Grey-Box Steganography*

Maciej Liśkiewicz, Rüdiger Reischuk, and Ulrich Wölfel

Institut für Theoretische Informatik, Universität zu Lübeck,
Ratzeburger Allee 160, 23538 Lübeck, Germany
{liskiewi,reischuk,woelfel}@tcs.uni-luebeck.de

Abstract. We propose a new model of steganography which combines partial knowledge about the type of covertext channel with machine learning techniques to learn the covertext distribution. Stegotexts are constructed by either modifying covertexts or creating new ones, based on the learned hypothesis. We illustrate our concept with channels that can be described by monomials. A generic construction is given showing that besides the learning complexity, the efficiency of secure grey-box steganography depends on the complexity of membership tests and suitable modification procedures. For the concept class monomials we present an efficient algorithm for changing a covertext into a stegotext.

1 Introduction

The aim of steganography is to hide secret messages in unsuspicious covertexts such that the mere existence of this message is concealed. The basic scenario assumes two communicating parties Alice (sender) and Bob (receiver) plus an adversary Eve, also called a "warden" due to Simmons' [22] scenario of secret communication among prisoners. Eve wants to find out whether Alice and Bob exchange hidden messages among their covertext communication.

A "useful" stegosystem should not only be *secure* (against Eve finding out about the presence of hidden communication), but also *reliable* (i.e. with high probability, encoded messages can be correctly decoded), *computationally efficient* (i.e. the time, space and oracle query complexities should be polynomial in the length of the hidden message) and *rate efficient* (i.e. the transmission rate should be close to the covertext entropy).

In the past few years significant advances have been achieved in developing a theoretical foundation of steganography [4,6,7,12,2,13,14,16,18]. Using notions from cryptography such as *indistinguishability* and adapting them to a steganographic scenario, Hopper et al. have constructed stegosystems that are provably secure against passive and active attacks [12,2]. Their constructions are based on the assumption that Alice and Bob know nothing about the covertext channel. They are only given access to a *black-box* oracle that samples according to the channel distribution. By repeatedly sampling based on a history of previously sampled covertexts these schemes try to find samples that already "contain"

* Supported by DFG research grant RE 675/5-1.

M. Ogihara and J. Tarui (Eds.): TAMC 2011, LNCS 6648, pp. 390–402, 2011.

the message bits to be embedded, hence this method has been named "rejection sampling". While Hopper et al. only embed one bit per covertext document, Le and Kurosawa [16] increase this rate by means of a coding scheme similar to arithmetic coding that they call "$\mathcal{P}$-Codes".

However, all black-box stegosystems suffer from several drawbacks. Lysyanskaya and Meyerovich first pointed out that sampling based on the full history might be too difficult and analysed under which conditions stegosystems that sample with restricted length histories become insecure [18]. Furthermore, Hundt et al. have shown that the construction of such a history-based sampling oracle, a core component of all black-box stegosystems, can lead to an intractable problem for practically relevant covertext channels [14]. Moreover, the scheme in [12] embeds only one bit per document, so each convertext consists of a large number of documents. In order to achieve a reasonable transmission rate, i.e. the average number of hiddentext bits per bit sent, one either has to choose documents of small size or embed more than one bit per document.

Dedić et al. have analysed a generalisation of the scheme in [12] to embed an arbitrary number of bits per document [7] They have shown that for a reliable and secure black-box stegosystem the number of sample documents drawn from the covertext channel grows exponentially in the number of bits embedded per document. Note that this exponential bound also holds for the construction by Le and Kurosawa [16] which uses black-box sampling, too.

In *white-box* steganography, on the other hand, the stegoencoder is assumed to have full knowledge about the covertext channel. The availability of a cumulative distribution function for the covertext channel enables Le and Kurosawa [16] to modify their encoding procedure for black-box sampling and turn it into a white-box stegosystem. Although this makes their construction much more efficient, it seems unlikely that in practice the cumulative distribution is known.

In our study we want to overcome the exponential sampling complexity of the black-box approach without assuming too much knowledge about the covertext channel, as in white-box steganography. The model that we propose here will be called *grey-box* steganography, as the encoder has *partial knowledge* of the covertext channel, making it lie between the black- and white-box scenarios. We will investigate whether efficient and secure grey-box steganography is possible and extract the different properties required for this purpose. Equipped with partial knowledge, the encoder still has to gather more information about the covertext channel to select as stegotexts only those documents that appear in the covertext channel. We will model this situation as an algorithmic learning problem (for an introduction to learning theory see [1]). A priori, Alice knows that the covertext channel belongs to some class of channels, but does not know which covertext documents lie in the support of the actual channel. This is where algorithmic learning comes into play: Alice considers covertext samples and computes a hypothesis that describes the support of the channel. Based on this hypothesis, she actively tries to construct suitable stegotexts that encode her hidden message instead of passively waiting for the sampling oracle to give her a covertext with the desired properties (i.e. using *rejection sampling*).

This construction can be done by modifying an existing covertext or creating a new one. In both cases the distribution of output stegotexts should look like "normal" samples from the oracle. We give a proof of concept with channels that can be described by monomials and concentrate on learning the support of such a channel. To avoid further complications of the learning process due to highly unbalanced distributions, a uniform distribution on the support will be assumed. A generic construction is given showing that apart from the learning complexity, the efficiency of grey-box steganography depends on the complexity of the membership test, and suitable covertext modification procedures. For the concept class monomials we present an efficient algorithm for changing a covertext into a stegotext. Obviously, membership tests for such concepts can be done fast. An additional feature of our construction is that only the sender needs access to the sampling oracle (to learn the concept class), while the receiver only decodes, as in [12,7] and unlike [16], where both sender and receiver require the sampling oracle (black-box) or the cumulative distribution function (white-box).

2 Basic Notation and Definitions

Let Σ be a finite alphabet and $\sigma := \log|\Sigma|$. As usual, Σ^ℓ denotes the set of strings of length ℓ over Σ, and $\Sigma^\star$ the set of strings of finite length over Σ. We denote the length of a string u by $|u|$ and the concatenation of two strings u_1 and u_2 by $u_1||u_2$, or by u_1u_2 if this does not lead to ambiguities.

Symbols $u \in \Sigma$ will be called *documents* and a finite concatenation of documents a *communication sequence* or *covertext*. Typically, the document models a piece of data (e.g. a digital image or fragment of the image) while the communication sequence $c \in \Sigma^\star$ models the complete message sent to the receiver in a single communication exchange.

If $\mathcal{P}$ is a probability distribution with finite support denoted by $\mathrm{supp}(\mathcal{P})$, we define the *min-entropy* of $\mathcal{P}$ as $H_\infty(\mathcal{P}) = \min_{x\in\mathrm{supp}(\mathcal{P})} - \log \Pr_{\mathcal{P}}[x]$. This notion provides a measure of the minimal amount of randomness present in $\mathcal{P}$.

Definition 1 (Channel). *A channel $\mathcal{C}$ is a function that takes a history $\mathcal{H} \in \Sigma^\star$ as input and produces a probability distribution $\mathcal{C}_\mathcal{H}$ on Σ. A history $\mathcal{H} = c_1c_2\ldots c_m$ is legal if each subsequent symbol is obtainable given the previous ones, i.e., $\Pr_{\mathcal{C}_{c_1c_2\ldots c_{i-1}}}[c_i] > 0$ for all $i \le m$. The min-entropy of $\mathcal{C}$ is the value $\min_\mathcal{H} H_\infty(\mathcal{C}_\mathcal{H})$ where the minimum is taken over all legal histories $\mathcal{H}$.*

This gives a very general definition of covertext distributions which allows dependencies between individual documents that are present in typical real-world communications. In order to embed additional information into covertexts, one has to assume that the covertext channel distribution has a sufficiently large min-entropy.

To get information about the covertext distribution *sampling oracles* can be used. $EX_\mathcal{C}(\mathcal{H})$ denotes an oracle that generates documents according to a channel $\mathcal{C}$ with history $\mathcal{H}$, i.e. each call of $EX_\mathcal{C}(\mathcal{H})$ returns a document c with probability $\Pr_{\mathcal{C}_\mathcal{H}}[c]$ and the responses are independent of each other.

A steganographic information transmission is thought of as taking a finite sequence $C_1, C_2, \ldots \in \Sigma^\star$ of covertexts and based on them to construct a stegotext $S \in \Sigma^\star$ such that the sequence additionally encodes an independent message M. This encoding is done by Alice who then sends the stegotext to the receiver Bob over a public channel. Let b denote the message encoding rate, i.e. a single stegodocument can encode up to b bits of M. Longer messages M have to be split into blocks of b bits each and for each block a separate stegodocument is generated. Their concatenation yields the stegotext.

Definition 2 (Stegosystem). *In the following, let $n = \ell \cdot b$ denote the length of the messages to be embedded, thus ℓ stegodocuments each hiding b bits are needed. A stegosystem $\mathcal{S}$ for the message space $\{0,1\}^n$ is a triple of probabilistic algorithms $[SK, SE, SD]$ with the following functionality:*

- *SK is the key generation procedure that on input 1^n outputs a key K of length κ, where κ is a security parameter that depends on n;*
- *SE is the encoding algorithm that takes as input a key $K \in \{0,1\}^\kappa$, a message $M \in \{0,1\}^n$ (called hiddentext), a channel history $\mathcal{H}$, and accesses the sampling oracle $EX_{\mathcal{C}}()$ of a given covertext channel $\mathcal{C}$ and returns a stegotext $S \in \Sigma^\ell$;*
- *SD is the decoding algorithm that takes K, S, and $\mathcal{H}$, and having access to the sampling oracle $EX_{\mathcal{C}}()$ returns a message M'.*

$\mathcal{S}$ is called a black-box stegosystem *if SE and SD have no a priori knowledge about the distribution of the covertext channel and can obtain information about it only by querying the sampling oracle.*

The application of *SK* is shared by Alice and Bob beforehand and its result is kept secret from an adversary. All further actions of Alice are specified by *SE*, those of Bob by *SD*. For all stegosystems discussed in this paper *SK* generates keys with a uniform distribution, thus when specifying a stegosystem we skip the description of *SK*.

The time complexities of the algorithms *SK*, *SE*, *SD* are measured with respect to n, κ, and the document size (specified formally by $\sigma = \log|\Sigma|$), where an oracle query is charged as one unit step. A stegosystem is *computationally efficient* if its time complexities are polynomially bounded. By convention, the running time of an algorithm includes the so called *description size* of that algorithm with respect to some standard encoding.

Ideally, one would expect that the encoder always succeeds in encoding the original message M and that the decoder always succeeds in extracting M from the stegotext. Since this may not always be possible, we define the unreliability of a stegosystem.

Definition 3 (Unreliability). *The unreliability $\mathtt{UnRel}_{\mathcal{C},\mathcal{S}}$ of $\mathcal{S}$ with respect to $\mathcal{C}$ is given by $\max_{M\in\{0,1\}^n,\mathcal{H}} \Pr_{K\leftarrow SK(1^n)}[SD(K, SE(K, M, \mathcal{H}), \mathcal{H}) \neq M]$.*

Next, let us measure the security of a stegosystem. How likely is it that an adversary, the warden W, can discover that the covertext channel is used for transmitting additional information? If we put no algorithmic restrictions on W

(i.e. information-theoretic security) it is necessary that (1) the stegotext S lies in the support of the covertext channel, otherwise W could test S for membership in $\mathrm{supp}(\mathcal{C})$, and (2) the probability of producing a stegotext S equals the probability of drawing S according to $\mathcal{C}$. Cachin has proposed the following information-theoretic model of steganographic security [6].

Definition 4 (Information-theoretic Security). *Let $\mathcal{C}$ be a covertext channel with distribution $P_{\mathcal{C}}$ and let $P_{\mathcal{S},\mathcal{C}}$ be the output distribution of the steganographic embedding function SE having access to the sampling oracle $EX_{\mathcal{C}}()$. The stegosystem $[SK, SE, SD]$ is called perfectly secure for the channel $\mathcal{C}$ (against passive adversaries) if the relative entropy satisfies $D(P_{\mathcal{C}}||P_{\mathcal{S},\mathcal{C}}) = 0$.*

To simplify the analysis, for the systems given later we will assume that the distribution on the support is uniform. Thus, we concentrate on the problem how the encoder can learn the support of the channel and then uniformly generate stegotexts. The constructions given below can be extended to a wider class of distributions using statistical learning techniques [15].

For a security analysis in the complexity-theoretic sense, W is assumed to be polynomially time-bounded. Thus, Alice has to make sure that an adversary cannot detect deviations from the two conditions above in polynomial time. However, the adversary may actively perform a *chosen hiddentext attack* [12,7]. Let $SE(K, M, \mathcal{H})$ with access to $EX_{\mathcal{C}}(\mathcal{H})$ be denoted by $SE^{\mathcal{C}}(K, M, \mathcal{H})$. In contrast, we define an oracle OC that for given message $M \in \{0,1\}^n$ and channel history $\mathcal{H}$ returns a truly random covertext $c_1 c_2 \ldots c_\ell$ of length $\ell = |SE^{\mathcal{C}}(K, M, \mathcal{H})|$ from the covertext channel $\mathcal{C}$ with the history $\mathcal{H}$, i.e. each c_i is drawn according to the probability distribution $\mathcal{C}_{\mathcal{H}||c_1 c_2 \ldots c_{i-1}}$.

Definition 5 (Warden, Chosen Hiddentext Attack (CHA)). *A probabilistic algorithm W is a (t, q, λ)-warden for the stegosystem $\mathcal{S} = [SK, SE, SD]$ if*

- *W runs in time t and accesses a reference oracle $EX_{\mathcal{C}}()$ that he can query for samples from the covertext channel $\mathcal{C}$ with a history $\mathcal{H}$;*
- *W can make a number of q queries of total length λ bits on a challenge oracle CH which is either $SE^{\mathcal{C}}(K, M, \mathcal{H})$ or $OC(M, \mathcal{H})$, where M and $\mathcal{H}$ can be chosen by W;*
- *the task of W is to determine the use of the stegosystem $\mathcal{S}$ with the help of the challenge oracle: $W^{\mathcal{C},CH} = 1$ means that W decides on "stegotext", resp. $W^{\mathcal{C},CH} = 0$ means that W decides on "covertext".*

We define the *advantage* of W over random guessing for a given covertext channel $\mathcal{C}$ as

$$\mathtt{Adv}^{\mathsf{cha}}_{\mathcal{C},\mathcal{S}}(W) \;:=\; \left|\Pr_{K \leftarrow SK(1^n)}[W^{\mathcal{C},SE^{\mathcal{C}}(K,\cdot,\cdot)} = 1] - \Pr[W^{\mathcal{C},OC(\cdot,\cdot)} = 1]\right| \;.$$

Note that in order to maximize the advantage, W may depend on the channel $\mathcal{C}$. In the most favourable case, W may possess a complete specification of $\mathcal{C}$, so that he even does not need to query the reference oracle. The amount of such information about $\mathcal{C}$ is part of the description size of W. This knowledge may put the adversary in a much better situation than the encoder.

Definition 6 (Steganographic Security against CHA). *The* insecurity *of a stegosystem $\mathcal{S}$ with respect to a covertext channel $\mathcal{C}$ and complexity bounds t, q, λ is defined by*

$$\mathtt{InSec}^{\mathsf{cha}}_{\mathcal{C},\mathcal{S}}(t,q,\lambda) \;:=\; \max_{W}\{\mathtt{Adv}^{\mathsf{cha}}_{\mathcal{C},\mathcal{S}}(W)\}\ ,$$

where the maximum is taken over all adversaries W working in time at most t and making at most q queries of total length λ bits to the challenge oracle CH.

Note that we do not explicitly mention the description size of the adversary, but assume this to be included in the running time t (W has to read this information at least once).

Below we recall some notions from cryptography required for the specification of the encoding function *SE*. Let $F : \{0,1\}^k \times \{0,1\}^l \to \{0,1\}^L$ be a function. Here $\{0,1\}^k$ is considered as the key space of F. For each key $K \in \{0,1\}^k$ we define the subfunction $F_K : \{0,1\}^l \to \{0,1\}^L$ by $F_K(x) = F(K,x)$. Thus, F specifies a family of functions, and is called a family of permutations if $l = L$ and for each key K the subfunction F_K is a permutation on $\{0,1\}^l$. For such an F we define the advantage of a probabilistic distinguisher D having access to a challenging oracle as

$$\mathtt{PRP\text{-}Adv}_F(D) \;=\; \left|\Pr_{K\in_R\{0,1\}^k}[D^{F_K(\cdot)}=1] - \Pr_{P\in_R PERM(l)}[D^{P(\cdot)}=1]\right|,$$

where $PERM(l)$ denotes the family of all permutations on $\{0,1\}^l$. The insecurity of a pseudorandom family of permutations F is given by $\mathtt{PRP\text{-}InSec}_F(t,q) = \max_D\{\mathtt{PRP\text{-}Adv}_F(D)\}$, where the maximum is taken over all probabilistic distinguishers D running in at most t steps and making at most q oracle queries. F is called a (t,q,ϵ)-pseudorandom family if $\mathtt{PRP\text{-}InSec}_F(t,q) \le \epsilon$. Let the length l grow polynomially with respect to k. A sequence $\{F_k\}_{k\in\mathbb{N}}$ of families $F_k : \{0,1\}^k \times \{0,1\}^l \to \{0,1\}^l$ is called pseudorandom if for all polynomially bounded distinguishers D, $\mathtt{PRP\text{-}Adv}_F(D)$ is negligible in k (for more formal definition of pseudorandom permutations see e.g. [5]).

3 A Grey-Box Model for Steganography

Previous steganographic models have considered computationally restricted adversaries W that possess *full knowledge* of the covertext channel. Dedić et al. [7] consider this "a meaningful strengthening of the adversary". We think that such strengthening is not appropriate to model Alice and Eve's basic knowledge about a covertext channel. In practice, encoders and wardens get an idea about typical covertexts by observing samples. They do not and likely will never possess any short advice that fully describes the channels they are looking at. Furthermore, there may be different families of channels (e.g. images, texts, audio-signals) and Alice may preselect one specific family from which the actual channel is then drawn without any outside influence. This more realistic setting strengthens the encoder and may be a chance to overcome the negative results for the black-box

scenario. We do not know any steganographic system used in practice that is based on rejection-sampling, instead stegotexts typically are derived by slight modifications of given covertexts.

In the grey-box model Alice has some *partial knowledge* about the covertext channel. Therefore, we use the notion of concept classes from machine learning and define a *channel family* $\mathcal{F}$ as a set of covertext channels that share some common characteristics, such as e.g. all pseudo-random sequences, digital photographs from a certain camera, or all English literary texts. In the context of pseudo-random sequences, a single channel $\mathcal{C}_i$ contains strings output by a specific pseudo-random number generator with a fixed seed and the channel family $F_{PRS} = \{\mathcal{C}_1, \mathcal{C}_2, \ldots\}$ contains channels with different seeds.

Note that both the encoder and the warden know the concept class, the family of channels. For the actual channel $\mathcal{C}$, one member is selected at random, which is unknown to the encoder. Depending on the modelled strength of the warden, W may also lack knowledge about $\mathcal{C}$ or he may have additional information about $\mathcal{C}$. Here, we do not investigate this question further and allow the adversary to have full knowledge. The decoder, on the other hand, is not involved in the learning process, he does not need any information about the concept class.

As before, the encoding SE may access the sampling oracle $EX_{\mathcal{C}}()$, but now we clearly differentiate between accesses to the oracle for learning purposes to construct a hypothesis for the covertext channel, and accesses to get a covertext that – using the hypothesis – can be modified into a stegotext.

Depending on the concept class, Alice may be able to derive a good hypothesis – an exact or very close description of the channel – or not. Even if the concept class is not known to be efficiently learnable it makes sense to consider a situation where a precise description of the channel is given to Alice for free. Still, even in this favourable case it is not clear how Alice can construct stegotexts. She must be able to efficiently modify covertexts and test the modifications for membership in the support of the channel.

Definition 7. *The* insecurity *and* unreliability *of a stegosystem $\mathcal{S}$ with respect to the channel family $\mathcal{F}$ are defined by*

$$\mathtt{InSec}^{\mathsf{cha}}_{\mathcal{F},\mathcal{S}}(t,q,\lambda) \;:=\; \max_{\mathcal{C}\in\mathcal{F}} \mathtt{InSec}^{\mathsf{cha}}_{\mathcal{C},\mathcal{S}}(t,q,\lambda) \;\; \textit{and} \;\; \mathtt{UnRel}_{\mathcal{F},\mathcal{S}} \;:=\; \max_{\mathcal{C}\in\mathcal{F}} \mathtt{UnRel}_{\mathcal{C},\mathcal{S}}\,.$$

We think this definition, which specifies the insecurity of stegosystems with respect to *families* of channels instead of *all* channels, corresponds better to real life intuition of insecurity than the commonly used definition. In fact, in real life steganalysis, our grey-box steganography model is already implicitly used to analyse the insecurity of particular stegosystems with respect to specific channel families. For example, it is easy to see that the steganographic algorithm F5 for JPEG images [23] is insecure with respect to the common insecurity definition, because $\mathtt{InSec}^{\mathsf{cha}}_{\mathcal{C},F5}$ is huge for almost all channels $\mathcal{C}$ deviating significantly from images compressed by JPEG. But this observation seems to be useless for a stegoanalyst, for whom a much more appropriate approach to analyse the insecurity of F5 would be to use our definition and restrict the channels to the family of JPEG-compressed images, like it was done e.g. in [10].

4 The Monomial Covertext Channels

In the rest of the paper we will present an example of a stegosystem showing that the issues discussed above are relevant and the grey-box model makes sense. In our study we consider a family of channels that can be described by monomials.

Consider a concept class over the document space $\Sigma = \{0,1\}^\sigma$ consisting of channels $\mathcal{C}$ where for each history $\mathcal{H}$, $\mathcal{C}_\mathcal{H}$ is a uniformly distributed subset of Σ that can be defined by a monomial. We denote such a channel family by `MONOM`.

A monomial over $\{0,1\}^\sigma$ will be represented by a vector $\mathbf{H} = (\mathbf{h}_1, \ldots, \mathbf{h}_\sigma) \in \{0,1,\times\}^\sigma$ and defines the subset of all 0-1-vectors, for which the i-th component is 0 if $\mathbf{h}_i = 0$, and 1 if $\mathbf{h}_i = 1$. The other components are called free variables. So, e.g. the monomial represented as "0×0×1" describes the set of strings $\{00001, 00011, 01001, 01011\}$. We denote the subset defined by a monomial $\mathbf{H}$ by $\boldsymbol{H}$.

One of the novel ingredients of our grey-box-stegosystems is a procedure called **Monomial-modify**, which for a given monomial $\mathbf{H}$ and a cover-document $c \in \boldsymbol{H}$, modifies c to a stego-document $s \in \boldsymbol{H}$ that encodes a b-bit message M in a way that preserves the uniform probability distribution over $\boldsymbol{H}$ to guarantee indistinguishability of stegotexts. This nontrivial task is described below.

Let, for short, $\sigma_b := \lfloor \sigma/b \rfloor$ and define for a permutation π of $\{1,2,\ldots,\sigma\}$ and $1 \le j \le b$ the subset $I_\pi(j)$ as follows: $I_\pi(j) := \{\pi(\sigma_b \cdot (j-1)+1), \pi(\sigma_b \cdot (j-1)+2), \ldots, \pi(\sigma_b \cdot j)\}$. These subsets partition a document $c = a_1 \ldots a_\sigma$ into b subsequences of length σ_b, where the j-th set contains all elements a_i with index i in $I_\pi(j)$. Let $FV_\pi(j)$ denote those indices in $I_\pi(j)$ that belong to free variables. Each subsequence embeds one bit of the message M as the parity of all its elements. If the parity does not match we want to flip at least one of these bits. If a free variable is chosen for this purpose it is guaranteed that the modified string still belongs to $\boldsymbol{H}$.

Procedure Monomial-modify(M, c, $\mathbf{H}$, K)
Input: hiddentext $M = m_1, m_2, \ldots, m_b \in \{0,1\}^b$; covertext document
$c = a_1 a_2 \ldots a_\sigma \in \{0,1\}^\sigma$; hypothesis monomial
$\mathbf{H} = \mathbf{h}_1 \mathbf{h}_2 \ldots \mathbf{h}_\sigma \in \{0,1,\times\}^\sigma$; private key K;
let π be the permutation specified by key K;
for $j := 1, \ldots, b$ **do**
if $[m_j \neq \bigoplus_{i \in I_\pi(j)} a_i$ **and** $FV_\pi(j) \neq \emptyset]$ **then** $a_{\nu_j} = 1 - a_{\nu_j}$, where $\nu_j := \min FV_\pi(j)$
end
Output: $s = a_1 a_2 \ldots a_\sigma$

The following procedure is used to decode a stegotext document.

Procedure Document-decode(s, K)
Input: stegotext document $s = a_1 a_2 \ldots a_\sigma \in \{0,1\}^\sigma$; private key K;
let π be the permutation specified by key K;
For $j := 1, \ldots, b$ **do** $m_j := \bigoplus_{i \in I_\pi(j)} a_i$;
Output: $m_1 m_2 \ldots m_b$

The crucial property of the procedure **Monomial-modify** says that if $\mathbf{H}$ and $\mathbf{C}$ are monomials (corresponding to a hypothesis, respectively to a concept) such that $\boldsymbol{H} \subseteq \boldsymbol{C}$ and if c is chosen randomly in $\boldsymbol{C}$, then **Monomial-modify** preserves the uniform probability distribution over $\boldsymbol{C}$. This is described formally by the following claim.

Lemma 1. *Let* $\mathbf{H}$ *and* $\mathbf{C}$ *be given monomials such that* $\boldsymbol{H} \subseteq \boldsymbol{C}$ *and let* K *be an arbitrary private key. Then for every* $s \in \boldsymbol{C}$ *it holds*

$$\Pr[\textbf{Monomial-modify}(M, c, \mathbf{H}, K) = s] \;=\; 1/|\boldsymbol{C}| \;,$$

where the probability is taken over random choices of $c \in \boldsymbol{C}$ *and* $M \in \{0,1\}^b$. *Moreover, for every* M, *every* $\mathbf{H}$ *with* φ *free variables, and* $c \in \boldsymbol{C}$ *the probability* $\Pr[\textbf{Document-decode}(\textbf{Monomial-modify}(M, c, \mathbf{H}, K), K) \neq M] \leq b \cdot e^{-\varphi/b+1}$ *over random choices of* K. *The time complexity of both procedures is linear in* σ.

5 Stegosystem for Monomial Channels

Now, we are ready to construct a stegosystem for monomial channels. The system, denoted by $\mathcal{S}$, is based on the following encoding and decoding procedures using families of permutations $F : \{0,1\}^k \times \{0,1\}^n \to \{0,1\}^n$. For a stegosystem $\mathcal{S}$ that is perfectly secure in the information-theoretic setting we choose $k = n$ and the function $F_K(x) = x \oplus K$. In the complexity-theoretic case a family F of efficiently computable pseudorandom permutations is used in order to prevent chosen hiddentext attacks. The following procedure is used by Alice to encode the message M.

Procedure Encode(M, K)
Input: hiddentext $M = m_1 \dots m_n \in \{0,1\}^n$; private key $K = K_0, \dots, K_{2\ell}$;
let $\mathcal{H}$ be a current history;
choose $T_0 \in_R \{0,1\}^n$ and let $T_1 := F_{K_0}(T_0 \oplus M)$;
parse T_0T_1 into $t_1t_2 \dots t_{2\ell}$, where $|t_i| = b$;
for $i := 1, \dots, 2\ell$ **do**
  $c_i := EX_{\mathcal{C}}(\mathcal{H})$;
  access $EX_{\mathcal{C}}(\mathcal{H})$ to learn a hypothesis $\mathbf{H}_i$ for $\mathcal{C}_{\mathcal{H}}$;
  $s_i :=$ **Monomial-modify**(t_i, c_i, $\mathbf{H}_i$, K_i);
  let $\mathcal{H} := \mathcal{H}||s_i$;
end
Output: $s_1 \dots s_{2\ell}$

The procedure below is used by Bob to decode a stegotext s.

Procedure Decode(s, K)
Input: stegotext $s = s_1 \dots s_{2\ell} \in \{0,1\}^{2n}$; private key $K = K_0, \dots, K_{2\ell}$;
for $i := 1, \dots, 2\ell$ **do**
  $t_i :=$ **Document-decode**(s_i, K_i);
end
$M := F^{-1}_{K_0}(t_{\ell+1} \dots t_{2\ell}) \oplus t_1 \dots t_\ell$;
Output: $M = m_1 \dots m_\ell$

Using the definition of perfect security according to Definition 4 and the security against chosen hiddentext attack given in Definition 6 we can now apply the new framework and state the following practical result.

Theorem 1. *Let the min-entropy of every channel $\mathcal{C}$ in* MONOM *be at least h. Let b denote the rate of the stegoencoding and n the length of the secret message to be embedded. Assume Alice has no a priori knowledge of $\mathcal{C}$, but both Alice and the warden have access to a sampling oracle $EX_{\mathcal{C}}()$. Then the stegosystem $\mathcal{S}$ is computationally efficient and achieves the following reliability and security:*

$$\mathtt{UnRel}_{\mathtt{MONOM},\mathcal{S}} \;\leq\; 2n \cdot e^{-h/b+1} + 2^{-n} \qquad \textit{and}$$

– *with encrypting function $F_K(x) = x \oplus K$ perfect security, and*
– *with a family F of pseudorandom permutations*

$$\mathtt{InSec}^{\mathsf{cha}}_{\mathtt{MONOM},\mathcal{S}}(t,q,\lambda) \;\leq\; 2 \cdot \mathtt{PRP\text{-}InSec}_F(p(t),\lambda/n) + \xi(\lambda,n)$$

where p is a fixed polynomial and the function $\xi(\lambda,n) := \left(\frac{\lambda^2}{n^2} - \frac{\lambda}{n}\right) 2^{-n}$ is a function related to the insecurity of the family F of pseudorandom permutations used in $\mathcal{S}$.

Note that this stegosystem is secure in both cases even if the adversary has complete knowledge of the channel.

A parity-based approach to steganography has previously been suggested by Anderson and Petitcolas [3]. They argue that the more bits are used for calculating the parity, the less likely one can distinguish the stegotext from an unmodified covertext. In our case, Alice produces stegotexts that are always consistent with her hypothesis and thus cannot be distinguished from covertexts by construction (modulo the error Alice makes when learning). Alice could also use a pseudo-random function f_K with key K instead of the parity, in which case she would eventually have to try changing different free variables before obtaining the desired value to be embedded, thus increasing the time complexity of her embedding algorithm.

Monomial concept classes may look too simple to describe covertexts in practice. However, in this setting we do not have to restrict the variables, in learning theory also called attributes, to properties of the physical medium. If one can efficiently implement a modification of a simple attribute, these attributes may also represent semantic properties of a document. For example, pictures may be classified according to their content – whether they were taken in summer or winter, contain objects like lakes, mountains, etc. Thus, in a simple way one can construct a secure system that may be called *semantic steganography*.

Recall the properties that were needed to achieve efficient and secure steganography for the concept class of monomials: 1.) monomials are efficiently learnable from positive examples, 2.) for each monomial $\mathbf{H}$ with enough entropy there is an efficient embedding function for the hiddentext on the support of $\mathbf{H}$, and 3.) one can efficiently compute a uniformly selected stegotext (in this case the procedure **Monomial-modify**). This generic construction can be applied to other concept classes fulfilling these properties.

For the class of monomials one actually does not need the modification procedure Monomial-modify to generate a stegotext from a given covertext. In this case, the hypothesis space even allows a direct generation of stegotexts by selecting for all, but one free variable in each group values at random.

6 Conclusions and Future Work

This paper introduces a new approach to modeling and analysing steganography. Previous models (e.g. [12], [7] or [16] either treat the covertext channel as a black-box – resulting in a sampling complexity exponential in the number of bits per covertext document – or assume a priori full knowledge about the covertext distribution, which seems unrealistic. We overcome this situation by allowing the encoder to *modify* covertexts, as done in most practical stegosystems. Our grey-box model is more realistic in the sense that we assume the encoder to have some partial knowledge about the channel.

Furthermore, a finer-grained distinction between the different ingredients for securely hiding information into covertexts provides new insights and helps in constructing stegosystems. We have shown that for efficiently learnable covertexts secure and efficient steganography is possible by presenting a construction for monomial channels, which are efficiently PAC-learnable. So far, our construction is restricted to monomial channels with the uniform distribution. For general distributions, note that the actual distribution on the support of the channels has to be learned in addition to the support in order to achieve information theoretic security. For arbitrary distributions this cannot be done efficiently. However, in the complexity theoretic setting we think that our construction can at least be generalized to the case where each free variables x_i independently of the others takes the value 1 with some arbitrary probability p_i, so called product distributions.

Even for channels that are hard to learn in the PAC-sense, assuming that by some other means the encoder can get hypotheses about the channel, one can design efficient stegosystems if the modification problem has an efficient solution.

Steganographic techniques like LSB-flipping for digital images can easily be expressed by this approach. They can be viewed as variants of **Monomial-modify**, with all but the last bits of each pixel being fixed and the least significant bit being a free variable. The support of the covertext channel for a given image I thus consists of all images that only differ in their least significant bits. However, digital images taken by modern cameras do not tend to generate truly random LSBs. Thus, representing the hypothesis as a monomial may be inappropriate for camera channels and the monomial stegosystem insecure. An important future task will be the implementation of grey-box steganography with practically relevant covertext channels.

In the grey-box setting there may still be a huge advantage for the adversary if he has complete knowledge of the covertext channel. As a next step one should investigate more carefully the case that the knowledge of the adversary is limited similar to the situation of the stegoencoder.

References

1. Angluin, D.: Computational Learning Theory: Survey and Selected Bibliography. In: Proc. STOC 1992, pp. 351–369. ACM, New York (1992)
2. von Ahn, L., Hopper, N.J.: Public-key steganography. In: Cachin, C., Camenisch, J.L. (eds.) EUROCRYPT 2004. LNCS, vol. 3027, pp. 323–341. Springer, Heidelberg (2004)
3. Anderson, R.J., Petitcolas, F.A.P.: On the limits of steganography. IEEE Journal of Selected Areas in Communications 16(4), 474–481 (1998)
4. Backes, M., Cachin, C.: Public-Key Steganography with Active Attacks. In: Kilian, J. (ed.) TCC 2005. LNCS, vol. 3378, pp. 210–226. Springer, Heidelberg (2005)
5. Bellare, M., Desai, A., Jokipii, E., Rogaway, F.: A Concrete Security Treatment of Symmetric Encryption. In: Proc. FOCS 1997, pp. 394–403. IEEE, Los Alamitos (1997), full paper available under `http://www-cse.ucsd.edu/~adesai/papers/pubs.html#BDJR97`
6. Cachin, C.: An information-theoretic model for steganography. Information and Computation 192(1), 41–56 (2004)
7. Dedić, N., Itkis, G., Reyzin, L., Russell, S.: Upper and lower bounds on black-box steganography. Journal of Cryptology 22(3), 365–394 (2009)
8. Denis, F.: PAC Learning from Positive Statistical Queries. In: Richter, M.M., Smith, C.H., Wiehagen, R., Zeugmann, T. (eds.) ALT 1998. LNCS (LNAI), vol. 1501, pp. 112–126. Springer, Heidelberg (1998)
9. Ehrenfeucht, A., Haussler, D.: Learning decision trees from random examples. Information and Computation 82(3), 231–246 (1989)
10. Fridrich, J.J., Goljan, M., Hogea, D.: Steganalysis of JPEG Images: Breaking the F5 Algorithm. In: Petitcolas, F.A.P. (ed.) IH 2002. LNCS, vol. 2578, pp. 310–323. Springer, Heidelberg (2003)
11. Haussler, D.: Bias, version spaces and Valiant's learning framework. In: Proc. of the 4th International Workshop on Machine Learning, pp. 324–336. University of California, Irvine (1987)
12. Hopper, N.J., Langford, J., von Ahn, L.: Provably secure steganography. In: Yung, M. (ed.) CRYPTO 2002. LNCS, vol. 2442, pp. 77–92. Springer, Heidelberg (2002)
13. Hopper, N.J.: On Steganographic Chosen Covertext Security. In: Caires, L., Italiano, G.F., Monteiro, L., Palamidessi, C., Yung, M. (eds.) ICALP 2005. LNCS, vol. 3580, pp. 311–323. Springer, Heidelberg (2005)
14. Hundt, C., Liśkiewicz, M., Wölfel, U.: Provably secure steganography and the complexity of sampling. In: Asano, T. (ed.) ISAAC 2006. LNCS, vol. 4288, pp. 754–763. Springer, Heidelberg (2006)
15. Kearns, M.: Efficient Noise-Tolerant Learning from Statistical Queries. In: Proc. STOC 1993, pp. 392–401. ACM, New York (1993)
16. Le, T.V., Kurosawa, K.: Bandwidth optimal steganography secure against adaptive chosen stegotext attacks. In: Camenisch, J.L., Collberg, C.S., Johnson, N.F., Sallee, P. (eds.) IH 2006. LNCS, vol. 4437, pp. 297–313. Springer, Heidelberg (2007)
17. Letouzey, F., Denis, F., Gilleron, R.: Learning From Positive and Unlabeled Examples. In: Arimura, H., Sharma, A.K., Jain, S. (eds.) ALT 2000. LNCS (LNAI), vol. 1968, pp. 71–85. Springer, Heidelberg (2000)
18. Lysyanskaya, A., Meyerovich, M.: Provably secure steganography with imperfect sampling. In: Yung, M., Dodis, Y., Kiayias, A., Malkin, T. (eds.) PKC 2006. LNCS, vol. 3958, pp. 123–139. Springer, Heidelberg (2006)

19. Quinlan, J.R.: Induction of decision trees. Machine Learning 1(1), 81–106 (1986)
20. Quinlan, J.R.: C4.5: Programs for Machine Learning. Morgan Kaufmann, San Mateo (1993)
21. Rogaway, P.: Nonce-based symmetric encryption. In: Roy, B., Meier, W. (eds.) FSE 2004. LNCS, vol. 3017, pp. 348–359. Springer, Heidelberg (2004)
22. Simmons, G.J.: The prisoners' problem and the subliminal channel. In: Crypto 1983, pp. 51–67. Plenum Press, New York (1984)
23. Westfeld, A.: F5-A Steganographic Algorithm. In: Moskowitz, I.S. (ed.) IH 2001. LNCS, vol. 2137, pp. 289–302. Springer, Heidelberg (2001)