The Square Root Law of Steganography: Bringing Theory Closer to Practice

Andrew D. Ker
Department of Computer Science
University of Oxford
Oxford OX1 3QD, UK
adk@cs.ox.ac.uk

ABSTRACT

There are two interpretations of the term 'square root law of steganography'. As a *rule of thumb*, that the secure capacity of an imperfect stegosystem scales only with the square root of the cover size (not linearly as for perfect stegosystems), it acts as a robust guide in multiple steganographic domains. As a *mathematical theorem*, it is unfortunately limited to artificial models of covers that are a long way from real digital media objects: independent pixels or first-order stationary Markov chains. It is also limited to models of embedding where the changes are uniformly distributed and, for the most part, independent.

This paper brings the theoretical square root law closer to the practice of digital media steganography, by extending it to cases where the covers are Markov Random Fields, including inhomogeneous Markov chains and Ising models. New proof techniques are required. We also consider what a square root law should say about adaptive embedding, where the changes are not uniformly located, and state a conjecture.

ACM Reference format:
Andrew D. Ker. 2017. The Square Root Law of Steganography: Bringing Theory Closer to Practice. In *Proceedings of IH&MMSec '17, June 20–22, 2017, Philadelphia, PA, USA*, , 13 pages.
DOI: http://dx.doi.org/10.1145/3082031.3083235

1 INTRODUCTION

The phrase 'square root law of steganography' was first coined in [26], to mean that steganographic payloads should scale with the square root of the size of the cover unless truly perfect steganography is available. It was inspired by small-scale experimental evidence in [16], heuristic theoretical predictions from [17], and a first mathematical theorem in [18].

A square root law has significant consequences for the practice of steganography, since it implies that an imperfect steganographic channel has zero 'rate': the more payload is sent, the more sparsely it must be spread. The covert channel never completely dries up, but transmission times grow quadratically with the total payload ([19] seems to be the only literature on how practically to live within a square root law). This is especially so since perfect steganography is confined to extremely low-bandwidth scenarios, and is not achieved in digital media [24].

The theoretical work underpinning the square root law [2, 11, 18, 20–23] applies to highly artificial models of steganography, where the covers are typically assumed to have independent and identically distributed elements, or which form a stationary Markov chain, and an embedding operation that applies independently and identically to each location. We will survey the literature more thoroughly in Sect. 2.

The first proper experimental validation of a square root law was performed in [26], which analyzed detection accuracy versus cover and payload size for then-leading steganography and steganalysis algorithms, in raw and JPEG images. Close adherence to a square root capacity law was observed, even though pixels in digital images are far from independent or 1st-order Markov. To our knowledge, this 2008 study has yet to be repeated for modern image steganography and steganalysis, or for adaptive steganography with a knowing attacker [6], but in the linguistic domain the same law *has* been exhibited recently for adaptive steganography [35]. On the square root law's robustness, the survey [24] said

> 'What is remarkable about the square root law is that, although both asymptotic and proved only for artificial sources, it is robust and manifests in real life. This is despite the fact that ... empirical sources do not match artificial models.'

Our aim is to widen considerably the scope of artificial models for which a square root law holds, bringing the theory closer to the practice of digital media steganography.

In Sect. 2 we survey and collect existing square root laws into a common notation. In Sect. 3 we propose a different information-theoretic approach to demonstrating square root laws, and prove a lemma. Section 4 contains our main results, a new square root law for a Markov Random Field (MRF) cover model: one half applies to very broad classes of MRFs, the other to a more restricted class, but wide enough (Sect. 5) to include both inhomogeneous Markov chains and Ising models. In Sect. 6 we discuss how a square root law should be modified for *adaptive embedding*, where the embedding changes are neither uniformly located, nor necessarily

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

IH&MMSec '17, June 20–22, 2017, Philadelphia, PA, USA
© 2017 Copyright held by the owner/author(s). Publication rights licensed to ACM. 978-1-4503-5061-7/17/06. . . $15.00
DOI: http://dx.doi.org/10.1145/3082031.3083235

independent. We conjecture on the result that might be shown using an extension of our techniques. In Sect. 7 we draw conclusions and suggest further research.

2 SQUARE ROOT LAWS

The first publication to analyze asymptotic steganography capacity [18] set the format for square root laws. They concern covers of size n and a payload size, dependent on the cover size, $m(n)$. They make certain assumptions about the probability distributions of cover and stego objects. Under such assumptions, the laws give a *critical rate* for $m(n)$, which we will call $r(n)$[1]. Then

(i) if m is above the critical rate, $m(n)/r(n) \to \infty$, then *an asymptotically perfect detector exists*;

(ii) if m is below the critical rate, $m(n)/r(n) \to 0$, then *every detector is asymptotically random.*

We call (i) the *upper bound*, and (ii) the *lower bound*, on asymptotic payload size, and will make the asymptotic detection notions more precise in a moment.

Typically, if n is the number of embedding locations in the cover, and steganography is performed without *source coding* [10], the critical rate is $r = \sqrt{n}$, hence the name *square root law.* Source coding can increase this to $r = \sqrt{n}\log n$, but not beyond: we will consider this case in Sect. 6.

Let us fix some notation that can unify the square root laws we discuss. They concern a binary classification model of steganalysis: the detector has an observation X_n drawn from either a cover distribution $\mathcal{P}_n$, or a stego distribution $\mathcal{Q}_n^m$; they wish to determine which distribution it came from with low false positive (mistaking $\mathcal{P}_n$ for $\mathcal{Q}_n^m$) and false negative (vice versa) error probabilities. Here X_n is some n-dimensional object consisting of perhaps n pixels, transform coefficients, frames, or other type depending on the medium (but in this paper we will just call them 'pixels'). We have emphasised that the stego distribution depends on the payload size m.

A *detector* has to be parameterized by the size of the object it is considering: think of a sequence of sets $(A_1, A_2, \ldots)$ determining the positive classifications for each cover size; if observation $X_n \in A_n$ then X_n is believed to come from $\mathcal{Q}_n^m$ rather than $\mathcal{P}_n$. To say that *an asymptotically perfect detector exists* means that there is such a sequence with

$$\mathrm{P}_{X_n \sim \mathcal{P}_n}[X_n \in A_n] + \mathrm{P}_{X_n \sim \mathcal{Q}_n^m}[X_n \notin A_n] \to 0 \text{ as } n \to \infty, \quad (1)$$

which is of course equivalent to requiring that both the false positive and false negative rates tend to zero. We do not enforce a condition on *how fast* they must tend to zero, which will depend on how much $m(n)$ exceeds the critical rate $r(n)$. In square root laws, (1) cannot hold if the cover and stego distributions are not different: we emphasise that square root laws do not apply to *perfect steganography* where $\mathcal{P}_n$ and $\mathcal{Q}_n^m$ are identical. Therefore they must impose what we shall call a *no free bits* condition,[2] enforcing some difference between $\mathcal{P}_n$ and $\mathcal{Q}_n^m$: this condition depends on the cover model being studied.

To say that *every detector is asymptotically random* means

$$\mathrm{P}_{X_n \sim \mathcal{P}_n}[X_n \in A_n] + \mathrm{P}_{X_n \sim \mathcal{Q}_n^m}[X_n \notin A_n] \to 1 \text{ as } n \to \infty, \quad (2)$$

i.e. the detector's advantage over random guessing tends to zero. Such a property cannot hold if there are regions with no uncertainty in the cover distribution, so square root laws must impose what we shall call a *no determinism* condition, the form of which depends on the cover model. It will ensure that every cover has a probability bounded away from zero.

2.1 Existing Square Root Laws

Let us survey the existing literature on square root laws, identifying the cover and embedding models along with the *no determinism* and *no free bits* conditions on them. Note that the asymptotic results have not always been stated exactly in the form of (1) and (2) but are equivalent, or can be adapted, to them.

The first publication [18] is an outlier. It concerns steganography in n independent objects and makes the strong assumption that the detector reduces each to a *one-dimensional* continuous observation; the effect of an embedding operation in one cover is to shift this observation, in proportion to the payload embedded (which enforces a *no free bits* condition). This is a simple abstraction of quantitative steganalysis [30], but not very realistic. To make proof in [18] work, the observations must be independent and the second derivative of their log density bounded below. The *no determinism* condition is that the support of these observations is infinite.

The simplest square root law of the type we study here, single objects with n pixels, is found in [20]. The cover model $\mathcal{P}_n$ consists of pixels which are independent and identically distributed discrete random variables with mass function $p(x)$. There are two common embedding models:

(a) replace a uniformly-chosen m out of n pixels, or

(b) independently with probability $\frac{m}{n}$ replace each pixel,

by random variables with mass function $q(x)$. (a) is a good model for simple uncoded embedding operations such as bit flipping; (b) is not a good model for a fixed payload, but it is rather easier to analyze because the stego pixels remain independent. The *no free bits* condition is that p and q differ. The *no determinism* condition is that $p(x)$ is nonzero for all x. Both cases (a) and (b) are proved in [20] (Thms. 2 and 1, respectively). The case (b) is extended to independent but not identically distributed pixels in [23, Thm. 2].

The proofs of these results depend critically on independence between the n elements in $\mathcal{P}_n$. This may be a reasonable model for batch steganography (the n elements are separate objects) with a fixed cover source, but it is absurd for pixels or transform coefficients in a digital image, or frames from audio or video. In [11] the cover distribution $\mathcal{P}_n$ is generalized to a stationary finite-valued Markov chain. It is required that its transition matrix contain no zeros, which functions as a conditional *no determinism* criterion. The embedding (termed 'mutually independent embedding') is of

[1] This is not a *rate* in the signal processing sense, since it is almost never linear in n.

[2] The payload bits are not *free* because each modification induces a change to the distribution, and therefore incurs a distortion *cost*.

the type (b), above, and the *no free bits* condition is that the second order co-occurrence probabilities are not preserved by the embedding process [11, Assumption 3].

A different generalization can be made to account for *source coding* in the embedding process [13], which every modern stegosystem should employ [24]. Source coding has two effects on square root laws: it may cause embedding changes that are not independent, and the number of embedding changes is sublinear in the size of the payload. We will discuss the former in Sect. 6. For the latter, the critical rate is increased from $r(n) = \sqrt{n}$ to $r(n) = \sqrt{n}\log n$ (rate-distortion bounds show that it cannot be increased further). A square root law for source coding was proved in [21], for the independent pixel model. The *no free bits* and *no determinism* conditions are the same as in [20].

The above results all assume that the detector has perfect knowledge of both (a) the cover distribution $\mathcal{P}_n$, and (b) the embedding process and payload size m, which gives perfect knowledge of $\mathcal{Q}_n^m$. It is not difficult to generalize them to the case where m is unknown (although such proofs have not been published) but the case where $\mathcal{P}_n$ is imperfectly known has proved more challenging, in part because the information theoretic machinery does not deal well with compound hypotheses. In the case of an extremely simple i.i.d. Bernoulli cover model, where the Bernoulli parameter is learned empirically, a modified square root law is proved in [22]: here, the critical rate depends on the amount of training data available to the detector, but the square root order is maintained as long as the amount of training data is at least linear in the amount of testing data. There is investigation of a highly artificial nonstationary version in [23]. To our knowledge there has been no further progress on square root laws where the detector has imperfect knowledge. In this work we will confine our attention to the perfect knowledge case, but return to it for future study.

Finally, there is a square root law for steganography in continuous noisy channels, [2]. There is not space to re-state the result here, but we note that what seem major differences with the previous models (the signal is continuous; there is no pre-existing cover; both the receiver and the detector receive the sender's signal subject to different, independent, additive Gaussian noise) are not so great on closer examination. The noise functions as a cover, the *no free bits* condition is enforced by additivity of the signal, and the *no determinism* condition by nonzero noise amplitude. The proof of the upper bound is a close continuous analogue of the i.i.d. discrete case. For the lower bound, as well as a continuous analogue of the information theoretic arguments for the discrete case, it is also necessary to construct a codebook with enough robustness to defeat the noise.

We should briefly discuss some non-square root capacity results in the literature. These are linear capacity laws for perfect steganography [3, 5, 31, 34]. Such systems are closer to *cover generation* than the *cover modification* paradigm we consider. If the cover source is completely known the embedder will mimic it, or otherwise they learn about it empirically. These systems have the opposite of a *no free bits* condition: either exactly or asymptotically, all payload bits are 'free', because stego objects match the cover distribution. However, we stress that systems are only perfect *for a particular cover model* (i.i.d. in the case of [3, 5, 34], k-order Markov in the case of [31]). If the true cover source deviates even slightly from the artificial model, this could be exploited (with exponentially vanishing error rates) by a knowledgeable detector. Perhaps this explains why (imperfect) cover modification is overwhelming dominant for digital media steganography, and square root laws are observed in practice.

We argue that the largest gap between theory of published square root laws and the practice of multimedia steganography is the cover model. The most complex so far analyzed is a first-order Markov source which, although having simple memory, is a poor model for images because of its one-dimensional nature. Although [11] claims that the result can be generalized to a Markov chain of overlapping patches, it is not clear that this is true: the *no determinism* condition bans zeros in the transition matrix, which would have to be present in a transition matrix of overlapping patches. In any case, it still would not model two-dimensional dependence as fully as, say, a Markov Random Field. At a stretch one might model a digital video as a one-dimensional chain of frames (a Markov version of the batch steganography model), but we know from video compression that inter-frame dependence is long-range and bidirectional.

The main focus of this work is to prove a square root law for a wide class of multidimensional covers, with almost arbitrary finite-range dependence between the elements. This includes inhomogeneous Markov chains and some $d{\geq}2$-dimension Ising models. We will do so in Sect. 4.

Furthermore, we would wish to combine such a cover model with an embedding operation that permits source coding. At present, the results of [21] and [11] cannot easily be fused. Now a *no determinism* cover condition seems too strong: if parts of the cover are deterministic (too risky to change) they will simply be avoided by the embedder, which is possible if source coding is employed. We should only require that *enough* cover locations satisfy a (conditional) no determinism condition. This must be part of a square root law for *adaptive steganography*, whose form we discuss briefly in Sect. 6.

3 TOTAL VARIATION

Typically the upper bound of a square root law (recall: if m is above the critical rate then an asymptotically perfect detector exists) is proved by construction. A detector is proposed, and its error rates bounded using tail inequalities. Typically the lower bound (recall: if m is below the critical rate then every detector is asymptotically random) is proved by showing that $D_{\mathrm{KL}}(X \sim \mathcal{P}_n, X \sim \mathcal{Q}_n^m) \to 0$, where

$$D_{\mathrm{KL}}(X \sim \mathcal{P}, X \sim \mathcal{Q}) = \sum_{x \in \mathcal{X}} \mathrm{P}_{\mathcal{P}}[X = x] \log\Big(\tfrac{\mathrm{P}_{\mathcal{P}}[X=x]}{\mathrm{P}_{\mathcal{Q}}[X=x]}\Big)$$

is the Kullback-Leibler Divergence (KLD) from distribution $\mathcal{P}$ to $\mathcal{Q}$. The sum over $\mathcal{X}$ denotes all possible (nonzero probability) values of X; in digital media objects such distributions will be finite-valued, and in this paper we will consider the finite case. We expect that generalizations to infinite discrete and continuous probability measures are also possible.

KLD has many notations in the literature; we have included name of the random variable X in case the distributions $\mathcal{P}$ and $\mathcal{Q}$ define other random variables not available to the detector. Note that KLD is only well-defined if X has the same support under distributions $\mathcal{P}$ and $\mathcal{Q}$.

Proving that $D_{\mathrm{KL}}(X \sim \mathcal{P}_n, X \sim \mathcal{Q}_n^m) \to 0$ (known as *convergence in KLD*) is sufficient to show that every detector is asymptotically random, something known since Cachin's early work on the information theory of steganography [3]. Cachin defined a stegosystem to be 'ϵ-secure' if $D_{\mathrm{KL}}(X \sim \mathcal{P}_n, X \sim \mathcal{Q}_n^m) < \epsilon$, and bounded the performance of detectors via an information processing inequality. It is likely that theoreticians have focused on KLD because of Cachin's example. Furthermore, KLD has useful connections with *error exponents* in the case of batch imperfect steganography in stationary independent covers with a constant embedding rate, although the square root law tells us that such a rate would be ill-advised.

However, convergence in KLD is a strictly stronger condition than every detector being asymptotically random. For example, imagine 'cover' and 'stego' objects of n independent pixels, distributed according to

$$\mathrm{P}_{\mathcal{P}_n}[X_i = 1] = \tfrac{1}{n^2}, \quad \mathrm{P}_{\mathcal{Q}_n}[X_i = 1] = e^{-n^2}. \tag{3}$$

It is routine to show $D_{\mathrm{KL}}\big((X_1, \ldots, X_n) \sim \mathcal{P}_n, (X_1, \ldots, X_n) \sim \mathcal{Q}_n\big) \to \infty$ as $n \to \infty$, yet asymptotically with probability one every cover object is entirely zeros, and so is every stego object, so any 'detector' that tries to discriminate between the two cases is asymptotically random. It follows that 'ϵ-security', for ϵ small, is a sufficient but not necessary condition for steganographic security.

Furthermore, KLD is not easy to work with. There is the strong requirement that $\mathcal{P}_n$ and $\mathcal{Q}_n$ have exactly the same nonzero probability events, even if those probabilities tend to zero and are therefore negligible for detection. The logarithm in the definition makes analysis complicated (see e.g. [21]). And, while easy to bound for independent pixels (it is additive across independent components), KLD can be extremely awkward to bound for dependent components: the square root law for Markov chains [11] uses a highly technical lemma, proved in [7], that requires difficult analytic arguments. Perhaps this is why square root laws have not yet been extended to more realistic cover distributions.

In this work we use a different information theoretic quantity. The Total Variation (TV) between $\mathcal{P}$ and $\mathcal{Q}$, again distributions defining a random variable X over the set $\mathcal{X}$, is

$$D_{\mathrm{TV}}(X \sim \mathcal{P}, X \sim \mathcal{Q}) = \tfrac{1}{2} \sum_{x \in \mathcal{X}} \Big| \mathrm{P}_{X \sim \mathcal{P}}[X = x] - \mathrm{P}_{X \sim \mathcal{Q}}[X = x] \Big|.$$

It is not necessary for the zero probability values of X to be identical under $\mathcal{P}$ and $\mathcal{Q}$. Note that some omit the constant factor. Other authors have briefly used TV in square root laws, but only by immediate appeal to Pinsker's inequality

$$2 D_{\mathrm{TV}}(X \sim \mathcal{P}, X \sim \mathcal{Q})^2 \leq D_{\mathrm{KL}}(X \sim \mathcal{P}, X \sim \mathcal{Q}). \tag{4}$$

An equivalent formula for TV is given by

$$D_{\mathrm{TV}}(X \sim \mathcal{P}, X \sim \mathcal{Q}) = \sup_{A \subseteq \mathcal{X}} \Big| \mathrm{P}_{X \sim \mathcal{P}}[X \in A] - \mathrm{P}_{X \sim \mathcal{Q}}[X \in A] \Big|,$$

which gives a strong connection with detection:

LEMMA 3.1. *An asymptotically perfect detector based on X exists if and only if $D_{\mathrm{TV}}(X \sim \mathcal{P}_n, X \sim \mathcal{Q}_n^m) \to 1$. Every detector based on X is asymptotically random if and only if $D_{\mathrm{TV}}(X \sim \mathcal{P}_n, X \sim \mathcal{Q}_n^m) \to 0$.*

The proof is elementary, for example see [27, Thm. 13.1.1]. In example (3), $D_{\mathrm{TV}}((X_1, \ldots, X_n) \sim \mathcal{P}_n, (X_1, \ldots, X_n) \sim \mathcal{Q}_n) \to 0$, proving asymptotic undetectability for this example, and demonstrating how TV can give a more refined analysis than KLD.

Total variation can still be extremely difficult to compute, and indeed there are few closed formulae for total variation between standard distributions, but it can be easier to bound than KLD. It is not additive, but is subadditive across independent components. Unlike KLD, it satisfies the triangle inequality, and in a moment we will prove a useful result about side information and TV.

3.1 Side Information for the Detector

As with KL divergence, TV cannot be decreased by the presence of side information. Because TV is more forgiving of impossible events in $\mathcal{P}$ that are not impossible in $\mathcal{Q}$ (as long as their probability tends to zero), we can prove a lemma that will be used in our main result.

We need a conditional version of total variation, which does not seem a widely-used concept (it appears in [32]):

Definition 3.2. Let S be a random variable that has the same distribution under $\mathcal{P}$ and $\mathcal{Q}$. Then

$$D_{\mathrm{TV}}(X \sim \mathcal{P}, X \sim \mathcal{Q} \,|\, S = s) = \\ \tfrac{1}{2} \sum_{x \in \mathcal{X}} \Big| \mathrm{P}_{(X,S) \sim \mathcal{P}}[X = x \,|\, S = s] - \mathrm{P}_{(X,S) \sim \mathcal{Q}}[X = x \,|\, S = s] \Big|.$$

LEMMA 3.3. *Let $\mathcal{S}$ denote the possible values of S. If $\mathcal{S}$ is partitioned into $\mathcal{S}_0$ and $\mathcal{S}_1$ then*

$$\begin{aligned} &D_{\mathrm{TV}}(X \sim \mathcal{P}, X \sim \mathcal{Q}) \\ &\quad \leq \mathrm{P}[S \in \mathcal{S}_0] + \max_{s \in \mathcal{S}_1} D_{\mathrm{TV}}(X \sim \mathcal{P}, X \sim \mathcal{Q} \,|\, S = s). \end{aligned} \tag{5}$$

PROOF. Abbreviating $X = x$ as simply X,

$$\begin{aligned} &D_{\mathrm{TV}}(X \sim \mathcal{P}, X \sim \mathcal{Q}) \\ &\quad = \tfrac{1}{2} \sum_{x \in \mathcal{X}} \Big| \sum_{s \in \mathcal{S}} \mathrm{P}_{\mathcal{P}}[S]\, \mathrm{P}_{\mathcal{P}}[X \,|\, S] - \sum_{s \in \mathcal{S}} \mathrm{P}_{\mathcal{Q}}[S]\, \mathrm{P}_{\mathcal{Q}}[X \,|\, S] \Big| \\ &\quad \overset{(i)}{\leq} \tfrac{1}{2} \sum_{s \in \mathcal{S}} \sum_{x \in \mathcal{X}} \mathrm{P}[S] \, \Big| \mathrm{P}_{\mathcal{P}}[X \,|\, S] - \mathrm{P}_{\mathcal{Q}}[X \,|\, S] \Big| \\ &\quad = \sum_{s \in \mathcal{S}} \mathrm{P}[S] \, D_{\mathrm{TV}}(X \sim \mathcal{P}, X \sim \mathcal{Q} \,|\, S) \\ &\quad \overset{(ii)}{\leq} \sum_{s \in \mathcal{S}_0} \mathrm{P}[S] + \max_{s \in \mathcal{S}_1} D_{\mathrm{TV}}(X \sim \mathcal{P}, X \sim \mathcal{Q} \,|\, S) \sum_{s \in \mathcal{S}_1} \mathrm{P}[S] \\ &\quad \leq \mathrm{P}[S \in \mathcal{S}_0] + \max_{s \in \mathcal{S}_1} D_{\mathrm{TV}}(X \sim \mathcal{P}, X \sim \mathcal{Q} \,|\, S). \end{aligned}$$

Above, (i) uses $\mathrm{P}_{\mathcal{P}}[S = s] = \mathrm{P}_{\mathcal{Q}}[S = s]$ and the triangle inequality. (ii) uses $D_{\mathrm{TV}}(X \sim \mathcal{P}, X \sim \mathcal{Q} \mid S = s) \leq 1$. □

We will think of S as a 'hint' for the detector. The set $\mathcal{S}_0$ are 'bad hints' that give too much information: we will ensure that the probability of a bad hint tends to zero. The hints in $\mathcal{S}_1$ are 'good hints' that give enough conditional information to the detector to make $D_{\mathrm{TV}}(X \sim \mathcal{P}, X \sim \mathcal{Q} \mid S = s)$ amenable to analysis, but little enough information so that it still tends to zero, for any $s \in \mathcal{S}_1$, below the critical rate.

We also state a simple result confirming that side information independent of the observed variables conveys nothing:

Lemma 3.4. *If X is independent of S under both $\mathcal{P}$ and $\mathcal{Q}$, and S has the same distribution under $\mathcal{P}$ and $\mathcal{Q}$, then*

$$D_{\mathrm{TV}}\big((X,S)\sim\mathcal{P}, (X,S)\sim\mathcal{Q}\big) = D_{\mathrm{TV}}(X \sim \mathcal{P}, X \sim \mathcal{Q}). \quad (6)$$

4 A SQUARE ROOT LAW FOR DEPENDENT PIXELS

Our square root law applies to covers whose pixels form a Markov Random Field (MRF) [14] with bounded degree. With only the addition of a *no asymptotic determinism* condition, we will prove the lower bound in Subsect. 4.1. Further assumptions (ensuring exponential decay of covariance) will be needed, as well as a *no free bits* condition, to prove the upper bound in Subsect. 4.2.

We will need to identify individual pixels within an n-element cover, which we write $\boldsymbol{X} = (X_1, \ldots, X_n)$[3]. A bounded-degree MRF may be concisely described as follows: each pixel X_i is allowed to depend directly only on a *neighbourhood* $\boldsymbol{N_i}$, which is of bounded size. Any dependence with X_i outside the neighbourhood is indirectly via $\boldsymbol{N_i}$.

More precisely, there is a universal constant D (not depending on n) and for each i there exists N_i such that

$$N_i \subset \{1, \ldots, n\}, \quad i \notin N_i, \quad |N_i| \leq D. \qquad \textbf{(C1)}$$

(We will label our assumptions, as above, in order to refer to them later.) We also require that neighbourhoods are symmetric in the sense that

$$i \in N_j \text{ if and only if } j \in N_i, \qquad \textbf{(C2)}$$

for all i and j.

We write $\boldsymbol{N_i} = (X_j \mid j \in N_i)$ and the non-neighbourhood as $\overline{\boldsymbol{N_i}} = (X_j \mid j \notin N_i, j \neq i)$. The local Markov property of MRFs is that, conditional on $\boldsymbol{N_i}$, X_i is independent of $\overline{\boldsymbol{N_i}}$. This is sometimes written

$$X_i \perp\!\!\!\perp \overline{\boldsymbol{N_i}} \mid \boldsymbol{N_i}. \qquad \textbf{(C3)}$$

An example of a two-dimensional model of this type is in Fig. 1, (an Ising model, of which more in Subsect. 5.2). The pixels are indexed (i,j) and the distribution of each depends on the values of its four immediate neighbours, so that there are both horizontal and vertical dependencies. Here $D = 4$ and $N_{(i,j)} = \{(i-1,j), (i+1,j), (i,j-1), (i,j+1)\}$. There will also be some boundary conditions, either periodic (pixels at the top/left and bottom/right are neighbours) or imposed via auxiliary (fixed) edge rows and columns.

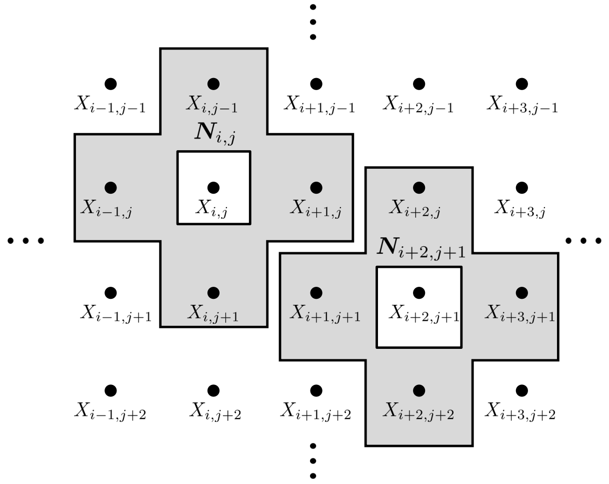

Figure 1: An example of a MRF cover model for which the square root law can hold. Each pixel X_{ij} is dependent on its immediate neighbours, and conditionally independent of the rest.

For embedding, we use the model of [20]. It is case (a) from Subsec. 2.1, where exactly $m = m(n)$ *embedding locations* are selected; they are changed by some random procedure[4], identical at each location and independent of other changes and the locations used. Unused locations are unchanged. In a square root law for adaptive embedding it makes more sense to consider case (b) from Subsec. 2.1, which is the *mutually independent embedding* model of [11], for which see Sect. 6. Our model seems more natural for uncoded embedding such as LSB or Ternary Embedding, where the number of changes is always bounded by a multiple of the payload size.

Formally, let $L = \{L_1, \ldots L_m\}$ be the embedding locations. We assume:

$$\{L_1, \ldots L_m\} \text{ is drawn uniformly from } \{1, \ldots, n\}. \qquad \textbf{(E1)}$$

Suppose that $\beta(x,y)$ is the probability that pixel value x is changed to y at each embedding location. Write $\boldsymbol{X}_{\backslash \boldsymbol{L}}$ for the vector of pixels not in L. Then the stego distribution is given by

$$\begin{aligned}&\mathrm{P}_{\mathcal{Q}_n^m}[\boldsymbol{X_L} = \boldsymbol{y}, \boldsymbol{X}_{\backslash \boldsymbol{L}} = \boldsymbol{z} \mid L] = \\ &\sum_{x_1} \cdots \sum_{x_m} \mathrm{P}_{\mathcal{P}_n}[\boldsymbol{X_L} = \boldsymbol{x}, X_{\backslash L} = \boldsymbol{z}] \prod_i \beta(x_i, y_i). \qquad \textbf{(E2)}\end{aligned}$$

We need one further condition, ensuring that no pixel value is made impossible by the embedding process. There should exist $\delta > 0$ such that

$$\text{for all } y \text{ there exists } x \text{ with } \beta(x,y) \geq \delta. \qquad \textbf{(E3)}$$

[3] This is slightly different from the notation of Sect. 2, where X_n represented an entire cover of size n.

[4] The randomness comes from the embedding key and payload. It may be assumed that both are indistinguishable from uniform random, which is a good model for compressed or encrypted data and well-chosen keys.

Together, (**E1**) and (**E2**) describe how the stego distribution $\mathcal{Q}_n^m$ is derived from the cover distribution $\mathcal{P}_n$. Observe that the stego distribution will contain dependencies not present in the cover MRF, and in general need not be a MRF of bounded degree, because of the non-local property that there are never more than m changes[5].

4.1 Lower Bound

The lower bound can be proved without further conditions on the cover and embedding models, except for a no determinism condition. It requires that each cover pixel value X_i is possible, conditional on its neighbourhood: this prevents a certain stego pattern from occurring that could never occur in covers, leading to a perfect detector. In fact, such a condition is usually included in the definition of a MRF (all probabilities must be positive) but we need a more uniform bound since we consider covers of growing size, and potentially anisotropic (nonstationary).

Thus we assume a *no asymptotic determinism* condition, also banning the likelihood of any pixel from tending to zero as $n \to \infty$ (it parallels the requirement in [23]). There exists a universal $\epsilon > 0$ (not depending on n) such that

$$P(X_i = x_i \,|\, \boldsymbol{N}_i = \boldsymbol{n_i}) \geq \epsilon \quad \text{for all } i,\ x_i,\ \boldsymbol{n_i}. \qquad (\mathbf{NAD})$$

THEOREM 4.1. *Assume* (**C1**), (**C2**), (**C3**), (**E1**), (**E2**), (**E3**), *and* (**NAD**). *If m is below the critical rate, $m(n)/\sqrt{n} \to 0$, then every detector is asymptotically random.*

PROOF. We will see that, with probability tending to one, no embedding change will lie in the neighbourhood of another. Conditional on their neighbourhoods, the pixels involved are independent. So if we give the detector side information about which pixels have been changed (which can only make their task easier) then we have reduced the problem to the lower bound of an independent pixel square root law.

However, we cannot give the detector the exact embedding locations, because this would effectively reduce n to m (part of the essence of a square root law is in the detector's uncertainty about where to look). Instead, we give them a *shortlist* of *possible* embedding locations, chosen so that conditional independence still holds, but of length $O(n)$ so that they still have plenty of confusion about the true embedding locations.

Of course, the embedding process does not really produce a shortlist of possible embedding locations, but we can create one by a method akin to those used in coupling arguments.

Identify a set of embedding locations $\{X_{l_1}, \dots, X_{l_k}\}$ by their indices $L = \{l_1, \dots, l_k\}$. Let $\mathcal{L}^k$ denote all possible sets of k embedding locations:

$$\mathcal{L}^k = \{L \subset \{1, \dots, n\} \,|\, |L| = k\}.$$

Where $i \in N_j$ we can say that the embedding locations i and j *interfere*, because their distributions are not conditionally independent. Let $\mathcal{L}_0^k$ be the sets of embedding locations where at least two interfere:

$$\mathcal{L}_0^k = \{L \in \mathcal{L}^k \,|\, i \in N_j \text{ for some } i, j \in L\},$$

and $\mathcal{L}_1^k = \mathcal{L}^k \setminus \mathcal{L}_0^k$. According to (**E1**), L is drawn uniformly from $\mathcal{L}^m$. If $L \in \mathcal{L}_0^m$, set $S = L$. Otherwise, set $s = \lceil n/2D \rceil$ and choose uniformly

$$S \in \{S \in \mathcal{L}_1^s \,|\, L \subseteq S\}.$$

By construction, this still chooses L uniformly $\mathcal{L}^m$, but the side information S, a shortlist of $s = O(n)$ non-interfering locations in which the true m locations can be found, can be generated *whether or not embedding takes place*. It will be used to bound the total variation between cover and stego objects. The cases where $L \in \mathcal{L}_0^m$ are special, since no non-interfering shortlist can contain L. The embedder may as well give up and confess guilt to the detector in such a case, because it has negligible probability:

$$\begin{aligned}
1 - P[L \in \mathcal{L}_0^m] &\overset{(i)}{=} \prod_{i=2}^{m} 1 - \frac{1}{n}\Big|\textstyle\bigcup_{j=1}^{i-1} N_{l_j} \cup \{l_j\}\Big| \\
&\overset{(ii)}{\geq} \prod_{i=2}^{m} 1 - \frac{(D+1)(i-1)}{n} \\
&\overset{(iii)}{\geq} \prod_{i=2}^{m} \exp\big(\tfrac{-2(D+1)(i-1)}{n}\big) \\
&\geq \exp\big(\tfrac{-(D+1)m^2}{n}\big)
\end{aligned}$$

For (i), consider adding l_i to non-interfering set $\{l_1, \dots, l_{i-1}\}$: its location is uniformly chosen, and we must avoid choosing from $l_1, \dots, l_{i-1}$ or their neighbourhoods. By (**C2**) this also ensures that none of $l_1, \dots, l_{i-1}$ lie in N_{l_i}. (ii) uses (**C1**). (iii) follows from $1 - x \geq \exp(-2x)$, at least for $x \leq \frac{1}{2}$: here $x \leq \frac{(D+1)m}{n}$, which certainly tends to zero if $m^2/n \to 0$. This establishes that, below the critical rate,

$$p_0 = \mathrm{P}[L \in \mathcal{L}_0^m] \to 0. \qquad (7)$$

Next, fix any nonempty shortlist $S = \{l_1, \dots, l_s\}$. Write $\boldsymbol{X_S} = (X_{l_1}, \dots, X_{l_s})$ for the pixels on the shortlist, and $\boldsymbol{N} = \bigcup_{i=1}^{s} \boldsymbol{N_{l_i}}$ for their neighbourhoods. Write $\boldsymbol{R} = \boldsymbol{X} \setminus \boldsymbol{N} \setminus \boldsymbol{X_S}$ (the remaining pixels). This decomposition is illustrated in Fig. 2. We can ignore the locations $\boldsymbol{R}$, since

$$\begin{aligned}
&D_{\mathrm{TV}}(\boldsymbol{X} \sim \mathcal{P}_n, \boldsymbol{X} \sim \mathcal{Q}_n^m \,|\, S, \boldsymbol{N}) \\
&\quad = D_{\mathrm{TV}}(\boldsymbol{X_S} \sim \mathcal{P}_n, \boldsymbol{X_S} \sim \mathcal{Q}_n^m \,|\, \boldsymbol{N}), \qquad (8)
\end{aligned}$$

because the MRF property of the cover (**C3**) implies $\boldsymbol{X_S} \perp\!\!\!\perp \boldsymbol{R} \mid \boldsymbol{N}$, and using (6). Conditional on $\boldsymbol{N}$, we have reduced the problem to a cover of independent random variables, since the $X_{l_i} \perp\!\!\!\perp X_{l_j} \mid \boldsymbol{N}$. The stego object does not have the same property, because choosing *exactly* m of n locations (conditional on the shortlist, m of s) introduces weak dependency, but this problem has already been attacked in [20].

Fix $\boldsymbol{N}$, and write $\mathcal{P}_s$ for the conditional cover distribution of $\boldsymbol{X_S}$ given $\boldsymbol{N}$. This is the independent product of the mass functions for each shortlisted cover element $p_i(k) = \mathrm{P}[X_{l_i} =$

[5] The results of this paper can be modified for the mutually independent embedding model, where each location is in L independently with probability m/n. A small addition is needed to the lower bound proof of Subsect. 4.1, to show that too many embedding locations has negligible probability. The upper bound proof in Subsect. 4.2 still applies, but it could be simplified since there are no long-range dependencies.

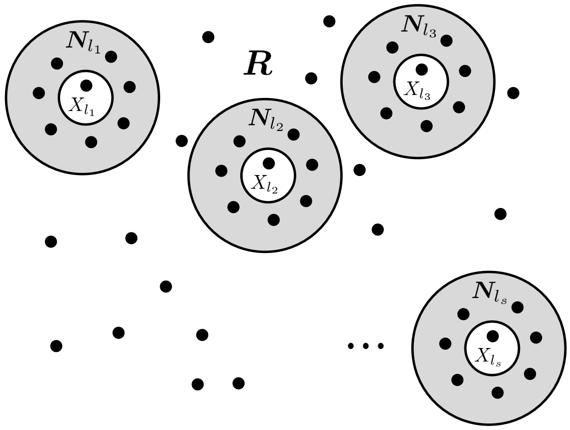

Figure 2: Decomposition of $\boldsymbol{X}$ into a non-interfering shortlist $(X_{l_1},\dots,X_{l_s})$, their neighbourhoods, and remaining pixels $\boldsymbol{R}$. Conditional on $\bigcup \boldsymbol{N_{l_i}}$, the X_{l_i} are independent of each other and $\boldsymbol{R}$. m of the s shortlisted locations are used for embedding.

$k \,|\, \boldsymbol{N_{l_i}}]$. Write $\mathcal{Q}_s^m$ for the conditional stego distribution of $\boldsymbol{X_S}$ given $\boldsymbol{N}$. By (**E2**),

$$\mathrm{P}_{\mathcal{Q}_s^m}[\boldsymbol{X_S} = (x_1 \dots x_s)] = \frac{1}{\binom{s}{m}} \sum_{\substack{L \subseteq \{1,\dots,s\},\\ |L|=m}} \prod_{i \notin L} p_i(x_i) \prod_{i \in L} q_i(x_i)$$

where

$$q_i(y) = \sum_x \beta(x,y) p_i(x)$$

is the mass function of element l_i if the embedding operation is applied to it.

Applying Pinsker's inequality (4),

$$\begin{aligned} 2D_{\mathrm{TV}}(\boldsymbol{X_S} \sim \mathcal{P}_n, &\boldsymbol{X_S} \sim \mathcal{Q}_n^m \,|\, \boldsymbol{N})^2 \\ &\le D_{\mathrm{KL}}(\boldsymbol{X_S} \sim \mathcal{P}_s, \boldsymbol{X_S} \sim \mathcal{Q}_s^m). \qquad (9) \end{aligned}$$

KL divergence between distributions of this form has been studied. In the proof of [20, Thm. 2(ii)], it is bounded above by Cm^2/s, for a constant C, subject to some conditions. We need to generalize the proof, however, because [20] does not permit the pixels to have different distribution. But the change is not difficult, and we spare the reader an almost identical proof. In [20] the distribution of $p(x)$ is required to have two properties. First, $p(x) > 0$ for all x. This must be modified to $p_i(x) > 0$ for all i and x, which follows from (**NAD**). Second, in the proof of the 'embedding probability lemma' in [20, Appendix A], the key line is

'there exists $c > 0$ such that, for all x, $p(x) \le cq(x)$.'

In the non-identical distribution case, we require a positive constant c such that, for all i and x, $p_i(x) \le cq_i(x)$. By (**E3**) and (**NAD**), $q_i(x) \ge \epsilon\delta$, so $c = 1/\epsilon\delta$ will do. Combining this bound with (8) and (9), we have

$$2D_{\mathrm{TV}}(\boldsymbol{X} \sim \mathcal{P}_n, \boldsymbol{X} \sim \mathcal{Q}_n^m \,|\, S, \boldsymbol{N})^2 \le Cm^2/s. \qquad (10)$$

Furthermore, because c is independent of S and $\boldsymbol{N}$, so is C.

Putting this together we have shown, when $m/\sqrt{n} \to 0$,

$$\begin{aligned} &D_{\mathrm{TV}}(\boldsymbol{X} \sim \mathcal{P}_n, \boldsymbol{X} \sim \mathcal{Q}_n^m) \\ &\quad \overset{(i)}{\le} p_0 + \max_{S \in \mathcal{L}_1} D_{\mathrm{TV}}(\boldsymbol{X} \sim \mathcal{P}_n, \boldsymbol{X} \sim \mathcal{Q}_n^m \,|\, S) \\ &\quad \overset{(ii)}{\le} p_0 + \max_{\boldsymbol{N} \in \mathcal{N}} \max_{S \in \mathcal{L}_1} D_{\mathrm{TV}}(\boldsymbol{X} \sim \mathcal{P}_n, \boldsymbol{X} \sim \mathcal{Q}_n^m \,|\, S, \boldsymbol{N}) \\ &\quad \overset{(iii)}{\le} p_0 + \sqrt{CDm^2/n} \\ &\quad \overset{(iv)}{\to} 0. \end{aligned}$$

Here (i) uses (5) with $\mathcal{S}_0 = \mathcal{L}_0^m$, $\mathcal{S}_1 = \mathcal{L}_1^s$; (ii) uses (5) again with $S_0 = \emptyset$, and $\mathcal{N}$ all possible values of $\boldsymbol{N}$ given S; (iii) uses (10) and $s \ge n/2D$; and finally (iv) uses (7). Thanks to Lemma 3.1, we have proved that every detector is asymptotically random below the critical rate. □

4.2 Upper Bound

The upper bound that we prove here is not as general, making further assumptions of the cover MRF: polynomial growth in neighbourhood size, and exponential decay of covariance.

Let us define a *distance* between locations to mean the distance in the neighbourhood graph of the MRF. That is, $d(i,j)$ is the smallest d with

$$l_0 = i,\ l_d = j,\ l_i \in N_{i-1} \text{ for } i = 1,\dots,d.$$

We require a polynomial $p(x)$ with the property that, for all i and $d \ge 0$,

$$\big|\{j \,|\, d(i,j) \le d\}\big| \le p(d). \qquad \textbf{(C4)}$$

This bans tree-like topologies, for which d-neighbourhoods can be exponentially large.

Since we will now focus on distributions of pixels jointly with their neighbours, it will convenient to abbreviate

$$R_i = \{i\} \cup N_i, \quad \boldsymbol{R_i} = (X_i, \boldsymbol{N_i}).$$

Observe that $R_i \cap R_j \ne \emptyset$ if and only if $d(i,j) \le 2$.

The other condition is that there exist positive constants c and C such that, for every set of indicators $I_1,\dots,I_n$ on $\boldsymbol{R_1},\dots,\boldsymbol{R_n}$,

$$\big|\mathrm{Cov}(I_i, I_j)\big| \le C\exp\big(-cd(i,j)\big). \qquad \textbf{(C5)}$$

Finally, there must be a *no free bits* condition, that the distribution of each $\boldsymbol{R_i}$ is altered by the embedding process. There should be a universal constant $\epsilon > 0$ such that

$$D_{\mathrm{TV}}(\boldsymbol{R_i} \sim \mathcal{P}_n, \boldsymbol{R_i} \sim \mathcal{Q}_n^m) \ge \epsilon \tfrac{m}{n} \text{ for all } i. \qquad \textbf{(NFB)}$$

THEOREM 4.2. *Assume* (**C1**)*,* (**C2**)*,* (**C3**)*,* (**C4**)*,* (**C5**)*,* (**E1**)*,* (**E2**)*, and* (**NFB**)*. If m is above the critical rate, $m(n)/\sqrt{n} \to \infty$, an asymptotically perfect detector exists.*

PROOF. By (**NFB**), for each i there is an indicator I_i with

$$\mathrm{P}_{I_i \sim \mathcal{Q}_m^n}[I_i = 1] - \mathrm{P}_{I_i \sim \mathcal{P}_n}[I_i = 1] \ge \epsilon \tfrac{m}{n}. \qquad (11)$$

An asymptotically perfect detector will be constructed from $\sum_i^n I_i$, similarly to proofs of upper bounds in independent and Markov chain square root laws, counting occurrences of local events more likely in stego than cover. The mean of this sum differs by at least $O(m)$ between case of cover and stego, thanks to (**NFB**). In the absence of independence, the exponential decay of cover correlation (**C5**) prevents its variance from growing more than $O(n)$ in covers, and weak dependence of embedding locations will prove the same for stego objects. Standard arguments will then construct a detector, asymptotically perfect above the critical rate.

To avoid many subscripts, let us write I_i for the indicator in the cover distribution, and J_i for the same indicator in the stego distribution $\mathcal{Q}_n^m$. Henceforth the distributions can remain implicit. Summing (11) over all locations,

$$\mathrm{E}\big[\textstyle\sum_i J_i\big] - \mathrm{E}\big[\textstyle\sum_i I_i\big] \geq \epsilon m. \tag{12}$$

To bound $\mathrm{Var}\big[\sum_i I_i\big]$,

$$\begin{aligned}\sum_j \mathrm{Cov}\big[I_i, I_j\big] &\overset{(i)}{\leq} \sum_j C\exp\big(-cd(i,j)\big)\\ &\overset{(ii)}{\leq} \sum_{d=0}^{\infty} C\exp(-cd)\,\big|\{j \mid d(i,j) \leq d\}\big|\\ &\overset{(iii)}{\leq} \sum_{d=0}^{\infty} C\exp(-cd)p(d)\\ &\overset{(iv)}{\leq} C_1,\end{aligned}$$

a constant independent of i. (i) is from (**C5**). (ii) enumerates the same terms (multiple times). (iii) is from (**C4**). (iv) is because the sum is convergent: the ratio of terms tends to $\exp(-c) < 1$, so d'Alembert's test applies. It follows that

$$\mathrm{Var}\big[\textstyle\sum_i I_i\big] = \sum_i \sum_j \mathrm{Cov}\big[I_i, I_j\big] \leq C_1 n. \tag{13}$$

We also need to bound the same variance in the stego case. Recall that the embedding process can introduce long-range (not exponentially diminishing) dependency, since exactly m locations are used. We now find a sufficient condition for such dependency to be negligible.

First, we can dispose of regions that overlap. Recall that this happens only when $d(i,j) \leq 2$. Then

$$\begin{aligned}\sum_i \sum_j \big|\mathrm{Cov}[J_i, J_j]\big| &= \sum_{d(i,j)\leq 2}\sum \big|\mathrm{Cov}[J_i, J_j]\big| + \sum_{d(i,j)>2}\sum \big|\mathrm{Cov}[J_i, J_j]\big|\\ &\leq np(2) + \sum_{d(i,j)>2}\sum \big|\mathrm{Cov}[J_i, J_j]\big|.\end{aligned} \tag{14}$$

The first term is $O(n)$, so it remains to bound the cases where $R_i \cap R_j = \emptyset$.

Let E_i indicate the event that at least one pixel is changed in region R_i. Let $I_{i_1}, \ldots I_{i_K}$ be indicators for the possible cover regions that can change to the region indicated by I_i. Let $C_{i_1}, \ldots, C_{i_K}$ be indicators for the event that each of these changes happens, given that at least one change is made. Exactly one of $C_{i_1}, \ldots, C_{i_K}$ will be one, the rest zero, and they can take the same distribution regardless of whether E_i is zero or one. With these indicators,

$$J_i = I_i(1 - E_i) + E_i \sum_{k=1}^{K} C_{i_k} I_{i_k} = I_i + E_i H_i$$

where $H_i = \sum_{k=1}^{K} C_{i_k} I_{i_k} - I_i$. Since $E_i \perp\!\!\!\perp I_i, \{I_{i_k}\}, \{C_{i_k}\}$, $E_i \perp\!\!\!\perp H_i$. Therefore

$$\begin{aligned}&\big|\mathrm{Cov}[J_i, J_j]\big|\\ &\quad= \big|\mathrm{Cov}[I_i + E_i H_i, I_j + E_j H_j]\big|\\ &\quad\overset{(i)}{=} \big|\mathrm{Cov}[I_i, I_j] + \mathrm{E}[E_i]\mathrm{Cov}[H_i, I_j] + \mathrm{E}[E_j]\mathrm{Cov}[I_i, H_j]\\ &\qquad + \mathrm{E}[E_i E_j]\mathrm{Cov}[H_i, H_j] + \mathrm{E}[H_i]\mathrm{E}[H_j]\mathrm{Cov}[E_i, E_j]\big|\\ &\quad\overset{(ii)}{\leq} \big|\mathrm{Cov}[I_i, I_j]\big| + \big|\mathrm{Cov}[H_i, I_j]\big| + \big|\mathrm{Cov}[I_i, H_j]\big|\\ &\qquad + \big|\mathrm{Cov}[H_i, H_j]\big| + \big|\mathrm{Cov}[E_i, E_j]\big|.\end{aligned} \tag{15}$$

(i) is a property of covariance[6] and (ii) is by the triangle inequality and because all the random variables have absolute value at most 1.

The first three terms of (15) are easily bounded by 1, 2, and 2 times $C\exp(-cd(i,j))$, respectively. For example

$$\begin{aligned}\big|\mathrm{Cov}[H_i, I_j]\big| &= \Big|\sum_k \mathrm{E}[C_{i_k}]\mathrm{Cov}[I_{i_k}, I_j] - \mathrm{Cov}[I_i, I_j]\Big|\\ &\leq 2C\exp(-cd(i,j))\end{aligned} \tag{16}$$

by $\{C_{i_k}\} \perp\!\!\!\perp \{I_{i_k}\}, I_j$ (which comes from (**E2**)), (**C5**) and the triangle inequality. The fourth term is similar; we omit the boring calculation but note that we need to take a step

$$\mathrm{Cov}[C_{i_k} I_{i_k}, C_{j_{k'}} I_{j_{k'}}] = \mathrm{E}[C_{i_k}]\mathrm{E}[C_{j_{k'}}]\mathrm{Cov}[I_{i_k}, I_{j_{k'}}]$$

which requires $\{C_{i_k}\} \perp\!\!\!\perp \{C_{j_{k'}}\}$; this is only necessarily true because $R_i \cap R_j = \emptyset$ (otherwise the decision on what to change in region R_i will constrain the decision in R_j).

For the fifth term of (15) it is more convenient to consider F_i, the complement of E_i, i.e. the probability that no changes occur in R_i. Note that $\mathrm{Cov}[E_i, E_j] = \mathrm{Cov}[F_i, F_j]$.

$$\begin{aligned}\big|\mathrm{Cov}[F_i, F_j]\big| &= \big|\mathrm{P}[F_i](\mathrm{P}[F_j \mid F_i] - \mathrm{P}[F_j])\big|\\ &\overset{(i)}{\leq} \mathrm{P}[F_j] - \mathrm{P}[F_j \mid F_i]\\ &\overset{(ii)}{=} \big(1 - \tfrac{m}{n}\big)\big(1 - \tfrac{m}{n-1}\big) \cdots \big(1 - \tfrac{m}{n-|R_j|+1}\big)\\ &\qquad - \big(1 - \tfrac{m}{n-|R_i|}\big) \cdots \big(1 - \tfrac{m}{n-|R_i|-|R_j|+1}\big)\\ &\overset{(iii)}{\leq} C_3/n,\end{aligned} \tag{17}$$

for a constant C_3. (i) is because $P[F_i] \leq 1$, and $P[F_j] > P[F_j \mid F_i]$ since we assumed $R_i \cap R_j = \emptyset$, so avoiding the locations in R_i can only make it less likely to also avoid those in R_j. (ii) imagines that the m embedding locations are fixed while regions R_i and R_j are chosen uniformly at random, and counts forbidden choices. (iii) is routine, tedious, and mostly omitted: some calculation shows that $\big(1 - \frac{m}{n-k}\big) < \big(1 - \frac{m}{n-D-k}\big) + \alpha/n$, for some constant α independent of m, and this implies the required result.

Finally,

$$\mathrm{Var}\big[\textstyle\sum J_i\big] \overset{(i)}{\leq} C_2 n + \displaystyle\sum_{d(i,j)\geq 2}\sum \big|\mathrm{Cov}[E_i, E_j]\big| \overset{(ii)}{\leq} (C_2 + C_3)n. \tag{18}$$

[6] If $X \perp\!\!\!\perp X'$ and $Y \perp\!\!\!\perp Y'$ then $\mathrm{Cov}[XX', YY'] = \mathrm{E}[XY]\mathrm{Cov}[X', Y'] + \mathrm{E}[X']\mathrm{E}[Y']\mathrm{Cov}[X, Y]$. If also $X' \perp\!\!\!\perp Y'$ then the first term vanishes.

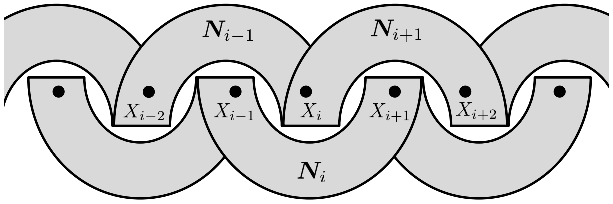

Figure 3: A Markov chain in the MRF cover model.

(i) combines (14), (15), inequalities of the form (16), and the same argument as (13). This step only depends on properties of the cover. (ii) is from (17), which bounds the effect of long-range dependencies in the embedding process.

Now we construct an asymptotically perfect detector using standard arguments. Define

$$I = \textstyle\sum_i I_i, \quad \mu = \mathrm{E}_{\mathcal{P}_n}[I]$$

and give a positive detection if $I > \mu + \frac{1}{2}m\epsilon$. The sum of false positive and negative errors is

$$\begin{aligned} &\mathrm{P}_{\mathcal{P}_n}[I > \mu + \tfrac{1}{2}m\epsilon] + \mathrm{P}_{\mathcal{Q}_n^m}[I \le \mu + \tfrac{1}{2}m\epsilon] \\ &\quad \overset{(i)}{\le} \frac{\mathrm{Var}_{\mathcal{P}_n}[I]}{(\frac{1}{2}m\epsilon)^2} + \frac{\mathrm{Var}_{\mathcal{Q}_n^m}[I]}{(\frac{1}{2}m\epsilon)^2} \\ &\quad \overset{(ii)}{\le} \frac{4(C_1+C_2+C_3)n}{m^2\epsilon^2} \to 0 \end{aligned}$$

above the critical rate. (i) is Chebyschev's inequality and (12), and (ii) uses (13) and (18). □

Note that $\sum_i \sum_j \bigl|\mathrm{Cov}[E_i, E_j]\bigr| = O(n)$ functions as the condition that long-range dependencies introduced by the embedding process must overall be weak.

5 EXAMPLES

We briefly discuss some cover models that meet the conditions for the generalized square root law. We assume the same embedding model, so that (**E1**), (**E2**), and (**E3**) hold.

5.1 Markov Chains

Let n pixels $(X_1, \ldots, X_n)$, $X_i \in \{1, \ldots, K\}$, be the realization of a Markov chain

$$\mathrm{P}[X_1 = i] = p_i, \quad \mathrm{P}[X_{k+1} = j \,|\, X_k = i] = p_{ij}.$$

The first state has distribution p_i and the transition matrix is $P = (p_{ij})$. This fits in the MRF cover model with $N_i = \{i-1, i+1\}$, as depicted in Fig. 3, with boundary cases $N_1 = \{2\}$ and $N_n = \{n-1\}$. Clearly $D = 2$, (**C1**) and (**C2**) hold. (**C3**) follows from the Markov property.

As long as

$$p_i > 0 \text{ and } p_{ij} > 0 \text{ for all } i \text{ and } j, \tag{19}$$

then the lower bound holds. (**NAD**) requires

$$\mathrm{P}[X_i = k \,|\, X_{i-1} = j, X_{i+1} = l] > \epsilon,$$

but this conditional probability is equal to

$$\frac{\mathrm{P}[X_i = k \,|\, X_{i-1} = j]\mathrm{P}[X_{i+1} = l \,|\, X_i = k]}{\mathrm{P}[X_{i+1} = l \,|\, X_{i-1} = j]}$$

which is bounded below by $(\min p_{ij})^2$.

For the upper bound, $d(i,j) = |i-j|$ and $p(x) = 2x+1$ works for (**C4**). (**C5**) follows from the exponential forgetting (also known as exponential or geometric ergodicity) property of Markov chains [28, Thm. 4.9]: there are positive constants C and c such that, for all values x and y,

$$D_{\mathrm{TV}}(X_{i+k} \,|\, X_i{=}x,\ X_{i+k} \,|\, X_i{=}y) \le C\exp(-ck). \tag{20}$$

In the case of (19) this can be proved by a simple coupling argument [28, Ex. 5.1 & Thm. 5.2]. It is not difficult to transform (20) into (**C5**), but for lack of space we will omit the elementary proof.

Our no free bits condition (**NFB**) is equivalent to that in [11]: the second-order co-occurrence probabilities are not preserved by the embedding process. This might be difficult to establish in practice, but thanks to [8] it has a simple equivalent condition: there exist two pairs (x,y), (x',y'), such that one can possibly change to the other by embedding $(\beta(x,x')\beta(y,y') > 0)$ and which are not given the same adjacent probability in covers:

$$\mathrm{P}[X_i = x, X_{i+1} = y] \ne \mathrm{P}[X_i = x', X_{i+1} = y']. \tag{21}$$

Thus we have verified all the conditions of Thms. 4.1 and 4.2.

We remark that this upper bound comes from essentially the same place as the proof in [11] – the exponential forgetting property of Markov chains – though our analysis of the stego variance is complicated by the long-range weak dependencies we allow in the embedding. Unlike [11], our lower bound does not use exponential forgetting; indeed it seems to be completely different from the technical uniform continuity proof used there, as well as more elementary.

Because we do not require even the existence of a stationary distribution, the square root law can now be generalized to *nonstationary* (inhomogeneous) Markov chains

$$\mathrm{P}[X_1 = i] = p_i, \quad \mathrm{P}[X_{k+1} = j \,|\, X_k = i] = p_{ij}^k,$$

with (19) strengthened to

$$\forall k, i, j.\ p_{ij}^k \ge \delta,$$

and (21) to

$$\begin{aligned} &\forall k.\ \exists (x,y),(x',y').\ \beta(x,x')\beta(y,y') > 0 \wedge \\ &\qquad \mathrm{P}[X_k = x, X_{k+1} = y] - \mathrm{P}[X_k = x', X_{k+1} = y'] \ge \delta, \end{aligned}$$

where δ is a positive constant not depending on n. The lower bound argument is exactly as above, and the upper bound follows because such a nonstationary chain is still exponentially forgetting, thanks to a small tweak to the coupling argument that proves (20).

5.2 Ising Models

Let $n = N^2$ pixels, $(X_{ij} \,|\, i,j \in \{1, \ldots, N\})$, be the realization of a 2-dimensional Ising model [14, Ex. 2.1], depicted in Fig. 1. In such a random field the pixels are binary, and we follow convention and assign them values $X_{ij} = \pm 1$[7]. The probability distribution is given by

$$\mathrm{P}[\boldsymbol{X} = \boldsymbol{x}] \propto \exp\Big(\beta H \sum_{i,j} x_{ij} + \beta J \sum_{(i',j')} \sum_{\in \mathcal{N}_{ij}} x_{ij} x_{i'j'}\Big) \quad (22)$$

where conventionally $\beta > 0$ is called the *inverse temperature*, $H \in \mathbb{R}$ the strength of an *external magnetic field*, and $J \in \mathbb{R}$ the *interaction strength* (we will impose some conditions on them in a moment). Higher positive values of βH bias the pixels more towards $+1$, and negative values towards -1; higher positive values of βJ bias neighbours to be equal more often, and negative values bias them to be unequal more often. The sum is over immediate neighbours in the grid, $N_{ij} = \{(i-1, j), (i, j-1), (i+1, j), (i, j+1)\}$. We will take the case of *toroidal* boundary conditions, where row (resp. column) 1 is considered adjacent to row (column) N.

This fits the MRF model: $D = 4$, and (**C1**,**C2**,**C3**) hold. The no asymptotic determinism condition (**NAD**) can be established by direct computation:

$$\mathrm{P}[X_{ij} = +1 \,|\, \boldsymbol{N_{ij}}] = \frac{1}{1 + e^{-2\beta H - 2\beta J (N^+ - N^-)}}$$

where N^+ (resp. N^-) is the number of $+1$ (-1) states in $\boldsymbol{N_{ij}}$; this conditional probability is bounded away from zero. We have established the lower bound of the square root law.

For the upper bound, $d\big((i,j),(i',j')\big) = |i-i'|+|j-j'|$ and the regular 2d grid gives $p(x) = 1 + 2x(x+1)$ for (**C4**). For (**C5**) we use some standard results from statistical physics: (22) satisfies *Dobrushin's uniqueness condition* at least for the cases $|H|$ sufficiently large (pixels biased away from uniform), or $H = 0$ and β sufficiently small (inter-pixel dependencies not too great) [4]. Random fields satisfying such a condition have many interesting properties, including [29] 'any finite volume covariance between two local functions f and g decays exponentially fast with the distance between their supports, with a rate that is uniform in ... the choice of f and g', see also [14, Thm. 2.1.3]. This is (**C5**).

Without solving the model, we can show that the no free bits condition (**NFB**) holds for any sublinear payload. The only exception is if $H = 0$ *and* $J = 0$ (i.e. a completely uniform independent field). To see why, consider any region $\boldsymbol{R_{ij}} = (X_{ij}, \boldsymbol{N_{ij}})$. Take $\boldsymbol{x}$ to be the (or, if there is more than one, a) most likely cover configuration on $\boldsymbol{R_{ij}}$, and let $\boldsymbol{x'}$ be strictly less likely and differ from $\boldsymbol{x}$ in one location. Write $K = \mathrm{P}_{\mathcal{P}_n}[\boldsymbol{R_{ij}} = \boldsymbol{x}] - \mathrm{P}_{\mathcal{P}_n}[\boldsymbol{R_{ij}} = \boldsymbol{x'}] > 0$. Then compute

$$\begin{aligned}
&\mathrm{P}_{\mathcal{P}_n}[\boldsymbol{R_{ij}} = \boldsymbol{x}] - \mathrm{P}_{\mathcal{Q}_n^m}[\boldsymbol{R_{ij}} = \boldsymbol{x}] \\
&\quad \overset{(i)}{=} \mathrm{P}_{\mathcal{P}_n}[\boldsymbol{R_{ij}} = \boldsymbol{x}] - \sum_{\boldsymbol{y}} \mathrm{P}_{\mathcal{P}_n}[\boldsymbol{R_{ij}} = \boldsymbol{y}] \alpha(\boldsymbol{y}, \boldsymbol{x}) \\
&\quad \overset{(ii)}{\geq} \alpha(\boldsymbol{x'}, \boldsymbol{x})\, K \\
&\quad \overset{(iii)}{=} \tfrac{m}{n} \tfrac{n-m}{n-1} \tfrac{n-m-1}{n-2} \tfrac{n-m-2}{n-3} \tfrac{n-m-3}{n-4}\, K.
\end{aligned}$$

(i) is by (**E2**), and where $\alpha(\boldsymbol{y}, \boldsymbol{x})$ is the probability that $\boldsymbol{y}$ is changed to $\boldsymbol{x}$ by the embedding process. (ii) is because $\sum_{\boldsymbol{y}} \alpha(\boldsymbol{y}, \boldsymbol{x}) = 1$ and using the fact that $\boldsymbol{x}$ was a most likely cover configuration. (iii) simply computes the probability that $\boldsymbol{x}$ is changed in one location to $\boldsymbol{x'}$. As long as $m/n \leq c < 1$, we have shown (**NFB**).

This result can be extended to other Ising models:

- Give fixed boundary conditions instead of toroidal periodicity. Only (**NFB**) needs more work, because the neighbourhoods are no longer identically distributed, and we must ensure that $\mathrm{P}_{\mathcal{P}_n}[\boldsymbol{R_{ij}} = \boldsymbol{x}] - \mathrm{P}_{\mathcal{Q}_n^m}[\boldsymbol{R_{ij}} = \boldsymbol{x}]$ cannot approach zero. The result follows because Dobrushin's condition ensures that influence of the boundary decays exponentially fast (there is some detail here that lack of space precludes).
- The conditions on H, J, and β can be relaxed somewhat [33]. Note that Dobrushin's condition is sufficient, but not necessary, for (**C5**).
- We can consider dimensions higher than 2, or models where interactions occur at bounded distance rather than only between immediate neighbours. The only part of the argument that needs to change is verification of Dobrushin's condition, which holds at least for $H = 0$, $J > 0$, and β sufficiently small (this follows from [14, Ex. 2.1.3]).

6 TOWARDS AN ADVERSARIAL SQUARE ROOT LAW

Having extended the square root law to a wider class of cover models, we now consider the embedding model. Until now, square root laws have applied to 'dumb' embedding models, which apply some fixed random function to either m uniformly chosen locations (as here), or to each location independently with probability m/n. This is only a good model for steganography without source coding, which (particularly syndrome coding [13]) is now prevalent in well-constructed steganography, for both reducing the number of embedding changes and choosing less-detectable change options.

Rate distortion arguments [13] show that a payload of at most $O(c \log \frac{n}{c})$ can be embedded while making c changes in an n-element cover; the critical rate should therefore rise to $\sqrt{n} \log n$. But there are two complications before a square root law can be proved.

First, codes vary: the bound is only achieved in trivial circumstances, and some codes do not even approach it as $n \to \infty$; of those that do, some are computationally infeasible. A result that only applies to a particular code could become redundant if new codes are discovered. Second, using a code means that certain combinations of changes will not happen, thus introducing long-range dependencies into the embedding process: can we be sure that this is not exploited by the detector?

We propose to abstract away details of the code, and concentrate only on the probabilities of change. Let us say that $\boldsymbol{\pi}(n)$ is an *embedding process* if, for any fixed cover of size n, it describes a probability distribution on the stego object (see also [9]): we ignore the coding itself. We write $H(\boldsymbol{\pi}(n))$ for the conditional entropy of the stego object given the cover, which is an upper bound for the payload size. Since useful codes exist that convey payload within a multiple of

[7] Non-binary Gibbs fields would be a valuable generalization.

this entropy [9], proving a square root law for embedding processes is sufficient to prove one for practical codes as well.

Codes that introduce long-range dependencies can be described by such a model. In practice we would not expect a good embedding process to introduce many strong dependencies, because this only reduces its capacity (entropy). Considering only capacity, embedding processes that induce *independent* changes are optimal: they maximize entropy. But the same is not necessarily true of security against a detector: it is reasonable to conjecture that optimal embedding noise might have similar covariances to the cover.

Note that a square root (here $\sqrt{n}\log n$) law cannot hold for *all* embedding processes, because there are some that are not asymptotically efficient (simple overwriting is an obvious example). It suffices for there to be *some* process that guarantees $o(\sqrt{n}\log n)$ asymptotically undetectable payload bits. On the other hand, we will require the upper bound to be inescapable by *any* embedding processes. Hence,

Conjecture 6.1. *Under similar assumptions to* (**C1**)*,* (**C2**)*,* (**C3**)*,* (**C4**)*,* (**C5**)*, no free bits* (**NFB**) *and no determinism in covers* (**NAD**) *or stego objects* (**E3**)*,*

(i) if $m(n)/\sqrt{n}\log n \to \infty$, then for every *embedding process $\boldsymbol{\pi}(n)$ with $H(\boldsymbol{\pi}(n)) \geq m(n)$, an asymptotically perfect detector exists;*

(ii) if $m(n)/\sqrt{n}\log n \to 0$, then there exists embedding processes $\boldsymbol{\pi}(n)$ with $H(\boldsymbol{\pi}(n)) = m(n)$ and such that every detector is asymptotically random.

Consider the detector's choice of detection statistic to be their *strategy*, and the embedder's choice of embedding process theirs. Add some payoff related to detectability and this is a game-theoretic formulation of steganography. The above result, which we call an *adversarial square root law*, proves that above the critical rate the detector has a winning strategy (regardless of the embedder's choice) and conversely below it. Unlike other game-theoretic analyses applicable to steganography [1, 15, 25], we are not locating an equilibrium, rather we prove something about its asymptotic behaviour[8].

What does this tell us of adaptive embedding, since it does not appear in the statement of the theorem? We will have shown that it cannot escape the $\sqrt{n}\log n$ capacity law: whether coding is used purely to improve embedding efficiency or whether it takes account of different embedding costs, as long as no costs are zero (which would violate *no free bits*) or infinite (violating *no determinism*) then the critical rate is the same. The costs can affect capacity up to constant multiples, but they do not affect the order of growth.

Part (ii) of Conjecture 6.1 should be the easier half, because we get to choose the embedding process. We cannot overwrite a fixed m locations, because this does not have enough entropy, but we can make use of the same trick as in [21]: break the cover into $o(\sqrt{n})$ blocks of size $\omega(\sqrt{n})$ and make exactly one change per block. This will have enough entropy for $o(\sqrt{n}\log n)$ payload bits, and we expect to be able to adapt the 'shortlist' idea from Subsect. 4.1 to the structure in this embedding process. Such an embedding code is well below optimal, of course, but for the asymptotic result it is enough to be within a constant factor of optimal.

Part (i) will be the more difficult, as the embedding process might have long-range dependencies. Either they will have to be bounded by assumption, or it will be necessary to show that too many dependencies that are too strong will reduce the entropy of the embedding process too much.

An advantage of source coding is that so-called 'wet' locations (in the language of [12]), where a change is considered perfectly detectable, can be avoided. Paralleling this, we expect to weaken the no determinism assumption (**NAD**), so that not all locations need satisfy it. We would then replace the cover size n by the number of 'dry' locations that do satisfy (**NAD**).

7 CONCLUSIONS

When embedding is perfect – a process that does not change the probability distribution of covers – or if the embedding learns a cover model so that the embedding tends to perfect, there is typically a *linear law of steganographic capacity*. When embedding can introduce groups of pixels that are impossible in cover objects, there is typically a *constant capacity* that does not grow with the cover size. Once we exclude the first case (with a *no free bits* condition) and the second (with a *no determinism* condition), it takes few additional assumptions to prove a *square root law*.

We extended the square root law to inhomogeneous Markov chains and a variety of Ising models, but more generally for covers subject to two main conditions: direct dependence of finite range, and exponential decay of covariance. We may reasonably expect these properties to be true for the acquisition chain of most digital media (for example: CCD leakage, demosaicking, and resampling all cause only local dependencies), but might not hold when there are macroscopic dependencies caused by scene content (for example: consistency of light sources). However, it is difficult to imagine a detector able to exploit such dependencies.

We do not claim that the sufficient conditions, from which the results of this paper have been proved, are always necessary. Consider for example a cover model without exponential decay of covariance: do strong cover interactions make detection more difficult, or easier? One might expect the latter[9], in which case it should be possible to drop the assumption. But consider binary pixels drawn from a Pólya Process, a MRF with unbounded dependence: it can be shown that the lower bound holds anyway; covariance does not decay (at all, let alone exponentially) and the upper bound does *not* hold. Space precludes further discussion of this fascinating example. Space also precludes discussion of infinite-range Ising models, which in some cases can be analyzed with extensions of the methods used here. Another generalization would be

[8] In [1] the payoff itself is concerned with asymptotic behaviour: the exponential rate at which detection tends to perfect in the case of a constant-rate payload.

[9] Adaptive steganography typically employs the following heuristic: cover locations that can be well-predicted from others are bad choices for embedding.

to weaken the no free bits assumption to a conditional rather than joint distributional difference.

We have only considered cases where the cover and stego distributions $\mathcal{P}_n$ and $\mathcal{Q}_n^m$ are known to the detector. Of course, in absence of this knowledge the lower bound still applies. When can the critical rate be raised? How much knowledge is required to keep the upper bound? These are questions for further research. Note that the model where a detector that learns about the cover source [22] is problematic when the cover is nonstationary, unless what is learned in the past gives information about the future.

In this paper we have omitted a third clause included in some square root laws: if $m(n)/r(n) \to c$, embedding *on* the critical rate, we can sometimes calculate or bound the KLD between cover and stego objects, proving ϵ-security for some value $\epsilon(c)$ (e.g. [11, 21]); equivalently, compute the Fisher Information of the embedding. In view of the discussion of KLD and TV in Sect. 3 the focus should probably be on TV instead, but the proof methods of this paper – inequalities rather than exact asymptotics – could only give loose bounds.

Finally, we have made a conjecture about a square root law for adaptive embedding, which is of an adversarial nature. Given that adaptive embedding is now dominant in image and video steganography, a proof should bring the theory even closer to practice.

REFERENCES

[1] M. Barni and B. Tondi. 2013. The Source Identification Game: An Information-Theoretic Perspective. *IEEE Transactions on Information Forensics and Security* 8, 3 (2013), 450–463.

[2] B. A. Bash, D. Goeckel, and D. Towsley. 2012. Square Root Law for Communication with Low Probability of Detection on AWGN Channels. In *Proc. International Symposium on Information Theory*. IEEE, Piscataway, NJ, 448–452.

[3] C. Cachin. 2004. An Information-Theoretic Model for Steganography. *Information and Computation* 192, 1 (2004), 41–56.

[4] J.-R. Chazottes, P. Collet, and F. Redig. 2016. On Concentration Inequalities and their Applications for Gibbs Measures in Lattice Systems. (2016). arXiv:1610.06502 (submitted for publication).

[5] P. Comesaña and F. Pérez-González. 2007. On the Capacity of Stegosystems. In *Proc. 9th Workshop on Multimedia & Security (MM&Sec)*. ACM, New York, NY, 15–24.

[6] T. Denemark, V. Sedighi, V.and Holub, R. Cogranne, and J. Fridrich. 2014. Selection-Channel-Aware Rich Model for Steganalysis of Digital Images. In *Proc. International Workshop on Information Forensics and Security (WIFS)*. IEEE, Piscataway, NJ, 48–53.

[7] T. Filler. 2008. *Important Properties of Normalized KL-Divergence under HMC Model.* Technical Report. DDE Lab, SUNY Binghamton. http://dde.binghamton.edu/filler/kl-divergence-hmc.pdf Techncial Report.

[8] T. Filler and J. Fridrich. 2009. Complete Characterization of Perfectly Secure Stego-systems with Mutually Independent Embedding Operation. In *Proc. International Conference on Acoustics, Speech, and Signal Processing*. IEEE, Piscataway, NJ, 1429–1432.

[9] T. Filler and J. Fridrich. 2010. Gibbs Construction in Steganography. *IEEE Transactions on Information Forensics and Security* 5, 4 (2010), 705–720.

[10] T. Filler, J. Judas, and J. Fridrich. 2011. Minimizing Additive Distortion in Steganography using Syndrome-Trellis Codes. *IEEE Transactions on Information Forensics and Security* 6, 3 (2011), 920–935.

[11] T. Filler, A. D. Ker, and J. 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.

[12] J. Fridrich, M. Goljan, D. Soukal, and P. Lisoněk. 2005. Writing on Wet Paper. *IEEE Transactions on Signal Processing* 53, 10 (2005), 3923–3935.

[13] J. Fridrich and D. Soukal. 2006. Matrix Embedding for Large Payloads. *IEEE Transactions on Information Forensics and Security* 1, 3 (2006), 390–394.

[14] X. Guyon. 1995. *Random Fields on a Network: Modeling, Statistics, and Applications.* Springer-Verlag, New York. Translated by C. Ludena.

[15] B. Johnson, P. Schöttle, A. Laszka, J. Grossklags, and R. Böhme. 2015. *Adaptive Steganography and Steganalysis with Fixed-Size Embedding.* Springer, Berlin, Heidelberg, 69–91.

[16] A. D. Ker. 2004. Improved Detection of LSB Steganography in Grayscale Images. In *Proc. 6th Information Hiding Workshop (LNCS)*, Vol. 3200. Springer, Berlin, Heidelberg, 97–115.

[17] A. D. Ker. 2006. Batch Steganography and Pooled Steganalysis. In *Proc. 8th Information Hiding Workshop (LNCS)*, Vol. 4437. Springer, Berlin, Heidelberg, 265–281.

[18] A. D. Ker. 2007. A Capacity Result for Batch Steganography. *IEEE Signal Processing Letters* 14, 8 (2007), 525–528.

[19] A. D. Ker. 2009. Locally Square Distortion and Batch Steganographic Capacity. *International Journal of Digital Crime and Forensics* 1, 1 (2009), 29–44.

[20] A. D. Ker. 2009. The Square Root Law Requires a Linear Key. In *Proc. 11th Workshop on Multimedia and Security.* ACM, New York, NY, 85–92.

[21] A. 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.

[22] A. D. Ker. 2010. The Square Root Law in Stegosystems with Imperfect Information. In *Proc. Information Hiding, 12th International Conference (LNCS)*, Vol. 6387. Springer, Berlin, Heidelberg, 145–160.

[23] A. 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.

[24] A. D. Ker, P. Bas, R. Böhme, R. Cogranne, S. Craver, T. Filler, J. Fridrich, and T. Pevný. 2013. Moving Steganography and Steganalysis from the Laboratory into the Real World. In *Proc. 1st Workshop on Information Hiding and Multimedia Security.* ACM, New York, NY, 45–58.

[25] A. D. Ker, T. Pevný, and P. Bas. 2016. Rethinking Optimal Embedding. In *Proc. 4th Workshop on Information Hiding and Multimedia Security.* ACM, New York, NY, 93–102.

[26] A. D. Ker, T. Pevný, J. Kodovský, and J. Fridrich. 2008. The Square Root Law of Steganographic Capacity. In *Proc. 10th Workshop on Multimedia and Security.* ACM, New York, NY, 107–116.

[27] E. L. Lehmann and J. P. Romano. 2005. *Testing Statistical Hypotheses* (3rd ed.). Springer-Verlag, New York.

[28] D. A. Levin, Y. Peres, and E. L. Wilmer. 2009. *Markov Chains and Mixing Times.* American Mathematical Society, Providence, RI.

[29] F. Martinelli. 2000. An Elementary Approach to Finite Size Conditions for the Exponential Decay of Covariances in Lattice Spin Models. In *In: On Dobrushins Way. From Probability Theory to Statistical Physics.* Translations Series 2, Vol. 198. American Mathematical Society, Providence, RI, 169–181.

[30] T. Pevný, J. Fridrich, and A. D. Ker. 2012. From Blind to Quantitative Steganalysis. *IEEE Transactions on Information Forensics and Security* 7, 2 (2012), 445–454.

[31] B. Ryabko and D. Ryabko. 2011. Constructing Perfect Steganographic Systems. *Information and Computation* 209, 9 (2011), 1223–1230.

[32] D. Ryabko. 2011. On the Relation Between Realizable and Non-Realizable Cases of the Sequence Prediction Problem. *Journal of Machine Learning Research* 12 (2011), 2161–2180.

[33] R. H. Schonmann and S. B. Shlosman. 1995. Complete Analyticity for 2D Ising Completed. *Communications in Mathematical Physics* 170, 2 (1995), 453–482.

[34] Y. Wang and P. Moulin. 2008. Perfectly Secure Steganography: Capacity, Error Exponents, and Code Constructions. *IEEE Transactions on Information Theory* 55, 6 (2008), 2706–2722.

[35] A. Wilson and A. D. Ker. 2016. Avoiding Detection on Twitter: Embedding Strategies for Linguistic Steganography. In *Media Watermarking, Security, and Forensics 2016.* IS&T, Article 9, 9 pages.