DIMACS Technical Report 2001-52
December 2001

Digital fingerprinting codes: Problems statements, constructions, identification of traitors [1]

by

Alexander Barg [2]
Math. Research
Bell Labs, Lucent Technologies
700 Mountain Ave.,
Murray Hil,. NJ 07974, USA
on leave from IPPI RAN, Moscow
E-mail abarg@research.bell-labs.com

G. Robert Blakley
Department of Mathematics
Texas A&M University
College Station, TX 77843-3368, USA
E-mail blakley@math.tamu.edu

Gregory Kabatiansky [3,4]
Project CODES
INRIA - Rocquencourt, France
on leave from IPPI RAN
Bol'shoj Karetnyj 19
Moscow 101447, Russia
E-mail kaba@iitp.ru.

[1]Presented in part at the IEEE International Symposium on Information Theory, June 2001, Washington, DC and at the Workshop on Coding and Cryptography, National University of Singapore, September 2001.
[2]Permanent Member
[3]Part of this research done while visiting DIMACS Center, Rutgers University, Piscataway, NJ, in June 2001.
[4]Supported in part by the Russian Foundation for Fundamental Research, Grant No. 99-01-00828.

DIMACS is a partnership of Rutgers University, Princeton University, AT&T Labs-Research, Bell Labs, NEC Research Institute and Telcordia Technologies (formerly Bellcore).

DIMACS was founded as an NSF Science and Technology Center, and also receives support from the New Jersey Commission on Science and Technology.

ABSTRACT

We consider a general fingerprinting problem of digital data under which coalitions of users can alter or erase some bits in their copies in order to create an illegal copy. Each user is assigned a fingerprint which is a word in a fingerprinting code of size M (the total number of users) and length n. We present binary fingerprinting codes secure against size-t coalitions which enable the distributor (decoder) to recover at least one of the users from the coalition with probability of error $\exp(-\Omega(n))$ for $M = \exp(\Omega(n))$. This is an improvement over the best known schemes that provide the error probability no better than $\exp(-\Omega(n^{1/2}))$ and for this probability support at most $\exp(O(n^{1/2}))$ users. The construction complexity of codes is polynomial in n. We also present versions of these constructions that afford $\mathsf{poly}(n) = \mathsf{polylog}(M)$ identification algorithms of a member of the coalition, improving over the best previously known complexity of $\Omega(M)$.

For the case $t = 2$ we construct codes of exponential size with even stronger performance, namely, the distributor can either recover both users from the coalition with probability $1 - \exp(\Omega(n))$, or identifies one traitor with probability 1.

1 Introduction

Let $\mathcal{Q}$ be a finite alphabet of size Q. Suppose a dealer D distributes copies of a long string Σ over $\mathcal{Q}$ available by subscription to registered users of the system. Let M be the total number of registered users. A copy assigned to the i-th user contains a substring $\mathbf{x}(i) = (x_1(i), \ldots, x_n(i))$ (a fingerprint); fingerprints of different users are different. The goal of inserting the fingerprint is to personalize the copy given out to the user, and to rule out redistribution. Clearly, an individual user cannot resell his copy of Σ without running the risk of being tracked down. However, several users may collude in order to produce an unregistered copy. In doing so they face the problem of supplying it with a fingerprint, which should, besides being different from each of their fingerprints, prevent D from identifying members of the colluding group. Thus the distributor D faces the problem of constructing a large set of fingerprints (a fingerprinting code) that enables him to locate at least one member of the colluding group. The algorithm of creating registered fingerprints is parameterized by a secret key k. We assume that the algorithm is publicly known; however, the particular value of k is kept secret by the distributor.

A group $U = \{u_1, \ldots, u_t\}$ of t registered users that intend to produce an illegal copy of Σ will be called a *coalition*. The goal of the coalition U is to create a fingerprint $\mathbf{y} \in \mathcal{Q}^n$ of the illegal copy so that D is unable to identify users from U. Following [5] we assume that the members of U can only alter those coordinates of the fingerprint in which at least two of their fingerprints differ, and refer to this as the *Marking Assumption* (for more detailed discussion and motivation see [5]). Thus it is known a priori that $y_m = x_m(u)$ for every $u \in U$ unless there is a pair $u, u' \in U$ such that $x_m(u) \neq x_m(u')$. What the users are allowed to do in the last case is a part of the formal description of the problem, made more precise later.

The distributor faces the task of identifying one or more members of the coalition U of traitors provided that $|U| = t$, where t is a parameter. The fingerprinting problem thus is to design a set $C = \{\mathbf{c}(1), \ldots, \mathbf{c}(M)\}$ (a t-fingerprinting code, or more precisely, an ensemble of codes) in such a way that no matter which coalition U of at most t users collude to produce an unregistered fingerprint and no matter which algorithm the members of U use, the distributor is always capable of tracing at least one of its members. This problem has been studied in different versions [2], [4]-[8], [12], [22]; [7], [8], see esp. [5] for an expanded informal discussion of the problem and literature overview. As customary in information and coding theory, we will address the existence question of families of t-fingerprinting codes with the number of codewords growing exponentially with the length of the code. In other words, to accommodate M users the distributor D inserts order $\log M$ fingerprint digits. The answer depends on the rules that the members of U (parents) are allowed to follow to create $\mathbf{y}$ (a descendant). As already remarked, when all the values $x_m(u)$ coincide, then y_m must be of the same value. If some $x_m(u)$ differ, then the following two basic strategies employed by the users appear in the literature. Under the narrow-case fingerprinting problem every coordinate of $\mathbf{y}$ can be chosen from the set of entries its parents have in this coordinate [2], [6]-[8], [12]. Under the general or wide-case problem, once there is a choice for y_m, it can be any letter from the alphabet. Both cases can be expanded by allowing the users to make some symbols unreadable, or erased. In particular, [5] is concerned with expanded wide-case problem; below we show that this case is equivalent to the (nonexpanded) wide case.

The narrow case fingerprinting problem was expressly formulated in [12], were it was also proved that for $t = 2$ and $Q \geq 3$ there exist codes with exponentially many codewords. For arbitrary $t \leq Q + 1$ this was proved in [2].

In the wide case it turns out [5] that exact identification of even one parent is generally impossible. Therefore we will allow some error rate, *i.e.*, with low probability ϵ the decision taken by the distributor can be wrong. Best known results in this problem were accomplished in [5], where trade-off between the size of a fingerprinting code and the probability ϵ was considered. The best decline rate of the probability ϵ accomplished in [5] is $\epsilon = \exp(-\sqrt{n})$, irrespective of the code size (*e.g.*, even for very small codes, for instance, of constant size). The question of attaining exponential code size and exponential decline of error rate with the length of the code n, natural from information-theoretic point of view, was left open in [5].

Furthermore, the construction of [5] in the part of identifying a guilty user relies upon complete (maximum likelihood) decoding of a random code. The best known algorithms for this (NP-hard) problem require complexity of order $\Omega(M)$. In this paper we will, relying on coding theory methods, present code families of size $\exp(\Omega(n))$ with $\epsilon = \exp(-\Omega(n))$. Both the complexity of constructing codes and of their decoding grow as $\mathsf{poly}(\log M)$.

In Sect. 2 we will give a precise statement of the problem with the aim of placing it in the standard information-theoretic context. This is worthwhile to do because rigorous statements of the digital fingerprinting problem do not appear in the literature, and moreover, this enables us to establish relations between different versions of the problem. In formulating the problem one faces a number of choices related to the worst vs. average case. While in information theory there is an accepted standard, in the field of digital fingerprinting the situation is unsettled. We give a version of the definition geared toward worst-case performance, which is consistent with the code constructions known in the literature and suggested below, and also corresponds well to the nature of cryptographic problems. We conclude Sect. 2 by discussing previous results and goals of our paper.

One of the new ideas in code construction introduced in this work is the use of separating codes, which we briefly discuss in Section 3, explaining also the reason for them to enter the fingerprinting problem. Section 4 is devoted to constructions of binary fingerprinting codes. The constructions employ concatenation of two codes which proved useful in this problem (see [5]): a long outer code W and a shorter inner code V. The code W is error-correcting with large minimum distance; the code V has a t-separating property (see [13] and the survey [20]) and is used to separate law-abiding users from those committing fraud. The distance properties of W enable us to amplify this separation. Section 5 gives identification algorithms of the distributor of complexity polynomial in the length of the codes constructed. This result is possible due to a surprising recent discovery of decoding algorithms of algebraic geometry ("evaluation") codes that in polynomial time correct far more errors than half the minimum distance of the code [11].

2 Statement of the problem and previous results

2.1 Problem statements

Let us assume some ordering of the users and write them as $\{1, \dots, M\}$. Let $\mathcal{Q}^n$ be the set of all Q-ary words of length n. Any subset $C \subset \mathcal{Q}^n$ is called a code of length n; if $|C| = M$ we say that C is a code of size M.and write $C(n, M)$. The number $R = R(C) := n^{-1} \log_Q M$ is called the *rate* of C. If Q is a prime power and C is a linear subspace of $\mathcal{Q}^n$, it is called a linear $[n, \ell, d]$ code, where $\ell = \dim(C)$ and d is the code distance (sometimes we omit it from the notation). By $\mathrm{dist}(\mathbf{x}, \mathbf{y})$ we denote the Hamming distance between words $\mathbf{x}$ and $\mathbf{y}$. The distance between a word $\mathbf{x}$ and a subset $\mathcal{H} \subset \mathcal{Q}^n$ is defined as $\operatorname{dist}(\mathbf{x}, \mathcal{H}) := \min_{\mathbf{h}\in\mathcal{H}} \operatorname{dist}(\mathbf{x}, \mathbf{h})$.

The distributor D assigns to the user i a fingerprint $\mathbf{x}(i) \in \mathcal{Q}^n$. The set of M fingerprints $\{\mathbf{x}(1), \dots, \mathbf{x}(M)\}$ forms a code $C \subset \mathcal{Q}^n$. As we show later (Propositions 2 and 3), in many important cases using a single code does not enable the distributor to solve the fingerprinting problem, namely, to locate a member of the coalition. Therefore, the distributor D uses a family of codes $\mathcal{C} = \{C_k\}$ of codes, choosing a particular code C_k with probability $\pi(k)$. More specifically, by the code C_k we mean an *ordered* set of M vectors $\{\mathbf{x}(1), \dots, \mathbf{x}(M)\}$, so that $\pi(k) = P_M(\mathbf{x}(1), \dots, \mathbf{x}(M))$. Here k ranges over some set of $\mathcal{K}$ possible keys. In the constructions below we will assume that $\pi(k) = \mathcal{K}^{-1}$. However, in the general problem statement of this section $\pi(k)$ can be an arbitrary distribution.

Having observed a fingerprint $\mathbf{y}$, the distributor D identifies user u as delinquent with some probability that depends both on $\mathbf{y}$ and on the specific code (key) used. Hence the most general decision rule of D can be described by a conditional probability distribution $P_D(u|\mathbf{y}, k)$. The distributions $\pi(k)$ and $P_D(u|\mathbf{y}, k)$ and the family of codes C_k are publicly known. The only information kept secret by D is the specific value of k.

Let $U = \{u_1, \dots, u_t\}$ be a coalition of t users that collude to create an unregistered fingerprint $\mathbf{y}$. Let $X = C_k(U) = (\mathbf{x}^1, \dots, \mathbf{x}^t)$ be the fingerprints assigned to the members of U (for brevity we write $\mathbf{x}^i$ instead of $\mathbf{x}(u_i)$). The word $\mathbf{y}$ is taken from a subset $\mathcal{E}(X) \subset \mathcal{Q}^n$, called the *envelope* of X.

Particular ways to form the envelope are discussed below in more detail; for the moment it can be assumed arbitrary. We assume that the members of U attempt at confusing the distributor by choosing a particular value of $\mathbf{y}$ with some conditional probability $P_U(\mathbf{y}|X)$, where X is the set of their fingerprints. Again, the distribution $P_U(\mathbf{y}|X)$ describes in the most general way the strategy that the members of U use to confuse the distributor.

The distributor's probability of error in identifying a member of U can be written in two ways:

$$p_e(U, P_U) = \sum_{u\notin U} \Big[\sum_{k=1}^{\mathcal{K}} \pi(k) \sum_{\mathbf{y}\in\mathcal{E}(C_k(U))} P_D(u|\mathbf{y}, k) P_U(\mathbf{y}|C_k(U)) \Big]$$

(total probability of identifying user u as guilty for all $u \notin U$), or

$$p_e(U, P_U) = \sum_{k=1}^{\mathcal{K}} \pi(k) \Big[\sum_{u\notin U} \sum_{\mathbf{y}\in\mathcal{E}(C_k(U))} P_D(u|\mathbf{y}, k) P_U(\mathbf{y}|C_k(U)) \Big]$$

(probability of error for a given code averaged over the choice of the code).

Let

$$p_e(U) = \max_{P_U} p_e(U, P_U)$$

be the error probability under the *optimal* strategy of the coalition (note that Definition IV.2 in [5] implies a weaker attack, namely, only the uniform distribution on $\mathcal{E}(C_k(U))$. Thus the maximum probability for D to incriminate an user that is not a member of U, equals

$$p_e = p_e(\{C_k\}, \pi(\cdot), P_D(\cdot)) = \max_{U,\ |U|=t} p_e(U).$$

This quantity still depends on the strategy $P_D(u|\mathbf{y}, k)$ of D, the choice of the family of codes $\{C_k\}$ and the probability $\pi(k)$ to choose a particular code C_k. It is natural to assume that the distributor performs these choices to minimize the value of p_e. The resulting probability is a function of the quadruple $(n, M, t; \mathcal{K})$ and the type of $\mathcal{E}$, and equals

$$\epsilon(\mathcal{E}; n, M, t; \mathcal{K}) = \min_{\{C_k\}, \pi(\cdot), P_D(\cdot)} p_e(\{C_k\}, \pi(\cdot), P_D(\cdot)). \tag{1}$$

If the number of keys (codes in the family $\{C_k\}$) is unrestricted, then we can introduce the value

$$\epsilon(\mathcal{E}; n, M, t) = \min_{\mathcal{K}} \epsilon(\mathcal{E}; n, M, t; \mathcal{K}) \tag{2}$$

Note that the total number of codes of size M is bounded above by 2^{nM}, which is also an upper bound on the number $\mathcal{K}$ of keys in this formula.

The general *fingerprinting problem* is to find $\epsilon(\mathcal{E}; n, M, t)$ and codes and the optimal strategy of D that achieve this error rate.

For a given $X \subset \mathcal{Q}^n$ we now define the envelope $\mathcal{E}(X)$. If $\mathbf{y} \in \mathcal{E}(X)$ then $\mathbf{y}$ is called in the literature a *descendant* of X and any $\mathbf{x} \in X$ is called a *parent* of $\mathbf{y}$. Several definitions of $\mathcal{E}$ have been discussed. We consider the narrow-sense and wide-sense envelopes. The *narrow-sense* envelope $e(X)$ is defined as follows:

$$e(X) = \{\mathbf{y} \in \mathcal{Q}^n | y_i \in \{x_i^1, \dots, x_i^t\}\}.$$

The fingerprinting problem thus defined was studied in [2], [6], [7], [12], [22] (called there codes with the identifying parent property or IPP codes). The *wide-sense* envelope is defined as $E(X) = \{\mathbf{y}\}$, where

$$\begin{aligned} y_i &= x_i^1 \quad \text{if } x_i^1 = x_i^2 = \dots = x_i^t \\ y_i &\in \mathcal{Q} \ \text{(any symbol of the alphabet)} \quad \text{otherwise.} \end{aligned}$$

Both $e(X)$ and $E(X)$ can be generalized to the expanded case under which the coalition is allowed to erase the symbol y_i if $|\{x_i^1, \dots, x_i^t\}| \ge 2$. In particular, paper [5] is concerned with the expanded wide-sense envelope $E^*(X)$. Formally $\mathbf{y} \in E^*(X)$ if

$$\begin{aligned} y_i &= x_i^1 \quad \text{if } x_i^1 = x_i^2 = \dots = x_i^t \\ y_i &\in \mathcal{Q} \cup \{*\} \quad \text{otherwise,} \end{aligned}$$

where $*$ denotes erasure. With some abuse of speech, a family of codes that enables D to identify at least one member of X as long as $|X| \le t$ (possibly, with error probability ϵ) will be called a *t-secure code.*

It is clear that

$$\epsilon(e; n, M, t; \mathcal{K}) \le \epsilon(E; n, M, t; \mathcal{K}) \le \epsilon(E^*; n, M, t; \mathcal{K}) \tag{3}$$

and

$$\epsilon(e; n, M, t; \mathcal{K}) \le \epsilon(e^*; n, M, t; \mathcal{K}) \le \epsilon(E^*; n, M, t; \mathcal{K}). \tag{4}$$

Note that these inequalities are also valid for the probabilities $\epsilon(\cdot; n, M, t)$ from (2).

2.2 Properties of different versions of the problem

In this section we establish some relations between fingerprinting problems formulated above. We begin with a result that shows equivalence of the expanded wide sense problem ($\mathcal{E} = E^*$) and the wide-sense one ($\mathcal{E} = E$).

Proposition 1 *For any family of codes $\{C_k\}$ and any probability distribution $\pi(k)$, the error rate of the distributor $\min_{P_D} p_e(\{C_k\}, \pi(k), P_D)$ optimized over its strategies $P_D(\cdot)$ is the same for the expanded fingerprinting problem and the wide-sense one.*

Note: This result contradicts the claim made in [5], p.1900, which is therefore invalid.
Proof : Clearly, the error rate achievable for the nonexpanded problem does not exceed the error rate for the expanded case. To show the converse, consider an arbitrary family of codes $\{C_k\}$ for the E-case together with some probability distribution $\pi(k)$ defined on it. Let $P_D(u|\mathbf{y}, k)$ be the optimal decision rule for these codes $\{(C_k)\}$ and this $\pi(k)$. Then for the expanded case, define the decision rule $P_D^*(\cdot)$ as follows: $P_D^*(u|\mathbf{y}', k) = P_D(u|\mathbf{y}, k)$, where $y_i = \alpha$ if $y_i' = *$ and $y_i = y_i'$ otherwise, and where α is a fixed element of the alphabet. The transformation $\tau : \mathbf{y}' \to \mathbf{y}$ establishes a mapping from the set of strategies $P_U^*(\mathbf{y}'|C_k(U))$ of the coalition U for the expanded wide-sense envelope E^* on the set of strategies for the wide-sense envelope E, defined as follows:

$$P_U(\mathbf{y}|C_k(U)) = \sum_{\mathbf{y}' \in \tau^{-1}(\mathbf{y})} P_U^*(\mathbf{y}'|C_k(U)).$$

Clearly, the error probability p_e of D for the expanded case under the strategy $P_D^*(\cdot)$ and any given strategy of U equals the error probability of D for the non-expanded case under the corresponding strategy of U. This proves the proposition. ∎

The next corollary follows on replacing an arbitrary code family and a probability distribution by the optimal ones for the E-case.

Corollary 1 $\epsilon(E; n, M, t; \mathcal{K}) = \epsilon(E^*; n, M, t; \mathcal{K})$.

Note that for the narrow-sense envelope an analogous claim does not hold since on the one hand, for the non-expanded case there exist t-secure zero-error codes of exponential size (for $t < Q$), and on the other hand, for the expanded case the size of any t-secure code is at most t (Prop 3). The argument of the proof of Prop. 1 fails in this case at the step of setting up the correspondence $\mathbf{y}' \to \mathbf{y}$. Nevertheless, in the binary case, the one most often encountered in the literature, the wide-sense envelope is the same as the narrow-sense one. So all the four statements of the general problem in this case coincide.

Special attention in the problem of digital fingerprinting was devoted to the case of exact recovery ($\epsilon = 0$.) Note that the zero-error requirement is a very stringent condition. Indeed, if a single code C is used and $t \geq Q$ then $|C| \leq Q$ even for $\mathcal{E} = e$, so zero-error t-secure codes of nonzero rate do not exist. If $t \leq Q - 1$, it is possible to construct zero-error t-secure codes of exponential size if $\mathcal{E} = e$ (see [12] for the case $t = 2$ and [2] for arbitrary t).

However, nontrivial zero-error codes do not exist for the three other types of envelopes considered above. Namely, we will establish that for the wide-sense envelope E and for the expanded narrow-sense envelope e^* not only the exact recovery $\epsilon = 0$, but also relatively small error probability ϵ cannot be achieved by a single code.

For a fixed code family $\{C_k\}$ we will write $p_e(\{C_k\})$ to refer to error probability of the distributor in identifying a guilty user under the minimax formulation:

$$p_e(\{C_k\}) = \min_{\pi(k)} \min_{P_D} \max_{U \subset C, |U|=t} p_e(U).$$

We begin with a simple technical remark that for any a wide-sense t-secure code C and its subcode $\mathcal{C} \subset C$

$$p_e(C) \geq p_e(\mathcal{C}). \tag{5}$$

Since the family consists of a single code, there is no distinction between a coalition and a subset of codewords, and the proof of (5) is obvious from the following inequality

$$p_e(C) = \min_{P_D} \max_{U \subset C, |U|=t} p_e(U) \geq \min_{P_D} \max_{U \subset \mathcal{C}, |U|=t} p_e(U) = p_e(\mathcal{C}).$$

Proposition 2 *Let C be a wide-sense t-secure code of size $|C| \geq 2t-1$. Then*

$$p_e(C) \geq \frac{t-1}{2t-1}. \tag{6}$$

Proof : We again identify the users and the codewords assigned to them. Let $\mathcal{C} = \{\mathbf{x}^1, \mathbf{x}^2, \dots, \mathbf{x}^{2t-1}\} \subset C$ be any subcode of C, chosen by D. Let U be a coalition of size t, which creates a fingerprint $\mathbf{y}$ using the following deterministic strategy: the ith coordinate y_i is set to a if there exist t vectors among $\{\mathbf{x}^1, \mathbf{x}^2, \dots, \mathbf{x}^{2t-1}\}$ whose ith coordinates equal a; otherwise, put $y_i = \alpha$, where α is some fixed element of the alphabet. It is easy to see that $\mathbf{y} = (y_1, \dots, y_n) \in E(U)$ for any $U \subset \mathcal{C}$, $|U| = t$. Let us restrict ourselves only to such coalitions; this can only increase the chances of D for successful identification.

In this situation, the best strategy of the distributor D is to choose u randomly from $\mathcal{C}$ with uniform probability, which ensures the distributor's success rate of $t/(2t-1)$. Indeed, consider any strategy of D, *i.e.*, a set of probabilities $P_D(u = \mathbf{x}^i|\mathbf{y}), 1 \leq i \leq 2t+1$. Suppose without loss of generality that

$$P_D(\mathbf{x}^1|\mathbf{y}) \leq P_D(\mathbf{x}^2|\mathbf{y}) \leq \dots \leq P_D(\mathbf{x}^{2t-1}|\mathbf{y}).$$

Then the probability of identifying a member of the coalition $U = (\mathbf{x}^1, \dots, \mathbf{x}^t)$ is

$$P_D(\mathbf{x}^1|\mathbf{y}) + \dots + P_D(\mathbf{x}^t|\mathbf{y}) \leq \frac{t}{2t-1}.$$

Hence,

$$p_e(\mathcal{C}) \geq 1 - \frac{t}{2t-1} = \frac{t-1}{2t-1}.$$

Conclude by (5). ∎

Corollary 2

$$\epsilon(E; n, M, t; \mathcal{K}) \geq \frac{t-1}{\mathcal{K}(2t-1)}.$$

Proof. Let $\{C_k, 1 \le k \le \mathcal{K}\}$ be a family of wide-sense t-secure codes. There is a key k_0 such that its probability $\pi(k_0) \ge \mathcal{K}^{-1}$. Consider the same strategy of the coalition as in the proof of Proposition 2, assuming that $k = k_0$. Then

$$P_e \ge \pi(k_0) P_e(C_{k_0}) \ge \frac{t-1}{(2t-1)\mathcal{K}}.$$

∎

Below we do not consider codes with $\le t$ codewords, which we call trivial. The proof of Proposition 2 also shows that for any nontrivial code C, the error rate $p_e(C) \ge 1/(t+1)$. Thus, we obtain the following result.

Corollary 3 *Let C be a wide-sense t-secure zero-error code. Then $|C| \le t$.*

Recall a result in [5] which shows that zero-error recovery is impossible with a single binary wide-sense t-secure code C with $|C| \ge 3$. The above sequence of results develops this by establishing a limiting tradeoff between the error rate and the size of the family for any alphabet size Q; in particular for a single code we have the estimate (6).

The same reasoning can be applied to the expanded narrow-sense problem.

Proposition 3 *The results of Proposition 2 and Corollaries 3 and 2 are valid if the wide-sense problem is replaced by the expanded narrow-sense problem.*

Proof : Let $\mathcal{C} = \{\mathbf{x}^1, \dots, \mathbf{x}^{2t-1}\}$ be a subcode of the code C. Let us modify the strategy of the coalition defined in the proof of Proposition 3, replacing α with $*$ and leaving the rest unchanged. Now the argument of that proof can be applied verbatim to the case considered. ∎

Thus we arrive at the following important conclusion. Both for the wide sense and expanded narrow-sense problem we have

$$\mathcal{K} \ge \frac{1}{(t+1)\epsilon}.$$

In other words: *to achieve exponential decline of the error rate the number of keys has to grow exponentially with the length of the fingerprint.* Note that this growth order of the number of keys will be achieved in the constructions of Sect. 4-5. Note also that the construction of [5] requires that $\mathcal{K} = \Omega(2^{nM})$, *i.e.*, that $\log \mathcal{K} = \exp(\Omega(n))$ for $M = \exp \Omega(n)$.

2.3 Previous results

The fingerprinting problem with expanded wide-sense envelope was first considered by Boneh and Shaw [5]. It is proved in [5] that it is not possible to solve the fingerprinting problem by using a fixed assignment of fingerprints to the users and suggested instead to use random choice of a code from a given code family. This enabled the authors of [5] to construct families of t-secure binary ($Q = 2$) fingerprinting codes. This construction is the best known in the literature from the point of view of code parameters. The codes have length n and size M and provide error probability $\epsilon > 0$, where $n = O(t^4(\log M + \log(1/\epsilon)) \log(1/\epsilon))$ [5, Thm.V.5]. Solving this for ϵ and putting $\log M = O(n^\alpha)$, we obtain $\log \epsilon = -\Omega(\min(\sqrt{n}, n^{1-\alpha}))$. Thus, in the construction of [5] ϵ cannot decrease faster than $\exp(-\Omega(\sqrt{n}))$ and then $M = 2^{O(\sqrt{n})}$. On the other hand, as proved in [5, Thm. VI.1], for any family of codes

$$\epsilon \ge \frac{1}{t} \exp(-\frac{2}{t-3} n),$$

and so it might be possible to construct codes with error probability falling exponentially with n. For this reason [5] puts forward the question of tightening the gap between the known bounds and constructions. The main goal of this paper is to resolve this question by constructing codes with exponentially small probability of identification error and the number of words growing exponentially with n, and to show that this performance is attainable with polynomial complexity of code construction and decoding.

Another parameter of the problem whose importance has not been singled out in previous works, but was emphasized in this section, is the number of keys in the fingerprinting scheme. It is desirable to have as few keys as possible because the distributor must store the key bits in order to manage the system. The best known scheme [5] does not invoke this parameter explicitly, referring instead to a random code of length $\Omega(n)$ and size M. Suppose that $M = \exp(\Omega(n))$, then the total number of keys is $\mathcal{K} = \Omega(2^{nM})$, *i.e.*, $\log \mathcal{K} = \exp(\Omega(n))$, and hence D must store $\exp(\Omega(n))$ bits of information. In contrast, we will obtain $\log \mathcal{K} = \Omega(n)$, which is the best possible order of magnitude.

3 Separating codes

Below we focus on the binary wide-sense fingerprinting problem. The fingerprinting codes that we construct are formed by concatenation of error-correcting codes and sets of binary vectors with a special property known as separating codes.

Generally, a Q-ary code C is called (t, t')-separating [13] if for any two disjoint subsets $X \subset C$, $Y \subset C$ such that $|X| = t, |Y| = t'$ there holds

$$e(X) \cap e(Y) = \emptyset.$$

In other words, for any two subsets X, Y with $|X| = t, |Y| = t'$ there is a coordinate i such that

$$\{x_i, \mathbf{x} \in X\} \cap \{y_i, \mathbf{y} \in Y\} = \emptyset.$$

Below we assume that $t + t' \geq 3$ because the case $t = t' = 1$ is trivial.

Separating codes were studied in [13], [17]-[20], [14], and also under the name of secure frameproof codes in [23] and of partially identifying codes in [8].

In this section we establish relationships between separating and fingerprinting. For a while we return to the general fingerprinting problem and will consider different models of envelopes.

Lemma 1 *For any single code C which is not (t, t)-separating, the identification error probability $p_e(C) \geq 1/2$ for the narrow-sense as well as the wide-sense envelope, expanded or not.*

Proof. If the (t, t)-separating property does not hold, then there exist two disjoint subsets $U, U' \subset C$ such that $|U| = t$, $|U'| = t$, and $e(U) \cap e(U') \neq \emptyset$. Let $\mathbf{y} \in e(U) \cap e(U')$. Suppose that the strategy of both coalition U and U' consists in choosing $\mathbf{y}$ as the fingerprint. Then irrespective of the strategy of D, the error probability of deciding between U and U', and thus the worst-case error $p_e(C)$ is bounded below by $1/2$. ∎

It was noted in the literature that separation is necessary (but not sufficient) for narrow-sense zero-error identification. In the context of probabilistic fingerprinting this remark is new.

Lemma 2 *For any single (t, t)-separating code C the identification error probability $p_e(C) \leq 1 - 1/t$ for the narrow-sense as well as for the wide-sense envelope, expanded or not.*

Proof : Consider any size-t coalition $U \in C$ and any strategy P_U of generating the fingerprint $\mathbf{y}$. The distributor's probability of correctly identifying a member of U is at least $1/t$, by the following strategy: D finds any coalition U' such that $\mathbf{y} \in E(C(U'))$ and identifies randomly a user $u \in U'$. Since C is (t,t)-separating, we have $|U' \cap U| \geq 1$; thus $u \in U$ with probability $\geq 1/t$. ∎

We are especially interested in the particular case $t = 2$, which received extensive coverage in the literature (see the survey [20]). In this case the identification error probability of any single $(2,2)$-separating (n,M) code C equals $p_e(C) = 1/3$ for narrow- or wide-sense envelope, expanded or not. In the three out of the four cases, this is optimal for a single fingerprinting code.

Lemma 3 *Let C be a $(2,2)$-separating (n,M) code. Then*

$$p_e(C) = \epsilon(e^*; n, M, 2; \mathcal{K}) = \epsilon(E; n, M, 2; \mathcal{K}) = \epsilon(E^*; n, M, 2; \mathcal{K}) = 1/3.$$

PROOF. By Propositions 2 and 3 we have $p_e(C) \geq 1/3$ for $\mathcal{E} = e^*, E$ or E^*. On the other hand, let $\mathbf{y}$ be a fingerprint observed by the distributor. Consider the set of all pairs $\mathbf{x}, \mathbf{x}'$ such that $\mathbf{y} \in E(\mathbf{x}, \mathbf{x}')$. By the $(2,2)$-separating property either all such pairs have a common element, or there are three pairs that form a triangle configuration $\{(\mathbf{x}, \mathbf{x}'), (\mathbf{x}', \mathbf{x}''), (\mathbf{x}'', \mathbf{x})\}$. In the first case D outputs the common element, and then $p_e(C) = 0$. In the last case case the strategy of D is to take a random element from the triple $\mathbf{x}, \mathbf{x}', \mathbf{x}''$, and then $p_e(C) = 1/3$. ∎

Existence bounds on $(2,2)$-separating codes were studied in [13], [17]; the best known bound [17] gives for the asymptotic rate of linear codes the value $R = \ell/n = (3 - \log_2 7)/3 \approx 0.0642$. On the other hand, it is known that asymptotically $R \leq 0.108$ for linear codes [20] and $R \leq 0.283$ for arbitrary codes (see [14], [20]). Paper [20] also contains tables of $(2,2)$-separating $[n,\ell]$ codes. For instance, the well-known $[7,3,4]$ code is readily seen to be $(2,2)$-separating. Below we use a $[35,6]$ code from [20] at the inner level to construct a 2-secure fingerprinting code with very low probability of identification error (see the Example after Theorem 2).

It is clear that binary linear (t,t')-separating codes do not exist for $\max(t,t') > 2$. In the unrestricted case the existence of (t,t')-separating codes is established by a standard probabilistic argument ("random coding with expurgation"); see [19],[14] for $t = t' = 2$. We have

Proposition 4 *Let $T = t + t'$. There exist binary (t,t')-separating codes of length m and size $\frac{1}{2} P_T^{-\frac{m}{T-1}}$, where $P_T = 1 - 2^{-(T-1)}$,* i.e., *codes of rate $R_T = -\frac{\log_2 P_T}{T-1} - 1/m$.*

Proof : Consider a random binary (m,L) code C and compute the expectation $\mathbf{E}$ of the number of pairs of subsets X, Y of the code C, $|X| = t$, $|Y| = t'$ that contradict the (t,t') separating property. Whenever $\mathbf{E} \leq L/2$ then $(m, L/2)$ codes with the (t,t') separating property exist.

The probability that a given pair X and Y violate the separating property is

$$\left(1 - 2^{-(T-1)}\right)^m = P_T^m.$$

The expectation of the number of pairs X, Y that violate the separating property is

$$\mathbf{E} \leq \binom{L}{t}\binom{L-t}{t'} P_T^m.$$

Take

$$L = \left(\frac{t!t'!}{2} P_T^{-m}\right)^{1/(T-1)},$$

then

$$\mathbf{E} < \frac{L^T}{t!t'!} P_T^m = \frac{L}{2}.$$

Finally note that $(t!t'!/2)^{(1/(T-1))} \geq 1$. ∎

Note that [13] gives an asymptotic bound $R_T \geq -\frac{\log_2 P_T}{T}$. which is somewhat weaker than this result.

4 Code constructions

4.1 The case of arbitrary t

We will be concerned with binary codes ($Q = 2$). To construct a family of fingerprinting codes $\{C_k\}$ we use the idea of concatenation. Recall that a binary concatenated code of length mN and size M is formed by a binary inner (m, q) code V and an q-ary outer code W of length N and size M based on a fixed 1-1 mapping of the q-ary alphabet to V. For every codeword $\mathbf{w} \in W$ the word of the concatenated code is obtained by replacing coordinates of $\mathbf{w}$ with the corresponding codewords of V.

Consider as the inner code V a binary (t,t)-separating code of length m and size q. We choose a q-ary $[N, K, D = \delta N]$ linear code as an outer code W. Since we need to obtain a family of codes $\{C_k\}$, our construction also involves N random bijections $\varphi_i : \mathbb{F}_q \to V$ $(i = 1, 2, \ldots, N)$, where $\mathbb{F}_q$ is the field of q elements. Restrictions on the parameters of the codes V and W will become clear when we analyze the construction.

Consider vector mappings $\phi = (\varphi_1, \varphi_2, \ldots, \varphi_N)$, where each $\varphi_i, 1 \leq i \leq N$ is an arbitrary bijection. Let us number these mappings from 1 to $(q!)^N$ in an arbitrary order and write $\phi^{(k)}$ to refer to the kth mapping. A typical codeword of C_k is obtained by taking a codeword $\mathbf{w} = (w_1, \ldots, w_N) \in W$ and computing the binary mN-vector $\phi^{(k)}(\mathbf{w}) = (\varphi_1^{(k)}(w_1), \varphi_2^{(k)}(w_2), \ldots, \varphi_N^{(k)}(w_N))$. The mapping $\phi^{(k)}$ is chosen by D with uniform distribution and constitutes the secret key. The length of the code C_k is mN, the size $M = q^K$. Each user is assigned a fingerprint given by a code vector in C_k. It will be convenient to identify the users with the code vectors of the code W (note that $|W| = |C_k|$). The (t,t)-separating property of V will be essential for the entire scheme.

Assume that a coalition $U = \{\mathbf{u}^1, \ldots, \mathbf{u}^t\} \subset W$ generates a fingerprint

$$\mathbf{y} = (\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_N) \in E(\phi^{(k)}(\mathbf{u}^1), \ldots, \phi^{(k)}(\mathbf{u}^t)),$$

where the subblocks $\mathbf{y}_i$ are length-m binary vectors. The decision algorithm of the distributor D first proceeds with "decoding" of every vector $\mathbf{y}_i$ with the inner code V and then decodes the result with the outer code W.

Algorithm 1:

1. The first stage is decoding of the code V. For a given i take an arbitrary t-set $\{\hat{\mathbf{v}}^1, \ldots, \hat{\mathbf{v}}^t\}$ of code vectors from V that can generate $\mathbf{y}_i$, *i.e.*, such that $\mathbf{y}_i \in E(\hat{\mathbf{v}}^1, \ldots, \hat{\mathbf{v}}^t)$. Define the set $H_i = \{h_i^1, \ldots, h_i^t\}$, where $\varphi_i^{(k)}(h_i^j) = \hat{\mathbf{v}}^j$. The result of the first-stage decoding is given by the set $\mathcal{H} = \{\mathbf{h} \in F_q^N : h_1 \in H_1, \ldots, h_n \in H_N\}$

2. The distributor D identifies a possible member of the coalition U as the user that corresponds to the code vector $\widehat{\mathbf{w}} \in W$ such that

$$s(\widehat{\mathbf{w}}, \mathcal{H}) = \max_{\mathbf{w} \in W} s(\mathbf{w}, \mathcal{H}),$$

where $s(\mathbf{w}, \mathcal{H}) := |\{i : w_i \in H_i\}| = N - \operatorname{dist}(\mathbf{w}, \mathcal{H})$.

We note that this algorithm has two nonstandard features. The first-stage decoding outputs a subset of the words. This differs from the standard decoding algorithm of concatenated codes (which is also Algorithm 2 of [5]), where it is assumed that this output is a single codeword. In the second stage, the algorithm looks for a nearest code vector of W to the subset $\mathcal{H}$ rather than to one "received" word.

The reason that this algorithm and its modifications below will work for the identification problem is that the number of occasions that coordinates of a code vector fall in the set H_i is a random variable whose average is greater when the code vector is a member of the coalition U than when it is not. Hence the probability that the distributor identifies incorrectly a code vector as a member of the coalition falls exponentially with the length N of the code W.

We will use the following standard bounds on large deviations. Let ξ_i be independent Bernoulli random variable equal 1 with probability p and 0 with probability $1-p$. Then the probabilities of the tails can be bounded as

$$\begin{aligned} \Pr\{\sum_{i=1}^{N} \xi_i \le N\sigma\} &\le 2^{-ND(\sigma\|p)} \quad \text{if } p > \sigma \\ \Pr\{\sum_{i=1}^{N} \xi_i \ge N\sigma\} &\le 2^{-ND(\sigma\|p)} \quad \text{if } p < \sigma, \end{aligned}$$

where $D(\sigma\|p) = \sigma \log_2(\sigma/p) + (1-\sigma)\log_2((1-\sigma)/(1-p))$ is the information divergence of two binomial distributions.

Another estimate, which is sometimes more convenient, is given by the Chernoff bounds.

Lemma 4 [1] *Let $X_i, i = 1, 2, \ldots, N$ be independent random variables such that $P(X_i = 1) = p_i$ and $P(X_i = 0) = 1 - p_i$. Let $a > 0$ and $p = (p_1 + \cdots + p_N)/N$. Then*

$$\Pr\{\sum_{i=1}^{N} X_i \ge \sum_{i=1}^{N} p_i + a\} \le e^{-2a^2/N}.$$

and

$$\Pr\{\sum_{i=1}^{N} X_i \le \sum_{i=1}^{N} p_i - a\} \le e^{-a^2/2Np}.$$

Theorem 1 *Let W be a q-ary linear $[N, K, D = \delta N]$ code with*

$$\delta > 1 - \frac{1}{t^2} + \frac{t-1}{t(q-1)}$$

and let V be a (t,t)-separating binary (m,q) code. The family of concatenated codes $\{C_k\}$ with inner code V, outer code W, and the set of all $(q!)^N$ bijections $\phi^{(k)}$, together with the decision rule defined by Algorithm 1, forms a binary fingerprinting code of length $n = mN$ with q^K code vectors (users) that identifies one traitor with error probability

$$p_e \le 2^{-nR(V)[(\log_2 q)^{-1} D(\sigma\| \frac{t-1}{q-1}) - R(W)]}, \tag{7}$$

where $\sigma = 1/t - (1-\delta)t$.

PROOF. Let $U_i = \{u_i^1, \dots, u_i^t\}$ be the multiset of the ith coordinates of the vectors in the coalition U. By $|U_i|$ we denote the number of distinct elements in U_i. Since for the sets H_i and U_i, their images under the bijection $\varphi_i^{(k)}$ can generate the same vector $\mathbf{y}_i$ and since the inner code V is (t,t)-separating, we have $H_i \cap U_i \neq \emptyset$. This implies that $\sum_{j=1}^t s(\mathbf{u}_j, \mathcal{H}) \ge N$, and hence $\max_j s(\mathbf{u}_j, \mathcal{H}) \ge N/t$.

On the other hand, for any $\mathbf{w} \notin U$, the element w_i can fall in the list H_i for one of the two reasons. The first possibility is that $w_i \in U_i$, and the number T of such positions is at most $t(N-D)$ since any two distinct vectors from W coincide in at most $N-D$ positions. The other option is that the mapping of w_i to V matches the random bijection $\varphi_i^{(k)}$. To compute the probability p_i of the last event note that by assumption we have $w_i \notin U_i$, thus $w_i \notin (U_i \cap H_i)$. By assumption the bijections are chosen randomly and uniformly, so we obtain

$$p_i := \Pr\{w_i \in H_i | w_i \notin U_i\} = \frac{t - |U_i \cap H_i|}{q - |U_i|} \le \frac{t-1}{q-1}. \tag{8}$$

Now let us bound above the probability that $s(\mathbf{w}, \mathcal{H}) \ge N/t$ for a particular vector $\mathbf{w}$. Let ξ_i be independent Bernoulli random variables equal 1 with probability p_i and 0 with probability $1 - p_i$. Then for $\mathbf{w} \notin U$

$$\begin{aligned} \Pr\{s(\mathbf{w}, \mathcal{H}) \ge N/t \mid \mathbf{w} \notin U\} &\le \Pr\{\sum_{i=1}^{N-T} \xi_i \ge \frac{N}{t} - T\} \\ &\le \Pr\{\sum_{i=1}^{N} \xi_i \ge \frac{N}{t} - t(N-D)\} \\ &\le 2^{-ND(\sigma\| \frac{t-1}{q-1})}, \end{aligned} \tag{9}$$

where in the last step we relied upon the inequality $\sigma > (t-1)/(q-1)$ implied by the condition on δ in the statement of the Theorem. Now

$$p_e \le \Pr\{\max_{\mathbf{w} \notin U} s(\mathbf{w}, \mathcal{H}) \ge N/t\} \le q^K \Pr\{s(\mathbf{w}, \mathcal{H}) \ge N/t \mid w \notin U\}.$$

Let us substitute the bound on the last probability. Taking the logarithm of the right-hand side and using the definitions $R(W) = K/N$ and $R(V) = \log_2 q/m$, we obtain:

$$p_e \le 2^{KmR(V)} 2^{-ND(\sigma\| \frac{t-1}{q-1})}$$

which is the same as (7). ∎

Remark. The theorem yields nontrivial results, *i.e.*, the exponential decline of the distributor's probability of incorrectly identifying a user as guilty, when the outer code W has large distance D. It is well known [18] (rediscovered in [23]) that concatenation of an inner separating code with an error-correcting outer code with large distance ($\delta > 1 - 1/t^2$) gives a longer separating code. We would like to stress that the separating property (renamed in crypto literature as secure frameproof property) is not sufficient for identifying members of the coalition. Therefore in the last theorem we perform a probabilistic analysis of the random ensemble of codes which amplifies the separating property to achieve a desired fingerprinting performance. Note also that concatenation of a fingerprinting inner code with an error correcting code used in [5] does not lead to asymptotically good fingerprinting codes.

Let us use the general construction of this theorem together with specific choices of codes to present a few families of easily constructible t-secure fingerprinting codes. These constructions will enable us to claim overall code rate separated from zero together with exponentially small probability of identification error for $t = const$. We will silently assume that for a fixed choice of the codes V and W the family of codes $\{C_k\}$ is obtained by applying N random mappings $\varphi_i^{(k)} : \mathbb{F}_q \to V, i = 1, 2, \dots, N$ to the coordinates of vectors in W. Below we focus only on the choice of the codes.

Take V to be a long binary (t,t)-separating code of length m and rate

$$R_{2t} = m^{-1} \log_2 q = -\frac{\log_2 P_{2t}}{2t-1}$$

(see Prop. 4) and W a $[q, K]$ extended Reed-Solomon (RS) code of rate $R(W) = K/q$ over $\mathbb{F}_q$. Substituting these parameters into Theorem 1 and taking into account that

$$(\log_2 q)^{-1} D\Big(\sigma \| \frac{t-1}{q-1}\Big) \sim \sigma$$

for t fixed and q growing, we obtain the following result.

Corollary 4 *For any fixed t and any rate $R < R_{2t}/t(t+1)$ a binary fingerprinting code of length n with 2^{Rn} code vectors (users) constructed by concatenating the codes V and W identifies one traitor with exponentially decreasing error probability*

$$p_e \le 2^{-n[t^{-1}R_{2t} - (t+1)R + o(1)]}.$$

Now let us take V to be a fixed binary (t,t)-separating (m,q) code, where q is an even power of a prime. This choice is possible due to Prop. 4. Let us take the outer code W to be a long algebraic geometry (AG) code from a maximal curve, whose parameters asymptotically approach the bound $R(W) = 1 - \delta - 1/(\sqrt{q} - 1)$ [24].

Corollary 5 *For any fixed t, any $q \ge (t^2+2)^2$ that is an even power of a prime, and any rate $R < R_{2t} R_0(W)$, where $R_0(W)$ is the root of the equation*

$$R(W) \log_2 q = D\Big(\frac{1}{t} - t\Big(R(W) - \frac{1}{\sqrt{q}-1}\Big) \Big\| \frac{t-1}{q-1}\Big) \tag{10}$$

a binary fingerprinting code of length n and size 2^{Rn} constructed by concatenating a fixed inner (m,q) code V with the (t,t)-separating property and algebraic geometry codes W of rate $R(W) \le R_0(W)$ and growing length N, identifies one traitor with exponentially falling error probability given by (7).

Proof : By Prop. 4, for any m there exists a binary (t,t)-separating code V of length m, size

$$|V| \ge (1/2)(1-2^{-(2t-1)})^{-m/(2t-1)},$$

and thus of rate $R(V) \approx R_{2t}$. Hence for a constant t and sufficiently large m it is possible to find a code of size q such that q satisfies the conditions of the corollary. Consider a q-ary AG code W of length N and rate $R(W) < R_0(W)$ such that for large N, its distance $D = \delta N$ satisfies $1-\delta \le R(W) + 1/(\sqrt{q}-1)$.

By Theorem 1, the error probability of identification will fall exponentially if $\sigma > (t-1)/(q-1)$ and

$$0 < R(W)\log_2 q < D(\sigma \| \frac{t-1}{q-1}). \tag{11}$$

The inequality $\sigma = \frac{1}{t} - t(R(W) + \frac{1}{\sqrt{q}-1}) > \frac{t-1}{q-1}$ is equivalent to the inequality $R(W) < A$, where

$$A = \frac{1}{t^2} - \frac{t-1}{t(q-1)} - \frac{1}{\sqrt{q}-1}.$$

It is immediate to verify that the condition $q \ge (t^2+2)^2$ implies that $A > 0$. Hence the upper bound (7) holds true if $0 < R(W) < A$.

Next, let us show that the segment of values of $R(W)$ that satisfy (11) is a proper subset of $(0, A)$. This is immediate since for $R(W) \in (0, A)$, $D(\sigma \| \frac{t-1}{q-1})$ is a positive decreasing function of $R(W)$ which reaches zero when $\sigma = (t-1)/(q-1)$, and the middle part of (11) increases from zero to $A \log_2 q$. Hence equation (10) has a (single) root $R_0(W) < A$, and for all $0 < R(W) < R_0(W)$, inequality (11) holds true. This completes the proof. ∎

Note that the codes of Corollary 5 are polynomially constructible. Indeed, by a recent result of [21] the construction complexity of the code W for them is $O((N \log_q N)^3)$, and the complexity of constructing the code V is constant, independent of N.

4.2 The case $t = 2$

A usual assumption in digital fingerprinting is that it is not possible to recover "the entire coalition since some of its members might be passive" (*e.g.*, [5, p.1899]). Therefore the fingerprinting problem was restricted to finding one guilty user with high probability. In the case of size-2 coalitions it is possible to construct a family of fingerprinting codes $\{C_k\}$ that have a stronger property than in the original definition. Namely, the distributor either recovers one member of the coalition $U = \{\mathbf{u}, \mathbf{u}'\}$ with probability one (*zero-error*), or *both* members with probability $1-\epsilon$.

To construct a family of fingerprinting codes $\{C_k\}$ in the case of $t = 2$ we use the same general idea of concatenation as for arbitrary t, with a somewhat different decision algorithm. The construction involves a binary $(2,2)$-separating $[m,\ell]$ linear code V [17], a q-ary $[N, K, D = \delta N]$-code, $q = 2^\ell$, and N random bijections $\varphi_i^{(k)} : \mathbb{F}_q \to V, i = 1, \dots, N$.

Encoding, or the fingerprinting assignment procedure, is the same as above. Namely, the fingerprint corresponding to a vector $\mathbf{w} = (w_1, \dots, w_N) \in W$ is obtained by computing $\mathbf{x} = (\varphi_1^{(k)}(w_1), \dots, \varphi_N^{(k)}(w_N))$. The length of the code C_k is $n = mN$ and the size is q^K. As above, we identify the users with the code vectors of the code W.

Assume that a coalition $U = \{\mathbf{u}, \mathbf{u}'\} \subset W$ generates a fingerprint $\mathbf{y} = (\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_N)$, where the subblocks $\mathbf{y}_i$ are length-m binary vectors.

Let us first describe inner decoding, which will be different from the general case. Recall that $Q = 2$, and so the narrow-sense envelope e and the wide-sense envelope E are the same. By definition, the subblock $\mathbf{y}_i$ is contained in the envelope $e(\{\mathbf{v}, \mathbf{v}'\})$, where $\mathbf{v} = \varphi_i^{(k)}(u)$ and $\mathbf{v}' = \varphi_i^{(k)}(u')$. Consider all possible pairs $(\mathbf{v}^1, \mathbf{v}^2)$ of code vectors from the code V such that $\mathbf{y}_i \in e(\{\mathbf{v}^1, \mathbf{v}^2\})$. By the $(2,2)$-separating property all these pairs should intersect. Hence either there is a vector $\mathbf{v}^*$ that is an element of every such pair (a *star* configuration), or there are three such pairs that form a triangle $\{\mathbf{v}, \mathbf{v}', \mathbf{v}''\}$ (cf. Lemma 3). In the first case the decoding result is given by $H_i = (\varphi_i^{(k)})^{-1}(\mathbf{v}^*) \in \mathbb{F}_q$; note that $H_i \in \{u_i, u_i'\}$. In the second case the result is given by the *list* $H_i = \{(\varphi_i^{(k)})^{-1}(\mathbf{v}), (\varphi_i^{(k)})^{-1}(\mathbf{v}'), (\varphi_i^{(k)})^{-1}(\mathbf{v}'')\}$, and $\{u_i, u_i'\} \subset H_i$.

Algorithm 2:

1. The N columns $\mathbf{y}_1, \ldots, \mathbf{y}_N$ of the observed fingerprint $\mathbf{y}$ are independently decoded with the code V. The decoding result H_i of the m-vector $\mathbf{y}_i$ is either a single element of $\mathbb{F}_q$ or a 3-subset of $\mathbb{F}_q$. The result of this step is the set $\mathcal{H} = \{\mathbf{h} \in \mathbb{F}_q^N : h_1 \in H_1, \ldots, h_N \in H_N\}$.

2. Denote by $\mathcal{I}$ the subset of coordinates in which $|H_i| = 1$ and let $\mathcal{L} = \{1, 2, \ldots, N\} \setminus \mathcal{I}$; $I = |\mathcal{I}|, L = |\mathcal{L}|$. If there is a vector $\mathbf{w} \in W$ such that the number of agreements
$$|\{i \in \mathcal{I} : w_i = H_i\}| > 2(N - D)$$
then the distributor identifies $\mathbf{w}$ as a member of U.

3. If the condition of the previous step is not satisfied, the distributor finds all code vectors $\mathbf{w} \in W$ such that $w_j \in H_j$ for all $j \in \mathcal{L}$ and identifies $\mathbf{u}$ and $\mathbf{u}'$ as any two of them.

This procedure enables D to find either both users from U with probability close to 1, or one user from U with probability 1.

Theorem 2 *The family of codes $\{C_k\}$ together with the decision rule given by Algorithm 2 forms a binary fingerprinting code of length $n = mN$ with $2^{Rn} = 2^{K\ell}$ code vectors (users) that either identifies one traitor with probability 1 or both with probability of error*
$$p_e \leq 2^{Rn}(2^\ell - 2)^{-N(1-6(1-\delta))}.$$

Proof. Let $\mathbf{y} = (\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_N)$ be a fingerprint generated by a coalition $U = \{u, u'\} \subset W$. Suppose that inner decoding of $\mathbf{y}$ results in a set $\mathcal{H}$. As we pointed out above, for all $i \in \mathcal{I}$ we have $H_i \in \{u_i, u_i'\}$. Hence for any $\mathbf{w} \notin \{\mathbf{u}, \mathbf{u}'\}$ the number of agreements
$$|\{i : w_i \in H_i\}| \leq |\{i : w_i = u_i\}| + |\{i : w_i = u_i'\}| \leq 2N(1 - \delta).$$
Therefore, Step 1 of outer decoding can yield only an element of U (or maybe both of them). If Step 1 gives no output, then $I \leq 4N(1 - \delta)$ because $I \leq |\{i : H_i = u_i\}| + |\{i : H_i = u_i'\}|$. In this case $L \geq N(1 - 4(1 - \delta))$, and both $u_i, u_i' \in H_i$.

Now observe that for $\mathbf{w} \notin \{\mathbf{u}, \mathbf{u}'\}$, its coordinate w_j can fall in the list H_j either because $w_j \in \{u_j, u_j'\}$, and the number of such positions is at most $2N(1 - \delta)$, or if $w_j \notin \{u_j, u_j'\}$, then it falls in H_j with probability $p \le \frac{1}{q-2}$ because the random bijection $\varphi_j^{(k)}$ is not known to the members of U, except for the (two) values $\varphi_j^{(k)}(u_j)$ and $\varphi_j^{(k)}(u'_j)$. Hence for all $j \in \mathcal{L}$ the probability

$$\Pr\{w_j \in H_j \,|\, j \in \mathcal{L}\} \le (q-2)^{-(L-2N(1-\delta))} \le (q-2)^{-N(1-6(1-\delta))}$$

for any vector $\mathbf{w} \in W \setminus \{\mathbf{u}, \mathbf{u}'\}$. Therefore the probability that at least one vector from the code W that is distinct from $\mathbf{u}, \mathbf{u}'$, will satisfy the condition of Step 2 is at most $(2^{K\ell}-2)(q-2)^{-N(1-6(1-\delta))}$. This proves the theorem. ∎

Example. Take a $(2,2)$-separating binary $[35,6]$-code [20] as inner code V, an extended Reed-Solomon $[65,7]$-code over $\mathbb{F}_{64}$ as outer code W and all $(64!)^{65}$ different vector mappings $\phi^{(k)} = (\varphi_1^{(k)}, \dots, \phi_{65}^{(k)})$, where the $\varphi_i^{(k)} : \mathbb{F}_{64} \to V$ are bijections. Then Theorem 2 states that the resulting binary fingerprinting code of length $n = 2275$ and size $M = 2^{42}$ either identifies one traitor with probability 1 or both with probability of error $p_e < 2^{42}62^{-29} < 10^{-39}$.

Let us use Theorem 2 to construct specific families of codes. Observe that, denoting $\alpha = 1 - 6(1-\delta)$, we can rewrite the bound for p_e as follows:

$$p_e \le 2^{-n(R(V)\alpha - R)}(1 - 2^{-\ell+1})^{-N\alpha} < 2^{-n(R(V)\alpha-R)}4^{2N\alpha/2^\ell}, \tag{12}$$

where the last step relies on the inequality $(1-1/x)^x > 1/4$ valid for all $x > 2$.

In particular, let us take W a Reed-Solomon code. We obtain

Corollary 6 *For any rate $R < (3 - \log_2 7)/21 \approx 0.0092$ there exists a family of fingerprinting codes C_k of length n with 2^{Rn} code vectors (users) that either identify one traitor with probability 1 or both with probability of error $p_e \le 2^{-n(R_4 - 7R - \beta)}$, where $R_4 = (3 - \log_2 7)/3 \approx 0.0642$ and $\beta = O(N^{-1})$.*

Proof : Take V a $(2,2)$ separating $[m, \ell]$ linear code. By Proposition 4 for large m there exists a code of rate arbitrarily close to R_4. Take W an extended Reed-Solomon $[N = 2^\ell, K]$ code over $\mathbb{F}_q, q = N$ of rate $R(W) < 1/7$. The estimate of p_e is obtained from (12) by direct substitution. ∎

Remark. Codes of Corollary 6 have a stronger property than codes of the previous section for the case of $t = 2$, namely, the error-free recovery of one user from the coalition or of both users with a small probability of error. This is achieved in exchange for a drop of the maximal achievable rate of codes by a factor of 6/7 (cf. Cor. 6 vs. Cor. 4).

Now let us concatenate a $(2,2)$-separating binary $[m, 2\ell]$ code V with a family of AG codes W from maximal curves over $\mathbb{F}_{2^{2\ell}}$. We assume that the rate $R(W)$ is fixed and let N grow. Then the parameters of W approach the bound $1 - \delta = R(W) + 1/(2^\ell - 1)$. The parameters of the resulting code family depend on the choice of the code V. For instance, we have the following result.

Corollary 7 *For any rate $R < \frac{121}{8001} \approx 0.015$ there exists a family of fingerprinting codes of length n and rate R that either identify one traitor with probability 1 or both with probability of error $p_e \le 2^{-n(0.105 - 7R)}$.*

Proof : Let V be the $(2,2)$-separating binary $[m = 126, 2\ell = 14]$ code of rate $R(V) = 1/9$ [8] and W be an $[N, K, \delta N]$ linear code from a family of maximal curves over $\mathbb{F}_{2^{14}}$. For large N we have $\alpha = (121/127) - 54R$. Using this in (12) gives the claimed bound on p_e. ∎

Note that construction complexity of codes given by Corollary 7 grows polynomially with the code length n.

5 Identification (decoding) algorithms

In this section we address the algorithmic side of the recovery problem of the users from the coalition. This question was essentially side-stepped in the literature. A notable exception is work on tracing traitors [6], [7] which discusses identification complexity for the narrow-case fingerprinting problem. However these papers, as well [5], only give algorithms with complexity linear in the number M of users of the system, *i.e.*, in our terms, exponential in the length of the fingerprinting code. Here we give algorithms of complexity $\mathsf{polylog}(M) = \mathsf{poly}(n)$ that ensure exponential decline of the error probability. Although the size of the codes discussed in this section is still exponential in n, we will have to introduce somewhat more stringent restrictions on the code rate R than in the previous section.

Conceptually we face a problem of decoding of two-level concatenated codes. In the case of error correction this problem has a long history in coding theory literature [10], [9]. Decoding can be accomplished by exhaustive search performed for the inner-level code and some algebraic algorithm for the outer code. Following this pattern, we analyze the decoding algorithm formed of the following steps.

Recall the parameters (m, q) of the inner code V and $[N, K, D = N\delta]$ of the outer code W which will be assumed an AG code. Suppose that a coalition $U = (\mathbf{u}^1, \dots, \mathbf{u}^t)$ generates a fingerprint $\mathbf{y} = (\mathbf{y}_1, \dots, \mathbf{y}_N)$. Upon inner decoding of all the $\mathbf{y}_i$'s we obtain a subset

$$\mathcal{H} = \{\mathbf{z} \in \mathbb{F}_q^N : z_1 \in H_1, \dots, z_N \in H_N\}.$$

of N-words over $\mathbb{F}_q$. The objective of outer decoding is to find a vector $\mathbf{w} \in W$ that minimizes the distance $\mathrm{dist}(\mathbf{w}, \mathcal{H})$ from the code to the "received" subset $\mathcal{H}$. A straightforward approach to this problem requires $|\mathcal{H}|$ runs of some conventional decoding algorithm, which results in exponential running time. Furthermore, until recently all known algebraic algorithms could correct only about $D/2$ errors (the so-called bounded distance decoding). Note that even if the set $\mathcal{H}$ is of size 1, it does not help to use a bounded distance decoding algorithm because we need to correct, roughly speaking, $N(1-1/t)$ errors. To show this, consider the following fingerprint vector $\mathbf{y} = (\mathbf{y}_1, \dots, \mathbf{y}_N)$, where (assuming that N/t is integer)

$$y_{rt+i} = \varphi^{(k)}_{rt+i}(u^i_{rt+i}), r = 0, 1, \dots, N/t - 1;\ i = 1, \dots, t$$

Then $H_{rt+i} = u^i_{rt+i}$, so $|H_{rt+i}| = 1$, and for every vector $\mathbf{u}^i$ we have $N - \mathrm{dist}(\mathbf{u}^i, \mathcal{H}) = N/t$. Fortunately, the recently found list decoding procedures of AG codes correct many more errors, outputting in polynomial time a small (polynomial-sized) list of code vectors. More precisely, we have

Guruswami-Sudan (GS) decoding [11]. Let W be a q-ary (one-point) AG code with parameters $[N, RN, \delta N]$. Then for any vector $\mathbf{x} \in F_q^N$ the decoding algorithm of [11] outputs a list $\mathcal{Y} \subset W$ formed of all the vectors $\mathbf{w} \in W$ such that for large values of N, $\mathrm{dist}(\mathbf{w}, \mathbf{x}) \le N(1-\sqrt{1-\delta})$. The size $|\mathcal{Y}|$ is bounded by a polynomial function of N. The implementation complexity of the algorithm is also polynomial in N.

Remarks. 1. GS algorithm is applicable to any AG code W constructed in a usual way from a (smooth projective absolutely irreducible) curve over $\mathbb{F}_q$ with N rational points $P_1, \dots, P_N$. Namely, let $F_1, \dots, F_K$ be a basis of the Riemann-Roch space of rational functions on X. Then the set of vectors $\{(F_i(P_1), \dots, F_i(P_N)), 1 \le i \le K\}$ in $\mathbb{F}_q$ forms a basis of the code W. In particular, taking

X the projective line over $\mathbb{F}_q$, we obtain the RS code. For this reason, all the codes obtainable in this way are sometimes called evaluation codes. Note that the GS algorithm is applicable to codes of any length N less than or equal to the number $\sharp(X/\mathbb{F}_q)$ of rational points of the curve X. Implementation details for one-point AG codes together with complexity estimates are worked out in [16].

2. Note that if our goal is to output one decoding result rather than a list, we choose a vector from $\mathcal{Y}$ such that its distance to $\mathbf{x}$ is the smallest among all the members of $\mathcal{Y}$. In coding terms, this algorithm enables one to correct, whenever possible, a number of errors greater than half the minimum distance of the code. Relying on this algorithm, we will show a way to find a member of the coalition as the nearest code vector, maintaining the overall polynomial complexity.

5.1 Identification in the case of arbitrary t

Consider the family of fingerprinting codes $\{C_k\}$ formed by concatenating a fixed binary (t,t)-separating (m,q) code V with outer codes q-ary $[N, R(W)N, \delta N]$ AG codes W of growing length N. Consider the following identification procedure for the observed fingerprint $\mathbf{y} = (\mathbf{y}_1, \dots, \mathbf{y}_N)$.

Algorithm $1'$:

1. Construct the set $\mathcal{H} = (\mathbf{h} \in \mathbb{F}_q^N : h_1 \in H_1, \dots, h_N \in H_N)$, where each H_i is the subset of symbols of $\mathbb{F}_q$ corresponding to an arbitrary t-tuple of vectors of V such that $\mathbf{y}_i$ is contained in their envelope. To accomplish this the distributor performs a lookup of at most $\binom{m}{t}$ t-tuples of vectors of V. (Same as Step 1 of Algorithm 1.)

2. Take a random vector $\hat{\mathbf{h}} \in \mathcal{H}$. Decode $\hat{\mathbf{h}}$ with respect to W by the GS algorithm to obtain a list $\mathcal{Y} \subset W$ of code vectors within distance $N(1-\sqrt{1-\delta})$ from it. If $\mathcal{Y} \neq \emptyset$, identify the vector from $\mathcal{Y}$ closest to $\mathbf{h}$ as a member of the coalition U.

Theorem 3 *There exists $R_0 > 0$ such that for the family of fingerprinting codes $\{C_k\}$ of rate $R < R_0$, Algorithm $1'$ has complexity $\mathsf{poly}(n)$ and identifies a member of the coalition with error probability $2^{-\Omega(n)}$.*

Proof. Let $\hat{\mathbf{u}} \in U \subset W$ be the vector of U closest to $\hat{\mathbf{h}}$. We shall prove that with probability $1 - 2^{-\Omega(N)}$

(i) $\hat{\mathbf{u}}$ is the closest codeword in W to $\mathbf{h}$, and
(ii) the distance $\operatorname{dist}(\hat{\mathbf{u}}, \hat{\mathbf{h}})$ is within the correction radius of GS decoding.

We begin with part (i). Let $\mathbf{w} \in W$ be any vector outside U and let us estimate the probability

$$P_{e,\mathbf{w}} = \Pr\{\operatorname{dist}(\mathbf{w}, \hat{\mathbf{h}}) < \operatorname{dist}(\hat{\mathbf{u}}, \hat{\mathbf{h}})\}.$$

Note that for any two random variables α and β and any real c we have

$$\begin{aligned}\Pr\{\alpha \geq \beta\} &\leq \Pr\{\alpha \geq \min(c, \beta)\}\\ &= \Pr\{\alpha \geq c, \beta \geq c\} + \Pr\{\alpha \geq \beta, c > \beta\}\\ &\leq \Pr\{\alpha \geq c\} + \Pr\{c \geq \beta\}.\end{aligned}$$

Hence

$$P_{e,\mathbf{w}} = \Pr\{s(\mathbf{w}, \hat{\mathbf{h}}) \geq s(\hat{\mathbf{u}}, \hat{\mathbf{h}})\} \leq \Pr\{s(\mathbf{w}, \hat{\mathbf{h}}) \geq c\} + \Pr\{s(\hat{\mathbf{u}}, \hat{\mathbf{h}}) \leq c\}.$$

Since $\hat{\mathbf{h}}$ is a random vector from $\mathcal{H}$, the quantity $\sum_{\mathbf{u}\in U} s(\mathbf{u},\hat{\mathbf{h}})$ is a random variable. The probability $\Pr\{\hat{h}_i \in U_i\} = |U_i \cap H_i|/|H_i|$, and therefore

$$\sum_{\mathbf{u}\in U} s(\mathbf{u},\hat{\mathbf{h}}) \geq \sum_{i=1}^{N} \xi_i,$$

where the ξ_i-s are independent Bernoulli random variables with $\Pr\{\xi_i = 1\} = 1/t, \Pr\{\xi_i = 0\} = 1 - 1/t$. Hence,

$$s(\hat{\mathbf{u}},\hat{\mathbf{h}}) = \max_{u\in U} s(\mathbf{u},\hat{\mathbf{h}}) \geq t^{-1}\sum_{i=1}^{N} \xi_i$$

and

$$\Pr\{s(\hat{\mathbf{u}},\hat{\mathbf{h}}) \leq c\} \leq \Pr\{\sum_{i=1}^{N} \xi_i \leq ct\} \leq 2^{-ND(\gamma t\|1/t)},$$

where we have put $c = N\gamma$ and assumed that $\gamma < t^{-2}$.

On the other hand, consider any vector $\mathbf{w} \in W, \mathbf{w} \notin U$ and let $\mathcal{F}(\mathbf{w}) = \{i : w_i \in U_i\}$. Then we have $\Pr\{\hat{h}_i = w_i\} = 1/t$ for $i \in \mathcal{F}(\mathbf{w})$ and

$$\Pr\{\hat{h}_i = w_i\} = \frac{1}{q - |U_i \cap H_i|} \leq \frac{1}{q-t}$$

for $i \notin \mathcal{F}(\mathbf{w})$. Since $|\mathcal{F}(\mathbf{w})| \leq t(N-D)$, we obtain the estimate

$$\Pr\{s(\mathbf{w},\hat{\mathbf{h}}) \geq c\} \leq \Pr\Big\{\sum_{i=1}^{tN(1-\delta)} \xi_i + \sum_{i=1}^{N(1-t(1-\delta))} \mu_i \geq c\Big\},$$

where every μ_i is a Bernoulli variable with $\Pr\{\mu_i = 1\} = 1/(q-t)$ and $P\{\mu_i = 0\} = 1 - 1/(q-t)$. To estimate this probability we use the Chernoff bound. This gives the estimate $\Pr\{s(\mathbf{w},\hat{\mathbf{h}}) \geq c\} \leq 2^{-NE_1}$, where

$$E_1 = 2(\log_2 e)\big(\gamma - \frac{(1-\delta)(q-2t)+1}{q-t}\big)^2,$$

which is valid as long as $\gamma > 1 - \delta + (1 - t(1-\delta))(q-t)$.

Overall we have

$$p_{e,W} = \Pr\{\max_{\mathbf{w}\notin U} s(\mathbf{w},\hat{\mathbf{h}}) \geq s(\hat{\mathbf{u}},\hat{\mathbf{h}})\} \leq q^{NR(W)} P_{e,\mathbf{w}} \leq 2\cdot 2^{-nR(E/m - R(W))}$$

where $E = \min\{D(\gamma t\|1/t), E_1\}$ and γ satisfies the conditions

$$1 - \delta + \frac{1 - t(1-\delta)}{q-t} \leq \gamma \leq \frac{1}{t}^2.$$

For large q these conditions are consistent if $1 - \delta < t^{-2}$. Therefore, for for relatively small rate $R(W)$ (more precisely, for $R(W) < \min\{E/m, 1 - \delta - 1/(\sqrt{q}-1)\}$) the error probability $P_{e,\mathbf{w}}$ decreases exponentially with n.

Let us examine the decoding condition (ii). The decoder may not find $\hat{\mathbf{u}}$ if $s(\hat{\mathbf{u}}, \hat{\mathbf{h}}) \le N\sqrt{1-\delta}$. The probability of this event satisfies

$$\Pr\{s(\hat{\mathbf{u}}, \hat{\mathbf{h}}) \le N\sqrt{1-\delta}\} \le \Pr\{t^{-1}\sum_{i=1}^{N} \xi_i \le N\sqrt{1-\delta}\} \le 2^{-NE'},$$

where $E' = D(t\sqrt{1-\delta}\|1/t)$, under condition that $\sqrt{1-\delta} < 1/t^2$.

Clearly, there exists $R_0 > 0$ such that the overall probability of incorrect identification

$$p_e \le 2 \cdot 2^{-n(E/m-R)} + 2^{-nE'/m}$$

decreases exponentially with n if the code rate $R = R(V)R(W) < R_0$.

The decoding complexity of GS algorithm is polynomial in N, and so the complexity of identification is polynomial in n. ∎

In the proof we made no attempt to optimize the maximum admissible value of the rate $R(W)$. It is possible to obtain a better result by replacing the Chernoff bound by the divergence-based estimates of the tails of the distributions, although this complicates the calculations.

5.2 Identification algorithm in the case $t = 2$

Codes secure against coalitions of size two were proven in Sect.4.2 to be able to restore either one user with probability one or both with a small error probability. Here we wish to show that this performance is attainable under a polynomial-time identification algorithm.

Consider the family of fingerprinting codes $\{C_k\}$ of length $n = Nm$ and size $2^{\ell N R(W)}$ formed by concatenating a fixed binary $(2,2)$-separating $[m, \ell]$ code V (ℓ is even) and a q-ary $[N, R(W)N, \delta N]$ AG code W ($q = 2^\ell$), For any subset $\mathcal{A}$ of coordinates and an N-vector $\mathbf{x}$ let $\mathbf{x}_{\mathcal{A}}$ be its projection on the coordinates in $\mathcal{A}$; let $W_{\mathcal{A}}$ be the the code W projected on the coordinates in $\mathcal{A}$. By Sect. 4.2, after inner decoding the distributor forms a set $\mathcal{H} = \{\mathbf{h} \in \mathbb{F}_q^N : h_1 \in H_1, \dots, h_N \in H_N\}$, where some subsets H_i are singletons (i.e., q-ary letters) and some 3-sets of q-ary letters. We recall the notation $\mathcal{I}$ for the subset of coordinates corresponding to the singletons and $\mathcal{L} = \{1, 2, \dots, N\} \backslash \mathcal{I}$ for the subset corresponding to the 3-sets; $I = |\mathcal{I}|, L = |\mathcal{L}|$. Then $\mathcal{H} = \mathbf{h}_{\mathcal{I}} \times \mathcal{H}_{\mathcal{L}}$, where $\mathbf{h}_{\mathcal{I}} = (H_i, i \in \mathcal{I})$ and $\mathcal{H}_{\mathcal{L}} = \{\mathbf{h}_{\mathcal{L}} : \mathbf{h} \in \mathcal{H}\}$.

Consider the following decoding procedure of the code C.

ALGORITHM $2'$:

1. Apply the same procedure as in Step 1 of Algorithm 2 to construct the set $\mathcal{H}$.

2. If $I > 4(N-D)$, apply GS decoding to the vector $\mathbf{h}_{\mathcal{I}}$ to find the nearest code vector $\hat{\mathbf{w}}_{\mathcal{I}} \in W_{\mathcal{I}}$ in the sphere of radius $I/2$ around $\mathbf{h}_{\mathcal{I}}$. Find the code vector $\hat{\mathbf{w}} \in W$ whose projection on $\mathcal{I}$ equals $\hat{\mathbf{w}}_{\mathcal{I}}$; identify $\hat{\mathbf{w}}$ as a member of the coalition.

3. If $I \le 4(N-D)$, take a random vector $\hat{\mathbf{h}}_{\mathcal{L}}$ in $\mathcal{H}_{\mathcal{L}}$. Using GS decoding, find two nearest code vectors $\hat{\mathbf{w}}_{\mathcal{L}}, \hat{\mathbf{w}}'_{\mathcal{L}} \in W_{\mathcal{L}}$ to $\hat{\mathbf{h}}_{\mathcal{L}}$ in the sphere of radius $L(1-\sqrt{1-\delta})$. Restore the corresponding code vectors $\hat{\mathbf{w}}, \hat{\mathbf{w}}'$ of W; identify the pair $\{\hat{\mathbf{w}}, \hat{\mathbf{w}}'\}$ as the coalition U.

Theorem 4 *For the family of fingerprinting codes $\{C_k\}$ the identification rule given by Algorithm $2'$ has complexity* $\mathsf{poly}(n)$ *and either identifies one traitor with probability* 1 *or both with probability of error with error probability* $2^{-\Omega(n)}$.

Proof. We will assume that (for large N) the parameters of W satisfy the relation $1-\delta = R(W)+1/(\sqrt{q}-1)$. Let us estimate the error probability p_e.

(a) If $I > 4(N-D)$ then the code vector $\hat{\mathbf{w}}_{\mathcal{I}}$ corresponds to a member of the coalition $U = \{\mathbf{u},\mathbf{u}'\}$ (see the proof of Thm. 2), and the Hamming distance $\operatorname{dist}(\hat{\mathbf{w}}_{\mathcal{I}},\mathbf{h}_{\mathcal{I}}) \le I/2$. In order to find $\hat{\mathbf{w}}_{\mathcal{I}}$ from $\mathbf{h}_{\mathcal{I}}$ we need to correct at most $I/2$ errors with the code $W_{\mathcal{I}}$ of length I. It is possible to correct $I/2$ errors with the GS algorithm as long as

$$1-\sqrt{1-\delta} \ge 1/2.$$

To satisfy this inequality it suffices that $R(W) \le 1/4 - 1/(\sqrt{q}-1)$. Note that in this case identification is error-free. Note also that it is possible to recover $\hat{\mathbf{w}}$ uniquely from $\hat{\mathbf{w}}_{\mathcal{I}}$ by solving the parity-check equations of the code W since the number of missing positions is $L \le 4D-3N \le D-1$.

(b) If the condition of (a) does not hold, *i.e.*, $L \ge N(1-4(1-\delta))$, then we take a random vector $\hat{\mathbf{h}}_{\mathcal{L}}$ in $\mathcal{H}_{\mathcal{L}}$. As in Theorem 3, we shall prove that with probability $1-2^{-\Omega(n)}$

(i), the vectors $\mathbf{u}_{\mathcal{L}},\mathbf{u}'_{\mathcal{L}}$ are the nearest code vectors to $\hat{\mathbf{h}}_{\mathcal{L}}$, *i.e.*, $\{\hat{\mathbf{w}}_{\mathcal{L}},\hat{\mathbf{w}}'_{\mathcal{L}}\} = \{\mathbf{u}_{\mathcal{L}},\mathbf{u}'_{\mathcal{L}}\}$, and
(ii), both distances $\operatorname{dist}(\mathbf{u}_{\mathcal{L}},\hat{\mathbf{h}}_{\mathcal{L}})$ and $\operatorname{dist}(\mathbf{u}'_{\mathcal{L}},\hat{\mathbf{h}}_{\mathcal{L}})$ are within the correction radius of the GS algorithm.

We begin with part (i) of the claim. Clearly,

$$s(\mathbf{x}_{\mathcal{L}},\hat{\mathbf{h}}_{\mathcal{L}}) = \sum_{i=1}^{L}\theta_i,$$

where $\mathbf{x}$ is any of the vectors $\mathbf{u},\mathbf{u}'$ and θ_i are independent Bernoulli variables with $P(\theta_i=1)=1/3$ and $P(\theta_i=0)=2/3$. By the Chernoff bound, if c is any real number such that $L/3 > c$, then

$$\Pr\{\min\{s(\mathbf{u}_{\mathcal{L}},\hat{\mathbf{h}}_{\mathcal{L}}),s(\mathbf{u}'_{\mathcal{L}},\hat{\mathbf{h}}_{\mathcal{L}})\} \le c\} \le 2\Pr\{\sum_{i=1}^{L}\theta_i \le c\} \le 2e^{-\frac{3}{2}(L/3-c)^2/L}.$$

On the other hand, let $\mathbf{w} \in (W \setminus U)$. Let us estimate the number of agreements between $\mathbf{h}_{\mathcal{L}}$ and $\mathbf{w}_{\mathcal{L}}$. Similarly to the proof of Theorem 2 we observe that for $\mathbf{w} \notin \{\mathbf{u},\mathbf{u}'\}$, its coordinate w_i can coincide with $\hat{h}_i$ in one of the two following cases. One option is that $w_i \in \{u_i,u'_i\}$; then the probability $\Pr\{w_i=\hat{h}_i\}=1/3$, and there can be at most $2(N-D)$ such positions. The other option is that $w_i \notin \{u_i,u'_i\}$, then $\Pr\{w_i=\hat{h}_i\} \le \frac{1}{q-2}$. Therefore, $s(\mathbf{w}_{\mathcal{L}},\hat{\mathbf{h}}_{\mathcal{L}})$ is a random variable that satisfies

$$s(\mathbf{w}_{\mathcal{L}},\hat{\mathbf{h}}_{\mathcal{L}}) \le \sum_{i=1}^{2N(1-\delta)}\theta_i + \sum_{i=1}^{L-2N(1-\delta)}\mu_i,$$

where every μ_i is a Bernoulli variable with $\Pr\{\mu_i=1\}=1/(q-2)$ and $P\{\mu_i=0\}=(q-3)/(q-2)$. Hence if we put $\bar{w}=(2/3)N(1-\delta)+L/(q-2)$ and assume that $c>\bar{w}$, then

$$\Pr\{s(\mathbf{w}_{\mathcal{L}},\hat{\mathbf{h}}_{\mathcal{L}}) \ge c\} \le e^{-2(c-\bar{w})^2/L}.$$

Summarizing, we have

$$\begin{aligned}
P_{e,\mathbf{w}} &= \Pr\{s(\mathbf{w},\hat{\mathbf{h}}) \ge \min\{s(\mathbf{u}_{\mathcal{L}},\hat{\mathbf{h}}_{\mathcal{L}}),s(\mathbf{u}'_{\mathcal{L}},\hat{\mathbf{h}}_{\mathcal{L}})\}\}\\
&\le \Pr\{s(\mathbf{w},\hat{\mathbf{h}}_{\mathcal{L}}) \ge c\} + \Pr\{\min\{s(\mathbf{u}_{\mathcal{L}},\hat{\mathbf{h}}_{\mathcal{L}}),s(\mathbf{u}'_{\mathcal{L}},\hat{\mathbf{h}}_{\mathcal{L}})\} \le c\}\\
&\le e^{-2(c-\bar{w})^2/L} + 2e^{-\frac{3}{2}(L/3-c)^2/L}.
\end{aligned}$$

An easy calculation shows that if $\delta > 5/6$ (for large q), then the conditions on c are consistent, and so the probability $P_{e,\mathbf{w}}$ falls exponentially in N if $R(W)$ is sufficiently small. Clearly, there exists a finite interval of rates $R(W) > 0$ such that the error probability $p_{e,W} \le q^{NR(W)} P_{e,\mathbf{w}}$ falls exponentially in N. This proves (i).

Let us examine the decoding condition (ii). The decoder may not find $\mathbf{u}_{\mathcal{L}}$ or $\mathbf{u}'_{\mathcal{L}}$ if

$$\min\{s(\mathbf{u}_{\mathcal{L}}, \hat{\mathbf{h}}_{\mathcal{L}}), s(\mathbf{u}'_{\mathcal{L}}, \hat{\mathbf{h}}_{\mathcal{L}})\} \le N\sqrt{1-\delta}.$$

By the Chernoff bound the probability of this event is at most $2e^{-\frac{3}{2}(L/3 - N\sqrt{1-\delta})^2/L}$ as long as $L/3 > N\sqrt{1-\delta}$. This inequality is satisfied for any $L > N(1-4(1-\delta))$ if $\delta > 15/16$.

Clearly, there exists $R_0 > 0$ such that the overall probability of incorrect identification decreases exponentially with n if the code rate $R = R(V)R(W) < R_0$. The decoding complexity is polynomial in n. ∎

6 Concluding remarks: related problems

We considered properties and constructions of codes for the wide-sense envelope, also called fingerprinting codes. A large group of problems related to the narrow-sense envelope, defined in Sect. 2.1, was recently overviewed in [22]. Among them we single out the following two problems.

1.1 *Codes with the identifying parent property* (t-IPP codes). As remarked in Sect. 2.1, codes for the narrow-sense fingerprinting problem is are also called IPP codes. This problem was suggested in [12], where it was also proved that there exist q-ary 2-IPP codes with positive rate for every $q \ge 3$. The question of code existence for other t was discussed in [22] and settled in [2]: there exist t-IPP codes with rate $R > 0$ if and only if $q \ge t+1$.

As remarked above, for IPP codes identification is possible with zero error rate. The problem of constructing explicit code families together with efficient identification procedures, as far as we know, has not been addressed in the literature (it was suggested as an open problem in [22]; as explained in the next remark, this problem is nontrivial only for $q \le t^2$). The technique introduced in Sect. 5 of the present paper enables us to obtain, for every $q \ge t+1$, explicit (polynomially constructible) code families that afford identification with complexity growing polynomially with the code length n, that gives in particular an affirmative question on the open problem from [22]. This result will be reported elsewhere [3].

1.2. *t-traceability codes.* A t-i.p.p. code C has the traceability property [6], [7] if for every fingerprint y created by the coalition U, the set of its nearest neighbors in C is contained in U. It is known [6] that a q-ary $[n,k,d]$ code possesses the t-traceability property if $d \ge n(1-1/t^2)$. It is clear from the properties of the GS algorithm that evaluation (AG or RS) codes with large distance possess the traceability property together with a polynomial identification algorithm.

On the other hand, to construct q-ary error-correcting codes of rate bounded away from zero, we must take $q > n/(n-d)$ (by the Plotkin bound). Hence, if traceability codes are constructed based on error correction properties only, then no matter what the identification algorithm is, nonzero code rate is obtained only if $q > t^2$. Thus, the existing technique does not enable us to solve a problem raised in [22] which calls for construction of traceability codes of nonzero rate over alphabets of size $q \le t^2$.

2. *Capacity of the fingerprinting channel.* Let us call the rate R an achievable rate of the "fingerprinting channel" if for any given $a > 0$ there exists a positive integer n_0 with the property

that for every $n > n_0$ the best attainable error probability of identification (2) satisfies

$$\epsilon(\mathcal{E}; n, q^{nR}, t) < a.$$

As usual, let us call the number

$$\mathcal{C}(\mathcal{E}; q, t) = \sup_{R \text{ achievable}} R$$

capacity of digital fingerprinting for the envelope $\mathcal{E}$. We have shown that $\mathcal{C}(E; 2, t) > 0$ for every given t and presented constructive code sequences with rate bounded away from zero. The question of finding the capacity of digital fingerprinting presents an interesting open problem.

References

[1] N. Alon, J. H. Spencer, *The probabilistic method*, John Wiley & Sons, Inc., New York, 1992.

[2] A. Barg, G. Cohen, S. Encheva, G. Kabatiansky, G. Zémor, "A hypergraph approach to the identifying parent property: the case of multiple parents," *SIAM J. Discrete Math.*, **14** (2001), pp. 423–431.

[3] A. Barg and G. Kabatiansky, "A class of i.p.p. codes with efficient identification," *Proc. 39th Annual Allerton Conf. Commun. Control Comput.*, Monticello, IL, Oct. 2001, to be published.

[4] G. R. Blakley, C. Meadows, and G. Purdy, "Fingerprinting long forgiving messages," *Proc. CRYPTO*, 1985, pp. 180–189.

[5] D. Boneh and J. Shaw, "Collusion-secure fingerprinting for digital data," *IEEE Trans. on Inf. Theory*, **44** (1998), pp. 1987–1905.

[6] B. Chor, A. Fiat and M. Naor, "Tracing traitors," *Crypto'94* LNCS 839 (1994), pp. 257–270.

[7] B. Chor, A. Fiat, M. Naor, and B. Pinkas, "Tracing traitors," *IEEE Trans. on Inf. Theory*, **46** (2000), pp. 893–910.

[8] G. Cohen, S. Encheva, and S. Litsyn, "Intersecting codes and partially identifying codes," *Proc. Int'l Workshop on Coding and Cryptography*, Paris, Ecoles de Coëtquidan, 2001, pp. 139–147.

[9] I. Dumer, *Concatenated codes and their multilevel generalizations*, Handbook of Coding Theory (V. Pless and W. C. Huffman, eds.), vol. 2, Elsevier Science, Amsterdam, 1998, pp. 1911–1988.

[10] G. D. Forney, Jr. *Concatenated Codes.* Cambridge, MA: MIT Press (1966).

[11] V. Guruswami and M. Sudan, "Improved decoding of Reed-Solomon and algebraic-geometry codes," *IEEE Trans. Inform. Theory*, **45**, no.6 (1999), pp. 1757–1767.

[12] H. D. Hollmann, J. H. van Lint, J.-P. Linnartz and L. M. G. M. Tolhuizen, "On codes with the identifiable parent property", *J. Combinatorial Theory*, Series A, **82** (1998) pp. 121–133.

[13] A. D. Friedman, R. L. Graham and J. D. Ullman, "Universal single transition time asynchronous state assignments," *IEEE Trans. Comput.*, **18** (1969), pp. 541–547.

[14] J. Körner and G. Simonyi, "Separating partition systems and locally different sequences," *SIAM J. Discrete Math.*, **1** (1988), pp. 355–359.

[15] F. J. MacWilliams and N. J. A. Sloane , *The Theory of Error-Correcting Codes*, North-Holland, Amsterdam (1977).

[16] R. R. Nielsen, *Decoding AG-codes beyond half the minimum distance*, Master Thesis, Dept. of Math., Technical University of Denmark (1998).

[17] M. S. Pinsker, Y. L. Sagalovich, "Lower bound on the cardinality of code of automata's states," *Probl. Inform. Trans.*, **8**, no. 3 (1972), pp. 59–66.

[18] Y. L. Sagalovich, "Concatenated codes of the states of an automaton," *Probl. Inform. Trans.*, **14**, no. 2 (1978), pp. 77–85.

[19] Y. L. Sagalovich, "Fully separated systems," *Probl. Inform. Trans.*, **18**, no. 2 (1982), pp. 74–82.

[20] Y. L. Sagalovich, "Separating systems," *Probl. Inform. Trans.*, **30**, no. 2 (1994) pp. 14–35.

[21] K. W. Shum, I. Aleshnikov, P. V. Kumar, H. Stichtenoth, and V. Deolalikar, "A low-complexity algorithm for the construction of algebraic-geometric codes better than the Gilbert-Varshamov bound," *IEEE Trans. Inform. Theory*, **47**, no 6 (2001), pp. 2225–2241.

[22] J. N. Staddon, D. R. Stinson, and R. Wei, "Combinatorial properties of frameproof and traceability codes," *IEEE Trans. Inform. Theory*, **47** (2000), pp. 1042–1049.

[23] D. R. Stinson, Tran V. T., and R. Wei, "Secure frameproof codes, key distribution patterns, group testing algorithms and related structures," *Journ. Stat. Plan. Inference*, **86** (1998), pp. 595–617.

[24] M. Tsfasman and S. Vlăduţ, *Algebraic-geometric codes*, Kluwer, Dordrecht, 1991.