Thesis Proposal:
Toward a Theory of Steganography

Nicholas Hopper

School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213

Abstract

Informally, *steganography* refers to the practice of hiding secret messages in communications over a public channel so that an eavesdropper (who listens to all communications) cannot even tell that a secret message is being sent. In contrast to the active literature proposing new ad hoc steganographic protocols and analyzing flaws in existing protocols, there has been very little work on formalizing steganographic notions of security, and none giving complete, rigorous proofs of security in any non-trivial model.

My thesis will initiate the study of steganography from a complexity-theoretic point of view. Modelling the communication between two parties as a probabilistic *channel*, I will introduce several different steganographic security properties, and for each of these properties, attempt to determine the necessary and sufficient conditions under which secure steganography is possible.

Furthermore, I propose to investigate the open question "How much information can be safely encoded by a stegosystem?"[3] by proving an upper bound MAX on the bit rate of any secure, universal stegosystem; and giving a protocol which achieves rate $\Omega(MAX)$, thus providing a universal stegosystem of approximately optimal rate.

1 Introduction

The scientific study of steganography in the open literature began in 1983 when Simmons [11] 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 "unusual" is going on. Notice how the goal of 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 [7]. 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 thesis is to give a rigorous, complexity-theoretic formulation of steganography. We define steganographic secrecy in terms of computational indistinguishability, and we explore steganographic robustness, which deals with the case of active wardens (ones that cross out innocent-looking sentences or modify the messages just to prevent successful transmission of secrets). The main contributions of this thesis will be:

- We will show that, relative to an oracle for a given channel distribution, secure private key steganography exists if and only if one-way functions exist; and in the plain model, steganography over a given channel is possible only if the channel is efficiently sampleable.
- We will show that, for any channel with sufficient minimum entropy, secure public key cryptography (information-theoretically impossible) exists if trapdoor one-way predicates on *dense domains* exist, and secure steganographic key exchange is possible if the Integer Decisional Diffie-Hellman (DDH) assumption holds.
- We will show that, in the private key setting, *robust* steganography against reasonably-bounded adversaries is possible for any efficiently sampleable channel; and in some cases, our protocol can be employed in a public-key setting.
- We will give an upper bound on the maximum ratio of hidden bits to transmitted bits achievable by a secure steganographic protocol, and a matching (up to a constant factor) lower bound, resolving an open question first posed by Anderson and Petitcolas[3].

2 Secret-Key Steganography

Secret-key steganography is the most basic setting for steganography: Alice and Bob possess a shared secret key and would like to use it to exchange hidden messages over a public channel so

that Ward cannot detect the presence of these messages. Despite the apparent simplicity of this scenario, there has been little work on giving a precise formulation of steganographic security. Our goal is to give such a formal description.

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. Surprisingly, though, very little work has attempted to formalize steganography, and most of the literature consists of heuristic approaches: steganography using digital images, steganography using video systems, etc. A few papers have given information theoretic models for steganography [4, 8, 10, 12], but these are limited in the same way that information theoretic cryptography is limited.

2.1 Preliminaries

A function $\mu : \mathbb{N} \to (0,1)$ 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. If X is a finite set, we let $U(X)$ denote the uniform distribution on X.

2.2 Cryptographic notions

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
- $\$(\cdot) = U_{|E_K(\cdot)|}$; 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 $q(k)$ queries to the oracle of at most $l(k)$ bits and run for $t(k)$ time steps. Define the CPA advantage of A against E as

$$\mathbf{Adv}^{\mathsf{cpa}}_{A,E}(k) = \left| \Pr[A^{E_K}(1^k) = 1] - \Pr[A^{\$}(1^k) = 1] \right|$$

where the probabilities are taken over the oracle draws and the randomness of A. Define the insecurity of E as

$$\mathbf{InSec}^{\mathsf{cpa}}_{E}(t, q, l, k) = \max_{A \in \mathcal{A}(t,q,l)} \left\{ \mathbf{Adv}^{\mathsf{cpa}}_{A,E}(k) \right\} .$$

Then E is (t, q, l, k, ϵ)-*indistinguishable from random bits under chosen plaintext attack* if $\mathbf{InSec}^{\mathsf{cpa}}_{E}(t, q, l, k) \leq \epsilon$. 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, $\mathbf{Adv}^{\mathsf{cpa}}_{A,E_k}(k)$ 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

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

We say that a function $f : X \to \{0,1\}$ is ϵ-*biased* if

$$\left| \Pr_{x \leftarrow \mathcal{C}}[f(x) = 0] - \frac{1}{2} \right| < \epsilon \ .$$

We say f is *unbiased* if f is ϵ-biased for ϵ a negligible function of the appropriate security parameter.

2.3 Channels

We seek to define steganography in terms of indistinguishability from a certain "usual" distribution on communications. In order to do so, we must characterize this distribution. We formalize this with the notion of a channel.

Definition. Let D be an efficiently recognizable, prefix-free set of strings, or *documents*. A *channel* is a distribution on sequences $s \in D^*$.

Any particular sequence in the support of a channel describes one possible outcome of all communication over this channel. Alternatively, communication on a channel can be regarded as iteratively drawing from the channel, that is, drawing a document from a distribution consistent with the history h of already drawn documents, obtained by an appropriate generalization of Bayes' Rule. 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.

Since anyone communicating over a channel can be seen as implicitly drawing from these marginal channel distributions, we will assume the existence of an probabilistic oracle capable of doing so. 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 marginal channel distribution conditioned on the history h of already drawn documents; we let $\mathcal{C}_h^l$ denote the marginal distribution on sequences of l documents conditioned on the history h. When we write "sample $x \leftarrow \mathcal{C}_h$" we mean that the oracle should be queried using history h.

We will require that the channel satisfy a minimum entropy constraint for all histories. Specifically, we require that there exist constants $b > 0$, $\alpha > 0$ such that

$$\forall h : \Pr_{\mathcal{C}}[h] = 0 \text{ or } H_\infty(\mathcal{C}_h^b) \geq \alpha \ .$$

If a channel does not satisfy this property, then it is possible for Alice to drive the information content of her channel to 0, so this is a reasonable requirement. If a channel satisfies this requirement, we say it is *always informative* (AINF). Note that always informativeness implies an additive-like property of minimum entropy for marginal distributions, specifically, $H_\infty(\mathcal{C}_h^{lb}) \geq l\alpha$.

2.4 Stegosystems

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*), and a message history $h \in D^*$. $SE(K, m, h)$ returns a sequence of documents $c_1, c_2, \ldots, c_l$ (the *stegotext*) from the support of $\mathcal{C}_h^l$. SD takes a key K, a sequence of documents $c_1, c_2, \ldots, c_l$, and a message history $h \in D^*$, and returns a hiddentext m. In addition, for all polynomials $p(k)$, there must be a negligible function $\mu(k)$ such that, for sufficiently large k, SE and SD also satisfy the relationship:

$$\forall m \in \{0,1\}^{p(k)} : \quad \Pr_K[SD(K, SE(K, m, h), h) = m] \geq 1 - \mu(k) \ ,$$

where the probability is also taken over any coin tosses of SE, SD, and the oracle for $\mathcal{C}$.

2.5 Steganographic Secrecy

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

1. W is given access to a probabilistic oracle which samples documents (one at a time) from the distribution $\mathcal{C}_h$.
2. W is given access to a second oracle which is sampled from one of two distributions: $ST(\cdot,\cdot)$ or $CT(\cdot,\cdot)$. Here a sample from ST draws a key K and responds to $ST(m,h)$ with $SE(K,m,h)$; a sample from CT responds to queries $CT(m,h)$ with samples from $\mathcal{C}_h^{|SE(K,m,h)|}$. Ward makes at most q oracle queries totalling l bits.
3. W outputs a bit.

We define W's advantage against a stegosystem S by

$$\mathbf{Adv}^{\mathsf{ss}}_{W,S,\mathcal{C}}(k) = \left|\Pr[W^{ST}(1^k)=1] - \Pr[W^{CT}(1^k)=1]\right| \ ,$$

where the probabilities are taken over the choice of oracle and the random bits of W. Define the insecurity of S by

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

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,k,ϵ) *steganographically secret against chosen hiddentext attack* for the channel $\mathcal{C}$ ((t,q,l,k,ϵ)-SS-CHA-$\mathcal{C}$) if $\mathbf{InSec}^{\mathsf{ss}}_{S,\mathcal{C}}(t,q,l) \le \epsilon$. A sequence of stegosystems $\{S_k\}_k$ is called steganographically secret against chosen hiddentext attack (SS-CHA) if $\mathbf{Adv}^{\mathsf{ss}}_{W,S_k,\mathcal{C}}(k)$ is negligible in k for every polynomial time W.

2.6 Construction: Steganographic Secrecy

We now give a protocol which is steganographically secret, for any channel which admits an unbiased function. The precise security bounds, and not the protocol itself, are the novel aspect of this work: similar protocol ideas have appeared in the literature[4, 3].

Let $f : D \to \{0,1\}$ be a public function which has maximum bias ϵ. 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. Assume that $|E_K(m)| \le |m| + \kappa$.

Construction 1. (Secret-key Steganography)

Procedure SS.Encode:
Input: key K, hiddentext m, history h
Let $s \leftarrow E_K(m)$
Parse s as $s_1^1||s_2^1||\cdots||s_l^1$
for $i = 1 \ldots l$ do
  repeat:
    $c_i \leftarrow \mathcal{C}_h$
  until $f(c_i) = s_i$ OR (too many failures)
  set $h = h||c_i$
Output: $c_1, c_2, \ldots, c_l$

Procedure SS.Decode:
Input: key K, stegotext $c \in D^*$
Let $c = (c_1, c_2, \ldots, c_l)$
for $i = 1 \ldots l$ do
  set $s_i = f(c_i)$
set $s = s_1||s_2||\cdots||s_l$.
Output: $D_K(s)$

Notice that, since any channel which is always informative can be compiled to a channel which admits a neglibly-biased function, this protocol gives a provably secure stegosystem for any AINF channel.

Theorem 1. $\mathbf{InSec}^{\mathsf{ss}}_{\mathsf{SS},\mathcal{C}}(t,q,l,k) \leq \mathbf{InSec}^{\mathsf{cpa}}_{E}(t+O(kl),q,l,k) + (l+q\kappa)\epsilon$.

Proof. Suppose we are given a program W for distinguishing between the output of construction 1 and the ordinary traffic of the channel, and an oracle for sampling blocks from that channel. We construct a program A which plays the CPA game — distinguishing an E_K oracle from a uniform \$ oracle — with the same advantage as W. A simply runs W, using the encoding procedure $\mathsf{SS.Encode}$ with its oracle in place of calls to E_K to respond to W's queries. Consider the following two cases:

- $s \leftarrow E_K(m)$. Then the stegotexts output by the encoding procedure will be identically distributed to stegotexts resulting from the normal use of construction 1.
- $s \leftarrow \$(m)$ is chosen uniformly from strings of appropriate length. Then the statistical distance between the stegotexts output by the encoding procedure and a history-dependent sample from the channel distribution $\mathcal{C}_h$ will be at most $(|m|+\kappa)\epsilon$. This follows by the fact that f has bias at most ϵ on $\mathcal{C}_h$ and the parsed substrings s_i are uniformly distributed on $\{0,1\}$.

Thus A can simply use the decision of W to gain advantage identical to that of W. More formally,

$$\begin{aligned}
\mathbf{Adv}^{\mathsf{cpa}}_{E,A}(k) &= \left|\Pr[A^{E_K}(1^k)=1] - \Pr[A^{\$}(1^k)=1]\right| \\
&= \left|\Pr[W^{ST}(1^k)=1] - \Pr[A^{\$}(1^k)=1]\right| \\
&\leq \left|\Pr[W^{ST}(1^k)=1] - \Pr[W^{CT}(1^k)=1]\right| + (l+q\kappa)\epsilon \\
&= \mathbf{Adv}^{\mathsf{ss}}_{S1,\mathcal{C}}(W) + (l+q\kappa)\epsilon
\end{aligned}$$

□

2.7 Other Results and Extensions

The results in this section were reported at CRYPTO 2002 [6]. In that work, we also prove that, relative to an oracle for the channel $\mathcal{C}$, secure steganographic protocols exist if and only if one-way functions exist. The protocol we outline there requires only the ability to draw two independent samples from (a distribution computationally indistinguishable from) $\mathcal{C}_h$ for any h. Notice that, in the plain model, this condition is also *necessary* for a secure stegosystem for $\mathcal{C}$, because any $\mathsf{SS\text{-}CHA}$-secure stegosystem in the plain model gives an algorithm capable of drawing samples from $\mathcal{C}_h$: choose a random key K and output (the first document of) $SE(K,1,h)$. This observation also holds for a stegosystem which is only secure against a weaker attack, in which the adversary may choose the initial history h and thereafter does not control either h or the hiddentexts m being encoded. My thesis will formally develop this weaker model and prove that in the plain model, the necessary and sufficient conditions for the existence of secret-key steganography for channel $\mathcal{C}$ are an efficient sampler for $\mathcal{C}$ and the existence of one-way functions.

3 Public-Key Steganography

The results described in the previous section assume that the sender and receiver share a secret, randomly chosen key. In the case that some exchange of key material was possible before the use of steganography was necessary, this may be a reasonable assumption. In the more general case, two parties may wish to communicate steganographically, without prior agreement on a secret key. We call such communication *public steganography*. Whereas previous work has shown that private-key steganography is possible – though inefficient – in an information-theoretic model, public steganography is information-theoretically *impossible*. Thus our complexity-theoretic formulation of steganographic secrecy is crucial to the security of the constructions in this section.

3.1 Related Work

Anderson and Petitcolas [2], and Craver [5], have both previously described ideas for public-key steganography. This work will differ from theirs in several significant ways:

1. [2] and [5] do not attempt to give rigorous definitions for security, and give only heuristic arguments for the security of their constructions. In contrast, we will give a rigorous definition and proof of security for public-key steganography.

2. [2] does not describe any mechanism for generating stegotexts, but simply assumes "the ability to manipulate some bits of the cover". Similarly, [5] assumes the existence of a "supraliminal function" F and the ability to generate a covertext which has $F(x) = y$ for arbitrary y. In contrast, our model does not assume the existence of a function with non-standard properties, and is constructive.

3. [2] confuses *decoding* with *detection* in its security argument. Thus they do not make clear what are the requirements on the underlying public-key cryptographic primitives. In contrast, we state exact requirements and give tight security bounds.

To the best of our knowledge, we are the first to provide a formal framework for public-key steganography and to *prove* that public key steganography is possible (given that standard cryptographic assumptions hold).

3.2 Public-key Steganography

Definition 3. (Stegosystem) A public-key steganographic protocol, or public-key stegosystem, is a triple of probabilistic algorithms $S = (SG, SE, SD)$. $SG(1^k)$ generates a key pair $(PK, SK) \in \mathcal{PK} \times \mathcal{SK}$. $SE : \mathcal{PK} \times \{0,1\}^* \times D^* \to D^*$ and $SD : \mathcal{SK} \times D^* \times D^* \to \{0,1\}^*$ retain the same behavior, except that SE is keyed by $\mathcal{PK}$ and SD is keyed by $\mathcal{SK}$. The stegosystem must satisfy the standard soundness condition: for every polynomial $p(k)$, there is a negligible function $\mu(k)$ such that for all sufficiently large k,

$$\forall m \in \{0,1\}^{p(k)} : \Pr_{(PK,SK) \leftarrow SG(1^k)} [SD(SK, SE(PK, m, h), h) = m] \geq 1 - \mu(k)$$

where the randomization is over any coin tosses of SE, SD, and SG.

Steganographic Secrecy

We can model steganographic secrecy in the public-key case analogously to steganographic secrecy for the private-key case; the only significant difference in our basic definitions would be that the warden should be given the public key generated by Bob. However, the public-key case also allows the possibility of stronger attacks. For example, the warden can detect the use of steganography by Bob simply by encoding a message, sending it to Bob and watching his reaction: if he reacts consistently with receiving the warden's message, then he is probably decoding messages. Thus the warden's goal should be to detect whether a specific pair, Alice and Bob are communicating steganographically. To protect against such an attack will require that Alice have some secret differentiating herself from the warden: we will allow Alice to publish a verification key for a signature scheme and keep the signing key secret. In this model, we will define additional attack games to the basic chosen-hiddentext attack: the Chosen Stegotext Only attack, the Chosen Exactly One Attack, and the Chosen Stegotext and Hiddentext attack.

In all of these models, we will continue to model a warden attacking a stegosystem as an efficient oracle machine which plays an oracle-distinguishing game:

1. W is given access to all public keys and to oracles which sample documents (one at a time) from the marginal channel distributions $\mathcal{C}_{P\to Q,h}$ for every pair of parties (P, Q). (This oracle allows W to learn the *covertext* distribution on communications between Alice and Bob.)

2. W is given access to a second oracle which is either ST_{atk} or CT_{atk}. The oracle ST will model the case that the pair Alice and Bob are communicating steganographically, while the oracle CT will model the case that they are not. The exact distributions over $ST_{\text{atk}}, CT_{\text{atk}}$ vary depending on the attack model. W makes at most q queries totaling l bits to this oracle.

3. W outputs a bit.

We define W's advantage against a stegosystem S and channel $\mathcal{C}$, in attack model atk by

$$\mathbf{Adv}^{\mathsf{atk}}_{W,S,\mathcal{C}}(k) = \left|\Pr[W^{ST_{\text{atk}}}(1^k) = 1] - \Pr[W^{CT_{\text{atk}}}(1^k) = 1]\right| ,$$

where the probabilities are taken over the choice of oracle and the random bits of W. Define the insecurity of S in channel $\mathcal{C}$ by

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

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

Chosen Hiddentext Attack (CHA). In a chosen hiddentext attack, W is allowed to supply hiddentexts to Alice, in hopes of detecting whether or not she is using steganography in her communications with Bob. Thus W hopes to distinguish, for as many chosen messages m as he needs, between the case that Alice's communications are drawn from $SE(PK_B, m, h)$ or from $\mathcal{C}_{A\to B,h}^{|SE(PK_B,m,h)|}$ (where all m and h are chosen by W). Formally, we define the following distributions on ST_{cha} and CT_{cha}:

- To sample from ST_{cha}: draw $(PK, SK) \leftarrow SG(1^k)$ once, and respond to all queries with $ST_{\text{cha}}(m, h) \leftarrow SE(PK, m, h)$.
- To sample from CT_{cha}: draw $(PK, SK) \leftarrow SG(1^k)$ once, and respond to all queries with $CT_{\text{cha}}(m, h) \leftarrow \mathcal{C}_h^{|SE(PK,m,h)|}$.

Thus, the key pair (PK, SK) is drawn once and fixed, and Ward is allowed to query his oracle with arbitrary (m, h) pairs. (Note that a chosen hiddentext attack is the steganographic equivalent of a chosen plaintext attack against an encryption scheme.)

Chosen Stegotext-Only Attack (CSO). A chosen stegotext-only attack models the situation where Ward wants to determine if Bob is receiving hiddentexts from Alice in particular. Ward is able to "ask" Bob to decode messages sent to him by observing Bob's reaction to stegotexts *generated by Ward only.* In this case, Ward's goal is to detect whether Bob is receiving steganography from Alice by watching Bob's reaction to stegotexts (allegedly from Alice) generated by Ward himself. If Bob is not receiving steganography from Alice, he will always output the distinguished message $\perp$, meaning "no reaction." If Bob is receiving steganography from Alice, he will decode the stegotexts. (Thus it follows that, for SS-CSO security, SD should normally output $\perp$ as well.) This game is modeled by giving Ward an oracle for SD (rather than an oracle for SE as in the previous case), and asking Ward to determine if SD is decoding messages that come from Alice. Formally, we define the following distributions for ST_{cso} and CT_{cso}:

- To sample from ST_{cso}: draw $(PK, SK) \leftarrow SG(1^k)$ once, and respond to all queries with $ST_{\text{cso}}(c, h) \leftarrow SD(SK, c, h)$.
- To sample from CT_{cso}: $CT_{\text{cso}}(c, h) = \perp$.

Note that SS-CSO security is extremely weak, and in general it can be achieved simply by appending signatures to messages sent in the clear. SS-CSO is thus mainly useful in combination with SS-CHA.

Chosen Exactly-One Attack (CXO). In a chosen exactly-one attack, Ward may both submit hiddentexts to Alice for encoding, and submit stegotexts to Bob for decoding, but *he is prohibited from accessing both at the same time* (i.e., with the same history). This may be the case if for example they both travel extensively. As in all of our attack models, Ward's goal is to determine if the specific pair of Alice and Bob are communicating steganographically. We formally define the oracle distributions $ST_{\text{cxo}}, CT_{\text{cxo}}$ as follows: First, draw $(PK, SK) \leftarrow SG(1^k)$, and set $\phi = \{\}$. Respond to queries using these programs:

$ST_{\text{cxo}}(b \in \{\texttt{enc}, \texttt{dec}\}, m, h)$	$CT_{\text{cxo}}(b \in \{\texttt{enc}, \texttt{dec}\}, m, h)$
if (b = `enc`) then:	if (b = `enc`) then:
Sample $c \leftarrow SE(PK, m, h)$	Sample $c \leftarrow \mathcal{C}_h^{\lvert SE(PK,m,h)\rvert}$
Set $\phi = \phi \cup \{h\}$	Set $\phi = \phi \cup \{h\}$
return c	return c
else	else
If $h \in \phi$ return ""	If $h \in \phi$ return ""
else return $SD(SK, m, h)$	else return $\perp$

Note that $\mathbf{InSec}^{\mathsf{cso}}_{S,\mathcal{C}}(t, q, l, k) \leq \mathbf{InSec}^{\mathsf{cxo}}_{S,\mathcal{C}}(t, q, l, k)$, since any CSO warden can be emulated by a CXO warden making only $(\texttt{dec}, c, h)$-queries. Similarly, $\mathbf{InSec}^{\mathsf{cha}}_{S,\mathcal{C}}(t, q, l, k) \leq \mathbf{InSec}^{\mathsf{cxo}}_{S,\mathcal{C}}(t, q, l, k)$.

Chosen Stegotext and Hiddentext Attack (CSH) We formally define the oracle distributions $ST_{\text{csh}}, CT_{\text{csh}}$ as follows: First, draw $(PK, SK) \leftarrow SG(1^k)$, and set $\phi = \{\}$. Respond to queries using these programs:

ST_{csh} $(b \in \{\texttt{enc}, \texttt{dec}\}, m, h)$
if(b = enc) then:
  Sample $c \leftarrow SE(PK, m, h)$
  Set $\phi = \phi \cup \{(c, h)\}$
  return c
else
  If $(m, h) \in \phi$ return ""
  else return $SD(SK, m, h)$

CT_{csh} $(b \in \{\texttt{enc}, \texttt{dec}\}, m, h)$
if(b = enc) then:
  Sample $c \leftarrow \mathcal{C}_h^{|SE(PK,m,h)|}$
  Set $\phi = \phi \cup \{(c, h)\}$
  return c
else
  If $(m, h) \in \phi$ return ""
  else return $\perp$

Thus, in a chosen-stegotext attack, Ward may ask Alice to encode any (message, history) pair of his choosing, as often as he likes, and may query Bob on any pair (s, h) where s was not a result of an encoding query for history h. Notice that if Ward can ask Bob to decode any message output by Alice for the same history it was encoded with, he can detect the use of steganography between Alice and Bob; this is why we do not allow Ward to query Bob on such stegotexts. This restriction is roughly analogous to the standard restriction that an adaptive chosen-ciphertext attacker may not query his decryption oracle on the challenge ciphertext. Advantage and insecurity for SS-CSH are defined analogously to SS-CXO, except that we count encoding and decoding queries separately (as q_e and q_d) as well as counting the number of queries made to random oracles.

3.3 Steganographic Key Exchange

A natural strategy for implementation of Public-key Steganography is to consider the possibility of *Steganographic Key Exchange*: Alice and Bob exchange a sequence of messages, indistinguishable from normal traffic on the channel, and at the end of this sequence, they are able to compute a shared key. So long as this key is indistinguishable from a random key to the warden, Alice and Bob can proceed to use their shared key in a secret-key stegosystem. In my thesis, I will formalize the notion of a Steganographic Key Exchange Protocol (SKEP).

3.4 Results and Extensions

All of the results for this section are work in progress, summarized by a recent submission to CRYPTO 2003 [1]. There we show that any "dense domain" semantically secure public-key cryptosystem can be used to construct a public-key cryptosystem which is indistinguishable from random bits. Combining this result with an appropriate signature scheme and Construction 1 gives a public-key stegosystem which is SS-CXO secure. Furthermore, we show that if the channel $\mathcal{C}$ is efficiently sampleable, we can construct a stegosystem which is SS-CSH secure in the random oracle model, if trapdoor one-way permutations on the domain $\{0,1\}^k$ exist. We also show that the Decisional Diffie-Hellman assumption implies the existence of secure SKEPs, by using a novel application of the Chinese Remainder Theorem to convert Diffie-Hellman (DH) triples over $\mathbb{Z}_P^*$ to DH triples over a prime-order subgroup.

I plan to investigate two extensions to this work:

SS-CSH security without Random Oracles. It would be nice to give a SS-CSH-secure stegosystem in the standard model. The main problem in constructing such a stegosystem seems to be constructing a public-key cryptosystem which is indistinguishable from random bits and non-malleable; I am unaware of a published cryptosystem with proof of security in the standard model which meets this requirement.

Tighter cryptographic assumptions. All of these constructions require a public-key cryptosystem which is indistinguishable from random bits; such cryptosystems can be constructed from

dense domain trapdoor predicates, which are (provably) a stronger assumption than the existence of public-key cryptography. It would be nice to give a tighter condition for the existence of such cryptosystems.

4 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 document 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 the constraint on Ward's power by a relation R which is constrained to not corrupt the channel too much. That is, if Alice sends document d, Bob must receive a document d' such that $(d, d') \in R$. This general notion of constraint is sufficient to include many simpler notions such as (for example) "only alter at most 1% of the bits".

Consider the question of what conditions on the relation R are necessary to allow communication to take place between Alice and Bob. Surely it should not be the case that $R = D \times D$. Also, in case there is some document d' and history h for which $\sum_{(d,d')\in R} \Pr_{\mathcal{C}_h}[d] = 1$ then when h has transpired, Ward can effectively prevent the transfer of information from Alice to Bob by sending the document d' regardless of the document transmitted by Alice. Since we model the attacker as controlling the history h, then, a necessary condition on R and $\mathcal{C}$ for robust communication is that

$$\forall h. \Pr_{\mathcal{C}}[h] = 0 \text{ or } \max_{y} \sum_{(x,y)\in R} \Pr_{\mathcal{C}_h}[x] < 1 \ .$$

We denote by $\Delta(R, D)$ the function $\max_y \sum_{(x,y)\in R} \Pr_D[x]$. We say that the pair $(R, \mathcal{C}_h)$ is δ-admissible if $\Delta(R, \mathcal{C}_h) \leq \delta$ and a pair $(R, \mathcal{C})$ is δ-admissible if $\forall h \ \ \Pr_{\mathcal{C}}[h] = 0$ or $\Delta(R, \mathcal{C}_h) \leq \delta$. Our necessary condition states that $(R, \mathcal{C})$ must be δ-admissible for some $\delta < 1$.

We model an R-bounded active warden W as an adversary which plays the following game against a stegosystem $S = (SE, SD)$:

1. W is given oracle access to the channel distribution $\mathcal{C}$.
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 documents $c = (c_1, c_2, \ldots c_\ell) \leftarrow SE(K, m, h)$, and produces a sequence $s = (s_1, \ldots, s_\ell) \in D^\ell$, where $(c_i, s_i) \in R$ for each $1 \leq i \leq \ell$.

Define the success of W against S by

$$\mathbf{Succ}^R_{W,S}(k) = \Pr[SD(K, s, h) \neq m] \ ,$$

where the probability is taken over the choice of K and the randomness of S and W. Define the failure rate of S by

$$\mathbf{Fail}^R_S(t, q, l_1, l_2, k) = \max_{W\in\mathcal{W}(R,t,q,l_1,l_2)} \left\{\mathbf{Succ}^R_{W,S}(k)\right\} \ ,$$

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

Definition 4. A sequence of stegosystems $\{S_k\}_{k\in\mathbb{N}}$ is called *steganographically robust* for $\mathcal{C}$ against R if it is steganographically secret for $\mathcal{C}$ and there is a negligible function $\mu(k)$ such that for every PPTM W, for all sufficiently large k, $\mathbf{Succ}^R_{S_k}(W) < \mu(k)$.

4.1 Construction: 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 $R \subset R'$. 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 $s \in \{0,1\}^k$ and history h produces a document distributed according to $\mathcal{C}_h$ (or statistically close to $\mathcal{C}_h$). 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 ℓ be a robustness parameter.

Construction 2. (Universally Robust Steganography)

Procedure SR.Encode:
Input: $K \in \mathcal{K}$, $m \in \{0,1\}$, $h \in D^*$
for $i = 1 \dots \ell$ do
  set $c_i = M(F_K(N, m), h)$
  increment N
  set $h = (h, c_i)$
Output: $c_1, c_2, \dots c_\ell$

Procedure SR.Decode:
Input: $K \in \mathcal{K}$, $c \in D^*$, $h \in D^*$
Let $c = (c_1, c_2, \dots, c_\ell)$
for $i = 1 \dots \ell$ do
  set $h_0 = h_1 = h$
  for $\sigma \in \{0,1\}$ do
    set $m_\sigma = M(F_K(N, \sigma), h_\sigma)$
    set $h_\sigma = (h_\sigma, m_\sigma)$
  increment N
if $(\forall j.(h_{0,j}, c_j) \in R')$
then $\sigma = 0$; else $\sigma = 1$
set $h = h_\sigma$
Output: σ

Suppose that instead of sharing a key to a pseudorandom function F, Alice and Bob shared two secret documents d_0, d_1 drawn independently from $\mathcal{C}_h$. Then Alice could send Bob the message bit σ by sending d_σ, and Bob could recover σ by checking to see if the document he received was related (by R') to d_0 or d_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 d_0, d_1 — would be at most δ. Intuitively, Construction 2 simply extends this notion to multiple bits by replacing the d_0, d_1 by draws from $M(\cdot)$ with shared pseudorandom inputs; and reduces the probability of decoding error to δ^ℓ by encoding each hiddentext bit ℓ times.

4.2 Results and Extensions

The results in this section were reported at CRYPTO 2002 [6]. For my thesis I plan to address several extensions:

Adversarial model. Currently we have only given a construction which is secure against an adversary which makes one-for-one substitutions under R. Thus an adversary could defeat our "robust" stegosystem by breaking one document into two documents of equivalent content, and such an attack might be reasonable for some document sets D. I plan to investigate the plausibility of more general adversarial models in which many-for-many substitutions are allowable; hopefully this will also address other attacks such as partial reordering.

Robust Public Key Steganography. It would be nice if Alice and Bob could robustly communicate (steganographically) even if they have not previously exchanged secrets. Unfortunately, it seems difficult to extend the algorithms from this section to a case where Alice and Bob have no shared secret. Assuming a global PKI for a Diffie-Hellman based public-key cryptosystem in which all parties use the same group G, Alice and Bob could use static Diffie-Hellman key exchange to derive a shared secret and under the Decisional Diffie-Hellman assumption use of this secret would be secure. We can also show how to undetectably "embed" a static Diffie-Hellman PKI into a "factoring-based" PKI, which would allow robust steganography among strangers who use (most of the popular) PKI systems currently available. For my thesis, I propose to formalize these results and investigate other solutions to this problem.

5 Steganographic Bit Rate

The *rate* of a stegosystem is defined by the (expected) ratio of hiddentext size to stegotext size. In general this is a function of the security parameter k, the channel $\mathcal{C}$, and the history h. For a given history h, the rate of the stegosystem described in Section 2 is upper-bounded by $E_{d \leftarrow \mathcal{C}_h}[1/|d|]$. The previously best known upper-bound on *any* stegosystem is 1; finding a tighter upper bound has been an open question in steganography [3]. I propose to give a tight bound for the case of a universal stegosystem, in the presence of a *bit-oracle* for the channel. In this section we will sketch a method to prove an upper bound for a stegosystem running in T steps for any channel. In particular, for messages m where $|m| \leq |\log T|$, we show that $H_\infty(\mathcal{C}_h^{|SE(m,h)|}) \geq |m|$; it is straightforward that if the only access a stegosystem has to $\mathcal{C}_h$ is black-box then the rate cannot improve for longer messages. We will also sketch a stegosystem which achieves rate $\Omega(MAX)$, bounding the optimal rate to within a constant factor.

5.1 Upper bound: $MAX = H_\infty(\mathcal{C}_h^b)/b$

We exhibit a chosen-history, chosen-hiddentext distinguisher for any black-box stegosystem (SE, SD) which encodes $\ell > H_\infty(\mathcal{C}_h^b)$ bits of hiddentext in b documents of stegotext non-negligibly often.

First, note that if SE is an efficient stegosystem (i.e. runs in time polynomial in k) has no information about $\mathcal{C}$ outside a bit oracle for $\mathcal{C}$, then we may restrict our attention to $\ell = O(\log k)$. This is because SE works by accessing $\mathcal{C}$ as a black box, so it cannot generate strings in support of $\mathcal{C}_h^b$ except by sampling, and cannot sample on criteria based on $\ell = \omega(\log n)$ bits in polynomial time.

Suppose we have a specific history h such that SE encodes $\ell + 1$ bits by samples from $\mathcal{C}_h^b$, and $H_\infty(\mathcal{C}_h^b) = \ell$. (If such histories occur non-negligibly often then we can find one by sampling from an oracle for SE; if they do not then the rate of the stegosystem doesn't exceed $H_\infty(\mathcal{C}_h^b)$) Since $H_\infty(\mathcal{C}_h^b) = \ell$, we know there is at least one b-document string, s^*, which has probability $2^{-\ell}$ under $\mathcal{C}_h^b$, and no other string has more probability mass than s^*. Now, if SE were deterministic, then we would have that

$$\Pr[SE(m) = s^* : m \leftarrow U_{\ell+1}] \leq 2^{-(\ell+1)} ,$$

by the unique decoding property. Even if SE is randomized, then for any fixed random bits r, we have

$$\Pr[SE(m;r) = s^* : m \leftarrow U_{\ell+1}] \leq 2^{-(\ell+1)} .$$

But then by an averaging argument, there must be some $m^* \in \{0,1\}^{\ell+1}$ with $\Pr[SE(m^*) = s^*] < 2^{-(\ell+1)}$. In contrast, a covertext oracle CT will have $\Pr[CT(m) = s^*] = 2^{-\ell}$, for any $m \in \{0,1\}^*$. This gap is wide enough to detect with high confidence, given $poly(2^\ell)$ chosen hiddentext samples. And since we have restricted to $\ell = O(\log k)$ this yields a polynomial-time distinguisher between a covertext oracle and a stegotext oracle.

It is straightforward to extend this attack to the case that SE is stateful rather than randomized, and to known-plaintext, known-history attacks.

5.2 Achieving rate $\Omega(MAX)$

Imagine we are given an algorithm L which, given r and h returns the least integer l satisfying $H_\infty(\mathcal{C}_h^l) \geq r$. Armed with this function we can construct a stegosystem which encodes $\log k$ bits of hiddentext in $L(4 \log k, h)$ stegotext documents, with probability $1 - \frac{1}{k}$, as follows. Assume F_K is a pseudorandom function. To encode the message $m \in \{0,1\}^{\log k}$ under history h, set $l = L(4 \log k, h)$, and draw $n = k^{3/2}$ samples $s_1, \ldots, s_n$ from $\mathcal{C}_h^l$. If there are no collisions in $s_1, \ldots, s_n$, send the lowest numbered s_i such that $F_K(s_i) = m$, or if there is none, draw a new sample $s \leftarrow C_h^l$ and send s. Likewise, if there are $i \neq j$ with $s_i = s_j$, choose a new $s \leftarrow C_h^l$ and transmit s. To decode the stegotext $s \in D^l$, compute $F_K(s)$.

Informally, this procedure is secure because, in the case that there are no collisions, the procedure which replaces F_K with a randomly chosen function has distribution identical to a procedure which randomly chooses a new function after each s_i is drawn; in such case the output distribution is identical to $\mathcal{C}_h^l$. In case there are collisions, the encoding procedure's output is again identical to the channel distribution. Thus a distinguisher for the encoding procedure is a distinguisher for F_K.

The procedure fails when either it finds no s_i such that $F_K(s_i) = m$, which happens for a random function with probability $(1 - \frac{1}{k})^{k^{3/2}} \leq e^{-\sqrt{k}} < 1/2k$, or when a collision occurs, which happens with probability at most $\binom{n}{2} 2^{-H_\infty(\mathcal{C}_h^l)} < \frac{n^2}{2k^4} = 1/2k$. Thus the procedure fails with probability at most $1/k$.

Finally, the function $L(r,h)$ can be efficiently estimated for $r = O(\log k)$ by repeated sampling from $\mathcal{C}_h^1, \mathcal{C}_h^2, \ldots, \mathcal{C}_h^{\lceil b/\alpha \rceil r}$. By encoding messages in a Reed-Solomon code over $GF(2^{\log k})$ with error bound $\sqrt{k}$ and using the counter techniques in [6] we can achieve rate $\frac{1-o(1)}{4} MAX$ with negligible error rate.

5.3 Further extensions

The results in this section represent work in progress. I plan to investigate several extensions:

Bound Tightening. The technique in the previous section can be extended to achieve rate $(\frac{1}{2} - \epsilon)MAX$ for any $\epsilon > 0$, but because of the need to avoid collisions, $\frac{1}{2}$ is an upper bound on the competitive ratio acheivable through this technique. I plan to investigate other techniques to further tighten these bounds.

Decoding without sampling. Our standard model assumes that Alice (and Ward) can sample from $\mathcal{C}$, but Bob cannot. Thus requiring Bob to do so is cumbersome. A possible approach is to

somehow encode the block lengths chosen by Alice in the stegotext, at some (hopefully constant factor) expense to the rate.

Improved upper-bound proof. The given upper-bound relies on the fact that an efficient stegosystem must have the same rate for messages of super-logarithmic length as for messages of logarithmic rate. It may be possible to give a more elegant proof of the upper bound using *extractor* [9] bounds. In particular, it is worth noting that the *decoding* algorithm for a stegosystem (SE, SD) acts as an extractor-like function for some distributions; in particular $SD_K(\cdot)$ extracts entropy from the distribution $SE_K(U)$. However, it is not immediately obvious how to extend this to a general extractor and apply the extractor bounds.

Non-universal stegosystems. Also of interest is the question of what rate is achievable given some (nonuniform) information about $\mathcal{C}$. For example, given a perfect compression scheme for sequences from $\mathcal{C}$, we can achieve the Shannon capacity of $\mathcal{C}$. Can we (efficiently) approximate this rate given an approximate compression scheme? What is the "cost of universality?"

6 Timeline

I propose to spend one year investigating the various extensions proposed in this document and completing the write-up of these results. This timeline calls for a defense in July 2004.

References

[1] Luis von Ahn and Nicholas J. Hopper. Public-Key Steganography. Submitted to CRYPTO 2003.

[2] Ross J. Anderson and Fabien A. P. Petitcolas. *On The Limits of Steganography.* IEEE Journal of Selected Areas in Communications, 16(4). May 1998.

[3] Ross J. Anderson and Fabien A. P. Petitcolas. *Stretching the Limits of Steganography.* In: *Proceedings of the first International Information Hiding Workshop.* 1996.

[4] Christian Cachin. *An Information-Theoretic Model for Steganography.* In: *Information Hiding – Second International Workshop, Preproceedings.* April 1998.

[5] S. Craver. *On Public-Key Steganography in the Presence of an Active Warden.* In: *Information Hiding – Second International Workshop, Preproceedings.* April 1998.

[6] Nicholas J. Hopper, John Langford, and Luis von Ahn. *Provably Secure Steganography.* CMU Tech Report CMU-CS-TR-02-149, 2002.

[7] D. Kahn. *The Code Breakers.* Macmillan 1967.

[8] T. Mittelholzer. *An Information-Theoretic Approach to Steganography and Watermarking* In: *Information Hiding – Third International Workshop.* 2000.

[9] Jaikumar Radhakrishnan and Amnon Ta-Shma. *Tight bounds for depth-two superconcentrators.* In: 38th Annual Symposium on Foundations of Computer Science, pages 585–594, 1997.

[10] J. A. O'Sullivan, P. Moulin, and J. M. Ettinger *Information theoretic analysis of Steganography.* In: *Proceedings ISIT '98.* 1998.

[11] G.J. Simmons. *The Prisoner's Problem and the Subliminal Channel.* In: *Proceedings of CRYPTO '83.* 1984.

[12] J. Zollner, H. Federrath, H. Klimant, A. Pfitzmann, R. Piotraschke, A. Westfield, G. Wicke, G. Wolf. *Modeling the security of steganographic systems.* In: *Information Hiding – Second International Workshop, Preproceedings.* April 1998.