Ensuring message embedding in wet paper steganography

Daniel Augot, Morgan Barbier, Caroline Fontaine

Ensuring message embedding in wet paper steganography

Daniel Augot[1], Morgan Barbier[1], and Caroline Fontaine[2]

[1] Computer science laboratory of École Polytechnique
INRIA Saclay – Île de France
[2] CNRS/Lab-STICC and Télécom Bretagne, Brest, France

Abstract. Syndrome coding has been proposed by Crandall in 1998 as a method to stealthily embed a message in a cover-medium through the use of bounded decoding. In 2005, Fridrich *et al.* introduced wet paper codes to improve the undetectability of the embedding by enabling the sender to lock some components of the cover-data, according to the nature of the cover-medium and the message. Unfortunately, almost all existing methods solving the bounded decoding syndrome problem with or without locked components have a non-zero probability to fail. In this paper, we introduce a randomized syndrome coding, which guarantees the embedding success with probability one. We analyze the parameters of this new scheme in the case of perfect codes.

Keywords: steganography, syndrome coding problem, wet paper codes.

1 Introduction

Hiding messages in innocuous-looking *cover-media* in a *stealthy* way, steganography is the art of stealth communications. The sender and receiver may proceed by cover selection, cover synthesis, or cover modification to exchange messages. Here, we focus on the cover modification scenario, where the sender chooses some *cover-medium* in his library, and modifies it to carry the message she wants to send. Once the cover-medium is chosen, the sender extracts some of its components to construct a *cover-data* vector. Then, she modifies it to embed the message. This modified vector, called the *stego-data*, leads back to the *stego-medium* that is communicated to the recipient. In the case of digital images, the insertion may for example consist in modifying some of the images components, *e.g.* the luminance of the pixels or the values of some transform (DCT or wavelet) coefficients. For a given transmitted document, only the sender and receiver have to be able to tell if it carries an hidden message or not [33]. This means that the *stego-media*, which carry the messages, have to be

statistically indistinguishable from original media [6,7]. But statistical detectability of most steganographic schemes increases with *embedding distortion* [24], which is often measured with the number of embedding changes. Hence it is of importance for the sender to embed the message while modifying as less components of the cover-data as possible.

In 1998, Crandall proposed to model the embedding and extraction process with the use of linear error correcting codes. He proposed to use Hamming codes, which are covering codes [9]. The key idea of this approach, called *syndrome coding*, or *matrix embedding*, is to modify the cover-data to obtain a stego-data lying in the *right coset* of the code, its *syndrome* being precisely equal to the message to hide. Later on, it has been showed that designing steganographic schemes is precisely equivalent to designing covering codes [3,22,23], meaning that this covering codes approach is not restrictive. Moreover, it has been shown to be really helpful and efficient to minimize the embedding distortion [3,22,23,4]. It has also been made popular due to its use in the famous steganographic algorithm F5 [36]. For all these reasons, this approach is of interest.

The process which states which components of the cover-data can actually be modified is called the *selection channel* [1]. Since the message embedding should introduce as little distortion as possible, the selection channel is of utmost importance. The selection channel may be arbitrary, but a more efficient approach is to select it dynamically during the embedding step, accordingly to the cover-medium and the message. This leads to a better undetectability, and makes attacks on the system harder to run, but in this context the extraction of the hidden message is more difficult as the selection channel is only known to the sender, and not to the recipient. *Wet Paper Codes* were introduced to tackle this non-shared selection channel, through the notions of *dry* and *wet* components [18]. By analogy with a sheet of paper that has been exposed to rain, we can still write easily on dry spots whereas we cannot write on wet spots. The idea is, adaptively to the message and the cover-medium, to *lock* some components of the cover-data — the wet components — to prevent them being modified. The other components — the dry components — of the cover-data remain free to be modified to embed the message.

Algorithmically speaking, syndrome coding provides the recipient an easy way to access the message, through a simple syndrome computation. But to embed the message, the sender has to tackle an harder challenge, linked with bounded syndrome decoding. It has been shown that if random codes may seem interesting for their asymptotic behavior, their use leads to solve really hard problems: syndrome decoding

and covering radius computation, which are proved to be NP-complete and Π_2-complete respectively [34,25]. Moreover, no efficient decoding algorithm is known, for generic, or random, codes. Hence, attention has been given on structured codes to design Wet Paper Codes: Hamming codes [9,21], Simplex codes [20], BCH codes [31,32,37,30,27], Reed-Solomon codes [14,15], perfect product codes [29,28], low density generator matrix codes [17,39,38,10], and convolutional codes [13,11,12].

Embedding techniques efficiency is usually evaluated through their relative payload (number of message symbols per cover-data (modifiable) symbol) and average embedding efficiency (average number of message symbols per cover-data modification). Today, we can find in the literature quasi-optimal codes in terms of average embedding efficiency and payload [17,39,38,16,10]. Nevertheless, we are interested here in another criterion, which is usually not discussed: the probability for the embedding to fail. In fact, the only case for which it never fails is when using perfect codes (a), without locking any component of the cover-data (b). But very few codes are perfect (namely the Hamming and Golay codes), and their average embedding efficiency is quite low. Moreover it is really important in practice to be able to lock some components of the cover-data. Hence, efficient practical schemes usually do not satisfy either condition (a) or condition (b), leading to a non-zero probability for the embedding to fail. And this probability increases with the number of locked components. More precisely, syndrome coding usually divides the whole message into fragments, that are separately inserted in different cover-data vectors (coming from one or several cover-medium). Inserting each fragment involves finding a low weight solution of a linear system which may not always have a solution for a given set of locked components. Consequently, the probability that the whole message can be embedded decreases exponentially with the number of fragments to hide and with the number of locked components [21]

Hence, we have to decide what to do when embedding fails. In the common scenario where the sender has to choose a cover-medium in a huge collection of documents, she can drop the cover-medium that leads to a failure and choose another one, iterating the process until finding a cover-medium that is adequate to embed the message. Another solution may be to cut the message into smaller pieces, in order to have shorter messages to embed, and a lower probability of failure. If none of these is possible, for example if the sender only has few pieces of content, she may unlock some locked components [13] to make the probability of failure decrease. But, even doing this modified embedding, and decreasing the

probability of failure, the sender will not be able to drop it to zero, except if she falls back to perfect codes without locked components.

In this paper, we consider the "worst case" scenario, where the sender does not have too much cover documents to hide his message in, and then absolutely needs embedding to succeed. This scenario is not the most studied one, and concerns very constrained situations. Our contribution is to propose an embedding scheme that will never fail, and does not relax the management of locked components of his cover-data to make embedding succeed. It is, to our knowledge, the first bounded syndrome coding scheme that manages locked components while guaranteeing the complete embedding of the message for any code, be it perfect or not. To do so, we modify the classical syndrome coding approach by using some part of the syndrome for randomization. Of course, as the message we can embed is now shorter than the syndrome, there is a loss in terms of embedding efficiency. We analyze this loss in the case of linear perfect codes. Moreover, inspired by the ZZW construction [39], we show how the size of the random part of the syndrome, which is dynamically estimated during embedding, can be transmitted to the recipient without any additional communication.

The paper is organized as follows. Basic definitions and notation on both steganography and syndrome coding are introduced in Section 2. The traditional syndrome coding approach is recalled at the end of this section. In Section 3, we show how to slightly relax the constraints on the linear system to make it always solvable, and also estimate the loss of embedding efficiency. We discuss the behavior of our scheme in the case of the Golay and Hamming perfect codes in Section 4. Finally, as our solution uses a parameter r that is dynamically computed during embedding, we provide in Section 5 a construction that enables to transmit r to the recipient through the stego-data itself, that is, without any parallel or side-channel communication. We finally conclude in Section 6.

2 Steganography and coding theory

2.1 Steganographic schemes

We define a *stego-system* (or a *steganographic scheme*) by a pair of functions, Emb and Ext. Emb embeds the message $\mathbf{m}$ in the cover-data $\mathbf{x}$, producing the stego-data $\mathbf{y}$, while Ext extracts the message $\mathbf{m}$ from the stego-data $\mathbf{y}$. To make the embedding and extraction work properly, these functions have to satisfy the following properties.

Definition 1 (Stego-System). *Let $\mathcal{A}$ a finite alphabet, $r, n \in \mathbb{N}$ such that $r < n$, $\mathbf{x} \in \mathcal{A}^n$ denote the* cover-data*, $\mathbf{m} \in \mathcal{A}^r$ denote the message to embed, and T be a strictly positive integer. A* stego-system *is defined by a pair of functions Ext and Emb such that:*

$$Ext(Emb(\mathbf{x}, \mathbf{m})) = \mathbf{m} \tag{1}$$

$$d(\mathbf{x}, Emb(\mathbf{x}, \mathbf{m})) \leq T \tag{2}$$

where $d(.,.)$ denoting the Hamming distance over $\mathcal{A}^n$.

Two quantities are usually used to compare stego-systems: the embedding efficiency and the relative payload, which are defined as follows.

Definition 2 (Embedding efficiency). *The* average embedding efficiency *of a stego-system, is usually defined by the ratio of the number of message symbols we can embed by the average number of symbols changed. We denote it by e.*

Definition 3 (Relative payload). *The* relative payload *of a stego-system, denoted by α, is the ratio of the number of message symbols we can embed by the number of (modifiable) symbols of covered data.*

For q-ary syndrome coding, the sphere-covering bound gives an upper bound for the embedding efficiency [16]. Note that it is usually stated for binary case, using the binary entropy function.

Proposition 1 (Sphere-covering bound). *For any q-ary stego-system $\mathcal{S}$, the* sphere-covering bound *gives*

$$e \leq \frac{\alpha}{\mathcal{H}_q^{-1}(\alpha)},$$

where $\mathcal{H}_q^{-1}()$ denotes the inverse function of the q-ary entropy $\mathcal{H}_q(x) = x\log_q(q-1) - x\log_q(x) - (1-x)\log_q(1-x)$ on $[0, 1-1/q]$, and α is the relative payload associated with $\mathcal{S}$.

2.2 From coding theory to steganography

This section recalls how coding theory may help embedding the message, and how it tackles the non-shared selection channel paradigm. In the rest of paper, the finite alphabet $\mathcal{A}$ is a finite field of cardinal q, denoted $\mathbb{F}_q$.

Here we focus on the use of linear codes, which is the most studied. Let $\mathcal{C}$ be a $[n, k, d]_q$-linear code, with parity check matrix H and covering

radius ρ — it is the smallest integer such that the balls of radius ρ centered on $\mathcal{C}$'s codewords cover the whole ambient space $\mathbb{F}_q^n$. A syndrome coding scheme based on $\mathcal{C}$ basically modify the cover-data $\mathbf{x}$ in such a way that the syndrome $\mathbf{y}H^t$ of the stego-data $\mathbf{y}$ will precisely be equal to the message $\mathbf{m}$. Determining which symbols of $\mathbf{x}$ to modify leads to finding a solution of a particular linear system that involves the parity check matrix H. This embedding approach has been introduced by Crandall in 1998 [9], and is called *syndrome coding* or *matrix embedding*.

We formulate several embedding problems. The first one addresses only Eq. (1) requirements, whereas the second one also tackles Eq. (2).

Problem 1 (Syndrome coding problem). Let $\mathcal{C}$ be an $[n, k, d]_q$ linear code, H be a parity check matrix of $\mathcal{C}$, $\mathbf{x} \in \mathbb{F}_q^n$ be a cover-data, and $\mathbf{m} \in \mathbb{F}_q^{n-k}$ be the message to be hidden in $\mathbf{x}$. The *syndrome coding problem* consists in finding $\mathbf{y} \in \mathbb{F}_q^n$ such that $\mathbf{y}H^t = \mathbf{m}$.

Problem 2 (Bounded syndrome coding problem). Let $\mathcal{C}$ be an $[n, k, d]_q$ linear code, H be a parity check matrix of $\mathcal{C}$, $\mathbf{x} \in \mathbb{F}_q^n$ be a cover-data, $\mathbf{m} \in \mathbb{F}_q^{n-k}$ be the message to be hidden in $\mathbf{x}$, and $T \in \mathbb{N}^*$ be an upper bound on the number of authorized modifications. The *bounded syndrome coding problem* consists in finding $\mathbf{y} \in \mathbb{F}_q^n$ such that $\mathbf{y}H^t = \mathbf{m}$, and $d(\mathbf{x}, \mathbf{y}) \leq T$.

Let us first focus on Problem 1, which leads to describing the stego-system in terms of syndrome computation:

$$\begin{aligned} \mathbf{y} = Emb(\mathbf{x}, \mathbf{m}) &= \mathbf{x} + D(\mathbf{m} - \mathbf{x}H^t), \\ Ext(\mathbf{y}) &= \mathbf{y}H^t, \end{aligned}$$

where D is the mapping associating to a syndrome $\mathbf{m}$, a vector whose syndrome is precisely equal to $\mathbf{m}$. The mapping D is thus directly linked to a decoding function $f_{\mathcal{C}}$ of $\mathcal{C}$ of arbitrary radius T_f, defined as $f_{\mathcal{C}} : \mathbb{F}_q^n \longrightarrow \mathcal{C} \cup \{?\}$, such that for all $\mathbf{y} \in \mathbb{F}_q^n$, either $f_{\mathcal{C}}(\mathbf{y}) =?$, or $d(\mathbf{y}, f_{\mathcal{C}}(\mathbf{y})) \leq T_f$.

The Hamming distance between vectors $\mathbf{x}$ and $\mathbf{y}$ is then less than or equal to T_f. Since decoding general codes is NP-Hard [2], finding such a mapping D is not tractable if $\mathcal{C}$ does not belong to a family of codes we can efficiently decode. Moreover, to be sure that the Problem 2 always has a solution, it is necessary and sufficient that $f_{\mathcal{C}}$ can decode up to the covering radius of $\mathcal{C}$. This means that solving Problem 2 with $T = \rho$ is precisely equivalent to designing a stego-system which find solutions to both Eqs. (1) and (2) requirements for any $\mathbf{x}$ and $\mathbf{m}$. In this context,

perfect codes, for which the covering radius is precisely equal to the error-correcting capacity ($\rho = \lfloor \frac{d-1}{2} \rfloor$), are particularly relevant.

Unfortunately, using perfect codes leads to an embedding efficiency which is far from the bound given in Prop. 1 [4]. Hence non-perfect codes have been studied (see the Introduction), even if they can only tackle Problem 2 for some T much lower than ρ. This may enable to force the system to perform only a small number of modifications.

As discussed in the introduction, *Wet paper* codes were introduced to improve embedding undetectability through the management of locked, or *wet*, components [18].

Problem 3 (Bounded syndrome wet paper coding problem). Let $\mathcal{C}$ be an $[n, k, d]_q$ linear code, H be a parity check matrix of $\mathcal{C}$, $\mathbf{x} \in \mathbb{F}_q^n$, $\mathbf{m} \in \mathbb{F}_q^{n-k}$, $T \in \mathbb{N}^*$, and a set of locked, or wet, components $\mathcal{I} \subset \{1, \ldots, n\}$, $\ell = |\mathcal{I}|$. The *Bounded syndrome wet paper coding problem* consists in finding $\mathbf{y} \in \mathbb{F}_q^n$ such that $\mathbf{y}H^t = \mathbf{m}$, $d(\mathbf{x}, \mathbf{y}) \leq T$, and $\mathbf{x}_i = \mathbf{y}_i$ for all $i \in \mathcal{I}$.

Of course, solving Problem 3 is harder and even perfect codes may fail here. More precisely, to deal with locked components, we usually decompose the parity check matrix H of $\mathcal{C}$ in the following way [18,19]:

$$\begin{aligned} \mathbf{y}H^t &= \mathbf{m}, \\ \mathbf{y}_{|\bar{\mathcal{I}}} H^t_{|\bar{\mathcal{I}}} + \mathbf{y}_{|\mathcal{I}} H^t_{|\mathcal{I}} &= \mathbf{m}, \\ \mathbf{y}_{|\bar{\mathcal{I}}} H^t_{|\bar{\mathcal{I}}} &= \mathbf{m} - \mathbf{y}_{|\mathcal{I}} H^t_{|\mathcal{I}}, \end{aligned}$$

where $\bar{\mathcal{I}} = \{1, \ldots, n\} \setminus \mathcal{I}$. The previous equation can only be solved if $rank(H_{\bar{\mathcal{I}}}) = n - k$. Since the potential structure of H does not help to solve the previous problem, we could as well choose H to be also a random matrix, which provides the main advantage to maximize asymptotically the average embedding efficiency [22,19].

Hiding a long message requires to split it and to repeatedly use the basic scheme. Let P_H the success probability for embedding $(n-k)$ symbols, then the global success probability P for a long message of length $L(n-k)$ is P_H^L. This probability decreases exponentially with the message length.

In order to bypass this issue, previous works propose either to take another cover-medium, or to modify some locked components. In this paper, we still keep unmodified the locked components, thus maintaining the same level of undetectability. Moreover, we tackle the particular case where the sender does not have a lot of cover-media available, and needs a successful embedding, even if this leads to a smaller embedding efficiency.

In the original Wet Paper Setting of [18], the embedding efficiency is not dealt with. In that case, we have a much easier problem.

Problem 4 (Unbounded wet paper Syndrome coding problem). Let $\mathcal{C}$ be an $[n,k,d]_q$ linear code, H be a parity check matrix of $\mathcal{C}$, $\mathbf{x} \in \mathbb{F}_q^n$, $\mathbf{m} \in \mathbb{F}_q^{n-k}$, and a set of locked components $\mathcal{I} \subset \{1,\ldots,n\}$, $\ell = |\mathcal{I}|$. The *Unbounded wet paper Syndrome coding problem* consists in finding $\mathbf{y} \in \mathbb{F}_q^n$ such that $\mathbf{y}H^t = \mathbf{m}$, and $\mathbf{x}_i = \mathbf{y}_i$, for all $i \in \mathcal{I}$.

In a random case setting, this problem can be discussed using a lower bound on random matrices, provided by [5].

Theorem 1. *Let M be a random $ncol \times nrow$ matrix defined over $\mathbb{F}_q$, such that $ncol \geq nrow$. We have:*

$$P\left(rank(M) = nrow\right) \geq \begin{cases} 0.288, & \text{if } ncol = nrow \text{ and } q = 2, \\ 1 - \frac{1}{q^{ncol-nrow}(q-1)}, & \text{otherwise.} \end{cases}$$

In a worst-case, or infallible, setting, the relevant parameter of the code is its *dual distance*.

Proposition 2. *Consider a q-ary wet channel on length n with at most ℓ wet positions, and that there exists a q-ary code C whose dual code $C^\perp$ has parameters $[n, k^\perp, d^\perp = \ell]_q$ with $k^\perp + d^\perp = n + 1 - g$. Then we can surely embed $n - \ell - g$ symbols using a parity check matrix of C.*

Proof. This can be derived from [26, Theorem 2.3].

This means that if the code is g far from the Singleton bound, then we loose g information symbols with respect to the maximum. In particular, if $n < q$, there exists a q-ary Reed-Solomon code with $g = 0$, and we can always embed $n - \ell$ symbols when there are ℓ wet symbols. Coding theory bounds tells us that the higher q, the smallest g can be achieved, eventually using Algebraic-Geometry codes [35].

3 Randomized (wet paper) syndrome coding

Since embedding a message has a non-zero probability to fail, we propose to relax the constraints in the following way:

Problem 5 (Randomized bounded syndrome coding problem for wet paper). Let $\mathcal{C}$ be an $[n,k,d]_q$ linear code, H be a parity check matrix of $\mathcal{C}$, r and T be two integers, $\mathbf{x} \in \mathbb{F}_q^n$, $\mathbf{m} \in \mathbb{F}_q^{n-k-r}$ be the message to embed, and

$\mathcal{I} \subset \{1, \ldots, n\}$ be the set of locked components, $\ell = |\mathcal{I}|$. Our *randomized syndrome coding problem for wet paper* consists in finding $\mathbf{y} \in \mathbb{F}_q^n$ and $\mathbf{R} \in \mathbb{F}_q^r$ such that (i) $\mathbf{y}H^t = (\mathbf{m}||\mathbf{R})$, and $||$ denotes the concatenation operator, (ii) $d(\mathbf{x}, \mathbf{y}) \leq T$, and (iii) $\mathbf{x}_i = \mathbf{y}_i$, for all $i \in \mathcal{I}$.

We thus randomize one fraction of the syndrome to increase the number of solutions. This gives a degree of freedom which may be large enough to solve the system. The traditional approach can then be applied to find $\mathbf{y}_{|\bar{\mathcal{I}}}$ and consequently $\mathbf{y}$. Using some random symbols in the syndrome was used in the signature scheme of Courtois, Finiasz and Sendrier [8]. While this reformulation allows to solve the bounded syndrome coding problem in the wet paper context without failure, we obviously lose some efficiency compared to the traditional approach.

We now estimate the loss in embedding efficiency for a given number of locked components. Let e denote the embedding efficiency of the traditional approach, and e' denote the efficiency of the randomized one. We obtain a relative loss of:

$$\frac{e - e'}{e} = \frac{r}{n-k},$$

while being assured that any $n - k - r$ message be embedded, as long as $r < n - k$.

Optimizing the parameter r is crucial, to ensure that our reformulated problem always has a solution, while preserving the best possible embedding efficiency. This is the goal on next Section.

4 Case of perfect linear codes

We discuss in this Section a sufficient condition on the size r of randomization, for our reformulated problem to always have a solution.

4.1 General Statement

The *syndrome function* associated with H, noted S_H, is defined by:

$$\begin{aligned} S_H : \mathbb{F}_q^n &\longrightarrow \mathbb{F}_q^{n-k} \\ \mathbf{x} &\longmapsto \mathbf{x}H^t. \end{aligned}$$

This function S_H is linear and surjective, and satisfies the following well-known properties. Let $\mathcal{B}(\mathbf{x}, T)$ denote the Hamming ball of radius T centered on $\mathbf{x}$.

Proposition 3. *Let $\mathcal{C}$ be an $[n,k,d]_q$-linear code, with covering radius ρ, H a parity check matrix of $\mathcal{C}$, and S_H the syndrome function associated with H. For all $\mathbf{x} \in \mathbb{F}_q^n$, the function S_H restricted to $\mathcal{B}(\mathbf{x}, \lfloor \frac{d-1}{2} \rfloor)$ is one-to-one, the function S_H restricted to $\mathcal{B}(\mathbf{x}, \rho)$ is surjective. When $\mathcal{C}$ is perfect, the syndrome function restricted to $\mathcal{B}(\mathbf{x}, \rho)$ is bijective.*

Now, we give a sufficient condition for upper-bounding r in Problem 5.

Proposition 4. *Given a $[n,k,d]$ perfect code with $\rho \frac{d-1}{2}$, if the inequality*

$$q^{n-k} + 1 \leq q^r + \sum_{i=0}^{\rho} (q-1)^i \binom{n-\ell}{i}, \tag{3}$$

is satisfied, then there exists a vector $\mathbf{y} \in \mathbb{F}_q^n$ and a random vector $\mathbf{R}$, which are solution of Problem 5. In this case, Problem 5 always has a solution $\mathbf{y}$.

Proof. Let N_1 —respectively N_2— be the number of different syndromes generated by the subset of $\mathbb{F}_q^n$ satisfying (i) of Problem 5 — respectively (ii) and (iii). If

$$N_1 + N_2 > q^{n-k}. \tag{4}$$

Then there exists $\mathbf{y}$ which fulfills conditions (i), (ii), and (iii). The number of different syndromes satisfying by the first constraint, for all $\mathbf{R}$, is q^r. Keeping in mind that ℓ components are locked and the syndrome function restricted to $\mathcal{B}(\mathbf{x}, \rho)$ is bijective, then

$$N_2 = \sum_{i=0}^{\rho} (q-1)^i \binom{n-\ell}{i}.$$

Combined with the sufficient condition (4) we obtain the result.

Next Section is devoted to the non trivial perfect codes: the Golay codes, and the (q-ary) the Hamming codes.

4.2 Golay codes

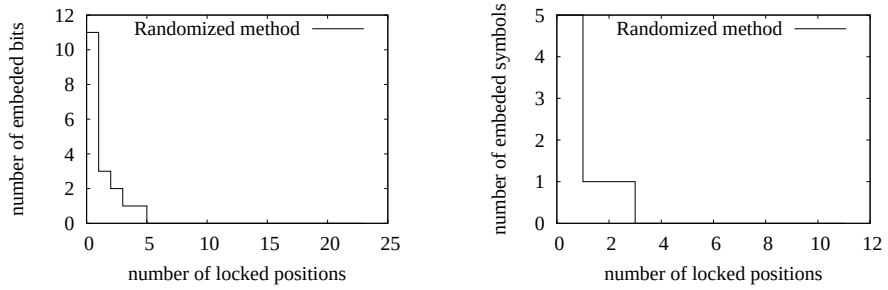
Binary Golay code We start by study the case of the binary $[23,12,7]_2$ Golay code, which is perfect. The inequality of the proposition 4 gives

$$r \geq \log_2 \left(1 + \frac{796}{3}\ell - \frac{23}{2}\ell^2 + \frac{1}{6}\ell^3 \right). \tag{5}$$

Ternary perfect Golay code The ternary Golay code has parameters $[11, 6, 5]_3$. Using the Proposition 4, we obtain:

$$r \geq \log_3 \left(1 + 44\ell - 2\ell^2\right). \tag{6}$$

Eqs 5 and 6 does not say much. We have plotted the results in Fig. 1, and we see that the number of available bits for embedding degrades very fast with the number of locked positions.



(a) Binary Golay code. (b) Ternary Golay code.

Fig. 1. Size of the random part for the two Golay codes. The number of remaining bits is plotted, in terms of the number of locked positions.

4.3 Hamming codes

We study the infinite family of Hamming codes. We find r, analyze the found parameters, and study its asymptotic behavior.

Computation of r Let $\mathcal{C}$ be a $[(q^p-1)/(q-1), n-p, 3]_q$ Hamming code over $\mathbb{F}_q$, for some p. Its covering radius is $\rho = 1$, and thus its embedding efficiency if p. We aim to minimize r, the length of the random vector $\mathbf{R}$. Since $q^{n-k} = q^p$, $(q^p - 1)/(q - 1) = n$, Proposition 4 gives:

$$r \geq \log_q \left(1 + (q-1)\ell\right). \tag{7}$$

Analysis of parameters In order to find an extreme case, it we maximize the number of locked components ℓ while still keeping $n - k - r \geq 1$.

A direct computation gives:

$$p-1 = \log_q((q-1)\ell+1),$$
$$\ell = \frac{q^{p-1}-1}{q-1} \approx \frac{n}{q}.$$

Therefore, using Hamming codes, we can embed at least one information symbol if no more than a fraction of $\frac{1}{q}$ of the components are locked. This is of course best for $q=2$. The minimum r which satisfies inequality (7) is $r = \lceil \log_q((q-1)\ell+1) \rceil$. In other words, for Hamming codes, the minimum number of randomized symbols needed to guarantee that the whole message can be embedded, is logarithmic in the number of locked components. Our randomized approach always solves successfully Problem 5 while traditional syndrome coding (including wet paper) exhibits a non-zero failure rate, when $\frac{\ell}{n} < \frac{1}{q}$.

Asymptotic behavior Now we evaluate the loss in embedding efficiency. Then, for a given ℓ, the relative loss of the embedding efficiency is given by:

$$\frac{\lceil \log_q((q-1)\ell+1) \rceil}{p}.$$

To conclude this section, we propose to focus on the normalized loss in symbols for the family of Hamming codes. We assume that the rate of ℓ, the number of locked components to compare to n, the length of the cover-data stays constant, i.e. $\ell = \lambda n$, for a given $\lambda \in [0, \frac{1}{q}($. Then the asymptotic of relative loss is

$$\frac{\log_q((q-1)\ell+1)}{p} \sim \frac{\log_q(n(q-1)\lambda)}{p} \sim 1 + \frac{\log_q \lambda}{p}.$$

This goes to 1 when p goes to infinity, i.e. all the symbols of syndrome are consumed by the randomization. It makes sense, since dealing with a given proportion λ of arbitrarily locked symbols in a long stego-data is much harder than dealing with several smaller stego-data with the same proportion λ of locked positions.

5 Using ZZW construction to embed dynamic parameters

In the approach given in previous Section, the sender and recipient have to fix in advance the value of r. Indeed the recipient has to know which

part of syndrome is random. This is not very compliant with the Wet Paper model, where the recipient does not know the quantity of wet bits. We propose in this Section a variant of ZZW's scheme [39], which enables to convey dynamically the value r, depending on the cover-data.

5.1 The scheme

We consider that we are treating n blocks of $2^p - 1$ bits, $\mathbf{x}_1, \ldots, \mathbf{x}_n$, for instance displayed as in Figure 2. Each block $\mathbf{x}_i$ is a binary vector of length $2^p - 1$, set as column, and we let $\mathbf{v} = (v_1, \ldots, v_n)$ be the binary vector whose i-th coordinate v_i is the parity bit of column $\mathbf{x}_i$. We use the (virtual) vector $\mathbf{v}$ to convey extra information, while at the same time the $\mathbf{x}_i$ are using for syndrome coding.

Our scheme is threefold : syndrome coding on the $\mathbf{x}_i$'s using the parity check H_1 of a first Hamming code, with our randomized method, then (unbounded wet paper) syndrome embedding on the syndromes $\mathbf{s}_i$'s of the $\mathbf{x}_i$'s. This second syndrome embedding see the $\mathbf{s}_i$ as q-ary symbols, and the matrix in use is the parity check matrix H_q of a q-ary Reed-Solomon code. We call the n first embeddings the H_1-embeddings, and the second one the H_q-embedding. Finally, we use $\mathbf{v}$ to embed dynamic information: the number r of random bits, and f the number of failure in the H_1-embeddings. We call this last embedding the H_2-embedding, where H_2 is the parity check matrix of a second, much shorter, binary Hamming code.

We assume that r is bounded by design, say $r \leq r_{\max}$. We shall see, after a discussion on all the parameters, that this is one design parameter of the scheme, together with o, which the precision, in bits, for describing real numbers $\in]\frac{1}{2}, 1]$.

Embedding

Inspect. Each column $\mathbf{x}_1, \ldots, \mathbf{x}_n$ is inspected, to find the number of dry bits in each. This enables to determine the size r of the randomized part, which shall be the same for all columns. This determines the columns $\mathbf{x}_i$'s where the H_1-embeddings are feasible. Let f be the number of $\mathbf{x}_i$'s where the H_1-embeddings fail.

Build the wet channel. For each of the $n - f$ columns $\mathbf{x}_i$'s where the H_1-embedding is possible, there is a syndrome s_i of p bits, where the last r bits are random, thus wet, and the $p - r$ first bits are dry. We consider these blocks of $p - r$ bits as a q-ary symbols, with $q = 2^{p-r}$. Thus we have a q-ary wet channel with $n - f$ dry q-ary symbols, and f wet q-ary symbols

Embed for the wet channel. Then, using a Reed-Solomon over the alphabet $\mathbb{F}_q$, we can embed $(n-f)$ q-ary symbols, using a $n \times (n-f)$ q-ary parity check matrix H_q of the code. Note that the number of rows of this matrix is dynamic since f is dynamic.

Embed dynamic data. We have to embed dynamic parameters r and f which are unknown to the recipient, using ZZW's virtual vector $\mathbf{v}$. For this binary channel, the dry bits v_i correspond to the columns $\mathbf{x}_i$ where the H_1-embedding has failed, and where there is at least one dry bit in $\mathbf{x}_i$. A second Hamming code is used with parity check H_2 for this embedding.

Recovery

H_2-extraction. First compute $\mathbf{v}$, and using the parity check matrix of the Hamming code H_2, extract r and f.

H_1-extraction Extract the syndromes of all the column $\mathbf{x}_i$'s using the parity check matrix H_1, and collect only the first $p-r$ bits in each column, to build q-ary symbols.

H_q-extraction Build the parity check matrix H_q of the q-ary $[n, f]_q$ Reed-Solomon code, with $q = 2^{p-r}$. Using this matrix, get the $(n-f)$ q-ary information symbols, which are the actual payload.

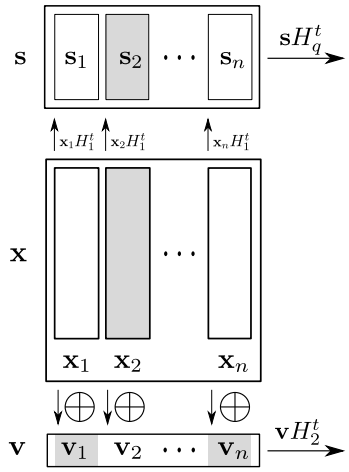


Fig. 2. A graphical view of our scheme inspired from ZZW. A syndrome $\mathbf{s}_i$ is considered wet for the H_q-embedding when the H_1-embedding is not feasible. Then the corresponding bit v_i in the vector $\mathbf{v}$ is dry for the H_2-embedding. Wet data is grey on the Figure.

5.2 Analysis

There are several constraints on the scheme.

First, for a Reed-Solomon code of length n to exist over the alphabet $\mathbb{F}_{2^{p-r}}$, we must have $n \leq 2^{p-r}$, for any r, i.e. $n \leq 2^{p-r_{\max}}$. We fix $n = 2^{p-r_{\max}} - 1$, and let us briefly denote $u = p - r_{\max}$.

Then the binary $[n = 2^u - 1, 2^u - u - 1]_2$ Hamming code, with parity check matrix H_2, is used for embedding in the vector $\mathbf{v}$, with f dry symbols. This a unbounded wet channel. From Proposition 2, we must have

$$f \geq 2^{u-1}, \tag{8}$$

which implies that some columns $\mathbf{x}_i$ may be artificially declared wet, for satisfying Eq. 8. Third, we also must have

$$u = \lceil \log r_{\max} \rceil + \lceil \log f_{\max} \rceil, \tag{9}$$

to be able to embed r and f. Since $f \leq 2^u - 1$, we have $\lceil \log f_{\max} \rceil = u$. Eq. 9 becomes $u = \lceil \log r_{\max} \rceil + u$, this is clearly not feasible. To remedy this, instead of embedding f, we embed its relative value $f_u = \frac{f}{2^u} \in [.5, 1]$, up to a fixed precision, say o bits, with o small. Then Eq. 9 is replaced by

$$u = \lceil \log r_{\max} \rceil + o, \tag{10}$$

$$p = r_{\max} + \lceil \log r_{\max} \rceil + o, \tag{11}$$

which is a condition easy to fulfill. It is also possible, by design, to use the all-one value of f_u as an out-of-range value to declare an embedding failure. The scheme is locally adaptive to the media: for instance, in a given image, r and f may take different values for different areas of the image.

In conclusion, the number of bits that we can embed using that scheme is bounded by $(n - f)(p - r) \leq 2^{u-1}(p - r)$, with dynamic r and f.

6 Conclusion

In this paper, we addressed the "worst-case" scenario, where the sender cannot accept embedding to fail, and does not want relax the management of locked components of his cover-data. As traditional (wet) syndrome coding may fail, and as the failure probability increases exponentially with the message length, we proposed here a different approach, which

never fails. Our solution is based on the randomization of a part of the syndrome, the other part still carrying symbols of the message to transmit. While our method suffers from a lost of embedding efficiency, we showed that this loss remains acceptable for perfect codes. Moreover, we showed how the size of the random part of the syndrome, which is dynamically estimated during embedding, may be transmitted to the recipient without any additional communication.

References

1. Anderson, R., Petitcolas, F.: On the limits of steganography. IEEE Journal on Selected Areas in Communications 16(4), 474 –481 (May 1998)
2. Berlekamp, E., McEliece, R., Van Tilborg, H.: On the inherent intractability of certain coding problems. IEEE Trans. on Information Theory 24(3), 384 – 386 (May 1978)
3. Bierbrauer, J.: On Crandall's problem. Personal communication (2001), `http://www.ws.binghamton.edu/fridrich/covcodes.pdf`
4. Bierbrauer, J., Fridrich, J.: Constructing good covering codes for applications in steganography. In: Shi, Y.Q. (ed.) Transactions on data hiding and multimedia security III. pp. 1–22. Springer Berlin / Heidelberg, Berlin, Heidelberg (2008)
5. Brent, R.P., Gao, S., Lauder, A.G.B.: Random Krylov spaces over finite fields. SIAM J. Discrete Math 16, 276–287 (2001)
6. Cachin, C.: An information-theoretic model for steganography. In: Information Hiding, 2nd International Workshop - IH 1998. Lecture Notes in Computer Science, vol. 1525, pp. 306–318. Springer-Verlag (1998)
7. Cachin, C.: An information-theoretic model for steganography. Information and Computation 192(1), 41–56 (2004)
8. Courtois, N., Finiasz, M., Sendrier, N.: How to achieve a McEliece-based digital signature scheme. In: Boyd, C. (ed.) Advances in Cryptology ASIACRYPT 2001, Lecture Notes in Computer Science, vol. 2248, pp. 157–174. Springer Berlin / Heidelberg (2001)
9. Crandall, R.: Some notes on steganography (1998), `http://os.inf.tu-dresden.de/~westfeld/crandall.pdf`, posted on the steganography mailing list.
10. Filler, T., Fridrich, J.: Wet ZZW construction for steganography. In: IEEE International Workshop on Information Forensics and Security - WIFS 2009. pp. 131–135 (2009)
11. Filler, T., Fridrich, J.: Minimizing additive distortion functions with non-binary embedding operation in steganography. In: IEEE International Workshop on Information Forensics and Security - WIFS 2010 (2010)
12. Filler, T., Judas, J., Fridrich, J.: Minimizing additive distortion in steganography unsing syndrome-trellis codes. IEEE Trans. on Information Forensics and Security (2011)
13. Filler, T., Judas, J., Fridrich, J.: Minimizing embedding impact in steganography using trellis-coded quantization. In: IS&T/SPIE International Symposium on Electronic Imaging 2010 - Media Forensics and Security II. Proceedings of the SPIE, vol. 7541. SPIE (2010)

14. Fontaine, C., Galand, F.: How can Reed-Solomon codes improve steganographic schemes. In: 9th Information Hiding – IH'07. Lecture Notes in Computer Science, vol. 4567, pp. 130–144. Springer-Verlag (2007)
15. Fontaine, C., Galand, F.: How Reed-Solomon codes can improve steganographic schemes. EURASIP J. Inf. Secur. 2009, 1–10 (2009)
16. Fridrich, J.: Asymptotic behavior of the ZZW embedding construction. IEEE Transactions on Information Forensics and Security 4(1), 151–153 (2009)
17. Fridrich, J., Filler, T.: Practical methods for minimizing embedding impact in steganography. In: IS&T/SPIE International Symposium on Electronic Imaging 2007 - Security, Steganography, and Watermarking of Multimedia Contents IX. Proceedings of the SPIE, vol. 6505. SPIE (2007)
18. Fridrich, J., Goljan, M., Lisonek, P., Soukal, D.: Writing on wet paper. IEEE Trans. on Signal Processing 53(10), 3923 – 3935 (October 2005)
19. Fridrich, J., Goljan, M., Soukal, D.: Wet paper codes with improved embedding efficiency. IEEE Trans. on Information Forensics and Security 1(1), 102 – 110 (March 2006)
20. Fridrich, J., Soukal, D.: Matrix embedding for large payloads. IEEE Trans. on Information Forensics and Security 1(3), 390 –395 (Sep 2006)
21. Fridrich, J.J., Goljan, M., Soukal, D.: Efficient wet paper codes. In: Information Hiding. pp. 204–218 (2005)
22. Galand, F., Kabatiansky, G.: Information hiding by coverings. In: Proc. ITW 2003. pp. 151–154 (2003)
23. Galand, F., Kabatiansky, G.: Coverings, centered codes, and combinatorial steganography. Problems of Information Transmission 45(3), 289–297 (2009)
24. Kodovský, J., Fridrich, J., Pevný, T.: Statistically undetectable jpeg steganography: Dead ends, challenges, and opportunities. In: Proc. of the ACM Multimedia and Security Workshop 2007. pp. 3–14. ACM (2007)
25. McLoughlin, A.: The complexity of computing the covering radius of a code. IEEE Trans. on Information Theory 30(6), 800–804 (1984)
26. Munuera, C., Barbier, M.: Wet paper codes and the dual distance in steganography. Advances in Mathematics of Communications (2011), to be published
27. Ould Medeni, M., Souidi, E.M.: A steganography schema and error-correcting codes. Journal of Theoretical and Applied Information Technology 18(1), 42–47 (2010)
28. Rifà, J., Ronquillo, L.: Product perfect Z2Z4-linear codes in steganography. In: International Symposium on Information Theory and its Applications - ISITA 2010 (2010)
29. Rifà-Pous, H., Rifà, J.: Product perfect codes and steganography. Digital Signal Processing 19(4), 764–769 (2009)
30. Sachnev, V., Kim, H., Zhang, R.: Less detectable jpeg steganography method based on heuristic optimization and BCH syndrom coding. In: ACM Multimedia & Security'09. pp. 131–139. ACM Press (2009)
31. Schönfeld, D., Winkler, A.: Embedding with syndrome coding based on BCH codes. In: Proceedings of the 8th workshop on Multimedia and security. pp. 214–223. ACM (2006)
32. Schönfeld, D., Winkler, A.: Reducing the complexity of syndrome coding for embedding. In: Proc. of the 10th International Worksop on Information Hiding. Lecture Notes in Computer Science, vol. 4567, pp. 145–158. Springer-Verlag (2007)
33. Simmons, G.: The prisoners' problem and the subliminal channel. In: Advances in Cryptology – CRYPTO'83. pp. 51–67. Plenum Press (1984)

34. Vardy, A.: The intractability of computing the minimum distance of a code. IEEE Trans. on Information Theory 43(6), 1757–1766 (1997)
35. Vladut, S., Nogin, D., Tsfasman, M.: Algebraic Geometric Codes: Basic Notions (Mathematical Surveys and Monographs). American Mathematical Society (September 2007)
36. Westfeld, A.: F5 - A steganographic algorithm. In: Moskowitz, I. (ed.) Information Hiding, Lecture Notes in Computer Science, vol. 2137, pp. 289–302. Springer Berlin / Heidelberg (2001)
37. Zhang, R., Sachnev, V., Kim, H.: Fast BCH syndrome coding for steganography. In: Katzenbeisser, S., Sadeghi, A.R. (eds.) Information Hiding. Lecture Notes in Computer Science, vol. 5806, pp. 48–58. Springer-Verlag (2009)
38. Zhang, W., Zhang, X., Wang, S.: Near-optimal codes for information embedding in gray-scale signals. IEEE Trans. on Information Theory 56(3), 1262–1270 (2010)
39. Zhang, W., Zhang, X., Wang, S.: Maximizing steganographic embedding efficiency by combining Hamming codes and wet paper codes. In: Proc. of the 10th International Worksop on Information Hiding. Lecture Notes in Computer Science, vol. 5284, pp. 60–71. Springer-Verlag (2008)