A Cryptographic Approach for Steganography

Jacques M. Bahi, Christophe Guyeux, and Pierre-Cyrille Heam*
FEMTO-ST Institute, UMR 6174 CNRS
Computer Science Laboratory DISC
University of Franche-Comté, France
{jacques.bahi, christophe.guyeux, pierre-cyrille.heam}@femto-st.fr
*Authors are cited in alphabetic order

***Abstract*—In this research work, security concepts are formalized in steganography, and the common paradigms based on information theory are replaced by another ones inspired from cryptography, more practicable are closer than what is usually done in other cryptographic domains. These preliminaries lead to a first proof of a cryptographically secure information hiding scheme.**

***Index Terms*—Information hiding; Steganography; Security; Cryptographic proofs.**

I. Introduction

The usual manner for preserving privacy when communicating over public channels is by using cryptographic tools. Users cipher the data and send them over possibly insecure networks. Even if a third party intercepts these data, he or she will not understand them without having the secret key for deciphering. In that well investigated scenario, anyone knows that a private message is transmitted through the public channel, but only authorized individuals (*i.e.*, owners of the secret key) can understand it.

A second approach investigated over two decades [6], and usually referred as information hiding or steganography [2], [9], aims at inserting a secret message into an innocent cover, in such a way that observers cannot detect the existence of this hidden channel (for instance, images sent through the Internet). The goal in this field is to appear as innocent as possible: observers should not think that something goes wrong with this public channel. It must not cross their mind that sometimes the public channel is used to transmit hidden messages. In that context, an attack is succeeded when the sleazy character of the channel is detected. Tools used in that field are mainly based on artificial intelligence. They are called steganalyzers, and their main objective is to detect whether a given communication channel is possibly steganographied, or if it only contains "natural" images. In case of detection, the unique countermeasure proposed by the literature is to stop the sleazy communication by closing the channel. To sum up, the steganography community currently only focuses on the ability to detect hidden channels, without investigating the consequences of this detection [3], [11].

However, observers have not necessarily the ability or the desire to stop the communication. For instance, who can switch off the Internet? Furthermore, by stopping the faked channel, attackers miss the opportunity to obtain more information about the secret message and the intended receiver. Finally, if attackers observe the communication, man can reasonably think that they already knew in advance that this channel is sleazy (if not, why they observe it?). The use of a steganalyzer on a channel only appears in the best situation as a reinforcement of their doubts or fears. In most operational contexts, only sleazy channels are observed, and the questions are finally to determine [4]: (1) when the hidden messages have been transmitted in this channel (among all the possibly faked images, how to determine the ones that really contain hidden information?), (2) what was the content of this message, and perhaps (3) who was the receiver among the observers. These questions make sense only within a stegranographic context, that is, when the channel is not ciphered. However, these important questionings have never been regarded by the information hiding community.

In this paper, authors provide a cryptographic theoretical framework to study this scenario related to steganography. Concrete illustrative examples of this framework of study are given thereafter. A first toy example is the hypothetical case of a dissident blogger in a totalitarian state, who posts regularly and publicly information in his or her blog, while being severely watched by the authorities. This blogger wants to transmit one day a secret message or a signal to an observer into confidence, without sounding the alarm in the authorities side. Another example is an individual who is invigilated, because he is correctly suspected to be a spy. This agent cannot be arrested on a simple presumption, or on the claim that the images he sent in his emails look sleazy. Despite this surveillance, this spy wants to transmit one day a message to his sponsor. The observers want to determine if an hidden message is really transmitted or not, to have a proof of such a transmission, together with the content of the message, the date of transmission, and the targeted receiver if possible. Obviously, these situations are related to both cryptography and steganography, however there is currently a lack of tool allowing their study. The key idea of this research work is to propose algorithms such that observers cannot switch from doubts (sleazy channels) to certainties or proofs.

The remainder of this article is organized as follows. In Section II, generalities from steganography are discussed. The key concepts and main results are presented in Section III. Finally, Section IV concludes this research work and details further investigations.

II. Notions and Terminologies in Information Hiding

In the following some common notions in the field of information hiding are recalled. We refer to [1] for a complete survey of this subject.

A. Information Hiding Security

Robustness and security are two major concerns in information hiding. These two concerns have been defined in [12] as follows. "Robust watermarking is a mechanism to create a communication channel that is multiplexed into original content [...]. It is required that, firstly, the perceptual degradation of the marked content [...] is minimal and, secondly, that the capacity of the watermark channel degrades as a smooth function of the degradation of the marked content. [...]. Watermarking security refers to the inability by unauthorized users to have access to the raw watermarking channel [...] to remove, detect and estimate, write or modify the raw watermarking bits."

In the framework of watermarking and steganography, security has seen several important developments since the last decade [5], [8], [13]. The first fundamental work in security was made by Cachin in the context of steganography [6]. Cachin interprets the attempts of an attacker to distinguish between an innocent image and a stego-content as a hypothesis testing problem. In this document, the basic properties of a stegosystem are defined using the notions of entropy, mutual information, and relative entropy. Mittelholzer, inspired by the work of Cachin, proposed the first theoretical framework for analyzing the security of a watermarking scheme [15].

These efforts to bring a theoretical framework for security in steganography and watermarking have been followed up by Kalker, who tries to clarify the concepts (robustness *vs.* security), and the classifications of watermarking attacks [12]. This work has been deepened by Furon *et al.*, who have translated Kerckhoffs' principle (Alice and Bob shall only rely on some previously shared secret for privacy), from cryptography to data hiding [10]. They used Diffie and Hellman methodology, and Shannon's cryptographic framework [17], to classify the watermarking attacks into categories, according to the type of information Eve has access to [8], [16], namely: Watermarked Only Attack (WOA), Known Message Attack (KMA), Known Original Attack (KOA), and Constant-Message Attack (CMA). Levels of security have been recently defined in these setups. The highest level of security in WOA is called stego-security [7], recalled below.

In the prisoner problem of Simmons [18], Alice and Bob are in jail, and they want to, possibly, devise an escape plan by exchanging hidden messages in innocent-looking cover contents. These messages are to be conveyed to one another by a common warden, Eve, who over-drops all contents and can choose to interrupt the communication if they appear to be stego-contents. The stego-security, defined in this framework, is the highest security level in WOA setup [7]. To recall it, we need the following notations:

- $\mathbb{K}$ is the set of embedding keys,
- $p(X)$ is the probabilistic model of N_0 initial host contents,
- $p(Y|K_1)$ is the probabilistic model of N_0 watermarked contents.

Furthermore, it is supposed in this context that each host content has been watermarked with the same secret key K_1 and the same embedding function e. It is now possible to define the notion of stego-security.

Definition 1 (Stego-Security): The embedding function e is *stego-secure* if and only if:

$$\forall K_1 \in \mathbb{K}, p(Y|K_1) = p(X).$$

This definition is almost always considered as not really tractable in practice, reasons explaining this mistrust are outlined in the following section. This is the reason why the information hiding community majorly focuses on the construction of steganalyzers, supposed to be able to determine whether a given communication channel appears to transmit steganographied messages or not.

B. Drawbacks of the Stego-Security Notion

Theoretically speaking, the stego-security notion matches well with the idea of a perfect secrecy in the WOA category of attacks. However, its concrete verification raises several technical problems difficult to get around. These difficulties impact drastically the effective security of the scheme.

For instance, in a stego-secure scheme, the distribution of the set of watermarked images must be the same than the one of the original contents, no matter the chosen keys. But *how to determine practically the distribution of the original contents?* Furthermore, claiming that Alice can constitutes her own subset of well-chosen images having the same "good" distribution is quite unreasonable in several contexts of steganography: Alice has not *always* the choice of the supports. Moreover, it introduces a kind of bias, as the warden can find such similarities surprising. Suppose however that Alice is in the best situation for her, that is, she has the possibility to constitute herself the set of original contents. How can she proceed practically to be certain that all media into the set follow a same distribution $p(X)$? According to the authors opinion, Alice has two possible choices:

1) Either she constitutes the set by testing, for each new content, whether this media has a same distribution than the ones that have been already selected.
2) Or she forges directly new images by using existing ones. For instance, she can replace all the least significant bits of the original contents by using a good pseudorandom number generator.

In the first situation, Alice will realize a χ^2 test, or other statistical tests of this kind, to determine if the considered image (its least significant bits, or its low frequency coefficients, etc.) has a same distribution than images already selected. In that situation, Alice does not have the liberty to choose the distribution, and it seems impossible to find a scheme being able to preserve any kind of distribution, for all secret

keys and all hidden messages. Furthermore, such statistical hypothesis testing are not ideal ones, as they only regard if a result is unlikely to have occurred by chance alone *according to a pre-determined threshold probability* (the significance level). Errors of the first (false positive) and second kind (false negative) occur necessarily, with a certain probability. In other words, with such an approach, Alice cannot design a perfect set of cover contents having all the same probability $p(X)$. This process leads to a set of media that follows a distribution Alice does not have access to.

The second situation seems more realistic, it will thus be further investigated in the next section.

III. Toward a Cryptographically Secure Hiding

In this section a theoretical framework for information hiding security is proposed, which is more closely resembling that of usual approaches in cryptography. It allows to define the notion of steganalyzers, it is compatible with the new original scenarios of information hiding that have been dressed in the previous sections, and it does not have the drawbacks of the stego-security definition.

A. Introduction

Almost all branches in cryptology have a complexity approach for security. For instance, in a cryptographic context, a pseudorandom number generator (PRNG) is a deterministic algorithm G transforming strings of length ℓ into strings of length M, with $M > \ell$. The notion of *secure* PRNG can be defined as follows [19].

Definition 2: Let $\mathcal{D} : \mathbb{B}^M \longrightarrow \mathbb{B}$ be a probabilistic algorithm that runs in time T. Let $\varepsilon > 0$. $\mathcal{D}$ is called a $(T, \varepsilon)-$distinguishing attack on pseudorandom generator G if

$$\left| Pr[\mathcal{D}(G(k)) = 1 \mid k \in_R \{0,1\}^\ell] - Pr[\mathcal{D}(s) = 1 \mid s \in_R \mathbb{B}^M] \right| \geqslant \varepsilon,$$

where the probability is taken over the internal coin flips of $\mathcal{D}$, and the notation "$\in_R$" indicates the process of selecting an element at random and uniformly over the corresponding set.

Let us recall that the running time of a probabilistic algorithm is defined to be the maximum of the expected number of steps needed to produce an output, maximized over all inputs; the expected number is averaged over all coin flips made by the algorithm [14]. We are now able to recall the notion of cryptographically secure PRNG.

Definition 3: A pseudorandom generator is $(T, \varepsilon)-$secure if there exists no $(T, \varepsilon)-$distinguishing attack on this pseudorandom generator.

Intuitively, it means that no polynomial-time algorithm can make a distinction, with a non-negligible probability, between a truly random generator and G.

Inspired by these kind of definitions, we propose what follows.

B. Definition of a stegosystem

Definition 4 (Stegosystem): Let $\mathcal{S}, \mathcal{M}$, and $\mathcal{K} = \mathbb{B}^\ell$ three sets of words on $\mathbb{B}$ called respectively the sets of supports, of messages, and of keys (of size ℓ).

A *stegosystem* on $(\mathcal{S}, \mathcal{M}, \mathcal{K})$ is a tuple $(\mathcal{I}, \mathcal{E}, inv)$ such that:

- $\mathcal{I}$ is a function from $\mathcal{S} \times \mathcal{M} \times \mathcal{K}$ to $\mathcal{S}$, $(s, m, k) \longmapsto \mathcal{I}(s, m, k) = s'$,
- $\mathcal{E}$ is a function from $\mathcal{S} \times \mathcal{K}$ to $\mathcal{M}$, $(s, k) \longmapsto \mathcal{E}(s, k) = m'$.
- inv is a function from $\mathcal{K}$ to $\mathcal{K}$, s.t. $\forall k \in \mathcal{K}, \forall (s, m) \in \mathcal{S} \times \mathcal{M}$, $\mathcal{E}(\mathcal{I}(s, m, k), inv(k)) = m$.
- $\mathcal{I}(s, m, k)$ and $\mathcal{E}(c, k')$ can be computed in polynomial time.

$\mathcal{I}$ is called the insertion or embedding function, $\mathcal{E}$ the extraction function, s the host content, m the hidden message, k the embedding key, $k' = inv(k)$ the extraction key, and s' is the stego-content. If $\forall k \in \mathcal{K}, k = inv(k)$, the stegosystem is symmetric (private-key), otherwise it is asymmetric (public-key).

C. Heading Notions

Definition 5 ($(T, \varepsilon)-$distinguishing attack): Let $S = (\mathcal{I}, \mathcal{E}, inv)$ a stegosystem on $(\mathcal{A}, \mathcal{M}, \mathcal{K})$, with $\mathcal{A} \subset \mathbb{B}^M$. A $(T, \varepsilon)-$distinguishing attack on the stegosystem S is a probabilistic algorithm $\mathcal{D} : \mathcal{A} \longrightarrow \{0, 1\}$ in running time T, such that there exists $m \in \mathcal{M}$,

$$\begin{aligned} &|Pr\left[\mathcal{D}\left(\mathcal{I}(s, m, k)\right) = 1 \mid k \in_R \mathcal{K}, s \in_R \mathcal{A}\right] \\ &\qquad - Pr\left[\mathcal{D}\left(x\right) = 1 \mid x \in_R \mathcal{A}\right]| \geqslant \varepsilon, \end{aligned}$$

where the probability is also taken over the internal coin flips of $\mathcal{D}$, and the notation $\in_R$ indicates the process of selecting an element at random and uniformly over the corresponding set.

Definition 6: A stegosystem is $(T, \varepsilon)-$undistinguishable if there exists no $(T, \varepsilon)-$distinguishing attack on this stegosystem.

Intuitively, it means that there is no polynomial-time probabilistic algorithm being able to distinguish the host contents from the stego-contents

D. A Cryptographically Secure Information Hiding Scheme

Theorem 1: Let

$$\mathcal{S} = \left\{ s_1^1, s_2^1, \ldots, s_{2^N}^1, s_1^2, s_2^2, \ldots, s_{2^N}^2, \ldots, s_1^r, s_2^r, \ldots, s_{2^N}^r \right\}$$

a subset of $\mathbb{B}^M = \mathcal{A}$. Consider $G : \mathbb{B}^L \longrightarrow \mathbb{B}^N$ a $(T, \varepsilon)-$secure pseudorandom number generator, and $\mathcal{I}(s_j^i, m, k) = s^i_{m \oplus G(k)}$. Assuming that r is a constant, and that from i, j one can compute the image s_j^i in time T_1, the steosystem is $(T - T_1 - N - 1, \varepsilon)$-secure.

Intuitively, $\mathcal{S}$ is built from r images containing N bits of low information. The image s_j^i corresponds to the i-th image where the N bits are set to j.

Proof 1: Assume there exists a (T', ε) distinguisher $\mathcal{D}'$for the stego-system. Therefore, there exists m_0 such that

$$\begin{aligned} &\left|Pr\left(\mathcal{D}'\left(\mathcal{I}(s, m_0, k)\right) = 1 \mid k \in_R \mathbb{B}^\ell, s \in_R \mathcal{S}\right)\right. \\ &\left. -Pr\left(\mathcal{D}'\left(x\right) = 1 \mid x \in_R \mathcal{S}\right)\right| \geqslant \varepsilon \end{aligned} \tag{1}$$

Choosing randomly and uniformly $s \in \mathcal{S}$ is equivalent to choose uniformly and randomly $i \in \{1, \ldots, r\}$ and $j \in$

$\{1,\ldots,2^N\}$. Therfore (1) is equivalent to

$$\begin{array}{l}\Big|Pr\left(\mathcal{D}'\left(s^i_{m_0\oplus G(k)}\right)=1 \mid k\in_R \mathbb{B}^\ell, i\in_R \{1,\ldots,r\}\right)\\ -Pr\left(\mathcal{D}'(x)=1 \mid x\in_R \mathcal{S}\right)\Big| \geqslant \varepsilon\end{array} \tag{2}$$

Let $\mathcal{D}$ be the distinguisher for G defined for $y\in\{0,1\}^N$ into $\{0,1\}$ by:

1) Pick randomly and uniformly $i\in\{1,\ldots,r\}$.
2) Compute $s=s^i_{m_0\oplus y}$.
3) Return $\mathcal{D}'(s)$.

The complexity of this probabilistic algorithm is 1 for the first step since r is a constant, T_1+N for the second step, and T' for the last one. Thus it works in thime $T'+T_1+1+N$.

Now we claim that $\mathcal{D}$ is a $(T'+T_1+1+N,\varepsilon)$-distinguisher for G. Indeed,

$$\begin{array}{l}Pr\left(\mathcal{D}(y)=1 \mid y\in_R \{0,2^N\}\right)\\ =Pr\left(\mathcal{D}'\left(s^i_y\right)=1 \mid y\in_R \{0,2^N\}, i\in_R \{1,\ldots,r\}\right)\\ =Pr\left(\mathcal{D}'(x)=1 \mid x\in_R \mathcal{S}\right).\end{array}$$

Moreover,

$$\begin{array}{l}Pr\left(\mathcal{D}(G(k))=1 \mid k\in_R \{0,1\}^\ell\right)\\ =Pr\left(\mathcal{D}'\left(s^i_{m_0\oplus G(k)}\right)=1 \mid k\in_R \{0,1\}^\ell, i\in_R \{1,\ldots,r\}\right).\end{array}$$

Therefore, using (2), one has

$$\begin{array}{l}\Big|Pr[\mathcal{D}(G(k))=1 \mid k\in_R \{0,1\}^\ell]\\ -Pr[\mathcal{D}(s)=1 \mid s\in_R \mathbb{B}^M]\Big| \geqslant \varepsilon,\end{array} \tag{3}$$

proving that $\mathcal{D}$ is a $(T'+T_1+1+N,\varepsilon)$-distinguisher for G, which concludes the proof.

IV. Conclusion

In this research work, a new rigorous approach for secure steganography, based on the complexity theory, has been proposed. This work has been inspired by the definitions of security that can usually be found in other branches of cryptology. We have proposed a new understanding for the notion of *secure hiding* and presented a first secure information hiding scheme. The intention was to prove the existence of such a scheme and to give a rigorous cryptographical framework for steganography.

In future work, we will investigate the situation where detection is impossible. In that case, we will consider both weak indistinguability (using a statistical or a complexity approach, with the cryptographically secure definition of PRNGs) and strong indistinguability (using the well known CC1 and CC2 sets). Additionally, we will reconsider and improve the definitions of security in the information hiding literature that are based on the signal theory. Among other thing, we will take into account a Shannon entropy that is not reduced to simple 1-bit blocs. Finally, we will show that tests using generators allow to attack information hiding schemes that are secure for the statistical approach, as LSB are not uniform in that situation.

References

[1] Eman Abdelfattah and Ausif Mahmood. Steganography and steganalysis: Current status and future directions. In Tarek Sobh and Khaled Elleithy, editors, *Emerging Trends in Computing, Informatics, Systems Sciences, and Engineering*, volume 151 of *Lecture Notes in Electrical Engineering*, pages 411–422. Springer New York, 2013.

[2] Jacques Bahi, Jean-François Couchot, and Christophe Guyeux. Steganography: a class of algorithms having secure properties. In *IIH-MSP-2011, 7-th Int. Conf. on Intelligent Information Hiding and Multimedia Signal Processing*, pages 109–112, Dalian, China, October 2011.

[3] Jacques Bahi, Nicolas Friot, and Christophe Guyeux. Lyapunov exponent evaluation of a digital watermarking scheme proven to be secure. In *IIH-MSP'2012, 8-th Int. Conf. on Intelligent Information Hiding and Multimedia Signal Processing*, pages 359–362, Piraeus-Athens, Greece, July 2012. IEEE Computer Society.

[4] Jacques Bahi and Christophe Guyeux. A new chaos-based watermarking algorithm. In *SECRYPT'10, Int. conf. on security and cryptography*, pages 455–458, Athens, Greece, July 2010. SciTePress.

[5] Mauro Barni, Franco Bartolini, and Teddy Furon. A general framework for robust watermarking security. *Signal Processing*, 83(10):2069–2084, 2003. Special issue on Security of Data Hiding Technologies, invited paper.

[6] Christian Cachin. An information-theoretic model for steganography. In *Information Hiding*, volume 1525 of *Lecture Notes in Computer Science*, pages 306–318. Springer Berlin / Heidelberg, 1998.

[7] F. Cayre and P. Bas. Kerckhoffs-based embedding security classes for woa data hiding. *IEEE Transactions on Information Forensics and Security*, 3(1):1–15, 2008.

[8] F. Cayre, C. Fontaine, and T. Furon. Watermarking security: theory and practice. *IEEE Transactions on Signal Processing*, 53(10):3976–3987, 2005.

[9] Nicolas Friot, Christophe Guyeux, and Jacques Bahi. Chaotic iterations for steganography - stego-security and chaos-security. In Javier Lopez and Pierangela Samarati, editors, *SECRYPT'2011, Int. Conf. on Security and Cryptography. SECRYPT is part of ICETE - The International Joint Conference on e-Business and Telecommunications*, pages 218–227, Sevilla, Spain, July 2011. SciTePress.

[10] T. Furon. Security analysis, 2002. European Project IST-1999-10987 CERTIMARK, Deliverable D.5.5.

[11] Christophe Guyeux, Nicolas Friot, and Jacques Bahi. Chaotic iterations versus spread-spectrum: chaos and stego security. In *IIH-MSP'10, 6-th Int. Conf. on Intelligent Information Hiding and Multimedia Signal Processing*, pages 208–211, Darmstadt, Germany, October 2010.

[12] T. Kalker. Considerations on watermarking security. pages 201–206, 2001.

[13] Andrew D. Ker. Batch steganography and pooled steganalysis. In Jan Camenisch, Christian S. Collberg, Neil F. Johnson, and Phil Sallee, editors, *Information Hiding*, volume 4437 of *Lecture Notes in Computer Science*, pages 265–281, Alexandria, VA, USA, July 2006. Springer.

[14] D. E. Knuth. *Seminumerical Algorithms*, volume 3. Addison-Wesley, Reading, MA, USA, third edition edition, 1997.

[15] Thomas Mittelholzer. An information-theoretic approach to steganography and watermarking. In Andreas Pfitzmann, editor, *Information Hiding*, volume 1768 of *Lecture Notes in Computer Science*, pages 1–16, Dresden, Germany, September 29 - October 1. 1999. Springer.

[16] Luis Perez-Freire, F. Prez-gonzalez, and Pedro Comesaa. Secret dither estimation in lattice-quantization data hiding: A set-membership approach. In Edward J. Delp and Ping W. Wong, editors, *Security, Steganography, and Watermarking of Multimedia Contents*, San Jose, California, USA, January 2006. SPIE.

[17] Claude E. Shannon. Communication theory of secrecy systems. *Bell Systems Technical Journal*, 28:656–715, 1949.

[18] Gustavus J. Simmons. The prisoners' problem and the subliminal channel. In *Advances in Cryptology, Proc. CRYPTO'83*, pages 51–67, 1984.

[19] Andrew C. Yao. Theory and application of trapdoor functions. In *Proceedings of the 23rd Annual Symposium on Foundations of Computer Science*, SFCS '82, pages 80–91, Washington, DC, USA, 1982. IEEE Computer Society.