# Quantitative Evaluation of Pairs and RS Steganalysis

Andrew D. Ker

Oxford University Computing Laboratory, Parks Road, Oxford OX1 3QD, England

## ABSTRACT

We give initial results from a new project which performs statistically accurate evaluation of the reliability of image steganalysis algorithms. The focus here is on the *Pairs* and *RS* methods, for detection of simple LSB steganography in grayscale bitmaps, due to Fridrich *et al.* Using libraries totalling around 30,000 images we have measured the performance of these methods and suggest changes which lead to significant improvements.

Particular results from the project presented here include notes on the distribution of the RS statistic, the relative merits of different "masks" used in the RS algorithm, the effect on reliability when previously compressed cover images are used, and the effect of repeating steganalysis on the transposed image. We also discuss improvements to the Pairs algorithm, restricting it to spatially close pairs of pixels, which leads to a substantial performance improvement, even to the extent of surpassing the RS statistic which was previously thought superior for grayscale images.

We also describe some of the questions for a general methodology of evaluation of steganalysis, and potential pitfalls caused by the differences between uncompressed, compressed, and resampled cover images.

**Keywords:** Steganography, Steganalysis

## 1. INTRODUCTION: STEGANOGRAPHY AND STEGANALYSIS

The aim of steganography is to transmit hidden data embedded in a cover medium, in such a way that no third party can determine that any data was embedded. The competing goal of steganalysis is to detect, as reliably as possible, the presence of hidden data. There is now a substantial body of literature on techniques for steganography, particularly in the case when the cover medium is formed of digital images, and an increasing amount on steganalytic techniques for uncovering the presence of steganography. However the latter methods are by their nature imperfect, and sufficiently scientific evaluation of their reliability is often lacking.

In this paper we outline a new project, involving a large image library and a distributed network of computers, to evaluate carefully the various steganalytic algorithms (for now restricting our attention to digital images). Our aim is to compare the performance of the different steganalysis techniques, but also to understand better their weaknesses so that they may be improved. The project is still at an early stage, and in this paper we give some of the first results from it, focussing on one particular type of steganography and two particular detection methods. Already we have found ways to improve their performance.

### 1.1. The Simple Classification Problem

We take on the role of an "information security officer", a hypothetical Warden whose job it is to scrutinise electronic communication and decide whether hidden data is being transmitted. For now we are not interested in more advanced analysis of suspect images (such as estimates of hidden message length[1–3]) although we discuss this issue in Sect. 1.2. This, then, is the "simple classification" question – to classify whether a given image has hidden data or not – and our project is currently focussed solely on the reliability of steganalysis methods to answer this question.

Each steganalysis method will be statistic (a function of the input image) designed to discriminate between the two cases. The values taken by the statistic will vary between images, but if it works well there will be two separated sets of values taken by the statistic, corresponding to the two cases of hidden or no hidden data. Thus we are looking for a hypothesis test, where the null hypothesis ($H_0$) is that no data is hidden, and the alternative

Further author information: E-mail: adk@comlab.ox.ac.uk, Telephone: +44 1865 283519

hypothesis ($H_1$) is that data is hidden. In some steganalysis literature[4, 5] the authors have implicitly designated the null hypothesis to be the case when data *is* hidden. This is contrary to the normal traditions of statistics, where the null hypothesis is taken to be the conservative case which we require evidence to reject. Furthermore, one usually makes the null hypothesis *simple* (not depending on unknown parameters) – this only applies to the case of no hidden data. When the null hypothesis is simple a proper analysis of the *p-value* of an observation (the probability of it occurring by chance when $H_0$ is true) can be carried out. Under our designation – the null hypothesis is the case of no hidden data – the p-value is a particularly important figure as it reflects the certainty with which a positive result can be declared.

Since we shall be simulating the distribution of the statistic we will require that the alternative hypothesis also be simple. So, for a particular experiment, we must fix not only the steganography method but also the size of the hidden data. Because images vary in size, we have chosen to specify message lengths as a proportion of the maximum available, for any particular image. In this way we obtain a simple null hypothesis (there is no hidden data so the cover image is drawn from the loosely defined set of "natural images") and alternative hypothesis (in which a particular method is used to embed a certain amount of hidden data, that amount being a fixed proportion of the maximum possible for each image). The price we pay is having to repeat experiments with many different amounts of hidden data and different steganography methods.

Usually the discriminating statistic will give higher values in the case of hidden data and lower values otherwise (or vice versa). In this case the Warden's only sensible strategy is to reject the null hypothesis (make a positive diagnosis of steganography) when the statistic exceeds a certain threshold. But in practice the distributions (histograms) of the statistic in the case of $H_0$ and $H_1$ will overlap so there is no threshold which will make the detector work perfectly. Given the choice of a number of imperfect thresholds, then, the question is which to use. By varying the threshold one obtains a *region of confidence* (ROC) curve, which plots the *reliability* (probability that the statistic gives a true positive result in the case of the alternative hypothesis) against the *false-positive rate* (probability that the statistic gives a false postitive result in the case of the null hypothesis). In statistical terminology, the former is the *power* of the hypothesis test using a particular threshold value, and the latter the *significance level*. The ROC curve can be computed easily from the distribution of the statistic under the null and alternative hypothesis.

Some authors avoid plotting a compelte ROC curve by choosing a threshold to equalise the probability of false positive and false negative results. But there is no reason to do this. Indeed, if steganography is expected to be a rare occurrence one should aim for a false positive rate much lower than the false negative rate, lest positives of the false type vastly outweigh positives of the true type. Our interest will generally be in steganalysis methods which lead to low false-positive rates, even if this is at the expense of some reliability.

## 1.2. Threshold-Free Statistics

We discuss briefly "threshold-free" statistics, which are found in some recent steganalysis literature and include the two statistics under consideration here. Threshold-free statistics usually provide an estimate of the length of the embedded message as opposed to simply trying to discriminate between the presence and absence of hidden data. In some sense this eliminates the need for a detection threshold because one simply reports the estimate instead of a yes/no output.

However we believe that this is a misleading direction to take: the primary function of a statistical detector is simple detection, and this is the primary concern of an information security officer. While undoubtedly a valuable addition to the Warden's toolbox, we feel that threshold-free statistics are not of themselves superior if they cannot answer the fundamental information security question – the classification question – reliably. One imagines that a Warden would first run simple tests to detect the presence of any hidden data, and only proceed to estimate its length (or, indeed, to try to recover the message itself) if a positive result occurs.

Perhaps the popularity of threshold-free statistics is a reflection of the desire for a level of certainty in the yes/no decision. But the p-value of an observation is a more precise measure of certainty, as it is the (un)likelihood of seeing a value as or more extreme by chance. Contrast the information that a particular image has an estimated hidden message length of 0.03 bits per pixel (without more information we cannot determine whether this is large enough warrant a positive result) against the fact that the probability of seeing a value that far from zero is less than 1%.

Both Pairs and RS are threshold-free and output the estimate of the hidden message length as their statistics. We can still evaluate these steganalytic methods by their reliability at answering the simple classification problem, because the estimate itself ought to be a reliable discriminator between images with no hidden data (when the statistic ought to be close to zero) and some hidden data (when it ought to be higher). We should accept, however, that our evaluation of these statistics is incomplete in as much as it does not take into account their additional uses as message length estimators. Furthermore one might ask whether focussing on the "wrong" question and using a statistic designed to estimate the hidden message length is suboptimal for the simple classification question, and whether there are better techniques designed for the latter*; further research is needed.

### 1.3. LSB Steganography, Pairs and RS Steganalysis

The results we present in this paper are limited to one type of steganography and two methods for detecting it. We describe them very briefly here only in as much detail is needed to understand the results, and refer the reader to the papers which first presented them for a fuller explanation.

Least significant bit (LSB) methods are ubiquitous in image steganography. After converting the hidden message into a stream of bits, one simply goes through the image replacing the least significant bits of pixel values with the hidden message. If, as in most of the investigations in this paper, the hidden message contains less bits than the cover image has pixels, it is best to spread the modifications randomly around the cover image – either by scanning through the image and leaving random gaps, or (better) by generating a random permutation of the image and using the permutation to decide the order of pixels to modify. In either case a key for generating the random gaps or permutation is presumed shared with the intended recipient of the stego image.

The methods of Pairs Analysis[3] and RS Analysis,[1] both due to Fridrich, are the two which we examine here. They have some features in common – in both cases there is a function of images which can be shown to be quadratic in the amount of embedded data when LSB replacement is used, and by making one assumption it is possible to obtain sufficient information to solve for that parameter.

Pairs Analysis first splits an image into a *colour cut*, scanning through and selecting only pixels which fall into each pair of values (0,1), (2,3), and so on. Concatenating the colour cuts into a single stream, one measures the homogeneity of the LSBs. Repeating with the alternatives pairs of values (255,0), (1,2), (3,4), etc, one can show that the function defined by the difference between the two homogeneity measures is quadratic in the amount of embedded data. Under the assumption that "natural images" have no difference in homogeneity, one can obtain enough information to deduce the amount of embedded data in an image, and this estimate forms the statistic we will use to distinguish the cases of hidden data present and absent. However the method is not reliable for images for which the assumption of equal homogeneity does not hold.

Pairs Analysis was designed with paletted images in mind, but there is no theoretical reason why it should not work for grayscale images and we will show it can be made to work well in this case also.

In RS Analysis the image is partitioned into groups of a fixed shape. Each group is classified as "regular" or "singular" depending on whether the pixel noise within the group (as measured by the mean absolute value of the differences between adjacent pixels) is increased or decreased after flipping the LSBs of a fixed set of pixels within each group (the pattern of pixels to flip is called the "mask"). The classification is repeated for a dual type of flipping. Some theoretical analysis and some experimentation show that that the proportion of regular and singular groups form curves quadratic in the amount of message embedded by the LSB method. Under a similar assumption to above, this time about the proportions of regular and singular groups with respect to the standard and dual flipping, sufficient information can be gained to estimate the proportion of an image in which data is hidden. The estimate can be extremely accurate (often within 1%), but fails when this assumption does not hold.

Other methods for the detection of LSB steganography exist (a notable early method was the Chi-square statistic due to Pfitzman and Westfeld,[4] which can be shown much less reliable than the above) but it is fair to say that RS and Pairs are the leading methods at the present time. There are other methods of steganographic embedding too, often much more sophisticated than simple LSB replacement, but they are beyond the scope of this paper.

*In work carried out subsequently to that reported here it became clear that the answer is yes. Results will be reported in a sequel.

## 2. A DISTRIBUTED SYSTEM FOR QUANTITATIVE EVALUATION OF RELIABILITY

We now outline our project for the evaluation of steganalysis algorithms. Since the project is at an early stage and subject to alteration we do not give all the details. We follow the description of the methodology with two important examples which will inform future experiments.

### 2.1. Methodology

Our comments in Sect. 1.1 show that, in order to evaluate the performance of a particular steganalysis algorithm against a particular method of steganography we should approximate closely the distributions of the discriminating statistic in the cases of hidden data and no hidden data (and we have already commented that we will make repeated experiments each with a fixed amount of data to hide). We will see in Sect 3.1 that we cannot find the distributions theoretically so we must take a large sample of images and compute the histograms.

We have negotiated with a number of image library owners to have use of their images, and also have been able to supply quite a large number of personally-owned photographs. We have been using around 30,000 images in our tests, although the selection of the image sets of testing should be done carefully (see Sects. 2.2 and 2.3). For details of the image sets we report results from in this paper, see Sect. 2.4.

The computations performed are simple: for each steganography algorithm under consideration, and each steganalysis method or variation being tested, with a number of different message sizes, we embed the message in each image and compute the discriminating statistic. But because there are many steganography methods we wish to test, against many steganalysis algorithms each with a large number of variations, a range of message sizes, and thousands of images to test (with, as we shall see, the possibility of subjecting them to pre-embedding compression), we expect to have to perform of the order of hundreds of millions of such calculations during the course of this project. Even with efficient implementations of steganography and steganalysis algorithms, it would take a prohibitive amount of time for one computer to do all the work. We use a small distributed network, mostly custom-built from cheap consumer components, to undertake the computations. We do not give all the details here, but they are fairly straightforward because the computation process consists of many unrelated calculations of steganalysis statistics, and these can be parallelised in an obvious way.

The experimental programme has a number of strands. The workhorse is a highly-optimised program dedicated to the simulation of steganographic algorithms and the computation of many different types of detection statistic; the message embedded is random (but reproducible), simulating compressed or encrypted data, and the size of the message, as a proportion of the maximum possible for each image, is a parameter we will vary. Written in C for speed, the code aims to be portable – the present version has been successfully compiled on a number of platforms including i386/windows, i386/linux, i386/sunos, and sparc/solaris, which between them made up the distributed network. Up to 50 machines have successfully been used at any one time, so far. The millions of results of these computations are stored in a MySQL database, with the distributed network also managed through the same database (where a queue of pending computations is kept). At the time of writing the database contained over 5 million rows and should scale smoothly to hundreds of millions. Perl scripts glue together the computation code and the database access, and at present the result analysis code is also written in Perl (since speed is not vital), with results exported to Excel or MATLAB for visualisation.

For each set of the following parameters we generate a “data set”: the image set, which (if any) pre-processing or pre-compression has been applied to the images, the steganography algorithm used (if any), the length of the embedded message, the detection statistic being computed. Each data set gives rise to a histogram of the values taken by the statistic, and from the histograms the ROC curves can be computed and graphed.

We conclude this section with a sample output from the project. We computed the standard RS statistic of Ref. 1 for 5000 JPEG images, before and after embedding a random message into the LSBs of 5% of the pixels in each image. The two histograms of those statistics are shown in Fig. 1, left. The RS statistic is giving results around 0 when there is no embedding, and results around 0.05 when there is embedding, so it is functioning correctly. But the message-length estimates are approximate, the two distributions overlap, and so any procedure which assigns a positive result to RS statistics in excess of a particular threshold value will be imperfect. Varying the threshold and counting the false-positive/false-negative rates gives rise to the ROC curve in Fig. 1, right, on

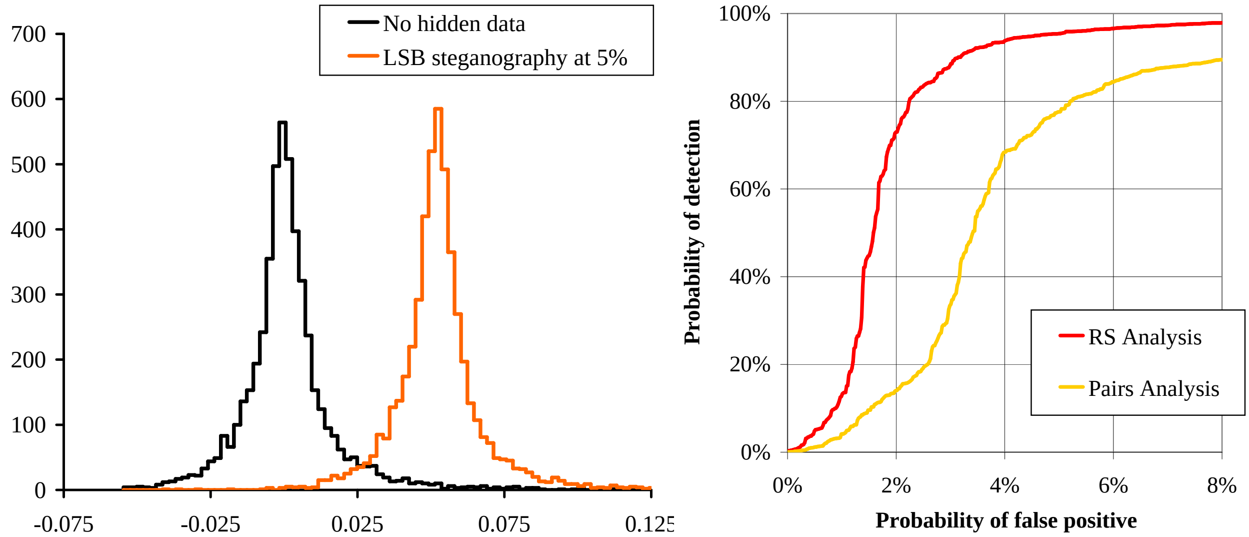

**Figure 1.** Sample output from the project. The histogram shows the distribution of the standard RS statistic over 5000 images before and after the embedding of LSB steganography in 5% of pixels. The graph shows the corresponding ROC curve, and the same ROC curve for the Pairs statistic.

which is also displayed the same ROC curve for the standard Pairs statistic of Ref. 3. One can observe that, as generally accepted, the RS statistic performs better than (i.e. more reliably then) the Pairs statistic on grayscale images (at least for this particular set of images and at this particular embedding rate).

There is a natural desire to reduce the ROC curve further, to a one-dimensional quantity, so that one can easily say that one statistic is "better than" another. We resist the urge, because quite often ROC curves will cross and it then depends on the desired false-positive rate, and hence the application, as to which is "better". In particular it is not often useful to fix a particular false-positive rate and compare the reliability rate at this point, as the curves are usually loosely sigmoid in shape and reliability tends to fall off very rapidly below a certain false positive point. A more reasonable one-dimensional measure of performance, and one we quote on occassion, is the level of false positives when the threshold is set for 50% reliability. We find that this often serves as a fair summary.

When choosing which ROC curves to graph we will focus on "interesting" cases – we will choose a steganography embedding rate so that the performance is neither too near perfect (in which case any differences are as likely due to chance as anything else) or too poor (because results of that nature are not interesting). For this paper this will mean rates of between about 1% and 20%. We will also scale the x-axis (false positive rate) so that the graph shows only areas of interest. The y-axis will always run over reliability rates of 0% to 100%.

## 2.2. First Warning Example

Inevitably, most of the images in our library have been subject to JPEG compression: most digital cameras store in JPEG format by default, and most image libraries supply photographs in the same way, if only for reasons of space. We will demonstrate that the performance of the RS and Pairs statistics is highly dependent on the type of images under investigation, with JPEG and uncompressed images having particularly notable differences. We use a set of 1200 uncompressed images, all 1024 × 768 pixels in size[†] and also JPEG compressed the same images using quality factor 75, a moderate level of compression. The graph on the left of Fig. 2 shows the ROC curves for the standard RS statistic for the original and compressed image set – the difference is quite

[†] All these images came from the same digital camera, which may have left characteristic traces in the images. Ideally we would have liked a larger and more heterogeneous collection, but obtaining uncompressed images in large numbers is quite difficult. For the purposes of this Section these issues will not be important.

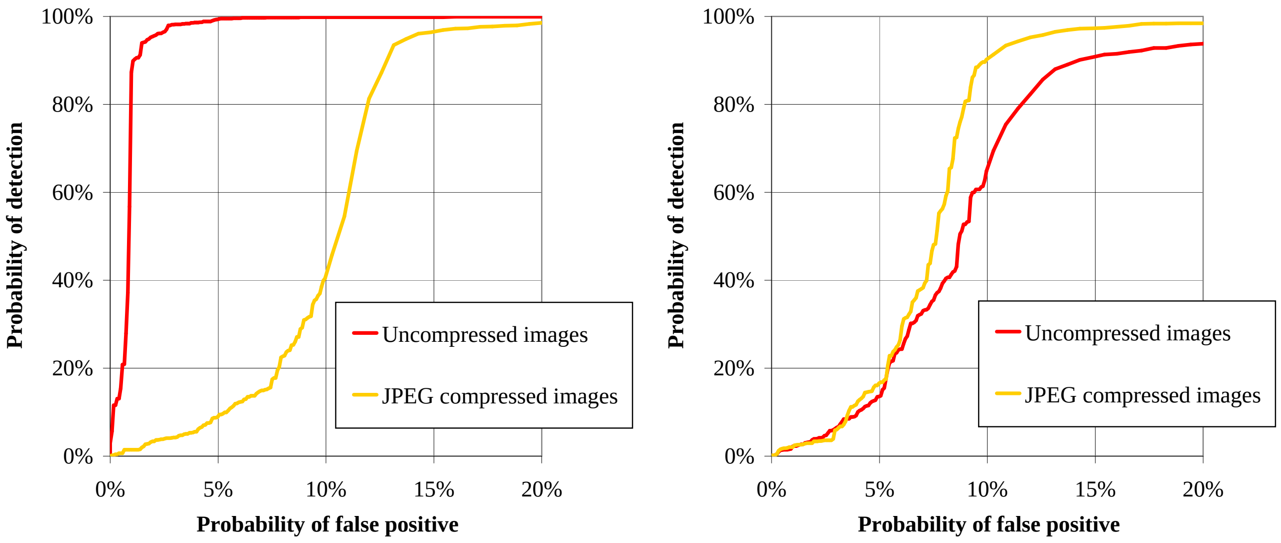

**Figure 2.** Left, ROC curves for RS steganalysis, for an uncompressed set of images and repeated with the same images first subjected to JPEG compression. 5% of the pixels were used for LSB steganography. Right, the same ROC curves after the images were reduced in size from 1024 × 768 to 800 × 600.

alarming. Repeating the experiment for Pairs Analysis (ROC graph not shown) shows less dramatic but still quite significant differences – although with Pairs Analysis the JPEG compressed image set gives rise to a better performance than the uncompressed set.

Any evaluation of the performance of RS or Pairs must take this factor into account – to only test against one type of image will give results which cannot be representative of the group as a whole and could be extremely misleading. We feel that many authors have fallen into this trap. Furthermore, this example shows that our aim of finding *the* ROC curve for these statistics is unattainable, because it is impossible to say what the proportion of JPEG compressed images should be in a "representative" set. Moreover we expect that there are other classes of images which give rise to particularly unusual results – images taken with digital cameras may leave traces due to the CCDs, images scanned from paper will have particular types of noise introduced by the scanning head, and images scanned from film will exhibit granularity from the film itself. We plan to test such images in due course, with the expectation that the results of most steganalytic algorithms will be affected.

The best we can do is to test separately against sets of uncompressed and compressed images, and any other types of images which arise, and report results for all of them. If some observation (such as better performance by one statistic over another, or undetectability by a certain statistic for steganography below a certain bit rate) is repeated for all types of images then we can be reasonably sure that it holds in general. Our preference, therefore, would be for a very large set of uncompressed images for which we can repeat experiments – after compressing the image set with various quality factors, simulating CCD noise or film granularity, and so on. But we have already commented that large numbers of uncompressed images are hard to come by, and in fact it is very difficult indeed to find an image which was not taken with a CCD or scanned in some way. (An image in nature is an analogue observation, so some D-to-A conversion must take place at some point to get it into a computer!)

The temptation is to "simulate" uncompressed images by manipulating JPEG images in some way: some authors have suggested that reducing JPEG images in size will suffice, because the compression artifacts are smoothed away by the resampling process (and one would hope that any artifacts due to particular CCDs or film granularity would be similarly removed). We urge caution: on the right of Fig. 2 we show the same ROC curves for the same two sets of images reduced down to 800 × 600. There is clearly still a difference between the two curves (although, curiously, the relative performances have swapped) – although not as great a difference as

**Table 1.** Shows whether a statistically significant difference in distribution of RS statistic is observed, between uncompressed and JPEG compressed images, after reduction in size. Kuiper's 2-Sample Test is used. The original image size was 1024 × 768.

| Quality factor of JPEG compression | p-value of test for distributional equality after reduction to | | | | | | |
|---|---|---|---|---|---|---|---|
| | 900 × 675 | 800 × 600 | 640 × 480 | 512 × 384 | 320 × 240 | 200 × 150 | 100 × 75 |
| 90 | $< 0.0001$ | $< 0.05$ | $< 0.05$ | Not sig. | Not sig. | Not sig. | Not sig. |
| 80 | $< 10^{-6}$ | $< 10^{-6}$ | $< 10^{-6}$ | $< 0.05$ | $< 0.05$ | Not sig. | Not sig. |
| 75 | $< 10^{-6}$ | $< 10^{-6}$ | $< 10^{-6}$ | $< 10^{-6}$ | $< 0.0001$ | $< 0.05$ | Not sig. |
| 50 | $< 10^{-6}$ | $< 10^{-6}$ | $< 10^{-6}$ | $< 10^{-6}$ | $< 10^{-6}$ | $< 10^{-6}$ | Not sig. |

before shrinking the images, it is still statistically significant and cannot be ignored. Thus we next investigate how far JPEG images should be reduced so that the effects of the JPEG compression can be removed.

Comparing ROC curves is difficult, especially as there is no simple statistical test for significant differences between two whole curves. Instead we return to the histograms which generate the ROC curve. Using the same 1200 images, and a number of reduced sizes (ranging from 900 × 675, a reduction by 12% in either dimenson, down to 100 × 75, reduction by more than a factor of 10), and a number of JPEG quality factors (we used quality factor 90 to represent mild compression, 80 and 75 to represent moderate compression, and 50 to represent harsh compression) we compared the distribution of the RS statistic for two samples: the original images reduced in size versus the JPEG compressed images reduced in size. (No data was hidden at first; repeating the test with different levels of steganography gave similar results). We emphasise that these were the same sets of pictures to start with so that any differences must be due to the JPEG compression. A suitable method for the comparison of the two distributions is Kuiper's 2-Sample Test.[6]

If the RS statistic has much the same distribution in either case then the reduction in size has wiped out any of the differences caused by JPEG compression, at least in as much as it will affect the results of the RS statistic. The results are shown in Table 1, where the order of magnitutude of the p-value of the test is shown in those cases where significant discrepancy was observed. We find these results startling: they show that mildly or moderately compressed JPEG files must be reduced in size very substantially before the artifacts of the compression are completely "washed out". Similar results arise if one compares the effects of different levels of JPEG compression. For example, differences in distribution between 1024 × 768 images compressed using quality factor 90 and quality factor 75 are significant down to reduction to 200 × 150, and between quality factor 80 and 75 are significant down to reduction to 512 × 384.

We must conclude that "simulation" of uncompressed images is a difficult aim. In the absence of large numbers of uncompressed images we have no option but to try it, but the examples in this section are a clear warning that results must be viewed cautiously.

Finally, we should point out that a steganographer performing LSB replacement on a JPEG cover image is opening themselves up to other attacks, including *JPEG compatability analysis*.[7] Also, we have found that steganalysis by Pairs or RS generally seems easy for the Warden when the cover set are compressed images (in contrast to the results of Fig. 2, left, which we can only guess are due to the characteristics of the cheap digital camera used to take the pictures and which were removed by the JPEG compression). But given that most large images are JPEG compressed, and most steganography software available on the internet involves LSB replacement, this combination is well worth examining in depth.

### 2.3. Second Warning Example

Our second warning example is related: this time we examine the effects of different methods of shrinking images. There are a number of different ways of performing the interpolation necessary to reduce a digital image, and

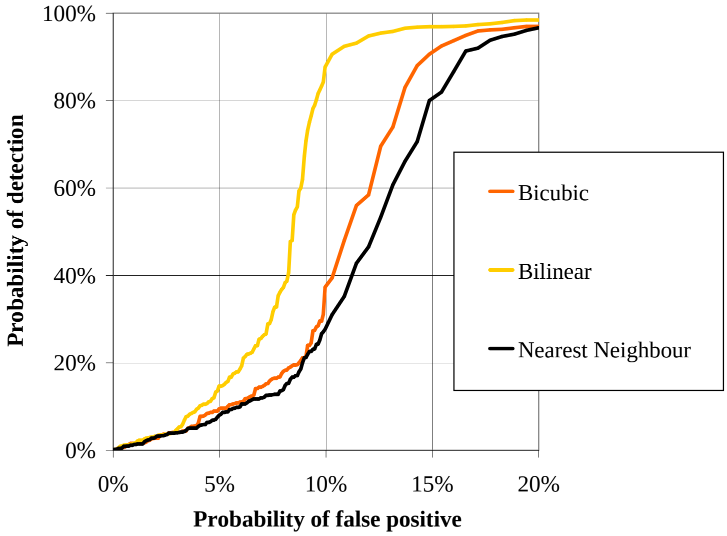

**Figure 3.** ROC curves for RS steganalysis; the same set of images is reduced to $800 \times 600$ using Adobe Photoshop's three different resampling algorithms. As before, 5% of the pixels were used for LSB steganography.

we will demonstrate that the choice of interpolation algorithm can have an effect on the results of reliability analysis.

We use the same set of 1200 uncompressed $1024 \times 768$ images, and reduced them to $800 \times 600$ using Photoshop's three resampling algorithms: bicubic and bilinear interpolation, and nearest-neighbour. Using the same steganography and steganalysis methods as before (LSB steganography on 5% of the pixels, standard RS steganalysis with mask $[0, 1, 1, 0]$) we plotted the ROC curves for the three data sets. The results are shown in Fig. 3.

In fact the differences between the bicubic and nearest-neighbour curves are not statistically significant, but the bicubic and bilinear resampling schemes have given quite different results. As before, we can return our attention to the histograms which give rise to the ROC curves – comparing the null distributions of the RS statistic for the image sets reduced by bicubic and bilinear algorithms, we see a statistically significant difference (the p-value of Kuiper's 2-Sample Test is about 0.001).

Similar results are observed if the uncompressed images are reduced by other amounts (although the differences in distribution become insignificant if the reduction is by a factor of more than about 3). We do not display the ROC curves for reasons of space, but if one substitutes the Pairs Analysis statistic for RS, the results are much stronger – it seems that Pairs is more sensitive to particular resampling methods than RS.

We must conclude that, even if it were possible to remove JPEG compression artifacts reliably by shrinking, the method used to shrink introduces its own artifacts which can affect the outcome of the reliability analysis.

### 2.4. Image Sets Used for Testing

These two warning examples show that, at this stage, it is impossible to produce a "representative" set of images. We have resolved to use a number of different sets of images, not only to cover all types of images but also to expose any differences in performance. One may hope that, in time, we can explain why one statistic performs better than another on a particular type of image and use this to make an improved statistic.

Our principal investigations will use the following:

**Image Set A:** 2200 simulated uncompressed images, all $512 \times 512$. The "simulation" of uncompressed images is performed by taking large and very mildy compressed JPEG files (original image sizes 2000-4000 pixels square, and JPEG compression corresponding to quality factors of at least 95) and reducing to $512 \times 512$. The reduction by a very large factor is in view of the first warning example, and because of the second we have used a variety of resampling algorithms in the reduction. Most of the images, before reduction in size, came from a collection of personal photographs taken by the author, and were of very high quality. In particular out-of-focus images and those with incorrect exposure have been removed from the set. It will be seen that this set of images is "difficult" for the steganalyst, in that the statistics generally perform less reliably than they do for the other sets. Since this set is uncompressed we can also repeat experiments by pre-compressing these images, giving a good measure of how the statistics' reliability varies with the amount of JPEG compression applied to the covers.

**Image Set B:** 5000 JPEG images, all sized 900 × 600. Each is compressed at JPEG quality factor 75. These came from a royalty-free image library purchased by the author. The photographs are of quite good quality in terms of exposure and focus, but they appear to have been scanned in from 35mm film and some show granularity. Some have a small black border. This set is generally "easier" than set A, in that statistics generally perform more reliably on it. This is undoubtedly partly due to JPEG compression, and the slightly larger image size but we have not yetattempted to analyze the reasons in full.

**Image Set C:** 10000 JPEG images of moderate quality, sizes varying between 890 × 560 and 1050 × 691. The JPEG compression levls also vary: about 85% of the images are quality factor 50 but others have compression equivalent to quality factors between 65 and 75. These images came from another royalty-free image library, but the quality of pictures is not as good as set B; some images are blurred or incorrectly exposed. This set turns out to be a little "easier" than set B, with the statistics being slightly more reliable, but the difference is not very great.

We will not use again the set of 1200 uncompressed images because it is rather small for accurate evaluation of reliability, and because there is a suspicion that the particular digital camera used to take them may effect the results. In our image library we also have a number of other sets of images with more widely varying size and quality, but we did not find that these sets told us anything markedly different from those above, and so do not report their results here.

## 3. RESULTS AND ANALYSIS

In the rest of this paper, notwithstanding our warning examples and the issues they raise, we detail some selected results from the project. They fall into four areas.

### 3.1. Distribution of the RS and Pairs Statistics

In Ref. 1 there appears a claim, based on a sample of 331 small images, that the RS estimate of hidden data length, when no data is actually hidden, is normally distributed. We have found compelling evidence that this is not the case, and so one should not use this approximation in the estimation of false positive rates. Although the centre of the distribution looks approximately normal (which could lead to the misapprehension that it is normal if a small sample is taken), the tails are much heavier. The same applies to the Pairs Analysis statistic.

In Fig. 4 we show the histogram of the RS statistic, calculated for the 5000 images in Image Set B, and when no steganography is performed. We also show the "best-fit" normal density. It appears that the RS distribution is more heavily-tailed than normal, and in particular there are many and more drastic outliers than one would expect. To be statistically precise, we can carry out a standard normality test, such as that of Anderson and Darling.[8] The p-value of this test is vanishingly small and thus it is certain that the RS distribution is not normal.

It is not necessary to use the relatively sophisticated technology of the Anderson-Darling test, however. A simple measure of the heaviness of the tails of a distribution is *kurtosis*; if $X$ is a random variable with mean $\mu$ then its kurtosis is defined by

$$\kappa = \frac{E[(X-\mu)^4]}{E[(X-\mu)^2]^2}. \tag{1}$$

For normal distributions $\kappa = 3$, and for heavier-tailed distributions it is greater than 3. The kurtosis of a sample is defined analagously, and the asymptotic variance of the kurtosis of normal samples is $24/n$, where $n$ is the sample size.[9] Although the sample kurtosis actually converges very slowly to its limit, the sample sizes in the

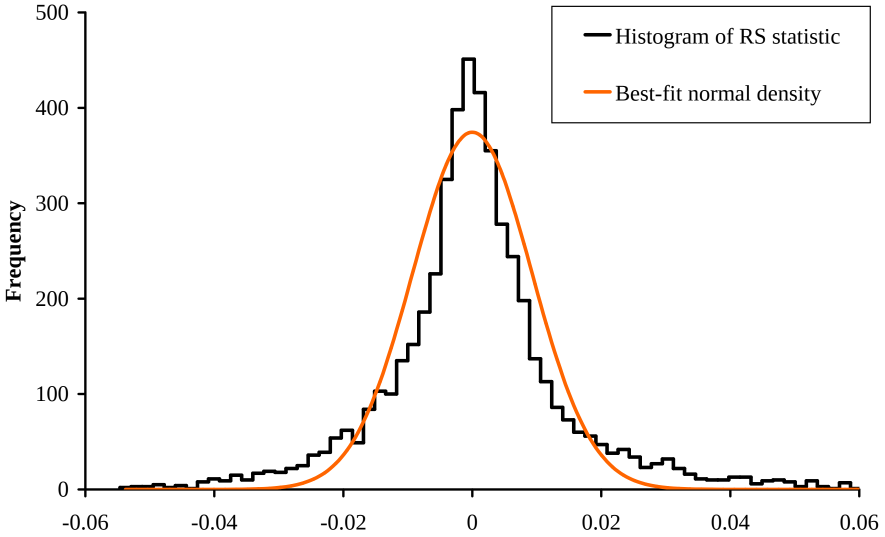

**Figure 4.** Histogram of distribution of RS statistic for 5000 JPEG images, with best-fit normal density superimposed. The observed distribution has substantially longer and heavier tails than the normal distribution.

thousands used here allow us to use it as a very suitable test for over-heavy tails. Table 2 shows the kurtosis for a number of image sets and both the Pairs and RS statistics, along with the significance level of the observation.

We deduce two things. Firstly, none of these samples are anywhere near normal in distribution. Second, their respective kurtosis values differ significantly between each other. This means that we have no use for finding an alternative model (e.g. generalised Gaussian) of these distributions: even if we did, it would have to have some parameter to explain the different in kurtosis, and for a given image we have no way of determining the parameter of the distribution it came from, and hence no way of calculating a p-value. We conclude that simulation of these distributions, by taking large samples of natural images, is the only way to proceed.

### 3.2. Varying the RS "Mask"

The results of RS Steganalysis depend on the size of the groups the image is partitioned into, and the "mask". The mask is the same size as the pixel groups and consists of zeros and ones, determining which pixels in each group are flipped as noise is measured. In Ref. 2 the authors use the flat mask $[0, 1, 0]$ and the $2 \times 2$ square mask $\begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$ but do not explain where they come from.

**Table 2.** Observed kurtosis and significance levels for the test against normality, for the various image sets.

| Statistic | RS Statistic | | | Pairs Statistic |
|---|---|---|---|---|
| Image Set | A | B | C | A |
| Sample kurtosis | 21.56 | 20.17 | 32.60 | 15.53 |
| Normal kurtosis | 3.0 | 3.0 | 3.0 | 3.0 |
| Theoretical standard deviation of sample kurtosis | 0.104 | 0.069 | 0.049 | 0.104 |
| p-value of result | $< 10^{-6}$ | $< 10^{-6}$ | $< 10^{-6}$ | $< 10^{-6}$ |

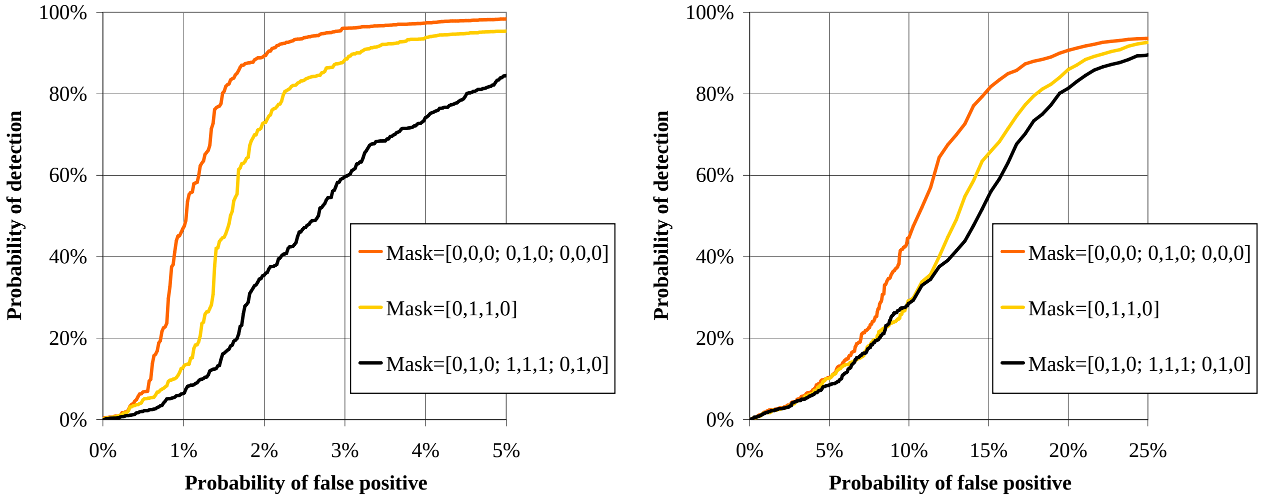

**Figure 5.** ROC curves for the RS statistic with three different masks. Left, 5000 quality factor 75 JPEG images (Image Set B); right, 2200 uncompressed images (Image Set A). 5% of the pixels are used for LSB stegangraphy in both cases.

We performed a number of investigations into the effect the mask has on the reliability of RS Steganalysis. Among the masks we tried were the flat masks $[0, 1]$, $[0, 1, 0]$, $[0, 1, 1, 0]$, $[0, 1, 1, 1, 0]$, $[0, 1, 0, 1, 0]$, and the squares

$$M_2 = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix},\ M_{3a} = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix},\ M_{3b} = \begin{bmatrix} 0 & 1 & 0 \\ 1 & 1 & 1 \\ 0 & 1 & 0 \end{bmatrix},\ M_{4a} = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 \\ 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix},\ M_{4b} = \begin{bmatrix} 0 & 1 & 1 & 0 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 0 & 1 & 1 & 0 \end{bmatrix}.$$

We generated ROC curves for each mask, for each of Image Sets A, B, and C, and with LSB steganography levels of between 1% and 20% of the maximum. Below 1% none of the statistics was performing with any useful level of reliability, and above 20% the statistics are functioning so near to perfectly that any differences in results are most likely due to chance. In the case of Image Set A we also pre-compressed the cover images using JPEG quality factors of 50, 75 and 90, as well as leaving the images uncompressed.

It would be tedious to include all the ROC curves here, so we summarise the results. In general there were no dramatic differences between the results from the different masks, however there were some small but significant differences which we report. We found that the masks $[0, 1, 0]$ and $M_{3a}$ gave somewhat better performance than the others, with the square mask slightly better than the flat mask in some cases. $M_2$, $[0, 1, 1, 0]$ and $[0, 1, 0, 1, 0]$ were the next-best performers. $[0, 1, 1, 1, 0]$, $M_{3b}$, $M_{4a}$ and $M_{4b}$ were somewhat worse in reliability. The results held no matter what level of steganography or Image Set was used.

We illustrate with the ROC curve for $M_{3a}$, $[0, 1, 1, 0]$, and $M_{3b}$, for Image Set B and 5% of the pixels used for steganography, see Fig. 5, left. One can see that $M_{3a}$ performs the best, but also that the difference with $[0, 1, 1, 0]$ is not overwhelming. Note also how well RS Steganalysis is performing in general: with only 5% of the capacity of the cover being used, the best mask is detecting 50% of stego messages with only 1% false positives. Contrast this with Fig. 5, right, noting the different x-axis labels. This is the same experiment performed on Image Set A. Here RS Steganalysis is not detecting the steganography with any useful level of reliability. Although this is in part due to Image Set A being formed of smaller images, that does not account for all of the difference in performance. Clearly there is further research needed to explain the cause.

We end this Section with some evidence regarding the claim in Ref. 1 that "for high quality images from scanners and digital cameras, we estimate that messages requiring less than 0.005 bits per pixel are undetectable using RS Steganalysis. Higher bit-rates are in the range of detectability using RS Steganalysis." The figure of 0.005 bits per pixel (bpp) is repeated in Ref. 2. Our results show that 0.005 is too low, even with the best performing mask found here, especially for high-quality uncompressed images.

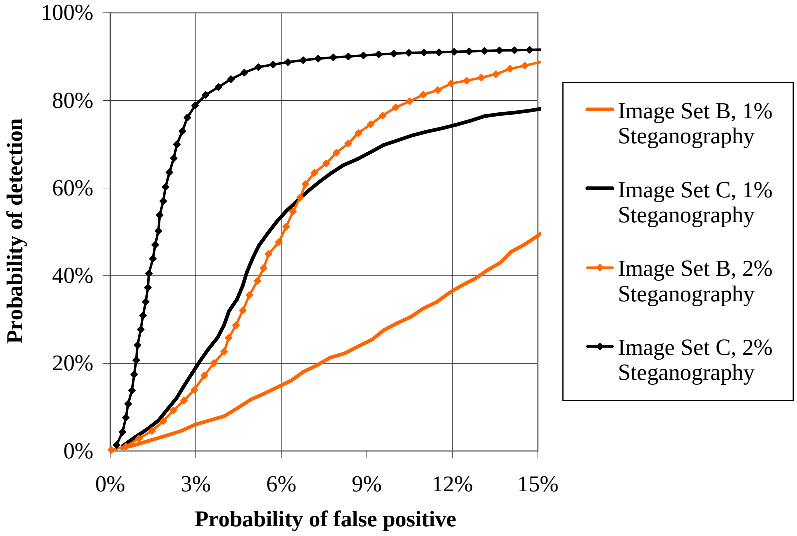

**Figure 6.** ROC curves for the RS statistic with what seems to be the best mask $M_{3a}$, testing the lower limit of detectability. 1% and 2% of the pixels are used for steganography, and Image Sets B and C are tested separately.

In Fig. 5, right, we already showed that RS steganalysis does not reliably detect messages using as much as 5% of the capacity (0.05 bpp) on Image Set A; for 2% the performance is barely better than a random declaration of hidden or no hidden data (ROC graphs not shown). In Fig. 6 we show how Image Sets B and C perform when 0.02 and 0.01 bpp are used. Whether one considers a detector to have useful reliability depends on the application, but at 0.01 bpp even the “easiest” Image Set C shows 5.5% false positives when the reliability rate is 50%, which is probably not acceptable. This figure also illustrates the difference in performance when tests are carried out on the two image sets; in fact, no matter what the test, Image Set C is constantly the “easier” set for the steganalyst, with lower false positives then in Set B. This is another lesson in the neccessity for wide testing.

So we suggest that, based on the evidence accumulated here, it is over-optimistic to hope that RS steganalysis will detect LSB steganography down to 0.005 bits per pixel. One should not have confidence in it for levels of below 0.01-0.02, and for high-quality images which have not been subject to JPEG compression (as in Image Set A) the figure is around 0.05.

The reason for false-positive results is the heavy tails in the null distribution. A few of our “natural” images produce a *bias* (estimated proportion of hidden data when no data is hidden) as high as 0.5, although such extremes are very rare. This shows why one should use a very large image library; if only a few hundred images are used there may, by chance, be no outliers at all in the sample.

## 3.3. Improving Pairs Analysis

The standard Pairs Analysis method creates a “colour cut” of a image by scanning row-by-row and picking out pixels matching certain conditions, repeating for different conditions, and concatenating the results. Although it is quite reasonable to assume some sort of correlation between adjacent pixels in an image[‡] , it seems unlikely that two pixels which are spatially well-separated, but happen to be consecutive when pulled out of the row-by-row scan for particular colours, would have the same properties. In particular we ask whether including such pixels (adjacent in the colour cut but spatially separated in the image) in the homogeneity calculation is introducing noise into the results.

[‡] In fact the method of Pairs does not rely on such a correlation *per se*, but it does require the number of adjacent pixels which differ by one to exceed the number which differ by two.

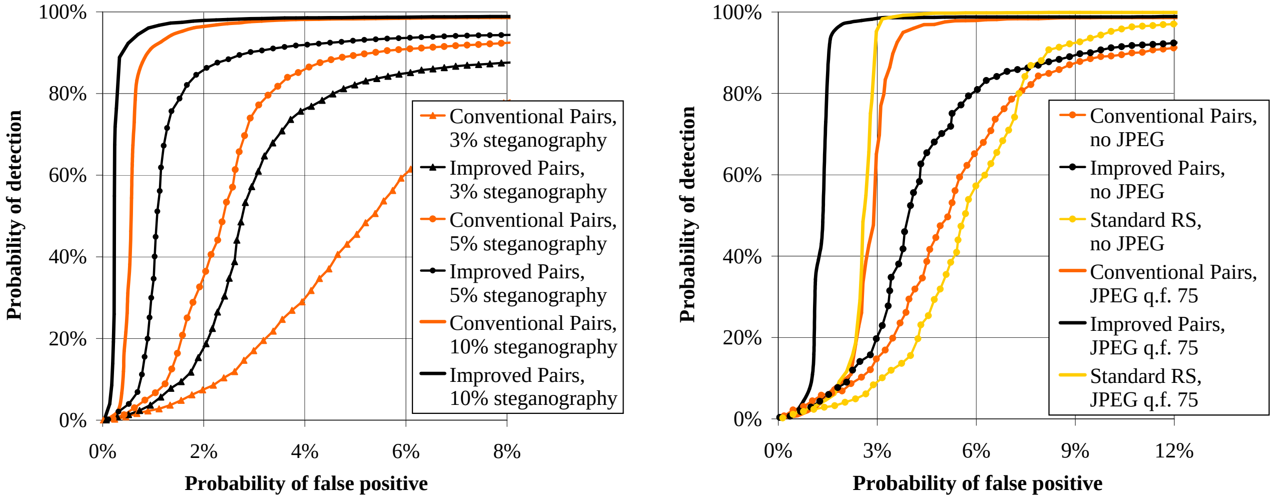

**Figure 7.** Comparison of conventional and improved methods for Pairs Analysis. Left, the ROC curves for the two methods arising from the combined set of 15000 JPEG images in Sets B and C, with 3%, 5% and 10% of pixels used for embedding. Right, the curves for the two Pairs methods and also the standard RS statistic, this time from Image Set A; 10% of pixels are used for embedding, and the experiment is repeated with the image set subject to JPEG compression (with quality factor 75) prior to embedding.

We implemented a modified form of Pairs Analysis in which non-adjacent pixels were excluded from the homogeneity calculation. In fact this turns out to be computationally simpler as one need not compute the full colour cuts at all; a single pass through the image is performed, keeping track of both homogeneity (counts of equal pixels) and the number of pixels connected both spatially and in the same pair of values; the revised homogeneity measure is the quotient of these two values. The rest of the algorithm is identical to the standard Pairs method (repeating for the alternative pairs of values and solving the same quadratic equation to find an estimate of hidden message length). We then tested extensively the conventional and modified Pairs methods to measure any improvement.

Figure 7 shows some of the results. The graph on the left shows the ROC curves for the conventional and improved statistics, at three different embedding rates (3%, 5%, 10%). In each case the results are from the 15000 images made up of Sets B and C together. The performance improvement is quite substantial: if set for 50% reliability the false positive rates for the different levels of embedding have dropped from 5% to 2.8% (at the 3% embedding rate), 2.4% to 1.1% (at the 5% embedding rate), and 0.5% to 0.2% (at the 10% embedding rate). We observed similar improvements with different image sets and other embedding rates – in most cases the improved Pairs statistic functions almost as well as the RS statistic. In Figure 7, right, we show a particularly interesting result. Here we compare the conventional and improved Pairs statistics with the standard RS statistic, at 10% embedding rate, for the "difficult" Image Set A, repeating with the same images first JPEG compressed (quality factor 75). In these cases the conventional Pairs method was not much worse, or was very marginally better than, the RS method, and the result for the improved Pairs method now surpasses that for RS.

### 3.4. Using the Second Dimension

What holds when scanning an image horizontally ought to hold, in general, when scanning vertically. Thus we are lead to wonder whether RS and our improved Pairs are missing out on some information because they focus on one dimension only. In the case of Pairs, we can try pooling the homogeneity between adjacent pixels (in the same pair of values) of the image scanning both horizontally and then vertically – the same calculations show that we should be able to obtain an estimate for the amount of steganography in exactly the same way as before. To some extent we already make use of the second dimension if we use a square mask for the RS statistic, but we might examine whether one can get away with the flat mask if we pool the counts of regular and singular groups when the pixels are affected by both the mask and a transposed copy of it.

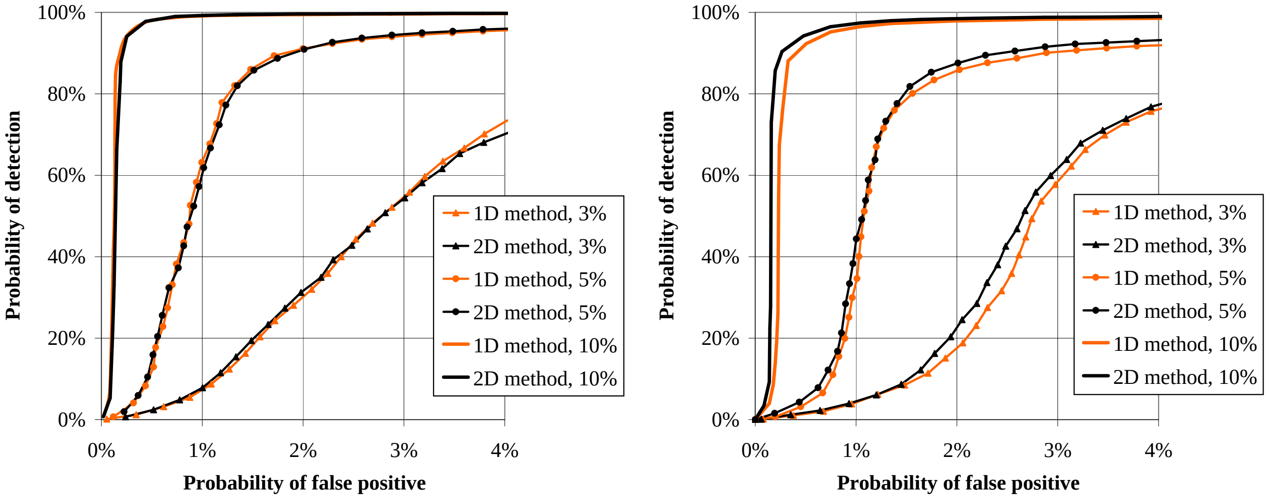

**Figure 8.** Left, ROC curves for the RS statistic, with 3%, 5%, and 10% of pixels used for embedding, comparing statistics using 1- and 2-dimensional masks. Right, the same results for the improved Pairs statistic. The results are shown for the combined set of 15000 JPEG images in Sets B and C.

The graph on the left of Fig. 8 compares the performance of the standard RS statistic (we used mask $[0, 1, 1, 0]$ here) with a two-dimensional version which also classifies pixel groups with the transposed mask. Varying levels of steganogaphy, using 3%, 5% and 10% of pixels, are shown. On the right we repeat for the improved Pairs statistic using 1- and 2-dimensional homogeneity measures. In the case of RS no improvement is apparent. There is some suggestion of a very minor benefit for the improved Pairs algorithm, but the difference is so very small that it may be due to chance (statistically it is of marginal significance). We repeated the experiments using Image Set A, both uncompressed and subjecting the covers to various levels of JPEG compression. The ROC curves are not included here, but again no difference was observed for RS and a tiny improvement for Pairs. Even supposing that the benefit to the Pairs method of using the second dimension is not a fluke, it is so small that the gain is not worth the extra complexity (doubling the number of calculations). We must conclude that using the second dimension in this way – which after all amounts only to using the same pixels again as evidence – does not usefully improve the reliability of either RS or the improved Pairs statistic.

## 4. CONCLUSIONS

The first results from our distributed steganalysis evaluation project teach us a number of important lessons. Firstly, we have seen the extent to which JPEG compression, or resampling artifacts, can affect the reliability of steganalysis algorithms. It should be no surprise to researchers in the image processing community that such artifacts exist, but our contribution is to show that they matter very much to the evaluation of steganalysis and must not be ignored by the information hiding community. More generally we have given examples of three image sets, each quite widely varied in content, which show very different performance when tested against steganalysis algorithms. Thus any conclusion based on experimental results will be incomplete and potentially very misleading unless it includes a number of different image sets and can replicate results across all of them.

Using a large image library and a network of computers we have performed investigations into the effectiveness of using the RS and Pairs methods to answer the simple classification question – the presence of absence of hidden data – in grayscale images. Despite both statistics being designed to answer a different question (the proportion of hidden data) and for different cover sets (RS was originally aimed at colour bitmaps and Pairs at palette images) this has proved a fruitful place to begin.

We have given strong evidence that the distributions of the RS and Pairs message length estimates are not normal. We have investigated the effects of various “masks” used in the RS algorithm, and found that the differences were not great but have identified the best performer. We have also investigated the general power of the RS method and found that LSB steganography using less than 0.01-0.02 bits per pixel was not reliably detectable, even for image set which seemed to show the best performance for steganalysis. For uncompressed images of the type in one of our image sets, the figure appears to be as high as 0.05 bits per pixel.

Pairs Analysis was designed with palette images in mind, but we have shown how it may be modified to work well on grayscale images; in most cases the improved statistic has performance approaching that of RS and in certain cases it proves to be even better. Further improvements to Pairs and RS will follow, as we gain more evidence as to their strengths and weaknesses from the project. Eventually we hope to use our results to make an optimal *combination* of Pairs, RS, and other statistics. Finally, we investigated whether there was anything to be gained by using the second dimension of images, scanning both horizontally and vertically through an image and pooling the resulting measures in both the Pairs and RS algorithms. The only improvement we found was very marginal.

In many respects this paper has raised more questions than it answers, because of the discrepancies between results for various types of images. We hope that future work, informed by the results presented here, will tie up the loose ends.

## ACKNOWLEDGMENTS

The author is a Royal Society University Research Fellow. Much of the work presented here was done during a Junior Research Fellowship at University College, Oxford with additional funding from DSTL. The author is grateful to Sue Haines and Rob Thompson at DSTL for many helpful conversations and suggestions on this research programme.

## REFERENCES

1. J. Fridrich, M. Goljan, and R. Du, “Reliable detection of LSB steganography in color and grayscale images,” *Proc. ACM Workshop on Multimedia and Security*, pp. 27–30, 2001.
2. J. Fridrich and M. Goljan, “Practical steganalysis of digital images – state of the art,” in *Security and Watermarking of Multimedia Contents IV*, E. J. Delp III and P. W. Wong, eds., *Proc. SPIE* **4675**, pp. 1–13, 2002.
3. J. Fridrich, M. Goljan, and D. Soukal, “Higher-order statistical steganalysis of palette images,” in *Security and Watermarking of Multimedia Contents V*, E. J. Delp III and P. W. Wong, eds., *Proc. SPIE* **5020**, pp. 178–190, 2003.
4. A. Westfeld and A. Pfitzmann, “Attacks on steganographic systems,” in *Proc. Information Hiding Workshop, Springer LNCS* **1768**, pp. 61–76, 1999.
5. A. Westfeld, “Detecting low embedding rates,” in *Proc. Information Hiding Workshop*, *Springer LNCS* **2578**, pp. 324–339, 2002.
6. N. H. Kuiper, “Tests concerning random points on a circle,” *Proc. Koninklijke Nederlandse Akademie van Wetenschappen* **63**, pp. 38–47, 1962.
7. J. Fridrich, M. Goljan, and R. Du, “Steganalysis based on JPEG compatability,” in *Multimedia Systems and Applications IV*, A. G. Tescher, B. Vasudev, and V. M. Bove, Jr, eds., *Proc. SPIE* **4518**, pp. 275–280, 2002.
8. R. W. Anderson and D. A. Darling, “Asymptotic theory of certain goodness-of-fit criteria based on stochastic processes,” *Ann. Math. Statist.* **23**, pp. 193–212, 1952.
9. K. O. Bowman and L. R. Shenton, “Omnibus test contours for departures from normality based on $\sqrt{b_1}$ and $b_2$,” *Biometrika* **62**, pp. 243–250, 1975.