

Practical Construction Against Theoretical Approach in Fingerprinting*

Fabien Galand
INRIA Rocquencourt - Projet CODES
BP 105
78153 Le Chesnay Cedex, France
Email: fabien.galand@inria.fr

Abstract

We consider fingerprinting under collusion attacks in the Hamming space. To model the attacks we use the framework of Somekh-Baruch and Merhav, *IEEE Trans. on Inf. Th.*, vol. 51(3), 2005. We construct a family of fingerprinting codes efficient against coalition of arbitrary size. Using this family, tracing dishonest users can be done without error and in polynomial time. The number of users is exponential in the length. The proposed construction relies on centered error correcting codes (Bassalygo and Pinsker, *Prob. of Inf. Trans.*, vol. 35, 1999). Our results have an amazing relation with an upper bound on the number of users derived by Somekh-Baruch and Merhav : dropping two assumptions that they used we construct codes beating their bound.

1 Introduction

Fingerprinting is used to prevent an illicit distribution of copyright protected data. More precisely, it allows to recover some legal users who participate in the illicit distribution.

To do so, for each legal user a particular copy of the data is created, by adding some side-information. Then, if some set of users, called a coalition, try to create an illicit copy, we want to be able to trace at least one of the members whenever the coalition has a size not more than some integer L .

*in Proceedings of IEEE International Symposium on Information Theory 2006, IEEE, pp. 2603 – 2606.

Now, the crucial point is “How does the members of a coalition create its illicit copy from their fingerprinted copies?”. Two important models have been used. The first one follows the *marking assumption* (see [1, 2] for discussion of this assumption): the forgery can be obtained from a set of fingerprinted copies only by changing parts in which at least two are different. In the second one [3], the only restriction is to keep the forgery close to at least one of the fingerprinted copies. In the sequel, we will address this latter model.

Among important problems studied for fingerprinting codes, there is the maximum number of users for a fixed coalition size L and data size n . A bound derived from information theory [3] under some conditions states the logarithm of this maximum scales with $O(1/L)$ and goes to zero if L grows linearly with n .

In this paper, we construct a family of binary fingerprinting codes with exponentially many users. Moreover, our fingerprinting codes have a very interesting property : they are effective against coalitions of arbitrary size. This implies that our codes beat the bound of [3]. Our fingerprinting codes also have an efficient polynomial time tracing algorithm. This algorithm only needs the forgery.

In section 2 we present the framework and our notations. Section 3 gives basic facts about the centered error correction codes. Section 4 explains how to use centered error correction codes to construct fingerprinting codes and we detail the achieved parameters in section 5. Section 6 briefly compares our results and results on the capacity of some fingerprinting systems obtained in [3]. Finally, we conclude pointing at an important difference between the model we use here and the marking assumption.

2 Formal problem

We have a set $\mathcal{V} \subset \mathbb{F}^n$ of original data, which can be seen as sequences of length n over some alphabet \mathbb{F} . In order to fingerprint some copyright protected data v , we allow to change at most Δ_o coordinates, that is a fingerprinted copy of v is in $B_{\Delta_o}(v) = \{y : d(v, y) \leq \Delta_o\}$ — d is the Hamming distance over \mathbb{F}^n —, the ball of center v and radius Δ_o .

An (n, M, Δ_o) fingerprinting code is a mapping, $E : \mathcal{V} \times [1, M] \rightarrow \mathbb{F}^n$, such that $i \mapsto E(v, i)$ is injective for all $v \in \mathcal{V}$.

A coalition U is a set of fingerprinted copies of some $v \in \mathcal{V}$, and a forgery f computed from this coalition must verify :

$$\exists c \in U \quad d(c, f) \leq \Delta_f ,$$

for some fixed value Δ_f . Thus, the set of forgeries computable from a set U is

$$\text{forg}_{\Delta_f}(U) = \bigcup_{c \in U} B_{\Delta_f}(c) .$$

A fingerprinting code is said to be (L, Δ_f) resistant if there exists a mapping $T : \mathbb{F}^n \rightarrow \mathbb{F}^n$ (called the tracing mapping), such that for any coalition U of size at most L ,

$$T(\text{forg}_{\Delta_f}(U)) \subset U .$$

In the sequel, we will set $\mathcal{V} = \mathbb{F}^n$, and \mathbb{F} will denote the finite field with 2 elements.

3 Centered Error Correcting Codes

Centered error correcting (CEC) codes, which were introduced in [4], are a generalization of the classical notion of error correcting codes.

Roughly speaking, we add a condition on the localization of the codeword c encoding a message : we have a pair (m, v) and c must be within the ball $B_T(v)$ of center v and radius T , where T is a new parameter of the code.

Definition 1 *A centered error correcting code of parameters $(n, M, T, 2t+1)$ is defined by an encoding mapping $E : \mathbb{F}^n \times \mathcal{M} \rightarrow \mathbb{F}^n$ such that*

1. $\forall (v, m)$ and $(v', m') \neq (v, m)$

$$B_t(E(v, m)) \cap B_t(E(v', m')) = \emptyset$$

2. $\forall (v, m) \quad E(v, m) \in B_T(v)$

The first condition is a classical one to allow correction of t Hamming errors.

As easily seen, setting $T = n$ leads to the usual definition of error correcting codes since the condition 2) is then always satisfied.

Another way, equivalent to the first one, to define CEC codes is by considering disjoint M coverings of radius T , distant from each other of at least $2t + 1$. The sets $E(\mathbb{F}^n, m)$, where $m \in \mathcal{M}$, are coverings of radius T : for all v , $E(v, m)$ is at distance at most T from v by 1. of Definition 1. Moreover, those sets are $2t + 1$ apart from each other : otherwise, there exist couples (v, m) and $(v', m') \neq (v, m)$ with $B_t(E(v, m)) \cap B_t(E(v', m')) \neq \emptyset$ which would contradict 2. of Definition 1. The reverse way can be proved as easily

as the previous one.

AN EXPLICIT CONSTRUCTION

We will use the last definition of CEC codes in the proof of the following result

Theorem 1 *Let \mathbf{C} be a $[T, k, 2t + 1]$ code over \mathbb{F}_{2^r} . There exist $((2^r - 1) \cdot T, 2^{k \cdot r}, T, 2t + 1)$ CEC codes.*

Proof.

Let \mathcal{C} be the direct sum of T Hamming codes of length $2^r - 1$, thus the length n of \mathcal{C} is $n = (2^r - 1) \cdot T$, and let h be a parity check matrix of a Hamming code of length $2^r - 1$. A parity check matrix of \mathcal{C} can be obtained from H as a block matrix by

$$H = \begin{pmatrix} h & & 0 \\ & \ddots & \\ 0 & & h \end{pmatrix} \quad (1)$$

Choose a basis of \mathbb{F}_{2^r} over \mathbb{F} , thus $\mathbb{F}_{2^r}^T$ and \mathbb{F}^{rT} can be identified. Denote by \bar{x} the element in $\mathbb{F}_{2^r}^T$ corresponding to $x \in \mathbb{F}^{rT}$.

We can now define our CEC code \mathcal{C}' , as the following set of cosets of \mathcal{C}

$$\left\{ y + \mathcal{C} : \overline{y \cdot H^t} \in \mathbf{C} \right\}$$

where H^t is the transpose of H . Since cosets have the same covering radius as the code \mathcal{C} , and \mathcal{C} has clearly a radius equal to T , we just have to prove that $y + \mathcal{C}$ and $z + \mathcal{C}$ are $2t + 1$ apart when $y \cdot H^t \neq z \cdot H^t$. If we have $y' = y + c_1$ and $z' = z + c_2$ where $c_1, c_2 \in \mathcal{C}$ then

$$\overline{(y' + z') \cdot H^t} = \overline{(y + z) \cdot H^t} \in \mathbf{C} \setminus \{0\}$$

But, since H is a block matrix, grouping the coordinates of y and z by r , that is writing $y = y_1 \dots y_T$ and $z = z_1 \dots z_T$, for some $y_i, z_i \in \mathbb{F}^{2^r - 1}$, we have

$$\overline{(y + z) \cdot H^t} = \left(\overline{(y_1 + z_1) \cdot h^t}, \dots, \overline{(y_T + z_T) \cdot h^t} \right)$$

On the other hand, \mathbf{C} has minimal distance $2t + 1$. Thus there exist at least $2t + 1$ coordinates of $\overline{(y + z) \cdot H^t}$ that are different from 0. Equivalently there exist $2t + 1$ values of i such that $(y_i + z_i) \cdot h^t = (y'_i + z'_i) \cdot h^t \neq 0$, i.e. such that $d(y'_i, z'_i) \geq 1$. Finally, we get

$$d(y', z') = \sum_{0 \leq j < T} d(y'_j, z'_j) \geq 2t + 1$$

□

Corollary 1 For $2^r > T > 2t$, there exist $((2^r - 1)T, 2^{r(T-2t)}, T, t)$ CEC codes.

Proof.

Since $q = 2^r$ is greater than T , we can use for \mathbf{C} a $[T, T - 2t, t]$ Reed-Solomon code over \mathbb{F}_{2^r} . □

Using algebraic geometric codes for \mathbf{C} allows to drop the upper bound condition on T . So we can fix $q = 2^r$ (r must be even), and still have the length of the CEC code growing. Asymptotically, we can have for \mathbf{C}

$$\frac{k}{T} > 1 - \frac{2t + 1}{T} - \frac{1}{\sqrt{q} - 1}$$

In order to use CEC codes for fingerprinting, we will need to perform decoding efficiently, i.e. to calculate the closest $E(v, m)$ to a given word x , for a fixed v . With the construction used in Theorem 1 it is enough to have a decoding algorithm for the 2^r -ary code \mathbf{C} .

Let's explain the decoding algorithm. We have two words z, v and we search one of the closest $E(v, m)$ to z . In fact, we won't need v for this. First, we compute $\bar{s} = z \cdot H^t$ and decode it in \mathbf{C} . Writing \bar{m} the result of this decoding, we just need a word e such that $e \cdot H^t = m - z \cdot H^t$. Then, $z + e$ will be a solution.

Write $e = e_1 \dots e_T$ and $m - z \cdot H^t = a_1 \dots a_T$, with $e_i \in \mathbb{F}^{2^r-1}$ and $a_i \in \mathbb{F}^r$. Since H is a block matrix of the form (1), where h is a parity check matrix of the Hamming code of length $2^r - 1$, the vectors e_i can be computed independently. The matrix h contains all the possible columns of dimension r , so e_i is the word with only one non zero coordinate, the one corresponding to the column equal to a_i .

This algorithm clearly shows that the most difficult part is the decoding of the code \mathbf{C} .

4 Construction of Fingerprinting Codes

Recall we want to distribute slightly different copies of some binary word $v \in \mathbb{F}^n$ to M different users. A fingerprinted copy c must satisfy the following distortion criterion :

$$d(c, v) \leq \Delta_o \tag{2}$$

The purpose is to trace illicit copies, computed from some fingerprinted copies where "to trace" means to find one of the fingerprinted copy used to create the forgeries. An illicit copy z computed from a set U of original copies must satisfy only a distortion criterion :

$$\min_{c \in U} d(z, c) \leq \Delta_f \quad (3)$$

If we have a CEC code \mathcal{C} with parameters

$$\left(n, 2^{\lceil \log(M) \rceil}, \Delta_o, \Delta_f \right)$$

it is possible to solve the fingerprinting problem using the following scheme : denote by E the encoding mapping of \mathcal{C} , give to the m -th user the fingerprinted copy $E(v, m)$. The criterion 2 is clearly satisfied. Now, if a coalition U creates an illicit copy z , then by (3), there exists m in U such that

$$d(E(v, m), z) \leq \Delta_f$$

But, \mathcal{C} allows to correct Δ_f errors. Thus, decoding z gives $E(v, m)$ and we can recover at least one member of U . Remark that the size of the coalition does not matter, which is a very interesting property.

To be practical, this scheme requires that the code \mathcal{C} has two important properties. On the one hand, \mathcal{C} must have an efficient encoding mapping. On the other hand, it must have an efficient decoding algorithm, which is far more restrictive than the previous property.

5 Achievable parameters

Putting the Corollary 1 and the construction given in the previous section together leads to (binary) fingerprinting codes with parameters

- length n
- logarithm of the number of users equal to

$$(\Delta_o - 2\Delta_f) \cdot \log \left(1 + \frac{n}{\Delta_o} \right)$$

- distortion for original fingerprinted copies Δ_o
- distortion for forgeries Δ_f

with $\Delta_o > 2\Delta_f$ and $\log(1 + n/\Delta_o) > \Delta_o$. Since Reed-Solomon codes are decodable in polynomial time, the tracing algorithm is also polynomial. As remarked in section 4, those fingerprinting codes are effective against coalitions of arbitrary size.

Let the distortions Δ_o and Δ_f grow linearly with n as in [3]. Thus we can write $\Delta_o = \delta_o \cdot n$ and $\Delta_f = \delta_f \cdot n$. The rate of users, defined by the logarithm of the number of users divided by the length, is

$$(\delta_o - 2\delta_f) \cdot \log \left(1 + \frac{1}{\delta_o} \right)$$

and is constant.

6 On the Capacity Game of Private Fingerprinting Systems

Roughly speaking, the capacity is the highest possible rate of users. For the model we consider in this paper, the capacity of some fingerprinting systems is derived in [3]. The fingerprinting systems considered fulfill two technical assumptions (see [3, Def. III.2 and III.3]). The first one (constant composition) is a constraint on fingerprinting codes and the second (“smoothness”) is a constraint on sequences of fingerprinting codes. A practical consideration is given in order to justify the constant composition assumption, essentially it allows efficient computation of the fingerprinted copies.

Basically the results are the following :

- for a fixed coalition size L and the distortions Δ_o and Δ_f growing linearly with the data length n , the expression derived for the capacity is of the form $O(1/L)$;
- when L grows linearly with n , then the capacity is 0.

At first glance, these results seem in contradiction with ours since in section 5 we construct codes with an asymptotic rate bounded away from zero and effective for every L . In fact, a possible reason to explain this point is that our construction does not fulfill the two technical assumptions discussed earlier.

At least, we can say that these assumptions have very important consequences and are not so mild as stated in [3] since they drastically reduce the set of achievable rates. Moreover, they do not lead to fingerprinting

codes with more practical properties than our codes since our codes fulfill the practical considerations, and much more, used to justify the constant composition assumption.

7 Conclusion

The model we consider in this paper was introduced in [3]. This model allows dishonest users to change any part in their own original copies as soon as they don't change too many bits compare with at least one of their copies. Recall another model, well known as the marking assumption [1], which allows to change positions in which at least two members of the coalition have different bits. Contrary to intuition, the new model is less favorable to dishonest users than the marking assumption model. Indeed, it is not possible to construct binary fingerprinting codes secure against coalitions even of size two without allowing some probability of error in the tracing algorithm (see [1, Th. IV.2]). Whereas with the model used in [3], we can do it. Namely, we have proved that binary fingerprinting codes without error in tracing exist.

In fact our proof leads to codes with a rate bounded away from zero with a new and interesting property : these codes resist to coalition of arbitrary size. This proves that the capacity derived in [3, Th. IV.1] for some fingerprinting systems doesn't hold for general ones.

Some heuristic explaining this strange, at first glance, effect is that the condition

$$\exists c \in U \quad d(c, f) \leq \Delta_f ,$$

is very restrictive since it means that only single member of the coalition, namely c , produces a forgery f . Hence, despite that formally the new model deals with coalitions of dishonest users, in fact it reduces to the case of a single user.

Acknowledgment

We wish to thank Grigory Kabatiansky for enlightening discussions, Caroline Fontaine for a careful reading of a previous version and the anonymous referees for helpful comments.

References

- [1] D. Boneh and J. Show, “Collusion-secure fingerprinting for digital data,” *IEEE Transactions on Information Theory*, vol. 44, no. 5, pp. 1897–1905, 1998.
- [2] B. Chor, A. Fiat, M. Naor, and B. Pinkas, “Tracing traitors,” *IEEE Transactions on Information Theory*, vol. 46, no. 3, pp. 893–910, May 2000.
- [3] A. Somekh-Baruch and N. Merhav, “On the capacity game of private fingerprinting systems under collusion attacks,” *IEEE Transactions on Information Theory*, vol. 51, no. 3, pp. 884–899, 2005.
- [4] L. Bassalygo and M. Pinsker, “Centered error-correcting codes,” *Problems of Information Transmission*, vol. 35, pp. 30–37, 1999.
- [5] A. Kuznetsov and B. Tsybakov, “Coding in a memory with defective cells,” *Problems of Information Transmission*, vol. 10, no. 2, pp. 132–138, 1974.
- [6] G. Cohen, I. Honkala, S. Listyn, and A. Lobstein, *Covering Codes*. North-Holland, 1997.
- [7] F. MacWilliams and N. Sloane, *The Theory of Error-Correcting Codes*, 3rd ed. North-Holland, 1996.
- [8] M. Tsfasman and S. Vladut, *Algebraic Geometric Codes*, ser. Mathematics and its Applications. Kluwer Academic Publishers, 1991.