The Ultimate Steganalysis Benchmark?

Andrew D. Ker
Oxford University Computing Laboratory
Parks Road
Oxford OX1 3QD, UK
adk@comlab.ox.ac.uk

ABSTRACT

We present a new benchmark for binary steganalysis methods, based on the asymptotic information (in the entropic sense) it gives about the presence of hidden data. The theoretical foundation is quite unlike ad hoc performance measures found in steganalysis literature that are based on false positive and negative rates. It is argued that this new metric is an application-independent long-run measure of true performance. There are some challenges to computing the benchmark empirically, and some suggested methods are presented, but no definitive answer emerges. As a simple case study, some steganalysis methods from the literature are evaluated using these techniques.

Categories and Subject Descriptors

D.2.11 [**Software Engineering**]: Software Architectures—*information hiding*; H.1.1 [**Models and Principles**]: Systems and Information Theory—*information theory*

General Terms

Performance, Security

Keywords

Steganalysis, Steganographic Capacity, Benchmarking

1. INTRODUCTION

Initially, the literature on the competing fields of steganography (hiding information undetectably in cover objects) and steganalysis (detecting the presence of hidden information) consisted of ad hoc methods for the embedding and detection of data. One would expect that, alongside the maturation of the techniques, methods for measuring their efficacy would become more refined. This has indeed been the case with steganography: a key innovation, adapted by Cachin [2] from prior work on authentication, was to use the Kullback-Leibler (KL) divergence [13] between the distribution of cover and stego objects to bound the detectability of hidden data. Despite some difficulty applying the theory (because of the need for a realistic "distribution of cover objects") this benchmark has aided further understanding of the fundamental aims of steganography, driving better embedding methods.

Benchmarks for the efficacy of steganalysis schemes, on the other hand, have not become very sophisticated. As we will briefly survey in Sect. 2, a number of different metrics have been used in the literature, all of which can be criticized for being relevant to only a limited range of applications. Further, it is not uncommon for the various benchmarks to rate steganalysis methods inconsistently.

The aim of this paper is to present a new benchmark for (binary) steganalysis, based on the *information* it provides as to the presence or absence of hidden data. It is based on principles (empirically-estimated KL-divergence and asymptotic behaviour) which are quite different from those of the benchmarks commonly used in the literature. It is argued that this measurement, which ultimately produces a single number by which steganalysis methods (applied to a particular type of cover) may be ranked, constitutes an application-independent measurement of the long-term detection capability. The principles from which the metric is derived are outlined in Sect. 2, the practical challenges in computing it are addressed, although not optimally solved, in Sect. 3, and a few current steganalysis methods are benchmarked by the new metric in Sect. 4, to provide some examples. Brief conclusions appear in Sect. 5.

2. PRINCIPLES OF AN APPLICATION-INDEPENDENT BENCHMARK

A vital first step is to state the problem precisely. Fix an embedding method E. We are benchmarking a steganalysis method by its ability to distinguish an innocent *cover object* from a *stego object* containing a payload embedded by E, i.e. given a single object to perform the hypothesis test

$$\begin{array}{ll} H_0: & \text{there is no payload} \\ H_1: & \text{there is some payload embedded by } E. \end{array} \tag{1}$$

We have restricted our attention to the commonly-considered case called *binary steganalysis*: there is only one embedding method E and it is known. There do exist multi-class steganalysers which attempt to determine E, but benchmarking them is a more difficult problem and not addressed here. Although the embedding method is considered fixed, the alternative H_1 remains a composite hypothesis because the size of payload, if any, is not known (it would be unrealistic to assume that it is known). It is quite common for

MM&Sec'07, September 20–21, 2007, Dallas, Texas, USA.
Copyright 2007 ACM 978-1-59593-857-2/07/0009 ...$5.00.

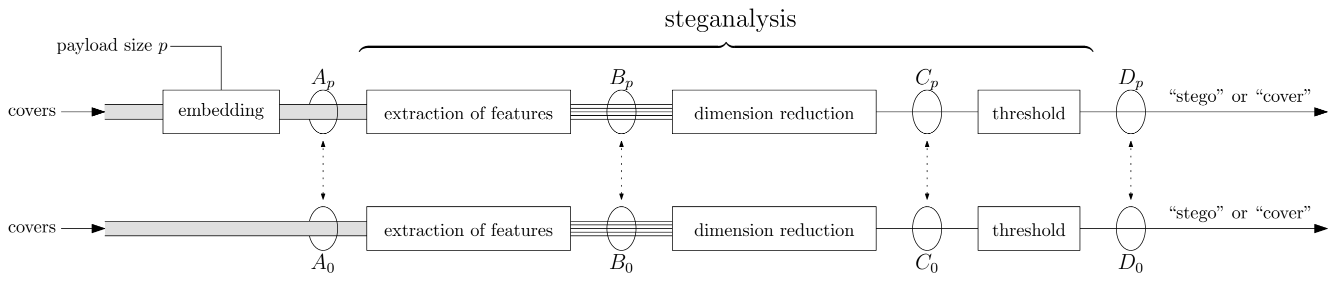

Figure 1: A conceptual breakdown of steganalysis into three stages, and differences in distribution at each point. Above, when a steganographic payload is embedded. Below, steganalysis with no payload.

steganalysis methods to make an estimate of the payload size, if any, (this is known as *quantitative steganalysis*) but we consider such estimation a secondary facet and not part of the benchmark.

In the literature there have arisen a number of standard benchmarks for the ability to perform hypothesis test (1). Simulating steganography on a set of cover objects, and then performing steganalysis, empirically-determined false positive (type I) and false negative (type II) errors are found, sometimes varying the detector's sensitivity to produce ROC curves. Because ROC curves often cross they do not usually allow competing steganalysis methods to be ranked, so they are commonly reduced further: to the area under the curve (AUC), or to the false positive rate for a fixed false negative rate (or vice versa). All of these measures are dependent on the size of embedded payload, which further complicates their use for comparison purposes. To condense to a one-dimensional metric, one can set a maximum tolerable false positive and false negative rate and determine experimentally the smallest payload size for which such detection is achieved.

These metrics are inevitably slanted towards certain applications. For example, if we measure sensitivity at a certain false positive rate, that only gives good information about steganalysis performance in applications with approximately the same false positive rate – and one can envisage applications ranging from intelligence (in which case false positives rates of a few percent are acceptable) to network scanning (in which case false positive rates must be many orders of magnitude lower, because of the number of tests carried out). Although it is common to fix a rate of, for example, 5%, it is impossible to justify such a choice objectively. Even the AUC is not application-independent because it gives equal weight to false positives and negatives: a very dubious principle for steganalysis. Only the full ROC family (one ROC for each possible payload size) could be called application-independent, and it is too large an amount of data to make for a usable benchmark.

Fundamentally, the aim of steganalysis is to give *information* about the presence or absence of payload. We propose a new measure which, like Cachin's framework for steganography, quantifies this information by the difference between the distributions of cover and stego objects, measured using KL-divergence: recall that, when f and g are the density functions for distributions F and G, $D_{\mathrm{KL}}(F,G)$ is defined as $\int f(x)\log\frac{f(x)}{g(x)}\,\mathrm{d}x$; the definition can be extended to non-continuous distributions with the appropriate abstract integral. Our logs will be taken to natural base.

We must measure the KL-divergence at the right stage of the steganalysis process. Let us breakdown a steganalysis method into three conceptual components. Consider Fig. 1, which displays the development of a cover object, through either embedding (with payload of size p) or not, and then subject to steganalysis. We argue that all steganalysis methods can be broken down into the three stages depicted: extraction of feature vectors from the object, use of some classification engine or payload estimation to reduce the dimensionality to one, and finally setting a simple threshold. If the one-dimensional value exceeds a threshold, the object is classified as a stego object, otherwise a cover. Sometimes, and often in steganalysis based on machine learning algorithms, the last two stages are not presented separately. Nonetheless, the author believes that such a breakdown can always be made because a classifier can (or, with extra work, can be made to) output a level of certainty as to its output, e.g. the distance of the observation from the classification boundary. The level of certainty becomes the one-dimensional value on which a sensitivity threshold can be set. (If we are interested only in KL-divergence then the exact form of this level of certainty is not important because KL-divergence is invariant under 1-1 transformations.)

This leads us to the first principle for the application-independent benchmark.

Principle 1. *Steganalysis performance should be measured by KL-divergence, at some intermediate stage before reduction to a binary decision, between covers and stego objects (as a function of payload size).*

In Fig. 1, the notations A_0, A_p, etc, represent the distribution of objects at each point. So A_p (parameterized by p) are the distributions of the stego objects with payload size p, B_p are the distributions of whatever feature vectors the steganalysis uses (a smaller dimensional distribution than A_p), C_p are the univariate distributions of the steganalysis prior to classification as cover or stego objects, and D_p are the binary distributions of the output. At which stage should we measure KL-divergence?

The value of $D_{\mathrm{KL}}(A_0, A_p)$ is certainly interesting, but it is a measure of how secure the embedding process is (cf. [2]), not the efficacy of the detector. At the other end, $D_{\mathrm{KL}}(D_0, D_p)$ is a poor measure because too much information is quantized away when a threshold is applied (indeed, it is the setting of a threshold which causes application-dependent benchmarks). It would be nice to evaluate the steganalysis features separately from the classifier by benchmarking on $D_{\mathrm{KL}}(B_0, B_p)$ but there are two problems. First, estimation of KL-divergence is much more difficult when there is large dimensionality, so practical considerations will usually preclude its measurement. Second, it would admit a nonsense result: the information processing theorem forces $D_{\mathrm{KL}}(B_0, B_p) \leq D_{\mathrm{KL}}(A_0, A_p)$, and therefore an "optimal" feature extraction method is the identity function! The paradox exists because high KL-divergence does not guarantee the existence of a good decision procedure. (When the null and alternative hypothesis are nonparametric the Neyman-Pearson Lemma gives an optimal decision procedure, but in the steganalysis problem the payload size is not known and there is often no uniformly most powerful test for composite hypothesis testing.) We should only consider steganalysis methods with a clear decision procedure.

This leaves $D_{\mathrm{KL}}(C_0, C_p)$, which is still not a simple benchmark because it depends on the payload size. We must decide on how "payload size" is to be quantified: we suggest that the size of the embedded data is, contrary to intuition, the wrong measure. All a steganalysis method can ever hope to do is detect *changes* in the cover, rather than the payload itself, and the existence of adaptive source-coding techniques [5] mean that the number of changes is not necessarily proportional to the payload size. Therefore we will measure payload by *the number of embedding changes* involved; this is acceptable as long as the embedding changes are of approximately equal magnitude: a common situation.

Now, return to the hypothesis test (1). It refers to the application of steganalysis to single objects, but a communications system does not consist of a single transmission. In the case of a stream of objects, steganalysis will be applied repeatedly, and most important is its *long run* performance. The application of steganography and steganalysis to multiple objects (so-called *batch steganography*) is considered in [9] and a result about capacity in the batch setting is found in [10]. We extract a key conclusion: if a steganographer is to act repeatedly, they *must* reduce the payload per object, as the number of objects increases. Whether or not you accept the hypotheses of the result in [10] it is hard to argue against the proposition that constant-rate embedding will cause evidence to build up over time, sufficient to guarantee detection eventually. Therefore we believe that long-term steganography must necessarily involve ever-decreasing payloads, leading to a second principle.

Principle 2. *For long-run performance it is sufficient to measure KL-divergence in the limit as embedding change rate tends to zero.*

The shape of such a limit is given by the following theorem (which is a much simplified statement of a result found in [13, §2.6]).

Theorem 1. *Suppose that $\{F_p \mid p \geq 0\}$ is a family of distributions indexed by p, satisfying Cramér's conditions [3]. Then $D_{\mathrm{KL}}(F_0, F_p)$ is locally quadratic at $p = 0$, i.e. the limit $\lim_{p\to 0} \frac{D_{\mathrm{KL}}(F_0,F_p)}{p^2}$ exists.*

Cramér's conditions are standard regularity conditions: the density function of the distribution family has to be differentiable three times, with the first two derivatives bounded by integrable functions and the third uniformly bounded in expectation. They allow Taylor expansion of the logarithm of the density function of F_p, and differentiation under the integral sign. More modern results are able to relax the conditions slightly, but we will not concern ourselves with their detail: it suffices to know that almost all standard nonpathological one-parameter distributions have the required properties, so we will assume that they hold for steganalysis outputs and omit the mathematics.

The theorem tells us that the information generated by steganalysis is governed by quadratic term in its power series. And we have argued that the limiting behaviour as $p \to 0$ is what matters to long-term repeated steganalysis. Finally, we have reached a position where steganalysis performance can be reduced to a single number:

$$Q = \lim_{p\to 0} \frac{D_{\mathrm{KL}}(C_0, C_p)}{p^2}$$

where p measures the number of embedding changes. Q (so-named because it is the coefficient of the quadratic term of $D_{\mathrm{KL}}(C_0, C_p)$) tells us everything about the amount of information, as to whether $p = 0$ or not, produced by the steganalysis method when considering its long-term sequential application, and it is the proposed new benchmark. In fact, Q is related to another fundamental statistical concept, which we shall discuss shortly.

3. COMPUTING THE BENCHMARK

How can the benchmark be computed? One possibility is to propose a model for cover objects and a model for steganography, then attempt to derive how p affects the distributions C_0 and C_p, hence computing Q, but there are many difficulties. First, realistic models for multimedia covers are difficult to find, and less-than-accurate models might well give wholly inaccurate answers: recall that it is precisely due to defects in cover models that many "undetectable" steganography schemes are broken (see e.g. [17]). Second, the feature extraction and dimension reduction stages are likely to be mathematically complex, so even if we know how p affects A_p, we may not be able to compute the consequence for C_p. One paper which makes some initial steps in this direction is [11], but complete analysis is likely to be so difficult that we will discard, for now, the possibility of deriving the benchmark theoretically.

The alternative is to use empirical data. Suppose that a large set of cover objects is selected and, in each of a randomly-selected half, a payload is embedded causing p changes. Applying steganalysis to each object, but stopping short of setting a threshold to give a binary classification, gives some experimental evidence about the distributions C_0 and C_p[1], and thus to an estimator for Q. There are then two challenges: to estimate $D_{\mathrm{KL}}(C_0, C_p)$ from finite samples drawn from the distributions C_0 and C_p, and to use this data to estimate the value of the limit $\lim_{p\to 0} D_{\mathrm{KL}}(C_0, C_p)/p^2$.

First, estimation of KL-divergence. Write $\hat{D}_{\mathrm{KL}}(F, G)$ for an estimator of the divergence $D_{\mathrm{KL}}(F, G)$. There is a fair amount of literature on this topic, with the simplest method being a *plug-in* approach, where kernel density estimation is performed separately for F and G, and the KL-divergence computed therefrom. A novel method by Wang et al. [16] is preferable because it demonstrates superior performance

[1] It is important that the set of test covers and test stego objects are distinct (i.e. the same covers are not re-used to create stego objects) to avoid correlations between the two data sets which might destroy the accuracy of the estimators.

(faster convergence) as well as an elementary description. We will not repeat the construction of this estimator; the reader is referred to [16] for all details.

Estimation of $\lim_{p\to 0} \hat{D}_{\mathrm{KL}}(C_0, C_p)/p^2$ is more difficult, since we must confine ourselves to a finite (preferably small) number of choices of p. A simple approach is to fix a "small" value of p, say ε, and use $\hat{D}_{\mathrm{KL}}(C_0, C_\varepsilon)/\varepsilon^2$ to approximate the limit. But determining a suitable value of ε is not easy: too large a value gives an answer too far from the limit, but too small a value means that $\hat{D}_{\mathrm{KL}}(C_0, C_\varepsilon)$ is itself too small to estimate with good accuracy unless the number of experiments is huge. All we can do is use initial data to hand-pick a sensible choice of ε; we will call this estimator $\hat{Q}_1$.

Or we could assume some slightly stronger regularity conditions than in Theorem 1 to conclude that $\hat{D}_{\mathrm{KL}}(C_0, C_p)/p^2$ can be fitted to a polynomial near $p = 0$ (we have found that a cubic is usually sufficient), then fit based on data for a small number of values near $p = 0$ to predict the limit at zero (again, there is no universal way to determine which data, but the answer should be less sensitive than with a single value of p). We will call this estimator $\hat{Q}_2$.

There is an alternative method to estimate Q. In [13, §2.6] is a continuation of Theorem 1, telling us that Q is in fact one half times *Fisher's Information*,

$$\int \left(\tfrac{\partial}{\partial p} \log f_p(x)\right)\Big|_{p=0} f_0(x)\,\mathrm{d}x \qquad (2)$$

(where f_p represents the density function of F_p). This is of no immediate help because Fisher's Information seem no easier to estimate than Q, but it becomes simpler if we make a further assumption about the effect of small payloads on steganalysis response.

Suppose that the effect of payload in C_p is locally linear in the number of embedding changes, i.e. if f_p is the density function of C_p then for p near zero and some constant λ,

$$f_p(x) \approx f_0(x - \lambda p). \qquad (3)$$

A steganalysis response with this property is called *quasi-linear* in [10], and it is argued that many or all steganalysis methods should have this property. In such a circumstance we can reduce (2) to a function of λ and f_0, giving:

$$Q = \frac{\lambda^2}{2} \int \frac{f_0'(x)^2}{f_0(x)}\,\mathrm{d}x. \qquad (4)$$

Estimating λ is straightforward (use the difference in mean between observations of zero and small payloads; a variety of payload sizes can be used) and it is possible to estimate the integral using a plug-in density estimator for f_0, based on the empirical observations from C_0: combining these estimates according to (4) gives an estimator we shall call $\hat{Q}_3$. It would be attractive if a data-dependent method similar to [16] could be used to estimate Fisher's Information but the author is not aware of such a technique in the literature. If (3) applies, we should expect to see consistent answers produced by all three estimators $\hat{Q}_1$, $\hat{Q}_2$, $\hat{Q}_3$.

We are under no illusions about the methods, for estimating Q, presented here. Although we believe that Q is in some sense an optimal benchmark, these methods for estimating it are far from optimal. They suffice as a temporary measure to allow Q itself to be investigated but further work is needed to find better estimators.

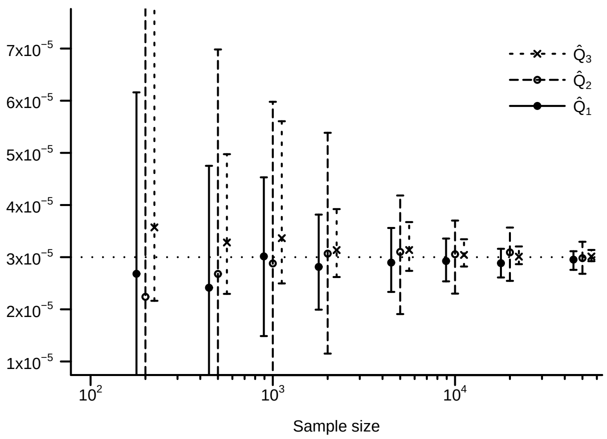

Figure 2: Comparison of three estimators for Q. Mean, and upper/lower 10% quantiles of 200 repeated experiments are shown, at each of 8 sample sizes (the estimators are shown staggered for clarity). The true Q-factor of 3×10^{-5} is indicated.

4. EXPERIMENTAL EXAMPLES

We include two types of experiments. First, some experiments on synthetic data, to try to determine which of the estimators $\hat{Q}_1$, $\hat{Q}_2$, $\hat{Q}_3$ gives the most accurate results. We will suppose a steganalysis method whose intermediate distribution C_p is from the Student t-distribution with 2 degrees of freedom, shifted to cause a mean of $0.01p$ (this decision is not arbitrary: up to a scaling factor, it is a well-fitted model for quantitative steganalysis found in [1]). It can be shown analytically that the true Q-factor for this family of distributions is 3×10^{-5}.

Eight sample sizes $N = 200$, 500, 1000, 2000, 5000, 10000, 20000, 50000 were tested, each repeated 200 times. Every sample was divided evenly between cover and stego objects. We tested values of p between 10 and 300 (some initial examples suggested taking $\varepsilon = 50$ as a reasonable trade-off between proximity to zero and robustness) and computed each estimator $\hat{Q}_1$, $\hat{Q}_2$, $\hat{Q}_3$ using the methods of Sect. 3. For each sample size, the observed mean and upper and lower 10% quantiles of the estimators are displayed in Fig. 2, as a summary of their accuracy.

From the displayed diagram we conclude that (for this particular shape of steganalysis response, at least) $\hat{Q}_3$ is the most accurate estimator (but recall that it makes an assumption about locally-linear steganalysis response) otherwise $\hat{Q}_1$ outperforms $\hat{Q}_2$, despite the latter's extra sophistication (trying to fit a polynomial to noisy data seems to accentuate the noise). The estimators give reasonable accuracy as long as a few thousand cover objects are available.

We now proceed to an example of genuine steganalysis. Consider LSB replacement embedding in bitmap images, and the Triples steganalysis method (also known as Triples/LSM) from [7]. Using a set of 10000 cover images (all approximately 0.6 Mpixels, full colour, bitmap images which had previously been subject to JPEG compression at quality factor 75) embedding changes were made in half of the images by flipping LSBs. Embedding change numbers of 50, 100, 150, …, 2500 were all tested, and $\hat{D}_{\mathrm{KL}}(C_0, C_p)/p^2$

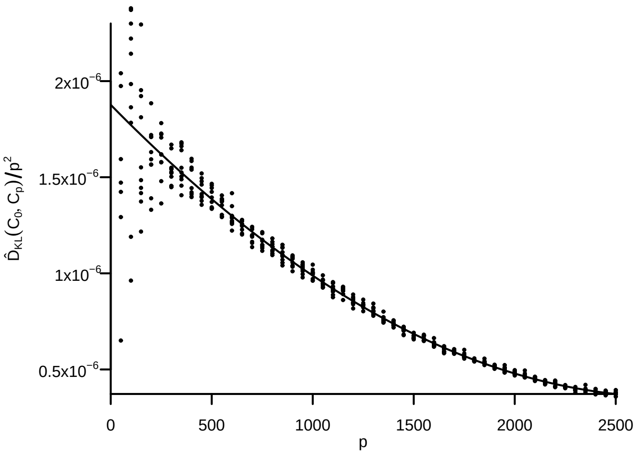

Figure 3: Experimental data from LSB Replacement detection in images (Triples steganalysis), showing how the quotient $\hat{D}_{\mathrm{KL}}(C_0, C_p)/p^2$ (y-axis) depends on the number of embedding changes p (x-axis); a polynomial fit is indicated.

was computed in each case. In order to make some assessment of the accuracy of the estimator for KL-divergence, each experiment was repeated 10 times with a different random allocation of the 10000 as cover and stego images. The results are plotted in Fig. 3, with a best-fit cubic displayed.

It does indeed appear that $D_{\mathrm{KL}}(C_0, C_p)$ is locally quadratic near $p = 0$, as Theorem 1 predicts, because the quotient appears to converge to a nonzero finite value. This figure also illustrates the tradeoffs of using small values of ε in estimating the limit: small values give $\hat{D}_{\mathrm{KL}}(C_0, C_\varepsilon)/\varepsilon^2$ closer to the limit at 0 but are more subject to estimation error. A compromise value of $\varepsilon = 300$ was therefore selected.

It is now possible to use the three techniques of Sect. 3 to estimate the Q-factor for this particular steganalysis, when applied to objects represented by this type of cover (for $\hat{Q}_3$ we used the plug-in density estimator of [15]). Similar experiments were also performed for three other steganalysis methods detecting the same type of steganography: SPA [4], SPA/LSM [14], and Triples/WLSM [12]. The three estimators for the Q-factor of each are displayed in Tab. 1. The benchmarking was performed both for the set of 10000 colour JPEGs already used, and another set of 3000 colour bitmap images which were never subject to compression, sized approximately 0.3 Mpixels. As an aside, note that the benchmark Q is not truly dimensionless. If all logarithms are to base e then KL-divergence is measured in so-called *nats* and then Q is measured in *nats per embedding change squared*; if the square-distortion principle of Theorem 1 is accepted, then this is the fundamental unit of steganalysis performance. In order to keep the values in the table readable, they are displayed in *nanonats per embedding change squared*.

Although the results are displayed to three significant figures, it is clear from Fig. 3 that the accuracy of estimation is lower than this, particularly as respects the polynomial fit for $\hat{D}_{\mathrm{KL}}(C_0, C_p)/p^2$. The results show that the estimators $\hat{Q}_1$, $\hat{Q}_2$, and $\hat{Q}_3$ do give similar results (as they ought). As measures of steganalytic performance they are broadly consonant with those found in [12], although they rate the newer detectors less highly. We caution the reader that these numbers should not, at this stage, be taken as a serious evaluation of the steganalysis methods in [4, 7, 12, 14]; the figures are to illustrate the benchmark and verify that the three estimators agree on an approximate value for Q.

Table 1: Estimates of Q for four detectors of LSB Replacement in images. All values in *nanonats per embedding change squared* and displayed to 3 sig. fig.

Steganalysis	Colour bitmaps			Colour JPEGs		
	$\hat{Q}_1$	$\hat{Q}_2$	$\hat{Q}_3$	$\hat{Q}_1$	$\hat{Q}_2$	$\hat{Q}_3$
SPA	16.1	16.5	17.8	28.3	33.5	38.7
SPA/LSM	12.1	13.8	12.3	161	183	174
Triples/LSM	20.7	17.6	14.6	1500	1760	1640
Triples/WLSM	16.1	16.3	17.3	1500	1600	1600

These experiments are all very well, but they involve quantitative steganalysis, in which case the steganalysis response has a particularly simple form: it estimates the payload, so the output should be proportional to p, with some error added; this makes (3) automatic. To widen the experiments, we now turn to a related but much more difficult steganalysis problem, that of detecting LSB Matching in bitmap images (the difference between LSB Matching and LSB Replacement is articulated in [8]). Not only is detection of this type of steganography harder (so we expect to see less information provided by the steganalysis) but also the steganalysis methods are usually not quantitative. We will consider just two from the literature, neither quantitative: the "HCF COM" detector due to Harmsen [6], and a modification of it [8] which calibrates the result by dividing by statistics of a downsampled image.

This time we used a set of 20000 grayscale bitmaps previously stored as JPEG: the larger base mitigates the problems of estimating very small KL-divergence values. We display the analogue of Fig. 3, for the HCF COM detector against LSB Matching embedding changes, in Fig. 4. Much larger numbers of embedding changes must be used because the information is so small: we tested $p = 6000, 8000, \ldots, 160000$ embedding changes, and could not test higher numbers because one should not change more than half of the cover pixels (on average, LSB steganography would not do so).

Again, the Q factor appears to be well-defined, but it is difficult to report its value with accuracy because of the noise in the estimators; at least its order of magnitude is clear. The analogue of Tab. 1 is Tab. 2. Again, the estimators generally agree, although the $\hat{Q}_3$ statistic for the HCF COM detector was out of line: further investigation showed that the plug-in density estimator had produced a bad fit. These results, despite larger potential estimation errors, confirm the general superiority of calibrated HCF COM steganalysis demonstrated in [8].

Because the information provided by the steganalysis is so low, Tab. 2 displays the estimated Q-factor in *piconats per embedding change squared*. One might ask: why, if detection is so difficult that small payloads are essentially undetectable (even micronats of information is negligible), is it reasonable to benchmark steganalysis by the limit as the

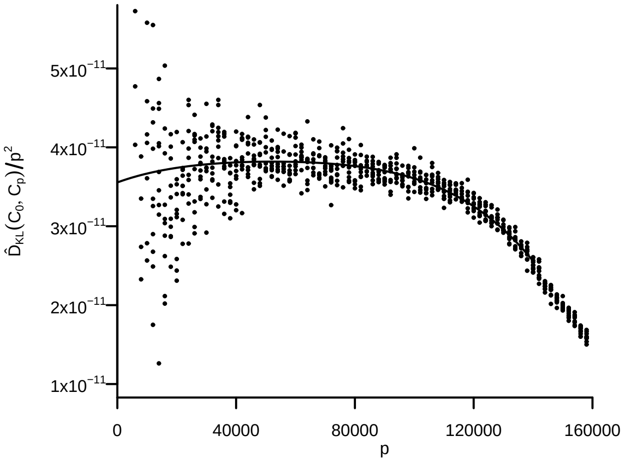

Figure 4: Experimental data from LSB Matching detection in images (HCF COM steganalysis).

payload size tends to zero? The answer is that, although small payloads are indeed practically undetectable in individual objects, if a steganographer continues to embed small payloads in many objects the evidence against them can be accumulated, as in [9], resulting in eventual detection. The benchmark measures how their risk increases with payload size, and number of stego objects.

As an illustration, suppose a steganalysis method producing 500 piconats per embedding change (approximately the performance of the calibrated HCF COM detector for LSB Matching, in our grayscale JPEG covers), and that a steganographer embeds payload requiring 10000 embedding changes per cover object. Assuming that this is low enough to be well-approximated by the limit, this means that the detector is producing $500 \times 10^{-12} \times 10000^2 = 0.05$ nats of information per object. If 100 stego objects are transmitted, the maximum information available to the detector totals 5 nats so, given an efficient method for pooling the evidence (which might be difficult [9]), and supposing that the detector wishes to equalize false positive and negative rates, the steganographer is at risk of detection with at least 0.63% false positive and negative rate. (In individual objects, the value of 0.05 nats implies the false positive and negative rates would be equalized at values at least 42.1%.)

To examine the applicability of the proposed benchmark, we tested a wide range of steganalysis methods, aimed at a number of different embedding mechanisms. In almost all cases the Q-factor appeared to be a well-defined measure (the KL-divergence did appear to be locally square in number of embedding changes) but in a few cases the information itself was too small to measure with any accuracy, and in a very small number of cases the divergence did not appear to be quadratic. In such cases it was observed that the quasi-linear assumption failed. Wider testing of the applicability of the Q-factor benchmark, and the mathematical hypotheses which validate it, is something for future research.

5. CONCLUSIONS

We have presented the rationale for a new steganalysis benchmark, based on principles of information theory. The Q-factor itself is application-independent inasmuch as it neither sets a particular error rate, nor assumes equivalency of false positives and negatives. As long as it is acceptable to consider performance of steganalysis as payload size tends to zero – and we have argued that this is the case for long-run performance – then it boils down a steganalysis method to a single statistic for each type of cover object. (Unavoidably, there are different classes of cover object in which certain steganalysis methods do better than others.) Thus there is a strong case for replacing or at least supplementing the traditional benchmarks with that presented here.

Table 2: Estimates of Q for two detectors of LSB Matching in bitmap images. All values in *piconats per embedding change squared*, displayed to 3 sig. fig.

Steganalysis	Grayscale JPEGs		
	$\hat{Q}_1$	$\hat{Q}_2$	$\hat{Q}_3$
HCF COM	36.9	38.0	152*
Calibrated HCF COM	435	504	491

* *density estimation failed*

We have not yet attempted to find optimal ways to estimate the Q-factor benchmark from empirical data. Three methods are suggested, but surely better choices exist (our methods require the selection, by hand, of informative payload sizes). The first direction for further work is to glean what we can from statistical literature on estimation of Fisher's Information.

We would like to have more certainty that Cramér's regularity conditions hold in practice. It is not possible to determine this using statistical tests on empirical data, but perhaps it would be worthwhile to fit models to steganalysis response and see whether the models meet the conditions.

As it stands, one requires at least a few thousand cover objects in order to make even a rough estimate of the new steganalysis benchmark. The key is the empirical estimation of KL-divergence, and perhaps there are more efficient methods, particularly for the low-divergence situations we encounter here, than [16]. That estimator suffers particularly from the curse of dimensionality, but there may be alternatives which can cope with higher dimensional data; if so, it might become possible additionally to benchmark on $D_{\mathrm{KL}}(B_0, B_p)/p^2$. This would inform feature vector selection. One might also examine the quotient $D_{\mathrm{KL}}(C_0, C_p)/D_{\mathrm{KL}}(A_0, A_p)$ as a measure of steganalysis *efficiency*, as opposed to absolute performance.

Other considerations include investigating whether it is optimal to split empirical data evenly between covers and stego objects (intuition suggests that it might be better to weight the evidence more towards covers), and whether estimator convergence can be accelerated by bootstrapping. The latter technique might also provide approximate confidence intervals for the Q-factors obtained.

Once these issues are addressed, it should be possible to produce a clear and simple procedure for computing the Q-factor benchmark so that authors can properly compare the efficiency of their steganalysis methods.

One problem we have not considered here is nonuniformity of the cover objects, and particularly the effect of cover size; it has been assumed that covers are equal in their capacity and ease of steganalysis, but is now well known that this rarely holds in practice. It would be particularly interesting to know how the size of the cover object affects detectability of payload (in some special cases [11] it is known that a square-root law applies, so that larger objects cannot contain proportionately larger payloads) and this would allow empirical estimation of the Q-factor using unequal covers. More generally, one might combine the benchmark with the framework of [1], to determine how other properties (including noise, saturation, etc) of cover objects affect the Q-factors of different steganalysis methods. But the scale of experimentation (and, particularly, the necessity for thousands of sample cover objects of each type) might make the computational costs prohibitive.

6. ACKNOWLEDGMENTS

The author is a Royal Society University Research Fellow.

7. REFERENCES

[1] R. Böhme and A. Ker. A two-factor error model for quantitative steganalysis. In *Security, Steganography and Watermarking of Multimedia Contents VIII*, volume 6072 of *Proc. SPIE*, pages 59–74, 2006.

[2] C. Cachin. An information-theoretic model for steganography. *Information and Computation*, 192(1):41–56, 2004.

[3] H. Cramér. *Mathematical Methods of Statistics*. Princeton University Press, 1946.

[4] S. Dumitrescu, X. Wu, and Z. Wang. Detection of LSB steganography via sample pair analysis. *IEEE Trans. Signal Processing*, 51(7):1995–2007, 2003.

[5] J. Fridrich and D. Soukal. Matrix embedding for large payloads. *IEEE Trans. Information Forensics and Security*, 1(3):390–395, 2006.

[6] J. Harmsen and W. Pearlman. Steganalysis of additive noise modelable information hiding. In *Security and Watermarking of Multimedia Contents V*, volume 5020 of *Proc. SPIE*, pages 131–142, 2003.

[7] A. Ker. A general framework for the structural steganalysis of LSB replacement. In *Proc. 7th Information Hiding Workshop*, volume 3727 of *Springer LNCS*, pages 296–311, 2005.

[8] A. Ker. Steganalysis of LSB matching in grayscale images. *IEEE Signal Processing Lett.*, 12(6):441–444, 2005.

[9] A. Ker. Batch steganography and pooled steganalysis. In *Proc. 8th Information Hiding Workshop*, volume 4437 of *Springer LNCS*, pages 265–281, 2006.

[10] A. Ker. A capacity result for batch steganography. *IEEE Signal Processing Lett.*, 14(8), 2007.

[11] A. Ker. Derivation of error distribution in least-squares steganalysis. *IEEE Trans. Information Forensics and Security*, 2(2):140–148, 2007.

[12] A. Ker. Optimally weighted least-squares steganalysis. In *Security, Steganography, and Watermarking of Multimedia Contents IX*, volume 6505 of *Proc. SPIE*, 2007.

[13] S. Kullback. *Information Theory and Statistics*. Dover, New York, 1968.

[14] P. Lu, X. Luo, Q. Tang, and L. Shen. An improved sample pairs method for detection of LSB embedding. In *Proc. 6th Information Hiding Workshop*, volume 3200 of *Springer LNCS*, pages 116–127, 2004.

[15] C. Stone, M. Hansen, C. Kooperberg, and Y. Truong. The use of polynomial splines and their tensor products in extended linear modeling (with discussion). *Annals of Statistics*, 25:1371–1470, 1997.

[16] Q. Wang, S. Kulkarni, and S. Verdu. Divergence estimation of continuous distributions based on data-dependent partitions. *IEEE Trans. Information Theory*, 51(9):3064–3074, 2005.

[17] Y. Wang and P. Moulin. Statistical modelling and steganalysis of DFT-based image steganography. In *Security, Steganography and Watermarking of Multimedia Contents VIII*, volume 6072 of *Proc. SPIE*, 2006.