Defining Security in Steganographic Systems

Stefan Katzenbeisser[a] and Fabien A.P. Petitcolas[b]

[a]Institute for Information Systems
Database and Artificial Intelligence Group
Vienna University of Technology
Favoritenstrasse 9-11/184-2, 1040 Wien, Austria

[b]Microsoft Research,
7 J. J. Thomson Avenue, Cambridge CB3 0FB
Cambridge (UK)

ABSTRACT

Intuitively, the security of a steganographic communication between two principals lies in the inability of an eavesdropper to distinguish cover-objects from stego-objects, that is objects which contain secret messages. A system should be already considered insecure, if an eavesdropper can suspect the presence of secret communication. Several definitions of steganographic security were proposed in the literature. However, they all consider only "perfectly secure" steganographic systems, where even a computationally unbounded observer cannot detect the presence of a secret message exchange. Second, it might be difficult to construct secure schemes usable in practice following these definitions. Third, they all require the knowledge of the probability distribution of "normal" covers; although it might be possible in certain cases to compute this probability, it will in general be infeasible to obtain.

In this paper, we propose a novel approach for defining security in steganographic systems. This definition relies on a probabilistic game between the attacker and a judge. Given the ability to observe the normal communication process and the steganographic system, the attacker has to decide whether a specific object (given to him by a judge) is in fact a plain cover or a stego-object. We discuss the applicability of this new definition and pose the open problem of constructing provably secure steganographic systems.

Keywords: steganography, security definition

1. INTRODUCTION

Simmons[6] introduced a classic scenario for invisible communication, the *prisoners problem.* Suppose two fictional characters named Alice and Bob are arrested for some crime and put in two different cells. In order to develop an escape plan, they have to communicate with each other. Unfortunately, all communication is arbitrated by a warden, named Wendy. If she notices any suspicious communication, she will suppress the exchange of messages at all. *Steganographic systems* allow to hide secret messages in un-suspicious objects, called *covers.* The aim is to exchange the secret message without raising suspicion of the warden.

In this paper, we consider only secret-key steganographic systems, i.e. systems in which both communication partners share one single (symmetric) stego-key, which will be used both in the embedding and extraction processes. The steganographic communication can be outlined as follows (see Figure 1). Alice chooses randomly a cover c and hides her secret message m in the cover by using the secret key k. The result of this operation is a stego-object s that is transmitted to Bob. He uses again the secret key k to extract the message m out of s.

Intuitively, the security of the system depends on the inability of a warden to distinguish covers (containing no valid secret information) from stego-objects. A system should already be called insecure if a warden can suspect the presence of secret communication. Thus, an eavesdropper is faced to solve what one might call *steganographic decision problem*: given any cover or stego-object, he must be able to guess (better than random) whether a secret message is actually contained in the object or not. For this purpose, he can compare his object with "common"

e-mail: `skatzenbeisser@acm.org`,`fabienpe@microsoft.com`. The work of the first author was supported by the Austrian Science Fund (FWF) under Program Nr. Z29-INF.

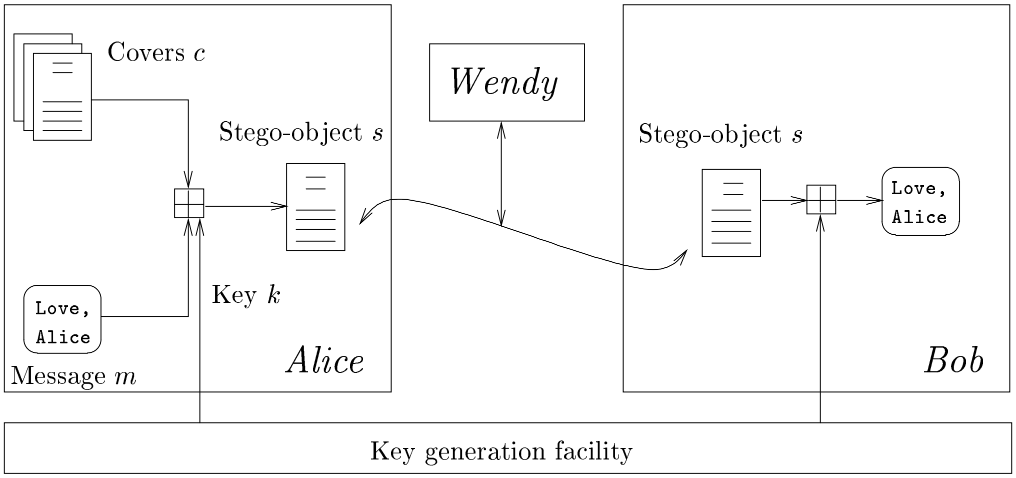

Figure 1. Schematic description of a steganographic channel: Alice randomly chooses a cover c and embeds the message m in c using a key k, creating the stego-object s which she passes on to Bob. Bob reconstructs m with the key k he shares with Alice.

objects Alice and Bob usually transmit during their communication. Based on a "history" of recently transmitted objects, an eavesdropper can evaluate and improve his decision strategy.

This paper surveys possible security definitions for steganographic systems. Instead of previous works, which suggested information-theoretic definitions, we propose to use a model that takes into account the limited computational power of the warden. Our definition of security is similar to security definitions in cryptography (so-called indistinguishability tests), which do not require a precise model of the communication channel; however, if such information is available, it can be incorporated in the decision process.

2. INFORMATION-THEORETIC SECURITY DEFINITIONS

Previous models for steganographic channels were mostly information-theoretic approaches. For example, Zöllner[7] proposed an information theoretic definition of steganographic security, in which the sets of messages M, covers C and keys K are seen as random variables. The output of the embedding process is again a random variable X. A steganographic system is secure in their model if the mutual information $I(M; X \wedge C)$ equals zero, i.e. if M is independent from X and C. In other words, knowledge of both random variables X and C does not reveal any information about M. As this definition is very similar to Shannon's definition of unconditional security of cryptographic systems, we may call any method satisfying this definition "unconditionally secure steganographic system".

There is a subtle issue in the definition of steganographic security. Requiring that the eavesdropper cannot get information about the hidden message implies that there *is* a hidden message. This is closer to cryptography than it is to steganography, where one focusses on the existence of the message. We believe that many previous definitions did not solve this problem satisfactory.

Another approach[1] uses the relative entropy between X and C as a measure for security; a stego-system is ε-secure under this definition, if $D(C\|X) \le \varepsilon$. We speak of a perfect steganographic system, if $\varepsilon = 0$. Mittelholzer[3] proposed an information-theoretic approach that allows to treat watermarking schemes and steganographic methods in a unified manner. Ettinger[2] proposed a game theoretic definition, which again needs knowledge of the distribution of covers in use.

However, it was first noted by Moskowitz et al.[4] that these models might not be appropriate to define security in steganographic systems formally. Their main argument goes as follows: "In steganography, the discovery of hidden information is not modeled in a continuous manner. We must readdress our old paradigms for secure systems to deal with discontinuities. Standard information theoretic models do not deal with *jumps*."

We agree with this proposal; in our opinion, information-theoretical models have the following main drawbacks:

- As in cryptography, it might not be easy to construct unconditionally secure steganographic systems (recall that in cryptography all known "perfect" systems, as the Vernam scheme, are indeed not practical). It turned out that most perfectly-secure steganographic systems were just some variant of the Vernam scheme under the previous security definitions.
- The probability distribution of C is *not* known in practice; although, for instance, some approximative models might be available for the set of all "meaningful" gray-scale images, it might be infeasible to compute an exact distribution. The problem is even more complicated by the fact that an attacker must find a model for covers that are "usually" sent between two principals (it might not be sufficient to work with a "general" model of e.g. grayscale images).
- If one works with an approximated probability distribution for covers, it might be possible that the modifications applied through the steganographic system are in fact smaller than the approximation error. In this case, the approximated distribution is useless in the decision process.
- It is reasonable to assume that an eavesdropper has only access to a computing device with *limited* computing power. As in cryptography, one might be satisfied if a steganographic system passes all probabilistic polynomial tests (assuming some standard model of computation) for solving the steganographic decision problem.

In the light of these severe problems, we propose to model steganographic security as a probabilistic game between the attacker and a judge. Given the ability to observe "normal" communications and to explore the steganographic system in use, the attacker has to decide whether a certain object (which is given to him by a judge) is in fact a plain cover or a stego-object.

3. CONDITIONAL SECURITY OF STEGANOGRAPHIC SYSTEMS

Let C be the set of possible covers (the only requirement is that there is a probabilistic polynomial-time algorithm that produces elements of C); for any $c \in C$ we denote with $\|c\|$ its length in bits. For the sake of simplicity, we assume that secret messages exchanged in the stego system are encoded as strings of zeroes and ones. Furthermore, denote with M the set of all possible messages; normally we let $M = \{0,1\}^*$, however more complicated message sets can be considered as well (as long as there is still a probabilistic polynomial-time algorithm that samples the set M).

Formally, a (symmetric) steganographic system can be defined by a triple $\langle G, E, D\rangle$ of probabilistic polynomial time algorithms. Algorithm G models the key generation process and outputs, on input 1^n (a string consisting of n ones), a random key $k \in \{0,1\}^n$, which will serve as a stego key. Note that the set of keys can be restricted to a subset of $\{0,1\}^n$. By following Kerckhoffs' principle, the security of a stego system should lie entirely in the stego keys (the longer the keys, the more difficult the detection of steganographic communication). Therefore, the length n of the stego key will be referred to as "security parameter".

Algorithm E represents the embedding process and produces on input $c \in C$, $m \in M$ and k (in the range of G), a stego object $s \in C$. Finally, algorithm D outputs, on input s and k, a string $m' \in \{0,1\}^*$, in case the algorithm succeeds. If the stego object s actually contained a secret message m, then $m' = m$. An eavesdropper trying to detect steganographic communication is faced to solve the steganographic decision problem:

Definition 3.1 (Steganographic Decision Problem). *Given $s \in C$, determine if there exists a $k \in \{0,1\}^*$ in the range of G and a message $m \in M$ such that $D(s,k) = m$.*

We can immediately draw an important consequence. A stego system that simply changes the least significant bits of pixels in an arbitrary image cannot be secure (i.e. it is always possible to answer the steganographic decision problem), as long as the set of messages is not structurally restricted, i.e. if $M = \{0,1\}^n$. Let us assume that the stego system operates in the following manner. On input n, G produces a permutation σ on n elements; E scrambles the message bits according to permutation σ and embeds the n message bits in n fixed bits of a cover. D reverses the process, i.e. it extracts n bits from well-known locations of a stego object and permutes the message bits according to σ^{-1} in order to reconstruct the secret message. Now, for any cover c and for every permutation σ' there exists some message m that seems to be embedded in c (simply run D on σ' and c to obtain m); normally, the obtained message will be completely random and non-sensical, but it is a valid message if the set of messages is not restricted. Thus, the answer to the steganographic decision problem is always “yes” in this system.

The problem stems from the intuitive definition of “security” that was adopted in many previous papers. When can we say that a warden “suspects” the presence of steganographic communication? Is this the case if he finds a totally random message that was allegedly exchanged by two communication partners (even if the “message” was probably created by accident) or must he able to find some “meaning” in the exchanged bits? Even worse, even if he has some suspicion that a secret message exchange is going on, this does not mean that he can prove his suspicion to a third person. We adopt a purely syntactical strategy, i.e. secret messages are constrained to have a specific form (which in turn implies that an attacker can actually prove his suspicion to a third person).

We model steganographic security as an interactive game between an eavesdropper and a judge. The eavesdropper can “observe” normal communication on a channel and get information about the stego system in use by retrieving stego-objects containing messages chosen by him. For this purpose, he is equipped with two oracles. One oracle repeatedly generates covers, whereas the second oracle issues the eavesdropper, on input $m \in M$ and $c \in C$, the corresponding stego-object containing m (the oracle acts like a black-box implementation of the stego embedding process for a fixed key, even if the key is unknown to the attacker). Whereas the first oracle simulates objects sent between two communication partners, the second oracle can be used by the eavesdropper to evaluate the internal structure of the steganographic algorithm. Note that both oracles are probabilistic; if the first oracle is queried several times for a cover, it will almost certainly return different objects.

The first oracle is called “steganographic oracle” and can be modeled by an infinite sequence of covers c_i; our security definitions will be given in terms of sets of steganographic oracles, thus avoiding the knowledge of a “true” probability distribution for covers. The oracle records the number of queries and always returns the next cover in the sequence.

Definition 3.2 (Steganographic Oracle). *A steganographic oracle U is an infinite sequence of covers $c_1, c_2, \ldots$, each cover drawn from the set C.*

The second oracle, called “structure evaluation oracle” can be defined as follows:

Definition 3.3 (Structure evaluation oracle). *Let $\langle G, E, D\rangle$ be a stego system and $k \in \{0,1\}^n$ be in the range of $G(1^n)$. A structure evaluation oracle V_k is a “black box” that returns, on input $m \in M$ and $c \in C$, an object $s \in C$ such that $E(c, m, k) = s$ and $D(s, k) = m$ (in case E is probabilistic, the oracle outputs one possible stego object $s \in E(c, m, k)$).*

Thus, a structure evaluation oracle can be used by the eavesdropper to obtain a stego-object containing an arbitrary chosen message m, *without* knowledge of the stego-key in use. By querying the oracle with a fixed message and some “special” cover like an image consisting of constant color, he might get some hints where the secret message will be embedded by the stego system.

The attack now proceeds as follows: an eavesdropper can repeatedly query both oracles (i.e. he can “observe” ordinary communications by using the first oracle and he can construct stego-objects by consulting the second oracle). There are no further restrictions on the computations done by the eavesdropper, except that the whole procedure must be polynomial in the security parameter, i.e. the length of the stego key, and in the maximal cover size. After he has finished his reasoning process, a judge gives him randomly (with probability 1/2) either a plain cover or a stego-object containing some secret message; both objects are produced by querying the first oracle. He is now faced to distinguish these two cases. If the eavesdropper has some *systematic* advantage in distinguishing these two cases after performing the interactive game (over a truly random decision), the stego system obviously leaks information. The “advantage” is defined as the probability of a correct guess minus 1/2. A stego system is said to be conditionally secure, if an eavesdropper can only guess the correct result with a negligibly better probability than random (i.e. his advantage is negligible, see Definition 3.4).

Formally, the attack model can be described by the following interactive game between the eavesdropper, two oracles U and V_k and a judge (we will refer to the following five steps as probabilistic game Z):

- **Step 1.** The judge runs $G(1^{k'})$ to construct a stego key k of length k' and gives the eavesdropper a structure evaluation oracle V_k implementing the embedding algorithm E under key k.

- **Step 2.** The eavesdropper performs polynomial computations. During these computations, he is allowed to query the oracle V_k with n_1 arbitrary messages $m_1, \ldots, m_{n_1}$ and covers $c_1, \ldots, c_{n_1}$, thereby retrieving the corresponding stego-objects $s_1, \ldots, s_{n_1}$, satisfying $E(c_i, m_i, k) = s_i$ and $D(s_i, k) = m_i$ for $1 \leq i \leq n_1$. Furthermore, he queries the oracle U exactly n_2 times to obtain covers $c_1, \ldots, c_{n_2}$. All oracle queries can be interwoven and the input of one query can be dependent on the output of the last oracle queries. The number of oracle queries n_1 and n_2 is not restricted; the only requirement is that the total computation time spent in the game is polynomial. Note that the input to the oracle V_k does not need to be generated by oracle U.

- **Step 3.** After the eavesdropper has finished his reasoning process, a judge selects two covers $c_1, c_2 \in C$ by querying the oracle U twice. Furthermore, he selects a message m and computes $s = E(c_2, m, k)$. He flips a coin and issues the eavesdropper either (i) the cover c_1 or (ii) the stego-object s.

- **Step 4.** The eavesdropper performs a probabilistic test in an attempt to decide whether he was given the stego object s or the plain cover c_1; he publishes his guess. The advantage for the eavesdropper is the probability of a correct guess minus $1/2$ (note that he can always make a random decision and succeed with probability $1/2$).

- **Step 5.** The stego system is secure for oracle U, if the advantage for the eavesdropper is negligible.

We adopt the notion of a "negligible sequence" that is used frequently in cryptography:

Definition 3.4 (Negligible Sequence). *A sequence n_i of non-negative real numbers is negligible, if for all polynomials p there exists an integer i_0 such that $n_i < 1/p(i)$ for all $i \geq i_0$.*

Now we are able to define steganographic security with respect to a fixed steganographic oracle U. For this purpose, we consider only steganographic systems with finite sets of covers that are smaller than some constant n, i.e. we require that all $c \in C$ satisfy $||c|| \leq n$. A stego-system is called U-secure, if for a randomly selected key k and for random decisions during the steps of the interactive game, an eavesdropper has no systematic advantage in winning the game (i.e. the advantage is a negligible sequence with respect to the security parameter k'). Formally:

Definition 3.5 (U-Security). *Let $S = \langle G, E, D \rangle$ be a steganographic system operating on a finite set of covers C such that $\forall c \in C : ||c|| \leq n$ for a fixed constant n. Furthermore, let U be any steganographic oracle, $k \in \{0,1\}^{k'}$ be a stego key in the range of $G(1^{k'})$ and V_k be a structure evaluation oracle implementing key k. We call S U-secure, if the advantage for an eavesdropper in step 5 of the probabilistic game Z is a negligible sequence $p(k')$ with respect to the length k' of the stego key. The probability is taken over all keys k and all internal coin tosses of game Z; the game must be polynomial in both n and k'.*

A stego system is secure for a set of oracles $\mathcal{C}$, if it is secure for each oracle contained in the set.

Definition 3.6 (Conditional Security). *A stego system $S = \langle G, E, D \rangle$ is conditionally secure for a set $\mathcal{C}$ of oracles, if for all steganographic oracles $U \in \mathcal{C}$, S is U-secure. A stego system is conditionally secure, if it is secure for all oracles.*

The term "conditionally" reflects the fact that such schemes are in generally *not* secure from an information-theoretic viewpoint.

Several variations of the definition could be possible. Instead of requiring a stego system to be U-secure for all oracles U, one might be satisfied in case the system is U-secure for all but finitely many oracles. Alternatively, one might require that a system is U-secure for infinitely many oracles. In order to get a definition for unconditional security of steganographic systems, one can remove the requirement that the game must be completed in polynomial time.

4. PROVABLE SECURITY

Given the definition of steganographic security detailed in the last section, one might be interested in finding an actual stego system S that satisfies this property. Unfortunately, it might be quite difficult to prove this property directly. However, one can try to base the security of S on some class of computational problems P that is believed to be intractable (e.g. on some cryptographic primitives that are believed to be secure). For this purpose, one constructs a "reduction" from P to the steganographic decision problem for S. Such a reduction can be outlined as follows. Assume that S is not secure in the sense of Definition 3.6 (for an arbitrary set of oracles), implying that there exists some probabilistic game Z between an eavesdropper and a judge that allows the eavesdropper to decide the steganographic decision problem for S with non-negligible probability. One has to show that under this assumption, instances of P can be solved as well (again with non-negligible probability), contradicting the intractability of P. To show this, one has to turn the interactive game Z into a randomized (non-interactive) algorithm Z' by replacing all oracle queries by (possibly randomized) computations; one can memorize this by "Z' has to answer all oracle queries itself".

Although the construction of provably secure schemes remains an open problem, we illustrate this procedure with a simple example of a stego system in a truly pseudorandom channel. Let n be an RSA modulus (i.e. a product of two distinct large primes p and q); in case more information on the RSA system is required, we refer to Katzenbeisser.[5] Assume that all messages that are sent in the communication channel are elements of $\mathbb{Z}_n$. We can describe a steganographic system $S = \langle G, E, D \rangle$ in the following way: let G be the key generation of the RSA public-key cryptosystem. Thus, algorithm G outputs, on input $1^{n'}$ a triple $\langle e, d, n \rangle$ of integers, where n is an RSA modulus of size n' and e and d satisfy

$$ed \equiv 1 \pmod{(p-1)(q-1)}.$$

The set of covers consists of all RSA-encrypted strings whose corresponding plaintext ends with a 0 in the binary expansion, whereas the set of stego-objects contains all strings whose corresponding plaintext ends with a 1. The embedding algorithm E adds a zero at the end of a secret message m, pads the message with random bits and encrypts it. The detection process D decrypts a potential stego-object and checks whether the LSB of the plaintext equals zero. If this is the case, the other bits correspond to the secret message, whereas the message is meaningless otherwise.

It is obvious that such a system cannot be unconditionally (information-theoretically) secure, as an eavesdropper can always try to break RSA by brute-force key search and decrypt all messages sent on the channel. However, the system can be seen as conditionally secure by the following argument. It is well-known that, under the so-called RSA assumption, computing the least significant bit is a hard-core predicate for the RSA function. In other words, any algorithm that guesses the least significant bit of a string, given only its RSA encrypted ciphertext, can be used as an oracle to break RSA. We will construct a reduction from guessing the LSB of an RSA-encrypted plaintext to the steganographic decision problem for S. As guessing the LSB of an RSA plaintext is computationally equivalent to breaking RSA as a whole, we would have invented a new way to attack the RSA scheme, which is believed to be computationally intractable.

Let us assume that the outlined stego-system S is not unconditionally secure, i.e. that there exists a game Z between an eavesdropper and a judge that allows the attacker to decide whether a given element of $\mathbb{Z}_n$ is a cover or a stego object. We will show that under this assumption there exists a probabilistic algorithm Z' that guesses the least significant bit of an RSA encrypted plaintext better than random, thereby contradicting the RSA hypothesis.

Let $x \in \mathbb{Z}_n$ be any ciphertext and $y = x^d \bmod n$ be the corresponding plaintext; we describe an algorithm Z' that decides whether the least significant bit of y equals one. Algorithm Z' simulates the game Z, but has to answer all oracle queries by the eavesdropper itself. If the eavesdropper asks for a cover c_i, Z' selects a string y_i with least significant bit zero randomly, encrypts it and returns the resulting string as oracle result. Conversely, if the eavesdropper asks for a cover that has a message m_i embedded, Z' appends a 1 as least significant bit, pads the message with random bits, encrypts the result and assumes the resulting string to be the oracle output. In step 4, Z' always returns x as result of the oracle. By assumption, the eavesdropper can now decide whether x has least significant bit zero or one with non-negligible probability; thus, also Z', who simulates the game, can make this decision. Thus, we have constructed a probabilistic algorithm Z' which decides the least significant bit of y, which in turn can be extended to an algorithm that breaks RSA itself, thereby violating the RSA hypothesis.

Note that the actual structure of the game Z is unknown; we just know that such a game Z exists and that it can be simulated by a probabilistic algorithm that has access to two oracles. When answering the oracle queries we have to be careful so that the simulated oracle answers remains correct, i.e. are a possible oracle output for the set of steganographic oracles (or structure evaluation oracles) we are working with.

It is possible to extend the scheme to a more practical one. Assume now that we have access to an embedding function E' that embeds a binary string m into a digital image o by modifying the least significant bits of o in such a way that the distribution of the least significant bits remains unchanged. Then we can construct a stego scheme in the following way: the key generation process remains unchanged. Covers for secret transmissions are images that are modified in the following way: choose any string x, append a zero, encrypt the result with RSA and embed the resulting string in the image using the operation E'. The stego embedding rule takes the secret message m, appends a 1, pads the string with random bits, encrypts the result and embeds the encrypted message in the image.

Thus, the main idea for the scheme is that one party *always* embeds some string in the cover; in case of a steganographic communication, this string resembles the secret message, otherwise it is just random. By using a similar reduction as outlined previously, the security of the scheme can be established. Instead of returning an encrypted string as oracle result, the string is embedded in some given image using E'. Again, if the eavesdropper is able to distinguish covers from stego objects, he is able to guess the LSB of an RSA-encrypted ciphertext.

The previous system makes another possible weakness of all definitions for steganographic security apparent. We modified every message sent in the communication channel steganographically and used cryptography to conceal this action. Any eavesdropper is faced to decide the steganographic decision problem soley with the knowledge of the covers used within this communication channel. Especially, in our model he is *not* allowed to use any "external information", like some "normal" images found outside the communication channel. Although this seems to be unrealistic, it is a problem present in all previous security definitions.

5. CONCLUSION AND FUTURE WORK

In this paper, we proposed a new definition for steganographic security. Instead of relying on information-theoretic approaches, we use an indistinguishability test to establish security. Knowledge of a "true" probability distribution for covers is substituted by an oracle mechanism that might be easier to handle when giving proofs of security. However, the constuction of practical provably secure steganographic schemes remains an open problem. Besides this issue, future work includes the investigation of steganalysis methods that conform to Definition 3.6 (one possible research direction would be applying Bayesian learning techniques). Furthermore, the implications of choosing special classes of steganographic oracles on the decision strategy has to be adressed.

REFERENCES

1. C. Cachin, "An Information-Theoretic Model for Steganography", in *Information Hiding: Second International Workshop*, vol. 1525 of LNCS, Springer, 1998, pp. 306–318.
2. J. M. Ettinger, "Steganalysis and Game Equilibria", in *Information Hiding: Second International Workshop*, vol. 1525 of LNCS, Springer, 1998, pp. 319–328.
3. T. Mittelholzer, "An Information-Theoretic Approach to Steganography and Watermarking", in *Information Hiding: Third International Workshop*, vol. 1768 of LNCS, Springer, 1999, pp. 1–16.
4. I. S. Moskowitz, G. E. Longdon, L. Chang, "A New Paradigm Hidden in Steganography", in *New Security Paradigms Workshop 2000, Proceedings*, ACM Press, pp. 41–50.
5. S. Katzenbeisser, *Recent Advances in RSA Cryptography*, Kluwer Academic Publishers, 2001.
6. G. Simmons, "The Prisoners' Problem and the Subliminal Channel", in *Advances in Cryptology, CRYPTO'83*, Plenum Press, 1984, pp. 51–67.
7. J. Zöllner et al., "Modeling the Security of Steganographic Systems", in *Information Hiding: Second International Workshop*, vol. 1525 of LNCS, Springer, 1998, pp. 344–354.