Detection of Hiding in the Least Significant Bit

O. Dabeer, K. Sullivan,
U. Madhow, S. Chandrasekharan,
B. S. Manjunath [1]
Department of ECE,
University of California,
Santa Barbara,
CA 93106, USA.
E-mail: onkar@ece.ucsb.edu

Abstract — **We consider the problem of detecting hiding in the least significant bit (LSB) of images. Since the hiding rate is not known, this is a composite hypothesis testing problem. We show that under a mild condition on the host probability mass function (PMF), the optimal composite hypothesis testing problem is solved by a related optimal simple hypothesis testing problem. We then develop practical tests based on the optimal test and exhibit their superiority over Stegdetect, a popular steganalysis method used in practice.**

I. Introduction

Steganography (data hiding) tools are easily available in the public domain (see [1], [2]) and there is a need for the design of steganalysis tools that detect the presence of hidden data. While research in steganography is well advanced, that in steganalysis is still in its infancy. The main reason for this is that in its full generality, steganalysis is an ill-posed problem: the original host data is unknown, the rate of hiding (if data is hidden) is not known, and the number of steganography schemes is large. A review of the few currently available steganalysis tools is given in [3]. Unfortunately, even the most promising existing approaches, such as Stegdetect ([1]) and the supervised learning framework ([4]), have drawbacks that limit their practical use.

1. Existing steganalysis methods are based on heuristics, and given a steganography method, there is no systematic approach for designing a steganalysis method.

2. Every steganalysis method has some parameters to be chosen, which determine the performance of the method. Ideally, the test parameters should be chosen purely on the basis of the data to meet the target performance. Such schemes are lacking in the literature.

3. Due to the lack of a theoretical foundation, it is not known how current steganalysis tests compare with ‘optimal’ tests.

In summary, many fundamental issues in steganalysis are yet to be understood. A promising approach to developing a systematic framework for steganalysis is the theory of hypothesis testing [5, 6]). We adopt this approach in our study, and investigate the detection of one of the simplest steganography methods–least significant bit (LSB) hiding. It is also popular in practice - recently, a web-browser based steganography application using LSB hiding was released by Hactivismo (see [2]). However, even in this simple case, several difficulties must be addressed. First, the host statistics are unknown, in general, so that the models under the hypotheses ”no data hidden” versus ”data hidden” are not known *a priori.* Second, the hypothesis ”data hidden” is a *composite* hypothesis, since the amount of data hidden may vary.

Our main results are as follows. In Section II, we provide a simple model for LSB hiding, assuming that the host symbols are independent and identically distributed (thus, we throw away information regarding correlations between host symbols, which could improve performance if accounted for). In Section III, we study the structure of optimal tests assuming that the host statistics are known, and identify conditions under which the composite hypothesis ”data hidden” can be reduced to the simple hypothesis ”data hidden at rate R_0.” The proof of the latter result is given in Appendix A. In Section IV, we provide estimators for the host statistics which work well under both hypotheses (as long as the host statistics are ”smooth enough”), and then plug in these estimates into the optimal detection rules in (a). This test comprehensively outperforms Stegdetect, and is less sensitive to choice of threshold than Stegdetect. Our conclusions are given in Section V.

II. Statistical model for LSB hiding

We consider the case of independent and identically distributed (i.i.d.) data samples. This model is commonly used in steganography ([7], [8]). Since the host samples are assumed to be i.i.d., without loss of generality we assume the data to be one dimensional. Suppose the i.i.d. host is $\{h_k\}_{k=1}^N$, where the intensity values h_k are represented by 8 bits, that is, $h_k \in \{0, 1, ..., 255\}$. We assume the hidden data $\{d_k\}_{k=1}^N$ is i.i.d. and,

$$P(d_k = 0) = \frac{R}{2}, \quad P(d_k = 1) = \frac{R}{2},$$
$$P(d_k = \text{NULL}) = (1 - R), \quad 0 < R \leq 1.$$

The hider does not hide in host sample h_k if $d_k = \text{NULL}$, otherwise the hider replaces the LSB of h_k with d_k. With this model for rate R LSB hiding, if the probability mass function (PMF) of h_k is $p(l), l = 0, 1, ..., 255$, then the PMF of the data

[1] This research was supported in part by a grant from ONR # NO0014-01-1-0380.

after LSB hiding at rate R is given by,

$$p_{R,2l} = \left(1 - \frac{R}{2}\right) p_{2l} + \frac{R}{2} p_{2l+1}, \tag{1a}$$

$$p_{R,2l+1} = \frac{R}{2} p_{2l} + \left(1 - \frac{R}{2}\right) p_{2l+1}, \quad l = 0, 1, ..., 127. \tag{1b}$$

For sake of convenience, we denote the PMF by the 256-dimensional vectors p, p_R, and we write $p_R = Q_R p$, where Q_R is a 256×256 matrix corresponding to the above linear operation.

III. Optimal Composite Hypothesis Testing

In this section we study optimal tests based on the knowledge of the host PMF; practical tests which do not assume knowledge of the host are given in the next section. Suppose we wish to decide between two possibilities: data is hidden at some rate R, where $R_0 \leq R \leq R_1$, or no data is hidden ($R = 0$). The parameters $0 < R_0 \leq R_1 \leq 1$ are specified by the user. We use H_R to represent the hypothesis that data is hidden at rate R. The steganalysis problem in this notation is to distinguish between H_0 and $K(R_0, R_1) := \{H_R : R_0 \leq R \leq R_1\}$. The hypothesis that data is hidden is thus *composite* while the hypothesis that nothing is hidden is *simple*. Suppose the observed data is $\{x_j\}_{j=1}^N$, where x_j are i.i.d. and take values in some alphabet $\mathcal{A}$. For grey-scale images, $\mathcal{A} = \{0, 1, ..., 255\}$. Mathematically, a detector δ is characterized by the acceptance region $A \in \mathcal{A}^N$ of hypothesis H_0:

$$\begin{aligned} \delta(x_1, ..., x_N) &= H_0, \text{ if } (x_1, ..., x_N) \in A, \\ &= K(R_0, R_1), \text{ otherwise.} \end{aligned}$$

In the absence of an apriori distribution on R when data is hidden, we use the Neyman-Pearson formulation of the optimal detection problem: for $\alpha > 0$ given, minimize

$$P\,(\text{Miss}) = \sup_{R_0 \leq R \leq R_1} P\left(\delta(x_1, ..., x_N) = H_0 | H_R\right)$$

over detectors δ which satisfy

$$P\,(\text{False alarm}) = P\left(\delta(x_1, ..., x_N) = K(R_0, R_1) | H_0\right) \leq \alpha.$$

One way to find the optimal detector is to find the least favorable distribution ([5, Theorem 7, pp. 91]). Intuitively, for steganalysis the worst case corresponds to the smallest hiding rate. The following proposition shows that this intuition is accurate for sufficiently large data lengths N, sufficiently small hiding rates, and under a 'smoothness' constraint on the host PMF p.

Proposition 1 *Suppose $p_l > 0$ for $l = 0, ..., 255$ and define $r_k := p_{2k+1}/p_{2k}$, $k = 0, 1, ..., 127$. We impose the following condition on the host PMF:*

$$U(p) := \sum_{k=0}^{127} \left\{ (p_{2k} + p_{2k+1}) \left(r_k + \frac{1}{r_k} - 2 \right) \right\} < 1. \tag{2}$$

Consider the composite hypothesis testing problem for distinguishing between H_0 and $K(R_0, R_1)$. We restrict our attention to detectors that operate in the region $P(\mathit{Miss}) \leq 0.5$, $P(\mathit{False\ alarm}) \leq 0.5$. Then there exists N_0, $R_* > 0$ such that for $N \geq N_0$, $R_1 \leq R_*$ the unique least favorable distribution is a unit mass at R_0. Therefore, if q denotes the empirical PMF (that is, normalized histogram) of the observed data, the optimal detector for $N \geq N_0$ and $R_1 < R_*$ is the corresponding likelihood ratio test (LRT), which accepts K if

$$S_{LLRT}(q) := D(q \| p_{R_0}) - D(q \| p) \leq T(\alpha),$$

where $T(\alpha)$ is a real-valued threshold chosen to obtain $P\,(\mathit{False\ alarm}) = \alpha$ and $D(p\|q)$ is the Kullback-Leibler divergence.

The proof is given in Appendix A. The main conclusion of the above result is that under (2) the composite hypothesis testing problem associated with steganalysis can be replaced by the simple hypothesis testing problem: test H_0 versus H_{R_0}. The condition (2) means that on an average, the ratio p_{2k+1}/p_{2k} is not too large or too small. This assumption would be satisfied for images whose histogram varies smoothly. We have verified that condition (2) is true for a database of 4000 DOQQ images. Based on this proposition, we restrict our attention to testing hypothesis H_0 versus H_R.

IV. Practical Tests

We note from Proposition 1 that we only need to develop tests for testing H_0 versus H_R, where R is the smallest rate amongst the possible rates the user is testing for. A problem with the optimal LLRT test is that we do not know the host PMF in practice. However, there are two factors that help us to develop good practical tests based on the optimal LLRT.

1. The hiding rate in practice is very low, and therefore, we can estimate the host PMF well. We found that a number of simple estimates of the host PMF based on the assumption that the host PMF is 'smooth' work well.

2. For the optimal LLRT, the threshold that minimize $aP(\text{Miss}) + (1 - a)P(\text{False alarm})$ for $a \in [0, 1]$ does not depend on the host.

With the above motivation, we propose to form an estimate $\hat{p}$ of the host PMF p and then form the decision statistic,

$$S(q) = D(q \| Q_R \hat{p}) - D(q \| \hat{p}).$$

We consider the following estimate for p; we have also tried other simple estimates, which are reported in our related paper [10]. For natural images the PMF is usually low pass. On the other hand, random LSB hiding introduces high frequency components in the histogram. Hence one simple estimate $\hat{p}$ is to pass the empirical PMF q though a low pass 2-tap FIR filter with taps $(0.5, 0.5)$. We note that normalization will be required after the filtering. We refer to this test as the approximate LLRT.

For a database of four thousand images from a DOQQ image set, we obtained the following results by simulation. In

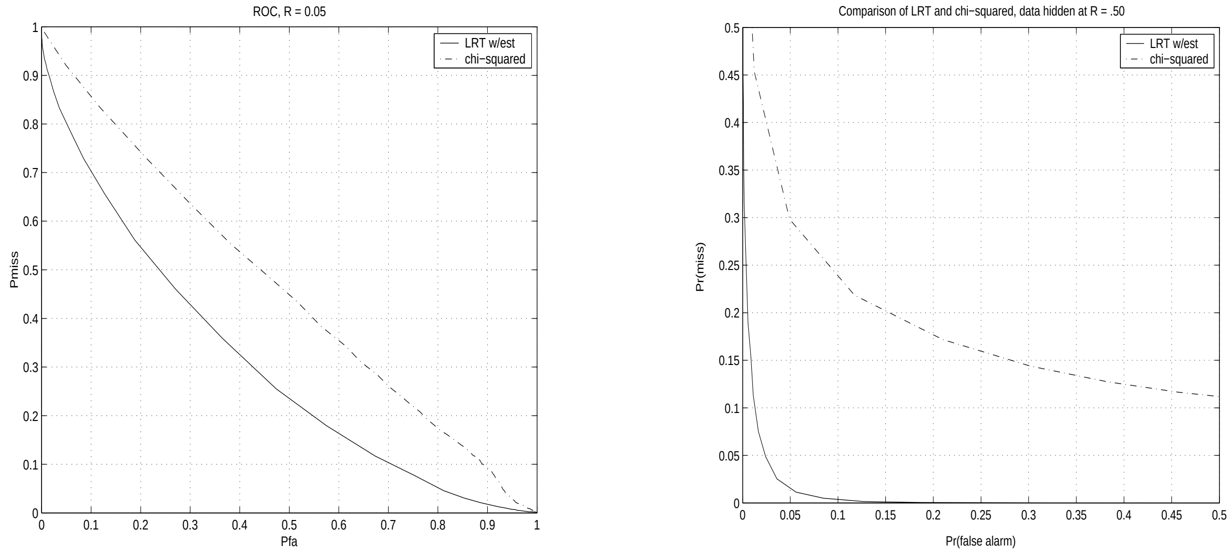

Figure 1: LLRT with half-half filter estimate versus Stegdetect: the approximate LLRT (with worst case rate $R = 0.05$) is superior at high as well as low actual hiding rates.

Figure 1 we compare the approximate LLRT test based on the half-half filter for estimating p with Stegdetect. For each point on the curve, the threshold has been fixed over the entire database. Clearly, our test outperforms Stegdetect for small as well as high rates. For the database of images we have used, the host PMF varies substantially from image to image. Thus these simulations suggest that Stegdetect is more sensitive to the choice of the threshold than our approximate LLRT test. This is not surprising since we know that to attain a target performance, the choice of the threshold in LLRT does not depend on the host PMF. For example, if we choose $T = 0$ for the approximate LLRT in the case when the hiding rate is 0.05, then we found the operating point to be $P(\text{Miss}) = 0.4043$ and $P(\text{False Alarm}) = 0.3219$. From Figure 1 we can verify that the tangent to the operating curve at this point is of slope approximately 1 as predicted by the theory.

V. Conclusion

We showed that under a mild smoothness assumption on the host PMF, the composite hypothesis testing problem associated with steganalysis is solved by the worst case simple hypothesis testing problem. We then demonstrated a practical test, which does not require the knowledge of the host PMF. This test comprehensively outperforms Stegdetect and is less sensitive to the choice of the threshold. More results can be found in our related paper [10].

While the results in this paper illustrate the power of the hypothesis testing approach to steganalysis, note that, by designing the test for i.i.d. host samples, we have thrown away information (e.g., regarding continuity of intensity values in typical images) that could be useful for steganalysis. An important area for future research is, therefore, to obtain compact mathematical models for such correlations to develop more powerful tests. More generally, since the hypothesis testing approach requires good models for the statistics under different hypotheses, modeling is the key challenge in applying this approach to the detection of various hiding strategies. Such modeling includes both the choice of model complexity (e.g., to what extent should correlations be modeled) and the estimation of the model parameters, which are typically unknown *a priori.*

A. Proof of Proposition 1

Consider the LLRT statistic

$$S_{LLRT}(q^{(N)}) = D(q^{(N)} \| p_{R_0}) - D(q^{(N)} \| p) = a^t q^{(N)},$$

where a is a column vector whose k^{th} entry is $\log(p_k / p_{R_0,k})$. In this proof we repeatedly use the following estimates for small R_0,

$$a_{2k} = R_0 \left(\frac{p_{2k} - p_{2k+1}}{2p_{2k}} \right) + O(R_0^2), \tag{3a}$$

$$a_{2k+1} = R_0 \left(\frac{p_{2k+1} - p_{2k}}{2p_{2k+1}} \right) + O(R_0^2). \tag{3b}$$

For simplicity, we represent $q^{(N)}$ by q. We note that the false alarm probability $P(S_{LLRT}(q) < T | H_0)$ depends only on p. Therefore to prove the result, we wish to show that for a given threshold T, $P(S_{LLRT}(q) > T | H_\theta)$ is decreasing for $R_0 \leq \theta \leq R_1$, so that the unit mass at R_0 is the least favorable distribution. To do so, we first obtain an approximation for $P(S_{LLRT}(q) > T | H_\theta)$ when θ is small and N is large. Let $F_N(t)$ be unity minus the distribution function of $S_{LLRT}(q)$. By the Berry-Esseen estimate for the rate of convergence in the central limit theorem ([11, Theorem 4.9, pp. 126]), under hypothesis H_θ,

$$\left| F_N(T) - Q\left(\sqrt{N} \frac{(T - \mu(\theta))}{\sigma(\theta)} \right) \right| \leq \frac{\text{constant}}{\sqrt{N}},$$

where the constant can be chosen independent of θ in our case. Now if $R_0, R_1 = O(1/\sqrt{N})$, then it is easy to check using

(3) that $\mu(\theta) = O(1/N)$ and $\sigma(\theta) = O(1/\sqrt{N})$. Therefore choosing $T = O(1/N)$, we get that

$$\gamma_N := \frac{\sqrt{N}(T - \mu(\theta))}{\sigma(\theta)} = O(1).$$

It follows that as $N \to \infty$, the ratio of $P(\text{Miss}|H_\theta) = F_N(T)$ and $Q(\gamma_N)$ approaches unity. In words, for sufficiently small R_0, R_1, and N sufficiently large, we can approximate $P(\text{Miss}|H_\theta)$ by $Q(\gamma_N)$. Since the approximation error is uniform in $\theta \in [R_0, R_1]$, to prove the theorem we now work with this approximation.

Since the Q-function is monotonically decreasing, we need to establish that $G(\theta) := \gamma_N/\sqrt{N} = (T - \mu(\theta))/\sigma(\theta)$ is increasing, that is, its derivative is non-negative for $\theta \in [R_0, R_1]$. Taking the derivative of $G(\theta)$,

$$G'(\theta) = -\frac{\mu'(\theta)\sigma^2(\theta) + (T - \mu(\theta))(\sigma^2(\theta))'}{\sigma^2(\theta)} =: -\frac{V(\theta)}{\sigma^2(\theta)}.$$

Our goal is to show that $V(\theta)$ is non-positive in the desired region. We begin by obtaining an expression for $V(\theta)$. Let w be the vector whose even components are of the form $p_{2k} - p_{2K+1}$ and whose odd components are of the form $p_{2k+1} - p_{2k}$. Then it is easy to see that $p_\theta = p - \theta w/2$, and hence $\mu(\theta) = \mu(0) + \beta\theta$, where $\beta = -a^t w/2$. Substituting for a and w we get that,

$$\beta = \frac{1}{2}\sum_{k=0}^{127} p_{2k}(r_k - 1)\log\left(\frac{(\theta/2) + (1-\theta/2)r_k}{(\theta/2)r_k^2 + (1-\theta/2)r_k}\right)$$

We note that each summand is negative, so that $\beta < 0$ and $\mu(\theta)$ is decreasing. Hence the restriction $P(\text{Miss}) \leq 0.5$ and $P(\text{False alarm}) \leq 0.5$ implies that

$$\mu(R_0) \leq T \leq \mu(0). \tag{4}$$

Similarly, we obtain,

$$\sigma^2(\theta) = \sigma^2(0) + b\theta + c\theta^2, \quad c = \beta^2$$
$$b = -\frac{1}{2}a^t(\text{diag}(w) + 2pw^t)a.$$

Putting $\gamma = \mu(0) - T$, we obtain,

$$V(\theta) = -\beta c\theta^2 - 2c\gamma\theta + (\beta\sigma^2(0) - \gamma b).$$

We note that $V'(\theta) = -2c(\mu(\theta) - T)$. Since $\beta < 0$, $\mu(\theta)$ is decreasing and hence $\mu(\theta) < \mu(R_0) \leq T$ from (4). Therefore $V'(\theta) \geq 0$, that is, $V(\theta)$ is increasing. Therefore it suffices to show that $V(1) \leq 0$, which is the same as the condition,

$$-\beta^3 + \beta\sigma^2(0) - \gamma(b + 2\beta^2) \leq 0, \text{ where } \gamma = \mu(0) - T \geq 0. \tag{5}$$

We note that

$$\begin{aligned} b &= -\frac{1}{2}a^t\text{diag}(w)a - (a^t p)\beta \\ &= -\frac{1}{2}\sum_{k=0}^{127}(p_{2k+1} - p_{2k})(a_{2k+1}^2 - a_{2k}^2) - D(p\|p_{R_0})\beta. \end{aligned}$$

Using (3), we know that $D(p\|p_{R_0}) = O(R_0^2)$, $\beta = O(R_0)$, and,

$$b = \frac{R_0^2}{8}\sum_{k=0}^{127}(p_{2k+1} - p_{2k})^4\frac{(p_{2k+1} + p_{2k})}{p_{2k}p_{2k+1}} + O(R_0^3).$$

Thus for R_0 sufficiently small, b is positive. Since $b + 2\beta^2$ is positive, to prove (5), we only need to show that $-\beta^3 + \beta\sigma^2(0) \leq 0$, that is, $\beta^2 \leq \sigma^2(0)$. Again by using (3), we have for $U(p)$ as in the statement of Proposition 1,

$$a = \frac{U(p)}{4}R_0^2 + O(R_0^4), \quad \beta^2 = \frac{U^2(p)}{4}R_0^2 + O(R_0^4).$$

Therefore for $U(p) < 1$, for R_0 sufficiently small, $\beta^2 \leq \sigma^2(0)$, and the proof is complete.

References

[1] N. Provos and P. Honeyman, "Detecting steganographic content on the internet," *ISOC NDSS'02*, San Diego, CA, 2002. See http://www.outguess.org/ for a reprint and source codes.

[2] http://hacktivismo.com/projects/camerashy/.

[3] J. Fridrich and M. Golijan, "Practical steganalysis of digital images - state of the art," in *Proceedings of SPIE*, 2002, vol. 4675.

[4] H. Farid, "Detecting stenographic messages in digital images," Tech. Rep., Dartmouth College, Computer Science, 2001.

[5] E. Lehmann, *Testing Statistical Hypothesis*, John Wiley, New York, 1959.

[6] H. V. Poor, *An introduction to signal detection and estimation*, Springer, NY, 1994.

[7] M.H.M. Costa, "Writing on dirty paper," *IEEE Trans. Info. Theory*, vol. IT-29, no. 3, pp. 439–441, May 1983.

[8] P. Moulin and J.A. O'sullivan, "Information-theoretic analysis of information hiding," preprint, Dec. 2001.

[9] N. Jacobsen, K. Solanki, U. Madhow, B. S. Manjunath a, and S. Chandrasekaran, "Image-adaptive high-volume data hiding based on scalar quantization," in *Proceedings of IEEE Military Communications Conference (MILCOM)*, Anaheim, CA, USA, October 2002.

[10] K. Sullivan, O. Dabeer, U. Madhow, B. S. Manjunath, and S. Chandrasekaran, "LLRT based detection of LSB hiding," submitted to ICIP 2003.

[11] R. Durrett, *Probability: Theory and examples*, Duxbury Press, Belmont, second edition, 1996.