Provably Secure Steganography (Extended Abstract)

Nicholas J. Hopper, John Langford, and Luis von Ahn

Computer Science Department, Carnegie Mellon University, Pittsburgh PA 15213, USA
{hopper,jcl,biglou}@cs.cmu.edu

Abstract. Informally, *steganography* is the process of sending a secret message from Alice to Bob in such a way that an eavesdropper (who listens to all communications) cannot even tell that a secret message is being sent. In this work, we initiate the study of steganography from a complexity-theoretic point of view. We introduce definitions based on computational indistinguishability and we prove that the existence of one-way functions implies the existence of secure steganographic protocols.

1 Introduction

The scientific study of steganography began in 1983 when Simmons [13] stated the problem in terms of communication in a prison. In his formulation, two inmates, Alice and Bob, are trying to hatch an escape plan. The only way they can communicate with each other is through a public channel, which is carefully monitored by the warden of the prison, Ward. If Ward detects any encrypted messages or codes, he will throw both Alice and Bob into solitary confinement. The problem of steganography is, then: how can Alice and Bob cook up an escape plan by communicating over the public channel in such a way that Ward doesn't suspect anything fishy is going on. (Notice how steganography is different from classical cryptography, which is about hiding the *content* of secret messages: steganography is about hiding the very existence of the secret messages.)

Steganographic "protocols" have a long and intriguing history that goes back to antiquity. There are stories of secret messages written in invisible ink or hidden in love letters (the first character of each sentence can be used to spell a secret, for instance). More recently, steganography was used by prisoners and soldiers during World War II because all mail in Europe was carefully inspected at the time [8]. Postal censors crossed out anything that looked like sensitive information (e.g. long strings of digits), and they prosecuted individuals whose mail seemed suspicious. In many cases, censors even randomly deleted innocent-looking sentences or entire paragraphs in order to prevent secret messages from going through. Over the last few years, steganography has been studied in the framework of computer science, and several algorithms have been developed to hide secret messages in innocent looking data.

The main goal of this paper is to put steganography on a solid complexity-theoretic foundation. We define steganographic secrecy in terms of computational indistinguishability, and we define steganographic robustness, which deals with the case of active wardens (ones that cross out innocent-looking sentences or modify the messages just to prevent secrets from going through). Our main result is a positive one: secret and robust steganographic protocols exist within our model, given that one-way functions exist.

Related Work

There has been considerable work on digital steganography. The first International Workshop on Information Hiding occurred in 1996, with five subsequent workshops, and even books have been published about the subject [9]. Surprisingly, though, very little work has attempted to formalize steganography, and most of the literature consists of heuristic approaches: steganography using digital images [9], steganography using video systems [9], etc. A few papers have given information theoretic models for steganography [3, 10, 11, 14], but these are limited in the same way that information theoretic cryptography is limited. We believe complexity theory is the right framework in which to view steganography and, to the best of our knowledge, this is the first paper to treat steganography from a complexity-theoretic point of view (and to achieve provably positive results).

Organization of the Paper

In section 2 we define the basic cryptographic quantities used throughout the paper, as well as the notions of a cover *channel* and a *stegosystem*. In section 3 we define steganographic secrecy and state protocols which are steganographically secret assuming the existence of one-way functions. In section 4 we define robust steganographic secrecy for adversaries with bounded power to perturb stegotext messages and state protocols which satisfy this definition. Section 5 closes the paper with a discussion of implications.

2 Definitions

2.1 Preliminaries

A function $\mu : \mathbb{N} \rightarrow (0, 1)$ is said to be *negligible* if for every $c > 0$, for all sufficiently large n, $\mu(n) < 1/n^c$. The concatenation of string s_1 and string s_2 will be denoted by $s_1||s_2$, and when we write "Parse s as $s_1^t||s_2^t||\cdots||s_l^t$" we mean to separate s into strings $s_1, \ldots s_l$ where each $|s_i| = t$, $l = \lceil |s|/t \rceil$, and $s = s_1||s_2||\cdots||s_l$. We will let $U(k)$ denote the uniform distribution on k bit strings, and $U(L, l)$ denote the uniform distribution on functions from L bit strings to l bit strings. If X is finite a set, we let $U(X)$ denote the uniform distribution on X.

2.2 Cryptographic notions

Let $F : \{0,1\}^k \times \{0,1\}^L \to \{0,1\}^l$ denote a family of functions. Let A be an oracle probabilistic adversary. Define the *prf-advantage of A over F* as

$$\mathbf{Adv}^{\mathrm{prf}}_F(A) = \left| \Pr_{K \leftarrow U(k), r \leftarrow \{0,1\}^*}[A_r^{F_K(\cdot)} = 1] - \Pr_{g \leftarrow U(L,l), r \leftarrow \{0,1\}^*}[A_r^g = 1] \right| .$$

where r is the string of random bits used by adversary A. Define the insecurity of F as

$$\mathbf{InSec}^{\mathrm{prf}}_F(t,q) = \max_{A \in \mathcal{A}(t,q)} \left\{ \mathbf{Adv}^{\mathrm{prf}}_F(A) \right\}$$

where $\mathcal{A}(t,q)$ denotes the set of adversaries taking at most t steps and making at most q oracle queries. Then F is a *(t,q,ϵ)-pseudorandom function* if $\mathbf{InSec}^{\mathrm{prf}}_F(t,q) \leq \epsilon$. Suppose that $l(k)$ and $L(k)$ are polynomials. A sequence $\{F_k\}_{k \in \mathbb{N}}$ of families $F_k : \{0,1\}^k \times \{0,1\}^{L(k)} \to \{0,1\}^{l(k)}$ is called *pseudorandom* if for all polynomially bounded adversaries A, $\mathbf{Adv}^{\mathrm{prf}}_{F_k}(A)$ is negligible in k. We will sometimes write $F_k(K,\cdot)$ as $F_K(\cdot)$.

Let $E : \mathcal{K} \times \mathcal{R} \times \mathcal{P} \to \mathcal{C}$ be a probabilistic private key encryption scheme, which maps a random number and an $|m|$-bit plaintext to a ciphertext. Consider a game in which an adversary A is given access to an oracle which is either:

- E_K for $K \leftarrow U(\mathcal{K})$; that is, an oracle which given a message m, uniformly selects random bits R and returns $E_K(R,m)$; or
- $g(\cdot) = U(|E_K(\cdot)|)$; that is, an oracle which on any query ignores its input and returns a uniformly selected output of the appropriate length.

Let $\mathcal{A}(t,q,l)$ be the set of adversaries A which make q queries to the oracle of at most l bits and run for t time steps. Define the CPA advantage of A against E as

$$\mathbf{Adv}^{\mathrm{cpa}}_E(A) = \left| \Pr_{K \leftarrow U(\mathcal{K}), s, r \leftarrow \{0,1\}^*}[A_r^{E_{K,s}} = 1] - \Pr_{g, r \leftarrow \{0,1\}^*}[A_r^g = 1] \right|$$

where $E_{K,s}$ denotes E_K with random bit source s. Define the insecurity of E as

$$\mathbf{InSec}^{\mathrm{cpa}}_E(t,q,l) = \max_{A \in \mathcal{A}(t,q,l)} \left\{ \mathbf{Adv}^{\mathrm{cpa}}_E(A) \right\} .$$

Then E is *(t,q,l,ϵ)-indistinguishable from random bits under chosen plaintext attack* if $\mathbf{InSec}^{\mathrm{cpa}}_E(t,q,l) \leq \epsilon$. A sequence of cryptosystems $\{E_k\}_{k \in \mathbb{N}}$ is called *indistinguishable from random bits under chosen plaintext attack* (IND\$-CPA) if for every PPTM A, $\mathbf{Adv}^{\mathrm{cpa}}_{E_k}(A)$ is negligible in k.

Let $\mathcal{C}$ be a distribution with finite support X. Define the *minimum entropy* of $\mathcal{C}$, $H_\infty(\mathcal{C})$, as

$$H_\infty(\mathcal{C}) = \min_{x \in X} \left\{ \log_2 \frac{1}{\Pr_{\mathcal{C}}[x]} \right\} .$$

2.3 Steganography

Steganography will be thought of as a game between the warden, Ward, and the inmate, Alice. The goal of Alice is to pass a secret message to Bob over a communication channel (known to Ward). The goal of Ward is to detect whether a secret message is being passed. In this and the following sections we will formalize this game. We start by defining a communication channel.

Definition. A *channel* is a distribution on bit sequences where each bit is also timestamped with monotonically increasing time value. Formally, a channel is a distribution with support $(\{0,1\}, t_1), (\{0,1\}, t_2), ...$, where $\forall i > 0 : \;\; t_{i+1} \geq t_i$.

This definition of a channel is sufficiently general to encompass nearly any form of communication. It is important to note that our protocols may depend upon the timing information as well as the actual bits sent on a channel. For example, it may be possible to do steganography over email using only the timing of the emails rather than the contents of the message. It may also be possible for an enemy to detect steganography via timing analysis.

Anyone communicating on a channel can be regarded as implicitly drawing from the channel, so we will assume the existence of an oracle capable of drawing from the channel. In fact, we will assume something stronger: an oracle that can *partially* draw from the channel a (finite, fixed length) sequence of bits. This oracle can draw from the channel in steps and at any point the draw is conditioned on what has been drawn so far. We let $\mathcal{C}_h$ be the channel distribution conditional on the history h of already drawn timestamped bits. We also let $\mathcal{C}_h^b$ be the marginal channel distribution over the next block of b timestamped bits conditional on the history h. Intuitively, $\mathcal{C}_h^b$ is a distribution on the next b timestamped bits conditioned on the history h.

Fix b. We assume the existence of an oracle which can draw from $\mathcal{C}_h^b$. We will call such a partial draw a "block". We will require that the channel satisfy a minimum entropy constraint for all blocks:

$$\forall h \text{ drawn from } \mathcal{C} : \;\; H_\infty(\mathcal{C}_h^b) > 1$$

This partial draw will be conditional on all past draws and so we can regard a sequence of partial draws as a draw from the channel. This notion of randomness is similar to Martingale theory where random variable draws are conditional on previous random variable draws (and we use Martingale theory in our analysis).

It is important to note that a "block" might (for example) contain timestamped bits which span multiple emails. We will overload the definition of the concatenation operator $||$ for sequences of timestamped bits. Thus $c_1||c_2$ will consist of the timestamped bits of c_1 followed by the timestamped bits of c_2.

One example of a channel might be electronic mail. We can map an email system allowing communication from Alice to Bob to a channel by considering all of the bits used to encode a *particular* email as a sequence of channel bits, each with the same timestamp. The timestamp of emailed bits would be the time of transmission. The complete channel consists of a distribution over sequences of emails.

Remark. In the remainder of this paper, we will assume that cryptographic primitives remain secure with respect to an oracle which draws from a channel distribution $\mathcal{C}_h^b$. Thus channels which can be used to solve the hard problems that standard primitives are based on must be ruled out. In practice this is of little concern, since the existence of such channels would have previously led to the conclusion that the primitive in question was insecure.

Definition 1. (Stegosystem) A steganographic protocol, or stegosystem, is a pair of probabilistic algorithms $S = (SE, SD)$. SE takes a key $K \in \{0,1\}^k$, a string $m \in \{0,1\}^*$ (the *hiddentext*), a message history h, and an oracle $M(h)$ which samples blocks according to a channel distribution $\mathcal{C}_h^b$. $SE^M(K, m, h)$ returns a sequence of blocks $c_1||c_2||\ldots||c_l$ (the *stegotext*) from the support of $\mathcal{C}_h^{l*b}$. SD takes a key K, a sequence of blocks $c_1||c_2||\ldots||c_l$, a message history h, and an oracle $M(h)$, and returns a hiddentext m. There must be a polynomial $p(k) > k$ such that SE^M and SD^M also satisfy the relationship:

$$\forall m, |m| < p(k) : \quad \Pr(SD^M(K, SE^M(K, m, h), h) = m) \geq \frac{2}{3}$$

where the randomization is over any coin tosses of SE^M, SD^M, and M. (In the rest of the paper we will use (SE,SD) instead of (SE^M,SD^M).)

Note that we choose a probability of failure for the stegosystem of 1/3 in order to include a wide range of possible stegosystems. In general, given a protocol with any reasonable probability of failure, we can boost the system to a very low probability of failure using error-correcting codes.

Although all of our oracle-based protocols will work with the oracle $M(h)$, we will always use it in a particular way. Consequently, it will be convenient for us to define the rejection sampling function $RS^{M,F} : \{0,1\}^* \times \mathbb{N} \to \{0,1\}$.

Procedure $\mathtt{RS}^{M,F}$:
Input: target x, iteration *count*
i = 0
repeat: $c \leftarrow M$; increment i
until $F(c) = x$ or $i = count$
Output: c

The function RS simply samples from the distribution provided by the sample oracle M until $F(M) = x$. The function will return c satisfying $F(c) = x$ or the *count*-th sample from M. Note that we use an iteration count to bound the worst case running time of RS and that RS may fail to return a c satisfying $F(c) = x$.

Comment. We have taken the approach of assuming a channel which can be drawn from freely by the stegosystem; most current proposals for stegosystems act on a single sample from the channel (one exception is [3]). While it may be possible to define a stegosystem which is steganographically secret or robust and works in this style, this is equivalent to a system in our model which merely makes a single draw on the channel distribution. Further, we believe that the lack of reference to the channel distribution may be one of the reasons for the failure of many such proposals in the literature.

It is also worth noting that we assume that a stegosystem has very little knowledge of the channel distribution—SE and SD may only *sample* from an oracle according to the distribution. This is because in many cases the full distribution of the channel has never been characterized; for example, the oracle may be a human being, or a video camera focused on some complex scene. However, our definitions do not rule out encoding procedures which have more detailed knowledge of the channel distribution.

Sampling from $\mathcal{C}_h^b$ might not be trivial. In some cases $M(h)$ is a human, and in others a simple randomized program. We stress that it is important to minimize the use of such an oracle, because oracle queries can be extremely expensive. In practice, this oracle is also the weakest point of all our constructions. We assume the existence of a *perfect* oracle: one that can perform independent draws, one that can be rewound, etc. This assumption can be justified in some cases, but not in others. If the oracle is a human, the human may not be able to perform independent draws from the channel as is required by the function RS. A real world Warden would use this to his advantage. We therefore stress the following cautionary remark: *our protocols will be shown to be secure under the assumption that the oracle is perfect.*

Finally, we mention that our decoding algorithm, SD, is defined to have access to the oracle $M(h)$. This is a general definition, and there are cases in which this access will not be necessary. Protocols in which SD needs no access to $M(h)$ are clearly preferred.

3 Steganographic Secrecy

A *passive* warden, W, is an adversary which plays the following game:

1. W is given access to an oracle $M(h)$ which samples blocks (one at a time) from the distribution $\mathcal{C}_h^b$, for past histories h drawn from the channel. W makes as many draws from $M(h)$ as it likes.
2. W is given access to a second oracle which is either $SE(K,\cdot,\cdot)$ or $O(\cdot,\cdot)$ defined by $O(m,h) \leftarrow \mathcal{C}_h^{|SE(K,m,h)|}$. Ward W makes at most q queries totaling l bits (of hiddentext) to this oracle.
3. W outputs a bit.

We define W's advantage against a stegosystem S by

$$\mathbf{Adv}^{\mathrm{SS}}_{S,\mathcal{C}}(W) = \left| \Pr_{K,r,M,SE}[W_r^{M,SE(K,\cdot,\cdot)} = 1] - \Pr_{r,M,O}[W_r^{M,O(\cdot,\cdot)} = 1] \right| ,$$

where the warden uses random bits r. Define the insecurity of S by

$$\mathbf{InSec}^{\mathrm{SS}}_{S,\mathcal{C}}(t,q,l) = \max_{W \in \mathcal{W}(t,q,l)} \left\{ \mathbf{Adv}^{\mathrm{SS}}_{S,\mathcal{C}}(W) \right\} ,$$

where $\mathcal{W}(t,q,l)$ denotes the set of all adversaries which make at most q queries totaling at most l bits (of hiddentext) and running in time at most t.

Definition 2. (Steganographic secrecy) A Stegosystem $S = (SE, SD)$ is called (t, q, l, ϵ) *steganographically secret against chosen hiddentext attack* for the channel $\mathcal{C}$ ((t, q, l, ϵ)-SS-CHA-$\mathcal{C}$) if $\mathbf{InSec}^{\mathrm{SS}}_{S,\mathcal{C}}(t, q, l) \leq \epsilon$.

Definition 3. (Universal Steganographic Secrecy) A Stegosystem S is called (t, q, l, ϵ)-*universally steganographically secret against chosen hiddentext attack* ((t, q, l, ϵ)-USS-CHA) if it is (t, q, l, ϵ)-SS-CHA-$\mathcal{C}$ for every channel distribution $\mathcal{C}$ that satisfies $H_\infty(\mathcal{C}^b_h) > 1$ for all h drawn from $\mathcal{C}$. A sequence of stegosystems $\{S_k\}_{k\in\mathbb{N}}$ is called *universally steganographically secret* if for every channel distribution $\mathcal{C}$ and for every PPTM W, $\mathbf{Adv}^{\mathrm{SS}}_{S_k,\mathcal{C}}(W)$ is negligible in k.

Note that steganographic secrecy can be thought of roughly as encryption which is indistinguishable from *arbitrary* distributions $\mathcal{C}$.

3.1 Construction: Stateful Universal Steganographic Secrecy

We now show a stateful protocol which is universally steganographically secret. This protocol is not new and can be found (up to small differences) in [1], for instance. The novel aspect is the proof of steganographic secrecy within our model.

Assume Alice and Bob share a channel and let $\mathcal{C}$ denote the channel distribution. We let $M(h)$ be an oracle which draws blocks (one at a time) from the conditional distribution $\mathcal{C}^b_h$. Let $F_K(\cdot,\cdot)$ denote a pseudorandom function family indexed by $k = |K|$ key bits which maps a pair consisting of a d-bit number and b timestamped bits to the set $\{0,1\}$, i.e., $F : \{0,1\}^k \times \{0,1\}^d \times (\{0,1\}, t)^b \to \{0,1\}$. Let Alice and Bob share a secret key $K \in \{0,1\}^k$ and a synchronized d bit counter N. Let Enc and Dec be the encoding and decoding procedures, respectively, for an error-correcting code which corrects up to a $\mu(k)$ fraction of errors (where $\mu(k)$ is defined below) and which has stretch function $\ell(\cdot)$. The following procedures allow Alice and Bob to encode and decode a hiddentext and to send it via their shared channel.

Construction 1. (Steganographic Secrecy)

Procedure S1.Encode	Procedure S1.Decode

Procedure `S1.Encode`:
Input: key K, hiddentext m', history h
Let $m = Enc(m')$
Parse m as $m^1_1 \| m^1_2 \| \cdots \| m^1_l$
for $i = 1 \ldots l$ do
    $c_i \leftarrow RS^{M(h),F_K(N,\cdot)}(m_i, \|K\|)$
    set $h = h \| c_i$
    increment N
Output: $c_1 \| c_2 \| \ldots \| c_l$
	Procedure `S1.Decode`:
	Input: key K, Stegotext c
	Parse c as $c^b_1 \| c^b_2 \| \ldots \| c^b_l$
	for $i = 1 \ldots l$ do
	    set $m_i = F_K(N, c_i)$
	    increment N
	let $m = m_1 \| m_2 \| \cdots \| m_l$
	Output: $Dec(m)$

The idea behind this construction is simple. First, the encoding algorithm splits the hiddentext m into single bits. Then it draws blocks from the channel until it finds c_1 such that $F_K(N, c_1) = m_1$. After that it finds c_2 such that $F_K(N +$

$1, c_2) = m_2$, and c_3, and so on. This continues until all bits of m have been encoded. The decoding algorithm simply applies $F_K(N+i-1,\cdot)$ to all the c_i's.

Note that the encoding procedure actually has a small probability of failure per bit. In particular, RS might not return a c_i such that $F_K(N+i-1, c_i) = m_i$, because RS "gives up" after $|K|$ timesteps. The probability of failure is dependent on the sampling oracle M and the evaluation oracle F.

Lemma 1. *If $\mathcal{C}_h^b$ has minimum entropy at least H and $|K| = k$, the probability of failure of RS is bounded above by fail$(k, \mathcal{C}_h^b) + \mathbf{InSec}_F^{prf}(O(kl), kl)$, where*

$$\mathit{fail}(k, \mathcal{C}_h^b) \leq \sum_{j=0}^{2^H} \frac{\binom{2^H}{j}}{2^{2^H}} \left(\frac{j}{2^H}\right)^k$$

Proof. See the full version of the paper [6].

It can be verified that $\text{fail}(k, \mathcal{C}_h^b)$ is exponentially small in $|K|$ and H. We will choose our error-correcting code to correct a $\mu(k) = (1+\epsilon) * \text{fail}(k, U(1))$ fraction of errors (note that $U(1)$ has minimum entropy 1). A Chernoff bound will then suffice to show that the probability of decoding failure is exponentially small in l. For a channel $\mathcal{C}_h^b$ with minimum entropy 1, we have $\text{fail}(k, \mathcal{C}_h^b) \approx \frac{1}{4}$. Consequently, we will at worst require a code with a stretch function $\ell(n)$ approximately $2n$. We will assume for simplicity that the running times of Enc and Dec are linear.

Theorem 1. *Let $k = |K|$. For any $l \leq 2^d$:*

$$\mathbf{InSec}_{S1,\mathcal{C}}^{ss}(t, q, l) \leq \mathbf{InSec}_F^{prf}(t + O(k\ell(l)), k\ell(l))$$

Proof. For any warden, W, running in time t and making q queries totaling l bits, we construct a corresponding PRF adversary A, where

$$\mathbf{Adv}_{S1,\mathcal{C}}^{\mathrm{ss}}(W) = \mathbf{Adv}_F^{\mathrm{prf}}(A)$$

The running time of A is the running time of warden W plus the time of rejection sampling (RS): $O(k\ell(l))$ in the worst case. The number of calls to the sampling oracle, $M(h)$, is at most $k\ell(l)$.

A^f simply runs W, emulating the encoding procedure `S1.Encode` using the function oracle f in place of $F_K(\cdot,\cdot)$. Note that when f is a uniformly chosen random function, the output of $RS^{M(h),f}(\cdot, k)$ is distributed identically to the channel distribution $\mathcal{C}_h^b$. Similarly, when f is chosen from $F_K(\cdot,\cdot)$, the output of $RS^{M(h),f}(\cdot, k)$ is distributed identically to the output of Construction 1, by the definition of the construction. So the advantage is:

$$\begin{aligned}
\mathbf{Adv}_F^{\mathrm{prf}}(A) &= \left| \Pr_{K \leftarrow U(k), r \leftarrow \{0,1\}^*}[A_r^{F_K(\cdot,\cdot)} = 1] - \Pr_{g, r \leftarrow \{0,1\}^*}[A_r^g = 1] \right| \\
&= \left| \Pr_{K,r,M,SE}[W_r^{M,SE(K,\cdot,\cdot)} = 1] - \Pr_{r,M,O}[W_r^{M,O(\cdot,\cdot)} = 1] \right| \\
&= \mathbf{Adv}_{S1,\mathcal{C}}^{\mathrm{ss}}(W)
\end{aligned}$$

And the theorem follows by the definition of insecurity.

Corollary 1. *If $F_K(\cdot,\cdot)$ is pseudorandom then construction 1 is universally steganographically secret against chosen-hiddentext attacks.*

3.2 Stateless Universal Steganographic Secrecy

The following protocol also satisfies our definition for universal steganographic secrecy. The proofs are omitted due to space considerations, but are similar to the proof of Theorem 1 in the previous section, and can be found in the full version of the paper Once again, this protocol (up to small differences) is not new and can be found in [1]; an information theoretic version of the protocol can also be found in [3].

We will say that a function $f : D \to R$ is an *unbiased function on a distribution* $\mathcal{C}$ if for all $r \in R, h$, $\Pr_{d \leftarrow \mathcal{C}_h^b}[f(d) = r] = \frac{1}{|R|}$. Let $f : (\{0,1\}, t)^b \to \{0,1\}$ be a public function which is unbiased on $\mathcal{C}$. Let $E_K(\cdot,\cdot)$ and $D_K(\cdot)$ denote the encryption and decryption functions for a cryptosystem which is indistinguishable from random bits under chosen plaintext attack (i.e., IND\$-CPA) [12]. Suppose Alice and Bob share a key $K \in \{0,1\}^k$. The following procedures allow encoding and decoding of messages in a manner which is steganographically secret under chosen hiddentext attack for the channel distribution $\mathcal{C}$.

Construction 2. (Stateless Steganographic Secrecy)

Procedure `S2.Encode`:
Input: key K, hiddentext m, history h
Let $s = E_K(m)$
Parse s as $s_1^1 || s_2^1 || \cdots || s_l^1$
for $i = 1 \ldots l$ do
  $c_i \leftarrow RS^{M(h), f(\cdot)}(s_i, |K|)$
  set $h = h || c_i$
Output: $c_1 || c_2 || \ldots || c_l$

Procedure `S2.Decode`:
Input: key K, Stegotext c
Parse c as $c_1^b || c_2^b || \ldots || c_l^b$
for $i = 1 \ldots l$ do
  set $s_i = f(c_i)$
set $s = s_1 || s_2 || \cdots || s_l$.
Output: $D(K, s)$

Note that the execution time for SD is essentially linear in l and the execution time for SE is an expected $O(l)$. Also note that our assumption of an unbiased function implies that an error correcting code as in construction 1 is unnecessary: the probability of failure here is $l/2^{|K|}$. Clearly, the assumption of an unbiased f is quite strong. In practice, perfect unbiasedness may be hard to meet and stronger assumptions may be required. For example, when f is a pseudorandom function, under the assumption of a large minimum entropy in the channel, f will be nearly unbiased.

Lemma 2. *For any warden $W \in \mathcal{W}(t, q, l)$, we can construct a corresponding adversary A where*

$$\mathbf{Adv}^{ss}_{S2,\mathcal{C}}(W) = \mathbf{Adv}^{cpa}_{E}(A) \ .$$

The running time of A is at most $t + O(kl)$, and A makes q encryption queries for a total of l bits of plaintext.

Theorem 2. $\mathbf{InSec}^{ss}_{S2,\mathcal{C}}(t, q, l) \leq \mathbf{InSec}^{cpa}_{E}(t + O(kl), q, l)$.

Generalization. The assumption that the balanced function, f, is unbiased can be weakened to the assumption of an ϵ-biased function where the probability of any value is within ϵ of uniform. The same proofs work with the insecurity increased by at most ϵ (however, error correcting codes might be necessary in this case).

A few easy corollaries follow from Theorem 2. If E is indistinguishable from random bits under chosen plaintext attack then construction 2 is SS-CHA-$\mathcal{C}$ secure. Additionally, if E is replaced by a public key cryptosystem which is indistinguishable from random bits under chosen plaintext attack, then construction 2 is a public key stegosystem which is steganographically secret under chosen hiddentext attack (under an appropriate generalization of our definitions to a public-key scenario).

4 Robust Steganography

4.1 Definitions for Robust Steganography

Robust steganography will be modelled as a game between Alice and Ward in which Ward is allowed to make some alterations to Alice's messages. Alice wins if she can pass a message with high probability, even when Ward alters her message. For example, if Alice passes a single bit per channel message and Ward is unable to change the bit with probability at least $\frac{1}{2}$, Alice can use error correcting codes to reliably transmit her message. It will be important to state the limitations we impose on Ward, since otherwise he can replace all messages with a new draw from the channel distribution, effectively destroying any hidden information. In this section we give a formal definition of robust steganography with respect to a limited adversary.

We will model Ward's power as defined by a relation R which is constrained to not corrupt the channel too much. This general notion of constraint is sufficient to include many simpler notions such as (for example) "only alter at most 1% of the bits".

Let $\mathcal{D}$ be a finite distribution with support X and let R be a relation between the set X and the set Y such that for every $x \in X$, there exists a $y \in Y$ where $(x, y) \in R$. Consider a single-player game of chance which is played as follows:

1. The player draws x according to $\mathcal{D}$.
2. The player chooses an arbitrary y such that $(x, y) \in R$.
3. The player makes an independent draw x' from $\mathcal{D}$.

The player wins if $(x', y) \in R$. Define the *obfuscation probability* of R for $\mathcal{D}$ by

$$\mathcal{O}(R, \mathcal{D}) = \max_{y} \sum_{(x',y)\in R} \Pr_{\mathcal{D}}[x'] .$$

This function represents an upper bound on the player's winning probability. In particular, for any y the player chooses in step 2, $\mathcal{O}(R, \mathcal{D})$ bounds the probability $\sum_{(x',y)\in R} \Pr_{\mathcal{D}}[x']$ of winning. Note that the $\log_2 \mathcal{O}(R, \mathcal{D})$ gives the minimum amount of conditional information retained about draws from $\mathcal{D}$ when they are substituted arbitrarily amongst possibilities which satisfy R. The obfuscation probability is therefore a worst-case conditional entropy (just as minimum entropy is a worst-case entropy), except that logarithms have been removed.

Now let R be an efficiently computable relation on blocks and let $R(x) = \{y : (x,y) \in R\}$. We say that the pair $(R, \mathcal{C}_h^b)$ is δ-admissible if $\mathcal{O}(R, \mathcal{C}_h^b) \leq \delta$ and a pair $(R, \mathcal{C})$ is δ-admissible if $\forall h \;\; (R, \mathcal{C}_h^b)$ is δ-admissible. An R-bounded active warden W can be thought of as an adversary which plays the following game against a stegosystem $S = (SE, SD)$:

1. W is given oracle access to the channel distribution $\mathcal{C}$ and makes as many draws as it likes.
2. W is given oracle access to $SE(K, \cdot, \cdot)$, and makes at most q queries totaling at most l_1 bits to SE.
3. W presents an arbitrary message $m \in \{0,1\}^{l_2}$ and history h.
4. W is then given a sequence of blocks $c = c_1||c_2||\ldots||c_u$ from the support of $\mathcal{C}_h^{(u*b)}$, and returns a sequence $c' = c'_1||c'_2||\ldots||c'_u$ where $c'_i \in R(c_i)$ for each $1 \leq i \leq u$. Here u is the number of blocks of stegotext output by $SE(K, m, h)$.

Define the success of W against S by

$$\mathbf{Succ}_S^R(W) = \Pr_{K \leftarrow U(k), r \leftarrow \{0,1\}^*, o \leftarrow \{0,1\}^*} [SD_o(K, W_r(SE_o(K, m, h)), h) \neq m]$$

Here, r and o are the random bits used by Ward and the protocol, respectively. Define the failure rate of S by

$$\mathbf{Fail}_S^R(t, q, l) = \max_{W \in \mathcal{W}(R,t,q,l)} \left\{ \mathbf{Succ}_S^R(W) \right\} ,$$

where $\mathcal{W}(R, t, q, l)$ denotes the set of all R-bounded active wardens that submit at most q queries of total length at most l_1, produce a plaintext of length at most $l_2 = l - l_1$ and run in time at most t.

Definition 4. (Robust Steganography) A stegosystem $S = (SE, SD)$ is called $(t, q, l, \epsilon, \delta)$ *steganographically robust against R-bounded adversaries* for the distribution $\mathcal{C}$ (denoted $(t, q, l, \epsilon, \delta)$-SR-CHA-$(\mathcal{C}, R)$) if the following conditions hold:

– (Secrecy): S is (t, q, l, ϵ)-SS-CHA-$\mathcal{C}$.
– (Robustness): $\mathbf{Fail}_S^R(t, q, l) \leq \delta$.

A stegosystem is called $(t, q, l, \epsilon, \delta)$ *steganographically robust* (SR-CHA) if it is $(t, q, l, \epsilon, \delta)$-SR-CHA-$(\mathcal{C}, R)$ for every δ-admissible pair $(\mathcal{C}, R)$.

Definition 5. (Universal Robust Steganography) A sequence of stegosystems $\{S_k\}_{k \in \mathbb{N}}$ is called *universally steganographically robust* if it is universally steganographically secret and there exists a polynomial $q(\cdot)$ and a constant $\delta \in [0, \frac{1}{2})$ such that for every PPTM W, every δ-admissible $(R, \mathcal{C})$, and all sufficiently large k, $\mathbf{Succ}_{S_k}^R(W) < 1/q(k)$.

4.2 Universally Robust Stegosystem

In this section we give a stegosystem which is Steganographically robust against any bounding relation R, under a slightly modified assumption on the channel oracles, and assuming that Alice and Bob know some efficiently evaluable, δ-admissible relation R' such that R' is a superset of R. For several reasons, this stegosystem appears impractical but it serves as a proof that robust steganography is possible for any admissible relation.

Suppose that the channel distribution $\mathcal{C}$ is efficiently sampleable, that is, there is an efficient algorithm M which, given a uniformly chosen string $s \in \{0,1\}^m$ and history h produces a block distributed according to $\mathcal{C}_h^b$ (or statistically close to $\mathcal{C}_h^b$). We will assume that Alice, Bob, and Ward all have access to this algorithm. Furthermore, we assume Alice and Bob share a key K to a pseudorandom function; and have a synchronized counter N. Let n be a robustness parameter.

Construction 3. (Universally Robust Steganography)

Procedure S3.Encode	Procedure S3.Decode
Procedure `S3.Encode`:	**Procedure** `S3.Decode`:
Input: K, m, h	**Input:** key K, stegotext c, history h
Parse m as $m_1^1 \| m_2^1 \| \cdots \| m_l^1$	Parse c as $c_1^b \| c_2^b \| \ldots \| c_{ln}^b$
for $i = 1 \ldots l$ do	for $i = 1 \ldots l$ do
for $j = 1 \ldots n$ do	set $h_0 = h_1 = h$
set $c_{i,j} = M(F_K(N, m_i), h)$	for $j = 1 \ldots n$ do
increment N	for $\sigma \in \{0,1\}$ do
set $h = h \| c_{i,j}$	set $m_\sigma = M(F_K(N,\sigma), h_\sigma)$
Output: $c_{1,1} \| c_{1,2} \| \ldots \| c_{l,n}$	set $h_\sigma = h_\sigma \| m_\sigma$
	increment N
	if $(\forall j.(h_{0,j}, c_{i,j}) \in R')$
	then $p_i = 0$; else $p_i = 1$
	set $h = h_{p_i}$
	set $p = p_1 \| p_2 \| \cdots \| p_l$.
	Output: p

Suppose that instead of sharing a key to a pseudorandom function F, Alice and Bob shared two secret blocks b_0, b_1 drawn independently from $\mathcal{C}_h^b$. Then Alice could send Bob the message bit σ by sending block b_σ, and Bob could recover σ by checking to see if the block he received was related (by R') to b_0 or b_1. Since the adversary is R bounded and $(\mathcal{C}, R')$ is δ-admissible, the probability of a decoding error — caused either by the adversary, or by accidental draw of b_0, b_1 — would be at most δ. Intuitively, Construction 3 simply extends this notion to multiple bits by replacing the b_0, b_1 by draws from $M(\cdot)$ with shared pseudorandom inputs; and reduces the probability of decoding error to δ^n by encoding each hiddentext bit n times.

Lemma 3. $\mathbf{InSec}^{ss}_{S3,\mathcal{C}}(t,q,l) \leq \mathbf{InSec}^{prf}_{F}(t + O(nl), nl)$.
Lemma 4. $\mathbf{Fail}^{R}_{S3}(t,q,l_1,l_2) \leq \mathbf{InSec}^{prf}_{F}(t + O(nl), nl) + l_2\delta^n$.

Proof. Let W be an active R-bounded (t, q, l_1, l_2) warden. We construct a PRF adversary A which runs in time $t + O(nl)$, makes at most nl PRF queries, and satisfies $\mathbf{Adv}_F^{\mathrm{prf}}(A) \geq \mathbf{Succ}_S^R(W) - l_2\delta^n$. A^f works by running W, using its function oracle f in place of $F_K(\cdot, \cdot)$ to emulate Construction 3 in responding to the queries of W. Let m, c' be the hiddentext and the stegotext sequence returned by W, respectively. Then A^f returns 1 iff $SD(K, c', h) \neq m$. Consider the following two cases for f:

- f is chosen uniformly from all appropriate functions. Then, for each i, j, the stegotexts $c_{i,j} = M(f(N_i + j, p_i), h_{i,j})$ are distributed independently according to $\mathcal{C}^b_{h_{i,j}}$. Consider the sequence of "alternative stegotexts" $d_{i,j} = M(f(N_i + j, 1 - p_i), h_{i,j})$; each of these is also distributed independently according to $\mathcal{C}^b_{h_{i,j}}$; and since W is never given access to the $d_{i,j}$, the $c'_{i,j}$ are independent of the $d_{i,j}$. Now SD will fail (causing A^f to output 1) only if the event $\forall j.(d_{i,j}, c'_{i,j}) \in R'$ occurs for some i. Because the d_i are independent of the actions of W, and because $(\mathcal{C}, R')$ is δ-admissible, each event $(d_{i,j}, c'_{i,j}) \in R'$ happens independently with probability at most δ. So for any fixed i, the probability of failure is at most δ^n. The union bound then gives
$$\Pr_{f \leftarrow U(b,n)}[A^f = 1] \leq l_2\delta^n .$$
- f is chosen uniformly from $F_K(\cdot, \cdot)$. Then A^F outputs 1 exactly when W succeeds against S, by the definition of S:
$$\Pr_{K \leftarrow U(k), r \leftarrow \{0,1\}^*}[A_r^{F_K} = 1] = \mathbf{Succ}_S^R(W) \ .$$

Taking the difference of these probabilities, we get:

$$\begin{aligned}\mathbf{Adv}_F^{\mathrm{prf}}(A) &= \Pr_{K \leftarrow U(k), r \leftarrow \{0,1\}^*}[A_r^{F_K} = 1] - \Pr_{f \leftarrow U(b,n), r \leftarrow \{0,1\}^*}[A_r^f = 1] \\ &\geq \mathbf{Succ}_S^R(W) - l_2\delta^n \ .\end{aligned}$$

Theorem 3. *If F is $(t + O(nl), nl, \epsilon)$-pseudorandom then Construction 3 is $(t, l_1, l_2, \epsilon, \epsilon + l_2\delta^n)$-SR-CHA.*

4.3 Robust Steganography for text-like channels

We provide a protocol which is steganographically robust against R-bounded adversaries for a *particular* class of admissible relations R on *particular* channels. For some channel distributions this class of relations may provide an accurate model of the limitations of real wardens; in particular it seems reasonable to suppose that a predominantly text-based channel might limit the power of an active warden by such a relation.

A text based channel (such as email) might consist of "documents" built out of many bits with the same timestamp (= sending time). Fix a symbol alphabet $\Sigma = \{0, 1\}^s$, and represent every document d in the support of $\mathcal{C}$ by a string of symbols $d = d_0 d_1 \cdots d_l$ and a single timestamp. The n-shingling of d, denoted by $ss_n(d)$ is the set $\{d_0 d_1 \cdots d_{n-1}, d_1 d_2 \cdots d_n, d_2 d_3 \cdots d_{n+1}, \ldots, d_{l-n} d_{l-n+1} \cdots d_l\}$. Define the family of relations R_δ^n for $0 \leq \delta \leq 1$ by

$$(x, y) \in R_\delta^n \Leftrightarrow \frac{|ss_n(x) \cap ss_n(y)|}{|ss_n(x) \cup ss_n(y)|} \geq 1 - \delta \ .$$

Let $F_K(\cdot, \cdot)$ denote a pseudorandom function family indexed by $|K|$ bits which maps n-shingles to the set $\{0,1\}^p$. Let E_K be a length-preserving, stateful (rather than randomized), IND\$-CPA encryption scheme which works by producing a pseudorandom stream s and outputs $E_K(m) = s \oplus m$, (such a scheme can be produced using a pseudorandom function in (stateful) CTR mode, for example) and let D be the corresponding decryption function. Let Enc and Dec be the encoding and decoding algorithms for an error-correcting code which can correct up to a $\delta + \xi$ fraction of symbol errors; and let $\ell(\cdot)$ be the stretch function of this code. We will assume Enc and Dec work in linear time for simplicity. Define the function $\max_f(d)$ as the shingle in d which maximizes f, that is

$$\max_f(d) = \mathrm{argmax}_{s \in ss_n(d)} \{f(s)\} \ .$$

Let $g : \{0,1\}^* \rightarrow \{0,1\}$ be unbiased for the distribution on shingles induced by $\max_{F_K}(\mathcal{C})$. Let Alice and Bob share secret keys K_f, K_e. The following procedures allow Alice and Bob to robustly encode and decode a message.

Construction 4. (Robust Steganographic Protocol)

Procedure `S4.Encode`:	Procedure `S4.Decode`:
Input: m, h, keys K_f, K_e	**Input:** stegotext c, keys K_f, K_e
Let $p = E_{K_e}(Enc(m))$	Parse c as $c_1^b \| c_2^b \| \ldots \| c_l^b$
Parse p as $p_1^1 \| p_2^1 \| \cdots \| p_l^1$	for $i = 1 \ldots l$ do
for $i = 1 \ldots l$ do	set $s_i = \max_{F_{K_f}(\cdot)}(c_i)$
set $c_i = RS^{M(h), g \circ \max_{F_{K_f}}}(p_i)$	set $p_i = g(s_i)$
set $h = h \| c_i$	set $p = p_1 \| p_2 \| \cdots \| p_l$.
Output: $c_1 \| c_2 \| \ldots \| c_l$	**Output:** $Dec(D_{K_e}(p))$

Note that it is important that encryption and bit errors commute here which holds for only some encryption techniques.
In the following, Let ℓ_q be the maximum size of q encoded strings with total length l_1 plus $\ell(l_2)$.

Lemma 5. $\mathbf{InSec}_{S4}^{ss}(t, q, l) \leq \mathbf{InSec}_E^{cpa}(t + O(k\ell_q), q, \ell_q)$.

Lemma 6. $\mathbf{Fail}_{S4}^{R_\delta^n}(t, q, l_1, l_2) \leq 2e^{-\xi^2 \ell(l_2)/2} + \mathbf{InSec}_F^{prf}(t + O(k\ell_q), k\ell_q)$.

Theorem 4. *If F is $(t + O(k\ell_q), k\ell_q, \epsilon)$-pseudorandom and E is $(t + \ell_q, q, \ell_q, \mu)$ - IND\$-CPA, then Construction 4 is $(t, l_1, l_2, \epsilon + \mu, 2e^{-\xi^2 \ell(l_2)/2} + \epsilon)$ - SR-CHA against R_δ^n-bounded adversaries.*

5 Discussion

5.1 Alternative security conditions

There are several conceivable alternatives to our security conditions; we will briefly examine these alternatives and justify our choices.

Find-Then-Guess: This is the standard model in which an attacker submits two plaintexts p_0 and p_1, receives $SE(p_b)$, and attempts to guess b. Security in our attack model implies find-then-guess security; moreover the essence of steganographic secrecy is not merely the inability to distinguish between messages (as in the find-then-guess model) but the inability to detect a message.

Fixed History: In this model the adversary may not submit alternate histories to the encryption model. Security under a chosen-history attacks implies security against a fixed-history attacks. This notion may be of interest however, especially because in many situations a chosen-history attack may not be physically realizable. Our attacks can be considered chosen-history attacks.

Integrity of Hiddentexts. Intuitively, Integrity of Hiddentexts requires that an active warden is unable to create a sequence of covertexts which decodes to a valid, new hiddentext. Suppose we amend the description of a stego system to allow the decoding algorithm to output the "fail" symbol $\perp$. Then suppose we give the adversary oracle access to SE and allow the adversary to make at most q queries $p_0, \ldots, p_q$ to $SE(K, \cdot, \cdot)$ totaling l bits. The adversary then produces a sequence of covertexts $c = c_1 || \ldots || c_m$. Denote the advantage of A against S by

$$\mathbf{Adv}^{\text{int}}_{S,\mathcal{C}}(A) = \Pr\left[SD(K, \boldsymbol{c}, h) \neq \perp \wedge \forall i. SD(K, \boldsymbol{c}, h) \neq p_i\right] \ ,$$

and denote the integrity failure of a stegosystem by

$$\mathbf{Fail}^{int}_{S,\mathcal{C}}(t, q, l) = \max_{A \in \mathcal{A}(t,q,l)} \left\{ \mathbf{Adv}^{\text{int}}_{S,\mathcal{C}}(A) \right\} \ .$$

A stegosystem has (t, q, l, ϵ) integrity of hiddentexts if $\mathbf{Fail}^{\text{int}}_{S,\mathcal{C}}(t, q, l) \leq \epsilon$.

Note that in practice this notion by itself is too weak because it allows the possibility for the warden to disrupt the communication between Alice and Bob. Finally, we note that if the symmetric encryption scheme E is INT-CTXT secure as defined by Bellare and Namprempre [2], then construction 2 also provides integrity of hiddentexts.

5.2 Complexity theoretic ramifications

Construction 1 gives a stegosystem which is steganographically secret for any channel distribution $\mathcal{C}$ which has minimum entropy greater than 1, assuming the existence of a pseudorandom function family. Goldreich *et al* [4] show how to construct a pseudorandom function from a pseudorandom generator, which in turn can be constructed from any one-way function, as demonstrated by Hastad *et al* [5]. Thus in an asymptotic sense, our constructions show that one-way functions are sufficient for steganography. Conversely, it is easy to see that a stegosystem which is steganographically secret for some $\mathcal{C}$ is a secure weak private key encryption protocol in the sense of Impagliazzo and Luby [7]; and they prove that the existence of such a protocol implies the existence of a one-way function. Thus the existence of secure steganography is equivalent to the existence of one-way functions.

Acknowledgments We are greatful to Manuel Blum for his suggestions and his unconditional encouragment. We also thank Steven Rudich and the anonymous CRYPTO reviewers for helpful discussions and comments. This work was partially supported by the National Science Foundation (NSF) grants CCR-0122581 and CCR-0085982 (The Aladdin Center). Nicholas Hopper is also partially supported by an NSF graduate research fellowship.

References

1. Ross J. Anderson and Fabien A. P. Petitcolas. *On The Limits of Steganography.* IEEE Journal of Selected Areas in Communications, 16(4). May 1998.
2. Mihir Bellare and Chanathip Namprempre. *Authenticated Encryption: Relations among notions and analysis of the generic composition paradigm.* In: *Advances in Cryptology – Asiacrypt '00.* December 2000.
3. C. Cachin. *An Information-Theoretic Model for Steganography.* In: *Information Hiding – Second International Workshop, Preproceedings.* April 1998.
4. Oded Goldreich, Shafi Goldwasser and Silvio Micali. *How to Construct Random Functions.* Journal of the ACM, 33(4):792 – 807, 1986.
5. Johan Hastad, Russell Impagliazzo, Leonid A. Levin, and Michael Luby. *A pseudorandom generator from any one-way function.* SIAM Journal on Computing, 28(4):1364–1396, 1999.
6. Nicholas J. Hopper, John Langford, and Luis von Ahn. *Provably Secure Steganography.* CMU Tech Report CMU-CS-TR-02-149, 2002.
7. Russell Impagliazzo and Michael Luby. *One-way Functions are Essential for Complexity Based Cryptography.* In: 30th FOCS, November 1989.
8. D. Kahn. *The Code Breakers.* Macmillan 1967.
9. Stefan Katzenbeisser and Fabien A. P. Petitcolas. *Information hiding techniques for steganography and digital watermarking.* Artech House Books, 1999.
10. T. Mittelholzer. *An Information-Theoretic Approach to Steganography and Watermarking* In: *Information Hiding – Third International Workshop.* 2000.
11. J. A. O'Sullivan, P. Moulin, and J. M. Ettinger *Information theoretic analysis of Steganography.* In: *Proceedings ISIT '98.* 1998.
12. Phillip Rogaway, Mihir Bellare, John Black and Ted Krovetz. *OCB: A Block-Cipher Mode of Operation for Efficient Authenticated Encryption.* In: *Proceedings of the Eight ACM Conference on Computer and Communications Security (CCS-8).* November 2001.
13. G.J. Simmons. *The Prisoner's Problem and the Subliminal Channel.* In: *Proceedings of CRYPTO '83.* 1984.
14. J Zollner, H.Federrath, H.Klimant, A.Pftizmann, R. Piotraschke, A.Westfield, G.Wicke, G.Wolf. *Modeling the security of steganographic systems.* In: *Information Hiding – Second International Workshop, Preproceedings.* April 1998.