RZ 3256 (# 93302) 07/10/00
Computer Science 15 pages

Research Report

An Information-Theoretic Model for Steganography

Christian Cachin*

IBM Research
Zurich Research Laboratory
8803 Rüschlikon
Switzerland

*Original work done at MIT Laboratory for Computer Science, supported by the Swiss National Science Foundation (SNF). Current address: IBM Research, Zurich Research Laboratory, 8803 Rüschlikon, Switzerland, cachin@acm.org.
A preliminary version of this work was presented at the 2nd Workshop on Information Hiding, Portland, USA, 1998, and appears in the proceedings (D. Aucsmith, ed., Lecture Notes in Computer Science, vol. 1525, Springer).

IBM Research
Almaden · Austin · Beijing · Delhi · Haifa · T.J. Watson · Tokyo · Zurich

An Information-Theoretic Model for Steganography

Christian Cachin*

IBM Research, Zurich Research Laboratory, 8803 Rüschlikon, Switzerland

Abstract

An information-theoretic model for steganography with passive adversaries is proposed. The adversary's task of distinguishing between an innocent cover message C and a modified message S containing a secret part is interpreted as a hypothesis testing problem. The security of a steganographic system is quantified in terms of the relative entropy (or discrimination) between P_C and P_S. It is shown that secure steganographic schemes exist in this model provided the covertext distribution satisfies certain conditions. A universal stegosystem is presented in this model for which the participants do not have to know the covertext distribution except that it consists of a series of independent experiments.

*Original work done at MIT Laboratory for Computer Science, supported by the Swiss National Science Foundation (SNF). Current address: IBM Research, Zurich Research Laboratory, 8803 Rüschlikon, Switzerland, `cachin@acm.org`.

A preliminary version of this work was presented at the 2nd Workshop on Information Hiding, Portland, USA, 1998, and appears in the proceedings (D. Aucsmith, ed., Lecture Notes in Computer Science, vol. 1525, Springer).

1 Introduction

Steganography is the art and science of hiding information such that its presence cannot be detected. Motivated by growing concern about the protection of intellectual property on the Internet and by the threat of a ban for encryption technology, interest in techniques for information hiding has been rising across the recent years [And96]. Two general directions can be distinguished within information hiding scenarios: protection only against the detection of a message by a passive adversary and hiding a message such that not even an active adversary can remove it. A survey of current steganography is given by Petitcolas et al. [PAK99].

Steganography with a *passive* adversary is perhaps best illustrated by Simmons' "Prisoners' Problem" [Sim84]. Alice and Bob are in jail and wish to devise an escape plan. All their communication is observed by the adversary (the warden), who will thwart their plan by transferring them to a high-security prison as soon as he detects any sign of a hidden message. Alice and Bob succeed if Alice can send information to Bob such that Eve does not become suspicious.

Hiding information from *active* adversaries is a different problem since the existence of a hidden message is publicly known, as for example in copyright protection schemes. Steganography with active adversaries can be divided into watermarking and fingerprinting. Watermarking supplies digital objects with an identification of origin; all objects are marked in the same way. Fingerprinting, conversely, attempts to identify individual copies of an object by means of embedding a unique marker in every copy that is distributed. If later an illegal copy is found, the copyright owner can identify the buyer by decoding the hidden information ("traitor tracing") [NFC94, PS96].

Since most objects to be protected by watermarking or fingerprinting consist of audio or image data, these data types have received most attention so far. A number of generic hiding techniques have been developed whose effects are barely perceptible for humans but can withstand tampering by data transformations that essentially conserve its contents [CKLS96, BGML96].

A common model and terminology for information hiding has been established at the 1996 Information Hiding Workshop [Pfi96]. An original, unaltered message is called covertext; the sender Alice tries to hide an embedded message by transforming the covertext using a secret key. The resulting message is called the stegotext and is sent to the receiver Bob. Similar to cryptography, it is assumed that the adversary Eve has complete information about the system except for a secret key shared by Alice and Bob that guarantees the security. However, the model does not include a formal notion of security.

Our Approach. In this paper, we propose that steganography with a passive adversary is a problem of *hypothesis testing* and introduce a corresponding information-theoretic notion of security. Upon observing a message sent by Alice, the adversary has to decide whether it is an original covertext C or contains an embedded message and is a stegotext S. This is the problem of distinguishing two different explanations for the observed data that is investigated in statistics and in information theory as "hypothesis testing." We follow the non-Bayesian approach between statistics and information theory advocated by Blahut [Bla87], based on the relative entropy function as the basic measure of the information contained in an observation. Thus, we use the relative entropy $D(P_C \| P_S)$ between P_C and P_S to quantify the security of a steganographic system (or stegosystem for short) against passive attacks. If the covertext and stegotext distributions are equal and $D(P_C \| P_S) = 0$, the stegosystem is perfectly secure and the adversary can have no advantage over merely guessing without observing a message. This parallels Shannon's notion of perfect secrecy for cryptosystems [Sha49].

However, some caution has to be exerted using this model: On one hand, information-theoretic methods have been applied with great success to the problems of information encoding and transmission, starting with Shannon's pioneering work [Sha48]. Messages to be transmitted are modeled as random processes and systems developed in this model perform well in practice, which can easily be verified. For information hiding, on the other hand, the relation between the model and its validity is more involved. A message encrypted under a one-time pad, for example, is indistinguishable from uniformly random bits and this is a perfectly secure stegosystem according to our notion of security. But no warden would allow the prisoners to use one-time pad encryption! Thus, the crucial issue for the validity of a formal treatment of steganography is the accuracy of the model for real data.

Nevertheless, we believe that our model provides insight in steganography. We hope that it also lays ground for further work to formalize active adversaries or computational security. A model for active adversaries is presented by Ettinger [Ett98] and uses a game-theoretic approach. A direct extension of our approach would be to model the covertext source as a stochastic process and consider statistical estimation and decision techniques.

Related Work. Other information-theoretic treatments of steganography have been developed by Zöllner et al. [ZFK$^+$98] and by Mittelholzer [Mit99]. A discussion of their models with respect to ours is included in Section 7. Another related work is a paper by Maurer [Mau96] on unconditionally secure authentication [Mas91], which shows how Simmons' bound [Sim85] and many other lower bounds in authentication theory can be derived and generalized using the powerful tools of hypothesis testing.

Organization of the Paper. Hypothesis testing is presented in Section 2 from an information-theoretic viewpoint. Section 3 contains the formal description of the model and the security definition. In Section 4, we provide some examples of unconditionally secure stegosystems. A universal information hiding scheme that requires no knowledge of the covertext statistics is presented in Section 5. Some extensions of the approach are sketched in Section 6 and the paper concludes with a discussion.

2 Review of Hypothesis Testing

We give a brief review of hypothesis testing and information-theoretic notions (cf. [Bla87, CT91]). Logarithms are to the base 2. The cardinality of a set $\mathcal{S}$ is denoted by $|\mathcal{S}|$. The *entropy* of a random variable X with probability distribution P_X and alphabet $\mathcal{X}$ is defined as

$$H(X) \;=\; -\sum_{x \in \mathcal{X}} P_X(x) \log P_X(x).$$

The *conditional entropy* of X conditioned on a random variable Y is

$$H(X|Y) \;=\; \sum_{y \in \mathcal{Y}} P_Y(y) H(X|Y=y)$$

where $H(X|Y=y)$ denotes the entropy of the conditional probability distribution $P_{X|Y=y}$. The *mutual information* between X and Y is defined as the reduction of entropy that Y provides about X, i.e., $I(X;Y) = H(X) - H(X|Y)$.

Hypothesis testing is the task of deciding which one of two hypotheses H_0 or H_1 is the true explanation for an observed measurement Q. In other words, there are two plausible probability distributions, denoted by P_{Q_0} and P_{Q_1}, over the space $\mathcal{Q}$ of possible measurements. If H_0 is true, then Q was generated according to P_{Q_0}, and if H_1 is true, then Q was generated according to P_{Q_1}. A *decision rule* is a binary partition of $\mathcal{Q}$ that assigns one of the two hypotheses to each possible measurement $q \in \mathcal{Q}$. The two errors that can be made in a decision are called a *type I error* for accepting hypothesis H_1 when H_0 is actually true and a *type II error* for accepting H_0 when H_1 is true. The probability of a type I error is denoted by α, the probability of a type II error by β.

A method for finding the optimum decision rule is given by the Neyman-Pearson theorem. The decision rule is specified in terms of a threshold parameter T; α and β are then functions of T. The theorem states that for a threshold $T \in \mathbb{R}$ and fixed maximal tolerable probability β of type II error, α can be minimized by assuming hypothesis H_0 for an observation $q \in \mathcal{Q}$ if and only if

$$\log \frac{P_{Q_0}(q)}{P_{Q_1}(q)} \;\geq\; T. \tag{1}$$

In general, many values of T must be examined to find the optimal decision rule. The term on the left hand side in (1) is called the *log-likelihood ratio.*

Relative Entropy. An important information measure in hypothesis testing is the *relative entropy* or *discrimination* between two probability distributions P_{Q_0} and P_{Q_1}, defined as

$$D(P_{Q_0} \| P_{Q_1}) \;=\; \sum_{q \in \mathcal{Q}} P_{Q_0}(q) \log \frac{P_{Q_0}(q)}{P_{Q_1}(q)} \tag{2}$$

(with the standard conventions that $0 \log 0 = 0 \log \frac{0}{0} = 0$ and $p \log \frac{p}{0} = \infty$ if $p > 0$).

The *conditional relative entropy* between P_{Q_0} and P_{Q_1} given a random variable V is defined as

$$D(P_{Q_0|V} \| P_{Q_1|V}) = \sum_{v \in \mathcal{V}} P_V(v) \sum_{q \in \mathcal{Q}} P_{Q_0|V=v}(q) \log \frac{P_{Q_0|V=v}(q)}{P_{Q_1|V=v}(q)}. \tag{3}$$

The relative entropy between two distributions is always nonnegative and is 0 if and only if the distributions are equal. Although relative entropy is not a true distance measure in the mathematical sense because it is not symmetric and does not satisfy the triangle inequality, it can be useful to think of it as a distance. The binary relative entropy $d(\alpha, \beta)$ is

$$d(\alpha, \beta) \;=\; \alpha \log \frac{\alpha}{1-\beta} + (1-\alpha) \log \frac{1-\alpha}{\beta}.$$

Relative entropy and hypothesis testing are linked through the Neyman-Pearson theorem above because the expected value of the log-likelihood ratio in (1) with respect to P_{Q_0} is equal to the relative entropy $D(P_{Q_0} \| P_{Q_1})$. The following standard result shows that deterministic processing cannot increase the relative entropy between two distributions.

Lemma 1. *Let P_{Q_0} and P_{Q_1} be probability distributions over $\mathcal{Q}$. For any function $f : \mathcal{Q} \to \mathcal{T}$, let $T_0 = f(Q_0)$ and $T_1 = f(Q_1)$. Then*

$$D(P_{T_0} \| P_{T_1}) \;\leq\; D(P_{Q_0} \| P_{Q_1}).$$

Because deciding between H_0 and H_1 is a special form of processing by a binary function, the type I and type II error probabilities α and β satisfy

$$d(\alpha,\beta) \;\leq\; D(P_{Q_0}\|P_{Q_1}). \tag{4}$$

This bound is typically used as follows: Suppose that $D(P_{Q_0}\|P_{Q_1}) < \infty$ and that there is a given upper bound on the type I error probability α. Then (4) yields a lower bound on the type II error probability β. For example, $\alpha = 0$ implies that $\beta \geq 2^{-D(P_{Q_0}\|P_{Q_1})}$.

If an experiment is repeated independently n times, an appropriate statistical test will cause the errors to decrease exponentially in n. Stein's Lemma, an asymptotic version of (4), shows that for a fixed upper bound on the type I error probability, the exponent of the type II error probability achieves $D(P_{Q_0}\|P_{Q_1})$ but cannot be made better.

Lemma 2 (Stein's Lemma). *Let $X_1,\ldots,X_n$ be independent and identically distributed according to P_X and consider the hypothesis test between the alternatives $P_X = P_{Q_0}$ or $P_X = P_{Q_1}$. For given α, let β_n^* be the smallest achievable type II error probability over all decision rules with the property that the type I error does not exceed α. Then*

$$\lim_{n\to\infty} \frac{1}{n} \log \beta_n^* = -D(P_{Q_0}\|P_{Q_1}).$$

Stronger results of this type can be shown, see the survey by Csiszár [Csi98]; it is also possible to weaken the independence assumption from memoryless sources to finite Markov chains [Nat85].

Allowing Side Information. The case is similar for a generalized hypothesis testing scenario, where the distributions P_{Q_0} and P_{Q_1} depend on knowledge of an additional random variable V. The probability distributions, the decision rule, and the error probabilities are now parameterized by V. In other words, the probability distributions are $P_{Q_0|V=v}$ and $P_{Q_1|V=v}$ for all $v \in \mathcal{V}$, the decision rule may depend on the value v of V, and the error probabilities are $\alpha(v)$ and $\beta(v)$ for each $v \in \mathcal{V}$. Let the average type I and type II errors be $\overline{\alpha} = \sum_{v\in\mathcal{V}} P_V(v)\alpha(v)$ and $\overline{\beta} = \sum_{v\in\mathcal{V}} P_V(v)\beta(v)$. It follows from the Jensen inequality and from (4) that

$$d(\overline{\alpha},\overline{\beta}) \;\leq\; D(P_{Q_0|V}\|P_{Q_1|V}). \tag{5}$$

Useful properties. We note the following two properties of relative entropy. The first one connects entropy, relative entropy, and the size of the alphabet for any random variable $X \in \mathcal{X}$: If P_U is the uniform distribution over $\mathcal{X}$, then

$$H(X) + D(P_X\|P_U) \;=\; \log|\mathcal{X}|. \tag{6}$$

The second one states that conditioning on derived information (with equal distribution) can only increase the discrimination: If there is a deterministic function $f : \mathcal{Q} \to \mathcal{V}$ such that the random variables $f(Q_0)$ and $f(Q_1)$ have the same distribution P_V, then [Bla87]

$$D(P_{Q_0}\|P_{Q_1}) \;\leq\; D(P_{Q_0|V}\|P_{Q_1|V}). \tag{7}$$

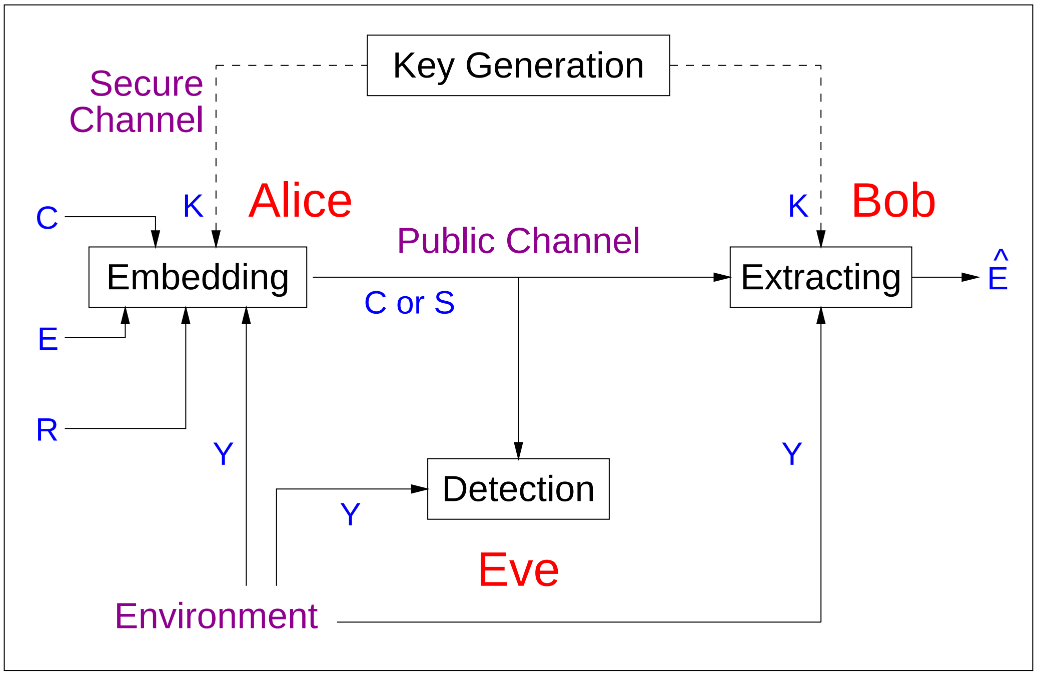

Figure 1: The model of a secret-key stegosystem with passive adversary. It shows the environment Y, the embedded message E, the covertext C, the stegotext S, Alice's private random source R, and the secret key K shared by Alice and Bob. Alice is either sending covertext C or stegotext S.

3 Model

Setting. Figure 1 shows our model of a stegosystem. Assume for the moment that the environment Y is fixed. Eve observes a message that is sent from Alice to Bob. She does not know whether Alice sends legitimate *covertext* C or *stegotext* S containing hidden information for Bob. We model this by letting Alice operate strictly in one of two modes: either she is active and her output is S or she is inactive and sends covertext C.

If Alice is active, she transforms C to contain an *embedded message* E using a *secret key* K. (Alternatively, Alice could also generate C herself.) Alice may use a *private random source* R for embedding. The output of the hiding process is the stegotext S. Bob's decoder outputs an estimate $\hat{E}$ for E using his knowledge of the stegotext S and from the key K; in order for the stegosystem to be effective, $\hat{E}$ must provide information about E.

Expressed in terms of entropy, the system satisfies:

1. $H(S|CEKRY) = 0$. The stegotext is determined uniquely by Alice's inputs.
2. $H(E|Y) > 0$. There is uncertainty about the embedded message.
3. $I(E;\hat{E}|SKY) > 0$. Bob must be able to get information about the embedded message.

If Alice is inactive, she sends covertext C and no embedding takes place. The embedding mechanism, E, K, and R can be thought of as absent.

It may be that C consists of multiple messages sent from Alice to Bob. We explicitly address the case where covertext and stegotext consist of *independently* repeated experiments; here, Alice is either active or passive in all repetitions. Otherwise, if Alice sends multiple *dependent* messages to Bob and at least one of them contains hidden information, she is considered active and S consists of the concatenation of all her messages.

The probability distributions are assumed to be known to all parties if not stated otherwise. In addition, Bob knows whether Alice is active or not.

Eve, upon observing the message sent by Alice, has to decide whether it was generated according to the distribution of the innocent covertext C or according to the modified distribution of the stegotext S, i.e., whether Alice is active. Since this task is a hypothesis testing problem, we quantify the security of a stegosystem in terms of the relative entropy distance between P_C and P_S.

Definition 1. A stegosystem as introduced above with covertext C and stegotext S is called *ϵ-secure against passive adversaries* if

$$D(P_C \| P_S) \;\leq\; \epsilon.$$

If $\epsilon = 0$, the stegosystem is called *perfectly secure.*

When covertext C and stegotext S consist of n *independently repeated experiments*, the security is measured in terms of the *normalized* relative entropy between P_C and P_S and the stegosystem is said to be ϵ-secure against passive adversaries whenever

$$\frac{1}{n} D(P_C \| P_S) \;\leq\; \epsilon.$$

It is sometimes appropriate to relax the above model of the embedding process and allow for a deterministic processing of C by Alice, resulting in an encoding Z, before the actual embedding takes place. In this case, Z is sent over the public channel and the relevant quantity is the relative entropy between P_Z and P_S. If C consists of a sequence of independent random variables, we still use normalized relative entropy even if Z has arbitrary distribution.

Bounds on Detection Performance. Consider Eve's decision process for a particular decision rule, given by a binary partition $(\mathcal{C}_0, \mathcal{C}_1)$ of the set $\mathcal{C}$ of possible covertexts. She decides that Alice is active if and only if the observed message c is contained in $\mathcal{C}_1$. Ideally, she would always detect a hidden message. (But this occurs only if Alice chooses an encoding such that valid covertexts and stegotexts are disjoint.) If Eve fails to detect that she observed stegotext S, she makes a type II error; its probability is denoted by β.

The opposite error, which usually receives less attention, is the type I error: Eve decides that Alice sent stegotext although it was a legitimate cover message C; this probability is denoted by α. An important special case is that Eve makes no type I error and never accuses Alice of sending hidden information when she is inactive ($\alpha = 0$). Such a restriction might be imposed on Eve by external mechanisms, justified by the desire to protect innocent users.

Lemma 1 imposes a bound on the achievable error probabilities by Eve. From (4) we obtain the following result.

Proposition 3. *In a stegosystem that is ϵ-secure against passive adversaries, the probability β that the adversary does not detect the presence of the embedded message and the probability α that the adversary falsely announces the presence of an embedded message satisfy*

$$d(\alpha, \beta) \leq \epsilon.$$

In particular, if $\alpha = 0$, then

$$\beta \;\geq\; 2^{-\epsilon}.$$

In a perfectly secure system we have $D(P_C \| P_S) = 0$ and therefore $P_C = P_S$; thus, Eve can obtain no information about whether Alice is active by observing the message.

Moreover, Stein's lemma can be applied to a stegosystem that consists of n independent repetitions of an ϵ-secure stegosystem (i.e., C and S are memoryless sources and Alice is either always active or always passive). It follows that the performance of an optimal statistical test by Eve is determined by the discrimination between the covertext and the stegotext and the type II error exponent is bounded by $-\epsilon$.

Proposition 4. *Let $C = (C_1, \dots, C_n)$ and $S = (S_1, \dots, S_n)$ denote the covertext and stegotext of an ϵ-secure stegosystem that consist of independently repeated experiments. Then for any fixed bound α on the probability that the adversary falsely detects an embedded message, the smallest achievable error probability β_n^* of not detecting an embedded message satisfies ${\beta_n^*}^{1/n} \geq 2^{-\epsilon - o(n)}$.*

An Example. Suppose Alice is given a digital image m that she is permitted to send to Bob. Using a perceptional model, she has determined a set $\mathcal{M}$ of equivalent images that are visually indistinguishable from m. Regardless of whether Alice is active or not, she will send a randomly chosen element of $\mathcal{M}$ and this defines the probability space underlying C. Note that in our model, the adversary knows at least $\mathcal{M}$ and possibly also m. Alice can use the techniques described below for embedding information; however, to achieve robustness against active adversaries who modify the image, more sophisticated coding methods are necessary, e.g. [CKLS96, BS98].

Incorporating the Environment. It may be the case that external events influence the covertext distribution; for example, a news report or the local weather if we think of the prisoners' problem. This external information is denoted by Y and known all participants. Our model and the security definition above are then as follows. All quantities involved are conditioned on knowledge of Y and we consider the average error probabilities $\overline{\alpha} = \sum_{y \in \mathcal{Y}} P_Y(y)\alpha(y)$ for the type I error and $\overline{\beta} = \sum_{y \in \mathcal{Y}} P_Y(y)\beta(y)$ for the type II error, where $\alpha(y)$ and $\beta(y)$ denote the type I and type II error probabilities for $Y = y$, respectively.

Definition 2. A stegosystem with external information Y, covertext C, and stegotext S is called *ϵ-secure against passive adversaries* if

$$D(P_{C|Y} \| P_{S|Y}) \;\leq\; \epsilon.$$

It follows from (5) that the average error probabilities satisfy $d(\overline{\alpha}, \overline{\beta}) \leq \epsilon$, similar to Proposition 3.

We now show that perfectly secure stegosystems exist for particular sources of covertext. We start with especially simple (or unrealistic) covertext distributions and consider arbitrary and unknown covertext statistics later.

4 Unconditionally Secure Stegosystems

The above model tells us that we obtain a secure stegosystem whenever the stegotext distribution is close to the covertext distribution without knowledge of the key. The embedding function depends crucially on knowledge about the covertext source. We assume first that the covertext distribution is known and design corresponding embedding functions.

One-Time Pad. If the covertext consists of independent and uniformly random bits, then the one-time pad provides a perfectly secure stegosystem. For completeness, we briefly describe this system formally.

Assume the covertext C is a uniformly distributed n-bit string for some positive n. The key generator chooses the n-bit key K with uniform distribution and sends it to Alice and Bob. The embedding function (if Alice is active) consists of the bitwise XOR of the particular n-bit message e and K, thus $S = e \oplus K$, and Bob can decode by computing $e = S \oplus K$. The resulting stegotext S is uniformly distributed in the set of n-bit strings and therefore $D(P_C \| P_S) = 0$. Thus, the one-time pad provides perfect steganographic security if the covertext is uniformly random.

As a side remark, we note that this one-time pad system is equivalent to the basic scheme of visual cryptography [NS95]. This technique hides a monochrome picture by splitting it into two random layers of dots. When these are superimposed, the picture appears. It is also possible to produce two innocent looking pictures such that both of them together reveal an embedded message.

General Distributions. For arbitrary covertext distributions, we now describe a system that embeds a one-bit message in the stegotext as an example. The extension to larger message spaces is straightforward, but requires even more accurate knowledge of the covertext distribution. Let the covertext C with alphabet $\mathcal{C}$ have distribution P_C. Alice constructs the embedding function from a partition of $\mathcal{C}$ into two parts such that both parts are assigned approximately the same probability under C. In other words, let

$$\mathcal{C}_0 \;=\; \min_{\mathcal{C}' \subseteq \mathcal{C}} \left| \sum_{c \in \mathcal{C}'} P_C(c) - \sum_{c \notin \mathcal{C}'} P_C(c) \right| \qquad \text{and} \qquad \mathcal{C}_1 \;=\; \mathcal{C} \setminus \mathcal{C}_0.$$

Alice and Bob share a one-bit key $K \in \{0, 1\}$. Define C_0 to be the random variable with alphabet $\mathcal{C}_0$ and distribution P_{C_0} equal to the conditional distribution $P_{C|C \in \mathcal{C}_0}$ and define C_1 similarly over $\mathcal{C}_1$. Then Alice computes the stegotext to embed a message $e \in \{0, 1\}$ as

$$S = C_{e \oplus K}.$$

Bob can decode the message because he knows that $e = 0$ if and only if $S \in \mathcal{C}_K$.

Theorem 5. *The one-bit message stegosystem described above is*

$$\frac{1}{\ln 2} \bigl(\Pr[C \in \mathcal{C}_0] - \Pr[C \in \mathcal{C}_1] \bigr)^2$$

secure against passive adversaries.

Proof. Let $\delta = \Pr[C \in \mathcal{C}_0] - \Pr[C \in \mathcal{C}_1]$. We show only the case $\delta > 0$. It is straightforward to verify that

$$P_S(c) \;=\; \begin{cases} P_C(c)/(1+\delta) & \text{if } c \in \mathcal{C}_0, \\ P_C(c)/(1-\delta) & \text{if } c \in \mathcal{C}_1. \end{cases}$$

It follows that

$$\begin{aligned}
D(P_C \| P_S) &= \sum_{c \in \mathcal{C}} P_C(c) \log \frac{P_C(c)}{P_S(c)} \\
&= \sum_{c \in \mathcal{C}_0} P_C(c) \log(1+\delta) + \sum_{c \in \mathcal{C}_1} P_C(c) \log(1-\delta) \\
&= \frac{1+\delta}{2} \cdot \log(1+\delta) + \frac{1-\delta}{2} \cdot \log(1-\delta) \\
&\leq \frac{1+\delta}{2} \cdot \frac{\delta}{\ln 2} + \frac{1-\delta}{2} \cdot \frac{-\delta}{\ln 2} \\
&= \delta^2 / \ln 2
\end{aligned}$$

using the fact that $\log(1+x) \leq x/\ln 2$. □

A remark on data compression techniques. Suppose the embedding as described above takes place before compression is applied to S (or C). Data compression is a deterministic process. Therefore, Lemma 1 applies and shows that if we start with an ϵ-secure stegosystem, the security of the compressed system is also at most ϵ. To put it another way, data compression can never hurt the security of a stegosystem and it does not make detection any easier for the adversary.

5 Steganography with Universal Data Compression

The stegosystems described in Section 4 assume that the covertext distribution is known to all parties. This seems not realistic for many applications. However, if we allow our covertext data to consist of independent repetitions of the same experiment, we can apply universal data compression algorithms that do not suffer from this limitation. We show how they can be modified for steganography and illustrate this for a particularly simple algorithm.

Traditional data compression techniques, such as Huffman coding, require a priori knowledge about the distribution of the data to be compressed. Universal data compression algorithms treat the problem of source coding for applications where the source statistics are a priori unknown or vary with time. A universal data compression algorithm achieves asymptotically optimal performance on every source in some large class of sources, characterized by an ergodicity condition. This is accomplished by learning the statistics of the data during operation as more and more data is observed. The best known examples of universal data compression are the practical algorithms by Lempel and Ziv [ZL77, WZW98, BCW90].

Throughout this section we assume the environment Y is fixed.

The Method of Types. One of the fundamental concepts of information theory is the *method of types* [CK81, Csi98]. It leads to simple proofs for the *asymptotic equipartition property (AEP)* and many other important results. The AEP states that the set of possible outcomes of n independent, identically distributed realizations of a random variable X can be divided into a typical set and a non-typical set, and that the probability of the typical set approaches 1 with $n \to \infty$. Furthermore, all typical sequences are almost equally likely and the probability of a typical sequence is close to $2^{-nH(X)}$.

Let x^n be a sequence of n symbols from $\mathcal{X}$. The *type* or empirical probability distribution U_{x^n} of x^n is the mapping that specifies the relative proportion of occurrences of each symbol $x_0 \in \mathcal{X}$ in x^n, i.e., $U_{x^n}(x_0) = \frac{N_{x_0}(x^n)}{n}$, where $N_{x_0}(x^n)$ is the number of times that x_0 occurs in the sequence x^n. The *set of types with denominator* n is denoted by $\mathcal{U}_n$ and for $U \in \mathcal{U}_n$, the *type class* $\{x^n \in \mathcal{X}^n : U_{x^n} = U\}$ is denoted by $\mathcal{T}(U)$.

The following standard result summarizes the basic properties of types.

Theorem 6 ([CK81, CT91]). *Let $X^n = X_1, \ldots, X_n$ be a sequence of n independent and identically distributed random variables with distribution P_X and alphabet $\mathcal{X}$ and let $\mathcal{U}_n$ be the set of types. Then*

1. *The number of types with denominator n is at most polynomial in n, more particularly $|\mathcal{U}_n| \leq (n+1)^{|\mathcal{X}|}$.*

2. *The probability of a sequence x^n depends only on its type and is given by $P_{X^n}(x^n) = 2^{-n(H(U_{x^n})+D(U_{x^n}\|P_X))}$.*

3. *For any $U \in \mathcal{U}_n$, the size of the type class $\mathcal{T}(U)$ is approximately $2^{nH(U)}$. More precisely, $\frac{1}{(n+1)^{|\mathcal{X}|}} 2^{nH(U)} \leq |\mathcal{T}(U)| \leq 2^{nH(U)}$.*

4. *For any $U \in \mathcal{U}_n$, the probability of the type class $\mathcal{T}(U)$ is approximately $2^{-nD(U\|P_X)}$. More precisely, $\frac{1}{(n+1)^{|\mathcal{X}|}} 2^{-nD(U\|P_X)} \leq \Pr[X^n \in \mathcal{T}(U)] \leq 2^{-nD(U\|P_X)}$.*

The Data Compression Algorithm. A universal coding scheme for a memoryless source X works as follows. Fix a rate $R < \log|\mathcal{X}|$ and let $R_n = R - |\mathcal{X}| \frac{\log(n+1)}{n}$. Define a set of sequences $A_n = \{x^n \in \mathcal{X}^n : H(U_{x^n}) \leq R_n\}$. The block code is given by an enumeration $0, \ldots, M-1$ of the elements of A_n. In other words, the encoder maps a sequence X^n to a codeword Z if the entropy of the *type* of X^n does not exceed R_n and to a default value Δ otherwise. Using the concept of types it is very easy to show using the first and the third property of types that $|A_n| \leq 2^{nR}$ and therefore $\lceil nR \rceil$ bits are sufficient to encode all $x^n \in A_n$ [CK81, CT91]. Moreover, if $H(X) < R$ then values outside A_n occur only with exponentially small probability and the error probability $p_e^{(n)} = \Pr[Z = \Delta]$ satisfies

$$p_e^{(n)} \;\leq\; (n+1)^{|\mathcal{X}|} 2^{-n \min_{U: H(U) > R_n} D(U\|P_X)}. \tag{8}$$

The following observation is needed below. Because codewords can be decoded uniquely, we have

$$H(X^n|Z) \;=\; \Pr[Z \neq \Delta] H(X^n|Z) + \Pr[Z = \Delta] H(X^n|Z) \;\leq\; p_e^{(n)} n H(X).$$

Together with $H(X^n) = H(X^n Z) = H(Z) + H(X^n|Z)$, it follows

$$H(Z) \;\geq\; nH(X)(1 - p_e^{(n)}). \tag{9}$$

A Universal Information Hiding Scheme. Suppose the covertext consists of n independent realizations of a random variable X. The universal hiding scheme applies data compression as described above and possibly embeds hidden information if Alice is active.

Given R and n, Alice maps the incoming covertext X^n to its index $Z \in \{0, \ldots, M-1\}$. W.l.o.g. assume Z is a binary m-bit string and $m = \lceil \log M \rceil$; further, let the key K be an ℓ-bit string and the particular message e to be embedded an ℓ-bit string with $\ell \leq m$.

If Alice is active, she outputs $S = Z \oplus (e \oplus K \| 0^{m-\ell})$, otherwise she outputs $C = Z$ unmodified (where $\|$ denotes the concatenation of bit strings).

Bob recovers the embedded message as $\hat{E} = Z \oplus K \| 0^{m-\ell}$.

Theorem 7. *Let covertext $C = (X_1, \ldots, X_n)$ consist of n identical and independently repeated experiments with distribution P_X and let $\epsilon > 0$. Then the algorithm above implements a universal stegosystem that is ϵ-secure against passive adversaries and hides an ℓ-bit message with $\ell \leq nH(X)$ for n sufficiently large.*

Proof. It is clear from the description that the scheme satisfies the first two conditions of a stegosystem. It remains to show that it is ϵ-secure and that $\hat{E}$ provides information about E.

Let $R = H(X) + \epsilon/2$. Then

$$m \;=\; \lceil nR \rceil \;\leq\; \lceil nH(X) + n\epsilon/2 \rceil. \tag{10}$$

The codeword Z is distributed according to the output of the compression algorithm. Let $T = Z_{[m-\ell+1,\ldots,m]}$ denote the suffix of Z of length $m - \ell$ bits. Because S has the same distribution as Z with the exception that it is uniform over all strings with a particular $m - \ell$-bit suffix, we have for all s

$$P_S(s) \;=\; 2^{-\ell}\, P_T(s_{[m-\ell+1,\ldots,m]}). \tag{11}$$

We now derive a bound on $D(P_Z \| P_S)$ according to the remarks after Definition 1. Since T has the same distribution regardless of whether Alice is active or not, we can apply (7), followed by (11) and (6) to obtain

$$\begin{aligned} D(P_Z \| P_S) &\leq D(P_{Z|T} \| P_{S|T}) \\ &= \sum_t P_T(t) D(P_{Z|T=t} \| P_{S|T=t}) \\ &= \sum_t P_T(t) \big(\ell - H(Z|T=t)\big) \\ &= \ell - H(Z|T). \end{aligned} \tag{12}$$

Because each Z uniquely determines its suffix T, which is an $\ell - m$-bit value, we have

$$H(Z|T) \;=\; H(ZT) - H(T) \;=\; H(Z) - H(T) \;\geq\; H(Z) - (m - \ell). \tag{13}$$

Combining (12) and (13) gives

$$D(P_Z \| P_S) \;\leq\; m - H(Z). \tag{14}$$

Now insert (10) and (9) into (14) to obtain

$$\begin{aligned} \frac{1}{n} D(P_Z \| P_S) &\leq \frac{1}{n}\Big(\lceil nH(X) + n\epsilon/2 \rceil - nH(X)(1 - p_e^{(n)}) \Big) \\ &\leq \frac{1}{n}\big(p_e^{(n)} nH(X) + n\epsilon/2 + 1 \big) \\ &= p_e^{(n)} H(X) + \epsilon/2 + \frac{1}{n}. \end{aligned}$$

Since R_n approaches R from below and $R > H(X)$, it follows that for all sufficiently large n, also $R_n > H(X)$ and the value $\min_{U: H(U) > R_n} D(U \| P_X)$ in the exponent in (8) is strictly positive. This implies that the last expression is smaller than ϵ for all sufficiently large n and that the stegosystem is indeed ϵ-secure.

It is easy to see that Bob recovers E from $\hat{E}$ whenever $Z \neq \Delta$, which occurs with probability at least $1 - p_e^{(n)}$. Thus, the probability of a decoding error is exponentially small by (8) and the remark above. □

6 Extensions

The presented information-theoretic model can be considered as one particular example of a statistical model. Other methods from statistics seem useful for a formal treatment of steganography as well. If one introduces valuations for the possible decisions, tools from statistical decision theory can be applied and allow for reasoning about the cost of the involved actions [Ber85]. As noted before, another extension more on the grounds of information theory would be to model the covertext source as an ergodic process.

Simmons' original scenario of the prisoners' problem includes authentication, that is, the secret key K shared by Alice and Bob is partially used for authenticating Alice's messages. The reason for this is that Alice and Bob want to protect themselves (and are allowed to do so) from a malicious warden that tries to fool Bob into accepting fraudulent messages as originating from Alice. This implies some changes to the model. Denote the part of the key used for authentication by V. Then, for every value v of V, there is a different covertext distribution $P_{C|V=v}$ induced by the authentication scheme in use. However, since the adversary Eve does not know V, the covertext distribution to consider for detection is P_C as the marginal distribution induced by P_{CV}. Note that this model differs from the general scenario with an active adversary; there, the adversary succeeds if she can destroy the embedded hidden information (as is the case in copyright protection applications, for example). Here, the prisoners are only concerned about hiding information in messages that may be authenticated to detect tampering.

7 Discussion

The approach of this paper is to view steganography with a passive adversary as a problem of hypothesis testing because the adversary succeeds if he merely detects the presence of hidden information.

Other information-theoretic models for steganography have been proposed in the literature and take a slightly different view:

- Zöllner et al. [ZFK$^+$98] correctly recognize that breaking a steganographic system means detecting the use of steganography to embed a message. However, they formally require only that knowledge of the stegotext does not decrease the uncertainty about an embedded message, similar to Shannon's notion of perfect secrecy for cryptosystems.

- Mittelholzer [Mit99] considers steganography (with a passive adversary) and watermarking (with an active adversary). A stegosystem is required to provide perfect secrecy for the embedded message in sense of Shannon, and an encoder constraint is imposed in terms of a distortion measure between covertext and stegotext. The expected mean squared error is proposed as a possible distortion measure.

Although these conditions may be necessary, they are not sufficient to guarantee undetectable communication, as can be seen from the following stegosystem.

> Let the covertext consist of an m-bit string with *even* parity that is otherwise uniformly random ($m \geq 2$). A ciphertext is computed as the XOR of the one-bit message and a one-bit random secret key; this is a random bit. Then the first bit of the covertext is replaced by the ciphertext and the last bit is adjusted such that the parity of the stegotext is *odd*.

Clearly, the scheme provides perfect secrecy for the message. The squared error distortion between covertext and stegotext is $1/m$ and vanishes as $m \to \infty$. Yet, an adversary can easily detect the presence of an embedded message *with certainty.* Our model from Section 3 reflects this fact adequately and considers such a scheme to be completely insecure (the discrimination is infinite).

As already mentioned in the introduction, the assumption of a fixed covertext distribution seems to render our model somewhat unrealistic for the practical purposes of steganography. But what are the alternatives? Should we rather study the perception and detection capabilities of human cognition since most cover data (images, text, sound) is ultimately intended for human receivers? Viewed this way, steganography could fall entirely into the realms of image, language, and audio processing or artificial intelligence in general. However, it seems that the information-theoretic model or other formal approaches will ultimately be more useful for deriving statements about the security of information hiding schemes—and a formal security notion is one of the main reasons for introducing a mathematical model of steganography.

References

[And96] R. Anderson (ed.), *Information hiding*, Lecture Notes in Computer Science, vol. 1174, Springer, 1996.

[BCW90] T. C. Bell, J. G. Cleary, and I. H. Witten, *Text compression*, Prentice Hall, 1990.

[Ber85] J. O. Berger, *Statistical decision theory and Bayesian analysis*, 2nd ed., Springer, 1985.

[BGML96] W. Bender, D. Gruhl, N. Morimoto, and A. Lu, *Techniques for data hiding*, IBM Systems Journal **35** (1996), no. 3 & 4, 313–336.

[Bla87] R. E. Blahut, *Principles and practice of information theory*, Addison-Wesley, Reading, 1987.

[BS98] D. Boneh and J. Shaw, *Collusion-secure fingerprinting for digital data*, IEEE Transactions on Information Theory **44** (1998), no. 5, 1897–1905.

[CK81] I. Csiszár and J. Körner, *Information theory: Coding theorems for discrete memoryless systems*, Academic Press, New York, 1981.

[CKLS96] I. J. Cox, J. Kilian, T. Leighton, and T. Shamoon, *A secure, robust watermark for multimedia*, Information Hiding, First International Workshop (R. Anderson, ed.), Lecture Notes in Computer Science, vol. 1174, Springer, 1996.

[Csi98] I. Csiszár, *The method of types*, IEEE Transactions on Information Theory **44** (1998), no. 6, 2505–2523.

[CT91] T. M. Cover and J. A. Thomas, *Elements of information theory*, Wiley, 1991.

[Ett98] M. Ettinger, *Steganalysis and game equilibria*, Information Hiding, 2nd International Workshop (D. Aucsmith, ed.), Lecture Notes in Computer Science, Springer, 1998, pp. 319–328.

[Mas91] J. L. Massey, *Contemporary cryptography: An introduction*, Contemporary Cryptology: The Science of Information Integrity (G. J. Simmons, ed.), IEEE Press, 1991, pp. 1–39.

[Mau96] U. M. Maurer, *A unified and generalized treatment of authentication theory*, Proc. 13th Annual Symposium on Theoretical Aspects of Computer Science (STACS) (C. Puech and R. Reischuk, eds.), Lecture Notes in Computer Science, vol. 1046, Springer, 1996, pp. 190–198.

[Mit99] T. Mittelholzer, *An information-theoretic approach to steganography and watermarking*, Information Hiding, 3rd International Workshop, IH'99 (A. Pfitzmann, ed.), Lecture Notes in Computer Science, vol. 1768, Springer, 1999, pp. 1–16.

[Nat85] S. Natarajan, *Large deviations, hypotheses testing, and source coding for finite markov chains*, IEEE Transactions on Information Theory **31** (1985), no. 3, 360–365.

[NFC94] M. Naor, A. Fiat, and B. Chor, *Tracing traitors*, Advances in Cryptology: CRYPTO '94 (Y. G. Desmedt, ed.), Lecture Notes in Computer Science, vol. 839, 1994.

[NS95] M. Naor and A. Shamir, *Visual cryptography*, Advances in Cryptology: EUROCRYPT '94 (A. De Santis, ed.), Lecture Notes in Computer Science, vol. 950, Springer, 1995, pp. 1–12.

[PAK99] F. A. Petitcolas, R. J. Anderson, and M. G. Kuhn, *Information hiding—a survey*, Proceedings of the IEEE **87** (1999), no. 7, 1062–1078.

[Pfi96] B. Pfitzmann, *Information hiding terminology*, Information Hiding, First International Workshop (R. Anderson, ed.), Lecture Notes in Computer Science, vol. 1174, Springer, 1996.

[PS96] B. Pfitzmann and M. Schunter, *Asymmetric fingerprinting*, Advances in Cryptology: EUROCRYPT '96 (U. Maurer, ed.), Lecture Notes in Computer Science, vol. 1233, Springer, 1996.

[Sha48] C. E. Shannon, *A mathematical theory of communication*, Bell System Technical Journal **27** (1948), 379–423, 623–656.

[Sha49] C. E. Shannon, *Communication theory of secrecy systems*, Bell System Technical Journal **28** (1949), 656–715.

[Sim84] G. J. Simmons, *The prisoners' problem and the subliminal channel*, Advances in Cryptology: Proceedings of Crypto 83 (D. Chaum, ed.), Plenum Press, 1984, pp. 51–67.

[Sim85] G. J. Simmons, *Authentication theory/coding theory*, Advances in Cryptology: Proceedings of CRYPTO 84 (G. R. Blakley and D. Chaum, eds.), Lecture Notes in Computer Science, vol. 196, Springer, 1985.

[WZW98] A. D. Wyner, J. Ziv, and A. J. Wyner, *On the role of pattern matchning in information theory*, IEEE Transactions on Information Theory **44** (1998), no. 6, 2045–2056.

[ZFK^{+}98] J. Zöllner, H. Federrath, H. Klimant, A. Pfitzmann, R. Piotraschke, A. Westfeld, G. Wicke, and G. Wolf, *Modeling the security of steganographic systems*, Information Hiding, 2nd International Workshop (D. Aucsmith, ed.), Lecture Notes in Computer Science, Springer, 1998, pp. 344–354.

[ZL77] J. Ziv and A. Lempel, *A universal algorithm for sequential data compression*, IEEE Transactions on Information Theory **23** (1977), no. 3, 337–343.