Perturbation Hiding and the Batch Steganography Problem

Andrew D. Ker

Oxford University Computing Laboratory, Parks Road, Oxford OX1 3QD, England
adk@comlab.ox.ac.uk

Abstract. The batch steganography problem is how best to split a steganographic payload between multiple covers. This paper makes some progress towards an information-theoretic analysis of batch steganography by describing a novel mathematical abstraction we call *perturbation hiding*. As well as providing a new challenge for information hiding research, it brings into focus the information asymmetry in steganalysis of multiple objects: Kerckhoffs' Principle must be interpreted carefully.

Our main result is the solution of the perturbation hiding problem for a certain class of distributions, and the implication for batch steganographic embedding. However, numerical computations show that the result does not hold for all distributions, and we provide some additional asymptotic results to help explore the problem more widely.

1 Introduction

The batch steganography problem was first posed in [1]. It supposes that a steganographer possesses a set of cover objects which, between them, are to conceal a covert payload. The aim is to split the payload into a number of parts and embed the parts, using standard steganographic methods, into some or all of the individual objects. The key question is whether the payload should be spread thinly amongst all the covers, whether a small number of covers should be filled to maximum capacity, or some intermediate choice.

This question is relevant to any scenario in which multiple covers are available, including covert communication and steganographic file systems. In fact, it is hard to imagine many scenarios in which only one cover is made available to the steganographer: if they have a plausible reason to send one cover communication, they almost certainly have a plausible reason to send more than one. Then it becomes important to know how to split the payload between the covers, to evade detection.

The initial analysis of the batch steganography problem in [1] includes only a few special cases, making very strong assumptions about the detector's behaviour, and subsequently there have been some results attacking other limited cases: [2] for a detector which counts observations exceeding a threshold, and [3] under the assumption that steganographic distortion is square in the number of embedding changes. An asymptotic capacity result is found in [4] but does not determine the best method of spreading the payload amongst multiple covers.

K. Solanki, K. Sullivan, and U. Madhow (Eds.): IH 2008, LNCS 5284, pp. 45–59, 2008.

Here we attack the general problem, not constraining the detector, and to do this we propose a mathematical abstraction which we call *perturbation hiding*. It is approached using tools of information theory, but the level of abstraction is different from the usual information-theoretic analysis of steganography [5, 6].

We will now summarise the batch steganography problem and point out some ambiguities in its statement; there follows a brief discussion of the appropriate interpretation of Kerckhoffs' Principle in steganography. In Sect. 2 we pose the perturbation hiding problem, discuss its connection with batch steganography, and solve the problem for a class of cover families: it is best to spread payload equally between the cover objects, even though this denies the opportunity to keep the enemy guessing as to payload distribution. However, this is not a general solution, as some numerical explorations show. Motivated by the numerical results, Sect. 3 suggests some asymptotic results (proved with rather less rigor than the solution of Sect. 2) which point towards more general conclusions. Finally, we will discuss the next steps in Sect. 4.

1.1 The Batch Steganography Problem

We take the role of a steganographer who, for reasons legitimate or not, wishes to conceal a payload in a number of cover objects. If the payload is spread thinly amongst all covers then there is little in each object; on the other hand, if a smaller number of objects are filled to capacity then the complete set of objects (all of which are transmitted) contains many genuinely innocent covers, which could confound the detector. The batch steganography problem is how best to balance those factors, but it is difficult to formalize. The security of an embedding process should be measured by the (un)reliability of detectors, but of course this depends on the choice of detector. The results in [1, 2] fix on a few particular cases of detector, and measure security by the number of false positive detections when the false negative rate is 50%. Such results are of limited applicability.

A detector-independent measure of security was suggested by Cachin in [5] and is now widely used in literature on the theory of steganographic security. Cachin postulated a distribution of covers, a corresponding distribution of stego objects, and considered the *Kullback-Leibler (KL) divergence* [7] between those distributions. KL divergence is nonnegative and zero only for equal distributions. Most importantly, there is a well-known connection with hypothesis testing: error rates for determining whether an observation is from distribution X or distribution Y are bounded below by a function of $\mathrm{D}_{\mathrm{KL}}(X \,\|\, Y)$[1].

Incidentally, the detector's task is not as simple as it may seem: they cannot simply test all the objects, knowing that they need only prove a single example of steganography, because this would compound their false positive errors. By measuring KL divergence we avoid discussion of the detector itself, bounding the

[1] Any detector must mistake an observation of X for Y with probability α, and vice versa with probability β, satisfying $\alpha \log \frac{\alpha}{1-\beta} + (1-\alpha) \log \frac{1-\alpha}{\beta} \leq \mathrm{D}_{\mathrm{KL}}(X \,\|\, Y)$.

performance of any detector (including detectors which choose to ignore some of the available evidence).

KL divergence is used to measure security of batch steganography in [3], but we now demonstrate a weakness in the formulation. For now, we will adopt the same notation as that paper, writing $X_i^{p_i}$ for the random variable corresponding to the i-th object, in which a payload has been embedded causing p_i embedding changes. If we may assume that the objects are independent random variables (for example, the covers should not be successive frames from a video, and the payloads should not be identical) then the additivity property means that the total KL divergence satisfies

$$\mathrm{D}_{\mathrm{KL}}\big((X_1^0,\ldots,X_n^0)\,\|\,(X_1^{p_1},\ldots,X_n^{p_n})\big)=\sum_{i=1}^{n}\mathrm{D}_{\mathrm{KL}}(X_i^0\,\|\,X_i^{p_i}) \tag{1}$$

and it is now, in principle, possible to select the number of changes $p_1,\ldots,p_n$ to minimize (1), thus minimizing the detector's reliability.

There are two problems. First, it is cumbersome to account for the relationship between the *size* of embedded payload and the *number of embedding changes* induced. Apart from the added layer of complexity, evident in [3], there are also implicit assumptions that all embedding changes are equally detectable, and that the number of changes depends deterministically on the payload size but not the cover object. Neither is correct: in digital media it is highly likely that some embedding changes are more obvious than others, and the number of changes varies with random correlations between cover and payload.

Second, and more seriously, there is a paradox in this analysis. Suppose that the steganographer has n (independent) covers drawn from the same distribution, and in just one object makes p embedding changes. The KL divergence between the random vector emitted and a vector of n unaltered objects is

$$\mathrm{D}_{\mathrm{KL}}(\boldsymbol{X}^0\,\|\,\boldsymbol{X}^p)=\mathrm{D}_{\mathrm{KL}}(X_1^0\,\|\,X_1^p)+\sum_{i=2}^{n}\mathrm{D}_{\mathrm{KL}}(X_i^0\,\|\,X_i^0)=\mathrm{D}_{\mathrm{KL}}(X_1^0\,\|\,X_1^p)$$

(regardless of which object is altered) and this is independent of n. This cannot be right: surely it is harder to detect one stego object in amongst many covers, than to tell a single stego object from a single cover? The explanation is that KL divergence is only appropriate for bounding the performance of a hypothesis test where both null and alternative are *simple*, involving no unknown parameters. When we use KL divergence as a metric we are assuming that the opponent knows everything except whether there is any payload or not, including the object which would be selected to carry the payload. This is surely unrealistic.

We do not use this example to claim that KL divergence is the wrong measure for security. It was an incorrect formulation of the (implicit) hypotheses which caused the paradox, and to avoid such problems we must take care about the information asymmetry in steganalysis scenarios.

1.2 Kerckhoffs' Principle in Steganography

So let us reconsider the security model for steganography. It is traditional, in analysis of cryptographic security, to assume the worst case: the opponent is granted almost omniscience regarding the cryptosystem, and (in the case of protocols) almost omnipotence as respects sabotage of transmitted messages. Such conservatism is justified by the possibility of traitors in the communications system.

This is known as *Kerckhoffs' Principle*, one of six desiderata for cryptosystems suggested by the Dutch cryptographer Auguste Kerckhoffs in 1883:

> Il faut qu'il n'exige pas le secret, et qu'il puisse sans inconvénient tomber entre les mains de l'ennemi [8]

or (approximately) that it must not be necessary to keep the system secret: it should not cause trouble were it to fall into enemy hands. Additionally to Kerckhoff's Principle, cryptographers consider the *chosen-plaintext attack*, when the opponent is given the ability to generate their own cyphertexts.

How should we interpret Kerckhoffs' Principle, and the chosen-plaintext attack, in the context of steganography? This issue is discussed in [9], whose authors point out that the principle is rarely mentioned in steganography literature. First, we can dispose of the full chosen- (or known-) plaintext model, which is not appropriate for covert communication if we assume that the steganographer uses an encryption scheme, secure against chosen-plaintexts, prior to embedding. This is analogous to the usual assumption of perfect cryptography added to the Dolev-Yao threat model for protocol security [10]. But even if the payload bits embedded in the cover are obscured by encryption, the same is probably not true of the number of such bits, i.e. the payload *size*.

So consider what we should grant the opponent. As well as knowledge of the steganographic embedding process for placing payload in individual objects, there seem to be four possibilities involving payload size in the batch situation:

(a) the steganalyst knows nothing about the payload being transmitted;
(b) the steganalyst knows the total payload size, but nothing of the steganographer's strategy for breaking it into components;
(c) the steganalyst knows the sizes of the individual payloads to be embedded in the covers, but does not know which object receives which payload size;
(d) the steganalyst knows the amount of payload in each object, they only lack knowledge of whether any embedding happens at all.

Option (a) is dangerously weak, clearly contradicting the spirit of Kerckhoffs' Principle. We should consider the possibility that the steganalyst might use a confederate to insert a payload of known size into the covert communication channel, or could compromise a recipient after the fact. Option (d) is probably too strong, for it is hard to see how the steganalyst could know so much information without also knowing for certain that the covert channel is being used[2]. The

[2] Note that estimating the payload in each object is not the same as knowing it.

correspondence between cover object and payload segment should be considered part of the steganographer's secret key shared with their intended recipient. We believe that reliance on option (d) is a significant weakness of [3].

This leaves (b) and (c), both sensible attack models for covert communication. (c) is the more conservative, but not unreasonably so: if the steganalyst were to obtain the steganographer's embedding software, they might learn the strategy for splitting payload between covers. More practically, option (b) seems difficult to analyse because it cannot be cast as hypothesis test without a compound alternative or an (unjustifiable) prior, and so KL divergence is not a good model for detection accuracy. In this paper, therefore, we focus on option (c).

1.3 Notation

In order to reduce complexity of presentation, we will use the following notational conventions throughout the paper. Random variables will always be given upper case letters, observations lower case, and distribution parameters will be Greek lower case. Vectors (of variables, random or otherwise, or parameters) will be boldface, $\boldsymbol{x} = (x_1, \ldots, x_n)$, and $\bar{x}$ will denote $\frac{1}{n}\sum_{i=1}^{n} x_i$. The set of permutations on n elements is S_n; its members will be denoted π and $\pi(\boldsymbol{x})$ means $(x_{\pi(1)}, \ldots, x_{\pi(n)})$. If D is a one-parameter family of distributions, with parameter λ, then $X \sim D(\lambda)$ indicates that the random variable X has this distribution with the given parameter. The expectation is denoted $\mathrm{E}[X]$ and, where the distribution of X needs clarification, it indicated by a subscript: $\mathrm{E}[X]_{X\sim D(\lambda)}$. With random vectors, $\boldsymbol{X} \sim \big(D(\lambda_1), \ldots, D(\lambda_n)\big)$ means that $X_i \sim D(\lambda_i)$ for each i, and also that the X_i are independent. Finally, $\boldsymbol{X} \sim D(\lambda)^n$ is used when the independent components of X are identically distributed.

2 The Perturbation Hiding Problem

We now present the perturbation hiding problem, draw the connection with batch steganography, and solve for a class of distribution families.

Suppose a fixed one-parameter family of probability distributions $D(\lambda)$[3] defined for $\lambda \geq 0$, an integer $n \geq 2$, and a positive constant l. We must choose a nonnegative vector of parameters $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_n)$ subject to the constraint $\bar{\lambda} = l$, with the aim of making the random vectors $\boldsymbol{X}$ and $\boldsymbol{Y}$, defined by

$$\begin{aligned} \boldsymbol{X} &\sim D(0)^n \\ \boldsymbol{Y} &\sim \big(D(\lambda_1), \ldots, D(\lambda_n)\big), \end{aligned}$$

as close to indistinguishable as possible: it should be difficult for an opponent to classify a realization as either $\boldsymbol{X}$ or $\boldsymbol{Y}$ accurately.

This is called *perturbation hiding* because we are required to choose the perturbation from zero in the n parameters defining the random vector. One could

[3] There is nothing in this paper which requires them to be one-dimensional random variables, and the same results will apply to random vectors.

imagine many variations of the problem, when the opponent has more or less knowledge about the choice of $\boldsymbol{\lambda}$, but we will fix on the version best aligned with batch steganography: we grant the opponent knowledge of the components of $\boldsymbol{\lambda}$ *but not their order*. Since the opponent has no information on the order of $\boldsymbol{\lambda}$, their observation is equivalent to one of $\boldsymbol{X}$ or $\boldsymbol{Y}$ with

$$\begin{aligned} &\boldsymbol{X} \sim D(0)^n \\ &\boldsymbol{Y} \sim \Pi\big(D(\lambda_1), \dots, D(\lambda_n)\big), \end{aligned} \quad \begin{aligned} &\text{where } \Pi \text{ is chosen uniformly from } S_n \\ &\text{independently of all other random variables.} \end{aligned}$$

The perturbation hiding problem is:

$$\text{Choose } \boldsymbol{\lambda}, \text{ subject to } \bar{\lambda} = l \text{ and all } \lambda_i \geq 0, \text{ to minimize } \mathrm{D}_{\mathrm{KL}}(\boldsymbol{X} \,\|\, \boldsymbol{Y}). \tag{2}$$

2.1 Connection with Batch Steganography

If we suppose that the steganographer has selected an embedding method, and uses a source of covers which are uniform (possessing the same characteristics as regards their potential for information hiding), then we can define $D(\lambda)$ as the distribution of objects with payload of size λ (bits). If the total payload is of size nl, there are n covers, the opponent knows everything about their strategy except which cover receives which payload size, and they know nothing more about the opponent so that KL divergence is the appropriate metric, then we have a direct correspondence with the perturbation hiding problem.

It is worthwhile to contrast this with other information theoretic analyses of steganography. Papers following Cachin [5] focus on optimizing the embedding method for individual objects to minimize $\mathrm{D}_{\mathrm{KL}}\big(D(0) \,\|\, D(\lambda)\big)$. An analysis of *perfectly secure* steganography, in which the KL divergence is zero, is described thoroughly in [6]. Such work considers a cover object to be a sequence of samples emitted by a source with known characteristics. Here, we are taking a different level of abstraction where the source emits entire cover objects, and furthermore our assumption is that a perfectly secure embedding is *not* used. This is reasonable because there are no known perfect schemes which work in genuine digital media, and mathematical models of such media do not accord closely with reality. In our setting, detection is possible; the question is how to minimize the reliability of detection, by allocating payload amongst multiple covers.

The perturbation hiding problem is attractive for a number of reasons. First, it seems to be an interesting mathematical challenge in its own right. Second, it allows use of KL divergence even in a situation when the opponent does not know which cover receives which payload, expressing the problem as a test between two simple hypotheses. This is at the cost of algebraic complexity. Third, it avoids the complications of [3] by folding the relationship between cover changes and transmitted payload into the parameterization of the distribution family $D(\lambda)$. Parameterization is important to this problem: as a simple example, the families $D(\lambda) \sim \mathrm{N}(\lambda, 1)$ and $D'(\lambda) \sim \mathrm{N}(e^\lambda - 1, 1)$ describe the same set of distributions as $\lambda \geq 0$ varies, but they correspond to different batch steganography problems,

the latter much less favourable for the steganographer because the distribution $D'(\lambda)$ moves away from $D'(0)$ much faster than $D(\lambda)$ from $D(0)$. (However, Theorem 2 demonstrates that the two problems have the same solution).

In the formulation of [3] the optimal solution – spread the payload equally between all covers – makes sense intuitively. But in the perturbation hiding problem the same solution is not so clearly optimal. For when the payload is spread equally between all covers, the opponent *does* know everything about the allocation of payload. Unevenly-spread payload has an apparent advantage of keeping the opponent guessing about its location.

2.2 Solution for Suitably Convex Exponential Families

Let us write $f(x;\lambda)$ for the density function of the distribution $D(\lambda)$. The solution of (2) can depend on f, but we will demonstrate that the symmetrical vector $\boldsymbol{\lambda} = (l, l, \ldots, l)$ is the solution for a certain class of functions f. Thus, for these distribution families, the disadvantage in allowing the opponent to know everything about the allocation of payload is outweighed by the advantage in having no object containing more payload than the necessary minimum.

Theorem 1. *A sufficient condition for $(l, l, \ldots, l)$ to be the solution to (2) is*

$$\sum_{i=1}^{n} \log \mathrm{E}\left[\frac{f(X;\lambda_i)}{f(X;\bar{\lambda})}\right]_{X\sim \mathrm{D}(0)} \;\leq\; 0 \tag{3}$$

for all choices of $\boldsymbol{\lambda}$.

Proof. Let us identify the distribution of observations at $\boldsymbol{\lambda} = (\bar{\lambda}, \ldots, \bar{\lambda})$: $\boldsymbol{Z} \sim D(\bar{\lambda})^n$. Assuming (3), we must show that $\mathrm{D}_{\mathrm{KL}}(\boldsymbol{X} \,\|\, \boldsymbol{Y}) \geq \mathrm{D}_{\mathrm{KL}}(\boldsymbol{X} \,\|\, \boldsymbol{Z})$ for all choices of $\boldsymbol{\lambda}$. Considering the difference,

$$\begin{aligned}
&\mathrm{D}_{\mathrm{KL}}(\boldsymbol{X} \,\|\, \boldsymbol{Y}) - \mathrm{D}_{\mathrm{KL}}(\boldsymbol{X} \,\|\, \boldsymbol{Z}) \\
&\quad= \mathrm{E}\left[-\log\left(\frac{1}{n!}\sum_{\pi\in S_n}\prod_{i=1}^{n}\frac{f(X_i;\lambda_{\pi(i)})}{f(X_i;0)}\right) + \log\left(\prod_{i=1}^{n}\frac{f(X_i;\bar{\lambda})}{f(X_i;0)}\right)\right]_{\boldsymbol{X}\sim D(0)^n} \\
&\quad= \mathrm{E}\left[-\log\frac{1}{n!}\sum_{\pi\in S_n}\prod_{i=1}^{n}\frac{f(X_i;\lambda_{\pi(i)})}{f(X_i;\bar{\lambda})}\right]_{\boldsymbol{X}\sim D(0)^n} \\
&\quad\overset{(1)}{\geq} -\log\left(\frac{1}{n!}\sum_{\pi\in S_n}\mathrm{E}\left[\prod_{i=1}^{n}\frac{f(X_i;\lambda_{\pi(i)})}{f(X_i;\bar{\lambda})}\right]_{\boldsymbol{X}\sim D(0)^n}\right) \\
&\quad\overset{(2)}{=} -\log \mathrm{E}\left[\prod_{i=1}^{n}\frac{f(X_i;\lambda_i)}{f(X_i;\bar{\lambda})}\right]_{\boldsymbol{X}\sim D(0)^n} \\
&\quad\overset{(3)}{=} -\log\prod_{i=1}^{n}\mathrm{E}\left[\frac{f(X;\lambda_i)}{f(X;\bar{\lambda})}\right]_{X\sim D(0)} \\
&\quad= -\sum_{i=1}^{n}\log \mathrm{E}\left[\frac{f(X;\lambda_i)}{f(X;\bar{\lambda})}\right]_{X\sim D(0)} \;\geq\; 0.
\end{aligned}$$

(1) is by Jensen's inequality[4] and linearity of expectation, (2) by identical distribution of the X_i, and (3) by their independence. ■

Now we demonstrate families of distributions for which the condition in Theorem 1 holds. Recall that a one-parameter family in λ is an *exponential family* [11] if the density function can be written in the form

$$f(x;\lambda) = h(x)\exp\bigl\{\eta(\lambda)T(x) - A(\lambda)\bigr\}$$

for functions h, η, T and A. Then $T(x)$ is a sufficient statistic, and $A(\lambda)$ is the normalizing constant determined by $\exp\bigl(A(\lambda)\bigr) = \int h(x)\exp\bigl\{\eta(\lambda)T(x)\bigr\}\,\mathrm{d}x$. When η is invertible the family can be re-parameterized to fit the form $f(x;\mu) = h(x)\exp\bigl\{\mu T(x) - \bar{A}(\mu)\bigr\}$ and in such cases μ is called the *natural parameter*. Even when the parameterization cannot be chosen (as in the perturbation hiding problem: we must solve the problem for the parameterization we are given) it is often cleaner to phrase results in terms of a natural parameter, as is the case here.

Theorem 2. *A sufficient condition for (3) is that f is a one-parameter exponential family for which a natural parameter exists, such that (a) η is convex nondecreasing, and (b) A'' is nondecreasing in the natural parameter.*

One such case is $D(\lambda_i) \sim \mathrm{N}\bigl(\phi(\lambda_i), \sigma^2\bigr)$, when ϕ is continuous and convex increasing, and σ^2 any positive constant.

Proof. We compute

$$\begin{aligned}\mathrm{E}\left[\tfrac{f(X;\lambda_i)}{f(X;\bar{\lambda})}\right] &= \int h(x)e^{T(x)\left(\eta(\lambda_i)-\eta(\bar{\lambda})+\eta(0)\right)-A(\lambda_i)+A(\bar{\lambda})-A(0)}\,\mathrm{d}x \\ &= \exp\Bigl\{A\circ\eta^{-1}\bigl(\eta(\lambda_i)-\eta(\bar{\lambda})+\eta(0)\bigr) - A(\lambda_i) + A(\bar{\lambda}) - A(0)\Bigr\}\end{aligned}$$

because of the relationship $\exp\bigl(A(\lambda)\bigr) = \int h(x)\exp\bigl\{\eta(\lambda)T(x)\bigr\}\,\mathrm{d}x$. If we write $\bar{A} = A\circ\eta^{-1}$, expressing A in terms of the natural parameter, the log of the expectation is equal to $g(\lambda_i)$ where

$$g(\theta) = \bar{A}\bigl(\eta(\theta) - \eta(\bar{\lambda}) + \eta(0)\bigr) - \bar{A}\bigl(\eta(\theta)\bigr) + \bar{A}\bigl(\eta(\bar{\lambda})\bigr) - \bar{A}\bigl(\eta(0)\bigr).$$

Note that $g(\bar{\lambda}) = 0$ and

$$g''(\theta) = \eta'(\theta)\bigl[\bar{A}''\bigl(\eta(\theta)-c\bigr) - A''\bigl(\eta(\theta)\bigr)\bigr] + \eta''(\theta)\bigl[\bar{A}'\bigl(\eta(\theta)-c\bigr) - A'\bigl(\eta(\theta)\bigr)\bigr]$$

where c is the constant $\eta(\bar{\lambda}) - \eta(0)$. Then use our assumptions: η' is nonnegative because η is nondecreasing; for the same reason $c \geq 0$ and so $\bar{A}''\bigl(\eta(\theta)-c\bigr) - A''\bigl(\eta(\theta)\bigr) \leq 0$ because $\bar{A}''$ is nondecreasing; η'' is nonnegative because η is convex; $\bar{A}'\bigl(\eta(\theta)-c\bigr) - A'\bigl(\eta(\theta)\bigr) \leq 0$ because $\bar{A}$ must be convex (this is always true for an exponential family). We deduce that $g'' \leq 0$.

[4] $\mathrm{E}[\phi(X)] \geq \phi(\mathrm{E}[X])$ for convex ϕ; note that $\phi(x) = -\log x$ is a convex function.

Therefore g is concave, hence

$$\sum \log \mathrm{E}\Big[\tfrac{f(X;\lambda_i)}{f(X;\bar{\lambda})}\Big]_{X\sim \mathrm{D}(0)} = \sum g(\lambda_i) \;\le\; g(\bar{\lambda}) + g'(\bar{\lambda}) \sum (\lambda_i - \bar{\lambda}) \;=\; 0.$$

For the Gaussian case mentioned, write the pdf of $N\big(\phi(\lambda), \sigma^2\big)$ in the exponential family form: $f(x;\lambda) = \frac{e^{-x^2/2\sigma^2}}{\sqrt{2\pi\sigma^2}} \exp\big\{\frac{\phi(\lambda)}{\sigma^2}x - \frac{1}{2\sigma^2}\phi(\lambda)^2\big\}$. ϕ must be invertible, so $\mu = \phi(\lambda)/\sigma^2$ is the natural parameter. By assumption, ϕ is convex increasing, and $A(\mu) = \mu^2\sigma^2/2$ satisfies A'' nondecreasing. ■

The given example is relevant because of adaptive source coding. Suppose, as a simple example, that the covers are Gaussian $\mathrm{N}(c, 1)$ where c is the number of embedding changes in a cover of size N. It is well-known [12] that the size of transmitted payload p satisfies $c \ge NH^{-1}\big(\frac{p}{N}\big)$ (H is the binary entropy function), a convex function of p. The relationship between bits transmitted and locations changed should always be convex in efficient codes, so this nonlinear relationship does not affect the conclusion that payload should be equally spread.

The conditions in Th. 2 seem natural. Monotonicity of η precludes the possibility that increasing payload is less detectable (in single objects). Some sort of convexity condition could be expected. And recall that, when A is expressed in terms of a natural parameter, A'' is the variance of the random variable: the condition that A'' is nondecreasing ensures that we do not have more certainty about larger payloads. But we should note that the conditions in Theorems 1 and 2 are stronger than necessary. It is possible to construct exponential families which do not satisfy the conditions of the latter, but do satisfy the former, and we will see next that there are distributions which do not form an exponential family at all, yet the solution to (2) is still (for some choices of n and l) the constant vector $(l, l, \ldots, l)$. We hope to widen the results in future work.

2.3 Explorations with Student *t*-Families

We now ask whether the preceding results apply more widely, when $D(\lambda)$ is not an exponential family. We might expect that the same conclusions should hold for random variables with exponentially-decaying tails, but for long-tailed distributions it might be optimal to concentrate the payload in a few cover objects. This would be plausible because, in the case of long tails, a small number of extreme observations would be expected even when no payload is present, so mimicking this could be a sensible embedding choice.

To explore this question, we performed some numerical computations. However, accurate estimation of the KL divergence

$$\mathrm{D}_{\mathrm{KL}}(\boldsymbol{X} \,\|\, \boldsymbol{Y}) = \mathrm{E}\left[-\log\left(\frac{1}{n!}\sum_{\pi\in S_n}\prod_{i=1}^{n}\frac{f(X_i;\lambda_{\pi(i)})}{f(X_i;0)}\right)\right]_{\boldsymbol{X}\sim D(0)^n} \tag{4}$$

– an integral over $\mathbb{R}^n$ of a function with $n!$ terms – represents a huge challenge unless n is very small. But even the cases $n = 2$ and $n = 3$ are suggestive.

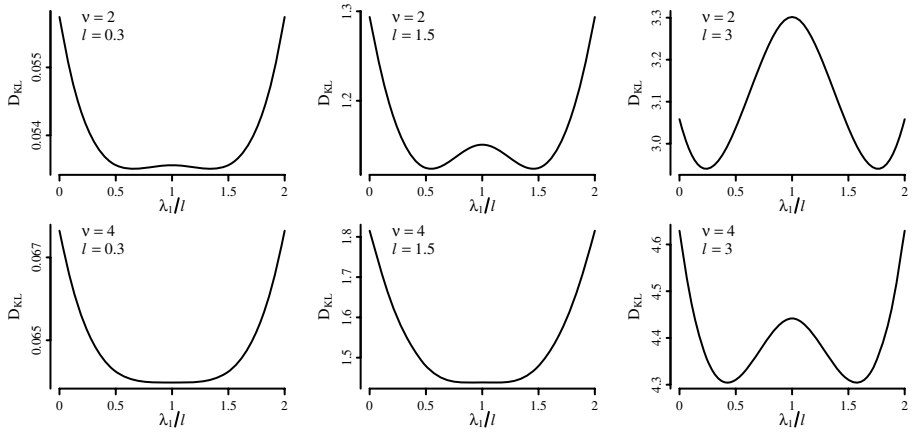

Fig. 1. The case $n = 2$. Numerically-computed $\mathrm{D}_{\mathrm{KL}}(\boldsymbol{X} \,\|\, \boldsymbol{Y})$ as a function of λ_1, when the parent distribution family is Student t with 2 d.f. (above) and 4 d.f. (below). Three different values of l are displayed.

We will concentrate on families of the Student t-distribution, determined by the *degrees of freedom* $\nu > 0$, and parameterized by location:

$$f(x;\lambda) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi}\,\Gamma\left(\frac{\nu}{2}\right)}\left(1+\frac{(x-\lambda)^2}{\nu}\right)^{-\frac{(\nu+1)}{2}}. \tag{5}$$

They were chosen because they never form an exponential family, and describe a continuum of distributions with varying tail weights: as $\nu \to \infty$ the distribution tends to Gaussian, and for $\nu = 1$ it is Cauchy, so heavily-tailed that even the expectation is not defined. Between these extremes, the density function $f(x;\lambda)$ tail decays as $|x|^{-(\nu+1)}$. We are interested in solutions to the perturbation hiding problem, with distribution family determined by ν and total payload by l.

First, we take the case $n = 2$, corresponding to splitting a steganographic payload between just two objects, with λ_1 in one object and λ_2 the other (the opponent does not know which is which), subject to $\lambda_1 + \lambda_2 = 2l$. For a number of different Student t-families $D(\lambda)$, determined by ν, and various values of l, we estimated the KL divergence, using quadrature, as λ_1 varies between 0 to $2l$. A selection of results are shown in Fig. 1, corresponding to the families $\nu = 2$ or $\nu = 4$ and $l = 0.3, 1.5, 3$. The figures are, of course, symmetrical because of symmetry between λ_1 and λ_2.

Regardless of ν, we observed that, as l grows large, the case of equally-spread payload $\lambda_1 = \lambda_2 = l$ eventually becomes the *worst* choice: the optimal choice of λ_1 is somewhere between 0 (concentrate payload) or l (spread equally), decreasing as l increases. Hence the conclusion of Theorem 1 cannot hold universally.

More interestingly, we observed distinct behaviour as $l \to 0$, depending on ν. As appears in Fig. 1, even for very small values of l the choice $\lambda_1 = \lambda_2 = l$ is not optimal for $\nu = 2$, but for l smaller than approximately 1.06 it *is* optimal for $\nu = 4$: although the curve for $\nu = 4$ and small l is very flat near the centre,

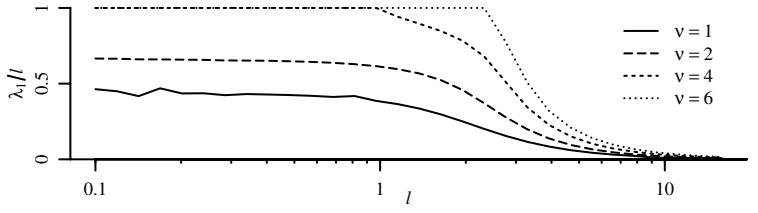

Fig. 2. Below, the numerically-determined optimal values of λ_1 as l varies, when the parent family is from the Student t-family. Four different d.f.s are displayed.

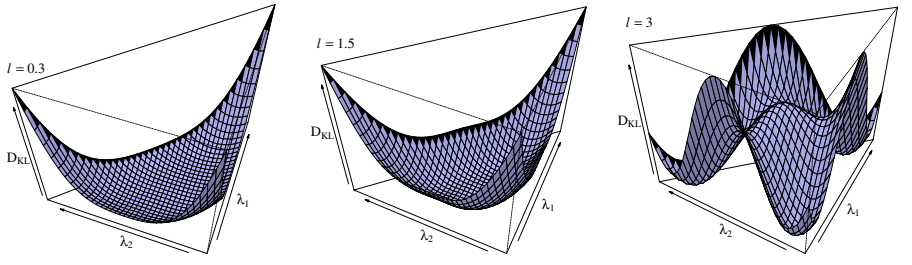

Fig. 3. The case $n = 3$. Numerically-computed $\mathrm{D_{KL}}(\boldsymbol{X} \,\|\, \boldsymbol{Y})$, as a function of λ_1 and λ_2, when the parent family is Student t with 4 d.f.

we do observe a genuine minimum at the central point. Figure 2 examines this further, showing how the optimal value of λ_1 (to minimize (4), assuming $\lambda_1 \leq \lambda_2$) depends on l and ν. For t-families with sufficiently light tails (apparently ν at least approximately 4), we can say that small enough total payloads are best spread equally between the covers, but we cannot say the same of heavily-tailed t-families (this is seen at least for $\nu \leq 3$). These results are only illustrative, but they motivate study of (4) as $l \to 0$; this will be performed in Sect. 3, where the critical change of behaviour near $\nu = 4$ will be explained.

We performed similar experiments for the case $n = 3$; charts for $\nu = 4$ and $l = 0.3, 1.5, 3$ are shown in Fig. 3, but others will not be included for reasons of space. The same features are apparent as for $n = 2$: when $\nu = 4$, for sufficiently small l (less than approximately 1.04), equally-spread payload $\lambda_1 = \lambda_2 = \lambda_3 = l$ gives the lowest KL divergence, but not when $\nu \leq 3$. Additional numerical explorations (of necessity not very thorough) show similar behaviour for $n = 4$ and, pushing our ability to compute (4) numerically to the limit, $n = 5$. The critical values of l do not seem to vary much with n, remaining a little over 1 in all observed cases, suggesting that the limiting result comes into play for larger payloads when there are more objects in which to hide.

3 Asymptotic Results

Motivated by the numerical results, we consider the asymptotics of the perturbation hiding problem. The results are presented briefly and some details are

omitted. We will assume considerable regularity without justifying it here. We will consider small payloads, and as $l \to 0$, $\boldsymbol{\lambda} \to \mathbf{0}$. Using the notation

$$\ell(x;\lambda) = \log f(x;\lambda), \quad \ell_\lambda(x) = \frac{\partial}{\partial\lambda} l(x;\lambda)|_{\lambda=0}, \quad \ell_{\lambda\lambda}(x) = \frac{\partial^2}{\partial\lambda^2} l(x;\lambda)|_{\lambda=0},$$

we can show:

Theorem 3. *Assuming sufficient regularity, as $\boldsymbol{\lambda} \to \mathbf{0}$ we have*

$$\mathrm{D}_{\mathrm{KL}}(\boldsymbol{X},\boldsymbol{Y}) \sim c_1\big(\textstyle\sum \lambda_i\big)^2 + c_2\big(\textstyle\sum \lambda_i\big)^3 + c_3\big(\textstyle\sum \lambda_i^2\big)\big(\textstyle\sum \lambda_i\big) + O(|\boldsymbol{\lambda}|^4) \qquad (6)$$

where

$$c_1 = \frac{1}{2n}\mathrm{E}\big[\ell_\lambda(X)^2\big], \quad c_2 = -\frac{1}{3n^2}\mathrm{E}\big[\ell_\lambda(X)^3\big], \quad c_3 = \frac{1}{2n}\mathrm{E}\big[\ell_\lambda(X)^3 + \ell_\lambda(X)\ell_{\lambda\lambda}(X)\big].$$

Proof. The full proof is laborious but routine, so we include only illustrative sections. In an effort to keep notation brief, let us write

$$F(\boldsymbol{x};\boldsymbol{\lambda}) = \prod_{i=1}^n f(x_i;\lambda_i), \quad F_i = \frac{\partial F}{\partial\lambda_i}, \quad F_{ij} = \frac{\partial^2 F}{\partial\lambda_i\partial\lambda_j}, \quad L = \log\Big(\tfrac{1}{n!}\sum_{\pi\in S_n} F\big(\boldsymbol{x};\pi(\boldsymbol{\lambda})\big)\Big).$$

Then

$$\begin{aligned}
\mathrm{D}_{\mathrm{KL}}(\boldsymbol{X}\,\|\,\boldsymbol{Y}) &= \mathrm{E}\bigg[-\log\bigg(\frac{\frac{1}{n!}\sum_{\pi\in S_n} F\big(\boldsymbol{X};\pi(\boldsymbol{\lambda})\big)}{F(\boldsymbol{X};\mathbf{0})}\bigg)\bigg]_{\boldsymbol{X}\sim D(0)^n} \\
&= \mathrm{E}\Big[-L(\boldsymbol{X};\boldsymbol{\lambda}) + L(\boldsymbol{X};\mathbf{0})\Big] \\
&= -\frac{1}{1!}\sum_{i=1}^n \lambda_i \mathrm{E}\big[\tfrac{\partial L}{\partial\lambda_i}\big|_{\boldsymbol{\lambda}=\mathbf{0}}\big] - \frac{1}{2!}\sum_{i,j=1}^n\sum \lambda_i\lambda_j \mathrm{E}\big[\tfrac{\partial^2 L}{\partial\lambda_i\lambda_j}\big|_{\boldsymbol{\lambda}=\mathbf{0}}\big] \\
&\quad -\frac{1}{3!}\sum_{i,j,k=1}^n\sum\sum \lambda_i\lambda_j\lambda_k \mathrm{E}\big[\tfrac{\partial^3 L}{\partial\lambda_i\lambda_j\lambda_k}\big|_{\boldsymbol{\lambda}=\mathbf{0}}\big] + O(|\boldsymbol{\lambda}|^4) \qquad (7)
\end{aligned}$$

where, at the last stage, we have assumed sufficient regularity to allow a Taylor expansion of L under the integral, in the second vector parameter, about $\boldsymbol{\lambda} = \mathbf{0}$. The expression will simplify because of the symmetry in L, and

$$\mathrm{E}\Big[\tfrac{F_s(\boldsymbol{X};\mathbf{0})}{F(\boldsymbol{X};\mathbf{0})}\Big] = \mathrm{E}\Big[\tfrac{F_{st}(\boldsymbol{X};\mathbf{0})}{F(\boldsymbol{X};\mathbf{0})}\Big] = 0 \qquad (8)$$

(in evaluating the expectation, the denominator cancels with the density function, and given sufficient regularity we can take the derivative of the numerator outside the integral; differentiating a constant gives zero).

Therefore for the first- and second-order terms in the Taylor expansion of L,

$$\mathrm{E}\bigg[\tfrac{\partial}{\partial\lambda_i} L(\boldsymbol{X};\boldsymbol{\lambda})\Big|_{\boldsymbol{\lambda}=\mathbf{0}}\bigg] = \mathrm{E}\Big[\tfrac{\sum_\pi F_{\pi^1(i)}(\boldsymbol{X};\pi(\boldsymbol{\lambda}))}{\sum_\pi F(\boldsymbol{X};\pi(\boldsymbol{\lambda}))}\big|_{\boldsymbol{\lambda}=\mathbf{0}}\Big] = \frac{1}{n}\sum_{s=1}^n \mathrm{E}\Big[\tfrac{F_s(\boldsymbol{X};\mathbf{0})}{F(\boldsymbol{X};\mathbf{0})}\Big] = 0,$$

$$\begin{aligned}
&\mathrm{E}\bigg[\tfrac{\partial^2}{\partial\lambda_i\partial\lambda_j} L(\boldsymbol{X};\boldsymbol{\lambda})\Big|_{\boldsymbol{\lambda}=\mathbf{0}}\bigg] \\
&\quad = \mathrm{E}\Big[\tfrac{\sum_\pi F_{\pi^1(i)\pi^1(j)}(\boldsymbol{X};\pi(\boldsymbol{\lambda}))}{\sum_\pi F(\boldsymbol{X};\pi(\boldsymbol{\lambda}))}\big|_{\boldsymbol{\lambda}=\mathbf{0}}\Big] - \mathrm{E}\Big[\tfrac{\sum_\pi F_{\pi^1(i)}(\boldsymbol{X};\pi(\boldsymbol{\lambda}))}{\sum_\pi F(\boldsymbol{X};\pi(\boldsymbol{\lambda}))}\tfrac{\sum_\pi F_{\pi^1(j)}(\boldsymbol{X};\pi(\boldsymbol{\lambda}))}{\sum_\pi F(\boldsymbol{X};\pi(\boldsymbol{\lambda}))}\big|_{\boldsymbol{\lambda}=\mathbf{0}}\Big]
\end{aligned}$$

$$= \left\{ \begin{array}{ll} \frac{1}{n(n-1)} \sum\limits_{s \neq t} \sum \mathrm{E}\Big[\frac{F_{st}(\boldsymbol{X};\boldsymbol{0})}{F(\boldsymbol{X};\boldsymbol{0})}\Big], & \text{if } i \neq j \\ \frac{1}{n} \sum\limits_{s} \mathrm{E}\Big[\frac{F_{ss}(\boldsymbol{X};\boldsymbol{0})}{F(\boldsymbol{X};\boldsymbol{0})}\Big], & \text{if } i = j \end{array} \right\} - \frac{1}{n^2} \mathrm{E}\Big[\sum_{s=1}^{n} \frac{F_s(\boldsymbol{X};\boldsymbol{0})}{F(\boldsymbol{X};\boldsymbol{0})}\Big]^2$$

$$= 0 - \tfrac{1}{n}\mathrm{E}[\ell_\lambda(X)^2]_{X \sim D(0)} = -2c_1$$

(At the final stage, we used (8) along with independence of $\frac{F_s(\boldsymbol{X};\boldsymbol{0})}{F(\boldsymbol{X};\boldsymbol{0})}$ and $\frac{F_t(\boldsymbol{X};\boldsymbol{0})}{F(\boldsymbol{X};\boldsymbol{0})}$ for $s \neq t$.) We observe that the first two terms of (7) together match the first term of (6). The third term of (7) reduces to the second and third terms of (6) for similar reasons, but the calculations are longer (because there are more types of mixed partial derivative at third order) and we omit them here. ■

We can draw some useful conclusions from Theorem 3, because the first two terms of (6) cannot be varied by choice of $\boldsymbol{\lambda}$, if $\bar{\lambda} = \frac{1}{n}\sum \lambda_i$ is constrained to equal l. Therefore, a) the steganographer's choice of $\boldsymbol{\lambda}$ can only affect the second-most significant term as $l \to 0$, and b) they should minimize $c_3\big(\sum \lambda_i^2\big)$. If c_3 is positive, the minimum is again at $\lambda_i = l$ (for all i), but if c_3 is negative then the minimum is found on the edge of the feasible region, where some λ_i are zero (we will not proceed to find the location of the minimum, in this paper). We have shown that the sign of $\mathrm{E}\big[\ell_\lambda(X)^3 + \ell_\lambda(X)\ell_{\lambda\lambda}(X)\big]_{X \sim D(0)}$, a constant depending on the distribution family and related to its skewness, determines the optimal strategy for sufficiently small l.

Does this explain the phenomena in Subsect. 2.3, where the degree of freedom parameter appeared critical to whether equal payload was optimal as $l \to 0$? Sadly not, because in a symmetrical distribution, parameterized by location, we always have $c_3 = 0$! We have proved that the effect of payload allocation is (at most) of order l^2 smaller than the leading term in (6) (hence the very flat-looking curves in Subsect. 2.3), but must continue the asymptotic analysis of Th. 3 to the fourth order to understand the asymptotically optimal strategy.

The calculations are of a similar type to those in the proof of Th. 3, but much more complex. We spare the reader the details, and simply state that the fourth-order terms in (6) are

$$c_4\big(\textstyle\sum \lambda_i\big)^4 + c_5\big(\sum \lambda_i^3\big)\big(\sum \lambda_i\big) + c_6\big(\sum \lambda_i^2\big)^2 + c_7\big(\sum \lambda_i^2\big)\big(\sum \lambda_i\big)^2$$

where

$$c_5 = \tfrac{1}{18n}d_2 - \tfrac{1}{6n}d_3, \quad c_6 = \tfrac{1}{4n(n-1)}d_1 + \tfrac{1}{24n}d_2 + \tfrac{1}{8n}d_3, \quad c_7 = \tfrac{(n-2)}{2n^2(n-1)}d_1 - \tfrac{1}{3n^2}d_2,$$
$$d_1 = \mathrm{E}\big[\ell_\lambda(X)^2\big]^2, \quad d_2 = \mathrm{E}\big[\ell_\lambda(X)^4\big], \quad d_3 = \mathrm{E}\big[\ell_{\lambda\lambda}(X)^2\big].$$

(c_4 is not relevant to the location of the minimum). Considering the Hessian at the central point, it can be shown that $\lambda_i = l$ is a (local) minimum if and only if

$$-6d_1 + d_2 + 3d_3 < 0 \tag{9}$$

(independently of l and n).

Finally, for the Student t-family (5) one can compute d_1, d_2 and d_3 in terms of the d.f. parameter ν: (9) turns out equivalent to

$$\nu^3 + 2\nu^2 - 15\nu - 20 > 0$$

which is true for $\nu > 3.6367\cdots$. This explains the behaviour seen in Subsect. 2.3.

4 Conclusions

The batch steganography problem is of importance to covert communication and storage, posing a fundamental question about the allocation of payload between multiple objects. Some other work has addressed special cases, but in this paper we have attacked the general problem. Perturbation hiding is a mathematical abstraction of the batch steganography problem – at a different level of abstraction to most of the literature on information-theoretic analyses of covert communication – and we have given some results about its solutions. It is likely that the results in Subsect. 2.2 can be extended to wider families of distributions, and the asymptotic results of Sect. 3 deserve a more rigorous analytical treatment. An asymptotic result as $n \to \infty$ would also be useful: perhaps Laplace's method can be applied.

We chose the problem formulation after a careful consideration of how much information should be granted to the steganalyst. The model should be seen as conservative: we do not necessarily believe that the steganalyst always knows the size of the individual payloads (without knowing their order), we merely *fear* that they might find out, perhaps by later compromising a recipient: such paranoia is in keeping with the spirit of Kerckhoffs' Principle. It may seem natural for the detector to try to gain information about payload allocation using a *quantitative* steganalysis (such payload size estimators are common) but the use of KL divergence as an insecurity measure limits the ability of *any* detector, including those who first apply estimators.

We would like to conclude that the steganographer's best choice is to spread payload equally between covers (as long as the covers are uniform), and thus the benefits of well-spread payload outweigh the drawbacks of the opponent having no uncertainty about the amount in each object. We have proved that this is so for suitably convex exponential distribution families, and for sufficiently small payloads if the critical value c_3, a constant depending on the distribution family, is positive. However it is not so for when c_3 is negative, or for large payloads. In order to inform the practice of covert communication, and its counterpart in steganalysis, it will be necessary to clarify circumstances under which these dichotomous situations occur. A first stage would be to relate c_3 to the tail behaviour of the family. It would be attractive if a simple test can be developed for genuine cover media, to determine the best embedding strategy.

For tractability and compactness, our analysis in this work has been limited to uniform covers. We note that uniformity does not necessary mean that the covers are truly uniform, merely that the parties do not know how, or do not choose, to take advantage of nonuniformity. Although it is folklore that more data can

securely be hidden in "noisier" covers there is not much literature quantifying this, so state-of-the-art steganography is not in a good position to make use of nonuniformity. More work on the perturbation hiding problem may be valuable here, perhaps producing a rule for allocating payload in nonuniform covers.

We have already performed a simple small-payload analysis of the perturbation hiding problem in nonuniform covers, but postpone it to a sequel. The results are quite interesting: it is in the steganographer's interest to distribute payload unevenly between the covers, but also to randomise the distribution: unlike in the uniform case, it does pay to keep the opponent guessing as to the distribution of payload.

Acknowledgements

The author is a Royal Society University Research Fellow.

References

1. Ker, A.: Batch steganography and pooled steganalysis. In: Camenisch, J.L., Collberg, C.S., Johnson, N.F., Sallee, P. (eds.) IH 2006. LNCS, vol. 4437, pp. 265–281. Springer, Heidelberg (2007)
2. Ker, A.: Batch steganography and the threshold game. In: Security, Steganography and Watermarking of Multimedia Contents IX. In: Proc. SPIE, vol. 6505, pp. 0401–0413 (2007)
3. Ker, A.: Steganographic strategies for a square distortion function. In: Security, Forensics, Steganography and Watermarking of Multimedia Contents X. In: Proc. SPIE, vol. 6819 (2008)
4. Ker, A.: A capacity result for batch steganography. IEEE Signal Processing Letters 14(8), 525–528 (2007)
5. Cachin, C.: An information-theoretic model for steganography. Information and Computation 192(1), 41–56 (2004)
6. Wang, Y., Moulin, P.: Perfectly secure steganography: Capacity, error exponents, and code constructions. IEEE Trans. Information Theory (to appear, 2008)
7. Kullback, S., Leibler, R.: On information and sufficiency. Annals of Mathematical Statistics 22, 79–86 (1951)
8. Kerckhoffs, A.: La cryptographie militaire. Journal des sciences militaires IX, 5–38, 161–191 (1883)
9. Cayre, F., Bas, P.: Kerckhoffs-based embedding security classes for WOA data-hiding. IEEE Trans. Information Forensics and Security (to appear, 2008)
10. Dolev, D., Yao, A.: On the security of public key protocols. IEEE Trans. Information Theory 29(2), 198–208 (1983)
11. Darmois, G.: Sur les lois de probabilité à estimation exhaustive. Comptes Rendus de l'Académie des Sciences 200, 1265–1266 (1935)
12. Fridrich, J., Soukal, D.: Matrix embedding for large payloads. IEEE Trans. Information Forensics and Security 1(3), 390–394 (2006)