An Information-Theoretic Model for Steganography

Christian Cachin [*]

MIT Laboratory for Computer Science
545 Technology Square
Cambridge, MA 02139, USA
cachin@acm.org

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. Several secure steganographic schemes are presented in this model; one of them is a universal information hiding scheme based on universal data compression techniques that requires no knowledge of the covertext statistics.

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, the interest in techniques for information hiding has been increasing over the recent years [1]. 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 can be found in [2].

Steganography with a passive adversary is perhaps best illustrated by Simmons' "Prisoners' Problem" [19]. 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, such as 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

[*] Research supported by the Swiss National Science Foundation (SNF).

David Aucsmith (Ed.): Information Hiding 1998, LNCS 1525, pp. 306–318, 1998.

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") [13,16,17].

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 [4,8].

A common model and terminology for information hiding has been established at the 1996 Information Hiding Workshop [15]. 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.

In this paper, we introduce an information-theoretic model for steganography with a passive adversary. We propose a security notion that is based on *hypothesis testing*: 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 information-theoretic (non-Bayesian) approach as presented by Blahut [6] using 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 even observing a message.

However, some caution has to be exerted using this model: On the 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 [18]. Messages to be transmitted are modeled as random processes and the systems developed in this model perform well in practice. 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 method is perfectly secure according to our notion of security. But no warden would allow the prisoners to exchange random-looking messages! 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 can serve also as a starting point for further work to formalize active adversaries or computational security. (A game-theoretic approach to information hiding with active adversaries is presented by Ettinger [10].) A first extension would be to model the covertext source as a stochastic process and consider statistical estimation and decision techniques. Another idea would be to value the possible decisions and use the methods of statistical decision theory [5].

Related to this work is a paper by Maurer [12] on unconditionally secure authentication [11,21]. It shows how Simmons' bound [20] and many other lower bounds in authentication theory can be derived and generalized using the hypothesis testing approach. Another information-theoretic approach to steganography is [24].

The paper is organized as follows. 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 and discuss the effects of data compression. A universal information hiding scheme that requires no knowledge of the covertext statistics is presented in section 5. It is based on a universal data compression algorithm, which is similar to the well-known Lempel-Ziv algorithms [3,23]. Some extensions and conclusions are given in section 6.

2 Review of Hypothesis Testing

We give a brief introduction to hypothesis testing and to information-theoretic notions (see [6,7]). 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 $P_{X|Y=y}$.

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 [6]. In other words, there are two possible 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 possible 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 any given threshold $T \in \mathbb{R}$ and a given 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.*

The basic information measure of hypothesis testing is the *relative entropy* or *discrimination* between two probability distributions P_{Q_0} and P_{Q_1}, defined as

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

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 defined as

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

The following relation 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

$$H(X) + D(P_X \| P_U) \;=\; \log |\mathcal{X}|. \tag{3}$$

Relative entropy and hypothesis testing are linked through the Neyman-Pearson theorem above: The expected value of the log-likelihood ratio in (1) with respect to P_{Q_0} is equal to the relative entropy $D(P_{Q_0} \| P_{Q_1})$ between P_{Q_0} and 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, 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 δ is an upper bound on $D(P_{Q_0} \| P_{Q_1})$ 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 $\beta \geq 2^{-\delta}$.

A similar result holds 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 $\overline{\beta} = \sum_{v \in \mathcal{V}} P_V(v)\beta(v)$.

The *conditional relative entropy* between P_X and P_Y (over the same alphabet $\mathcal{X}$) conditioned on a random variable Z is defined as

$$D(P_{X|Z} \| P_{Y|Z}) = \sum_{z \in \mathcal{Z}} P_Z(z) \sum_{x \in \mathcal{X}} P_{X|Z=z}(x) \log \frac{P_{X|Z=z}(x)}{P_{Y|Z=z}(x)}. \tag{5}$$

It follows from the Jensen inequality and from (4) that

$$d(\overline{\alpha}, \overline{\beta}) \ \leq \ D(P_{Q_0|Z} \| P_{Q_1|Z}). \tag{6}$$

3 Model and Definition of Security

Fig. 1 shows our model of a stegosystem. 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 inactive (sending 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 must be able to recover E from his knowledge of the stegotext S and from the key K. Expressed in terms of entropy, the system satisfies:

1. $H(S|CEKR) = 0$. The stegotext is determined uniquely by Alice's inputs.
2. $H(E) > 0$. There is uncertainty about the embedded message.
3. $H(E|SK) = 0$. Bob must be able to decode the embedded message uniquely.

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.

Repetition is not considered in this model; it encompasses everything sent from Alice to Bob. For example, if Alice sends multiple 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.

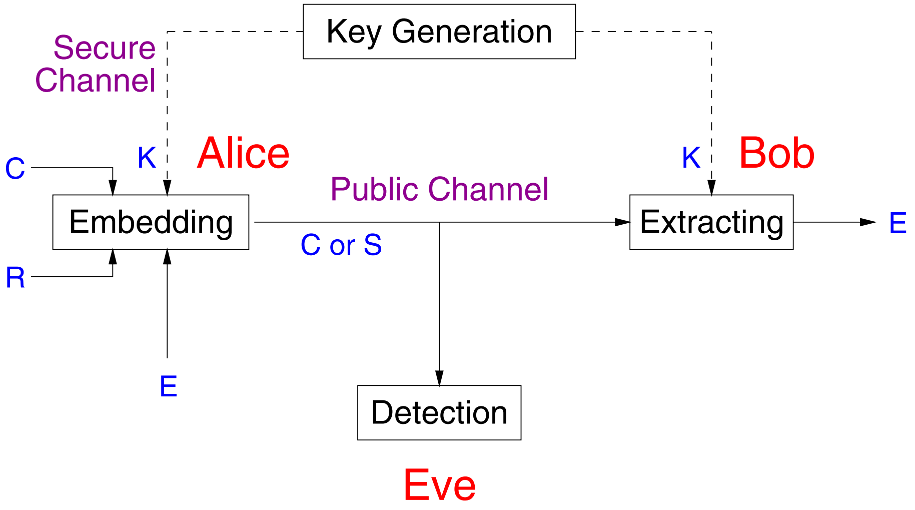

Fig. 1. The model of a secret-key stegosystem with passive adversary. It shows the embedded text E, the covertext C, the stegotext S, Alice's private random source R, and the secret key K shared by Alice and Bob. Alice is either sending covertext C or stegotext S.

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 defined 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*.*

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. As a special case, one can assume that Eve never makes a type I error and accuses Alice of sending hidden information when she is inactive ($\alpha = 0$). Such a restriction could be imposed on Eve by external mechanisms.

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

Theorem 1. *In a stegosystem that is ϵ-secure against passive adversaries, the probability β that the adversary does not detect a hidden message and the probability α that the adversary falsely detects a hidden 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.

As 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 (see e.g. [8]).

It may be the case that external events influence the covertext distribution; for example, news reports 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 can be modified accordingly. The quantities involved will be 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.

The modified 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 (6) that the average error probabilities satisfy $d(\overline{\alpha}, \overline{\beta}) \leq \epsilon$, similar to Theorem 1.

In the next section, we show that perfectly secure stegosystems exist for particular sources of covertext. We start with especially simple (or unrealistic) covertext distributions and then consider arbitrary covertext statistics and the effects of data compression. A universal stegosystem that includes data compression and does not rely on knowledge of the covertext distribution is presented in section 5.

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.

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 [14]. 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.

For general covertext distributions, we now describe a system that embeds a one-bit message in the stegotext. The extension to larger message spaces is straightforward. Let the covertext C with alphabet $\mathcal{C}$ have an arbitrary 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 mass 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 2. *The one-bit message stegosystem described above is*

$$\frac{1}{\ln 2} \left(\mathrm{P}[C \in \mathcal{C}_0] - \mathrm{P}[C \in \mathcal{C}_1] \right)^2$$

secure against passive adversaries.

Proof. Let $\delta = \mathrm{P}[C \in \mathcal{C}_0] - \mathrm{P}[C \in \mathcal{C}_1]$. 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 word 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 make detection 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 extend the model of a stegosystem to stochastic processes and consider the covertext as an ergodic source, its distribution can be estimated by observing the source output. This is precisely what universal data compression algorithms do for the purpose of source coding. We now show how they can be modified for information hiding.

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 unknown a priori or vary with time. A universal data compression universal algorithm achieves asymptotically optimal performance on every source in some large class of possible sources. Essentially, 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 algorithms by Lempel and Ziv [3,23].

We describe a universal data compression algorithm based on the concept of repetition times due to Willems [22], which is related to Elias' interval length coding [9]. Then we modify the algorithm to illustrate that a stegosystem can be constructed without knowledge of the covertext distribution. The performance of Willems' algorithm is inferior to the Lempel-Ziv algorithms for most practical data but it is simpler to describe and to analyze. We assume that covertext and stegotext in the model according to ection 3 are stationary stochastic processes. This corresponds to the ergodicity assumptions that are made for many data compression algorithms.

The Repetition Times Compression Algorithm: The algorithm is described for binary sources but can easily be generalized to arbitrary alphabets. The parameters of the algorithm are the blocklength L and the delay D. Consider a stationary binary source X producing $\{X_t\} = X_1, X_2, \ldots$ with values in $\{0, 1\}$. The source output is divided into blocks $Y_1, Y_2, \ldots$ of length L bits each. Encoding of a block Y_t operates by considering its *repetition time*, the length of the interval since its last occurrence. Formally, the repetition time Δt_y of the block $Y_t = y$ satisfies $Y_j \neq Y_t$ for $1 \leq j < \Delta t_y$ and $Y_{t-\Delta t_y} = Y_t$. If $\Delta t_y < 2^D$, the encoder outputs $C(\Delta t_y)$, using a particular variable-length encoding C of Δt_y described below. If $\Delta t_y \geq 2^D$, however, the block y is retransmitted literally. The distinction between repetition time encoding and literal data is marked by a single bit in the output stream.

Repetition time is encoded using the following code for integers between 1 and $2^D - 1$. Let $B_l(k)$ denote binary representation of the integer k using l digits and let $d = \lceil \log D \rceil$. The encoding C is

$$C(t) \;=\; B_d(\lfloor \log t \rfloor) \parallel B_{\lfloor \log t \rfloor}\left(t - 2^{\lfloor \log t \rfloor}\right),$$

where $\parallel$ denotes the concatenation of the bit strings. Thus, $C(t)$ contains first the binary length of t encoded using fixed length d and then the remaining bits of t except for the most significant bit. For initialization, a block y that occurs for the first time is encoded as if it had occurred at time $t = -y$.

The encoder and decoder maintain a buffer of the last 2^D blocks of the source. In addition, the encoder maintains an array indexed by L-bit blocks y that contains the position t_y (modulo 2^D) where y last occurred (the time buffer). Encoding and decoding therefore take only a constant number of operations per block. The formal analysis of the scheme [22] using $D = L$ shows that for $L \to \infty$, the encoding rate (the average number of code bits per source word) converges to the entropy rate of the source X.

The Modification for Information Hiding: The stegosystem based on Willems' algorithm exploits the fact that the average repetition time of a block $Y_t = y$ yields an estimate of its probability since it will converge to $P_Y(y)^{-1}$. If the block y is replaced with another block y' close to y in average repetition time (and therefore in probability), the source statistics are only slightly altered. Information is only hidden in blocks with low probability, as determined by a stego rate parameter $\rho > 2^{-D}$. Alice and Bob share an m-bit secret key K and Alice wants to hide an m-bit message E.

Here, both the encoder and decoder maintain a time buffer indexed by blocks. In addition to the index t of the last occurrence of block y, each entry contains also its average repetition time $\overline{\Delta t_y}$ and the number of its occurrences so far, n_y. For each encoded block y with repetition time Δt_y, the average repetition time $\overline{\Delta t_y}$ is replaced by $(n_y \overline{\Delta t_y} + \Delta t_y)/(n_y + 1)$. In addition, n_y is increased, but never beyond 2^D. Let $r(y)$ denote the rank function of blocks that associates with a block y the rank of $\overline{\Delta t_y}$, considering the current values of the average repetition times.

Information hiding takes place if the encoder or the decoder encounters a block y such that $\overline{\Delta t_y} \geq \frac{1}{\rho}$ (before updating buffers). If this is the case, bit j of the message m is embedded in y' according to

$$y' \;=\; r^{-1}\bigl(r(y) + (m_j \oplus K_j)\bigr)$$

and encoding proceeds as before with y' replacing y. In other words, y' is either equal to y or to the block immediately following y in the average repetition time ranking, depending on the embedded bit. The decoder computes the average repetition times in the same way and can thus detect the symbols containing hidden information and decode E similarly.

Compared to data compression, the storage complexity of the encoding and decoding algorithms is increased by a constant factor, but their computational complexity grows by a factor of about L due to the maintenance of the ranking.

The resulting stegosystem achieves asymptotically perfect security since the distance between the probabilities of the exchanged blocks vanishes. The formal statement of this will be given in the full version of the paper.

6 Extensions

The presented information-theoretic model for steganography can be considered as one particular example of a statistical model. We propose to consider also other approaches from statistical decision theory. As noted before, an immediate extension would be to model the covertext source as a stochastic process.

Simmons' original scenario of the prisoners' problem includes authentication, that is, the secret key K shared by Alice and Bob can partially be 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 Z. Then, for every value z of Z, there is a different covertext distribution $P_{C|Z=z}$ induced by the authentication scheme in use. However, since the adversary Eve does not know Z, the covertext distribution to consider for detection is P_C, the marginal distribution of P_{CZ}. We 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.

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 (text, sound, images) is ultimately intended for human receivers? Viewed this way, steganography could fall entirely into the realms of image, audio, and speech processing or artificial intelligence. However, it seems that the information-theoretic model and other statistical 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.

Acknowledgment

I am grateful to Victor Boyko, Mark Ettinger, Ron Rivest, and Michael Waidner for interesting discussions on this subject.

References

1. R. Anderson, ed., *Information Hiding*, vol. 1174 of *Lecture Notes in Computer Science*, Springer, 1996.
2. R. J. Anderson and F. A. Petitcolas, "On the limits of steganography," *IEEE Journal on Selected Areas in Communications*, vol. 16, May 1998.
3. T. C. Bell, J. G. Cleary, and I. H. Witten, *Text Compression.* Prentice Hall, 1990.
4. W. Bender, D. Gruhl, N. Morimoto, and A. Lu, "Techniques for data hiding," *IBM Systems Journal*, vol. 35, no. 3 & 4, 1996.
5. J. O. Berger, *Statistical Decision Theory and Bayesian Analysis.* Springer, 2. ed., 1985.
6. R. E. Blahut, *Principles and Practice of Information Theory.* Reading: Addison-Wesley, 1987.
7. T. M. Cover and J. A. Thomas, *Elements of Information Theory.* Wiley, 1991.
8. I. J. Cox, J. Kilian, T. Leighton, and T. Shamoon, "A secure, robust watermark for multimedia," in *Information Hiding* (R. Anderson, ed.), vol. 1174 of *Lecture Notes in Computer Science*, Springer, 1996.
9. P. Elias, "Interval and recency rank source coding: two on-line adaptive variable-length schemes," *IEEE Transactions on Information Theory*, vol. 33, pp. 3–10, Jan. 1987.
10. M. Ettinger, "Steganalysis and game equilibria," in *Proc. 2nd Workshop on Information Hiding* (D. Aucsmith, ed.), Lecture Notes in Computer Science, Springer-Verlag, 1998.
11. J. L. Massey, "Contemporary cryptography: An introduction," in *Contemporary Cryptology: The Science of Information Integrity* (G. J. Simmons, ed.), ch. 1, pp. 1–39, IEEE Press, 1991.

12. U. M. Maurer, “A unified and generalized treatment of authentication theory,” in *Proc. 13th Annual Symposium on Theoretical Aspects of Computer Science (STACS)* (C. Puech and R. Reischuk, eds.), vol. 1046 of *Lecture Notes in Computer Science*, pp. 190–198, Springer, 1996.
13. M. Naor, A. Fiat, and B. Chor, “Tracing traitors,” in *Advances in Cryptology: CRYPTO '94* (Y. G. Desmedt, ed.), vol. 839 of *Lecture Notes in Computer Science*, 1994.
14. M. Naor and A. Shamir, “Visual cryptography,” in *Advances in Cryptology: EUROCRYPT '94* (A. De Santis, ed.), vol. 950 of *Lecture Notes in Computer Science*, pp. 1–12, Springer, 1995.
15. B. Pfitzmann, “Information hiding terminology,” in *Information Hiding* (R. Anderson, ed.), vol. 1174 of *Lecture Notes in Computer Science*, Springer, 1996.
16. B. Pfitzmann and M. Schunter, “Asymmetric fingerprinting,” in *Advances in Cryptology: EUROCRYPT '96* (U. Maurer, ed.), vol. 1233 of *Lecture Notes in Computer Science*, Springer, 1996.
17. B. Pfitzmann and M. Waidner, “Anonymous fingerprinting,” in *Advances in Cryptology: EUROCRYPT '97* (W. Fumy, ed.), vol. 1070 of *Lecture Notes in Computer Science*, Springer, 1997.
18. C. E. Shannon, “A mathematical theory of communication,” *Bell System Technical Journal*, vol. 27, pp. 379–423, 623–656, July, Oct. 1948.
19. G. J. Simmons, “The prisoners' problem and the subliminal channel,” in *Advances in Cryptology: Proceedings of Crypto 83* (D. Chaum, ed.), pp. 51–67, Plenum Press, 1984.
20. G. J. Simmons, “Authentication theory/coding theory,” in *Advances in Cryptology: Proceedings of CRYPTO 84* (G. R. Blakley and D. Chaum, eds.), vol. 196 of *Lecture Notes in Computer Science*, Springer, 1985.
21. G. J. Simmons, “An introduction to shared secret and/or shared control schemes and their application,” in *Contemporary Cryptology: The Science of Information Integrity* (G. J. Simmons, ed.), pp. 441–497, IEEE Press, 1991.
22. F. M. Willems, “Universal data compression and repetition times,” *IEEE Transactions on Information Theory*, vol. 35, pp. 54–58, Jan. 1989.
23. J. Ziv and A. Lempel, “A universal algorithm for sequential data compression,” *IEEE Transactions on Information Theory*, vol. 23, pp. 337–343, May 1977.
24. J. Zöllner, H. Federrath, H. Klimant, A. Pfitzmann, R. Piotraschke, A. Westfeld, G. Wicke, and G. Wolf, “Modeling the security of steganographic systems,” in *Proc. 2nd Workshop on Information Hiding* (D. Aucsmith, ed.), Lecture Notes in Computer Science, Springer-Verlag, 1998.