Batch Steganography and Pooled Steganalysis

Andrew D. Ker

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

Abstract. Conventional steganalysis aims to separate cover objects from stego objects, working on each object individually. In this paper we investigate some methods for pooling steganalysis evidence, so as to obtain more reliable detection of steganography in large sets of objects, and the dual problem of hiding information securely when spreading across a batch of covers. The results are rather surprising: in many situations, a steganographer should *not* spread the embedding across all covers, and the secure capacity increases only as the square root of the number of objects. We validate the theoretical results, which are rather general, by testing a particular type of image steganography. The experiments involve tens of millions of repeated steganalytic attacks and show that pooled steganalysis can give very reliable detection of even tiny proportionate payloads.

1 Introduction

The classic definition of *steganography* involves an actor (Steganographer) aiming to communicate with a passive conspirator over an insecure channel, and an eavesdropper (or Warden) monitoring the channel. The Steganographer hides his communication inside some other medium by taking a seemingly-innocent *cover object* and making changes, hopefully imperceptible to the Warden, which convey the secret information to the recipient. The Warden's aim is not to decode the hidden information but merely to deduce its presence. This is *steganalysis*: the creation of hypothesis tests which can distinguish cover objects from so-called stego objects in which a payload has been embedded. Such language assumes that each cover object is treated in isolation by both the embedder and the eavesdropper, and in the literature the focus is almost exclusively on single cover objects (usually individual digital images, but also sometimes audio files, movies, or more unusual digital objects). In this paper we begin to ask about large groups of cover objects, and how the methods for both embedding into, and steganalysis of, individual pieces can be applied to the groups as a whole.

There are two good reasons for doing so. First, we contend that practical applications of steganalysis inevitably will involve multiple objects: the Warden will surely have intercepted more than one communication from the Steganographer, and the Steganographer will surely have access to more than one cover. Second, even given state-of-the-art steganalysis and weak steganography, very high reliability steganalysis (in which false positive rates are as low as, say,

J. Camenisch et al. (Eds.): IH 2006, LNCS 4437, pp. 265–281, 2007.

10^{-5}) is simply not possible with the small amount of evidence obtained from a single cover (except for deeply flawed steganography which leaves a particular signature, or perhaps enormous objects such as entire digital movies).

In this paper we assume that an imperfect method of statistical steganalysis already exists for individual cover objects, and investigate how the set of detection statistics computed over a group can be combined by the Warden into an overall detector for steganography for the whole group. This gives information on the opposite problem, where the Steganographer has to decide how best to spread secret information amongst a batch of covers. The answers to this latter question, at least for some of the pooled detectors suggested here, are rather surprising. There seems to be little literature on this problem: Trivedi [1] has used sequential hypothesis tests to repeat steganalysis, but only in the context of locating a hidden message embedded sequentially within a single image.

In Sect. 2 we formulate more precisely the competing aims of *batch steganography* and *pooled steganalysis*. In this work, which only scratches the surface of what appears to be a complex topic, we allow certain assumptions (which are not implausible) about the steganalysis methods for individual objects which we aim to combine; they are discussed in Sect. 2. In Sect. 3 we suggest three possible pooling strategies for the Warden, analysing them for performance and deriving the Steganographer's best tactic to avoid detection. In Sect. 4 we move away from the abstract nature of the first part of the paper, and focus on Least Significant Bit Replacement in digital images, a well-studied problem; for this embedding method, and a popular detection algorithm, we perform millions of simulations to benchmark the strategies of Sect. 3, confirming the theoretical results. Briefly, we return to our assumptions about steganalysis response – there is a sting in the tail here. Finally, Section 5 suggests avenues for further work.

2 Problem Formulation

The scenario we have in mind, which motivates this paper, is the following. Suppose that a criminal wishes to hide information on his computer, deniably, using steganography. He already has a large number of innocent cover pictures on his hard disk. To be quite sure of hiding his secret information well, he might split it into many small pieces and hide a little in each of a selection of the pictures, believing that this is more secure than the alternative of filling a smaller number of images to maximum capacity.

When the authorities impound his computer, they are faced with a dilemma: how do they know which pictures to examine? Even possessing state-of-the-art steganalysis, they still observe fairly large false positive rates, and so if they test every picture on his computer they will inevitably turn up a lot of positive diagnoses – even if he is not a steganographer at all. They must run their statistical detector on every picture individually, and then find some way to combine the detection statistics into an overall "pooled" steganalysis for the presence of data, possibly spread across all the images.

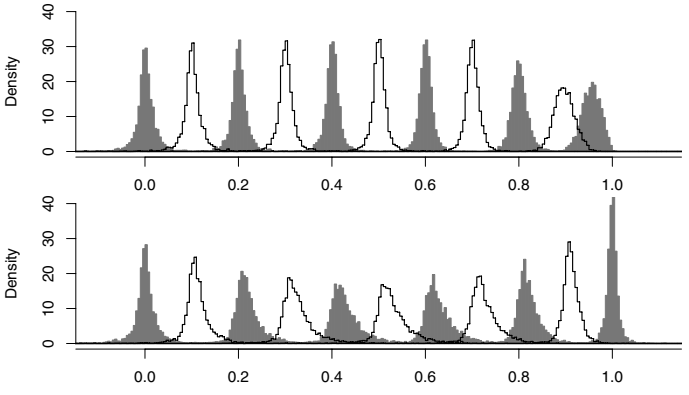

Fig. 1. Histograms of detector response; two detectors for LSB Replacement in digital images, calculated on each of 3000 never-compressed grayscale bitmap images with embedding at $p = 0, 0.1, \ldots, 1$. Above, "Sample Pairs" detector. Below, "WS" detector.

2.1 The Shift Hypothesis and Other Assumptions

In this work we will assume that the Warden already possesses a *quantitative* detector for whatever type of steganography the Steganographer is using, an estimator for the length of hidden message in an individual stego object as a proportion of the maximum. We will call this *the component detector*. We assume that it suffers from random errors due to properties of the cover objects, or the hidden messages. The Warden aims to detect steganography in a set of objects by combining the *component statistics* – the values of the component detector on each object in the set.

We write ψ_p for the density function of the component estimator, when proportion p of the maximum is embedded in a cover object; we expect that it is unbiased i.e. $\int x\psi_p(x)\,\mathrm{d}x \approx p$ if the detector is any good. In this analysis we go further, assuming what we call the *shift hypothesis*. This is that

$$\psi_p(x) = \psi_0(x - p),$$

i.e. the distribution of the detector response only depends on p in the form of a shift, so that the (additive) estimation error is independent of the true value. Our primary reason is to reduce the analyses of pooled steganalysis to tractable problems. Since in this case all ψ_p are determined by p and ψ_0 we write ψ for ψ_0, and Ψ for the corresponding cumulative distribution function.

Before we continue, we ask whether the shift hypothesis is plausible. In Fig. 1 we display histograms for two particular quantitative steganalysis methods for the detection of LSB Replacement in digital images (the Sample Pairs (SPA) detector of [2] and the detector now known as WS from [3]). These histograms are the observed detector response for a set of 3000 images, with the experiments repeated at 10 embedding rates. We see that the shift hypothesis holds approximately for the SPA detector, for embedding rates of less than 0.8, but

there is both a shape change and a negative bias for higher rates[1]. For the WS detector, the shift hypothesis seems less apt, but for medium values of p there is still evidence of a constant distributional shape. We view these histograms as evidence that we should be able to develop detectors which are not far away from satisfying the shift hypothesis.

Finally, although we try to keep the first part of this paper abstract, we may need some assumptions about the functional form of ψ itself. We will certainly want that ψ is symmetric about 0. In a detailed analysis of the response of detectors for LSB Replacement in images [4], we found that the Student t-family provides a good model, up to a scaling constant. The density function is

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

where $\lambda > 0$ is the scale factor and $\nu > 0$ the *degrees of freedom* parameter. An advantage of this family is that it can model a wide range of unimodal symmetric distributions, including the cases of finite and infinite variance, and it seems to have uses in all types of steganalysis. When $\nu = 1$ it is the Cauchy distribution, when $\nu \leq 2$ it has infinite variance, but as $\nu \to \infty$ it tends to the Gaussian.

When we need to make assumptions about the shape of ψ, therefore, we will suppose that it is a t-density. We have found that quantitative LSB detectors are often well-modelled by t-distributions with approximately 2 degrees of freedom (numbers both above and below 2 are observed, depending on the type of detector and also the type of cover, indicating that finite variance is a possibility but cannot be guaranteed) and a scale factor of the order of 0.01. These will be useful figures for the generation of some synthetic data in Subsect. 3.3.

2.2 Batch Steganography and Pooled Steganalysis

We now pose precisely the problem of *batch steganography*. Given a number of cover objects N, the Steganographer hides data in rN of them, using proportionate capacity p of each, and leaves the other covers alone. We assume that a) the number of cover objects is fixed, b) all cover objects have the same capacity, c) the Steganographer has no control over the objects themselves, and chooses which to embed in at random. We also assume that the Steganographer wants to embed a fixed total amount of secret data, BNC, where C is the capacity of each cover object and $B < 1$ is the proportionate *bandwidth*. He therefore must ensure that $rp = B$. The conditions $r \leq 1$ (he cannot embed in more objects than are present) and $p \leq 1$ (each object has a maximal capacity) give the dual conditions $p \geq B$ and $r \geq B$. (We assume that $B < 1/N$ so that r as low as B is a meaningful option.) Within this range, he can vary r or p and should do so to try to minimize the chance of overall detection.

[1] In fact we believe that the SPA detector can be corrected to remove the negative bias at high embedding rates, but that is not our current topic of study.

The Warden's task is *pooled steganalysis*: given the N objects treated by the Steganographer, the Warden's aim is to detect whether any steganography at all has been performed. That is, to perform the hypothesis test

$$\begin{aligned} H_0 &: r = 0 \\ H_1 &: p, r > 0 \end{aligned} \tag{1}$$

with best reliability. We *do not* assume that the Warden wants to estimate the values of r and p, or B, or tries to determine which of the objects do contain steganography, although these are certainly secondary aims for future study.

We assume that the Warden applies a component steganalysis method satisfying the assumptions of the previous subsection. Then its density function, on a randomly selected object output by the Steganographer, is

$$f(x) = (1 - r)\psi(x) + r\psi(x - p).$$

This is a *mixture model*, of a simple kind. Finite mixtures are quite well-studied, although there is much more literature on Gaussian mixtures than mixtures of longer-tailed distributions we expect from steganalysis. For a good survey, see [5].

Finally, we emphasise that we are assuming that the number of cover objects N is fixed from the start, and known to both the Steganographer and Warden. In practice, it seems likely that the Steganographer will gain access to new covers over time, and that in some cases (such as network monitoring) the Warden will obtain more evidence over time. This suggests the study of *sequential* embedding and tests. We reserve this topic for future work; the theory of sequential hypothesis tests is rather different from that of standard tests of finite samples, and we believe that the practice of sequential steganography and steganalysis may be rather different from the fixed-size batch problems studied here.

2.3 Performance Metric

We choose the following as a measure of performance, for the Warden's task of pooled steganalysis. Fix a detection threshold so that the overall false negative rate (the probability of type-II error for (1)) is 50%; then measure the false positive rate (the probability of type-I error), which we will denote p_f.

This is a rather unusual measure, but suitable for two reasons. Firstly, it allows for tractable analysis (many other measures of performance do not). Second, we have found it to be a good summary of the performance of detectors in general. Although it might be preferable to plot a full *Receiver Operating Characteristics* (ROC) curve, this results in too much information to display concisely. And in almost all cases the ROC curve takes the form of a sudden jump from a very low detection rate of almost 0% to a high detection rate well above 50%, as the false positives increase past a certain point. Therefore the most important information is to be found in the location of that jump, which is well-measured by the false positive rate when the false negative rate is 50%.

While the false negative rate is a measure of a Steganographer's chance to evade detection, the false positive rate shows *how certain* the Warden is that they have caught the right person: fundamental from the latter's point of view.

3 Possible Strategies for Warden

We examine three strategies which the Warden might use to detect batch steganography, given a component detector satisfying the assumptions of Subsect. 2.1. Of course there are other possible pooling strategies, but we have included a range of methods: simple nonparametric tests for median, the average component statistic, and a more sophisticated method based on a generalized maximum likelihood ratio test.

At the end of the section we summarise by asking the following questions of each detection strategy. Given a bandwidth B, how can the Steganographer best avoid detection, by trading p against r? In the Steganographer's best case, how does the false positive rate for the Warden depend on N and B (assuming a threshold set for 50% false negatives)? And what must be the relationship between N and B, if the error rate is to be held constant?

3.1 Count Positive Observations (Sign Test)

The first method for pooled steganalysis is the simplest: the Warden should simply count the number of positive values produced by the component estimator, which we will denote by the random variable $\#P$. If there is steganography in some of the objects, we expect that $\#P > \frac{1}{2}N$. This is simply the traditional *sign test* for whether the median of a distribution, from which we have a sample, is greater than zero. It has the advantage of being nonparametric: its distribution under the null hypothesis does not depend on ψ, being $\#P \sim \mathrm{Bi}(N, \frac{1}{2})$.

Under the alternative hypothesis, $\#P \sim \mathrm{Bi}\big((1-r)N, \frac{1}{2}\big) + \mathrm{Bi}\big(rN, \Psi(p)\big)$[2] of which the median is $(1-r)N\frac{1}{2} + rN\Psi(p) = \frac{1}{2}N + rN(\Psi(p) - \frac{1}{2})$. Making the Gaussian approximation to the binomial distribution (valid for even moderately large values of N) the probability of false positive for the Warden when the false negative rate is 50% and the Steganographer's bandwidth is B, is therefore

$$p_f = 1 - \Phi\left(2\sqrt{N}B\big(\tfrac{\Psi(p)-\frac{1}{2}}{p}\big)\right). \tag{2}$$

where Φ is the normal distribution function. Let us consider the Steganographer's best strategy. He wants to maximize the false positive probability, and therefore to minimize $\frac{\Psi(p)-\frac{1}{2}}{p}$. Provided that ψ is a nonincreasing function on $[0, \infty)$, this is achieved by maximizing p. Therefore the Steganographer will choose $p = 1$ and $r = B$, hiding maximal amounts of data in as few cover objects as possible. The makes intuitive sense, because the sign test is no more sensitive to large positive detection values than small positive values.

In Sect. 4 we will additionally test the more powerful nonparametric test known as the *Wilcoxon signed rank test*. Here the component values are ranked by absolute value, and the test statistic used is the sum of the ranks of the

[2] This sum of distributions notation indicates the *independent* sum of random variables.

positive observations. For reasons of space we do not attempt to analyse the behaviour of this statistic. One might expect superior performance to the sign test, given parallel results in standard hypothesis testing, but we shall see that the improvement is not very substantial in this application.

3.2 Average Detection Statistic

The main weakness of the sign test is that it ignores all information except the sign of each observation. An alternative method without this drawback, and seemingly a simple one, is to take the component statistics and compute their mean: $\bar{X} = \frac{1}{N}\sum X_i$ where X_i is the component detector response for object i of the batch. It is immediate that median$(\bar{X}) \approx \mathbf{E}[\bar{X}] = rp$ if the expectation exists at all and, given the shift hypothesis, the distributional shape of $\bar{X}$ does not depend on r or p. Therefore, the Steganographer has no reason to select any particular value of r and p, as long as they multiply to his bandwidth constraint B.

The complexity here is in computing the distribution of $\bar{X}$ under the null hypothesis. If the distribution density ψ has thin enough tails, the variance of the X_i is finite and the Central Limit Theorem applies. However, we have already noted that in practice the variance of the detector response may be infinite. Thankfully, there is a generalized from of the Central Limit Theorem, which is presented in detail in [6], from which we extract the following result:

Lemma 1. *Suppose that all X_i are independent and identically distributed, and the tail index of X_i is $\nu > 1$, i.e. $P(|X_i| > x) \sim cx^{-\nu}$ as $x \to \infty$, for some constant c. Then $\mathbf{E}[X_i]$ exists and*

(i) if $1 < \nu < 2$ then $\bar{X} \xrightarrow{\mathrm{d}} \mathbf{E}[X_i] + kN^{\frac{1}{\nu}-1}Z$ where Z has a standardized Symmetric Stable distribution with index of stability ν, and k is a constant (depending on the dispersion of the X_i, and ν).

(ii) if $\nu > 2$ then $\bar{X} \xrightarrow{\mathrm{d}} \mathbf{E}[X_i] + kN^{-\frac{1}{2}}Z$ where Z has a standard Normal distribution, and k is a constant (depending on the variance of the X_i).

There is a more complex case when ν is exactly 2, but we will not concern ourselves with it because it is unlikely that the tail index will be precisely 2. Note that the t-distribution with ν degrees of freedom has tail index ν.

The median value of $\bar{X}$ under the null hypothesis is approximately B. Therefore

$$p_f = \begin{cases} 1 - \Phi(N^{\frac{1}{2}}B/k), & \text{if } \nu > 2 \\ 1 - F_\nu(N^{1-\frac{1}{\nu}}B/k), & \text{if } 1 < \nu < 2. \end{cases} \tag{3}$$

where F_ν is the distribution function for the Symmetric Stable distribution with stability index ν and k is some constant. As long as $\nu > 1$, evidence does accumulate, but the rate at which this happens, as N increases, depends critically on whether ν is greater, or less, than 2.

3.3 Generalized Likelihood Ratio Test

A general and powerful tool for hypothesis testing is the likelihood ratio test. In the case of two simple hypotheses this takes the form of the quotient of the likelihood of observations given the null and alternative hypothesis (and, according to the Neyman-Pearson Lemma, is the optimal test); when one or both of the hypotheses is composite we use the generalized likelihood ratio test instead. The statistic is computed as $\ell = \log\left(\frac{L(X_1,\ldots,X_n;\hat{\theta}_1)}{L(X_1,\ldots,X_n;\hat{\theta}_0)}\right)$ where L is the likelihood function, $\hat{\theta}_0$ denotes the maximum likelihood estimator (MLE) for unknown parameter(s) θ when constrained by the null hypothesis, and $\hat{\theta}_1$ the MLEs when constrained by the alternative hypothesis. The test rejects the null hypothesis for large values of ℓ.

We apply the test to the problem of pooled steganalysis as follows. Suppose that the component response ψ is known. Then we can compute

$$\ell = \log\left(\frac{L(X_1,\ldots,X_N;\hat{r},\hat{p})}{L(X_1,\ldots,X_N;r{=}0,p{=}0)}\right) \tag{4}$$

where L is the likelihood function for the mixture pdf $f(x) = (1-r)\psi(x) + r\psi(x-p)$ and $\hat{r}$ and $\hat{p}$ are MLEs for r and p given the observations.

The test is powerful, but there are two pitfalls here. The first is that it is not so easy to compute MLEs for r and p. In the case of mixture distributions when ψ is nontrivial the maximization problem admits no closed form solutions, and the likelihood function contains more than one local maximum. Therefore we are forced to use numerical optimization techniques, and it is quite possible to find the wrong local maximum and hence to mis-estimate r and p. This is true whether one uses standard numerical methods to maximize $L(X_1,\ldots,X_N;r,p)$, or whether specialised techniques such as *Expectation Maximization* are applied. See [7] for some discussion of this problem. Our solution is the computationally-expensive one of using a coarse grid search to find good starting values for r and p, and then using a standard iterative optimization procedure [8] to hone the answer (while accepting that this leaves the possibility of convergence to the wrong root, but with low probability).

The second pitfall is more subtle. It is commonly stated that the log-likelihood ratio statistic ℓ has a particular asymptotic null distribution: up to a scaling factor, the χ^2 distribution with $\dim(\theta_1) - \dim(\theta_0)$ degrees of freedom. This is Wilks' Theorem [9], but we have omitted a vital hypothesis – we require some regularity conditions of the likelihood function *and that the null hypothesis be contained within the interior of the alternative hypothesis*. In the case of the hypothesis test (1) this is not so: the null hypothesis is on the boundary of the alternative.

Nonetheless, there is some recent work which generalizes Wilks' Theorem and shows that the conclusions often remain true, perhaps with modified scaling constants or degrees of freedom parameter, even when some of the hypotheses are violated (see e.g. [10]). Alternatively, there are scoring methods which modify the likelihood ratio statistic into one with a known null distributions [11]. Rather

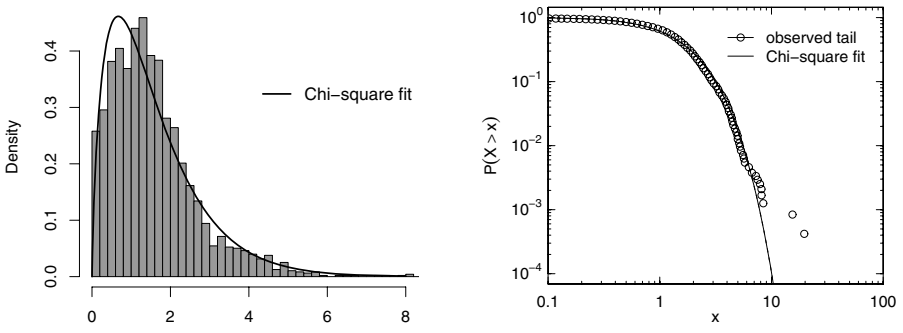

Fig. 2. Histogram and logarithmic tail plot of observed log-likelihood ratio statistic. 2500 samples of $N = 500$ observations were generated from a t-distribution ($\nu = 2$, $\lambda = 0.01$) and fitted to a scaled χ^2 distribution.

than be diverted into a discussion of this issue, we generated 2500 samples of $N = 500$ synthetic data points from a t-distribution (using $\nu = 2$ and $\lambda = 0.01$ so as to be a good model for steganalysis statistics) and compared the distribution of ℓ, from (4), with the χ^2 family. We found that a scale factor of 0.45 and degrees of freedom 3.45 were the best fit (this df parameter is higher than predicted by Wilks' Theorem), which is displayed in Fig. 2. There is a good match and it appears that a modified form of Wilks' Theorem is still approximately valid.

We now examine the likelihood ratio under the alternative hypothesis.

Lemma 2. *Let ψ be any probability density, assumed known, such that*

a) ψ is symmetric about 0,

b) $\frac{\psi(x+p)-\psi(x)}{p\psi(x)}$ is bounded for $x \in \mathbb{R}$ and $p \in [0,1]$,

c) $\int \frac{\psi(x+p)^2-\psi(x)^2}{p^2\psi(x)}\,\mathrm{d}x$ is an increasing function of p, for $p \in [0,1]$,

along with the usual regularity conditions sufficient for MLEs of the mixture parameters r and p to be consistent. Then, in the limit as $N \to \infty$, and for sufficiently small B, the Steganographer's best strategy to evade detection by the generalized likelihood ratio test is to take $p = B$ ($r = 1$), in which case the expectation of the generalized log-likelihood ratio statistic is asymptotically

$$\mathbf{E}[\ell] \sim \frac{NB^2}{2} \int \frac{\psi'(x)^2}{\psi(x)} + \psi''(x)\,\mathrm{d}x \;+\; NO(B^3) \tag{5}$$

The proof may be found in the Appendix. Conditions a) and b) are quite easy to check, but c) is more difficult to establish. However they all seem to hold for common density functions ψ, including the t-distribution pdf with zero location parameter and $\nu > 1$[3].

[3] To match exactly our performance metric we should be considering the *median* of the statistic under the alternative hypothesis, rather than the *mean* as here. However the former is much harder to examine, and commonly in generalized likelihood ratio tests the mean and median differ only by a small constant.

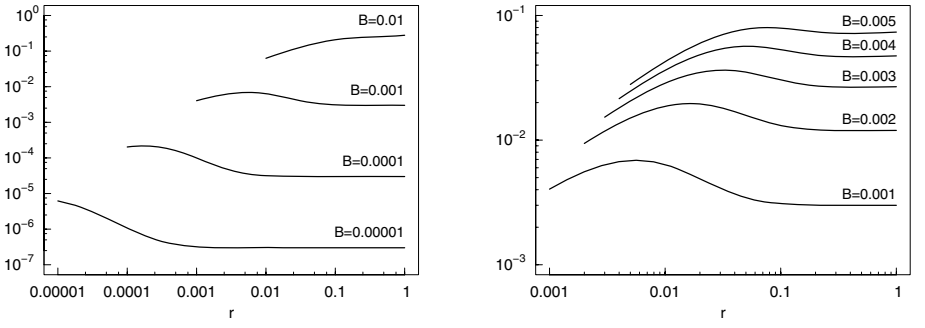

Fig. 3. Expected value of log-likelihood ratio statistic (*y-axis*) according to (7) without the scaling factor N, plotted against r (*x-axis*). Some different values of B are displayed.

Finally, in the case of a t-distribution with ν degrees of freedom and scale factor λ, we have

$$\frac{NB^2}{2}\int \frac{\psi'(x)^2}{\psi(x)} + \psi''(x)\,\mathrm{d}x = \frac{NB^2}{2\lambda^2}\frac{\nu+1}{\nu+3}. \tag{6}$$

This indicates that the parameter ν is not vitally important to the accuracy of the likelihood ratio detector (as a function of ν, for ν around 2, (6) varies rather slowly and $\nu = 2$ is not a critical value).

Now the statement of Lemma 2 is not ideal in that it only informs the Steganographer what to do in the limiting case of small B. On further examination, it seems that the best strategy is exactly the reverse – $p = 1$ and $r = B$ – for B above a certain point and then switches to the proven optimal method for B below. Using the pdf of a t-distribution with 2 degrees of freedom and scale factor 0.01 we plot numerically-computed values of (7) (which, in the Appendix, is shown to be the asymptotic expectation of ℓ even when B is not small), without the scaling factor N, in Fig. 3. On the left we see how the magnitude of B influences the choice of r: for small B there is a definite disadvantage to the Steganographer in using small values of r, because detection becomes easier. On the right we examine more closely the values of B around which the best strategy switches from $r = 1$ to $p = 1$. We observe that there is no internal minimum on any of the displayed curves, so that the best strategy is switches directly from one extreme (spread the payload as thinly as possible for small B) to the other (concentrate the payload in as few covers as possible for large B).

Finally, we can also apply a generalized likelihood ratio test when the distribution ψ has unknown parameters, forming MLEs for the unknowns in the usual way. This may be a useful application, because the distribution parameters of the steganalysis error distribution may well depend on the source of covers, of which the Warden might be unaware. But the theory of Lemma 2 does not apply directly in this case and we will not consider it further in this paper.

3.4 Summary of Warden's Strategies

For the sign test, and the likelihood ratio test for large bandwidths, we see that the Steganographer's best strategy is to take p=1 and r=B, concentrating all the payload in as few covers as possible. We find this result rather counterintuitive, as there is a natural inclination to spread hidden data thinly.

The total amount of data hidden by the Steganographer is proportional to BN. Consider equations (2), (3), and (5). To fix the risk of detection, we must have $B \propto N^{-\frac{1}{2}}$ (sign or likelihood ratio test, or average if $\nu > 2$) or $B \propto N^{-1+\frac{1}{\nu}}$ (average if $\nu < 2$). Therefore the "capacity" (of the undetectable kind) for the Steganographer appears to grow as $N^{\frac{1}{2}}$ or $N^{\frac{1}{\nu}}$ – not proportionately to N.

We can compare the false positive rates of the tests by looking at the tail probabilities for the Normal, Symmetric Stable, and χ^2 distributions. It can be shown that, as long as $\nu > 2$, the false positive rate is of the form $p_f \sim a(NB^2)^b \exp(-cNB^2)$ for each of the three tests, where a, b and c are constants. The parameter c is most important to the shape here with larger c meaning that evidence is gained more quickly as N increases. For the sign test, c is no larger than 2. For the average and likelihood ratio tests, it is generally much larger, being inversely proportional to the square of the dispersion of ψ.

Although the results for the other tests are not changed when $\nu < 2$, in this case the discrimination of the average statistic increases only as a power law, with $p_f \sim aN^{1-\nu}B^{-\nu}$. Ensuring that steganalysis detectors have finite variance appears to be important. However we emphasise that there is no sudden discontinuity at $\nu = 2$. When ν is slightly above 2, the standard Central Limit Theorem applies but convergence to the asymptotic distribution is extremely slow. The key for the steganalyst is to keep the tails of the component steganalysis estimator as light as possible.

4 Case Study: LSB Replacement in Images

We now move away from the abstract setting and select a particular type of cover object, steganographic embedding, and component steganalysis. We choose LSB Replacement in bitmap grayscale images, because this problem is extremely well-studied and also because it is a poor enough form of steganography for fairly good component steganalysis methods to be known. There are many to choose from, and we have selected the one known as *Sample Pairs Analysis* (SPA) [2]. It was not selected because it is the best – there are now newer methods such as [12] which are more sensitive – but because it shares the advantages of computational simplicity (highly desirable given the scale of our experiments) with an approximate validation of the shift hypothesis (see Fig. 1).

We aim to benchmark the pooled steganalysis methods of Sect. 3. Because we want to consider large samples (up to $N = 4000$) we need a particularly large corpus of cover images, a random selection of which is presented to the Steganographer for each trial. Also, to keep as closely as possible to the shift hypothesis and the other assumptions of Sect. 2, we want all the images to be

the same size and of similar "character" (each image should have macroscopic characteristics which indicate similar sensitivity to steganography). We used a set of 14000 images, selected for size and image quality out of 20000 on a stock photo CD[4]. The images selected were all 640×416 pixels and had been stored as colour JPEG images (at quality factor 58), later converted to grayscale. This is probably representative of the type of images which a Steganographer can gain access to in large quantity: big, never-compressed images are more scarce.

The distribution of the SPA estimator, when no data is hidden in these images, is well modelled by a t-distribution with $\nu = 1.61$ and $\lambda = 0.00956$ (and we will use these values to compute approximate likelihoods for benchmarking the pooling method of Subsect. 3.3). These values indicate that, even with maximal embedding, for single images there will be a false positive rate of approximately 2×10^{-4} when the false negative rate is 50%. For more reliable detection, or for smaller bandwidths, the Warden has no option but to gather multiple images as evidence and apply pooled steganalysis.

4.1 Empirical Performance

The majority of our experiments were performed with the following parameters: $B \in \{0.01, 0.003, 0.001\}$ and three out of $N \in \{10, 100, 1000, 4000\}$, depending on B. This covers the interesting range of possibilities, as detection moves from easy to very difficult.

For each N we took 10000 samples of N detection statistics when $p = 0$ and then fitted the data according to the theoretical models[5]: Normal distributions for the sign and signed rank statistics, Symmetric Stable distributions for the average statistic, and scaled χ^2 distributions for the likelihood ratio statistic. The fits were good, and the fitted parameters were in the expected range.

For each B, up to 11 different pairs of r and p were chosen, subject to $1 \geq r \geq 1/N, B$ and $rp = B$. Repeating 1000 times, N covers were picked at random (for technical reasons it is necessary to sample *with replacement*) and random data embedded at rate p in Nr of these. Values of the SPA estimator were computed for all N objects, and combined using each of the pooling methods described in Sect. 3. The false positive rate, at which detection rates are 50%, was computed for each statistic at each value of r.

In total, therefore, the results shown here come from tens of millions of steganography and steganalysis computations. Figure 4 shows the results. We see that for small N or B none of the methods gives reliable detection. Once detection becomes possible, observe that the sign and signed rank tests very effectively punish the Steganographer if he uses a high value of r, but otherwise they are poor performers. The averaging method is generally superior for the values of

[4] "20,000 photos", published in 2002 by Focus Multimedia Ltd, Staffordshire, UK. `http://www.focusmm.co.uk`.

[5] Although we would prefer to avoid fitting a distribution which might not be exact, we can only draw conclusions about false positive rates less than $1/S$, where S is the number of repeated simulations, by doing so.

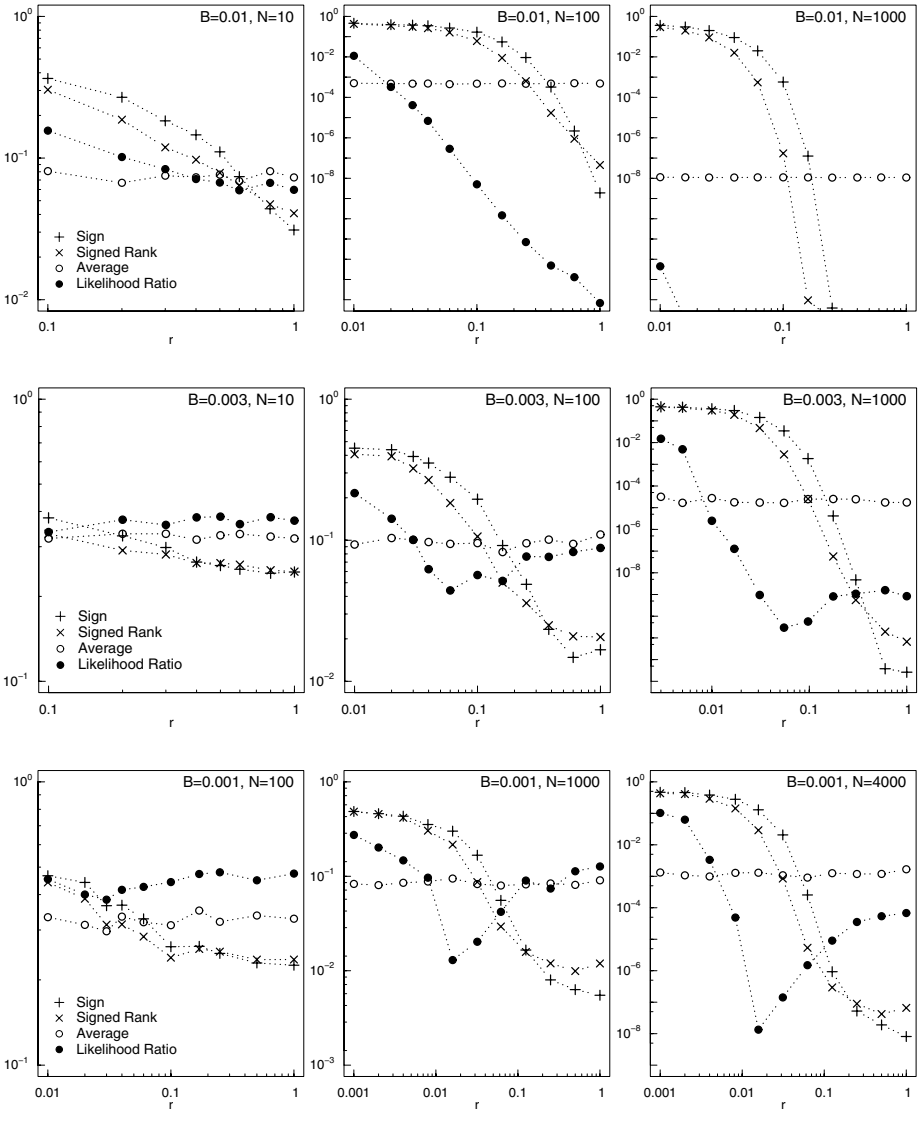

Fig. 4. The false positive detection rate (*y-axis*) when the false negative rate is 50%, for varying r (*x-axis*) with $p = B/r$, of pooled steganalysis using the SPA detector on grayscale images. Four pooling methods are evaluated, the null distributions being fitted to the theoretically-predicted shapes based on 10000 simulations, and the medians of the alternative distributions computed from 1000 simulations for each r. From top to bottom, the bandwidths of the embedding are $B = 0.01$, $B = 0.003$, $B = 0.001$. Increasing values of N, from left to right. The pooled detectors have no discrimination power for $(B, N) = (0.003, 10)$ or $(0.001, 100)$ and poor reliability for $(0.01, 10)$, $(0.003, 100)$, $(0.001, 1000)$, but then reliability increases rapidly with N.

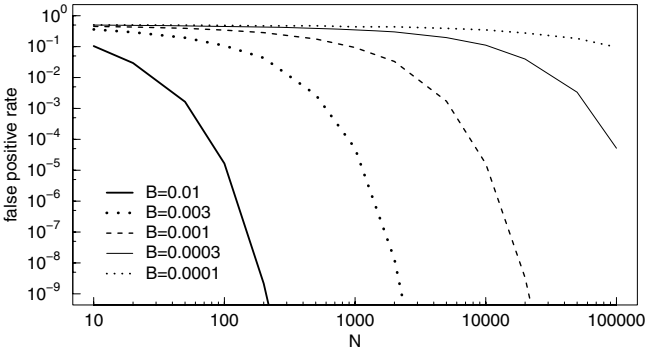

Fig. 5. False positive rate (*y-axis*) when false negative rate is 50%, for varying sample sizes N (*x-axis*), observed in 1000 experiments when data is embedded at $p = 1$, $r = B$

N and B tested here, but given sufficiently large values of N it looks possible that the likelihood ratio test will become the better test (some other results, not displayed, emphatically confirm this). Also, the average method cannot punish the Steganographer for using a suboptimal embedding strategy.

We observe that concentrating the steganography is the best strategy against the likelihood ratio test, but that for the smallest value of B tested a second peak is forming near $r = 1$. If we had been able to test lower B we would expect to see the results of Lemma 2, with $r = 1$ (spreading the payload thinly across all covers) the least detectable embedding method. But we cannot test lower B without substantially increasing N (else there is no detection power), and we cannot do this because of the size of the corpus of images from which we sample.

Finally, focussing only on the averaging pooling method, we tested a larger range of B and N in Fig. 5. Because performance of the average detector does not depend on r, we can pick only a single value of r and plot the false positive rate as N varies. This indicates that, once N passes a point where detection takes place with reasonable reliability, false positive rates drop extremely fast with a few extra observations.

4.2 Assumptions Revisited

We also tested pooled steganalysis performance with alternative types of cover image, and noticed a possible problem. Of the assumptions of Sect. 2, the simplest and apparently mildest is that the steganalysis detector response is symmetric about zero, hence unbiased. Indeed, for the 14000 cover images we tested in Sect. 4, this is fairly accurate: the observed bias was 0.00018. However in certain other types of cover, for example colour never-compressed images, it is common to observe a systematic and significant bias of as much as 0.005 or higher. Whether an artifact of the SPA method or the covers themselves is irrelevant; the effect is to destroy the reliability of the pooled detectors, flooding them with false positives for any large value of N. This must be so, because a

systematic bias of b is indistinguishable from a batch Steganographer's behaviour with $p = b$ and $r = 1$.

If we train the detector first on covers of the right type (effectively removing the bias) then the problem goes away. The difficulty is that the bias varies, depending on the source of covers (and the problem is not limited to the SPA component detector). We see two options for dealing with this difficult issue. On one hand, we could set a maximum bias which we believe is "possible" in natural images, and alter detection thresholds for all the pooled detectors by, for example, subtracting this value before diagnosis. Of course, this reduces their performance and makes it impossible to detect small bandwidths. The alternative is to modify the likelihood ratio test to include a location parameter in both null and alternative hypotheses (possibly constrained). This will detect a Steganographer using small r, as distinct clusters are observed, but not $r = 1$.

In previous work we, and other authors, have not considered a small detector bias (say as small as 0.001) to be significant. For pooled steganalysis, where small bandwidths are in principle detectable given large enough N, removing the bias becomes the best way to improve pooled performance.

5 Conclusions and Directions for Further Research

We have motivated and defined the problems of batch steganography and pooled steganalysis, presenting a menu of techniques for the latter and examining the implications for the former. The pooled steganalysis methods have been benchmarked for a particular type of steganography, with results in line with the theoretical predictions.

The conclusion, that in many cases the Steganographer should cluster the embedding in a small number of cover objects, seems rather counterintuitive. We emphasise that it cannot apply to *every* type of pooling algorithm. For example, the number of observations greater than 1 is quite sensitive to embedding at $p = 1$, although worthless for other batch parameters. There are many other possible pooling algorithms, and some advanced techniques based on mixture modelling, which should be the first priority for further study. In this work we have deliberately avoided methods which are parametric for the Warden – for example the pooling method of counting observations greater than some threshold T – because this leads to a game theoretical setup which can be intractable.

We have only modelled the simplest type of batch steganography. In future work we should consider allowing the Steganographer to vary the amount of data embedded in each object (this results in a larger mixture), and to deal with objects of varying capacity. We must also consider other pooling strategies; we did test, briefly, a pooled detector which simply stitches together N images and applies one steganalysis method to the entire montage, but omitted it from this paper because the performance is similar but a little inferior to the average method. More information on the individual steganalysis response, as it depends on object size and other object parameters, will be needed here. Finally, we would like to prove a general result on how steganographic capacity increases with N.

Given the assumptions we made in this paper, including all covers being the same size, there is an additional possibility we have not explored. If we may assume that all, or most, of the steganography is performed using *the same secret key*, this may imply that the same pixels in each cover would be used for steganography (depending on the embedding method). If so, there could be an amplification of the statistical traces, which we might exploit.

More speculatively, consider whether these results give information about adaptivity and steganography in individual images. If it is optimal to cluster data amongst a set of objects, it is not implausible to suggest also clustering stego noise within each single object, although the analogy is not perfect.

Acknowledgements

The author is a Royal Society University Research Fellow.

References

1. Trivedi, S., Chandramouli, R.: Active steganalysis of sequential steganography. In: Delp III, E.J., Wong, P.W. (eds.): Security and Watermarking of Multimedia Contents V. vol. 5020 of Proc. SPIE, pp. 123–130 (2003)
2. Dumitrescu, S., Wu, X., Wang, Z.: Detection of LSB steganography via sample pair analysis. In: Petitcolas, F.A.P. (ed.) IH 2002. LNCS, vol. 2578, pp. 355–372. Springer, Heidelberg (2003)
3. Fridrich, J., Goljan, M.: On estimation of secret message length in LSB steganography in spatial domain. In: Delp III, E.J., Wong, P.W. (eds.): Security, Steganography, and Watermarking of Multimedia Contents VI. vol. 5306 of Proc. SPIE, pp. 23–34 (2004)
4. Böhme, R., Ker, A.: A two-factor error model for quantitative steganalysis. In: Delp III, E.J., Wong, P.W. (eds.): Security, Steganography and Watermarking of Multimedia Contents VIII. vol. 6072 of Proc. SPIE, pp. 59–74 (2006)
5. Everitt, B., Hand, D.: Finite Mixture Distributions. Chapman and Hall, Sydney, Australia (1981)
6. Gnedenko, B., Kolmogorov, A.: Limit Distributions for Sums of Independent Random Variables. Addison-Wesley, London, UK (1954)
7. Marin, J., Mengersen, K., Robert, C.: Bayesian modelling and inference on mixtures of distributions. In: Dey, D., Rao, C. (eds.) Handbook of Statistics, vol. 25, Elsevier, Amsterdam (2006)
8. Byrd, R., Lu, P., Nocedal, J., Zhu, C.: A limited memory algorithm for bound constrained optimization. SIAM J. Scientific Computing 16, 1190–1208 (1995)
9. Wilks, S.: The large-sample distribution of the likelihood ratio for testing composite hypotheses. Ann. Mathematical Statistics 9, 60–62 (1938)
10. Fan, J., Zhang, C., Zhang, J.: Generalized likelihood ratio statistics and Wilks phenomenon. Annals of Statistics 29, 153–193 (2001)
11. Pilla, R., Loader, C., Taylor, C.: New technique for finding needles in haystacks: Geometric approach to distinguishing between a new source and random fluctuations. Phys. Review Letters, vol. 95 (2005)
12. Ker, A.: A general framework for the structural steganalysis of LSB replacement. In: Barni, M., Herrera-Joancomartí, J., Katzenbeisser, S., Pérez-González, F. (eds.) IH 2005. LNCS, vol. 3727, pp. 296–311. Springer, Heidelberg (2005)

Appendix: Proof of Lemma 2

Let L be the likelihood function, and $\hat{r}$ and $\hat{p}$ the MLEs for r and p given the observations. The log-likelihood ratio statistic is

$$\ell = \log\left(\frac{L(X_1,\ldots,X_n;\hat{r},\hat{p})}{L(X_1,\ldots,X_n;0,0)}\right) = \sum \log\left(\left(\frac{(1-\hat{r})\psi(X_i)+\hat{r}\psi(X_i-p)}{\psi(X_i)}\right)\right)$$
$$= \sum \log\left(1+\hat{r}\left(\frac{\psi(X_i-\hat{p})-\psi(X_i)}{\psi(X_i)}\right)\right)$$

Let Y_i be independent random variables with pdf ψ. We know that $(1-r)N$ of the X_i are distributed as Y_i, and the others as Y_i+p, and all are independent. We may also assume, as N grows large, that $\hat{r}\sim r$ and $\hat{p}\sim p$ (by consistency of the MLEs). Therefore we have

$$\mathbf{E}[\ell] \sim \mathbf{E}\Bigg[\sum_{\text{terms}}^{rN} \log\Big(1+r\big(\tfrac{\psi(Y_i)-\psi(Y_i+p)}{\psi(Y_i+p)}\big)\Big) + \sum_{\text{terms}}^{(1-r)N} \log\Big(1+r\big(\tfrac{\psi(Y_i-p)-\psi(Y_i)}{\psi(Y_i)}\big)\Big)\Bigg]$$
$$= N\int r\log\Big(1+r\big(\tfrac{\psi(y)-\psi(y+p)}{\psi(y+p)}\big)\Big)\psi(y)\,\mathrm{d}y$$
$$+ N\int (1-r)\log\Big(1+r\big(\tfrac{\psi(y-p)-\psi(y)}{\psi(y)}\big)\Big)\psi(y)\,\mathrm{d}y$$
$$= N\int \left(\log\Big(1+\tfrac{B}{p}\big(\tfrac{\psi(x+p)}{\psi(x)}-1\big)\Big)\right)\Big(1+\tfrac{B}{p}\big(\tfrac{\psi(x+p)}{\psi(x)}-1\big)\Big)\psi(x)\,\mathrm{d}x. \qquad (7)$$

(In the last step substituting $x=-y-p$ in the first integral, and $x=-y$ in the second, making use of condition a), that $\psi(x)=\psi(-x)$, and writing $r=\frac{B}{p}$.) Condition b) allows us to use the Taylor expansion $\log(1+z)\sim z-\frac{z^2}{2}+O(z^3)$ in the knowledge that, here, z is of order B: given sufficiently small B and neglecting terms of order B^3, we have

$$\mathbf{E}[\ell] \sim N\int \Big(\tfrac{B}{p}\big(\tfrac{\psi(x+p)}{\psi(x)}-1\big)\Big)\Big(1-\tfrac{1}{2}\tfrac{B}{p}\big(\tfrac{\psi(x+p)}{\psi(x)}-1\big)\Big)\Big(1+\tfrac{B}{p}\big(\tfrac{\psi(x+p)}{\psi(x)}-1\big)\Big)\psi(x)\,\mathrm{d}x.$$

Parts of this integral can be removed, using the fact that $\int\psi(y)\,\mathrm{d}y=1$, leaving

$$\mathbf{E}[\ell] \sim \frac{NB^2}{2}\int \frac{\psi(x+p)^2-\psi(x)^2}{p^2\psi(x)}\,\mathrm{d}x.$$

The Steganographer wants to minimize $\mathbf{E}[\ell]$, to reduce his chance of detection. By condition c) this integral is an increasing function of p, so the Steganographer should minimize p, taking $r=1$ and $p=B$. Given that B is small, we may now use a Taylor expansion for ψ. We need the first three terms:

$$\mathbf{E}[\ell] \sim \frac{N}{2}\int \frac{\big(\psi(x)+B\psi'(x)+\frac{B^2}{2}\psi''(x)+O(B^3)\big)^2-\psi(x)^2}{\psi(x)}\,\mathrm{d}x.$$

The constant terms (in B) cancel, the term in B is zero (ψ' is an odd function), and the term in B^2 is $\frac{\psi'(x)^2}{\psi(x)}+\psi''(x)$. This leads to the stated result.