Available at
www.ElsevierComputerScience.com
POWERED BY SCIENCE DIRECT®
Journal of Complexity 20 (2004) 137–147

Journal of
COMPLEXITY

http://www.elsevier.com/locate/jco

# A class of I.P.P. codes with efficient identification

Alexander Barg[a,*,1] and Gregory Kabatiansky[b,2]

[a] *DIMACS, Rutgers University, 96 Frelinghuysen Rd., Piscataway, NJ 08854, USA*
[b] *IPPI RAN, Bol'shoj Karetnyj 19, Moscow 101447, Russia*

Received 23 October 2002; accepted 12 August 2003

**Abstract**

Let $\mathscr{C}$ be a code of length $n$ over a $q$-ary alphabet. An $n$-word $y$ is called a descendant of a set of $t$ codewords $x^1, \dots, x^t$ if $y_i \in \{x_i^1, \dots, x_i^t\}$ for all $i = 1, \dots, n$. A code is said to have the *t-identifying parent property* ($t$-i.p.p.) if for any $n$-word $y$ that is a descendant of at most $t$ parents it is possible to identify at least one of them.

An explicit construction is presented of $t$-i.p.p. codes of rate bounded away from zero, for which identification can be accomplished with complexity $\mathsf{poly}(n)$.

*Keywords:* Code concatenation; Digital finger printing; Identifiable parent property; List decoding; Traceability

## 1. Introduction: I.P.P. and traceability codes

Consider the set of $n$-words over an alphabet $\mathscr{Q}$ of size $q$. Let $\mathrm{d}(a, b)$ denote the Hamming distance and $s(a, b) = n - \mathrm{d}(a, b)$ denote the number of agreements between the vectors $a$ and $b$. A subset $\mathscr{C}$ of $\mathscr{Q}^n$ is called a code. Let $d = \mathrm{d}(\mathscr{C})$ be the distance of $\mathscr{C}$, i.e., the minimum distance between distinct codewords. We write $\mathscr{C}(n, M, d)$ to denote a code of length $n$, size $M$, and distance $d$, and $\mathscr{C}(n, M)$ if the

[*] Corresponding author. Department of Electrical and Computer Engineering, University of Maryland, 2361 A.V. Williams Bldg., College Park, MD 20742, USA.

*E-mail addresses:* abarg@ieee.org (A. Barg), kaba@iitp.ru (G. Kabatiansky).

[1] Research supported in part by the Binational Science Foundation (USA–Israel), Grant 1999099, and by NSF, Grant CCR 0310961.

[2] Research supported in part by the Russian Foundation for Fundamental Research under the RFFI Grant 03-01-00098.

0885-064X/$ - see front matter
doi:10.1016/j.jco.2003.08.012

value of the distance is of no importance. By $R = R(\mathscr{C})$ we denote the code rate

$$R = (1/n)\log_q M.$$

Finally, if $\mathscr{Q}$ forms a finite field, we use the notation $\mathscr{C}[n,k,d]$ for a linear code of length $n$, dimension $k$ and distance $d$ (occasionally omitting the distance).

For a subset $\mathscr{X} = \{x^1, \dots, x^t\} \subset \mathscr{Q}^n$ define its *envelope* $e(\mathscr{X})$ by

$$e(\mathscr{X}) = \{x \in \mathscr{Q}^n : x_i \in \{x_i^1, \dots, x_i^t\}, i = 1, \dots, n\}.$$

Elements of the envelope $e(\mathscr{X})$ will be called descendants of $\mathscr{X}$. If $y \in e(\mathscr{X})$, we will call any of the elements of $\mathscr{X}$ parents of $y$.

The union of the envelopes of all the subsets of $\mathscr{C}$ of size $\leqslant t$ is called the *t-envelope* of the code

$$E_t(\mathscr{C}) = \bigcup_{\mathscr{X} \subset \mathscr{C},\, |\mathscr{X}| \leqslant t} e(\mathscr{X}).$$

The parent–descendant relation gives rise to several classes of codes that have recently gained considerable attention in the literature.

**Definition.** An $(n, M)$ code $\mathscr{C}$ has the *t-identifying parent property* if for any $y \in E_t(\mathscr{C})$,

$$P_{\mathscr{C}}(y) := \bigcap_{\substack{\mathscr{X} \subset \mathscr{C} \\ y \in e(\mathscr{X}),\, |\mathscr{X}| \leqslant t}} \mathscr{X} \neq \emptyset. \tag{1}$$

For brevity we call such codes $t$-i.p.p. codes. Informally, given a $t$-i.p.p. code it is possible for every vector $y \in E_t(\mathscr{C})$ to identify with certainty at least one of its parents.

$t$-i.p.p. codes and codes for related cryptographic problems received considerable attention in the recent years [1–8,10] (paper [10] introduced i.p.p. codes for the case $t = 2$), [13,14].

For $q \geqslant 3$, existence of 2-i.p.p. codes with rate $R$ bounded away from 0 for $n \to \infty$ was proved in [10]. For the case of general $t$, existence of i.p.p. codes of nonzero rate for any $q \geqslant t + 1$ was first proved in [4]. These papers also provided a characterization of 2- and 3-i.p.p. codes, respectively.

Let $R_q(t)$ denote the maximum asymptotically attainable rate of $q$-ary $t$-i.p.p. codes:

$$R_q(t) = \liminf_{n \to \infty} \max_{\mathscr{C} \subset \mathscr{Q}^n,\, \mathscr{C} \text{ is } t\text{-i.p.p.}} R(\mathscr{C}).$$

The following results are known about $R_q(t)$.

**Theorem 1.1** (Barg et al. [4]). *Let* $u = \lfloor (t/2 + 1)^2 \rfloor$. *Then*

$$R_q(t) \geqslant \frac{1}{u-1} \log_q \frac{(q-t)!q^u}{(q-t)!q^u - q!(q-t)^{u-t}}.$$

This result was proved by introducing $(t, u)$ hashing codes (functions) [4] and using the probabilistic method to derive an existence bound on their number. In particular, for $t = 2$ Theorem 1.1 coincides with an earlier result of Hollmann et al. [10]. Generally, we see that for $t$-i.p.p. codes to exist it is sufficient that $t \leqslant q - 1$. This is also a necessary condition because for $q \leqslant t$ there are only trivial identifying codes: the maximal size of a $t$-i.p.p. code, $t \geqslant q$, is $M = t$.

The probabilistic argument for $(t, u)$ hashing was refined in [1] resulting in the following estimate which improves the result of Theorem 1.1 for $q = t + 1$ and all $t$ except $t = 2, 3$.

**Theorem 1.2** (Alon et al. [1]).

$$R_{t+1}(t) \geqslant \frac{1}{u-1} \frac{t!(u-t)^{u-t}}{u^u \ln(t+1)}.$$

The procedure of finding a codeword $x$ in an i.p.p. code $\mathscr{C}$ which necessarily is a parent of a given word $y$, i.e., $x \in P_{\mathscr{C}}(y)$, is called *identification* (or *tracing traitors*) [7]. It is clear that for any $y \in E_t(\mathscr{C})$, we can identify at least one of its parents by examining all the $\binom{M}{t}$ subsets of $\mathscr{C}$ of size $t$. For codes of large size this requires prohibitively long computation. Therefore, the following open problem was raised in [13]:

> Given a $t$-i.p.p. code, traceability can be done in time $O(\binom{M}{t})$. Can this be improved, perhaps, for certain subclasses of $t$-i.p.p. codes?

Before stating our results, let us isolate the range of parameters where this problem can be addressed by a straightforward application of error-correcting properties of codes. For that note that there is a subclass of i.p.p. codes, called *t-traceability* codes, for which identification can be performed more efficiently than in the general case. For a code $\mathscr{C}$ and a vector $y \in \mathscr{Q}^n$ let

$$\mathrm{d}(y, \mathscr{C}) = \min_{x \in \mathscr{C}} \mathrm{d}(y, x).$$

For a vector $y \in \mathscr{Q}^n$ let

$$S_{\mathscr{C}}(y) = \{c \in \mathscr{C} : \mathrm{d}(y, c) = \mathrm{d}(y, \mathscr{C})\}$$

be the set of the nearest neighbors in $\mathscr{C}$ of $y$.

**Definition.** A $t$-i.p.p. code is said to have the *t-traceability property* if for every $y \in E_t(\mathscr{C})$, there holds

$$S_{\mathscr{C}}(y) \subset P_{\mathscr{C}}(y), \tag{2}$$

i.e., if $c$ is a nearest (in $\mathscr{C}$) codevector to $y$ then it is also one of its parents. Hence, in a traceability code a parent of a given word $y \in E_t(\mathscr{C})$ can be identified by finding a closest codeword to $y$. This requires examining at most $M$ codewords; thus, identification for traceability codes can be performed faster than for i.p.p. codes in general.

One simple sufficient condition for a code to possess the $t$-traceability property was given in [7].

**Proposition 1.3** (Chor et al. [7]). *Let $\mathscr{C}$ be a $q$-ary code with minimum code distance $d$, where $d > n(1 - 1/t^2)$ for some $t > 0$. Then $\mathscr{C}$ is a $t$-traceability code. In particular, for any vector $y \in E_t(\mathscr{C})$ there exists a vector $x \in S_{\mathscr{C}}(y)$ such that $s(y,x) \geqslant n/t$; any such vector also satisfies $x \in P_{\mathscr{C}}(y)$.*

**Proof.** Let $\mathscr{X} = \{x^1, \ldots, x^t\} \subset \mathscr{C}$ be a $t$-coalition and let $y \in e(\mathscr{X})$. By the definition of the envelope, for some $x \in \mathscr{X}$,

$$s(x,y) \geqslant n/t.$$

On the other hand, for any codevector $w$ outside $\mathscr{X}$,

$$s(y,w) \leqslant \sum_{j=1}^{t} s(w,x^j) = \sum_{j=1}^{t} (n - \mathrm{d}(w,x^j)) < n/t. \qquad \square$$

This proposition together with some known results on error-correcting codes leads to a more accurate formulation of the above problem. Namely, it is clear that a solution to it can be obtained if we can construct a sequence of $t$-traceability codes equipped with an efficient algorithm of finding a nearest codeword to a vector $y \in \mathscr{Q}^n$ (hence, in particular, to a vector $y \in E_t(\mathscr{C})$). By Proposition 1.3 such a sequence can be obtained from any family of error-correcting codes with large minimum distance. A natural candidate is a family of $q$-ary $[N, RN, \delta N]$ linear algebraic–geometric (AG) codes whose relative distance $\delta$ can be made arbitrarily close to one for sufficiently large $q$ [15]. Moreover, a polynomial-time list decoding algorithm of Guruswami and Sudan [9] enables us to find a nearest codeword $x$ to a given point $y$ provided that $s(y,x) \geqslant N\sqrt{1-\delta}$. Choosing $\delta$ sufficiently close to 1, we obtain a sequence of efficient $t$-i.p.p. codes with the traceability property.[3]

However, to construct codes of large size following this approach, one has to employ codes over an alphabet of a fairly large size. Namely, recall that by the Plotkin upper bound (see [15]) for any $q$-ary $(n, M)$ code with distance $d$ its cardinality

$$M \leqslant \frac{qd}{qd - (q-1)n}.$$

Substituting $d > n(1 - 1/t^2)$, we see that for $q < t^2$ the denominator is a positive integer, and therefore, the size of the code $M \leqslant qn$. Thus, for $q < t^2$ it is not possible to construct codes whose rate $R$ remains bounded away from zero as $n$ grows. Therefore, the nontrivial part of the above problem can be formulated as follows:

> For $t^2 \geqslant q \geqslant t + 1$ construct a sequence of $q$-ary $t$-i.p.p. codes with rate $R > 0$ and an identification algorithm of complexity polynomial in the code length $n$.

A solution to this problem is given in this paper. Moreover, the family of codes which we present is also polynomial-time constructible, and affords a

[3] After presenting the results of this work at a conference in October 2001 [5] we became aware of the paper [12] (published in December of the same year) that works out the details of this idea.

polynomial-time algorithm of assigning fingerprints to registered users. We will call a code family with these properties a *polynomial-complexity* code family. Our main result is given by the following theorem.

**Theorem 1.4.** *For any $q \geqslant t+1$ and any rate*

$$0 \leqslant R < \frac{1}{t^2} R_q(t),$$

*there exists a polynomial-complexity family of q-ary $(n, q^{Rn})$ t-i.p.p. code.*

## 2. Proof: a polynomial-complexity family of $t$-i.p.p. codes

Now we will construct a family of $q$-ary codes $\mathscr{C}(n, M)$ by concatenating a $q$-ary $t$-i.p.p. code $\mathscr{V}(m, q^{mR(\mathscr{V})})$ and a $Q$-ary linear $[N, K, D]$ code $\mathscr{W}, Q = q^{mR(\mathscr{V})}$. The parameters of the code $\mathscr{C}$ are: length $mN$, rate $R(\mathscr{C}) = R(\mathscr{V})R(\mathscr{W})$. By carefully choosing the parameters of the constituent codes we will ensure that $R(\mathscr{C}) \geqslant \text{const} > 0$ for an infinite sequence of values of the code length. The value of $q$ is fixed, independent of the code length. Finally, $t$ can be any number such that $t+1 \leqslant q$.

First, we recall the concatenated construction. There are two codes: the inner $q$-ary $(m, M)$ code $\mathscr{V}$ and the outer $Q$-ary $[N, K]$ code $\mathscr{W}$, where $Q \leqslant M = q^{mR(\mathscr{V})}$. For the reasons which will become clear later we take $Q$ to be the largest even power of a prime less than or equal to $q^{mR(\mathscr{V})}$. Let us fix an arbitrary injective mapping $\phi$ from $F_Q$ to $\mathscr{V}$. The code $\mathscr{C}$ can be viewed as a composite mapping $\mathscr{C} : (F_Q)^K \to (F_q)^{mN}$ formed of the following two mappings. First, a $K$-vector from $F_Q$ is encoded by the code $\mathscr{W}$ into a vector $w$ in $(F_Q)^N$. Next, every coordinate of $w$ is "encoded" with the code $\mathscr{V}$, i.e., mapped by $\phi$ to a vector in $F_q$. This results into a codeword in $(F_q)^{Nm}$, which is a word of the concatenated code $\mathscr{C}$. It is convenient to have in mind a representation of this codeword as a matrix with $N$ columns and $m$ rows. Clearly, $\mathscr{C}$ is an $(n, Q^K)$ code with $n = mN$, $Q^K \approx q^{mNR(\mathscr{V})R(\mathscr{W})}$ and hence of rate $R \approx R(\mathscr{V})R(\mathscr{W})$.

Let $t \leqslant q-1$ be fixed. $\mathscr{V}$ is taken to be a $t$-i.p.p. $(m, q^{mR(\mathscr{V})})$ code whose existence is proved in Theorems 1.1 and 1.2. We also take $\mathscr{W}$ to be a $Q$-ary one-point AG code with the parameters $[N, K = R(\mathscr{W})N, D = \delta N]$. To obtain as high rate $R(\mathscr{C})$ as possible, we take AG codes from asymptotically maximal curves so that their parameters for large $N$ attain the bound [15]

$$\delta \geqslant 1 - R(\mathscr{W}) - (\sqrt{Q} - 1)^{-1} + o(1). \tag{3}$$

Let $y \in E_t(\mathscr{C})$ be an $n$-word, which is represented as an $(m \times N)$ matrix over $F_q$. Let $y^1, \ldots, y^N$ denote the columns of $y$, where $y^i \in E_t(\mathscr{V}), i = 1, \ldots, m$. A column $y^i$ can be "decoded" with the code $\mathscr{V}$, where by decoding we mean finding a parent of $y^i$ in $\mathscr{V}$. Our decoding algorithm of the code $\mathscr{C}$ is a two-stage procedure which in the initial stage employs $N$ parallel, independent decodings of the columns in $y$ with the code $\mathscr{V}$, and uses the Guruswami–Sudan (GS) algorithm [9] in the second stage to find a codeword from $P_{\mathscr{C}}(y)$.

Guruswami–Sudan decoding [9]. Let $\mathscr{W}$ be a $Q$-ary Reed–Solomon or an (one-point) AG code with the parameters $[N, RN, \delta N]$. Then for any vector $z \in F_Q^N$ the algorithm outputs a subset (also called a *list*) $Y_z \subset \mathscr{W}$ formed of all the vectors $w \in \mathscr{W}$ such that the distance $\mathrm{d}(w, z)$ satisfies $\mathrm{d}(w, z) \leqslant N(1 - \sqrt{1-\delta})$ or, equivalently the number of agreements $s(w, z) \geqslant N\sqrt{1-\delta}$. The size $|Y_z|$ of this list is bounded above by a polynomial function of $N$. The implementation complexity of the algorithm is also polynomial in $N$.

Given a number $T \geqslant N\sqrt{1-\delta}$ and an input vector $y$, the GS algorithm can be trivially modified to output a list $Y_z(T) = \{w \in \mathscr{W}\colon\ s(w, z) \geqslant T\}$. This can be accomplished either by discarding from the full list codevectors that are farther away from $y$ than $N - T$ or by an appropriate modification (and simplification) of the original GS algorithm. We will call this reduced decoding procedure the *reduced GS algorithm*.

Relying on $t$-i.p.p. codes and GS decoding, we will show that under certain conditions, the concatenated code $\mathscr{C}$ admits an efficient identification algorithm described as follows.

Identification Algorithm for the code $\mathscr{C}$.

- *Input*: An $(mN)$-vector $y = (y^1, y^2, \ldots, y^N) \in E_t(\mathscr{C})$, where $y^i = (y_1^i, y_2^i, \ldots, y_m^i)$ for $i = 1, 2, \ldots, N$.
- *Inner decoding*: For every $i, 1 \leqslant i \leqslant N$ find a vector $v^i \in P_{\mathscr{V}}(y^i)$. Form a vector $z = (z_1, z_2, \ldots, z_N) \in F_Q^N$, where $z_i = \phi^{-1}(v^i)$.
- *Outer decoding*: Decode $z$ with the code $\mathscr{W}$ using the reduced GS algorithm producing the list $Y_z(N/t)$.
- *Output*: Take an arbitrary vector $w$ from the list $Y_z(N/t)$ (we will show that the list is not empty). Encode $w$ with the map $\phi$ as described in the beginning of this section to obtain a vector $x \in \mathscr{C}$. Output $x$.

**Proposition 2.1.** *Consider a code $\mathscr{C}(n, q^{Rn})$ constructed by concatenating a q-ary t-i.p.p. $(m, q^{mR(\mathscr{V})})$ code $\mathscr{V}$ and a Q-ary $[N, R(\mathscr{W})N, \delta N]$ code $\mathscr{W}, \delta > 1 - 1/t^2$. The code $\mathscr{C}$ is t-i.p.p., and for every $y \in E_t(\mathscr{C})$ the Identification Algorithm will output a vector $x \in P_{\mathscr{C}}(y)$.*

**Proof.** Let $\mathscr{X} = \{x^1, \ldots, x^t\} \subset \mathscr{C}$ be the coalition which generated $y$, i.e., $y \in e(\mathscr{X})$. Denote by $w^1, \ldots, w^t$ the codewords of $\mathscr{W}$ that correspond to $x^1, \ldots, x^t$, respectively.

Denote by $x^{i,j} \in \mathscr{V}$ the $i$th column of the codeword $x^j, 1 \leqslant j \leqslant t$. Then

$$y^i \in e(x^{i,1}, \ldots, x^{i,t}), \quad 1 \leqslant i \leqslant N.$$

Furthermore, since $\mathscr{V}$ is a $t$-i.p.p. code, for each $1 \leqslant i \leqslant N$ there exists an algorithm (for instance, exhaustive search) that outputs $v^i$ which is one of the parent codewords of the column $y^i$. Let $j(i)$ be a number such that $v^i = x^{i,j(i)}$. Then

$z = (w_1^{j(1)}, \dots, w_N^{j(N)})$ and hence

$$\sum_{j=1}^{t} s(z, w^j) \geqslant N.$$

Therefore there exists at least one vector $\hat{w} \in \{w^1, \dots, w^t\}$, such that

$$s(z, \hat{w}) \geqslant N/t. \tag{4}$$

This vector $\hat{w}$ will be included in the list $Y_z(N/t)$ produced by the outer (reduced GS) decoding since $\delta > 1 - 1/t^2$. On the other hand, repeating the corresponding part of the proof of Proposition 1.3 shows that for any vector $w \in \mathscr{W} \backslash \{w^1, \dots, w^t\}$, the number of agreements with $z$ will be smaller than $N/t$ and hence such vectors are not contained $Y_z(N/t)$.

This proves that the Algorithm always produces a vector in $P_{\mathscr{C}}(y)$, and therefore also the fact that $\mathscr{C}$ is $t$-i.p.p. □

The rate of the concatenated code $\mathscr{C}$ equals $R(\mathscr{C}) = m^{-1}(\log_q Q) R(\mathscr{W}) = R(\mathscr{V}) R(\mathscr{W})$. If the inner and outer codes are chosen as described, then it is possible to construct a code whose rate $R(\mathscr{C})$ is

$$R(\mathscr{C}) \geqslant (R_q(t) - \varepsilon_1)\left(1 - \delta - \frac{1}{\sqrt{Q} - 1} - \varepsilon_2\right) \geqslant t^{-2} R_q(t) - \left(\varepsilon_1 + \frac{1}{\sqrt{Q} - 1} + \varepsilon_2\right),$$

where both $\varepsilon_1$ and $\frac{1}{\sqrt{Q-1}}$ tend to zero when $m$ tends to infinity, and $\varepsilon_2$ tends to zero when $N$ grows. Next, given the value of $R(\mathscr{C}) = R$ we choose the length $m$ of the inner code $\mathscr{V}$ sufficiently large (but fixed) so that $\varepsilon_1 + \frac{1}{\sqrt{Q-1}} + \varepsilon_2 \leqslant 1/2(\frac{1}{t^2} R_q(t) - R)$. This proves the estimate on the rate $R(\mathscr{C})$ from Theorem 1.4.

It remains to prove the claims about the complexity of the codes $\mathscr{C}$. The construction, encoding, and decoding of the code $\mathscr{V}$ whose length is fixed, are of constant complexity. Therefore, the construction and the encoding complexity of the codes $\mathscr{C}$ up to a constant factor are equal to the construction (resp., encoding) complexity of the sequence of AG codes, which is known to be polynomial (in fact, a very recent result of Shum et al. [11] provides for them a construction algorithm of low complexity $(N \log_Q N)^3$). The complexity of identification is governed by the decoding complexity of the GS algorithm which is known to be polynomial [9]. This concludes the proof of Theorem 1.4.

## 3. Tracing traitors

This section is devoted to an application of the above result to the “tracing traitors problem” [7,8]. Following these papers, we give a short description of how i.p.p. codes fit into a broader context of data distribution schemes that have to deal with pirate decoders. A more complete description is given in the introductory part of the paper [8].

Informally speaking, a traitor tracing scheme "helps trace the source of leaks when sensitive or proprietary data is made available to a large set of parties" [8]. More precisely, information is distributed in an encrypted form over a public channel and "any captured pirate decoder (which decrypts with success probability which is better than the probability of breaking the encryption scheme that is used) will correctly identify a traitor and will protect the innocent user if up to $t$ traitors collude and combine their keys" [8].

Denote by $M$ the number of users of the system. The information can be decrypted using a secret key $s$ which is transmitted over the same channel in an encrypted form. For encryption of the vector $s$ the vectors $s_1, \dots, s_{n-1}$ are chosen randomly by the distributor, and the vector $s_n$ is determined by the following condition: $s = \sum_{i=1}^{n} s_i (\text{mod } 2)$. Then the vectors $s_i, i = 1, \dots, n$ are encrypted in an enabling block (which is sent along with the encrypted information through the channel) in such a way that registered users can access them using their personal keys $k^j, j = 1, \dots, M$, where $k^j = (k_1^j, \dots, k_n^j)$. The personal keys of the users are kept secret and known only to the distributor. The $j$th user gets an access to the enabling block and uses the $i$th segment $k_i^j$ of the key $k^j$ to decipher $s_i$. Finally, the user recovers the secret key since $s = \sum_{i=1}^{n} s_i \ (\text{mod } 2)$. It is convenient to think of this procedure as applying decoder to the enabling block and obtaining $s$ as a result.

A strategy of the coalition of traitors can be to produce a pirate decoder relying on the knowledge of all their personal keys in an attempt to hide their identities. The task of the distributor is to design these keys in such a way that as long as the size of the coalition of traitors does not exceed a certain value $t$, it is always possible to retrieve at least one member of the coalition.

Let us enumerate the segments $k_i^j$ (for a given $i$) by elements of some finite alphabet $\mathcal{Q}$ of size $q$. For every personal key $k^j = (k_1^j, \dots, k_n^j)$ we substitute a $q$-ary vector $c^j = (c_1^j, \dots, c_n^j)$, where $c_i^j \in \mathcal{Q}$. Denote by $\mathcal{C}$ the code formed by all the vectors $c^j, j = 1, \dots, M$. A tracing traitor scheme is called *open* if the corresponding code $\mathcal{C}$ is known to all the users. In order to create a working pirate decoder the coalition should provide it with some key $k = k_1, \dots, k_n$ with no position $i$ left blank (otherwise the decoder cannot recover the corresponding $s_i$ and hence, $s$). Recall that despite the code $\mathcal{C}$ being public, the members of the coalition of users $j_1, \dots, j_t$ know only their own keys $k_i^j$. Therefore, every key segment $k_i$ in the pirate decoder should be contained in the set $\{k_i^{j_1}, \dots, k_i^{j_t}\}$. This means that the key of a pirate decoder lies in the envelope of the secret keys of the coalition members.

Thus, to construct an open $t$-resilient traceability scheme it is sufficient to construct a $t$-i.p.p. code over an alphabet of a suitably chosen size $q$. Traceability schemes are characterized by the key length $n$ and the length $r = nq$ of the enabling block (and, generally, the number of decryptions performed by a user, although in our context it always equals $n$). Such schemes are optimized to achieve small values of $n$ and $r$ as functions of the number of users $M$.

The paper [8] presents two open fully $t$-resilient tracing traitor schemes, i.e. schemes that enable the distributor to locate with certainty at least one traitor (and a few more schemes that allow for a nonzero probability of misidentification). The approach to traitor tracing described above corresponds to the first of these two schemes (the second scheme is a bit more complicated and does not corresponds directly to i.p.p. codes)

The parameters of the traitor tracing scheme in [8] are obtained by an application of the probabilistic method. In our notation they are summarized in the following theorem.

**Theorem 3.1** (Chor et al. [8]). *There exists an open t-traceability scheme with* $n = 4t^2 \ln M$ *and* $q = 2t^2$.

Proposition 1.3 together with a relation to i.p.p. codes outlined above implies that any code with a large minimum Hamming distance defines a traceability scheme. Therefore, it makes sense to compare the existence result of Theorem 3.1 to the parameters of the scheme obtained by choosing a code with minimum distance at least $n(1 - 1/t^2)$ whose rate attains the Gilbert–Varshamov bound. In this way we obtain a $t$-traceability scheme with $q = 2t^2$ and $n = 2t^2 \ln M/(2 \ln 2 - 1)$, which is only by a factor $(4 \ln 2 - 2)^{-1} \approx 1.29$ inferior to Theorem 3.1.

We will use $t$-i.p.p. codes to present constructive alternatives to the existence results just cited.

In coding terms the problem can be formulated as follows: *given t and M construct a q-ary t-i.p.p. code of size M so that the length n and the value r = qn are as small as possible*. One obvious avenue to follow is again to take a code with a stronger property, a $t$-traceability code. For instance, let us take $Q$-ary AG codes with relative distance $\delta > 1 - 1/t^2$ assuming that for large length $N$ the code parameters attain bound (3). For $Q = (2t^2)^2$ we obtain

$$1 - \frac{1}{t^2} < \delta = 1 - R - \frac{1}{2t^2 - 1} - o(1).$$

Omitting small terms, we obtain the following rate: $R = 1/(2t^2)$, or

$$\ln M = \ln(Q^{nR}) = \frac{n}{t^2} \ln 2t^2.$$

Therefore, $n = t^2 \ln M/\ln 2t^2$, and the length of the enabling block equals $r = 4t^6(\ln M)/(\ln 2t^2)$.

**Proposition 3.2.** *There is a constructive t-traceability scheme with* $n = t^2 \ln M/\ln 2t^2$ *and* $r = 4t^6 \ln M/\ln 2t^2$ *and with an identification algorithm of complexity polynomial in n.*

The comparison of the parameters of this scheme with Theorem 3.1 shows that from AG traceability codes we obtain a shorter key length but a longer enabling block. Recall that Theorem 3.1 is only an existence result. The parameters of constructive schemes cited in [8] are $n = O(t^6 \ln M), r = O(t^8 \ln M)$, so Proposition 3.2 gives a substantial improvement of that result.

Now let us use $t$-i.p.p. codes constructed in this paper to construct traceability schemes. Let

$$v = \frac{(q-t)!q^u}{(q-t)!q^u - q!(q-t)^{u-t}}.$$

Let us take a code of length $n$ and rate $R$ such that $R = (\log_q v)/t^2(u-1)$ obtained by combining Theorems 1.1 and 1.4. We have

$$n = (u-1)t^2 \ln M / \ln v.$$

Assuming that $q \gg t$, after a series of simplifications, we obtain

$$n = t^2(u-1)(e^{tu/q} - 1)\ln M.$$

Choosing $q = t^3$, we get for the code length the inequality $n \leqslant 2t^2(u-1)\ln M$ and consequently, $r \leqslant 2t^5(u-1)\ln M$. Again the parameters of this scheme are better than constructive schemes previously known.

## 4. Concluding remarks

(1) *Traceability codes*: As discussed above, traceability codes form a subclass of i.p.p. codes. Similarly to the problem addressed in this paper, one can ask if there exist $t$-traceability codes with $t^2 \geqslant q \geqslant t+1$ and code rate bounded away from zero for growing length $n$. This problem so far remains open.

(2) *Digital fingerprinting*: The technique of this paper can be employed to construct codes for a more difficult problem, that of digital fingerprinting. Codes for digital fingerprinting satisfy the same condition (1) as i.p.p. codes except that the envelope $e(\mathscr{X})$ is formed under a more relaxed rule: if $x_i^1 = x_i^2 = \cdots = x_i^t$, then also $y_i$ takes on the same value; however, if $|\{x_i^1, \ldots, x_i^t\}| \geqslant 2$, then $y_i$ can be any letter of the alphabet $\mathscr{Q}$. This problem was suggested in [6], where it was also observed that unless we allow some probability $p_e$ of identification error, there are no binary fingerprinting codes of rate $R > 0$. Binary fingerprinting codes with positive rate $R$ and $p_e = \exp(-\Omega(n))$ are constructed in [2]. These codes afford a $\mathsf{poly}(n)$-time identification of a vector from the set of parents of a given vector $y$. The decoding (identification) algorithm of Barg et al. [2] also relies upon list decoding algorithms of one-point AG codes, although some additional arguments are involved to estimate the probability $p_e$ of identification error.

The construction [2] makes use of a more powerful version of the GS algorithm associated with soft decision, or weighted decoding of the outer codes. Weighted decoding is proposed in the same paper [9] as the version of list decoding (the so-called hard decoding) employed in this paper. The difference between weighted decoding and hard decoding in our context can be briefly explained as follows. In the first step of our identification algorithm, for a given column $i$, we find an element $z_i$ of the field $F_Q$ that corresponds to a member of the coalition $\mathscr{X}$. The vector $z$ thus formed is then submitted to the decoding algorithm of the outer code. Note however that to find $z_i$ we first find a $t$-subset of vectors of the inner code (or of elements of the field) that supposedly forms the coalition. This $t$-subset may contain more information about the parents of $y$ than just one element. The purpose of weighted decoding is to make use of just this information. In other words, instead of decoding from one vector $z$ we attempt to decode from a subset of vectors of $(F_Q)^N$ obtained by taking a product of the $N$ $t$-subsets that correspond to the coordinates $i = 1, \dots, N$.

While in the problem of digital fingerprinting the use of weighted decoding yields an improvement in the rate and error bounds of the codes constructed, it does not seem to improve over hard decision decoding in the problem of identification with i.p.p. codes.

## References

[1] N. Alon, G. Cohen, M. Krivelevich, S. Litsyn, Generalized hashing and applications to digital fingerprinting, Proceedings of the 2002 IEEE International Symposium on Information Theory, Lausanne, Switzerland, p. 436.

[2] A. Barg, G.R. Blakley, G. Kabatiansky, Digital fingerprinting codes: problem statements, constructions, identification of traitors, IEEE Trans. Inform. Theory 49 (4) (2003) 852–865.

[3] A. Barg, G.R. Blakley, G. Kabatiansky, On digital fingerprinting codes, Proceedings of the 2001 IEEE International Symposium on Information Theory, Washington, DC, p. 161.

[4] 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) 423–431.

[5] A. Barg, G. Kabatiansky, A class of i.p.p. codes with efficient identification, Proceedings of the 39th Annual Allerton Conference on Communications Control and Computing, Monticello, IL, October 2001, pp. 885–890.

[6] D. Boneh, J. Shaw, Collusion-secure fingerprinting for digital data, IEEE Trans. Inform. Theory 44 (1998) 1905–1987.

[7] B. Chor, A. Fiat, M. Naor, Tracing traitors, Crypto'94, Lecture Notes in Computer Science, Vol. 839, 1994, pp. 257–270.

[8] B. Chor, A. Fiat, M. Naor, B. Pinkas, Tracing traitors, IEEE Trans. Inform. Theory 46 (2000) 893–910.

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

[10] H.D.L. Hollmann, J.H. van Lint, J.-P. Linnartz, L.M.G.M. Tolhuizen, On codes with the identifiable parent property, J. Combin. Theory Ser. A 82 (1998) 121–133.

[11] K.W. Shum, I. Aleshnikov, P.V. Kumar, H. Stichtenoth, V. Deolalikar, A low-complexity algorithm for the construction of algebraic-geometric codes better than the Gilbert–Varshamov bound, IEEE Trans. Inform. Theory 47 (6) (2001) 2225–2241.

[12] A. Silverberg, J. Staddon, J.L. Walker, Efficient traitor tracing algorithm using list decoding, ASIACRYPT 2001, Lecture Notes in Computer Science, Vol. 2248, 2001, pp. 175–192.

[13] J.N. Staddon, D.R. Stinson, R. Wei, Combinatorial properties of frameproof and traceability codes, IEEE Trans. Inform. Theory 47 (2001) 1042–1049.

[14] D.R. Stinson, R. Wei, Combinatorial properties and constructions of traceability schemes and frameproof codes, SIAM J. Discrete Math. 11 (1998) 41–53.

[15] M. Tsfasman, S. Vlăduţ, Algebraic-Geometric Codes, Kluwer, Dordrecht, 1991.