii)}{\geq} \frac{1}{K} (\sum_i c_{ii}) \log \left( \frac{\sum_i c_{ii}}{\sum_i C_i} \right) + \frac{1}{K} (\sum_{i \neq j} c_{ij}) \log \left( \frac{\sum_{i \neq j} c_{ij}}{\sum_{i \neq j} C_j} \right) \\ &\stackrel{(iii)}{=} \frac{1}{K} \bar{C} \log \bar{C} + \frac{1}{K} (K - \bar{C}) \log \frac{K - \bar{C}}{K - 1} \\ &\stackrel{(iv)}{=} \frac{1}{K} (1 + K\alpha) \log(1 + K\alpha) + \frac{K-1}{K} \left( 1 - \frac{K}{K-1} \alpha \right) \log \left( 1 - \frac{K}{K-1} \alpha \right) \end{aligned} \quad (10)$$

where (i) is the information-processing inequality; the subsequent equalities apply the definitions and unpack the double sum into

the case  $i = j$  and  $i \neq j$ ; (ii) is the log-sum inequality; (iii) uses  $\sum_i C_i = \sum_{i,j} c_{ij} = K$  and  $\sum_{i \neq j} c_{ij} = K - \bar{C}$ ; finally, (iv) uses  $\text{Adv}_+(K, N, M) = \sum_k \frac{1}{K} P[\mathcal{D}(\mathbf{X}) = k | H_k] - \frac{1}{K} = \frac{\bar{C}}{K} - \frac{1}{K}$ .

Lemma A.1 is able to separate this equation into terms in  $\alpha$  and  $K$ . Putting them all together, we have:

CLAIM 3.

$$\begin{aligned} (i) \text{ Adv}_+(K, N, M)^2 &\leq 2D \frac{M^2}{N \log K}, \text{ for sufficiently small } M/N. \\ (ii) \text{ Adv}_+(K, N, M)^2 &\lesssim D \frac{M^2}{N} \frac{K-2}{K \log(K-1)}, \text{ as } M/N \rightarrow 0. \end{aligned}$$

PROOF. For (i),

$$\begin{aligned} \text{Adv}_+(K, N, M)^2 \log K &\stackrel{(30)(ii)}{\leq} \frac{1}{K} (1 + K\alpha) \log(1 + K\alpha) + \frac{K-1}{K} \left( 1 - \frac{K}{K-1} \alpha \right) \log \left( 1 - \frac{K}{K-1} \alpha \right) \\ &\stackrel{(10)}{\leq} D_{\text{JS}}(\mathbf{X}^1, \mathbf{X}^2, \dots, \mathbf{X}^K) \\ &\stackrel{(8)}{\leq} 2D \frac{M^2}{N}. \end{aligned}$$

and (ii) is identical but uses the tighter asymptotic results: (9) instead of (8), and (30)(i) instead of (30)(ii).  $\square$

### 3.3 Interpretation

The statements of Claims 2 and 3 give the headline conclusion: compared to critical rate  $O(\sqrt{N})$  in the classical Square-Root Law (in the absence of source coding), a steganographer acting in a crowd of  $K$  has a critical rate of  $O(\sqrt{N \log K})$ . To ensure that  $\text{Adv}_+(K, N, M) \leq \Delta$ , if this represents the risk they are willing to take, they should not transmit a payload of more than

$$M \lesssim \sqrt{\Delta (\log K) N / D}. \quad (11)$$

The extra factor of  $\sqrt{\log K}$  is certainly slower-growing than the square-root factor of  $N$ , but this shows that the crowd of size  $K$  gives as much extra security (or capacity) as does a multiple of  $\log K$  in the cover size.

The constant  $D$  is a measure of the *detectability* of the embedding operation itself, given the cover and stego pixel distributions: the same constant appears in the classical Square-Root Law as stated in [10], and  $D = \frac{(q-p)^2}{p(1-p)}$  is in fact the  $\chi^2$ -divergence between distribution of cover and stego pixels (see Subsect. 5.1). Larger  $D$  forces a lower payload.

We should, though, be careful that Claim 3 is an *asymptotic* result: a bound for sufficiently small  $M/N$ , or a slightly tighter asymptotic bound as  $M/N \rightarrow 0$ . If  $K$  is fixed and  $N \rightarrow \infty$  then (11) forces  $M/N \rightarrow 0$ , but what if *both*  $N$  and  $K$  grow? Asymptotic results in the presence of more than one variable can be tricky, but we may still conclude (11) as long as  $K$  does not grow exponentially in  $N^1$ , as then  $\text{Adv}_+(K, N, M) \leq \Delta$  still forces  $M/N \rightarrow 0$ .

Finally, we briefly turn our attention to the multiplicative measure of Warden's advantage. Since

$$\text{Adv}_\times(K, N, M) = K \text{ Adv}_+(K, N, M) + 1,$$

<sup>1</sup>Further analysis shows that the same result still holds even in this strange case, but we omit it from the paper as an unnecessary complication.

the critical rate becomes  $O(\sqrt{N \log K/K})$ . This is true in any regime where  $N$  and/or  $K \rightarrow \infty$ . It implies that the steganographer must hide *less* information as  $K$  grows, in order to keep their multiplicative risk below a threshold. This is simply because they are demanding ever smaller risks of detection in larger crowds, and our new capacity law tells us that the crowd alone does not provide this.

## 4 HETEROGENEOUS ACTORS WITH HETEROGENEOUS PIXELS

We now examine the effect of heterogeneity amongst the actors, of which we simultaneously consider two types. First, the binary ‘pixels’ that form the actors’ covers need not be identically distributed, instead varying both by location and by actor; this will have little effect on the conclusions of the previous section, as long as we ban certain pathological asymptotic behaviour, except that the detectability constant  $D$  differs between actors. Second, we assume that the actors have covers of different sizes, and that the size of the payload embedded by the steganographer may optionally also depend on the cover size; the results will tell a more complex story here.

The model still consists of  $K$  independent actors. We will suppose that actor  $k$  sends  $N_k$  independent Bernoulli pixels, and that  $p_k^i$  is the probability that pixel  $i$  from actor  $k$  is 1 in covers. If actor  $k$  is the steganographer then we model a payload of size  $M_k$  using each of their pixels with probability  $M_k/N_k$ , and if used then the pixel parameter changes to  $q_k^i$ , with  $r_k^i = q_k^i - p_k^i$ . Thus

$$\begin{aligned} H_k : X_k^i &\sim \text{Ber}(p_k^i + r_k^i \frac{M_k}{N_k}), \quad i = 1, \dots, N_k, \\ X_l^i &\sim \text{Ber}(p_l^i), \quad l \neq k, \quad i = 1, \dots, N_l, \end{aligned} \quad (12)$$

and so

$$\bar{H} : X_k^i \sim \text{Ber}(p_k^i + r_k^i \frac{M_k}{KN_k}), \quad i = 1, \dots, N_k, \quad k = 1, \dots, K.$$

The reader may alternatively wish to imagine a model where the  $N_k$  locations for actor  $k$  represent well-separated potential changes in a larger image, with  $p_k^i$  the probability that it contains its most likely value conditional upon its local neighbourhood, and  $q_k^i$  the probability of that same value after embedding. This is a good model for sparse changes in a correlated cover, against a detector based on local statistics.

We make the conservative assumption that all the parameters –  $M_1, \dots, M_K, N_1, \dots, N_K$  and all the  $p_k^i$  and  $q_k^i$  – are known to the detector. It is essential to our analysis that the values of  $N_k$  do not, in themselves, convey suspicion: the steganographer is not more likely to send larger numbers of images, for example.  $N_k$  is merely a parameter of the problem. (We will mention briefly how one might incorporate suspicions about the value of  $N_k$  into a detection model in Sect. 5.)

In the homogeneous model we required that  $p \neq 0, 1$  (deterministic covers) and  $p \neq q$  (perfect embedding). In the heterogeneous case we must also ban *asymptotic* determinism or perfection where, for example  $p_k^i \rightarrow 0$  or  $p_k^i - q_k^i \rightarrow 0$  as  $i \rightarrow \infty$ . For simplicity we will require that there is  $\delta > 0$  such that

$$\delta \leq p_k^i, q_k^i \leq 1 - \delta \text{ and } |p_k^i - q_k^i| \geq \delta, \text{ for all } i \text{ and } k. \quad (13)$$

There is no real need for this bound to be uniform in  $k$ , or indeed to hold for all  $i$ , but it complicates the algebra in unenlightening ways to make weaker assumptions. It is cleaner to assume that the  $p_k^i$  and  $q_k^i$  are generated by some stationary process (potentially different for each  $k$ ) such that  $|p_k^i - q_k^i| \geq \delta$ , so that there is no long-term dependence on  $i$ .

Having stated the model, we can give asymptotic bounds on the attackers advantage. However, just as we had to be careful about asymptotic results in both  $N$  and  $K$  in Sect. 3, here we have variables  $K, N_1, \dots, N_K$ , any or all of which might tend to infinity. We will demonstrate that the steganographer’s risk is bounded below by

$$\min_k \frac{M_k^2}{N_k},$$

and above by

$$\frac{1}{K \log K} \sum_{k=1}^K D_k \frac{M_k^2}{N_k},$$

where  $D_k$  is a detectability factor for actor  $k$ ’s embedding process that functions similarly to  $D$  in Sect. 3. What this means for the practice of steganography in a crowd will be examined at the end of the section.

### 4.1 Asymptotically Perfect Detection

Plugging (12) into (3) tells us the optimal detector for the Warden is now:

$$\begin{aligned} \mathcal{D}(X) &= \arg \max_k \frac{\prod_{i=1}^{N_k} (p_k^i + r_k^i \frac{M_k}{N_k})^{X_k^i} (1 - p_k^i - r_k^i \frac{M_k}{N_k})^{1-X_k^i}}{\prod_{i=1}^{N_l} (p_l^i)^{X_k^i} (1 - p_l^i)^{1-X_k^i}} \\ &= \arg \max_k \sum_{i=1}^{N_k} X_k^i \log \left( 1 + \frac{r_k^i}{p_k^i} \frac{M_k}{N_k} \right) + (1 - X_k^i) \log \left( 1 - \frac{r_k^i}{1-p_k^i} \frac{M_k}{N_k} \right) \\ &= \arg \max_k S_k, \end{aligned} \quad (14)$$

say, where  $S_k$  is a weighted score for actor  $k$ , that functions analogously to  $Y_k$  in Sect. 3: pixels more likely to be 1 in stego than cover are given higher positive weights, with negative weights for pixels more likely to be 1 in cover than stego.

We will need to show that the score of the steganographer separates from the scores of the other actors. We highlight some intuition about the asymptotics: if the steganographer  $k$  and at least one innocent actor  $l$  both send few pixels then we should *not* expect asymptotically perfect detection because, even if the other  $K - 2$  actors rule themselves out of suspicion by sending many pixels matching their cover distributions, the detector will be unable to distinguish  $k$  and  $l$  perfectly.

To begin the analysis we use Lem. A.2(i) to bound

$$E[S_k | H_k] = \sum_{i=1}^{N_k} D_{\text{KL}}(p_k^i + r_k^i \frac{M_k}{N_k}, p_k^i) \geq \sum_{i=1}^{N_k} 2(r_k^i \frac{M_k}{N_k})^2 \geq 2\delta^2 \frac{M_k^2}{N_k}$$

and

$$E[S_l | H_k] = \sum_{i=1}^{N_l} -D_{\text{KL}}(p_l^i, p_l^i + r_l^i \frac{M_l}{N_l}) \leq -\sum_{i=1}^{N_l} 2(r_l^i \frac{M_l}{N_l})^2 \leq -2\delta^2 \frac{M_l^2}{N_l},$$

and it remains to show that the scores of the steganographer and the others concentrate on positive and negative values, respectively. We

can do this because the terms in the sums making up  $S_k$  and  $S_l$  are bounded: they are of the form  $\log(1 + \frac{M_k}{N_k} \frac{q-p}{p})$  and  $\log(1 - \frac{M_k}{N_k} \frac{q-p}{1-p})$ , where  $p$  and  $q$  are bounded away from 0, 1, and each other, by  $\delta$ . Thus, by Lem. A.4, every term in  $S_i$  is bounded within a range of  $C_i = 2 \frac{M_i}{N_i} \frac{1-2\delta}{\delta}$ . Now set

$$\eta = 0, \quad \theta = \frac{\delta^3}{1-2\delta} \min_i \frac{M_i}{\sqrt{N_i}}.$$

Then, crudely bounding the probability that  $\mathcal{D}(\mathbf{X}) = k$ , given  $H_k$ ,

$$\begin{aligned} & \mathbb{P}[\arg \max_i S_i = k \mid H_k] \\ & \geq \mathbb{P}[S_k > \eta \wedge \forall l \neq k. S_l < \eta \mid H_k] \\ & \stackrel{(i)}{\geq} \mathbb{P}[S_k > \mathbb{E}[S_k] - \theta C_k \sqrt{N_k} \wedge \forall l \neq k. S_l < \mathbb{E}[S_l] + \theta C_l \sqrt{N_l} \mid H_k] \\ & \stackrel{(ii)}{\geq} \left[1 - \exp(-2\theta^2)\right] \cdot \prod_{l \neq k} \left[1 - \exp(-2\theta^2)\right] \\ & \stackrel{(iii)}{\geq} 1 - K \exp(-2\theta^2), \end{aligned} \quad (15)$$

where (i) is because we have chosen  $\theta$  to ensure that, for each  $i$ ,

$$\theta C_i \sqrt{N_i} \leq |\eta - \mathbb{E}[S_i \mid H_k]|; \quad (16)$$

(ii) is Hoeffding's inequality Lem. A.3(ii); (iii) is again the Weierstrass product inequality.

This establishes, denoting the Warden's advantage with parameters  $K, N_1, \dots, N_K, M_1, \dots, M_K$  as  $\text{Adv}_+(K, \mathbf{N}, \mathbf{M})$ ,

CLAIM 4.

$$\text{Adv}_+(K, \mathbf{N}, \mathbf{M}) \rightarrow 1 - \frac{1}{K}, \quad \text{if } \frac{M_i^2}{N_i \log K} \rightarrow \infty \text{ for all } i.$$

PROOF. The hypothesis ensures that  $\frac{\theta^2}{\log K} \rightarrow \infty$ , and the result follows from (15).  $\square$

The reader may notice, however, that it is not necessary for *all*  $M_i^2/N_i \log K$  to tend to infinity. For example, let  $k$  be the steganographic actor. Setting

$$\eta = \frac{1}{2} \mathbb{E}[S_k], \quad \theta = \frac{\frac{1}{2} \delta^3}{1-2\delta} \frac{M_k}{\sqrt{N_k}}$$

also satisfies (16) and proves asymptotically perfect detection given  $M_k^2/N_k \log K \rightarrow \infty$ , regardless of the asymptotics of the quotients for other  $i$ . Alternatively, setting

$$\eta = \frac{1}{2} \max_{i \neq k} \mathbb{E}[S_i], \quad \theta = \frac{\frac{1}{2} \delta^3}{1-2\delta} \min_{i \neq k} \frac{M_i}{\sqrt{N_i}}$$

again satisfies (16), proving asymptotically perfect detection if  $M_i^2/N_i \log K \rightarrow \infty$  for all  $i \neq k$ .

That is, asymptotically perfect separation of the scores occurs if *either* the steganographer or *all* the innocent actors have their  $M_i$  exceed the critical rate of  $O(\sqrt{N_i \log K})$ . Furthermore we could prove that the attacker's advantage tends to at least  $\frac{L}{K}$  if  $L$  of the innocent actors exceed the critical rate. These results do not fit neatly within our definition of attacker's advantage, which is founded on the steganographer being a uniformly random actor (and therefore not identified by a particular index  $k$ ), but it illustrates

some of the complications when there are many variables, only some of which tend to infinity.

## 4.2 Asymptotically Zero Advantage

Here the steps are identical to Subsect. 3.2. If the parameters  $p_k^i$  and  $q_k^i$  are generated by some stationary process, and bounded away from each other and zero, then for each actor  $k$

$$\frac{1}{N_k} \sum_{i=1}^{N_k} \frac{(r_k^i)^2}{p_k^i(1-p_k^i)} \rightarrow D_k,$$

for some finite positive constant  $D_k$  as  $N_k \rightarrow \infty$ . Furthermore, by (13),

$$4\delta^2 \leq D_k \leq \frac{(1-2\delta)^2}{\delta(1-\delta)}. \quad (17)$$

Then for sufficiently large  $N_k$  and small  $M_k/N_k$ ,

$$\begin{aligned} D_{\text{KL}}(\mathbf{X}_k \mid H_k, \bar{H}) &= \sum_{i=1}^{N_k} D_{\text{KL}}(p_k^i + r_k^i \frac{M_k}{N_k}, p_k^i + r_k^i \frac{M_k}{N_k}) \\ &\leq 2D_k \frac{(K-1)^2 M_k^2}{K^2 N_k}, \text{ or} \\ &\sim D_k \frac{(K-1)^2 M_k^2}{K^2 N_k} \text{ as } N_k \rightarrow \infty \text{ and } \frac{M_k}{N_k} \rightarrow 0, \end{aligned} \quad (18)$$

and similarly

$$\begin{aligned} D_{\text{KL}}(\mathbf{X}_l \mid H_k, \bar{H}) &= \sum_{i=1}^{N_l} D_{\text{KL}}(p_l^i + r_l^i \frac{M_l}{N_l}, p_l^i + r_l^i \frac{M_l}{N_l}) \\ &\leq 2D_l \frac{M_l^2}{K^2 N_l}, \text{ or} \end{aligned} \quad (20)$$

$$\sim D_l \frac{M_l^2}{K^2 N_l} \text{ as } N_l \rightarrow \infty \text{ and } \frac{M_l}{N_l} \rightarrow 0. \quad (21)$$

Thus

$$\begin{aligned} D_{\text{JS}}(\mathbf{X} \mid H_1, H_2, \dots, H_K) &= \frac{1}{K} \sum_{k=1}^K D_{\text{KL}}(\mathbf{X} \mid H_k, \bar{H}) \\ &= \frac{1}{K} \sum_{k=1}^K \left( D_{\text{KL}}(\mathbf{X}_k \mid H_k, \bar{H}) + \sum_{l \neq k} D_{\text{KL}}(\mathbf{X}_l \mid H_k, \bar{H}) \right) \\ &\stackrel{(18,20)}{\leq} \frac{1}{K} \frac{(K-1)^2}{K^2} \sum_k 2D_k \frac{M_k^2}{N_k} + \frac{K-1}{K} \sum_k 2D_k \frac{M_k^2}{N_k} \\ &= \frac{K-1}{K^2} \sum_k 2D_k \frac{M_k^2}{N_k} < \frac{1}{K} \sum_k 2D_k \frac{M_k^2}{N_k}, \end{aligned}$$

or alternatively using (19) and (21),

$$D_{\text{JS}}(\mathbf{X} \mid H_1, H_2, \dots, H_K) \sim \frac{K-1}{K^2} \sum_k D_k \frac{M_k^2}{N_k}.$$

Then following the same steps as Claim 3,

CLAIM 5.

$$(i) \text{Adv}_+(K, \mathbf{N}, \mathbf{M})^2 \leq \frac{2}{K \log K} \sum_k D_k \frac{M_k^2}{N_k},$$

*for sufficiently small  $\max M_k/N_k$  and large  $\min N_k$ .*

$$(ii) \text{Adv}_+(K, \mathbf{N}, \mathbf{M})^2 \lesssim \frac{K-2}{K^2 \log(K-1)} \sum_k D_k \frac{M_k^2}{N_k},$$

*if all  $M_k/N_k \rightarrow 0$  and  $N_k \rightarrow \infty$ .*

### 4.3 Interpretation

Once again, we see that the size of the payload may scale as a factor of approximately  $\sqrt{\log K}$  if the Warden's advantage is to be fixed. But the possibility of some  $N_k \rightarrow \infty$ , and some not, complicates the exposition. Concentrating on the upper bound on  $\text{Adv}_+(K, \mathbf{N}, \mathbf{M})$ , suppose that the steganographer wishes to ensure that

$$\frac{1}{K \log K} \sum_{k=1}^K D_k \frac{M_k^2}{N_k} \leq \Delta, \quad (22)$$

to stay within their risk tolerance. Let us suppose that  $K$  does not increase exponentially with  $\sum N_k$  so that we must indeed have  $M_i/N_i \rightarrow 0$  for each  $i$  where  $N_i \rightarrow \infty$ .

We may identify three types of embedder, each of whom chooses  $M_k$  as a different function of  $N_k$ :

*Constant payload embedder.*  $M_i = M$  for some  $M$ . This situation might arise if a detector finds exfiltrated data, and looks back in network logs to try to identify which actor transmitted it: the size of the payload is fixed, but the identity of the actor is not. In order to meet (22) the steganographer should require

$$M \leq \sqrt{\Delta \log K \frac{K}{\sum \frac{D_k}{N_k}}}.$$

This gives the surprising result that the payload capacity is proportional to the square root of the *harmonic mean*<sup>2</sup> of the number of pixels transmitted by each actor: a weighted harmonic mean with the  $D_k$  providing the weights. The harmonic mean of a set of real numbers is, recall, dominated by the *smallest* members of that set.

How should we understand this phenomenon? In the case where  $M$  is fixed, innocent actors sending few pixels rule themselves strongly *out* of suspicion: for if they were steganographers, they would have to use a high proportion of pixels to fit  $M$  payload bits into their small cover, and would have attained a high score. Thus the detector can using negative information about some of the innocent actors (those with few images who did not attain a high score are definitely innocent) in order to raise the risk of detection of the steganographer.

*Constant rate embedder.*  $M_i = rN_i$  for some *rate*  $r$ . This simulates a steganographer who is ignorant of the Square-Root Law, or uses steganographic software with a fixed embedding rate. From the detector's side this might arise if they see such software being downloaded onto their network. In order to meet (22) the steganographer

should set

$$r \leq \sqrt{\Delta \log K \frac{K}{\sum N_k D_k}}.$$

Here the steganographer's capacity is proportional to the inverse square-root of the arithmetic mean number of pixels transmitted (the mean weighted by the  $D_k$ ). The mean of a set tends to be dominated by the largest values. Here we understand that the detector is again able to use negative information to rule *out* some of the innocent actors, in this case those who transmitted many pixels. For if they were steganographers then they would have given plenty of evidence by transmitting large payloads, and thus attained a high score. Recall that the Warden will have asymptotically positive advantage if they can rule out even one innocent actor.

*Root-rate embedder.* A steganographer who has read about the Square-Root Law should decide that  $M_i = r\sqrt{N_i}$  for some *root-rate*  $r$ . If all actors would behave in the same manner, in order to meet (22) the steganographer should set

$$r \leq \sqrt{\Delta \log K \frac{K}{\sum D_k}}.$$

In this case the steganographer's payload has been scaled correctly, and the capacity root-rate depends neither on the number of pixels transmitted by the steganographer *nor on the number transmitted by the innocent actors*. It is determined by  $\sqrt{\log K}$  and the inverse square-root of the average detectability of all actors' covers. The steganographer is still hiding within the noise of the other actors' pixels (and the noisier those pixels, the higher the  $D_k$ 's, so the more the steganographer can hide) but they are indifferent to the number of pixels sent by the rest of the crowd. Given that the steganographer may very well not know the other  $N_k$  at the point they transmit, this seems a desirable property. (Whereas, if they are unable to estimate  $D_k$ , they can at least use the bound (17).)

So returning to the question at the end of Sect. 1: does the amount of information transmitted by the innocent actors affect the capacity of the steganographer? Are the innocent actors ruling themselves out of suspicion as important as the steganographer's positive evidence in their pixel distribution?

If the steganographer does not scale their payload with the correct square-root relationship to their own cover size, the answer is yes. And *how* the capacity is affected – more affected by innocent actors sending more images, or innocent actors sending few images – depends on that very scaling.

## 5 EXTENSIONS

As with Square-Root Laws for the binary steganalysis problem, the results can be further extended. We sketch some such extensions here, postponing details to an extended version of this paper.

### 5.1 Multicolour Pixels

Moving away from binary cover elements is straightforward. Indeed, we could have started with this case if we were willing to tolerate the density of the mathematical notation.

Suppose that the cover elements are taken from an alphabet  $\{1, 2, \dots, J\}$ , and  $p_k^{i,j}$  (resp.  $q_k^{i,j}$ ) is the probability that pixel  $i$  from

<sup>2</sup>The harmonic mean of  $\{x_1, \dots, x_n\}$  is  $n(\sum x_i^{-1})^{-1}$ , and a weighted harmonic mean is  $(\sum w_i)(\sum w_i x_i^{-1})^{-1}$ .

actor  $k$  takes value  $j$  in covers (resp. pixels used by the steganographer), with  $\sum_{j=1}^J p_k^{i,j} = \sum_{j=1}^J q_k^{i,j} = 1$  for all  $i$  and  $k$ . We gather together each  $(p_k^{i,1}, \dots, p_k^{i,J})$  into a vector  $\mathbf{p}_k^i$ , (resp.  $\mathbf{q}_k^i$ ).

The proofs turn out to be identical to those in Sect. 4, with the key property being that

$$D_{\text{KL}}(\mathbf{p}, \mathbf{p} + \frac{M_k}{N_k}(\mathbf{q} - \mathbf{p})) \sim \frac{M_k^2}{2N_k^2} D_{\chi^2}(\mathbf{p}, \mathbf{q}),$$

as  $\frac{M_k}{N_k} \rightarrow 0$ , where  $D_{\chi^2}$  represents the  $\chi^2$ -divergence between probability distributions:

$$D_{\chi^2}(\mathbf{p}, \mathbf{q}) = \sum_{i=1}^J \frac{(p_i - q_i)^2}{p_i}.$$

Then the KL divergence can be bounded above and below to match the steps used in proving Claims 4 and 5. The bound (13) can be replaced with the rather more elegant requirement that

$$\lim_{N_k \rightarrow \infty} \frac{1}{N_k} \sum_{i=1}^{N_k} D_{\chi^2}(\mathbf{p}_k^i, \mathbf{q}_k^i) = D_k$$

is bounded and nonzero for each  $k$ . It is not even necessary for  $J$  to be finite, as long as the sums converge. The conclusions are identical to Subsect. 4.3

## 5.2 Two Steganographers

A more interesting extension is to change the scenario so that there are two steganographers in the crowd, rather than one. Now we have a  $\binom{K}{2}$ -class classification problem with hypotheses

$$H_{\{k,l\}} : \text{Actors } k \text{ and } l \text{ are the steganographers.}$$

How does the mathematics change?

The optimal detector, by Bayes' criterion, is now

$$\mathcal{D}(\mathbf{X}) = \arg \max_{\{k,l\}} P[\mathbf{X} | H_{\{k,l\}}] = \arg \max_{k \neq l} \frac{P_S[\mathbf{X}_k]}{P_C[\mathbf{X}_k]} \frac{P_S[\mathbf{X}_l]}{P_C[\mathbf{X}_l]},$$

which means (unsurprisingly) that the detector simply picks the two steganographers with the highest scores according to (14). Claim 4 has an unchanged conclusion: if all actors have  $M_i^2/(N_i \log K) \rightarrow \infty$  then *both* steganographers will have scores that are almost surely positive, while all innocent actors' scores will be almost surely negative. Thus asymptotically perfect detection occurs.

For the other bound we calculate the Jensen-Shannon divergence using

$$\bar{H} : X_k^i \sim \text{Ber}(p_k^i + r_k^i \frac{2M_k}{KN_k}), i = 1, \dots, N_k, k = 1, \dots, K,$$

and, after a careful counting of the triple sums involved, find that

$$\begin{aligned} D_{\text{JS}}(\mathbf{X} | \{H_{\{k,l\}} | 1 \leq k \neq l \leq K\}) &\leq \sum_k \frac{2(K-2)}{K^2} D_k \frac{M_k^2}{N_k} \\ &\sim \frac{2}{K} \sum_k D_k \frac{M_k^2}{N_k}. \end{aligned} \quad (23)$$

It will be convenient to write  $\kappa = 1/\binom{K}{2}$  for the probability of guessing both steganographers at random. We could continue to define

$$\text{Adv}_+ = \max_{\mathcal{D}} \sum_{k,l} \kappa P[\mathcal{D}(\mathbf{X}) = \{k,l\} | H_{\{k,l\}}] - \kappa$$

and in this case the same results would follow, except that the factor of  $\sqrt{\log K}$  becomes  $\sqrt{\log \binom{K}{2}}$  because there are more hypotheses to distinguish. However, it is more interesting to consider the probability that the Warden correctly identifies *at least one* of the steganographers, which we measure by their *partial advantage*:

$$\text{Padv}_+ = \max_{\mathcal{D}} \sum_{k,l} \kappa P[\mathcal{D}(\mathbf{X}) \cap \{k,l\} \neq \emptyset | H_{\{k,l\}}] - \gamma,$$

where

$$\gamma = \frac{\binom{K}{2} - \binom{K-2}{2}}{\binom{K}{2}} = \frac{4K-6}{K(K-1)} \quad (24)$$

is the probability of guessing at least one correct steganographer. We can prove an upper bound on  $\alpha = \text{Padv}_+(K, \mathbf{N}, \mathbf{M})$  using the following reasoning. For  $I, J \in \{\{k,l\} | 1 \leq k \neq l \leq K\}$  write  $c_{IJ} = P[\mathcal{D}(\mathbf{X}) = J | H_I]$  and  $C_I = \sum_J c_{IJ}$ , for the  $\binom{K}{2} \times \binom{K}{2}$  confusion matrix of the detector and its column sums. Define  $C = \{(I, J) | I \cap J \neq \emptyset\}$  for those entries of the matrix corresponding to at least partial success, identifying at least one steganographer correctly. We will need to compute:

$$c_1 = \sum_{(I,J) \in C} c_{IJ} = (\gamma + \alpha)/\kappa, \quad (25)$$

$$c_2 = \sum_{(I,J) \notin C} c_{IJ} = (1 - \gamma - \alpha)/\kappa, \quad (26)$$

$$c_3 = \sum_{(I,J) \in C} C_I = \sum_{(I,J) \in C} \sum_{J'} c_{IJ'} = \gamma/\kappa^2, \quad (27)$$

$$c_4 = \sum_{(I,J) \notin C} C_I = \sum_{(I,J) \notin C} \sum_{J'} c_{IJ'} = (1 - \gamma)/\kappa^2. \quad (28)$$

Finally,

$$\begin{aligned} D_{\text{JS}}(\mathbf{X} | \{H_{\{k,l\}} | 1 \leq k \neq l \leq K\}) &\stackrel{(i)}{\geq} D_{\text{JS}}(\mathcal{D}(\mathbf{X}) | \{H_{\{k,l\}} | 1 \leq k \neq l \leq K\}) \\ &= \kappa \sum_I \sum_J c_{IJ} \log\left(\frac{c_{IJ}}{\kappa C_J}\right) \\ &\stackrel{(ii)}{=} \kappa \sum_{(I,J) \in C} c_{IJ} \log\left(\frac{c_{IJ}}{\kappa C_I}\right) + \kappa \sum_{(I,J) \notin C} c_{IJ} \log\left(\frac{c_{IJ}}{\kappa C_J}\right) \\ &\stackrel{(iii)}{\geq} \kappa c_1 \log\left(\frac{c_1}{\kappa c_3}\right) + \kappa c_2 \log\left(\frac{c_2}{\kappa c_4}\right) \\ &\stackrel{(iv)}{=} (\gamma + \alpha) \log\left(\frac{\gamma + \alpha}{\gamma}\right) + (1 - \gamma - \alpha) \log\left(\frac{1 - \gamma - \alpha}{1 - \gamma}\right) \\ &\stackrel{(v)}{\geq} \alpha^2 \log(1/\gamma) \sim \alpha^2 \log\left(\frac{K}{4}\right). \end{aligned} \quad (29)$$

where (i) is the information-processing inequality; (ii) breaks down the case of at least partial success and no success; (iii) is the log-sum inequality; (iv) uses (25)-(28); (v) uses Lemma A.1(ii) with  $K = 1/\gamma$ , and (24).

Connecting (23) and (29) gives the upper bound (neglecting factors of  $(K-2)/K$ ):

$$\text{Padv}_+(K, \mathbf{N}, \mathbf{M})^2 \lesssim \frac{2}{K \log\left(\frac{K}{4}\right)} \sum_k D_k \frac{M_k^2}{N_k}$$

which is of the same order as the one-steganographer case, but with a higher constant. We postpone the extension to arbitrary  $L$



steganographers (where  $L$  is presumed known to the detector) to future work. We will simply state that the multiplicative constant becomes  $L/\log(\frac{K}{L^2})$ , so that the steganographers derive less protection from the crowd when there are more of them in it. It is unsurprising that the result breaks down for  $L > \sqrt{K}$  because then the Birthday Paradox means that pure guessing should identify one of them by chance.

### 5.3 One or Zero Steganographers

Let us return to the case of a single steganographer. Another way to widen the scenario is to add one additional hypothesis:

$H_0$  : There is no steganographer.

Analysis of this  $(K + 1)$ -class classification problem turns out to be rather easy. (14) shows that hypothesis  $H_0$  should be assigned score

$$S_0 = 0,$$

and in fact all of the mathematics follows through with a dummy actor for whom  $M_0 = N_0 = 0$ . The JS divergence is unchanged and the conclusions of Subsect. 4.3 remain valid, with  $\log(K + 1)$  replacing  $\log K$  because there are now  $K + 1$  hypotheses. The same trick also allows us to consider zero, one, or two steganographers, by adding *two* dummy actors to the situation in Subsect. 5.2, and so on.

However, at this point we start to question the initial assumption of a uniform prior on the hypotheses. While it may be quite reasonable to assign a uniform prior to  $K$  actors, since they were numbered in some arbitrary fashion, it seems unreasonable to insist that the no-steganographer hypothesis has likelihood  $\frac{1}{K+1}$ . To examine this case properly we need to permit a prior distribution  $\pi_0, \dots, \pi_K$  on the hypotheses. This would also give a mechanism for the Warden to assign suspicion to actors who send too many pixels, as their prior could be adjusted according to how unusual was their observed cover size. The Warden's advantage should be defined as the Bayes risk,

$$\max_{\mathcal{D}} \sum_k \pi_k P[\mathcal{D}(X) = k | H_k] - \max_k \pi_k.$$

Dealing with non-uniform priors is beyond the scope of this paper. A difficulty is that a *weighted Jensen-Shannon divergence* exists [13] but it is known that its bounds on multiclass accuracy are weak as  $K \rightarrow \infty$ . Some new ideas will be needed here.

## 6 CONCLUSIONS

These new results quantify the intuition that we discussed in Sect. 1: a single steganographer in a crowd of  $K$  can increase their payload size by a factor  $O(\sqrt{\log K})$ . We have shown this to be true for a homogeneous binary pixel model and then generalized to heterogeneous models,  $J$ -colour pixels, and have sketched results showing that one can weaken the assumption of exactly one steganographer. Capacity laws are always determined by a measure of detection accuracy, and we have focused on the additive measure (1) because of its links with Bayes risk. The intuitively paradoxical result that larger crowds force smaller payloads when accuracy is measured multiplicatively (2) is one reason why we generally eschew this measure.

A natural extension would be to permit (suitably-bounded) dependence between the cover 'pixels', as was done for the classical Square-Root Law in [11]. However this may be challenging, and since bounded dependence did not affect the asymptotic capacity law for the binary steganalysis problem it is reasonable to expect the same for the multiclass problem. A more significant extension would be to permit source coding by the steganographer: the basic embedding model studied in this paper – 'use' each pixel with probability  $M/N$  – models naïve methods such as LSB overwriting, whereas modern steganography almost inevitably employs some kind of syndrome code to minimize a distortion metric [4]: in the classical case this turns the  $O(\sqrt{N})$  critical rate into  $O(\sqrt{N} \log N)$ , and we will need to examine the interaction with  $K$  when there is a crowd. A challenge here is that the embedder's code could introduce potentially unbounded dependencies between the changes, and to our knowledge there is (as yet) no proof that the number of changes necessarily still respects the Square-Root critical rate. Preliminary work has shown that, when the embedder uses a code that minimizes an additive distortion, the optimal embedding probabilities depend on  $K$ : this is another respect in which a steganographer needs to take into account the size of the crowd.

Recall that we conservatively assume that the Warden knows everything about the stegosystem, including the method that the steganographer uses to scale their payload with their cover size. (This is a plausible assumption, for example in case there is a traitor in the stegosystem who leaks the details to the Warden, or in case the Warden has read the same publications as the steganographer.) Aside from the headline result about the effect of the crowd on the size of the payload, Subsect. 4.3 shows that proper scaling of the payload is essential, and choices other than  $M = O(\sqrt{N})$  permit the Warden to use negative information about the innocent actors better to locate the steganographer.

Our discussion that adds a zero-steganographer hypothesis has illustrated the need to generalize these results to a nonuniform prior. This is a direction for future work. Another aim for further work is to narrow the upper and lower bounds on the Warden's advantage. Consider embedding at or very close to the critical rate  $\frac{M^2}{N \log K} = r + o(1)$ . Claims 3 and 5 are able to give concrete upper bounds on the Warden's advantage, but Claims 2 and 4 give a very weak lower bound that is generally uninformative. That is because our approximations in (5) and (15) are crude. A delicate analysis of the maximum of  $K$  random variables in the regime where their means differ only slightly – even variables as well-behaved as the scores  $S_k$  which, it can be shown, have asymptotic normality under the hypotheses of this paper – is notoriously difficult because the maximum occurs when a variable is *not* near its mean and hence concentration results are useless.

More generally, we hope that this research will stimulate *practical* approaches to detecting a steganographer amongst multiple actors. This is an area which has relatively little attention [12, 15], compared with binary detection applied to a single image. Just as the single-actor, single-image Square-Root Law – a theoretical result about unrealistic probabilistic models – has been observed robustly in genuine image steganography [7], we predict that a well-designed steganographer detector should exhibit the  $O(\sqrt{\log K})$  law derived here.

## REFERENCES

- [1] Ross J. Anderson. 1996. Stretching the Limits of Steganography. In *Proc. 1st Information Hiding Workshop (Lecture Notes in Computer Science)*, Vol. 1174. Springer, 39–48. <https://doi.org/10.5555/647594.731520>
- [2] Herman Chernoff. 1952. A Measure of Asymptotic Efficiency for Tests of a Hypothesis Based on the Sums of Observations. *Annals of Mathematical Statistics* 23, 4 (1952), 409–507. <https://doi.org/10.1214/aoms/1177729330>
- [3] Rémi Cogranne and Jessica Fridrich. 2015. Modeling and Extending the Ensemble Classifier for Steganalysis of Digital Images Using Hypothesis Testing Theory. *IEEE Transactions on Information Forensics and Security* 10, 12 (2015), 2627–2642. <https://doi.org/10.1109/TIFS.2015.2470220>
- [4] Tomáš Filler, Jan Judas, and Jessica Fridrich. 2011. Minimizing Additive Distortion in Steganography using Syndrome-Trellis Codes. *IEEE Transactions on Information Forensics and Security* 6, 3 (2011), 920–935. <https://doi.org/10.1109/TIFS.2011.2134094>
- [5] Tomáš Filler, Andrew D. Ker, and Jessica Fridrich. 2009. The Square Root Law of Steganographic Capacity for Markov Covers. In *Media Forensics and Security XI (Proc. SPIE)*, Vol. 7254. SPIE, Article 08, 11 pages. <https://doi.org/10.1117/12.805911>
- [6] Dario Garcia-Garcia and Robert C. Williamson. 2012. Divergences and Risks for Multiclass Experiments. In *Proceedings of the 25th Annual Conference on Learning Theory (Proceedings of Machine Learning Research)*, Vol. 23. PMLR, Edinburgh, Scotland, 28.1–28.20.
- [7] Quentin Giboulot and Jessica Fridrich. 2019. Payload Scaling for Adaptive Steganography: An Empirical Study. *IEEE Signal Processing Letters* 26, 9 (2019), 1339–1343. <https://doi.org/10.1109/LSP.2019.2929435>
- [8] Wassily Hoeffding. 1963. Probability Inequalities for Sums of Bounded Random Variables. *J. Amer. Statist. Assoc.* 58, 301 (1963), 13–30. <https://doi.org/10.2307/2282952>
- [9] Andrew D. Ker. 2010. The Square Root Law Does Not Require a Linear Key. In *Proc. 11th Workshop on Multimedia and Security*. ACM, New York, NY, 213–223. <https://doi.org/10.1145/1854229.1854267>
- [10] Andrew D. Ker. 2011. A Curiosity Regarding Steganographic Capacity of Pathologically Nonstationary Sources. In *Media Watermarking, Security, and Forensics XIII (Proc. SPIE)*, Vol. 7880. SPIE, Article 0E, 12 pages. <https://doi.org/10.1117/12.871885>
- [11] Andrew D. Ker. 2017. The Square Root Law of Steganography: Bringing Theory Closer to Practice. In *Proc. 5th Workshop on Information Hiding and Multimedia Security*. ACM, New York, NY, 33–44. <https://doi.org/10.1145/3082031.3083235>
- [12] Andrew D. Ker and Tomas Pevný. 2014. The Steganographer is the Outlier: Realistic Large-Scale Steganalysis. *IEEE Transactions on Information Forensics and Security* 9, 9 (2014), 1424–1435. <https://doi.org/10.1109/TIFS.2014.2336380>
- [13] Jianhua Lin. 1991. Divergence Measures Based on the Shannon Entropy. *IEEE Transactions on Information Theory* 37, 1 (1991), 145–151. <https://doi.org/10.1109/18.61115>
- [14] Nigel P. Smart. 2015. *Cryptography Made Simple*. Springer. <https://doi.org/10.1007/978-3-319-21936-3>
- [15] Mingjie Zheng, Sheng-hua Zhong, Songtao Wu, and Jianmin Jiang. 2017. Steganographer Detection via Deep Residual Network. In *2017 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, Piscataway, NJ, 235–240. <https://doi.org/10.1109/ICME.2017.8019320>

## A TECHNICAL LEMMAS

LEMMA A.1. For all  $k > 1$ ,  $0 \leq \alpha \leq 1 - \frac{1}{k}$ ,

$$\begin{aligned} \frac{1}{k}(1 + k\alpha) \log(1 + k\alpha) + \frac{k-1}{k} \left(1 - \frac{k}{k-1}\alpha\right) \log\left(1 - \frac{k}{k-1}\alpha\right) \\ \geq \alpha^2 \frac{k}{k-2} \log(k-1) \stackrel{(ii)}{\geq} \alpha^2 \log k. \end{aligned} \quad (30)$$

PROOF. The case  $\alpha = 0$  can be disposed of immediately, as all terms are zero. It is then helpful to reparameterize, writing  $\beta = \alpha k$  and  $l = k - 1$ , so that  $l > 0$  and  $0 < \beta \leq l$ . The claim is then

$$\begin{aligned} \frac{1}{l+1}(1 + \beta) \log(1 + \beta) + \frac{l}{l+1} \left(1 - \frac{\beta}{l}\right) \log\left(1 - \frac{\beta}{l}\right) \\ \stackrel{(i)}{\geq} \beta^2 \frac{1}{(l-1)(l+1)} \log l \stackrel{(ii)}{\geq} \beta^2 \frac{1}{(l+1)^2} \log(l+1). \end{aligned}$$

Since  $\beta^2 > 0$  we may divide through and multiply by  $l + 1$  to obtain the equivalent inequalities

$$\frac{(1 + \beta) \log(1 + \beta) + (l - \beta) \log\left(1 - \frac{\beta}{l}\right)}{\beta^2} \stackrel{(i)}{\geq} \frac{\log l}{l - 1} \stackrel{(ii)}{\geq} \frac{\log(l + 1)}{l + 1}.$$

Call the function on the left  $g(\beta, l)$ : it is convex in  $\beta$  and

$$\frac{\partial g}{\partial \beta} = \frac{(-2 - \beta) \log(1 + \beta) + (\beta - 2l) \log\left(1 - \frac{\beta}{l}\right)}{\beta^3} = 0$$

if and only if  $\beta = l - 1$ , which must be the minimum with respect to  $\beta$ . Thus

$$g(\beta, l) \geq g(l - 1, l) = \frac{\log l}{l - 1},$$

establishing (i). The simple inequality (ii) comes from considering  $\phi(l) = (l + 1) \log l - (l - 1) \log(l + 1)$ , which is an increasing function with  $\phi(1) = 0$ , establishing (ii) for both  $l < 1$  and  $l > 1$ .  $\square$

LEMMA A.2. (i)  $D_{\text{KL}}(p + \epsilon, p) \geq 2\epsilon^2$  for all  $p$  and  $\epsilon$ .

(ii)  $D_{\text{KL}}(p + \delta, p + \epsilon) \leq \frac{(\epsilon - \delta)^2}{p(1-p)}$  for sufficiently small  $\delta$  and  $\epsilon$ .

(iii)  $D_{\text{KL}}(p, p + \epsilon) \sim \frac{\epsilon^2}{2p(1-p)}$  as  $\epsilon \rightarrow 0$ .

PROOF. (iii) comes from the Taylor expansion of  $D_{\text{KL}}(p, p + \epsilon)$ , and then (i) and (ii) follow using continuity and  $\frac{1}{p(1-p)} \geq 4$ .  $\square$

LEMMA A.3. Some concentration inequalities:

(i) (Chernoff bounds) If  $X_i = 0, 1$  and  $E[X_i] = p$ ,

$$P\left[\sum_{i=1}^N X_i \leq N(p - \epsilon)\right] \geq 1 - \exp(-ND_{\text{KL}}(p - \epsilon, p)), \text{ and}$$

$$P\left[\sum_{i=1}^N X_i \geq N(p + \epsilon)\right] \geq 1 - \exp(-ND_{\text{KL}}(p + \epsilon, p)).$$

(ii) (Hoeffding's inequality) If  $a_i \leq X_i \leq b_i$ ,  $b_i - a_i < C$  for all  $i$ , and  $\sum E[X_i] = E$ ,

$$P\left[\sum_{i=1}^N X_i \leq E + t\right] \geq 1 - \exp(-2t^2/NC^2), \text{ and}$$

$$P\left[\sum_{i=1}^N X_i \geq E - t\right] \geq 1 - \exp(-2t^2/NC^2).$$

PROOF. Standard results [2, 8].  $\square$

LEMMA A.4. If  $\delta \leq p, q \leq 1 - \delta$ ,  $|p - q| \geq \delta$ , and  $0 \leq \alpha \leq 1$ , then

$$\left|\log\left(1 + \alpha \frac{q-p}{p}\right)\right| \leq \alpha \frac{1-2\delta}{\delta}.$$

By symmetry, the same is true for  $\left|\log\left(1 - \alpha \frac{q-p}{1-p}\right)\right|$ .

PROOF. Let  $z = \log\left(1 + \alpha \frac{q-p}{p}\right)$ . If  $q > p$  then  $z$  is positive and maximized when  $p = \delta$  and  $q = 1 - \delta$ , and then

$$z = \log\left(1 + \alpha \frac{1-2\delta}{\delta}\right) \leq \alpha \frac{1-2\delta}{\delta}.$$

If  $q < p$  then  $z$  is negative and minimized when  $p = 1 - \delta$  and  $q = \delta$ , and then

$$z = \log\left(1 - \alpha \frac{1-2\delta}{1-\delta}\right) \geq \frac{-\alpha \frac{1-2\delta}{1-\delta}}{1 - \alpha \frac{1-2\delta}{1-\delta}} \geq \frac{-\alpha \frac{1-2\delta}{1-\delta}}{1 - \frac{1-2\delta}{1-\delta}} = -\alpha \frac{1-2\delta}{\delta}.$$

The symmetrical result follows by taking  $p' = 1 - p$  and  $q' = 1 - q$ .  $\square$