Journal of Combinatorial Theory, Series A 117 (2010) 1049–1057

Contents lists available at ScienceDirect

Journal of Combinatorial Theory, Series A

www.elsevier.com/locate/jcta

Traceability codes ☆

Simon R. Blackburn [a], Tuvi Etzion [b], Siaw-Lynn Ng [a]

[a] *Department of Mathematics, Royal Holloway, University of London, Egham, Surrey TW20 0EX, United Kingdom*
[b] *Computer Science Department, Technion, Israel Institute of Technology, Haifa 32000, Israel*

ARTICLE INFO

Article history:
Received 19 February 2009
Available online 11 March 2010

Keywords:
Traceability code
Probabilistic construction
IPP code

ABSTRACT

Traceability codes are combinatorial objects introduced by Chor, Fiat and Naor in 1994 to be used in traitor tracing schemes to protect digital content. A k-traceability code is used in a scheme to trace the origin of digital content under the assumption that no more than k users collude. It is well known that an error correcting code of high minimum distance is a traceability code. When does this 'error correcting construction' produce good traceability codes? The paper explores this question.

Let ℓ be a fixed positive integer. When q is a sufficiently large prime power, a suitable Reed–Solomon code may be used to construct a 2-traceability code containing $q^{\lceil \ell/4 \rceil}$ codewords. The paper shows that this construction is close to best possible: there exists a constant c, depending only on ℓ, such that a q-ary 2-traceability code of length ℓ contains at most $cq^{\lceil \ell/4 \rceil}$ codewords. This answers a question of Kabatiansky from 2005.

Barg and Kabatiansky (2004) asked whether there exist families of k-traceability codes of rate bounded away from zero when q and k are constants such that $q \leqslant k^2$. These parameters are of interest since the error correcting construction cannot be used to construct k-traceability codes of constant rate for these parameters: suitable error correcting codes do not exist when $q \leqslant k^2$ because of the Plotkin bound. Kabatiansky (2004) answered Barg and Kabatiansky's question (positively) in the case when $k = 2$. This result is generalised to the following: whenever k and q are fixed integers such that $k \geqslant 2$ and $q \geqslant k^2 - \lceil k/2 \rceil + 1$, or such that $k = 2$ and $q = 3$, there exist infinite families of q-ary k-traceability codes of constant rate.

☆ This research was partially supported by E.P.S.R.C. Grant EP/F056486/1.
E-mail addresses: s.blackburn@rhul.ac.uk (S.R. Blackburn), etzion@cs.technion.il (T. Etzion), s.ng@rhul.ac.uk (S.-L. Ng).

0097-3165/$ – see front matter
doi:10.1016/j.jcta.2010.02.009

1. Introduction

Traceability codes were first introduced by Chor, Fiat and Naor [7] in order to construct traitor tracing schemes. We need to introduce some notation before defining these codes.

Let F be a finite set of cardinality q. For q-ary words $\mathbf{x}, \mathbf{y} \in F^\ell$ of length ℓ, we write $d(\mathbf{x}, \mathbf{y})$ for the (Hamming) distance between $\mathbf{x}$ and $\mathbf{y}$. For a code $\mathcal{C} \subseteq F^\ell$, we write $d(\mathcal{C})$ for the minimum distance of $\mathcal{C}$. The *rate* of a q-ary code $\mathcal{C}$ of length ℓ is defined to be $(\log_q |\mathcal{C}|)/\ell$.

Let $P \subseteq F^\ell$ be a set of q-ary words of length ℓ. We define the set $\operatorname{desc}(P)$ of *descendants* of P to be the set of words whose components are chosen from the corresponding components of words in P:

$$\operatorname{desc}(P) = \left\{ \mathbf{w} \in F^\ell \;\middle|\; \forall i \in \{1, 2, \ldots, \ell\}\ \exists \mathbf{x} \in P\colon\ w_i = x_i \right\}.$$

For example, if $P = \{1111, 1231\}$ then

$$\operatorname{desc}(P) = \{1111, 1211, 1131, 1231\}.$$

We often abuse notation by writing $\operatorname{desc}(\mathbf{x}, \mathbf{y}, \ldots, \mathbf{z})$ for $\operatorname{desc}(\{\mathbf{x}, \mathbf{y}, \ldots, \mathbf{z}\})$.

Let k be an integer such that $k \geqslant 2$. Let $\mathcal{C} \subseteq F^\ell$ be a code. For a word $\mathbf{w} \in F^\ell$, we say that a codeword $\mathbf{x} \in \mathcal{C}$ is a (*possible*) *parent of* $\mathbf{w}$ if there exists a set $P \subseteq \mathcal{C}$ of k or fewer codewords such that $\mathbf{x} \in P$ and $\mathbf{w} \in \operatorname{desc}(P)$.

A code $\mathcal{C}$ is a *k-traceability code* (or a k-TA code) if the following condition is satisfied. For all words $\mathbf{w} \in F^\ell$, the set of codewords at minimum distance to $\mathbf{w}$ is contained in every set $P \subseteq \mathcal{C}$ with $|P| \leqslant k$ and $\mathbf{w} \in \operatorname{desc}(P)$. This condition means that if we are given a word $\mathbf{w}$ that is a descendant of an unknown set P of k or fewer codewords, we can deduce some information about P: the codewords at minimum distance to $\mathbf{w}$ all lie in P. The following example of a 2-traceability code of length 3 is simple to define, and seems to be new:

Example 1. Let $q = 2r + 1$ where r is a positive integer, and let $F = \{0, 1, \ldots, 2r\}$. Define $\mathcal{C} = \mathcal{C}_1 \cup \mathcal{C}_2 \cup \mathcal{C}_3$, where

$$\mathcal{C}_1 = \left\{ (0, i, i)\colon\ 1 \leqslant i \leqslant r \right\},$$
$$\mathcal{C}_2 = \left\{ (i, 0, r + i)\colon\ 1 \leqslant i \leqslant r \right\},$$
$$\mathcal{C}_3 = \left\{ (r + i, r + i, 0)\colon\ 1 \leqslant i \leqslant r \right\}.$$

Then $\mathcal{C}$ is a q-ary 2-traceability code of length 3, containing $3r = (3/2)(q - 1)$ codewords.

An error correcting code of high minimum distance is a k-traceability code. More precisely, the following result is due to Chor, Fiat and Naor [7] (a proof can also be found in Blackburn [4]):

Theorem 1. *Let $\mathcal{C}$ be a q-ary error correcting code of length ℓ. If $d(\mathcal{C}) > \ell - \lceil \ell/k^2 \rceil$ then $\mathcal{C}$ is a k-traceability code.*

This theorem is tight for MDS codes: see Jin and Blaum [11]. Fernandez, Cotrina, Soriano and Domingo [8] show that a linear code which satisfies a weaker condition than high minimum distance is a k-traceability code, but do not give any examples of codes meeting this weaker condition.

Most examples of k-traceability codes known to the authors are (explicitly or implicitly) constructed by exhibiting an error correcting code and then applying Theorem 1. This is certainly true for the traceability codes in Staddon, Stinson and Wei [16] and van Trung and Martirosyan [17]. An exception is a construction due to Lindkvist, Löfvenberg and Svanström [14]: they construct q-ary codes $T(M, q)$ that have M codewords and are of length $\binom{M}{q-1}$ whenever $M \geqslant q + 1 \geqslant 4$. They prove that $T(M, q)$ is a k-traceability code whenever

$$\frac{k-1}{k}\left(\binom{M}{q-1} - \binom{M-k}{q-1}\right) < \binom{M-1}{q-2} + \binom{M-k-1}{q-k-1},$$

an inequality that is always satisfied when $k = 2$. Note however that these codes are small: their rates tend to zero very rapidly. Example 1 is unusual in that it is a traceability code of short length that cannot be constructed using Theorem 1. Indeed, the code is larger than any traceability code constructed using Theorem 1: to see this, note that Theorem 1 constructs 2-traceability codes of length 3 from error correcting codes with minimum distance 3, and so codes constructed using Theorem 1 contain at most q codewords.

Codes such as Example 1 open up the possibility that there might exist traceability codes that are much larger than the error correcting codes of high minimum distance required by Theorem 1. So when are the codes constructed by Theorem 1 good? This paper explores this question.

When q is a sufficiently large prime power, q-ary Reed–Solomon codes give codes of length ℓ and minimum distance $\ell - \lceil \ell/4 \rceil + 1$ containing $q^{\lceil \ell/4 \rceil}$ codewords. Thus, by Theorem 1, there exists a q-ary 2-traceability code of length ℓ with $q^{\lceil \ell/4 \rceil}$ codewords. Blackburn [4, Open problem 1] and Kabatiansky [13] asked whether the rate of such codes (approximately $1/4$) is the best possible for 2-traceability codes. We prove the following upper bound on the size of a 2-traceability code which answers this question:

Theorem 2. *Let ℓ be a positive integer. Then there exists a constant c, depending only on ℓ, with the following property. A 2-traceability code $\mathcal{C}$ of length ℓ has at most $cq^{\lceil \ell/4 \rceil}$ codewords.*

We can interpret Theorem 2 as implying that Theorem 1 is a good way of constructing 2-traceability codes when q is large, as it produces 2-traceability codes with an optimal number of codewords, up to a constant (though possibly large) factor.

Traceability codes are a special class of IPP codes: see Hollmann, van Lint, Linnartz and Tolhuizen [9], and Staddon, Stinson and Wei [16]. Blackburn [4] contains a survey of these codes (and related objects such as frameproof codes and secure frameproof codes). Barg, Blakley and Kabatiansky [2] discusses analogues of IPP codes with a more general notion of descendant. We note that Hollmann et al. [9] proved that a q-ary 2-IPP code of length ℓ contains at most $3q^{\lceil \ell/3 \rceil}$ codewords, but their methods do not extend to prove Theorem 2.

This same issue, the question of whether error correcting codes of high minimum distance give good codes, is at the core of the following question due to Barg and Kabatiansky [3]:

Question 1. Let k and q be such that $k + 1 \leqslant q \leqslant k^2$. Do there exist infinitely many q-ary k-traceability codes whose rate is bounded away from zero?

(It is not difficult to show that when $q \leqslant k$ a q-ary k-traceability code has at most q codewords, and so the rate cannot be bounded away from zero in this situation. This explains the lower bound on q in Question 1.) Theorem 1 cannot be used to answer Question 1, since the Plotkin bound (see van Lint [15, p. 67], for example) forbids the existence of codes with minimum distance large enough and of rate bounded away from zero. Kabatiansky [12,13] answered Barg and Kabatiansky's question (in the affirmative) by showing that such codes exist in the case when $k = 2$ and $q = 3$. His argument is probabilistic, and shows that randomly chosen linear ternary codes give rise to 2-traceability codes of non-zero rate. We show this phenomenon is not restricted to the case $k = 2$, by proving the following result:

Theorem 3. *Let k and q be integers such that $k \geqslant 2$. When*

$$k^2 - \lceil k/2 \rceil + 1 \leqslant q$$

or when $k = 2$ and $q = 3$, the following statement holds. There exists a positive constant R (depending on q and k) and a sequence of q-ary k-traceability codes $\mathcal{C}_1, \mathcal{C}_2, \ldots$ with the property that $\mathcal{C}_\ell$ has length ℓ and $|\mathcal{C}_\ell| \sim q^{R\ell}$ as $\ell \to \infty$.

One interpretation of Theorem 3 is that the codes constructed by Theorem 1 are far from optimal when q is fairly small and ℓ is large.

$$\begin{array}{rccc} & \overbrace{}^{A} & \overbrace{}^{I} & \overbrace{}^{D} \\ \mathbf{x}\in\mathcal{C}' & 0000\cdots00 & 0000\cdots00 & 0000\cdots00 \\ \mathbf{y}\in\mathcal{C}' & 0000\cdots00 & 1111\cdots11 & 1111\cdots11 \\ \mathbf{z}\in\mathcal{C} & ****\cdots** & 0000\cdots00 & ****\cdots** \\ \mathbf{w}\in\operatorname{desc}(\mathbf{y},\mathbf{z}) & \underbrace{0000\cdots00}_{|A|=t+\delta} & \underbrace{0000\cdots00}_{|I|=t} & \underbrace{1111\cdots11}_{|D|=2t-\delta} \end{array}$$

Fig. 1. When $t<\delta<2t$.

Like Kabatiansky, we use a probabilistic method. However, the details of our argument are different. (Indeed, we do not restrict ourselves to linear codes, and the condition our code needs to satisfy differs from Kabatiansky's.)

The structure of the rest of this paper is as follows. We prove Theorem 2 in Section 2, and we prove Theorem 3 in Section 3. Finally, we conclude with some open problems in Section 4.

2. An upper bound on 2-traceability codes

We aim to prove Theorem 2 in this section. The following lemma is easy to prove.

Lemma 1. *Let* $\mathbf{x}$, $\mathbf{y}$ *and* $\mathbf{w}$ *be words. Then* $\mathbf{w}\in\operatorname{desc}(\mathbf{x},\mathbf{y})$ *if and only if*

$$d(\mathbf{x},\mathbf{w})+d(\mathbf{w},\mathbf{y})=d(\mathbf{x},\mathbf{y}).$$

For a code $\mathcal{C}$ of length ℓ, a codeword $\mathbf{x}\in\mathcal{C}$ and a subset $I\subseteq\{1,2,\ldots,\ell\}$ of positions, define

$$F_{\mathcal{C}}(\mathbf{x},I)=\bigl|\{\mathbf{y}\in\mathcal{C}\colon\ x_i=y_i \text{ for all } i\in I\}\bigr|.$$

Lemma 2. *Let* t *be a fixed positive integer, and let* $\ell=4t$. *There exists a constant* c' (*depending only on* ℓ) *with the following property. Suppose that* $\mathcal{C}$ *is a* q*-ary* 2*-traceability code of length* ℓ *containing two or more codewords. Then there is a set* X *of at most* $c'q^t$ *codewords such that the subcode* $\mathcal{C}'=\mathcal{C}\setminus X$ *of* $\mathcal{C}$ *has* $d(\mathcal{C}')\geqslant d(\mathcal{C})+1$.

Proof. Suppose that $d(\mathcal{C})>\ell-t$. The Singleton bound (see van Lint [15, p. 67], for example) implies that $|\mathcal{C}|\leqslant q^t$, and so we may take $X=\mathcal{C}$ and $\mathcal{C}'=\emptyset$ in this case. Thus we may assume that $d(\mathcal{C})\leqslant\ell-t=3t$.

Suppose that $d(\mathcal{C})\leqslant t$. Define a subcode $\mathcal{C}'$ of $\mathcal{C}$ by removing all codewords in $\mathcal{C}$ that possess t positions that are not shared with another codeword. So

$$\mathcal{C}'=\bigl\{\mathbf{x}\in\mathcal{C}\colon\ F_{\mathcal{C}}(\mathbf{x},I)>1 \text{ for all } t\text{-subsets } I\subseteq\{1,2,\ldots,\ell\}\bigr\}.$$

Note that $|X|=|\mathcal{C}\setminus\mathcal{C}'|\leqslant\binom{\ell}{t}q^t$. We claim that there are no distinct codewords $\mathbf{x},\mathbf{y}\in\mathcal{C}'$ with $d(\mathbf{x},\mathbf{y})=d(\mathcal{C})$. Assume, for a contradiction, that such a pair exists. Let I be a t-subset of positions that contains all positions where $\mathbf{x}$ and $\mathbf{y}$ disagree. Note that I exists, since $d(\mathcal{C})\leqslant t$. Let $\mathbf{z}\in\mathcal{C}\setminus\{\mathbf{x}\}$ be such that $x_i=z_i$ for $i\in I$. Note that a choice for $\mathbf{z}$ exists, since $F_{\mathcal{C}}(\mathbf{x},I)\geqslant 2$ by the definition of $\mathcal{C}'$. But then $\mathbf{x}\in\operatorname{desc}(\mathbf{y},\mathbf{z})$, which contradicts the fact that $\mathcal{C}$ is a 2-traceability code. Thus $d(\mathcal{C}')>d(\mathcal{C})$, and so the lemma follows in this case. Thus we may assume that $d(\mathcal{C})>t$.

Write $d(\mathcal{C})=\ell-(t+\delta)$ for some integer δ. The previous two paragraphs show that we may assume that $0\leqslant\delta<2t$.

Define $\mathcal{C}'$ by

$$\mathcal{C}'=\left\{\mathbf{x}\in\mathcal{C}\colon\ F_{\mathcal{C}}(\mathbf{x},I)>\binom{\ell-t}{\delta+1} \text{ for all } t\text{-subsets } I\subseteq\{1,2,\ldots,\ell\}\right\}.$$

Note that

$$
\begin{array}{rcccc}
 & \overbrace{}^{A} & \overbrace{}^{I} & \overbrace{}^{J} & \overbrace{}^{D\setminus J} \\
\mathbf{x}\in\mathcal{C}' & 0000\cdots00 & 0000\cdots00 & 0000\cdots00 & 0000\cdots00 \\
\mathbf{y}\in\mathcal{C}' & 0000\cdots00 & 1111\cdots11 & 1111\cdots11 & 1111\cdots11 \\
\mathbf{z}\in\mathcal{C} & ****\cdots** & 0000\cdots00 & 2302\cdots05 & ****\cdots** \\
\mathbf{w}\in\operatorname{desc}(\mathbf{y},\mathbf{z}) & \underbrace{0000\cdots00}_{|A|=t+\delta} & \underbrace{0000\cdots00}_{|I|=t} & \underbrace{2302\cdots05}_{|J|=t-\delta} & \underbrace{1111\cdots11}_{|D\setminus J|=t}
\end{array}
$$

Fig. 2. When $0\leqslant\delta\leqslant t$.

$$\left|\mathcal{C}\setminus\mathcal{C}'\right|\leqslant\binom{\ell-t}{\delta+1}\binom{\ell}{t}q^t\leqslant 2^{2\ell}q^t.$$

We claim that there are no distinct codewords $\mathbf{x},\mathbf{y}\in\mathcal{C}'$ with $d(\mathbf{x},\mathbf{y})=d(\mathcal{C})$. To prove the lemma, it is sufficient to prove this claim. Assume, for a contradiction, that such a pair exists. Let A be the set of positions where $\mathbf{x}$ and $\mathbf{y}$ agree. So $|A|=t+\delta$. Let I be a t-subset of positions disjoint from A, so $x_i\neq y_i$ for all $i\in I$. Note that such a subset exists, since $d(\mathcal{C})\geqslant t$. Write D for the set of positions not in $A\cup I$. So $|D|=2t-\delta$. See Fig. 1 for an illustration of our notation. The minimum distance of $\mathcal{C}$ implies that a codeword is specified uniquely once $t+\delta+1$ of its components have been given. Thus there are at most $\binom{\ell-t}{\delta+1}$ codewords $\mathbf{c}\in\mathcal{C}$ such that $c_i=x_i$ for all $i\in I$ and such that $c_i=y_i$ for $\delta+1$ or more of the positions $i\in A\cup D$. Since $F_{\mathcal{C}}(\mathbf{x},I)>\binom{\ell-t}{\delta+1}$, there is at least one choice for $\mathbf{z}\in\mathcal{C}$ such that $z_i=x_i$ for $i\in I$ and such that $\mathbf{z}$ and $\mathbf{y}$ agree in at most δ positions. In particular, $d(\mathbf{z},\mathbf{y})\geqslant\ell-\delta$ and $\mathbf{z}\neq\mathbf{x}$.

Assume that $t<\delta<2t$. Define $\mathbf{w}\in\operatorname{desc}(\mathbf{y},\mathbf{z})$ by $w_i=z_i$ when $i\in I$ and $w_i=y_i$ otherwise. Note that $d(\mathbf{w},\mathbf{y})=t$, since for all $i\in I$ we have $w_i=z_i=x_i\neq y_i$. Moreover, by Lemma 1,

$$d(\mathbf{w},\mathbf{z})=d(\mathbf{y},\mathbf{z})-d(\mathbf{w},\mathbf{y})\geqslant\ell-\delta-t>t$$

since $\delta<2t$. So $\mathbf{w}$ is at distance t from its nearest parent. But $w_i=z_i=x_i$ whenever $i\in I$, and $w_i=y_i=x_i$ in the $t+\delta$ positions i where $x_i=y_i$. Thus $d(\mathbf{w},\mathbf{x})\leqslant\ell-t-(t+\delta)=2t-\delta<t$. Since $\mathbf{x}$ is not a parent, this contradicts the traceability property of the code, as required.

Finally, assume that $\delta\leqslant t$. At most δ positions in D are such that $y_i=z_i$, and so there are at least $2(t-\delta)$ positions $i\in D$ such that $y_i\neq z_i$ and $x_i\neq y_i$. Choose a set J of these positions of size $t-\delta$. (Note that this makes sense since $\delta\leqslant t$.) See Fig. 2 for an illustration of our situation. Define a descendent $\mathbf{w}\in\operatorname{desc}(\mathbf{y},\mathbf{z})$ by $w_i=z_i$ for $i\in I\cup J$ and $w_i=y_i$ otherwise. Note that $d(\mathbf{w},\mathbf{y})=2t-\delta$, since whenever $i\in I$ we have $w_i=z_i=x_i\neq y_i$ and whenever $i\in J$ we have that $w_i=z_i\neq y_i$ by our choice of J. Moreover, by Lemma 1,

$$d(\mathbf{w},\mathbf{z})=d(\mathbf{y},\mathbf{z})-d(\mathbf{w},\mathbf{y})\geqslant(\ell-\delta)-(2t-\delta)=2t\geqslant 2t-\delta,$$

so $\mathbf{w}$ is at distance $2t-\delta$ from its nearest parent. Note that $w_i=z_i=x_i$ when $i\in I$, and $w_i=y_i=x_i$ when $i\in A$. Thus

$$d(\mathbf{w},\mathbf{x})\leqslant\ell-(t+\delta)-t=2t-\delta.$$

Since $\mathbf{x}$ is not a parent, this contradicts the traceability property of the code, as required. □

Proof of Theorem 2. Write $\ell=4t-r$, where $t\in\mathbb{Z}$ and $0\leqslant r\leqslant 3$. By concatenating all codewords with the word 0^r, we may realise $\mathcal{C}$ as a traceability code of length $4t$. So we may assume that ℓ is divisible by 4.

Let $d=d(\mathcal{C})$. By applying Lemma 2 at most $\ell-d$ times, we obtain a code $\mathcal{C}'$ which has at most one codeword. We have removed at most $(\ell-d)c'q^t$ codewords to obtain $\mathcal{C}'$, and so $|\mathcal{C}|\leqslant(\ell-d)c'q^t+1\leqslant cq^t$ where we define $c=\ell c'$. So the theorem follows. □

3. Probabilistic existence results

The aim of this section is to establish Theorem 3. An outline of our proof is as follows. We pick a code at random. We define a 'bad' event to be a set $\{\mathbf{x}\}\cup P$ of $k+1$ codewords that contradicts the

k-traceability property: there is a descendent of P that is closer to another codeword $\mathbf{x}$ than to any of the codewords in P. We show that only a small number of codewords are involved in a bad event, and so once these codewords are removed we obtain a k-traceability code.

We will use the following consequence of the Chernoff bound, due to Janson [10] (see Bollobás [6, p. 12]). Recall that $\mathrm{Bin}(n,p)$ is the random variable taking the Binomial distribution with n trials and success probability p, so $\Pr(\mathrm{Bin}(n,p)=i)=\binom{n}{i}p^i(1-p)^{n-i}$ for $0\leqslant i\leqslant n$.

Lemma 3. *Let $p\in[0,1]$ and n be a positive integer. Then for all non-negative ϵ,*

$$\Pr\bigl(\mathrm{Bin}(n,p)\leqslant(p-\epsilon)n\bigr)\leqslant\exp\left(-\frac{\epsilon^2 n}{2p}\right).$$

Lemma 4. *Let $\mathbf{x},\mathbf{y}_1,\mathbf{y}_2,\ldots,\mathbf{y}_k\in F^\ell$ be chosen uniformly and independently at random. Let D be the random variable defined by*

$$D=\min\bigl\{d(\mathbf{x},\mathbf{z})\colon\ \mathbf{z}\in\mathrm{desc}(\mathbf{y}_1,\mathbf{y}_2,\ldots,\mathbf{y}_k)\bigr\}.$$

Define $\mu_0=(1-q^{-1})^k$. Then for any positive real number ϵ,

$$\Pr\bigl(D\leqslant(\mu_0-\epsilon)\ell\bigr)\leqslant\exp\left(-\frac{\epsilon^2\ell}{2\mu_0}\right).$$

Proof. For $i\in\{1,2,\ldots,\ell\}$, write D_i for the random variable defined to be 1 if $\mathbf{x}$ disagrees with all of $\mathbf{y}_1,\mathbf{y}_2,\ldots,\mathbf{y}_k$ in their ith positions, and is defined to be 0 otherwise. Note that $D_1,D_2,\ldots,D_\ell$ are independent, and each takes the value 1 with probability μ_0. Since $D=\sum_{i=1}^{\ell}D_i$, we find that $D=\mathrm{Bin}(\ell,\mu_0)$ and so the lemma follows by Lemma 3. □

Lemma 5. *Let $\mathbf{y}_1,\mathbf{y}_2,\ldots,\mathbf{y}_k\in F^\ell$ be chosen uniformly and independently at random, and let $P=\{\mathbf{y}_1,\mathbf{y}_2,\ldots,\mathbf{y}_k\}$. Let X be the maximum distance that any $\mathbf{z}\in\mathrm{desc}(P)$ can be from the set P. So X is the random variable defined by*

$$X=\max\Bigl\{\min\bigl\{d(\mathbf{z},\mathbf{y}_i)\colon\ i\in\{1,2,\ldots,k\}\bigr\}\colon\ \mathbf{z}\in\mathrm{desc}(P)\Bigr\}.$$

Define $\mu_1=\frac{k-1}{k}(1-q^{-(k-1)})$. Then for any positive real number ϵ,

$$\Pr\bigl(X\geqslant(\mu_1+\epsilon)\ell\bigr)\leqslant\exp\left(-\frac{k^2q^{k-1}\epsilon^2\ell}{2(k-1)^2}\right).$$

Proof. For words $\mathbf{y}_1,\mathbf{y}_2,\ldots,\mathbf{y}_k\in F^\ell$, define $f(\mathbf{y}_1,\mathbf{y}_2,\ldots,\mathbf{y}_k)$ to be the number of components where all of $\mathbf{y}_1,\mathbf{y}_2,\ldots,\mathbf{y}_k$ are equal. We claim that for any $\mathbf{z}\in\mathrm{desc}(\mathbf{y}_1,\mathbf{y}_2,\ldots,\mathbf{y}_k)$

$$\min\bigl\{d(\mathbf{z},\mathbf{y}_i)\colon\ i\in\{1,2,\ldots,k\}\bigr\}\leqslant\frac{k-1}{k}\bigl(\ell-f(\mathbf{y}_1,\mathbf{y}_2,\ldots,\mathbf{y}_k)\bigr).$$

To see this, let I be the set of positions where $\mathbf{y}_1,\mathbf{y}_2,\ldots,\mathbf{y}_k$ are not all equal, so $|I|=\ell-f(\mathbf{y}_1,\mathbf{y}_2,\ldots,\mathbf{y}_k)$. The definition of descendent implies that there exists a parent $\mathbf{y}_j$ that agrees with $\mathbf{z}$ on at least $1/k$ of the positions in I (and so disagrees with $\mathbf{z}$ on at most $\frac{k-1}{k}$ of the positions in I). Moreover, $\mathbf{y}_j$ clearly agrees with $\mathbf{z}$ on all positions not in I. Hence

$$\min\bigl\{d(\mathbf{z},\mathbf{y}_i)\colon\ i\in\{1,2,\ldots,k\}\bigr\}\leqslant d(\mathbf{z},\mathbf{y}_j)\leqslant\frac{k-1}{k}\bigl(\ell-f(\mathbf{y}_1,\mathbf{y}_2,\ldots,\mathbf{y}_k)\bigr),$$

and so our claim follows.

Define the random variable Y by

$$Y=\frac{(k-1)\ell}{k}-\frac{k-1}{k}f(\mathbf{y}_1,\mathbf{y}_2,\ldots,\mathbf{y}_k).$$

The argument in the paragraph above shows that

$$\Pr\big(X \geqslant (\mu_1+\epsilon)\ell\big) \leqslant \Pr\big(Y \geqslant (\mu_1+\epsilon)\ell\big),$$

and so it suffices to show that

$$\Pr\big(Y \geqslant (\mu_1+\epsilon)\ell\big) \leqslant \exp\left(-\frac{k^2 q^{k-1}\epsilon^2\ell}{2(k-1)^2}\right).$$

For $i \in \{1,2,\ldots,\ell\}$, define Y_i to be the random variable which is equal to 1 when all of $\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_k$ are equal at position i, and 0 otherwise. Clearly $Y_i = 1$ with probability $q^{-(k-1)}$, and $Y = \frac{(k-1)\ell}{k} - \frac{k-1}{k}\sum_{i=1}^{\ell} Y_i$. Since the random variables Y_i are independent, $\sum_{i=1}^{\ell} Y_i = \mathrm{Bin}(\ell, q^{-(k-1)})$, and so the definition of μ_1 implies that

$$\Pr\big(Y \geqslant (\mu_1+\epsilon)\ell\big) = \Pr\left(\mathrm{Bin}\big(\ell, q^{-(k-1)}\big) \leqslant \left(q^{-(k-1)} - \frac{k}{k-1}\epsilon\right)\ell\right).$$

The lemma now follows, by Lemma 3. □

Before we prove the main theorem of the section, we state the following technical lemma.

Lemma 6. *Let k and q be positive integers such that $k \geqslant 2$ and $q \geqslant 2$. Then*

$$(k-1)q\big(q^{k-1}-1\big) < k(q-1)^k \tag{1}$$

if and only if either $k = 2$ and $q = 3$ or

$$k^2 - \lceil k/2 \rceil + 1 \leqslant q. \tag{2}$$

Our proof of this lemma is straightforward, but is detailed and not especially illuminating. A brief outline of the proof is as follows. The lemma is easy to prove when $k = 2$, so we may assume that $k \geqslant 3$. Let β be a real number. To prove the lemma, define the real number q by $q = k^2 - (1/2)k + \beta$. It is sufficient to show that

$$\left(\frac{q}{q-1}\right)^k - \left(\frac{q}{(q-1)^k}\right) < \frac{k}{k-1} \tag{3}$$

holds when $\beta = 1/2$, but does not hold when $\beta = 0$. Expanding both sides of (3) as power series in k^{-1}, we find that the coefficients of k^{-i} agree when $i = 0, 1, 2$. The coefficient of k^{-3} on the left hand side of (3) is less than the coefficient of k^{-3} on the right hand side if and only if $\beta > 5/12$. This establishes our lemma whenever k is sufficiently large. Crude estimates for the absolute values of the $O(k^{-4})$ terms on both sides of (3) show that in fact the lemma holds for $k > 1000$. Finally some simple computations (we used Mathematica) verify that the inequalities are equivalent for $3 \leqslant k \leqslant 1000$.

Proof of Theorem 3. Let $\mathbf{x}, \mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_k \in F^{\ell}$ be chosen uniformly and independently at random. Let T be the event that there exists $\mathbf{z} \in \mathrm{desc}(\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_k)$ such that

$$d(\mathbf{x}, \mathbf{z}) \leqslant \min\big\{d(\mathbf{z}, \mathbf{y}_j):\ j \in \{1,2,\ldots,k\}\big\},$$

and define $p_0 = \Pr(T)$. We claim that there exists a positive constant R (depending only on q and k) such that

$$p_0 = o\big(q^{-kR\ell}\big) \tag{4}$$

as $\ell \to \infty$. Proving this claim is sufficient to establish the theorem, as the following argument shows.

Let ℓ be fixed. Define $M = \lfloor q^{R\ell} \rfloor$. Choose M codewords $\mathbf{c}_1, \mathbf{c}_2, \ldots, \mathbf{c}_M \in F^\ell$ uniformly and independently at random. For a sequence of distinct indices $i_0, i_1, \ldots, i_k \in \{1, 2, \ldots, M\}$, let $T_{(i_0,i_1,\ldots,i_k)}$ be the 'bad' event that there exists a descendant $\mathbf{z} \in \operatorname{desc}(\mathbf{c}_{i_1}, \mathbf{c}_{i_2}, \ldots, \mathbf{c}_{i_k})$ such that

$$d(\mathbf{c}_{i_0}, \mathbf{z}) \leqslant \min\bigl\{d(\mathbf{z}, \mathbf{c}_{i_j}) \colon\ j \in \{1, 2, \ldots, k\}\bigr\}.$$

(We call such an event bad, since the code $\{\mathbf{c}_i \colon\ i \in \{1, 2, \ldots, M\}\}$ is a k-traceability code of cardinality M if and only if none of the events $T_{(i_0,i_1,\ldots,i_k)}$ occur.)

Note that $\Pr(T_{(i_0,i_1,\ldots,i_k)}) = p_0$. By linearity of expectation, the expected number of bad events is $(M!/(M-k-1)!)p_0$, and so there is a choice of $\mathbf{c}_1, \mathbf{c}_2, \ldots, \mathbf{c}_M$ so that at most $\lfloor (M!/(M-k-1)!)p_0 \rfloor$ bad events occur. These bad events involve at most $(k+1)\lfloor (M!/(M-k-1)!)p_0 \rfloor$ codewords, and so by removing these codewords we obtain a k-traceability code $\mathcal{C}_\ell$ with M' codewords, where

$$M' \geqslant M - (k+1)\bigl\lfloor \bigl(M!/(M-k-1)!\bigr)p_0 \bigr\rfloor \geqslant M - (k+1)M^{k+1}p_0.$$

Our claim (4) implies that $(k+1)M^{k+1}p_0 = o(M)$ and so $M' \sim M \sim q^{R\ell}$. Thus the theorem follows once we have established our claim.

Define $\mu_0 = (1-q^{-1})^k$ and $\mu_1 = \frac{k-1}{k}(1-q^{-(k-1)})$. Our assumption on k and q together with Lemma 6 implies that $\mu_1 < \mu_0$. Let ϵ be a positive constant chosen so that $\mu_1 + \epsilon < \mu_0 - \epsilon$. Recall the definitions of $\mathbf{x}, \mathbf{y}_1, \ldots, \mathbf{y}_k$ and the event T from the first paragraph of the proof. Define the random variables D and X as in Lemmas 4 and 5. Note that

$$\begin{aligned}
\Pr(T) &\leqslant \Pr(D \leqslant X) \\
&\leqslant \Pr\bigl(D \leqslant (\mu_0 - \epsilon)\ell\bigr) + \Pr\bigl(X \geqslant (\mu_1 + \epsilon)\ell\bigr) \\
&\leqslant \exp\left(-\frac{\epsilon^2\ell}{2\mu_0}\right) + \exp\left(-\frac{k^2q^{k-1}\epsilon^2\ell}{2(k-1)^2}\right) \\
&\quad \text{(by Lemmas 4 and 5)} \\
&= o\bigl(q^{-kR\ell}\bigr)
\end{aligned}$$

where R is any constant such that

$$0 < R < \min\left\{\frac{\epsilon^2}{2k\mu_0 \log q}, \frac{kq^{k-1}\epsilon^2}{2(k-1)^2 \log q}\right\}.$$

Thus our claim (4) is established, and the theorem follows. □

4. Open problems

Can the bound of Theorem 2 be extended to k-traceability codes? (For IPP codes, the corresponding bound due to Hollmann et al. [9] does indeed generalise: see Alon and Stav [1] and Blackburn [5].) The following generalisation is the most natural one.

Question 2. Let k and ℓ be fixed positive integers such that $k \geqslant 2$. Does there exist a constant c (depending only on k and ℓ) such that the number of codewords in a q-ary k-traceability code of length ℓ is bounded above by $cq^{\lceil \ell/k^2 \rceil}$?

We believe this generalisation is true. It might be possible to extend the methods of Theorem 2 to settle this question, but we cannot currently see how this can be done.

Question 3. What is the best possible constant c in Theorem 2?

We see that we must have $c \geqslant 1$, by using a suitable MDS code and Theorem 1. Moreover, Example 1 shows that $c > 1$ in some situations. The constant implicit in the proof of Theorem 2 is exponential in ℓ: is this actually the case, or is this an artifact of our proof?

The following question is a very natural and interesting one, now some cases of Barg and Kabatiansky's question have been settled:

Question 4. For which values of q and k such that $k \geqslant 3$ and

$$k + 1 \leqslant q \leqslant k^2 - \lceil k/2 \rceil$$

is it the case that there exists an infinite family of q-ary k-traceability codes of rate bounded away from zero?

In particular, does there exist an infinite family of q-ary 3-traceabilty codes of rate bounded away from zero, when $4 \leqslant q \leqslant 7$? We do not see how the probabilistic methods of Theorem 3 can be used to answer this question; indeed, perhaps there exists a 'Plotkin bound' for traceability codes that forbids the existence of such codes.

Acknowledgments

The authors are grateful to G. Kabatiansky, and to an anonymous referee, for drawing our attention to Refs. [12] and [13].

References

[1] N. Alon, U. Stav, New bounds on parent identifying codes: the case of multiple parents, Combin. Probab. Comput. 13 (2004) 795–807.
[2] A. Barg, G.R. Blakley, G.A. Kabatiansky, Digital fingerprinting codes: problem statements, constructions, identification of traitors, IEEE Trans. Inform. Theory 49 (2003) 852–872.
[3] A. Barg, G. Kabatiansky, A class of I.P.P. codes with efficient identification, J. Complexity 20 (2004) 137–147.
[4] S.R. Blackburn, Combinatorial schemes for protecting digital content, in: C.D. Wensley (Ed.), Surveys in Combinatorics 2003, in: London Math. Soc. Lecture Note Ser., vol. 307, Cambridge University Press, Cambridge, UK, 2003, pp. 43–78.
[5] S.R. Blackburn, An upper bound on the size of a code with the k-identifiable parent property, J. Combin. Theory Ser. A 102 (2003) 179–185.
[6] B. Bollobás, Random Graphs, 2nd edition, Cambridge University Press, Cambridge, UK, 2001.
[7] B. Chor, A. Fiat, M. Naor, Tracing traitors, in: Y.G. Desmedt (Ed.), Advances in Cryptology – CRYPTO '94, in: Lecture Notes in Comput. Sci., vol. 839, Springer, Berlin, 1994, pp. 257–270.
[8] M. Fernandez, J. Cotrina, M. Sorario, N. Domingo, A note about the traceability properties of linear codes, in: K.-H. Nam, G. Rhee (Eds.), Information Security, Cryptology – ICISC 2007, in: Lecture Notes in Comput. Sci., vol. 4817, Springer-Verlag, Berlin, 2007, pp. 251–258.
[9] 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.
[10] S. Janson, On concentrations of probability, in: B. Bollobás (Ed.), Contemporary Combinatorics, in: Bolyai Soc. Math. Stud., vol. 10, János Bolyai Mathematical Society, Budapest, 2002, pp. 289–301.
[11] H. Jin, M. Blaum, Combinatorial properties for traceability codes using error correcting codes, IEEE Trans. Inform. Theory 53 (2007) 804–808.
[12] G.A. Kabatiansky, Good ternary 2-traceability codes exist, in: Proc. IEEE Symp. Inform. Theory, Chicago, IL, 2004, p. 203.
[13] G.A. Kabatiansky, Codes for copyright protection: the case of two pirates, Probl. Inf. Transm. 41 (2005) 182–186.
[14] T. Lindkvist, J. Löfvenberg, M. Svanström, A class of traceability codes, IEEE Trans. Inform. Theory 48 (2002) 2094–2096.
[15] J.H. van Lint, Introduction to Coding Theory, 3rd edition, Springer, Berlin, 1999.
[16] J.N. Staddon, D.R. Stinson, R. Wei, Combinatorial properties of frameproof and traceability codes, IEEE Trans. Inform. Theory 47 (2001) 1042–1049.
[17] T. van Trung, S. Martirosyan, On a class of traceability codes, Des. Codes Cryptogr. 31 (2004) 125–132.