# An Information-Theoretic Model for Steganography*

Christian Cachin†

March 4, 2004

**Abstract**

An information-theoretic model for steganography with a passive adversary is proposed. The adversary's task of distinguishing between an innocent cover message $C$ and a modified message $S$ containing hidden information is interpreted as a hypothesis testing problem. The security of a steganographic system is quantified in terms of the relative entropy (or discrimination) between the distributions of $C$ and $S$, which yields bounds on the detection capability of any adversary. 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 that needs no knowledge of the covertext distribution, except that it is generated from independently repeated experiments.

## 1 Introduction

Steganography is the art and science of communicating in such a way that the presence of a message cannot be detected. This paper considers steganography with a *passive* adversary. The model is perhaps best illustrated by Simmons' "Prisoners' Problem" [16]: Alice and Bob are in jail, locked up in separate cells far apart from each other, and wish to devise an escape plan. They are allowed to communicate by means of sending authenticated messages via trusted couriers, provided they do not deal with escape plans. The couriers are agents of the warden Eve (the adversary) and will leak all communication to her. If Eve detects any sign of conspiracy, she will thwart the escape plans by transferring both prisoners to high-security cells from which nobody has ever escaped. Alice and Bob are well aware of these facts, so that before getting locked up, they have shared a secret codeword that they are now going to exploit for adding a hidden meaning to their seemingly innocent messages. Alice and Bob succeed if they can exchange information allowing them to coordinate their escape and Eve does not become suspicious.

Of course, Eve knows what constitutes a legitimate communication among prisoners; such a communication is called *covertext*. Eve also knows about the tricks that prisoners apply to add a hidden meaning to a seemingly innocent message, thereby generating so-called *stegotext*. Following the approach of information theory, we capture this knowledge by a *probabilistic model*, and view Eve's task of detecting hidden messages as a problem of *hypothesis testing*. We define

*To appear in *Information and Computation*. 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).

†Original work done at MIT Laboratory for Computer Science, supported by the Swiss National Science Foundation (SNF). Current address: IBM Research, Zurich Research Laboratory, Säumerstr. 4, CH-8803 Rüschlikon, Switzerland, `cachin@acm.org`.

the security of a steganographic system in terms of the *relative entropy* (or *discrimination*) between the distributions of the covertext and the stegotext. A stegosystem is called *perfect* if this relative entropy is zero. The model is presented in Section 2.

The consequence of our security notion for the detection performance of an adversary is investigated in Section 3, following a brief review of the theory of hypothesis testing. Two elementary stegosystems with information-theoretic security are described in Section 4 for illustrating the definition.

In Section 5, a universal stegosystem is presented that requires no knowledge of the covertext distribution for its users; it works by estimating the distribution and then simulating a covertext by sampling a stegotext with a similar distribution. A discussion of our model and a comparison to related work are given in Section 6, and conclusions are drawn in Section 7.

## 2 Model

**Preliminaries.** We define the basic properties of a stegosystem using the notions of entropy, mutual information, and relative entropy [2, 3].

The *entropy* of a probability distribution $P_X$ over an alphabet $\mathcal{X}$ is defined as $H(X) = -\sum_{x\in\mathcal{X}} P_X(x)\log P_X(x)$. When $X$ denotes a random variable with distribution $P_X$, the quantity $H(X)$ is simply called the *entropy of the random variable* $X$ (with the standard convention $0\log 0 = 0$ and logarithms to the base 2). Similarly, the *conditional entropy* of a random variable $X$ given 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 entropy of any distribution satisfies $0 \leq H(X) \leq \log|\mathcal{X}|$, where $|\mathcal{X}|$ denotes the cardinality of $\mathcal{X}$.

The *mutual information* between $X$ and $Y$ is defined as the reduction of entropy that $Y$ provides about $X$, i.e., $I(X;Y) = H(X) - H(X|Y)$. It is symmetric in $X$ and $Y$, i.e., $I(X;Y) = I(Y;X)$, and always non-negative.

The *relative entropy* or *discrimination* between two probability distributions $P_{Q_0}$ and $P_{Q_1}$ is defined as $D(P_{Q_0}\|P_{Q_1}) = \sum_{q\in\mathcal{Q}} P_{Q_0}(q)\log\frac{P_{Q_0}(q)}{P_{Q_1}(q)}$ (with $0\log\frac{0}{0} = 0$ 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 $V$ defined in both probability spaces is $D(P_{Q_0|V}\|P_{Q_1|V}) = \sum_{v\in\mathcal{V}} P_V(v)\sum_{q\in\mathcal{Q}} P_{Q_0|V=v}(q)\log\frac{P_{Q_0|V=v}(q)}{P_{Q_1|V=v}(q)}$.

The relative entropy between two distributions is non-negative and it is equal to 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 is useful to think of it as a distance.

**Stegosystems.** We use the standard terminology of information hiding [14]. There are two parties, Alice and Bob, who are the *users* of the stegosystem. Alice wishes to send an innocent-looking message with a hidden meaning over a public channel to Bob, such that the presence of hidden information goes unnoticed by a third party, the *adversary* Eve, who has perfect read-only access to the public channel.

Alice operates in one of two modes. In the first case, Alice is *inactive* and sends an innocent, legitimate message containing no hidden information, called *covertext* and denoted by $C$; it is generated according to a distribution $P_C$ known to Eve. One may imagine that the covertext is generated by a source to which only Alice has access. In the second case, Alice is *active* and sends *stegotext* $S$ with distribution denoted by $P_S$. The stegotext is computed from an

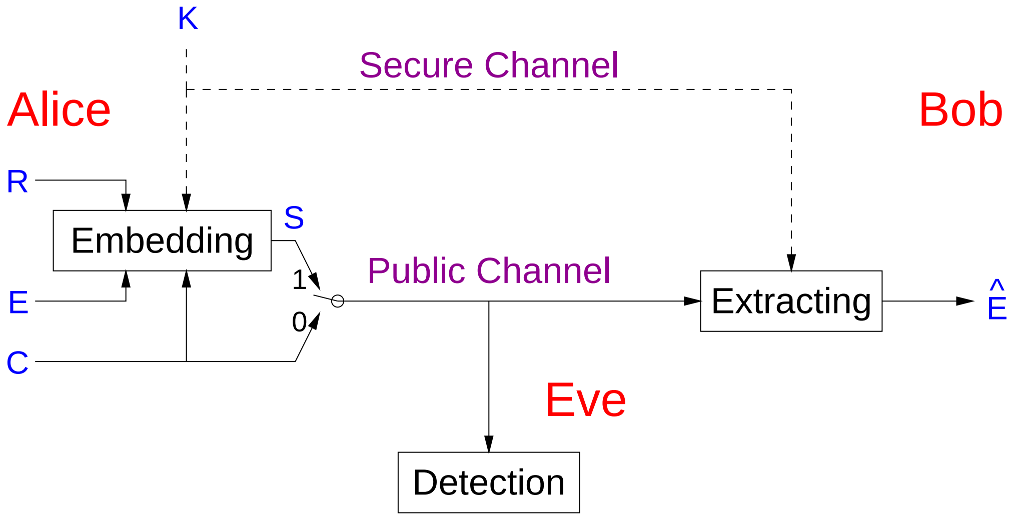


Figure 1: The model of a secret-key stegosystem.

*embedding function* $\mathcal{F}$ and contains an *embedded message* $E$ intended for Bob. The message is a random variable drawn from a *message space* $\mathcal{E}$.

Alice's embedding algorithm may access a *private random source* $R$ and a *secret key* $K$, which is shared by Alice and Bob. We assume that $R$ is independent of $E$ and $C$ and known only to Alice, and that $K$ is unknown to Eve. The key has been chosen at random and communicated over a secure channel prior to the use of the stegosystem—in any case before the message $E$ that Alice wants to communicate to Bob becomes known. Thus, we assume that $K$ is independent of $E$, $R$, and $C$.

The embedding function $\mathcal{F}$ and the distributions of all random variables are known to Eve. Hence, the model respects the prudent tradition known as "Kerckhoffs' principle" in cryptology, which places the security of a system only in the secrecy of a key but never in the secrecy of the design.

Figure 1 shows the model of a stegosystem in more detail. The switch at Alice's end of the public channel determines if Alice is active or not.

- In the first case (switch in position 0), Alice is inactive and sends only legitimate covertext $C$ to Bob over the public channel. The covertext is generated by a covertext source; no embedding takes place. The adversary Eve observes $C$.

- In the second case (switch in position 1), Alice is active and is given a message $E$ that she "embeds" into the given covertext $C$ using the embedding function $\mathcal{F}$. This is an algorithm that takes $C$, the shared key $K$, and private randomness $R$ as inputs and produces stegotext $S$. The stegotext is sent to Bob over the public channel. The adversary Eve and the receiver Bob observe $S$. Using his *extracting algorithm* $\mathcal{G}$, Bob extracts a decision value $\hat{E}$ from $S$ and $K$, in the hope that this gives him some information about $E$.

We assume that the covertext and stegotext distributions are known to Alice and Bob and thus the embedding algorithm may exploit knowledge about the covertext distribution (this will be relaxed in Section 5). However, we require that given a covertext distribution, the embedding function $\mathcal{F}$ is universal for information embedding, i.e., it works for any distribution $P_E$ of the message $E$. Thus, $\mathcal{F}$ must not depend on knowledge of $P_E$. This makes the stegosystem robust in the sense that the legitimate users do not have to worry about the adversary's knowledge of $E$.

Furthermore, we assume that Bob has an *oracle* that tells him if Alice is active or not. This is a strong assumption, and we make it here in order to focus on the security properties of a stegosystem. Removing it does not hurt the security of a stegosystem with respect to Eve's detection capability—if Bob was trying to extract an embedded message from the covertext when Alice is inactive, he would merely obtain garbage. As discussed in remark 5 below, the oracle does not open the way to trivial stegosystems, and in Section 4, Example 2, we demonstrate how to remove this assumption.

From the point of view of Eve, who does *not* know if Alice is active, the two cases above look similar: she observes data that is sent from Alice to Bob over the public channel. If Alice is not active, the data was generated according to $P_C$ and if she is active, it was generated from $P_S$. These are the two explanations that Eve has for the observation, which faces her with a problem of *hypothesis testing* [2, 3].

We quantify the security of the *stegosystem* in terms of the *relative entropy* $D(P_C \| P_S)$ between $P_C$ and $P_S$.

**Definition 1.** Fix a covertext distribution $C$ and a message space $\mathcal{E}$. A pair of algorithms $(\mathcal{F}, \mathcal{G})$ is called a *stegosystem* if there exist random variables $K$ and $R$ as described above such that for all random variables $E$ over $\mathcal{E}$ with $H(E) > 0$, it holds $I(\hat{E}; E) > 0$.

Moreover, a stegosystem is called *perfectly secure (against passive adversaries)* if

$$D(P_C \| P_S) \;=\; 0;$$

and a stegosystem is called *$\epsilon$-secure (against passive adversaries)* if

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

This model describes a stegosystem for *one-time use*, where Alice is always active or not. If Alice sends multiple dependent messages to Bob and at least one of them contains hidden information, she is considered to be active at all times and $S$ consists of the concatenation of all her messages.

Some remarks on the definition.

1. In a *perfectly secure* stegosystem, Eve cannot distinguish the two distributions and has no information at all about the presence of an embedded message. This parallels Shannon's notion of perfect secrecy for cryptosystems [15].

2. The condition in the definition of a stegosystem, $I(\hat{E}; E) > 0$, implies that a stegosystem is "useful" in the sense that Bob obtains at least some information about $E$. We chose not to model "useless" stegosystems.

3. Our model differs from the scenario sometimes considered for steganography, where Alice uses a covertext that is *known* to Eve and modifies it for embedding hidden information. Such schemes can only offer protection against adversaries with limited capability of comparing the modified stegotext to the covertext (otherwise, they are trivially breakable). For instance, this applies to the popular use of steganography on visual images, where a stegoimage may be perceptually indistinguishable from the coverimage for humans, but not for an algorithm with access to the coverimage.

4. It would be natural to require explicitly that a *perfectly* secure stegosystem provides also *perfect secrecy* for $E$ in the sense of Shannon [15] by demanding that $S$ and $E$ are statistically independent (as for example in the definition of Mittelholzer [12]). However, this is not necessary since we required the embedding algorithm to work without knowledge of the distribution $P_E$. This guarantees perfect secrecy for $E$ against Eve as follows. Fix a covertext distribution and an embedding function $\mathcal{F}$. For any distribution of $E$, algorithm $\mathcal{F}$ must produce $S$ with the same distribution as $C$. Since a concrete message value corresponds to a particular distribution of $E$ but the distribution of $S$ is the same for all values, $S$ is statistically independent from $E$.

   Analogously, we do not impose a secrecy constraint on $E$ for non-perfect stegosystems. The implications for the secrecy of $E$ are more involved and not investigate here; however, it is easy to construct stegosystems with perfect secrecy also in this case (see the stegosystem for general distributions in Section 4).

5. In our definition of a stegosystem, Bob knows from an oracle if Alice is active or not. Hence, one might be tempted to construct the following "perfect" stegosystem that exploits this knowledge for transmitting hidden information without using a shared secret key. W.l.o.g. consider an embedding algorithm $\mathcal{F}$ consisting of an ideal source encoder that manages to compress some message $E_1$ into stegotext $S_1$, which consists of independent and uniformly random bits. If the covertext $C$ is a sequence of independent and uniformly random bits of the same length, the two distributions are the same and Eve cannot distinguish a compressed message from covertext. In this case, Bob obtains $E_1$ without any secret key. His advantage to distinguish stegotext from covertext stems entirely from the oracle, and one might conclude that assuming such an oracle allows for trivial stegosystems.

   However, this conclusion does not hold because the described stegosystem is not perfectly secure according to Definition 1. Recall that $\mathcal{F}$ is deterministic and is required to work for *any* message distribution, so it must work also for some $E_2$ with strictly less entropy than $E_1$—for instance, when Eve has partial knowledge of the message. Let $S_2 = \mathcal{F}(E_2)$. Then it is intuitively clear that the deterministic $\mathcal{F}$ will not output enough random bits and the distributions of $C$ and $S_2$ are different.

   Formally, this can be seen by expanding the mutual information between the message and the stegotext in two ways. Since the encoder is deterministic *and* perfect, we have $H(S_1) = H(E_1)$ from expanding $I(E_1; S_1)$. The same encoder applied to $E_2$ also uniquely determines $S_2$, and therefore $H(S_2) = H(E_2) - H(E_2|S_2) \leq H(E_2)$ from expanding $I(E_2; S_2)$. Hence, $H(S_2) \leq H(E_2) < H(E_1) = H(S_1)$ by the assumption on $E_2$, which implies that the distributions of $S_1$ and $S_2$ differ and this contradicts the assumption that the stegosystem is perfect.

**Stochastic processes.** It is often appropriate to model an information source as a stochastic process. For example, the covertext may be generated from independent repetitions of the same experiment. In the model above, Eve observes the complete covertext, but it also makes sense to consider a restricted adversary who has only access to a subset of a long covertext sequence.

Let all random variables in the model above be extended to stochastic processes and let $n$ denote the number of repetitions. Assume that the covertext is generated by a *stationary* information source. Hence, the *normalized* relative entropy between the covertext and stegotext processes determines the security in cases where Eve is restricted to see a finite part of the covertext sequence.

**Definition 2.** A stegosystem for stochastic processes with stationary covertext is called *perfectly secure on average (against passive adversaries)* whenever

$$\lim_{n\to\infty} \frac{1}{n} D(P_C \| P_S) \;=\; 0.$$

Analogously, a stegosystem for stochastic processes is called *$\epsilon$-secure on average (against passive adversaries)* whenever

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

Notice that Alice is still either active or inactive during the entire experiment, and the stegotext distribution will not be ergodic in general.

# 3 Detection Performance

This section analyzes Eve's capabilities of detecting an embedded message. Basic bounds on her performance are obtained from the theory of hypothesis testing. A brief review of hypothesis testing is given first, following Blahut [2] (see also [11]).

**Hypothesis testing.** 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 $Q$. 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 $\alpha$, the probability of a type II error by $\beta$.

A basic property in hypothesis testing is that *deterministic processing* cannot increase the relative entropy between two distributions. For any function $f : \mathcal{Q} \to \mathcal{T}$, if $T_0 = f(Q_0)$ and $T_1 = f(Q_1)$, then $D(P_{T_0} \| P_{T_1}) \;\leq\; D(P_{Q_0} \| P_{Q_1})$.

Let $d(\alpha, \beta)$ denote the *binary relative entropy* of two distributions with parameters $(\alpha, 1-\alpha)$ and $(1-\beta, \beta)$, respectively, $d(\alpha,\beta) \;=\; \alpha \log \frac{\alpha}{1-\beta} + (1-\alpha) \log \frac{1-\alpha}{\beta}$.

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 $\alpha$ and $\beta$ satisfy $d(\alpha, \beta) \;\leq\; D(P_{Q_0} \| P_{Q_1})$. This inequality can be used as follows: Suppose that $D(P_{Q_0} \| P_{Q_1}) < \infty$ and that an upper bound $\alpha^*$ on the type I error probability is given. Then the above inequality yields a lower bound on the type II error probability $\beta$. For example, $\alpha^* = 0$ implies that $\beta \geq 2^{-D(P_{Q_0} \| P_{Q_1})}$.

**Bounds on the detection performance.** Consider Eve's decision process for detecting a hidden message in a stegosystem as a hypothesis testing problem. Any particular decision rule is 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 $\beta$.

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 $C$; its probability is denoted by $\alpha$. 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.

The deterministic processing property bounds the detection performance achievable by Eve. The following result is immediate from the discussion above.

**Theorem 1.** *In a stegosystem that is $\epsilon$-secure against passive adversaries, the probability $\beta$ that the adversary does not detect the presence of the embedded message and the probability $\alpha$ 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, the observed message does not give Eve any information about whether Alice is active or not.

## 4 Secure Stegosystems

According to our model, we obtain a secure stegosystem whenever the stegotext distribution is close to the covertext distribution for an observer with no knowledge of the secret key. The embedding function depends crucially on the covertext distribution. We assume in this section that the covertext distribution is known to the users Alice and Bob, and describe two basic stegosystems.

**Uniform covertext distributions.** The following is a simple example of a perfectly secure stegosystem.

*Example* 1. In the prisoner's scenario, suppose Alice and Bob both have a copy of the Bible in their cells. The adversary allows them to make a reference to any verse of the Bible in a message. All verses are considered to occur equally likely in a conversation among prisoners and there is a publicly known way to associate codewords with Bible verses. W.l.o.g. let the set of verses be $\{v_0, \ldots, v_{m-1}\}$. Furthermore, Alice and Bob share a uniformly random secret key $K$ in $\mathbb{Z}_m$. If Alice is active, she may embed a message $E \in \mathbb{Z}_m$ by mentioning $S = v_{(K+E) \bmod m}$. Bob obtains $E$ from $S$ and $K$ easily. Since we assume the distribution of a verse reference to be uniform, covertext and stegotext distributions are equal.

Likewise, the *one-time pad* is a perfectly secure stegosystem whenever the covertext consists of uniformly random bits. Assuming such a covertext would be rather unrealistic, but we describe it here briefly in order to illustrate the model.

*Example* 2. Assume the covertext $C$ is a uniformly distributed $n$-bit string for some positive $n$ and let Alice and Bob share an $n$-bit key $K$ with uniform distribution. The embedding function (if Alice is active) consists of applying bitwise XOR to the $n$-bit message $E$ and $K$, thus $S = E \oplus K$; Bob can decode this by computing $\hat{E} = S \oplus K$. The resulting stegotext $S$ is uniformly distributed in the set of $n$-bit strings and therefore $D(P_C \| P_S) = 0$.

We may remove the assumption that Bob knows if Alice is active as follows. Let the embedded message be $k < n$ bits long and take a binary linear code with $k$ information bits and block length $n$. Then Alice uses the message to select a codeword and embeds it in place of $E$ using the one-time pad stegosystem. Bob checks if the vector extracted from the one-time pad is a codeword. If yes, he concludes that Alice is active and decodes it to obtain the embedded message.

Incidentally, the one-time pad stegosystem is equivalent to the basic scheme of visual cryptography [13]. This technique hides a monochrome picture by splitting it into two random layers of dots. When these are superimposed, the picture appears. Using a slight modification of the basic scheme, it is also possible to produce two innocent-looking pictures such that both of them together reveal a hidden embedded message that is perfectly secure against an observer who has only one picture. Hence, visual cryptography is an example of a perfectly secure stegosystem.

**General distributions.** We now describe a system that embeds a one-bit message for arbitrary covertext distributions. The extension to larger message spaces is straightforward and omitted.

*Example* 3. Given a covertext $C$, Alice constructs the embedding function from a binary partition of the covertext space $\mathcal{C}$ such that both parts are assigned approximately the same probability under $P_C$. In other words, let

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

Alice and Bob share a uniformly distributed one-bit secret key $K$. 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 on $\mathcal{C}_1$. 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 $S \in \mathcal{C}_K$. Note that the embedding provides perfect secrecy for $E$.

**Theorem 2.** *The one-bit stegosystem in Example 3 has security $\delta^2/\ln 2$ against passive adversaries for $\delta = \Pr[C \in \mathcal{C}_0] - \Pr[C \in \mathcal{C}_1]$.*

*Proof.* We show only the case $\delta > 0$. It is straightforward but tedious 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 $\log(1+x) \leq x/\ln 2$. $\square$

In general, determining the optimal embedding function from a covertext distribution is an NP-hard combinatorial optimization problem. For instance, if we find an efficient embedding algorithm for the above one-bit stegosystem that achieves perfect security whenever possible, we have solved the NP-complete PARTITION problem [8].

## 5 Universal Stegosystems

The stegosystems described above require that the covertext distribution is known to its users. This seems not realistic for many applications. In this section, we describe a method for obtaining a *universal* stegosystem where such knowledge is not needed. It works for a covertext signal that is produced by a sequence of independent repetitions of the same experiment. Alice applies a *universal data compression* scheme to compute an approximation of the covertext distribution. She then produces stegotext with the approximate distribution of the covertext from *her own* randomness and embeds a message into the stegotext using the method of the one-time pad. Eve may have complete knowledge of the covertext distribution, but as long as she is restricted to observe only a finite part of the covertext sequence, this stegosystem achieves perfect average security asymptotically.

There are many practical universal data compression algorithms [1], and most encoding methods for perceptual data rely on them in some form. It is conceivable to combine them with our universal stegosystem for embedding messages in perceptual coverdata such as audio or video.

**The method of types.** One of the fundamental concepts of information theory is the *method of types* [6, 5]. 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 [6, 3] summarizes the basic properties of types.

**Lemma 3.** *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 on the order of $2^{nH(U)}$. 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,* $\frac{1}{(n+1)^{|\mathcal{X}|}} 2^{-nD(U\|P_X)} \leq \Pr[X^n \in \mathcal{T}(U)] \leq 2^{-nD(U\|P_X)}$.

**A universal data compression scheme.** A universal coding scheme $(\mathcal{E}, \mathcal{D})$ for a memoryless source $X$ works as follows. Fix a rate $\rho < \log|\mathcal{X}|$ and let $\rho_n = \rho - |\mathcal{X}|\frac{\log(n+1)}{n}$. Define a set of sequences $A_n = \{x^n \in \mathcal{X}^n : H(U_{x^n}) \leq \rho_n\}$. The block code is given by an enumeration $\mathcal{A} = \{1, \ldots, |\mathcal{A}|\}$ of the elements of $A_n$. The encoder $\mathcal{E}$ maps a sequence $X^n$ to a codeword in $\mathcal{A}$ if the entropy of the *type* of $X^n$ does not exceed $\rho_n$ and to a default value $\Delta$ otherwise. Let $Z$ denote the output of $\mathcal{E}$. Given a value $S \in \mathcal{A} \cup \{\Delta\}$, the decoder $\mathcal{D}$ returns the appropriate sequence in $A_n$ if $S \neq \Delta$ or a default sequence $x_0^n$ otherwise.

Lemma 3 implies that $|A_n| \leq 2^{n\rho}$ and therefore $\lceil n\rho \rceil$ bits are sufficient to encode all $x^n \in A_n$ [6, 3]. Moreover, if $H(X) < \rho$ then values outside $A_n$ occur only with exponentially small probability and the error probability $p_e^{(n)} = P_Z(\Delta)$ satisfies

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

The following observation is needed below. Write

$$\begin{aligned}
H(X^n) &= H(X^n Z) && (2)\\
&= P_Z(\Delta) H(X^n Z | Z = \Delta) + \big(1 - P_Z(\Delta)\big) H(X^n Z | Z \neq \Delta) && (3)\\
&\leq P_Z(\Delta) H(X^n) + \big(1 - P_Z(\Delta)\big)\big(H(Z|Z \neq \Delta) + H(X^n | Z, Z \neq \Delta)\big) && (4)\\
&\leq P_Z(\Delta) H(X^n) + H(Z | Z \neq \Delta), && (5)
\end{aligned}$$

where (2) follows because $Z$ is determined uniquely by $X^n$, (3) follows from rewriting, (4) holds because $Z$ is uniquely determined by $X^n$ and by rewriting, and (5) follows because codewords $Z \neq \Delta$ can be decoded uniquely. Rewriting this as

$$H(Z|Z \neq \Delta) \;\geq\; nH(X)\big(1 - p_e^{(n)}\big), \tag{6}$$

we see that the codeword $Z$ carries almost all information of $X^n$.

**A universal stegosystem.** Suppose the covertext, which is given as input to Alice, consists of $n$ independent realizations of a random variable $X$. Our universal stegosystem applies the above data compression scheme to the covertext. If Alice is active, she generates stegotext containing hidden information using the derived encoder and her private random source.

More precisely, given $\rho > H(X)$ and $n$, $\mathcal{F}$ maps the incoming covertext $X^n$ to its encoding $Z = \mathcal{E}(X^n)$. W.l.o.g. assume the output of the encoder is a binary $m$-bit string for $m = \lceil \log|\mathcal{A}| \rceil$ (or the special symbol $\Delta$) and the shared key $K$ is a uniformly random $\ell$-bit string with $\ell \leq m$; furthermore, let the message $E$ to be embedded be an $\ell$-bit string and let Alice's random source $R$ generate uniformly random $(m - \ell)$-bit strings.

If $\mathcal{E}$ outputs $Z = \Delta$, Alice sends $S = X^n$ and no message is embedded. Otherwise, she computes the $m$-bit string

$$T \;=\; (E \oplus K) \| R,$$

where $\|$ denotes the concatenation of bit strings, and sends $S = \mathcal{D}(T)$.

Bob extracts the embedded message from the received stegotext $S$ as follows. If $\mathcal{E}(S) = \Delta$, he declares a transmission failure and outputs a default value. Otherwise, he outputs

$$\hat{E} \;=\; \mathcal{E}(S)_{[1,\ldots,\ell]} \oplus K,$$

where $Z_{[1,\ldots,\ell]}$ stands for the prefix of length $\ell$ of a binary string $Z$.

Note that this stegosystem relies on Alice's private random source in a crucial way.

**Theorem 4.** *Let the covertext consist of a sequence $(X_1, \ldots, X_n)$ of $n$ independently repeated random variables with the same distribution $P_X$ for $n \to \infty$. Then given any $\epsilon > 0$, the algorithm above implements a universal stegosystem that is $\epsilon$-secure on average against passive adversaries and hides an $\ell$-bit message with $\ell \leq nH(X)$, for $n$ sufficiently large.*

*Proof.* It is easy to see that the syntactic requirements of a stegosystem are satisfied because the embedding and extraction algorithms are deterministic. For the information transmission property, it is easy to see from the given universal coding scheme $(\mathcal{E}, \mathcal{D})$ that, whenever $\mathcal{E}(S) \neq \Delta$, we have

$$\hat{E} \;=\; \mathcal{E}(S)_{[1,\ldots,\ell]} \oplus K \;=\; \mathcal{E}(\mathcal{D}(T))_{[1,\ldots,\ell]} \oplus K \;=\; T_{[1,\ldots,\ell]} \oplus K \;=\; E.$$

But this happens with overwhelming probability as shown below. Hence, $I(\hat{E}; E) \geq H(E|\mathcal{E}(S) \neq \Delta) > 0$ as required. It remains to show that the stegosystem is $\epsilon$-secure on average.

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

$$m \;=\; \lceil n\rho \rceil \;\leq\; \lceil nH(X) + n\epsilon/2 \rceil. \tag{7}$$

Define a binary random variable $V$ as follows:

$$V \;=\; \begin{cases} 0 & \text{if } Z \neq \Delta, \\ 1 & \text{if } Z = \Delta. \end{cases}$$

We now bound the relative entropy between covertext and stegotext. It is well-known that conditioning on derived information (side information, which has the same distribution in both cases) can only increase the discrimination between two distributions. Namely, given two random variables $Q_0$ and $Q_1$ over $\mathcal{Q}$, and a function $f : \mathcal{Q} \to \mathcal{V}$ such that the random variables $f(Q_0)$ and $f(Q_1)$ have the same distribution $P_V$, it holds $D(P_{Q_0} \| P_{Q_1}) \;\leq\; D(P_{Q_0|V} \| P_{Q_1|V})$ [2, Thm. 4.3.6]. Hence,

$$\begin{aligned}
D(P_C \| P_S) \;&\leq\; D(P_{C|V} \| P_{S|V}) && (8)\\
&=\; P_V(0) D(P_{C|V=0} \| P_{S|V=0}) + P_V(1) D(P_{C|V=1} \| P_{S|V=1}) && (9)\\
&\leq\; D(P_{C|V=0} \| P_{S|V=0}) && (10)\\
&\leq\; D(P_{Z|V=0} \| P_T) && (11)\\
&=\; m - H(Z|V=0), && (12)
\end{aligned}$$

where (9) follows from the definition of conditional relative entropy. The second term in (9) vanishes because the covertext and stegotext distributions are the same whenever $V = 1$, and $P_V(0) \leq 1$, hence we obtain (10). Because $C$ and $S$ in the case $V = 0$ are obtained from $Z$ and $T$, line (11) follows from the deterministic processing property. Since $T$ is uniformly distributed, the next step (12) follows from the fact that for any random variable $X$ with alphabet $\mathcal{X}$, if $P_U$ denotes the uniform distribution over $\mathcal{X}$, then $H(X) + D(P_X \| P_U) \;=\; \log |\mathcal{X}|$.

Using the fact that the events $V = 0$ and $Z \neq \Delta$ are the same, insert (6) and (7) into (12) to obtain

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

Since $\rho_n$ approaches $\rho$ from below and $\rho > H(X)$, it follows that for all sufficiently large $n$, also $\rho_n > H(X)$ and the value $\min_{U:H(U)>\rho_n} D(U\|P_X)$ in the exponent in (1) is strictly positive. This implies that the last expression is smaller than $\epsilon$ for all sufficiently large $n$ and that the stegosystem is indeed $\epsilon$-secure on average. $\square$

## 6 Discussion

**Limitations.** The adequacy of our information-theoretic model for real-world steganographic applications depends crucially on the assumption that there is a probabilistic model of the covertext. Moreover, the users of a stegosystem need at least some way to access or to sample the covertext distribution.

The use of probabilistic models is common practice in engineering today, but their application to steganography is of a somewhat different nature, since the security of a stegosystem cannot be demonstrated as easily as the performance of a data compression algorithm, for example. A secure stegosystem requires that the users and the adversary share the same probabilistic model of the covertext. As Example 2 shows, *if* the covertext distribution consists of uniformly random bits, *then* encrypting a message under a one-time pad results in a perfectly secure stegosystem according to our notion of security. But no reasonable warden would allow the prisoners to exchange randomly looking messages in the Prisoners' Problem! Thus, the validity of a formal treatment of steganography is determined by the accuracy of a probabilistic model for the real world.

Assuming the existence of a 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 the human cognition since most coverdata (images, text, and sound) is ultimately addressed to humans? Viewed in this way, steganography could fall entirely into the realms of image, language, and audio processing. However, it seems that an information-theoretic model, or any other formal approach, is more useful for deriving statements about the security of steganography schemes—and a formal security notion is one of the main reasons for introducing a mathematical model of steganography.

**Related work.** Most existing formal models for information hiding have not addressed steganography but the more general problem of *hiding information with active adversaries* in watermarking and fingerprinting applications. This is different from steganography because the existence of a hidden message is known publicly.

Since most objects to be protected by watermarking and fingerprinting consist of audio, image, or video data, these domains have received the most attention so far. A large number of hiding techniques and domain-specific models have been developed for robust, imperceptible information hiding [4]. Ettinger [7] models active adversaries with game-theoretic techniques.

We are aware of only two related information-theoretic models for steganography.

Zöllner et al. [17] define steganography using information-theoretic methods and mention 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, analogous to Shannon's notion of perfect secrecy for cryptosystems.

Mittelholzer [12] defines steganography (with a passive adversary) and watermarking (with an active adversary) using an information-theoretic model. 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 the security conditions from both models may be necessary, they are not sufficient to guarantee undetectable communication, as can be seen from the following insecure stegosystem.

*Example* 4. Let the covertext consist of an $m$-bit string with *even* parity that is otherwise uniformly random ($m \geq 2$). Let a ciphertext bit be computed as the XOR of a 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 bit and the last bit is adjusted such that the parity of the resulting 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*. In the sense of Definition 1, such a scheme is completely insecure since the discrimination is infinite.

Another related work is a paper of Maurer [11] on unconditionally secure authentication in cryptography, which demonstrates the generality of the hypothesis testing approach.

## 7 Conclusion

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

Simmons' original formulation of the prisoners' problem includes explicit authentication, that is, the secret key $K$ shared by Alice and Bob is partially used for authenticating Alice's messages. The reason is that Alice and Bob want to protect themselves from the adversary and from malicious couriers (and they are allowed to do so), which may give rise to a subliminal channel in the authentication scheme. It would be interesting to extend our model for this scenario.

Another possible extension, taken up by Katzenbeisser and Petitcolas [10] and by Hopper, Langford, and von Ahn [9], is to model steganography with the complexity-theoretic security notions of modern cryptography, and to define a secure stegosystem such that the stegotext is computationally indistinguishable from the covertext.

## Acknowledgment

I am grateful to Thomas Mittelholzer for interesting discussions and for suggesting a crucial improvement to Section 5.

## References

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

[2] R. E. Blahut, *Principles and Practice of Information Theory.* Reading: Addison-Wesley, 1987.

[3] T. M. Cover and J. A. Thomas, *Elements of Information Theory.* Wiley, 1991.

[4] I. J. Cox, M. L. Miller, and J. A. Bloom, *Digital Watermarking.* Morgan Kaufmann, 2002.

[5] I. Csiszár, "The method of types," *IEEE Transactions on Information Theory*, vol. 44, pp. 2505–2523, Oct. 1998.

[6] I. Csiszár and J. Körner, *Information Theory: Coding Theorems for Discrete Memoryless Systems.* New York: Academic Press, 1981.

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

[8] M. R. Garey and D. S. Johnson, *Computers and Intractability — A Guide to the Theory of NP-Completeness.* W. H. Freeman, 1979.

[9] N. J. Hopper, J. Langford, and L. von Ahn, "Provably secure steganography," in *Advances in Cryptology: CRYPTO 2002* (M. Yung, ed.), vol. 2442 of *Lecture Notes in Computer Science*, pp. 77–92, Springer, 2002.

[10] S. Katzenbeisser and F. A. P. Petitcolas, "Defining security in steganographic systems," in *Security and Watermarking of Multimedia Contents IV* (E. J. Delp and P. W. Won, eds.), vol. 4675 of *Proceedings of SPIE*, pp. 260–268, 2002.

[11] U. Maurer, "Authentication theory and hypothesis testing," *IEEE Transactions on Information Theory*, vol. 46, no. 4, pp. 1350–1356, 2000.

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

[13] 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.

[14] B. Pfitzmann, "Information hiding terminology," in *Information Hiding, First International Workshop* (R. Anderson, ed.), vol. 1174 of *Lecture Notes in Computer Science*, pp. 347–350, Springer, 1996.

[15] C. E. Shannon, "Communication theory of secrecy systems," *Bell System Technical Journal*, vol. 28, pp. 656–715, Oct. 1949.

[16] 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.

[17] J. Zöllner, H. Federrath, H. Klimant, A. Pfitzmann, R. Piotraschke, A. Westfeld, G. Wicke, and G. Wolf, "Modeling the security of steganographic systems," in *Information Hiding, 2nd International Workshop* (D. Aucsmith, ed.), vol. 1525 of *Lecture Notes in Computer Science*, pp. 344–354, Springer, 1998.