The Boneh-Shaw Fingerprinting Scheme is Better than We Thought

Hans Georg Schaathun

27th November 2003

With a digital fingerprinting scheme a vendor of digital copies of copyrighted material marks each individual copy with a unique fingerprint. If an illegal copy appears, it can be traced back to one or more guilty pirates, due to this fingerprint.

Boneh and Shaw have deviced a classic fingerprinting scheme, and several recent papers have designed improvements. In the present paper we make a new error analysis of Boneh and Shaw's original scheme, and we prove that it is far better than assumed and in fact better than the improvements in some respects.

Keywords
collusion-secure fingerprinting, copyright protection, traitor tracing

1 Introduction

The problem of digital fingerprinting was introduced in [19], studied in [2], and given increasing attention following [3, 4]. A vendor selling digital copies of copyrighted material wants to prevent illegal copying. Digital fingerprinting is supposed to make it possible to trace the guilty user (pirate) when an illegal copy is found. This is done by embedding a secret identification mark, called a fingerprint, in each copy, making every copy unique.

The fingerprint must be embedded in such a way that it does not disturb the information in the data file too much. It must also be impossible for the user to remove or damage the fingerprint, without damaging the information contents beyond any practical use. In particular, the fingerprint must survive any change of file format (e.g. gif to tiff) and any reasonable compression including 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 [6, 7]. Whereas fingerprinting protects the digital data in themselves, traitor tracing protects broadcast encryption keys. The fingerprinting literature is most often interested in probabilistically collusion-secure coding, where the vendor shall be able to trace a pirate with probability at least $1-\epsilon$ for some small error rate ϵ. In the traitor tracing literature, combinatorially collusion-secure codes is the norm, where the tracing is required to succeed with probaility 1. Still, in principle, there is no obvious reason not to use combinatorial codes for fingerprinting and probabilistic ones for traitor tracing. Other important variants of the problems are dynamic traitor tracing (e.g. [14]) and anonymous fingerprinting [13].

In this paper we make a new error-analysis to show that the Boneh and Shaw scheme from [3, 4] is better than previously known. The lengths can be made shorter than previously assumed. In particular the scheme yields asymptotic classes of codes with positive rate and exponentially decreasing error rate, a property first proved for the BBK scheme [1]. In the case with two pirates, we also present modifications which greatly reduce the required length.

2 Preliminaries

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, uses an alphabet of q symbols, and requires n such symbols embedded in the digital file. The Hamming distance between two words $\mathbf{x}$ and $\mathbf{y}$ is denoted $d(\mathbf{x}, \mathbf{y})$, and the minimum distance of a code C is denoted $d(C)$ or just d. The normalised minimum distance is $\delta = d/n$. The code book C is a matrix where the rows are the codewords of C. The rate of the code is $R = (\log M)/n$.

Closest neighbour decoding is any algorithm which takes a word $\mathbf{x}$ and returns a word $\mathbf{c} \in C$ such that $d(\mathbf{c}, \mathbf{x})$ is minimised. This can always be performed in $O(M)$ operations, and for some codes it may be faster.

Concatenation is a standard technique from coding theory, and it has proven extremely useful in fingerprinting.

Definition 1 (Concatenation)
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.

For the error analysis, we will use the well known Chernoff bound as given in the following theorem. See e.g. [10] for a proof. The relative entropy is defined as

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

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

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

We write $\mathcal{B}(n,p)$ for the binomial distribution with n trials with probability p. If X is distributed as $\mathcal{B}(n,p)$, we write $X \sim \mathcal{B}(n,p)$.

3 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 (The Marking Assumption)
Let $P \subseteq C$ be the set of fingerprints held by a coallition of pirates. The pirates can produce a copy with a false fingerprint $\mathbf{x}$ for any $\mathbf{x} \in F_C(P)$, 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 from the embedding of the fingerprint in the digital data. Constructing appropriate embeddings, is non-trivial, though it is not theoretically impossible [4]. Alternative assumptions have been proposed, and some overview of this can be found in [1].

A *tracing algorithm* for the code C is any algorithm A which takes a vector $\mathbf{x}$ as input and outputs a set $L \subseteq C$. If $\mathbf{x}$ is a false fingerprint produced by some coallition $P \subseteq C$. 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, wheras Type II means accusing an innocent user. Let ϵ_1 and ϵ_2 denote the probabilites 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_1 higher than we can accept ϵ_2.

A code is said to be said to be combinatorially t-secure, if it has a tracing algorithm which succeeds with probability 1 when there are at most t pirates. It is said to be t-secure with ϵ-error if the probability of error (of either type) is at most ϵ when there are at most t pirates.

Fingerprinting is a cryptographic problem; it is a problem of identification. The pirates should be prevented from hiding their identity when they make and distribute copies of a file. According to Kerchoff's principles it is important to have a clear understanding of what is secret information and what is public information.

A binary fingerprinting scheme consists of a binary (n, M) code C, a tracing algorithm A, and a mapping ι between C and the set of users. The tracing algorithm A is public information. The code C is 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*.

The fingerprinting scheme should be evaluated for each M, according to the code length n, the error probabilities ϵ_1 and ϵ_2, the running time of A, and the amount of secret information which has to be stored (key size). This is a lot of parameters, so we do not expect one scheme to be better than any other in every way.

In the next main section, we will study and improve the classic concatenated scheme due to Boneh and Shaw (BS-CS) [4], but first we will briefly present some other fingerprinting schemes from the literature. We will need most of the key ideas from those schemes in our discussion later on.

It is well known that any code with $\delta > 1 - t^{-2}$ is a socalled t-traceability code, which is combinatorially t-secure using closest neighbour decoding. Unfortunately, this large minimum distance is only possible when the alphabet is large. A binary code cannot be combinatorially collusion-secure.

3.1 Boneh and Shaw replication scheme (BS-RS)

BS-RS is used as the inner code in the Boneh and Shaw concatenated scheme (BS-CS). It uses a binary $(r(M-1), M)$ code which is M-secure with ϵ-error. 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 $(1 \le i \le M)$ user has zeroes in the first $i-1$ types and a one in the rest. We can see 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 in both these types. If r is large enough we can use statistics to test the null hypothesis that user i be innocent. The output is a list of users for which the null hypothesis may be rejected.

Theorem 2 (Boneh and Shaw)
The BS-RS with replication factor r is M-secure with ϵ-error whenever $r = 2M^2 \log(2M/\epsilon)$.

The tracing complexity of BS-RS is clearly $O(n)$. The key space consists of all the permutations of the columns of the code book, so the key size in bits is

$$K = \log \frac{(r(M-1))!}{(r!)^{M-1}}.$$

3.2 General binary schemes

Barg, Blakley, and Khabatiansky introduced a new scheme, which we call the BBK scheme, in [1]. They use socalled separating codes as inner codes, and codes with large distance as outer codes. The outer code distance must be larger than what is required for traceability codes, because the codes have to correct tracing errors from inner decoding in addition to the tracing. An important idea behind this scheme is that the inner code can have a very high error rate, because the outer code can be made powerful enough to correct it. We shall see that this idea applies to the Boneh and Shaw scheme as well.

The BBK scheme exhibits asymptotically non-zero rate and exponentially declining error rate with increasing code size. They suggest to use algebraic geometry codes as outer codes, and to decode them with the GS list decoding algorithm [9]. Though the running time for inner decoding may be heavy, the asymptotical complexity is polynomial in the code length n.

The BBK scheme uses a key much shorter than that of other known schemes. Only the mapping from the outer code alphabet onto the inner code must be kept secret. On the other hand, this mapping must be chosen at random for each block. Thus the key size is $n_2 \log Q! = O(\log M)$ bits, where n_2 is the outer code length and Q is the outer code alphabet size.

Another scheme was proposed in [12] with $n = O(\log M - \log \epsilon)$, but the paper only states asymptotic bounds on the lengths besides a few lengths against three pirates.

3.3 Against two or three pirates

In addition to the general t-secure schemes, there exist a few 2- or 3-secure codes. Simplex codes were proved to be 2-secure with ϵ-error in [11]. Small simplex codes are very good, and closest neighbour decoding can be used. However, the asymptotic rate of these codes is zero. A similar idea was employed in [16], where an asymptotically good family $(2,2)$-separating codes was proven to be 2-secure with ϵ-error, where ϵ tends to zero with increasing code size.

Scattering codes were introduced in [18], and by concatenating scattering codes and simplex codes, 3-secure codes with ϵ-error are obtained. This scheme also works well for small M, but the asymptotic rate is zero.

4 Concatenated schemes

Two combinatorially t-secure codes can simply be concatenated to obtain a larger t-secure code. The fingerprints can be viewed alternately as words of the outer code C_O or of the concatenated code C. Decoding starts with an element of the feasible set with respect to C. Successful inner decoding of a block gives an outer code symbol which is seen by one of the pirates; thus inner decoding yields a vector in the feasible set with respect to the outer code, which can be decoded.

Having two probabilistically t-secure codes, concatenation is non-trivial, but it can still be done, as it is for BS-CS. Suppose we want to construct a t-secure code C with ϵ-error. BS-CS uses closest neighbour decoding, and Boneh and Shaw chose parameters such that inner decoding succeeds in every position with probability $1-\epsilon/2$, and and such that outer decoding, given perfect inner decoding, succeeds with probability $1-\epsilon/2$. Thus the total error probability is less than ϵ.

BS-CS is actually far better than proved by Boneh and Shaw. Demanding that inner decoding be correct in every position is a strong requirement, because its probability declines exponentially in the code length. This requirement is not necessary. A small fraction of failures from inner decoding will only slightly increase the error probability in outer decoding and improve the overall error rate significantly. This observation was put to use in the BBK scheme, but it should be remembered for any concatenated scheme.

We suggest to decode the outer code with list decoding. Apart from the obvious advantage of allowing us to trace more than one pirate in many cases, it also makes the error analysis simpler, and it becomes clear how to adapt the error analysis for other choices for inner and outer codes in the scheme. Even though an error analysis for closest neighbour decoding can be made, it is not expected to give better error bounds.

4.1 List decoding of concatenated codes

Let C_I be an (n_1,q) inner code which is t-secure with ϵ_{in}-error, and C_O an $(n_2,M)_q$ outer code. Let R_I and R_O denote the rates of C_I and C_O respectively.

Our decoding algorithm works as follows. Let P be a pirate coallition of size at most t, and $\mathbf{x} \in F_C(P)$. First each block is decoded with respect to the inner code, to produce a q-ary vector $\mathbf{y}$ of length n_2. The algorithm returns the set L of codewords $\mathbf{c} \in C_O$ at a distance $d(\mathbf{c},\mathbf{y}) \leq D$, for some decoding threshold D.

Let F be the number of positions where inner decoding is incorrect. Clearly, $F \sim \mathcal{B}(n,\epsilon_{\text{in}})$. The pirates match $\mathbf{y}$ in at least $(n-F)/t$ positions on average, which means that if $F \leq tD-(t-1)n_2$, then at least one guilty pirate is caught. The following theorem follows by the Chernoff bound.

Theorem 3
Using a concatenated code of an (n_1,q) t-secure inner code with ϵ_{in}-error, and an (n_2,M) outer code, with outer list decoding with threshold $D=n_2\Delta$, the probability

of identifying no guilty user is

$$\epsilon_1 \leq P(F \geq (1-t+t\Delta)n_2), \quad F \sim \mathcal{B}(n_2, \epsilon_{\text{in}}),$$

and

$$\epsilon_1 \leq 2^{-n_2 D(1-t+t\Delta||\epsilon_{\text{in}})}, \quad \text{if} \quad \epsilon_{\text{in}} < 1-t+t\Delta.$$

Corollary 1
If $D(1-t+t\Delta||\epsilon_{\text{in}}) > 0$, then the probability of Type I error tends to zero with increasing code length n_2.

Note that the bound on ϵ_1 is valid for any codes, and it depends only on n_2, Δ, t, and ϵ_{in}. The Type II error rate ϵ_2 will depend on the design of the outer code.

4.2 Random codes (RC)

Random codes for fingerprinting were introduced in [5], and they are used as outer codes in BS-CS. Let C_O be a $(n_2, M)_q$ code, where each symbol in each codeword is chosen uniformly at random from the alphabet.

The security of random codes for fingerprinting depends on the random code being kept secret, which gives a large key for the vendor to store. Thus the key for the random code scheme is $M \cdot n_2 \cdot \log q$ bits, not counting the keys required by the inner code.

Theorem 4
If a random code is used as outer code for concatenation and $1/q < 1-\Delta$, the probability of including a given innocent user $\mathbf{c}$ in the output list is bounded as

$$P(\mathbf{c} \in L) \leq 2^{-n_2 D(1-\Delta||1/q)},$$

and the total Type II error rate is bounded as

$$\epsilon_2 \leq 2^{n_2(R_O \log q - D(1-\Delta||1/q))}.$$

Proof: Consider the output $\mathbf{y}$ from inner decoding and an innocent user $\mathbf{c} \not\in P$. Let $X = n_2 - d(\mathbf{c}, \mathbf{y})$. Clearly X is a stochastic variable with distribution $B(n_2, 1/q)$, and $P(\mathbf{c} \in L) = P(X \geq n_2 - D)$. The error probability is bounded as

$$\epsilon_2 \leq \sum_{\mathbf{c} \in C \setminus P} P(\mathbf{c} \in L) \leq M \cdot P(X \geq n_2(1-\Delta)),$$

and the theorem follows by Chernoff's bound. □

Corollary 2
The Type II error rate tends to zero with increasing length if $R_O < D(1-\Delta||1/q)/\log q$.

One great advantage of random codes is that they can be made for any number of users quite trivially. Observing the error bounds, we note that ϵ_1 is unaltered, and ϵ_2 degrades gracefully when M increases.

4.3 Replication scheme with random codes

Suppose we use an (n_1, q) BS-RS as an inner code, as Boneh and Shaw suggested. Let r denote the replication factor, such that $n_1 = r(2t-1)$. This scheme has several control parameters which may be used to tune the performance of the system. The inner code cardinality q is the trickiest one. Most of the time we will follow Boneh and Shaw and set $q = 2t$. Obviously n_2 and r control a trade-off between code length and error rate. Finally, we have Δ to control the trade-off between the two error types.

Theorem 5
If we use

$$q = 2t, \quad \Delta = \frac{t}{t+1}, \quad \epsilon_{\text{in}} = \frac{1}{2t},$$

then RS-RC is a t-secure fingerprinting scheme with ϵ-error accomodating M users requiring length

$$n = (2t-1)\left\lceil 8t^2(3+2\log t)\right\rceil n_2,$$

where

$$n_2 = \frac{\max\{-\log\epsilon_1, \log M - \log\epsilon_2\}}{D(\frac{1}{t+1}||\frac{1}{2t})}.$$

Asymptotically, the length is

$$n = \Theta\big(t^4(\log t)(\log M - \log\epsilon)\big).$$

In this theorem, Δ is made only slightly greater than the minimum value of $(t-1)/t$. By Corollary 1 we require $\epsilon_{\text{in}} < 1/(t+1)$, but to make n_2 linear in t, ϵ_{in} must in fact be much smaller than $1/(t+1)$.
Proof: Theorems 3 and 4 give two bounds on n_2, so we get

$$n_2 = \max\left\{\frac{-\log\epsilon_1}{D(\frac{1}{t+1}||\frac{1}{2t})}, \frac{\log M - \log\epsilon_2}{D(\frac{1}{t+1}||\frac{1}{2t})}\right\}.$$

It can be shown that $D(1/(t+1)||1/(2t)) = \Theta(t^{-1})$, and hence

$$n_2 = \Theta(t(\log M - \log\epsilon)).$$

For the inner code, we have

$$n_1 = (q-1)2q^2(\log(2q) - \log\epsilon_{\text{in}}) = (2t-1)8t^2(3+2\log t) = \Theta(t^3\log t).$$

The theorem follows since $n = n_1 n_2$. □

For comparison, we include the original theorem from [4].

$t = \log M$	BS-CS	RS-RC
10	$6.64 \cdot 10^8$	$3.14 \cdot 10^8$
15	$3.91 \cdot 10^9$	$1.82 \cdot 10^9$
20	$1.40 \cdot 10^{10}$	$6.56 \cdot 10^9$
25	$3.80 \cdot 10^{10}$	$1.80 \cdot 10^{10}$
30	$8.68 \cdot 10^{10}$	$4.15 \cdot 10^{10}$

Table 1: Some lengths when $t = \log M$.

Theorem 6 (Boneh and Shaw)
BS-CS with replication factor r and $q = 2t$ users for the inner code, is a t-secure (n, M) code with ϵ-error, where

$$n_2 = \left\lceil 2t \log \frac{2M}{\epsilon} \right\rceil, \quad r = \left\lceil 8t^2 \log \frac{8tn_2}{\epsilon} \right\rceil,$$

$$n = n_2 r(2t-1) \approx 16t^3(2t-1)\Big(\log \frac{2M}{\epsilon}\Big)\Big(\log \frac{8tn_2}{\epsilon}\Big).$$

The decoding complexity was $\Theta(n+M)$.

The most interesting point in BS-CS is that $r = \Theta(\log n_2)$, such that n grows faster than linearly in n_2. Since n_2 depends on M and on ϵ, the length of BS-CS is much more dependent on ϵ and M than is our scheme. In Table 1 we see some real sample lengths for these codes, with our and Boneh and Shaw's formulæ.

Considering asymptotic classes of codes, Δ can be made smaller. The following theorem gives the better rates.

Theorem 7
There exists an asymptotic class of fingerprinting codes with exponentially declining error rate for any rate R satisfying

$$R < \frac{D(\frac{1-2q2^{-r/(2q^2)}}{t}||1/q)}{r(q-1)}, \tag{2}$$

if q and r are natural numbers such that $(1-2q2^{-r/(2q^2)})/t > 1/q$.

Proof: Asymptotically, ϵ_{in} can be taken arbitrarily close to $1-t+t\Delta$, or in other words

$$\Delta \approx \frac{t-1+\epsilon_{\text{in}}}{t} = \frac{t-1+2q2^{-r/2q^2}}{t}.$$

By Theorem 3, the outer rate can be chosen arbitrarily close to $D(1-\Delta||1/q)/\log q$. We get the following component code rates

$$R_O \approx \frac{D(\frac{1-2q2^{-r/2q^2}}{t}||1/q)}{\log q}, \quad R_I = \frac{\log q}{r(q-1)},$$

	RS-RC			BBK	
t	q	r	Rate	C_I	Rate
2	4	238	$2.42 \cdot 10^{-4}$	$(126, 2^{14})$	0.0172
3	5	410	$3.62 \cdot 10^{-5}$	$(2\,046, 2^{7})$	$3.98 \cdot 10^{-4}$
4	7	847	$9.62 \cdot 10^{-6}$	$(32\,766, 2^{10})$	$1.82 \cdot 10^{-5}$
5	9	1457	$3.53 \cdot 10^{-6}$	$(1\,048\,572, 2^{12})$	$4.36 \cdot 10^{-6}$
7	13	3223	$8.04 \cdot 10^{-7}$	$(10^{28} - 1, 2^{12})$	$0.116 \cdot 10^{-8}$

Table 2: Asymptotic rates and maximising values of q and r for the RS-RC codes for some numbers of pirates.

which gives the total rate as stated in the theorem. □

In Table 2, we can see some asymptotic rates for our codes. The BBK codes given are the best we could find using constructible inner codes from the literature, namely duals of BCH codes [17]. Better codes are known to exist but they have not been constructed yet. We can see that BBK is better for few pirates, but for larger t we could not find (t,t)-separating codes which are good enough. It is also interesting to note that $2t$ is not the maximising value of q asymptotically, except for $t = 2$.

5 Fighting two pirates

We mentioned that the BS replication codes may not be the ideal choice for inner codes. For two pirates we have good alternatives, which we consider now.

Definition 3
A (t,u)-separating code or (t,u)-SS has the property for any two disjoint sets $\boldsymbol{T}$ and $\boldsymbol{U}$ of respectively t and u codewords, there is one coordinate position where every codeword of $\boldsymbol{T}$ is di*ff*erent from any codeword of $\boldsymbol{U}$.

Separating codes have been applied in various fields for more than three decades, see [15] for a survey. It is known that any $(2,2)$-SS is 2-secure with 1/3-error [1], and that the $[126,14]$ punctured dual of the two-error correcting BCH code is $(2,2)$-separating [8].

Of course, an error rate of $1/3$ in the inner code is a lot, but with proper threshold Δ this may be compensated. Furthermore 2^{14} codewords means that $1/q$ in the calculation of ϵ_2 is very small.

Theorem 8
By concatenating the $[126,14]$ punctured dual of the two-error-correcting BCH code with a random code, we get an infinite class of 2-secure codes with ϵ-error and rate $\boldsymbol{R}$, for any $\boldsymbol{R} < 0.0297$ and exponentially declining error rates given as

$$\epsilon_1 \le 2^{-n\frac{D(2\Delta-1||1/3)}{126}} \quad \text{and} \quad \epsilon_2 \le 2^{n(R-D(1-\Delta||2^{-14})/126)},$$

$\log M$	BS-CS	RS-RC	Simplex	SX-RC
10	759 330	299 889	1 023	1 305
15	848 085	334 359	32 767	1 455
20	937 440	367 359	1 048 575	1 545
25	1 026 684	401 001	$2^{25}-1$	1 605
30	1 116 408	435 471	$2^{30}-1$	1 695

Table 3: Code lengths against two pirates for 1000 to a billion users and error rate $\epsilon \leq 10^{-10}$.

where Δ may be chosen freely in the interval $2/3 < \Delta < 1 - 2^{-14}$.

The best asymptotic rate offered in [1] was 0.015, and [16] offers a rate of 0.026, so we have an improvement. Similarly, any $(3,3)$-SS is 3-secure with $4/7$-error, and using the $(4092, 2^{12})$ subcode of the dual of BCH(3) presented in [17] we can construct an asymptotic class of codes which are 3-secure with ϵ-error and rate $2.74 \cdot 10^{-4}$, where ϵ vanishes.

Another possible choice is to use simplex codes as analysed in [11], where it was shown that the $[2^k - 1, k]$ simplex code is 2-secure with ϵ-error where $\epsilon \leq 2^{k-2^{k-1}}$. We introduce the SX-RC scheme, with the $[15, 4, 8]$ codes as inner codes, random codes for outer codes, and list decoding.

Theorem 9
The SX-RC scheme forms an infinite class of 2-secure codes with ϵ-error and rate R, for any $R < 0.062$, and exponentially declining error rates given as

$$\epsilon_1 \leq 2^{-n\frac{D(2\Delta-1||1/16)}{15}} \quad \text{and} \quad \epsilon_2 \leq 2^{n(R-D(1-\Delta||1/16)/15)},$$

where Δ may be chosen freely in the interval $17/32 < \Delta < 15/16$.

Corollary 3
The SX-RC codes are probabilistically t-secure with length

$$n = 15\left\lceil \max\left\{ \frac{\log \epsilon_1}{D(2\Delta-1||1/16)}, \frac{\log \epsilon_2 - \log M}{D(1-\Delta||1/16} \right\} \right\rceil,$$

for any Δ such that $17/32 < \Delta < 15/16$.

This is a second improvement on the record code rate in the two-pirate case. In Table 3, we present code lengths for 1000 to a billion users with the schemes we know. The RS-RC codes are computed with $q = 2t$, $\epsilon_{\text{in}} = 0.002$, and $\Delta = 0.525$. Here there is probably room for improvement. The error rates were set such that both ϵ_1 and ϵ_2 both are less than $10^{-10}/2$. We used $\Delta = 0.655$ for 2^{10} users, $\Delta = 210/320$ for 2^{15} users, $\Delta = 52/80$ for 2^{20} users, $\Delta = 41/64$ for 2^{25} users, and $\Delta = 203/320$ for 2^{30} users.

Unfortunately, [12], [16], and [1] do not give explicit formulæ for the length for a given, finite code size, and therefore these three schemes are not represented in our table. Note that the simplex codes will have much better error rate than the 10^{-10} that we require.

6 Conclusion and open problems

We have made a new error analysis of the Boneh-Shaw fingerprinting scheme, and proved that it actually exhibits some of the advantages introduced by 'improving' schemes in recent years. It yields asymptotic classes of codes with constant rate and exponentially declining error rate. The length of the codes can be made significantly shorter than previously proved. The Boneh-Shaw style codes also have the advantage that they can be constructed easily for any number of users, any number of pirates, and any error rate.

Using list decoding facilitates the error analysis, in addition to making it possible to trace more than one pirate most of the time. Either inner codes or outer codes may be replaced, and modifying the error analysis should be fairly easy. It is particularly interesting to make constructions with AG codes with long distance as outer codes, for which list decoding can be done in time linear in n. The problem with such constructions is that they require larger alphabets than do random codes, at least $q > t^2$, and thus they are not very efficient with BS-RS as inner codes.

We have pointed out the control parameters in the scheme, and these may be used to tune the performance of the scheme to actual applications. Good and general statements on optimal choices of control parameters is still an open problem.

In the two-pirate case, we replaced the original inner codes by simplex codes in order to get a further improvement. This gave the impressive length of only 1695 for one billion users at an error rate of 10^{-10}. It is probably possible to make similar improvements against three pirates by using the scattering codes and simplex codes from [18] as inner codes. Both BS-CS and RS-RC use q-secure inner codes with $q = 2t$ when only a t-secure code is needed. A most interesting open problem is to construct finite $(n, q)_2$ t-secure codes for $q >> t > 3$, which can be used as inner codes and improve the overall rate.

Barg, Blakley, and Khabatiansky [1] ask whether it is possible to compute a channel capacity for the fingerprinting problem. As we are able to construct schemes with higher and higher rates, it is of increasing interest to know the theoretical capacity limit.

7 Acknowledgements

Thanks to dr. Christian Elsholz for his recommendation to read [10], and to prof. Tor Helleseth for his advice and encouragement.

References

[1] A. Barg, G. R. Blakley, and G. A. Kabatiansky. Digital fingerprinting codes: Problem statements, constructions, identification of traitors. *IEEE Trans. Inform. Theory*, 49(4):852–865, April 2003. 1, 3, 3.2, 5, 5, 5, 6

[2] G. R. Blakley, C. Meadows, and G. B. Purdy. Fingerprinting long forgiving messages. In *Advances in cryptology—CRYPTO '85 (Santa Barbara, Calif., 1985)*, volume 218 of *Lecture Notes in Comput. Sci.*, pages 180–189. Springer, Berlin, 1986. 1

[3] Dan Boneh and James Shaw. Collusion-secure fingerprinting for digital data. In *Advances in Cryptology - CRYPTO'95*, volume 963 of *Springer Lecture Notes in Computer Science*, pages 452–465, 1995. 1

[4] Dan Boneh and James Shaw. Collusion-secure fingerprinting for digital data. *IEEE Trans. Inform. Theory*, 44(5):1897–1905, 1998. Presented in part at CRYPTO'95. 1, 3, 4.3

[5] Yeow Meng Chee. *Turán-type problems in group testing, coding theory and cryptography*. PhD thesis, 1996. 4.2

[6] B. Chor, A. Fiat, and M. Naor. Tracing traitors. In *Advances in Cryptology - CRYPTO '94*, volume 839 of *Springer Lecture Notes in Computer Science*, pages 257–270. Springer-Verlag, 1994. 1

[7] B. Chor, A. Fiat, M. Naor, and B. Pinkas. Tracing traitors. *IEEE Trans. Inform. Theory*, 46(3):893–910, May 2000. 1

[8] Gérard D. Cohen, Sylvia B. Encheva, Simon Litsyn, and Hans Georg Schaathun. Intersecting codes and separating codes. *Discrete Applied Mathematics*, 128(1):75–83, 2003. 5

[9] Venkatesan Guruswami and Madhu Sudan. Improved decoding of Reed-Solomon and algebraic-geometry codes. *IEEE Trans. Inform. Theory*, 45(6):1757–1767, 1999. 3.2

[10] Torben Hagerup and Christine Rüb. A guided tour of Chernoff bounds. *Information Processing Letters*, 33:305–308, 1990. 2, 7

[11] J Herrera-Joancomarti and Josep Domingo-Ferrer. Short collusion-secure fingerprints based on dual binary Hamming codes. *Electronics Letters*, 36:1697–1699, September 2000. 3.3, 5

[12] Tri Van Le, Mike Burmester, and Jiangyi Hu. Short c-secure fingerprinting codes. In *Proceedings of the 6th Information Security Conference*, October 2003. Available at `http://websrv.cs.fsu.edu/~burmeste/`. 3.2, 5

[13] Birgit Pfitzmann and Michael Waidner. Anonymous fingerprinting. In *Advances in cryptology—EUROCRYPT '97*, volume 1233 of *Lecture Notes in Comput. Sci.*, pages 88–102. Springer, Berlin, 1997. 1

[14] Reihaneh Safavi-Naini and Yejing Wang. Sequential traitor tracing. In *Advances in cryptology—CRYPTO 2000 (Santa Barbara, CA)*, volume 1880 of *Lecture Notes in Comput. Sci.*, pages 316–332. Springer, Berlin, 2000. 1

[15] Yu. L. Sagalovich. Separating systems. *Problems of Information Transmission*, 30(2):105–123, 1994. 5

[16] Hans Georg Schaathun. Fighting two pirates. In *Applied Algebra, Algebraic Algorithms and Error-Correcting Codes*, volume 2643 of *Springer Lecture Notes in Computer Science*, pages 71–78. Springer-Verlag, May 2003. 3.3, 5, 5

[17] Hans Georg Schaathun and Tor Helleseth. Separating and intersecting properties of BCH and Kasami codes. Springer Lecture Notes in Computer Science. Springer-Verlag, December 2003. To be presented at IMA'03, Cirencester, Dec. 2003. 4.3, 5

[18] Francesc Sebé and Josep Domingo-Ferrer. Scattering codes to implement short 3-secure fingerprinting for copyright protection. *Electronics Letters*, 38:958–959, August 2002. 3.3, 6

[19] Neal R. Wagner. Fingerprinting. In *Proceedings of the 1983 Symposium on Security and Privacy*, 1983. 1