PAPER Special Section on Information Theory and Its Applications

Soft Decision Decoding of Boneh-Shaw Fingerprinting Codes*

Hans Georg SCHAATHUN[†], *Nonmember and* **Marcel FERNANDEZ**[††a)], *Member*

SUMMARY Collusion-secure codes are used for digital fingerprinting and for traitor tracing. In both cases, the goal is to prevent unauthorized copying of copyrighted material, by tracing at least one guilty user when illegal copies appear. The most well-known collusion-secure code is due to Boneh and Shaw (1995/98). In this paper we improve the decoding algorithm by using soft output from the inner decoder, and we show that this permits using significantly shorter codewords.

key words: *collusion-secure fingerprinting, copyright protection, traitor tracing, soft-decision decoding*

1. Introduction

Unauthorised copying and distribution of copyrighted material has received increasing attention over many years, both in research communities and in the daily press. Authors and artists depend on their income from legal sales, and unauthorised copying is often seen as a threat to these sales. For the movie or music industry, this is a question of big money.

Estimates of the losses due to illegal copying are generally disputable. There is no generally accepted method to estimate the sales that would have been achieved without illegal copying. For example, it is sometimes claimed that illegally distributed copies have a promotional effect which actually increase sales. Still, it is clear that big money are at stake and the issue receives interest from many different angles. Several countries these days change their legislation to deal more effectively with illegal distribution in new media.

Digital fingerprinting was introduced in [1], and given increasing attention following [2], [3]. Each copy of the copyrighted material is made unique by embedding a fingerprint identifying the buyer. When illegal copies are found, the fingerprint can be extracted and used to identify the guilty pirate.

The embedding of the fingerprint should not make perceptible changes in the file, so that the fingerprinted copy has the same practical value as the original. It must also be impossible for the user to remove or damage the fingerprint, without also destroying the information contents. In particular, the fingerprint must survive any change of file format (e.g. gif to tiff) and any reasonable lossy compression. This embedding problem is essentially the same as the problem of watermarking.

If a single pirate distributes unauthorised copies, they will carry his fingerprint. If the vendor discovers the illegal copies he can trace them back to the pirate and prosecute him. If several pirates collude, they can to some extent tamper with the fingerprint. When they compare their copies they see some bits (or symbols) which differ and thus must be part of the fingerprint. Identified bits may be changed, and thus the pirates create a hybrid copy with a false fingerprint.

A collusion-secure code is a set of fingerprints which enables the vendor to trace pirates even when they collude, given that there are no more than t pirates for some threshold t. Collusion-secure coding is also employed in traitor tracing [4], [5]. Whereas fingerprinting protects the digital data in themselves, traitor tracing protects broadcast encryption keys.

The most well-known collusion-secure code is the probabilistic scheme due to Boneh and Shaw [2], [3]. A handful of other schemes have also appeared over the years [6]–[8]. A new analysis of the error probability for the Boneh-Shaw scheme was made in [9], showing that the codewords could be made much shorter than initially assumed. In this paper, we make further improvement by using soft output from the inner decoding. The error analysis follows the approach of [9]. The major novelty of this paper is to find a good output parameter from the inner decoding. Soft decision decoding has also previously been applied for other collusion-secure codes, in [10] among others.

2. On Collusion-Secure Codes

2.1 Some Coding Theory

We use notation and terminology from coding theory. The set of fingerprints is an $(n, M)_q$ code, which provides for up to M buyers (size of the code), uses an alphabet of q symbols, and requires n such symbols embedded in the digital file. The code book is the matrix formed by taking the codewords (fingerprints) as rows. The Hamming distance $d(\vec{x}, \vec{y})$ between two words $\vec{x}$ and $\vec{y}$ is the number of positions where the two words differ, and the minimum distance of a code C is denoted d. The normalised minimum distance is $\delta = d/n$.

Manuscript received February 6, 2006.
Manuscript revised April 12, 2006.
Final manuscript received June 8, 2006.

[†]The author is with the Department of Informatics, University of Bergen, PB. 7800 N-5020 Bergen, Norway.
[††]The author is with the Departament of Telematics Engineering, Universitat Politècnica de Catalunya, C/Jordi Girona 1 i 3, Campus Nord, Mod C3, 08034 Barcelona, Spain.

*This work has been supported in part by the Spanish Research Council (CICYT) Project TSI2005-07293-C02-01 (SECONNET).

a) E-mail: marcel@entel.upc.es

DOI: 10.1093/ietfec/e89–a.10.2603

The rate of the code is $\rho = (\log M)/n$. We define concatenated codes as follows.

Definition 1: Let C_1 be a $(n_1, Q)_q$ and let C_2 be an $(n_2, M)_Q$ code. Then the concatenated code $C_1 \circ C_2$ is the $(n_1 n_2, M)_q$ code obtained by taking the words of C_2 and mapping every symbol on a word from C_1. Each set of n_1 symbols corresponding to one word of the inner code will be called a *block*.

Concatenated codes are often decoded by first decoding each block using some decoding algorithm for the inner code, so that a word of symbols from the outer code alphabet is obtained. This word can finally be decoded with a decoding algorithm designed for the outer code.

2.2 The Fingerprinting Problem

To understand the fingerprinting problem, we must know what the pirates are allowed to do. This is defined by the Marking Assumption.

Definition 2 (Marking Assumption): Let $P \subseteq C$ be the set of fingerprints held by a coalition of pirates. The pirates can produce a copy with a hybrid fingerprint $\vec{x}$ for any $\vec{x} \in F_C(P)$, where

$$F_C(P) = \{(c_1, \ldots, c_n) : \forall i, \exists (x_1, \ldots, x_n) \in P, x_i = c_i\}.$$

We call $F_C(P)$ the feasible set of P with respect to C.

The Marking Assumption defines the requirements for the embedding of the fingerprint in the digital data. Constructing appropriate embeddings is non-trivial, though it is not theoretically impossible [3]. Alternative assumptions have been proposed, and some overview of this can be found in [8].

A *tracing algorithm* for the code C is any algorithm A which takes a vector $\vec{x}$ as input and outputs a set $L \subseteq C$. If $\vec{x}$ is a false fingerprint produced by some coalition $P \subseteq C$, then A is successful if L is a non-empty subset of P. We say that we have an error of Type I if $L \cap P = \emptyset$ and an error of Type II if $L \backslash P \neq \emptyset$. A Type I error means that we do not find any guilty pirate, whereas Type II means accusing an innocent user. Let ϵ_{I} and ϵ_{II} denote the probabilities of Type I and Type II errors respectively. Given our juridical system, Type II is clearly a graver error than Type I, so we might accept ϵ_{I} higher than we can accept ϵ_{II}.

A code is (t, ϵ)-secure if the probability of error (of either type) is at most ϵ when there are at most t pirates.

A binary fingerprinting scheme consists of a binary (n, M) code C, a tracing algorithm A, and a bijection ι between C and the set of users. The tracing algorithm A is public information. The code C may be secret information, but it is drawn at random from some probability distribution which is publicly known. The mapping ι may be secret or public. The ensemble of secret information is called the *key*.

Our challenge is, for a given number of users M and a maximum number of pirates t, to find a code with the shortest possible length n and the best possible error rate ϵ. It is also advantageous if the tracing algorithm A is efficient.

3. The Code

3.1 The Concatenated Construction

The fingerprinting code of [3] is a concatenated code, that consists of an inner code that we call the BS code, and an outer code.

Using the notation in Sect. 2.1, the inner code is the binary $(r(q-1), q)$ code from [3]. The code consists of q codewords of length $r(q-1)$ and has rate $\rho_I = \log q/(r(q-1))$.

The outer code C_O is an $(n, M)_q$ code of rate $\rho_O = \log M/n$. In the original Boneh-Shaw scheme it was a random code (RC). We will consider both random codes and algebraic codes with large minimum distance.

The obtained concatenated code is a binary code of length $nr(q-1)$ that consists of M codewords. Note that the rate of the concatenated code ρ_t, can be expressed as $\rho_t = \rho_I \cdot \rho_O$.

Note that since there is a one-to-one correspondence between words in the outer code and words in the concatenated code, we can identify each user with codewords of either code. By abuse of language, we will refer to both words of C_O and words of the concatenated code, as fingerprints.

3.2 On the Inner Code

The fingerprinting code of [3] is a concatenated code, of an inner code which we call the BS code, and a random outer code. The BS code is a binary $(r(M-1), M)$ code which is (M, ϵ)-secure. The code book has $M-1$ distinct columns replicated r times. A set of identical columns will be called a type. Every column has the form $(1 \ldots 10 \ldots 0)$, such that the i-th $(1 \leq i \leq M)$ user has zeroes in the first $i-1$ types and a one in the rest. Before embedding, the columns of the inner code are reorder using a random, secret permutation. This ensures that unless user i is a pirate, the pirates cannot distinguish between the $(i-1)$-th and the i-th type. Hence they have to use the same probability of choosing a 1 for both these types.

A hybrid fingerprint is characterised by the number F_i of ones for each column type i. Let $F_0 = 0$ and $F_q = r$ by convention (as if there were a column type 0 with all zeroes, and a type q with all ones). The F_i are stochastic (random) variables with distributions depending on the pirate strategy. If user i be innocent, the pirates cannot distinguish between column types i and $i-1$, and consequently $F_i \sim F_{i-1}$, where $\sim$ indicates that the two stochastic variables are identically distributed.

The decoding algorithm of the original Boneh-Shaw scheme makes a hypothesis test of the null hypothesis $F_i \sim F_{i-1}$ for each i. If this null hypothesis is rejected, user i is assumed to be guilty. This gives a threshold such that if $|F_i - F_{i-1}|$ is sufficiently high, then user i can be accused. This provides hard input to the outer decoding algorithm.

Our idea is to return soft information, i.e. a function of $F_i - F_{i-1}$, to be used by the outer decoding algorithm.

We have played with many variants, but most of them have been very difficult to work through the error analysis. The following might not be optimal, but it does work well. The output is a vector $\vec{v} = (v_1, \ldots, v_q)$, given as

$$v_j = \frac{F_j - F_{j-1}}{r}. \tag{1}$$

Observe that all the v_j sum to 1 and $v_j \in [-1, 1]$ for all j.

Furthermore, if the pirate j is innocent, then $E(v_j) = 0$, where $E(\cdot)$ denotes the expectation of an stochastic variable. When decoding one block of the concatenated code, j being innocent means that none of the pirates see symbol j in this block.

Example For $q = 8$ and $r = 3$, the code book of the BS code is defined as

$$C = \begin{pmatrix} \overbrace{111}^{F_1} & \overbrace{111}^{F_2} & \overbrace{111}^{F_3} & \overbrace{111}^{F_4} & \overbrace{111}^{F_5} & \overbrace{111}^{F_6} & \overbrace{111}^{F_7} & 1 \\ 000 & 111 & 111 & 111 & 111 & 111 & 111 & 2 \\ 000 & 000 & 111 & 111 & 111 & 111 & 111 & 3 \\ 000 & 000 & 000 & 111 & 111 & 111 & 111 & 4 \\ 000 & 000 & 000 & 000 & 111 & 111 & 111 & 5 \\ 000 & 000 & 000 & 000 & 000 & 111 & 111 & 6 \\ 000 & 000 & 000 & 000 & 000 & 000 & 111 & 7 \\ 000 & 000 & 000 & 000 & 000 & 000 & 000 & 8 \end{pmatrix}$$

If the pirates are associated with codewords $2, 4, 6$, then

	$\overbrace{111}^{v_3}$	$\overbrace{111}^{v_5}$		
000	111	111	111 111	$p_1 = 2$
000	000	111	111 111	$p_2 = 4$
000	000	000	111 111	$p_3 = 6$

We note that, since codeword 3 does not belong to the pirate coalition the columns in v_3 are indistinguishable by the pirates and therefore, since $v_3 = \frac{F_3 - F_2}{r}$, we have that $E(v_3) = 0$.

3.3 On the Outer Code

Boneh and Shaw suggested to concatenate the BS code with a random $(n, M)_q$ code which would be decoded using closest neighbor decoding. In the improved error analysis [9], this was replaced by list decoding, returning all codewords within a certain distance of the hybrid word after inner decoding. It has also been suggested to use codes with large minimum distance, typically AG codes, but unfortunately such codes need much larger alphabets than random codes, and the inner Boneh-Shaw code is bad in that case, having length linear in q. We will return to outer codes with a large distance in a later section.

After inner decoding of all the blocks, we form the $q \times n$ reliability matrix $R = [r_{i,j}]$ where the j-th column is the vector $\vec{v}$ from inner decoding of the j-th block. We employ the common approximation that the pirates make independent decisions in each column, such that $F_i \sim B(r, p_i)$ for some probability p_i, in other words, F_i is a Binomial stochastic variable with r Bernoulli trials and probability p_i to get 1 (success).

The outer decoding algorithm takes the $q \times n$ reliability matrix R as input and returns all codewords $\vec{c} = (c_1, \ldots, c_n)$ that satisfy

$$W(\vec{c}) = \sum_{i=1}^{n} r_{i,c_i} \geq \Delta n.$$

We call $W(\vec{c})$ the weight of $\vec{c}$. The decoding can always be made in time $O(n \cdot M)$.

It is an important property that the terms r_{i,c_i} of the sum are stochastically independent. Each term is also bounded in the interval $[-1, 1]$ and has a fairly simple distribution. This will allow us to use the well-known Chernoff bound in the error analysis.

Theorem 1 (Chernoff): Let $X_1, \ldots, X_t$ be bounded, independent, and identically distributed stochastic variables in the range $[0, 1]$. Let x be their (common) expected value. Then for any $0 < \delta < 1$, we have

$$P\left(\sum_{i=1}^{t} X_i \leq t\delta\right) \leq 2^{-tD(\delta\|x)}, \quad \text{when } \delta < x,$$

$$P\left(\sum_{i=1}^{t} X_i \geq t\delta\right) \leq 2^{-tD(\delta\|x)}, \quad \text{when } \delta > x,$$

where

$$D(\sigma\|p) = \sigma \log \frac{\sigma}{p} + (1 - \sigma) \log \frac{1 - \sigma}{1 - p}.$$

For an understanding of the proof of this bound, we recommend to read [11].

We define three schemes in this paper, all using the BS inner code and soft decision decoding. We call the scheme BS-RC-SD, using a random outer code; BS-RS-SD using a Reed-Solomon outer code; and BS-AG-SD using an (asymptotically good) family of Algebraic Geometry codes.

4. Error Analysis

In this section, we shall bound the error probability for concatenated codes with Boneh-Shaw inner codes and soft decision decoding as defined in the previous section. The probability of failing to accuse any guilty user is independent of the outer code used. We study this error probability in the first subsection. The probability of accusing innocent users does depend on the outer code, and this will be studied for random codes in Sect. 4.2 and for codes with large distance in Sect. 4.3.

4.1 Probability of Failing

The probability ϵ_1 that the decoding algorithm outputs no guilty user, is bounded as

$$\epsilon_1 \leq P\left(\frac{1}{t}\sum_{i=1}^{n}\sum_{\vec{c}\in P} r_{i,c_i} \leq \Delta n\right) = P\left(\sum_{i=1}^{n} Y_i \leq \Delta n\right)$$

where

$$Y_i = \sum_{\vec{c}\in P} \frac{r_{i,c_i}}{t} = \frac{1}{t} \sum_{\vec{c}\in P} \frac{F_{c_i} - F_{c_i-1}}{r}.$$

Obviously

$$\sum_{\gamma\in Q} \frac{F_\gamma - F_{\gamma-1}}{r} = 1,$$

and $E(F_\gamma - F_{\gamma-1}) = 0$ when γ is not seen by the pirates. Hence we get $E(Y_i) = 1/t$. Observe that $-1 \le Y_i \le 1$. In order to get a stochastic variable in the range $[0, 1]$, we set $X_i = (1 + Y_i)/2$. Thus

$$E(X_i) = \bar{x} = \frac{t+1}{2t},$$

and we get

$$\epsilon_{\mathrm{I}} \le P\left(\sum_{i=1}^{n} X_i \le \frac{1+\Delta}{2} n\right).$$

If $1/t > \Delta$, the Chernoff bound is applicable, giving the following theorem.

Theorem 2: Using the concatenated code with a BS inner code and soft input list decoding with threshold $\Delta < 1/t$ for the outer code, the probability of failing to accuse any guilty user is given as

$$\epsilon_{\mathrm{I}} \le 2^{-nE}, \text{ where } E = D\left(\frac{1+\Delta}{2} \Big\| \frac{t+1}{2t}\right). \tag{2}$$

This bound is independent of the choice of outer code.

4.2 Random Outer Codes

In this section we study the BS-RC-SD scheme. The outer code is constructed by drawing each symbol for each codeword independently and uniformly at random from the alphabet. This random code is kept secret by the vendor. The bound on ϵ_{I} from Theorem 2 is still valid. In this section we bound ϵ_{II}.

Let $\vec{c} \notin P$ be an innocent user. The probability of accusing $\vec{c}$ is

$$\pi(\vec{c}) = P\left(\sum_{i=1}^{n} r_{i,c_i} \ge \Delta n\right).$$

Clearly $E(r_{i,c_i}) = 1/q$. Like in the last section, we make a stochastic variable in the $[0, 1]$ range,

$$X_i = \frac{1 + r_{i,c_i}}{2},$$
$$E(X_i) = \frac{q+1}{2q},$$

and

$$\pi(\vec{c}) = P\left(\sum_{i=1}^{n} X_i \ge \frac{1+\Delta}{2} n\right).$$

Table 1 Some constructions with random outer codes and $\epsilon \le 10^{-10}$.

$\log M$	t	q	Δ	n_O	n
10	10	35	0.06667	42400	1 441 600
10	50	153	0.01370	1 200 000	182 400 000
10	32	100	0.02128	480 000	47 520 000
20	20	66	0.03448	200 000	13 000 000
20	100	335	0.006897	4 918 000	1 642 612 000
20	1000	3350	0.0006897	$4.95{\cdot}10^8$	$1.657755{\cdot}10^{12}$
30	30	99	0.02353	494 000	48 412 000
30	150	500	0.004701	12 230 000	6 102 770 000

Table 2 Asymptotic rates for some constructions of RS-AG with soft decision decoding.

	Random codes		AG codes		Old record
t	q	Rate	q	Rate	Rate
2	5	0.0180	9^2	$6.79 \cdot 10^{-4}$	0.0688 [9]
3	8	0.00466	19^2	$6.14 \cdot 10^{-5}$	0.00102 [8]
4	11	0.00187	32^2	$1.10 \cdot 10^{-5}$	$1.01 \cdot 10^{-4}$
5	14	0.000930	49^2	$2.89 \cdot 10^{-6}$	$1.25 \cdot 10^{-5}$
6	17	0.000529	64^2	$9.42 \cdot 10^{-7}$	$1.78 \cdot 10^{-6}$
7	20	0.000330	91^2	$3.80 \cdot 10^{-7}$	$2.76 \cdot 10^{-7}$

Theorem 3: Concatenating a $(r(q-1), q)$ BS code with a random outer code using soft input list decoding with threshold $\Delta > 1/q$ for the outer code, the probability of accusing an innocent user is given as

$$\epsilon_{\mathrm{II}} \le 2^{(\rho_O \log q - E)n}, E = D\left(\frac{1+\Delta}{2} \Big\| \frac{q+1}{2q}\right).$$

Interestingly, both the bounds on ϵ_{I} and ϵ_{II} are independent of r, and hence we are going to choose $r = 1$ to minimise the length.

In Table 1, we show some constructions of RS-RC-Soft. The parameters have been found by trial and error, and cannot be expected to be optimal. However, major improvements appear to be impossible.

Theorem 4: For any $q > t$, there is an asymptotic class of (t, ϵ)-secure codes with $\epsilon \to \infty$ and rate given by

$$\rho_t \approx \frac{D\left(\frac{t+1}{2t} \| \frac{q+1}{2q}\right)}{q-1}.$$

Proof: For asymptotic codes, $\epsilon_{\mathrm{I}} \to 0$ if $\Delta < 1/t$, so we can take $\Delta \approx 1/t$. Likewise, $\epsilon_{\mathrm{II}} \to 0$ if $\Delta > 1/q$ and

$$\rho_O < \frac{D\left(\frac{t+1}{2t} \| \frac{q+1}{2q}\right)}{\log q}.$$

Since $\rho_I = \log q/(q-1)$, we get the theorem. □

Unfortunately, we cannot see any nice expression for the optimal value of q. Clearly, we require $q = \Omega(t)$, and if $q = \Theta(t)$, we get $\rho_t = O(t^{-3})$. The only other known scheme with $\rho_t = \Omega(t^{-3})$ is the Tardos scheme with $\rho_t = \Theta(t^{-2})$. Table 2 presents asymptotic rates for some constructions against few pirates.

4.3 Outer Code with Large Distance

Suppose now that we use an outer code with large minimum distance δ, typically a Reed-Solomon (RS) code in the finite case or an algebraic geometry (AG) code asymptotically. This leads to the BS-RS-SD and BS-AG-SD schemes. We want to bound the probability of accusing $\vec{c}$ when $\vec{c}$ is innocent, i.e. to bound the probability

$$\pi(\vec{c}) \leq P\left(\sum_{i=1}^{n} r_{i,c_i} \geq \Delta n\right).$$

An innocent user $\vec{c}$ can match a given pirate in at most $(1-\delta)n$ positions. Thus there are at most $t(1-\delta)n$ positions where $\vec{c}$ matches some pirate. For the purpose of a worst case analysis, we assume that $r_{i,c_i} = 1$ whenever c_i matches a pirate. There are at least $N = [1 - t(1-\delta)]n$ positions $i_1, \ldots, i_N$, where $r_{i,j} = v_j$ is given by (1) with $F_j \sim F_{j-1}$. Thus we get

$$\pi(\vec{c}) \leq P\left(\sum_{j=1}^{N} r_{i_j,c_{i_j}} \geq \tau N\right),$$

$$N = [1 - t(1-\delta)]n,$$

$$\tau = \frac{\Delta - t(1-\delta)}{1 - t(1-\delta)}.$$

Clearly, τ increases in δ as well as in Δ.

When $F_j \sim F_{j-1}$, we have $E(F_j - F_{j-1}) = 0$. Setting $Y_j = (1 + r_{i_j,c_{i_j}})/2$, we get $E(Y_j) = 1/2$ and

$$\pi(\vec{c}) \leq P\left(\sum_{j=1}^{N} Y_j \geq \frac{1+\tau}{2} N\right),$$

This results in the following theorem.

Theorem 5: Concatenating a $(r(q-1), q)$ BS code with a $(n, 2^{\rho_O n}, \delta n)$ outer code using soft input list decoding with threshold Δ for the outer code, the probability of accusing an innocent user is given as

$$\epsilon_{\mathrm{II}} \leq 2^{(\rho_O \log q - [1-t(1-\delta)]D(\sigma\|1/2))n},$$

provided $\Delta > t(1-\delta)$, and

$$\sigma = \frac{1}{2} + \frac{\Delta - t(1-\delta)}{2(1 - t(1-\delta))}.$$

Again, we see that the error rate is independent on r, so $r = 1$ for maximum rate. In Table 3, we show some good constructions of RS-RS-SD. These lengths are better than those of RS-RC, but RS-RC appears to be better for large t. In particular, RS-RS-SD requires $q > t^2$. This is also illustrated by the fact that we got a shorter length for $t = 100$ when $M = 2^{30}$ than when $M = 2^{20}$ in the table.

In Table 4, we compare lengths for the different variants. We observe that random codes provide the shortest lengths, but for moderate t, even Reed-Solomon codes with soft decision beat the old random codes with hard decision.

Table 3 Some good finite constructions with Reed-Solomon outer codes ($q = n_O$). To get $\epsilon \leq 10^{-10}$, the code is also concatenated with an $[m, 1]$ repetition code.

$\log M$	t	$[n_O, k_O]$	m	Δ	n
10	10	[1024,1]	22	0.053	23 046 144
20	20	[1024,2]	252	0.0364	263 983 104
20	100	[2^{20},1]	3	0.006	3 298 531 737 600
30	30	[2^{15},2]	8	0.02	8 589 672 448
30	100	[2^{15},2]	169	0.00707	181 456 830 464
30	150	[2^{15},2]	1861	0.005781	1 998 172 553 216

Table 4 Comparison of finite constructions of the two new schemes and the original Boneh-Shaw code with improved error analysis.

		Hard dec. [9]	Random code	Large distance
$\log M$	t	n	n	n
10	10	306 548 964	1 441 600	23 046 144
10	50	$0.233 \cdot 10^{12}$	182 400 000	555 202 560
10	32	$0.265 \cdot 10^{11}$	47 520 000	227 318 784
20	20	$6.44 \cdot 10^{9}$	13 000 000	263 983 104
20	100	$5.10 \cdot 10^{12}$	1 642 612 000	3 298 531 737 600
20	10^3	$7.02 \cdot 10^{16}$	$1.657755 \cdot 10^{12}$	255086454374400
30	30	$4.09 \cdot 10^{10}$	48 412 000	8 589 672 448
30	150	$1.38 \cdot 10^{23}$	6 102 770 000	1 998 172 553 216

4.3.1 Asymptotic Codes

For asymptotic codes, $\epsilon_{\mathrm{I}} \to 0$ if $\Delta < 1/t$, so we can take $\Delta \approx 1/t$. Likewise, $\epsilon_{\mathrm{II}} \to 0$ if both $\Delta > t(1-\delta)$ and

$$\rho_O < \frac{1 - t(1-\delta)}{\log q} D(\sigma\|1/2).$$

Using AG codes with

$$\rho = 1 - \delta - \frac{1}{\sqrt{q} - 1},$$

where q is an even prime power, we can get codes with ρ_O solving the following

$$\rho_O = \frac{1 - t\left(\rho_O + \frac{1}{\sqrt{q}-1}\right)}{\log q} D\left(\frac{1+\alpha}{2}\middle\|\frac{1}{2}\right), \tag{3}$$

$$\alpha = \frac{1 - t^2\left(\rho_O + \frac{1}{\sqrt{q}-1}\right)}{t - t^2\left(\rho_O + \frac{1}{\sqrt{q}-1}\right)} > 0. \tag{4}$$

The total rate is $\rho_t(q) = \rho_I \cdot \rho_O$ where

$$\rho_I = \frac{\log q}{q-1}.$$

The number of pirates t, is a property of the resulting codes, whereas q is a control parameter chosen so as to maximise ρ_t. We have computed some asymptotic rates in Table 2, by choosing q by trial and error, and solving (3) by fix point iteration.

5. On Complexity, Conclusions, Open Problems

We have constructed two new collusion-secure coding scheme with good rates; the codewords are significantly shorter than for the popular Boneh-Shaw scheme. Using Koetter-Vardy decoding for BS-RS-SD gives decoding complexity polynomial in n. The other scheme, BS-RC-SD, has better code rate, but complexity $O(n \cdot M)$, which is the same as the original Boneh-Shaw scheme.

We know of three other schemes which are secure under the Marking Assumption. The BBK scheme [8] has decoding complexity polynomial in n, like BS-RS-SD, but exponential in t. Also the code length increases rapidly in t. This makes BBK a very good scheme against few pirates, for more pirates decoding is intractable and the code rate becomes inferior to BS-RS-SD, cf. Table 2.

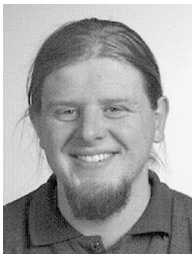
The Tardos [7] and LBH [6] schemes both have decoding complexity $O(n \cdot M)$. The LBH scheme has $n = \Theta(2^t)$, and is thus uninteresting except for very few pirates. The Tardos scheme however has length only $n = 100t^2 \log(M/\epsilon)$, which is better than either of our schemes.

The main contribution of this paper is the BS-RS-SD and BS-AG-SD schemes, having efficient decoding in terms of t as well as in terms of M.

We also note that the error bound ϵ is valid for any pirate coalition of size t, and thus the conditional error probability given the pirates observation of compared copies is also bounded by ϵ, and any pirate coalition not willing to risk being caught with probability $1-\epsilon$ will be deterred. For the Tardos scheme only the unconditional error probability was explicitly bounded, which means that a pirate coalition willing to risk being caught if the probability is slightly less than $1-\epsilon$, can compare their copies first, and they might not be deterred.

References

[1] N.R. Wagner, "Fingerprinting," Proc. 1983 Symposium on Security and Privacy, pp.18–22, 1983.
[2] D. Boneh and J. Shaw, "Collusion-secure fingerprinting for digital data," Advances in Cryptology — CRYPTO'95, LNCS 963, pp.452–465, Springer, 1995.
[3] D. Boneh and J. Shaw, "Collusion-secure fingerprinting for digital data," IEEE Trans. Inf. Theory, vol.44, no.5, pp.1897–1905, 1998, Presented in part at CRYPTO'95.
[4] B. Chor, A. Fiat, and M. Naor, "Tracing traitors," Advances in Cryptology — CRYPTO'94, LNCS 839, pp.257–270, Springer-Verlag, 1994.
[5] B. Chor, A. Fiat, M. Naor, and B. Pinkas, "Tracing traitors," IEEE Trans. Inf. Theory, vol.46, no.3, pp.893–910, May 2000, Presented in part at CRYPTO'94.
[6] T. Van Le, M. Burmester, and J. Hu, "Short c-secure fingerprinting codes," Proc. 6th Information Security Conference, Oct. 2003.
[7] G. Tardos, "Optimal probabilistic fingerprint codes," Proc. 35th Annual ACM Symposium on Theory of Computing, pp.116–125, 2003.
[8] A. Barg, G.R. Blakley, and G.A. Kabatiansky, "Digital fingerprinting codes: Problem statements, constructions, identification of traitors," IEEE Trans. Inf. Theory, vol.49, no.4, pp.852–865, April 2003.
[9] H.G. Schaathun, "The Boneh-Shaw fingerprinting scheme is better than we thought," Technical Report 256, Dept. of Informatics, University of Bergen, 2003.
[10] M. Fernández and M. Soriano, "Fingerprinting concatenated codes with efficient identification," Information Security (ISC'02), LNCS 2433, pp.459–470, Springer, 2002.
[11] T. Hagerup and C. Rüb, "A guided tour of Chernoff bounds," Inf. Process. Lett., vol.33, pp.305–308, 1990.



Hans Georg Schaathun was born in Bergen, Norway, in 1975. He is Cand.Mag. 1996, Cand.Scient. 1999, and Dr.Scient. 2002, all from the University of Bergen, Norway. He was lecturer in coding and cryptography at the University of Bergen 2002 and Post.Doc. 2003–2006. He started as a permanent lecturer at the University of Surrey, England, February 2006. His main research field is in multimedia security, and he is also intersted in algebraic coding theory and cryptography.

Marcel Fernandez received the M.S. degree in Telecommunications Engineering (1997) and a Ph.D. in Telematics Engineering (2003) both from the Universitat Politènica de Catalunya (UPC). In 2002, he joined the Information Security Workgroup within the Telematics Services Research Group at the Department of Telematics Engineering of the UPC. Currently he works as an assistant professor at the Telecommunications Engineering School in Barcelona-ETSETB. His research interests include error correcting codes and its applications to copyright protection of digital objects.