On codes with the identifiable parent property

Henk D.L. Hollmann Jack H. van Lint Jean-Paul Linnartz
Ludo M.G.M. Tolhuizen*

September 1, 1997

*The authors are with Philips Research Laboratories, Prof. Holstlaan 4, 5656 AA Eindhoven, The Netherlands. Email: (hollmann linnartz tolhuizn)@natlab.research.philips.com . Jack H. van Lint is also with Eindhoven University of Technology, Dept. of Mathematics and Computer Science, P.O. Box 513, 5600 MB Eindhoven, The Netherlands.

Proposed running head: On codes with IPP.

Corresponding author:
Henk D.L. Hollmann
Philips Research Laboratories, Room WY 8.56
Prof. Holstlaan 4, 5656 AA Eindhoven, The Netherlands.
Email: hollmann@natlab.research.philips.com

Abstract

If C is a q-ary code of length n and $\mathbf{a}$ and $\mathbf{b}$ are two codewords, then $\mathbf{c}$ is called a descendant of $\mathbf{a}$ and $\mathbf{b}$ if $c_i \in \{a_i, b_i\}$ for $i = 1, \ldots, n$. We are interested in codes C with the property that, given any descendant $\mathbf{c}$, one can always identify at least one of the 'parent' codewords in C. We study bounds on $F(n, q)$, the maximal cardinality of a code C with this property, which we call the *identifiable parent property*. Such codes play a rôle in schemes that protect against piracy of software.

1 Introduction.

In this paper, we consider a code C of length n over an alphabet Q with $|Q| = q$ (i. e. $C \subset Q^n$). For any two words $\mathbf{a}$, $\mathbf{b}$ in Q^n we define the *set* of *descendants* $D(\mathbf{a}, \mathbf{b})$ by

$$D(\mathbf{a}, \mathbf{b}) := \{ \quad \mathbf{x} \in Q^n \quad | \quad x_i \in \{a_i, b_i\}, i = 1, 2, \ldots, n \quad \}. \tag{1}$$

Note that among the descendants of $\mathbf{a}$ and $\mathbf{b}$ we also find $\mathbf{a}$ and $\mathbf{b}$ themselves.

For a code C, we define the descendant code C^* by

$$C^* := \bigcup_{\mathbf{a} \in C, \mathbf{b} \in C} D(\mathbf{a}, \mathbf{b}). \tag{2}$$

For example, if C is the binary repetition code, then $C^* = \mathbf{F}_2^n$. Similarly, if C is the ternary Hamming code of length 4, then $C^* = \mathbf{F}_3^4$, since it is obvious that all words in a ball of radius 1 around a codeword are descendants of some pair containing that codeword. (For background information on coding theory, see e.g. [5].)

If $\mathbf{c} \in C^*$ is an element of $D(\mathbf{a}, \mathbf{b})$, with $\mathbf{a} \in C$, $\mathbf{b} \in C$, then we call $\mathbf{a}$ and $\mathbf{b}$ *parents* of $\mathbf{c}$. In general, an element of C^* has several pairs of parents. A trivial example are words of C themselves. We say that C has the "*identifiable parent property*" (IPP) if, for every descendant in C^*, at least one of the parents can be identified. In other words, for each $\mathbf{c} \in C^*$ there is a codeword $\pi(\mathbf{c})$ in C such that each parent pair of $\mathbf{c}$ must contain $\pi(\mathbf{c})$.

Example 1 Consider the ternary Hamming code C of length 4, which has size 9. Since every pair of distinct codewords has distance 3, any descendant $\mathbf{c}$ in C^* has distance ≤ 1 to exactly one of the parents in a parent pair. There cannot be two codewords with distance 1 to $\mathbf{c}$, so the unique codeword with distance ≤ 1 to $\mathbf{c}$ is the identifiable parent. For the other parent there are then three choices if $\mathbf{c} \notin C$ (and of course eight choices if $\mathbf{c} \in C$).

We leave it to the reader to verify the following.

Lemma 1 *A code $C \subseteq Q^n$ has IPP iff*

IPP1: $\mathbf{a}, \mathbf{b}, \mathbf{c}$ *distinct in* $C \Rightarrow a_i, b_i, c_i$ *distinct in* Q *for some* i,

IPP2: $\mathbf{a}, \mathbf{b}, \mathbf{c}, \mathbf{d} \in C$ *with* $\{\mathbf{a}, \mathbf{b}\} \cap \{\mathbf{c}, \mathbf{d}\} = \emptyset \Rightarrow \{a_i, b_i\} \cap \{c_i, d_i\} = \emptyset$ *for some* i.

Remark that the condition [IPP1] states that the code is *trifferent*, see [2],[3],[4].

We are interested in the *maximal size* of a code with the identifiable parent property. We define

$$F(n, q) := \max\{ \quad |C| \quad | \quad C \subseteq Q^n, C \text{ has IPP}, |Q| = q \quad \}.$$

Trivially, a code of cardinality 2 has IPP. If $q = 2$, a code of cardinality ≥ 3 does not have IPP. This follows from IPP1, but can also be seen directly: consider three distinct binary words $\mathbf{a}_1$, $\mathbf{a}_2$, $\mathbf{a}_3$. For $i = 1, \ldots, n$, the i-th coordinate of $\mathbf{c}$ is determined by a majority

vote over the corresponding coordinates of the three given words. Then $\mathbf{c}$ is clearly a descendant of any pair taken from the three words $\mathbf{a}_j$. So from now on we assume $q \geq 3$.

As trivial cases we have $F(1,q) = q$, $F(2,q) = q$. (If x_i, $i = 1, 2$, is a symbol that occurs twice as i-th coordinate, then (x_1, x_2) has no identifiable parent.)

Example 2 Take $n = 3$. Let $m := \lfloor \frac{q-1}{2} \rfloor$, $Q = \{0, 1, \ldots, q-1\}$. The code C consists of the following words :

(i) $(0, 0, 0)$,

(ii) $(0, i, i)$ with $1 \leq i \leq m$,

(iii) $(i, 0, i+m)$ with $1 \leq i \leq m$,

(iv) $(i, i, 0)$ with $m+1 \leq i \leq q-1$.

Clearly C has $q + m$ words. In each position, every non-zero symbol occurs in at most one codeword. So, if $\mathbf{c} \in C^*$ is not $\mathbf{0}$, then a unique parent is identifiable (possibly even both). If $\mathbf{c} = \mathbf{0}$, then $\mathbf{0}$ must be one of the parents.

We have shown that

$$F(3,q) \geq q + \lfloor \frac{q-1}{2} \rfloor. \tag{3}$$

We give a better bound in the next example. For the sake of simplicity, we do not treat the general case but assume that q has the form $q = r^2 + 1$.

Example 3 Let $q = r^2 + 1$, $Q = \{0, 1, \ldots, q-1\}$. We define two codes C_L and C_H as follows :

$$C_L := \{(a, b, ra + b) \quad | \quad 0 \leq a, b < r\},$$
$$C_H := \{(x, x, q-1) \quad | \quad r \leq x \leq q-1\}.$$

In C_L all words start with two *low* values ($< r$) and in C_H all words start with two *high* values ($\geq r$). The maximal value of the third coordinate in a word of C_L is

$$r(r-1) + r - 1 = r^2 - 1 = q - 2.$$

Therefore $C_L \cap C_H = \emptyset$ and furthermore, the third coordinate in a word of C_L clearly uniquely determines that word. Let $C := C_L \cup C_H$. We have just observed that a word in C^* with a third coordinate $< q-1$ has an identifiable parent in C_L. So, assume $\mathbf{c} \in C^*$ has the form $(c_1, c_2, q-1)$. If both c_1 and c_2 have low values, they uniquely determine a parent in C_L; if one of c_1, c_2 is high, say x, then $(x, x, q-1)$ must be a parent. So C has the identifiable parent property. We have shown

$$F(3, r^2+1) \geq 2r^2 - r + 1. \tag{4}$$

For the general case, a similar argument would lead to

$$F(3,q) \geq (\lfloor \sqrt{q-1} \rfloor)^2 + q - \lfloor \sqrt{q-1} \rfloor. \tag{5}$$

In subsequent sections, we investigate the behaviour of $F(n,q)$ for $n = 3$, $n = 4$, and general n. Our main results are Theorems 5 and 6, which provide upper and lower exponential bounds on $F(n,q)$. The problem of obtaining tight bounds for $F(4,q)$ remains open.

2 The case $n = 3$.

From Example 3, we first conjectured that $F(3,q)$ would behave roughly like $2q$ for $q \to \infty$. The following construction shows that in fact $F(3,q)$ increases at least like $3q$ (roughly). Again, we prefer simplicity and to achieve that, we assume that q has a special form, in this case $q = r^2 + 2r$.

Example 4 We divide the alphabet $Q := \{1, 2, \ldots, q\}$ into three disjoint classes S, M, and L, where
$S := \{1, 2, \ldots, r\}$, (the *small* numbers);
$M := \{r+1, r+2, \ldots, 2r\}$, (the *medium* numbers);
$L := \{2r+1, \ldots, r^2+2r\}$, (the *large* numbers).

The code C will be the union of three subcodes C_i $(i = 1, 2, 3)$, where

$$\begin{aligned} C_1 &:= \{(s_1, s_2, rs_1 + s_2 + r) \mid s_1 \in S, s_2 \in S\}; \\ C_2 &:= \{(m, sr + m, s) \mid m \in M, s \in S\}; \\ C_3 &:= \{(rm_1 + m_2 - r^2, m_1, m_2) \mid m_1 \in M, m_2 \in M\}. \end{aligned}$$

Observe that $C_1 \subset S \times S \times L$, $C_2 \subset M \times L \times S$, $C_3 \subset L \times M \times M$. So the codes C_i are disjoint and $C := C_1 \cup C_2 \cup C_3$ has cardinality $3r^2$. It is easy to see that C has the identifiable parent property by the following argument.

If a word $\mathbf{c} \in C^*$ has a large coordinate, then this coordinate uniquely determines a parent by the position of the large coordinate and the fact that in each C_i the large coordinates all occur exactly once. If none of the coordinates of $\mathbf{c}$ is large, then clearly the two subcodes from which the parents must come are determined. For one of these two subcodes we know two coordinates of the parent in that subcode. Again, these uniquely determine that parent. For example, if $\mathbf{c} = (c_1, c_2, c_3) \in M \times S \times S$, then the parents are in C_1 and C_2 and the pair $c_1 \in M$, $c_3 \in S$ uniquely determines the parent in C_2.

Therefore

$$F(3, r^2 + 2r) \geq 3r^2. \tag{6}$$

In a similar way one can treat the general case to show that $F(3,q)$ grows at least as fast as $3q - 12\sqrt{q}$.

We now aim to show that the previous result is essentially best possible, i. e. that $F(3,q)$ is roughly $3q$.

Consider a code C of length 3 over $Q := \{0, 1, \ldots, q-1\}$ with the identifiable parent property. We assume that $|C| > q$. We will consider a graph (which we shall also call C)

with the words of C as vertices. We join the codewords $\mathbf{c}_k$ and $\mathbf{c}_l$ by an edge of "*color*" i $(i = 1, 2, 3)$ if $\mathbf{c}_k$ and $\mathbf{c}_l$ have the same i-th coordinate.

Lemma 2 *In the graph C, no two vertices are joined by more than one edge.*

Proof. If the assertion is false, then w. l. o. g. we have $\mathbf{c}_1 = (0, 0, 0)$, $\mathbf{c}_2 = (0, 0, 1)$ joined by edges of color 1 and color 2. Let $\mathbf{c}_3 = (x, y, a)$. Then the descendant $(0, 0, a)$ shows that $a \notin \{0, 1\}$ and that no other codeword has a as third coordinate. In that case $|C| \leq q$, a contradiction. □

For the remainder of the discussion, we distinguish the alphabets that are used for the first, second, and third coordinate (say alphabet Q_i of size q_i for $i = 1, 2, 3$). We call C a (Q_1, Q_2, Q_3)-code. As an immediate corollary of Lemma 2 we have

$$|C| \leq q_i q_j \qquad (1 \leq i < j \leq 3). \tag{7}$$

This follows from the pigeonhole principle.

Clearly, for each of the colors i considered separately, C is a union of disjoint cliques. As a consequence of the identifiable parent property, there are two *forbidden subgraphs* in C.

Lemma 3 *(i) C does not contain a triangle $\{\mathbf{c}_1, \mathbf{c}_2, \mathbf{c}_3\}$ with edges of three different colors; (ii) C does not contain a chain $\mathbf{c}_1 \sim \mathbf{c}_2 \sim \mathbf{c}_3 \sim \mathbf{c}_4$, where the three edges have three different colors.*

Proof. (i) A triangle with edges of three different colors would imply that C has a subset of type

$$\{(a, b, x), (a, y, c), (z, b, c)\}$$

and then (a, b, c) is a descendant of any two of these words.

(ii) A chain with three different colors would imply that C has a subset of type

$$\{(a, x, y), (a, b, z), (u, b, c), (v, w, c)\}$$

and then (a, b, c) is a descendant of both the first and third and of the second and fourth element of this subset. □

We now consider the graph C, disregarding the colors for a moment. Pick a *connected* component S and then reconsider the colors occurring in S. We distinguish three cases, depending on the number of different colors in S. Clearly every alphabet Q_i can be split into two disjoint subsets Q_i' and Q_i'', such that S is a (Q_1', Q_2', Q_3')-code and $C \backslash S$ involves the other subalphabets.

Case (i) : All edges in S have color 1. Then S is a clique of color 1 and furthermore the second, respectively third coordinates of words in S are all different and (as observed above) do not occur in any other codeword of C. Here we obviously have $|Q_1'| = 1$.

Case (ii) : Two colors occur in S, say 1 and 2. In that case all words in S have a different third coordinate that does not occur in $C\backslash S$.

Case (iii) : If all three colors occur in S, then the forbidden configurations show that S must be the union of three cliques, of colors 1,2,3, respectively, that have exactly one common vertex. (To see this, first show that there exists a point incident with edges of all three colors.)

We can now estimate the cardinality of S in each of the three cases. In case (i) we trivially have

$$|S| = |Q_2'| = |Q_3'|. \tag{8}$$

In case (ii) it is again trivial that

$$|S| = |Q_3'|. \tag{9}$$

Case (iii) is more difficult to analyze. Let $\mathbf{c}$ be the common vertex of the three cliques. Let these cliques have cardinality s_1, s_2, s_3. All vertices different from $\mathbf{c}$ in the cliques of color 2, respectively color 3, have different first coordinates, also differing from the common first coordinate in the clique of color 1. Therefore

$$|Q_1'| = 1 + (s_2 - 1) + (s_3 - 1),$$

and similarly for Q_2' and Q_3'. Since $|S| = s_1 + s_2 + s_3 - 2$, we find

$$|S| = \frac{|Q_1'| + |Q_2'| + |Q_3'| - 1}{2}. \tag{10}$$

In all three cases, we have

$$|S| \leq |Q_1'| + |Q_2'| + |Q_3'| - 1. \tag{11}$$

This argument shows that C is the union of disjoint codes, all of which satisfy (11), hence the cardinality of C satisfies

$$|C| \leq |Q_1| + |Q_2| + |Q_3| - 1. \tag{12}$$

In our case, the three alphabets Q_i all have size q, so we conclude the following.

Theorem 1

$$F(3, q) \leq 3q - 1.$$

3 The case $n = 4$

We begin with an example showing that $F(4, q)$ behaves (roughly) at least like $q\sqrt{q}$. Again for simplicity, we assume that q is a square, say $q = r^2$. As letters of our alphabet Q (of size q) we take all pairs (a, b) from R^2, where $R := \{0, 1, \ldots, r-1\}$. We use addition mod r in R.

We define

$$C := \left\{ ((a_1, a_2), (a_1, a_3), (a_2, a_3), (a_1 + a_2, a_3)) \quad | \quad (a_1, a_2, a_3) \in R^3 \right\}.$$

Clearly $|C| = r^3 = q\sqrt{q}$. Note that a word in C is uniquely determined if two of its coordinates are known.

Let $(\alpha, \beta, \gamma, \delta) = ((\alpha_1, \alpha_2), (\beta_1, \beta_2), (\gamma_1, \gamma_2), (\delta_1, \delta_2))$ be a descendant in C^*. We distinguish two cases :

(i) Among α, β, γ two *obviously* are from different parents, say w. l. o. g. α and β. So $\alpha_1 \neq \beta_1$ and the parents look like

$$((\alpha_1, \alpha_2), (\alpha_1, x), (\alpha_2, x), (\alpha_1 + \alpha_2, x))$$

and

$$((\beta_1, y), (\beta_1, \beta_2), (y, \beta_2), (\beta_1 + y, \beta_2)).$$

If there is doubt about the parent that yielded γ, then we must have $\alpha_2 = \gamma_1 = y$, $x = \gamma_2 = \beta_2$. Then, since $\alpha_1 \neq \beta_1$, the coordinate δ uniquely determines one of the parents.

(ii) If we are not in case (i), then we must have $\alpha_1 = \beta_1$, $\alpha_2 = \gamma_1$, $\beta_2 = \gamma_2$. Two of the coordinates α, β, γ must come from the same parent and they uniquely determine this parent. Fortunately, each of the three possible pairs determine the *same* parent, namely $((\alpha_1, \alpha_2), (\alpha_1, \beta_2), (\alpha_2, \beta_2), (\alpha_1 + \alpha_2, \beta_2))$, which is therefore one of the parents. The other parent is one of the words ending in (δ_1, δ_2).

We have shown that C is a code of length 4 and cardinality $q\sqrt{q}$ that has IPP.

Alternative proof:
Note that C has minimum distance $d_H(C) = 3$. Hence IPP1 is trivially satisfied. Let $c(\alpha) = ((\alpha_1, \alpha_2), (\alpha_1, \alpha_3), (\alpha_2, \alpha_3), (\alpha_1 + \alpha_2, \alpha_3))$. Suppose that $\{c(\alpha)_i, c(\beta)_i\} \cap \{c(\gamma)_i, c(\delta)_i\} \neq \emptyset$, for $i = 1, \ldots, 4$. W.l.o.g we may assume that $c(\alpha)_1 = c(\gamma)_1$, whence $\alpha_1 = \gamma_1$ and $\alpha_2 = \gamma_2$. Since $d_H(C) = 3$, in the non-trivial case there is a permutation i_2, i_3, i_4 of $2, 3, 4$ such that $c(\alpha)_{i_2} = c(\delta)_{i_2}$, $c(\beta)_{i_3} = c(\gamma)_{i_3}$, and $c(\beta)_{i_4} = c(\delta)_{i_4}$. This implies $\alpha_3 = \delta_3$, $\beta_3 = \gamma_3$, and $\beta_3 = \delta_3$, whence also $\alpha_3 = \gamma_3$, so $\alpha = \gamma$ and $c(\alpha) = c(\gamma)$. Hence IPP2 is also satisfied, and C has IPP.

It is not possible to extend this code without losing IPP. To show this, assume that

$$\mathbf{x} = (x_1, x_2, x_3, x_4) = ((\alpha_1, \alpha_2), (\beta_1, \beta_2), (\gamma_1, \gamma_2), (\delta_1, \delta_2)) \notin C^*.$$

Consider $C' := C \cup \{\mathbf{x}\}$. Choose ξ such that $\xi \neq \alpha_2$, $\xi \neq \alpha_1 + \alpha_2 - \beta_1$.

The code C contains the following three *distinct* codewords :

$$\mathbf{u} = (u_1, u_2, u_3, u_4) = ((\alpha_1, \alpha_2), (\alpha_1, \beta_2), (\alpha_2, \beta_2), (\alpha_1 + \alpha_2, \beta_2)),$$

$$\mathbf{v} = (v_1, v_2, v_3, v_4) = ((\beta_1, \xi), (\beta_1, \beta_2), (\xi, \beta_2), (\xi + \beta_1, \beta_2)),$$

$$\mathbf{c} = (c_1, c_2, c_3, c_4) = ((\xi + \beta_1 - \alpha_2, \alpha_2), (\xi + \beta_1 - \alpha_2, \beta_2), (\alpha_2, \beta_2), (\xi + \beta_1, \beta_2)).$$

(These codewords are distinct because of the two restrictions that we made on ξ.) Now, both the pair $\{\mathbf{u}, \mathbf{v}\}$ and the pair $\{\mathbf{x}, \mathbf{c}\}$ have as descendant

$$(u_1, v_2, u_3, v_4) = (x_1, x_2, c_3, c_4).$$

We have shown that C is maximal with respect to IPP but of course not that $|C|$ is maximal. Indeed, in Example 1 we discussed a ternary code of size 9 with IPP. We consider this code as a code over an alphabet of size four (in which one of the letters is not used). Adding a word containing the remaining letter in each coordinate does not destroy IPP, hence we see that $F(4,4) \geq 10$, while the above construction with $q = 4$ produces a code of size 8.

The best upper bound that we could obtain (see the next section) shows that $F(4,q) = O(q^2)$. It would be interesting to try and close this gap.

4 Some comments on the general case

The form of the conditions IPP1 and IPP2 suggests the following construction by concatenation. Let $C \subseteq Q^n$, $|Q| = q$, and $D \subseteq R^m$ both have IPP, and suppose that $|C| = |R|$. By identifying C and R, we can consider the code D as a code of length nm over Q, and by applying IPP1 and IPP2 twice we see that this code over Q again has IPP. So we have proved the following result.

Theorem 2 *$F(nm, q) \geq F(m, F(n, q))$.*

For certain classes of codes, it is easy to see that IPP holds. We start with *equidistant* codes.

Theorem 3 *If C is an equidistant code of length n over an alphabet of size q and with distance d, then C has the identifiable parent property if d is odd or if d is even and $n < \frac{3}{2}d$.*

Proof. If $\mathbf{a} \in C$, $\mathbf{b} \in C$, and $\mathbf{c} \in D(\mathbf{a}, \mathbf{b})$, then clearly

$$d(\mathbf{a}, \mathbf{c}) + d(\mathbf{b}, \mathbf{c}) = d.$$

If d is odd then one of the words $\mathbf{a}$, $\mathbf{b}$ is the *unique* codeword with distance $< \frac{1}{2}d$ to $\mathbf{c}$. If d is even and there is doubt about the parents of a word $\mathbf{c} \in C^*$, then $\mathbf{c}$ must have distance $\frac{1}{2}d$ to at least three codewords. From this one immediately finds $n \geq \frac{3}{2}d$. □

To make other general statements, we first analyze what it means that a code C does not have IPP. One of two things can happen:

(i) There is a word $\mathbf{c} \in C^*$ such that each pair from $\{\mathbf{u}, \mathbf{v}, \mathbf{w}\}$ is a parent pair, where $\mathbf{u}$, $\mathbf{v}$, $\mathbf{w}$ are in C.

(ii) There is a word $\mathbf{c} \in C^*$ and four distinct words $\mathbf{u}$, $\mathbf{v}$, $\mathbf{x}$, $\mathbf{y}$ in C such that both $\{\mathbf{u}, \mathbf{v}\}$ and $\{\mathbf{x}, \mathbf{y}\}$ are parent pairs of $\mathbf{c}$.

We analyze case (i). Let d be the minimum distance of C. Let $d(\mathbf{u}, \mathbf{v}) = d_1$, $d(\mathbf{u}, \mathbf{w}) = d_2$, and $d(\mathbf{v}, \mathbf{w}) = d_3$. We now must have

$$n \leq (n - d_1) + (n - d_2) + (n - d_3),$$

i.e. $3d \leq 2n$.

In case (ii) we find in the same way that $4d \leq 3n$. It follows that if $d \geq \frac{3n+1}{4}$, then C has IPP. This implies the following theorem.

Theorem 4 *Let q be a prime power. If $q \geq n - 1$ then a (shortened, extended, or doubly extended) Reed-Solomon code over $\mathbf{F}_q$ with parameters $[n, \lceil \frac{n}{4} \rceil, n - \lceil \frac{n}{4} \rceil + 1]$ exists and has IPP.*

Corollary 1 *If $q \geq n - 1$ and q is a prime power, then $F(n, q) \geq q^{\lceil \frac{n}{4} \rceil}$.*

For example, if $q \geq 4$, there is a [5,2,4] MDS code C over $\mathbf{F}_q$ and hence $F(5, q) \geq q^2$. Indeed, for every word $\mathbf{c} \in C^*$ there is at least one word $\mathbf{a}$ in C (namely one of the parents) with distance at most 2 to $\mathbf{c}$. This must be a parent, since otherwise there would be two other parents and one of these would then have distance ≤ 3 to $\mathbf{a}$.

If we consider r-tuples of symbols from the alphabet Q of size q as symbols from the alphabet Q^r, then a code C of length $n = mr$ over Q that has IPP is also a code of length m over Q^r with IPP. This immediately implies the following theorem as a consequence of Theorem 1.

Theorem 5 *We have that $F(n, q) \leq 3q^{\lceil \frac{n}{3} \rceil}$.*

So, for instance, we see from this and Corollary 1 that for a prime power $q \geq 4$ we have

$$q^2 \leq F(5, q) \leq 3q^2.$$

(In fact, using (12), it is easy to sharpen this result to $q^2 \leq F(5, q) \leq 2q^2 + q - 1$.)

We shall prove a lower bound using the *Lovász Local Lemma* (cf. [1], see also [7], [6]).

Let $A_1, A_2, \ldots, A_n$ be events in a probability space and assume that $Pr(A_i) \leq p$ for each i. A graph G on the vertices $1, 2, \ldots, n$ is called a *dependency graph* for the events A_i $(i = 1, 2, \ldots, n)$ if for each i, the event A_i is independent of *every* subset of $\{A_j : \{i, j\} \notin E(G)\}$. One version of the Lovász Local Lemma states that if each vertex of G has degree $\leq d$ $(d \geq 1)$ and $4dp < 1$, then $Pr(\bar{A}_1 \wedge \bar{A}_2 \wedge \ldots \wedge \bar{A}_n) \neq 0$. For a proof by induction see [7], where a stronger version is given.

We now consider N words $\mathbf{c}_1, \mathbf{c}_2, \ldots, \mathbf{c}_N$ of length n over an alphabet of size q, where each word is chosen randomly. For each 4-tuple $X = \{i, j, k, l\}$ from $\{1, 2, \ldots, N\}$ the event A_X is : "The code $\{\mathbf{c}_i, \mathbf{c}_j, \mathbf{c}_k, \mathbf{c}_l\}$ does *not* have IPP or contains two equal words." If we define the graph G on the 4-tuples from $\{1, 2, \ldots, N\}$ by $X \sim X'$ if and only if $X \cap X' \neq \emptyset$, then G is a dependency graph for the events A_X. As we saw above, $Pr(A_X)$ is the probability that the 4-tuple contains a *bad triple*, i.e., a triple such that the three pairs from this triple have a common descendant, or that the 4-tuple is a *bad 4-tuple*, i.e., the 4-tuple can be split into two pairs that have a common descendant, or that the 4-tuple contains a *bad pair*, i.e., a pair of two equal words.

For the first of these, note that there are four ways to choose the triple and that for each coordinate position the probability that two or three of the codewords have the same coordinate in that position is less than $\frac{3}{q}$. For the second and third possibilities, we can argue similarly and thus find

$$Pr(A_X) \leq 4(\frac{3}{q})^n + 3(\frac{4}{q})^n + 6(\frac{1}{q})^n =: p.$$

Each vertex of G has degree

$$\begin{aligned} d &:= \binom{N}{4} - \binom{N-4}{4} \\ &= \frac{1}{24}\{16N^3 - 168N^2 + 632N - 840\}. \end{aligned}$$

Asymptotically, the condition $4dp < 1$ yields

$$N \lesssim \frac{1}{2}(\frac{q}{4})^{\frac{n}{3}}.$$

Application of Lovász's Local Lemma shows that if $4dp < 1$ then $Pr(\cap \bar{A}_X) > 0$, which means that a code with IPP exists. Since $p \leq 5(\frac{4}{q})^n$ and $d \leq \frac{2}{3}N^3$, $n \geq 3$, we have proved the following theorem.

Theorem 6 *For $n \geq 3$, there is a constant c such that*

$$F(n, q) \geq c \left(\frac{q}{4}\right)^{\frac{n}{3}}.$$

From our calculations above, it follows that we could take $c = 0.4$. For large q, Theorem 6 is better than Corollary 1.

Remark 1: J. Körner (private communication) suggested an alternative proof of Theorem 6 by means of the "expurgation method". Here, the idea is the following. Again, we choose at random N words of length n from an alphabet of size q. By linearity of expectation, the average number E of bad pairs, triples, and 4-tuples is

$$E \approx (\frac{1}{q})^n \binom{N}{2} + (\frac{3}{q})^n \binom{N}{3} + 3(\frac{4}{q})^n \binom{N}{4} \approx \frac{1}{8}N^4(\frac{4}{q})^n. \tag{13}$$

Now choose N such that

$$E \leq (1-\delta)N, \tag{14}$$

where δ will be specialized later. If we remove a word from each bad pair, triple, or 4-tuple, then the remaining collection C of words has IPP. We conclude that there is a collection C of size $|C| \geq \delta N$ that has IPP. Combining (13) and (14) shows that we may take

$$N \approx 2(1-\delta)^{1/3}(\frac{q}{4})^{n/3},$$

whence

$$|C| \geq 2(\delta^3(1-\delta))^{1/3}(\frac{q}{4})^{n/3}.$$

The above lower bound is optimal when $\delta = 3/4$, in which case we obtain that

$$|C| \geq (27/32)^{1/3}\,(\frac{q}{4})^{n/3}.$$

Remark 2: A more careful calculation of the relevant probabilities used in both proofs will show that

$$F(n,q) \geq c(q^3/(4q^2-6q+3))^{n/3}.$$

It follows that

$$f(3) := \liminf_{n\to\infty} n^{-1}\log F(n,3) \geq \log(q/(4q^2-6q+3)^{1/3}).$$

Remark 3: In both of the above constructions, it might be interesting to start with an alphabet R where the letters are identified with the codewords of a code with IPP of length m and size $|R| = F(m,q)$ over an alphabet Q of size q. Then the observation at the beginning of this section shows that the code that is obtained, when considered as a code over Q, again has IPP. (Cf. [2].) For example, when $q = 3$, we have that $F(4,3) = 9$, attained by the ternary Hamming code of length 4. Now by applying one of the above constructions with an alphabet size of 9 and using the result from Remark 2 we may conclude in this way that

$$f(3) \geq 12^{-1}\log(3^5/91).$$

Unfortunately, this is slightly worse than the bound $f(3) \geq 3^{-1}\log(9/7)$ obtained by a direct application of the result in Remark 2. Compare this with [2], where this idea leads to the best known lower bound for trifference. (Further details are left to the reader.)

5 Discussion

Multi-media publishers can "fingerprint" images by changing perceptually insignificant aspects in order to be able to trace violation of copyright restrictions. Here, the idea is

that if different customers receive a version of an image with different fingerprint, then the customer who illegally redistributes his or her version of the image can be traced. This paper investigates sets of "fingerprint codewords" with the property that if two users create a new image by combining parts of their images, then the new image reveals the identity of at least one of the source images. Our results show that for fixed alphabet size the maximal size of such codes grows exponentially with the codeword length.

6 Acknowledgements

We wish to thank our colleagues C.P.M.J. Baggen and A.G.C. Koppelaar for their interest in the problem. Also, we thank J. Körner for some helpful suggestions.

References

[1] P. Erdös and L. Lovász, Problems and results on 3-chromatic hypergraphs and some related questions, *in* "Infinite and finite sets" (A. Hajnal, R. Rado, and V. Sós, Eds), Coll. Math. Soc. J. Bolyai, Vol. 11, Budapest, 1973, 609–627.

[2] J. Körner and K. Marton, New bounds for perfect hashing via information theory, *Eur. J. Comb.* **9** (1986), 523–530.

[3] J. Körner and M. Lucertini, Compressing inconsistent data, *IEEE Trans. Inform. Theory* **40** (1994), 706–715.

[4] J. Körner and G. Simonyi, Trifference, *Studia Sci. Math. Hung.* **30** (1995), 95–103.

[5] F.J. MacWilliams and N.J.A. Sloane, "The Theory of Error-Correcting Codes", Amsterdam: Elsevier, 1977.

[6] J. B. Shearer, On a problem of Spencer, *Combinatorica* **5** (1985), 241–245.

[7] J. Spencer, Probabilistic methods, *Graphs and Combinatorics* **1** (1985), 357–382.