An Upper Bound on the Size of a Code with the k-Identifiable Parent Property

Simon R. Blackburn
Department of Mathematics
Royal Holloway, University of London
Egham, Surrey TW20 0EX
United Kingdom

July 19, 2002

Abstract

The paper gives an upper bound on the size of a q-ary code of length n that has the k-identifiable parent property. One consequence of this bound is that the optimal rate of such a code is determined in many cases when $q \to \infty$ with k and n fixed.

1 Introduction

The concept of a code with the identifiable parent property was introduced by Hollmann, van Lint, Linnartz and Tolhuizen [3] in 1998, motivated by an application in fingerprinting digital multimedia. Staddon, Stinson and Wei [4] generalised this concept to codes having the k-identifiable parent property (or k-IPP codes for short); we define k-IPP codes as follows.

Let Q be a finite set of size q and let n be a positive integer. For a word $x \in Q^n$, we write x_i for the ith component of x. Let $C \subseteq Q^n$ be a code, and let $X \subseteq C$ be a set of codewords. The *set of descendants* of X, written $\mathrm{desc}(X)$, is defined by

$$\mathrm{desc}(X) = \{d \in Q^n : \text{ for all } i \in \{1, 2, \ldots, n\}, d_i = x_i \text{ for some } x \in X\}.$$

(If the elements of X are thought of as the DNA strings of a closed population of organisms, then $\mathrm{desc}(X)$ is the set of possible DNA strings of a descendant of this population, assuming no mutations occur). A set $X \subseteq C$ is said to be a *parent set* of a word $d \in Q^n$ if $d \in \mathrm{desc}(X)$. For $d \in Q^n$, we write $\mathcal{H}_k(d)$ for the set of parent sets $X \subseteq C$ of d such that $|X| \leq k$.

A code C of length n over Q is said to be a *k-IPP code* if for all $d \in Q^n$, either $\mathcal{H}_k(d) = \emptyset$ or

$$\bigcap_{X \in \mathcal{H}_k(d)} X \neq \emptyset.$$

In other words, a code has the k-identifying parent property if whenever d is a descendant of k (or fewer) codewords, at least one of the parents of d may be identified.

This paper aims to prove a good upper bound on the maximal size of a q-ary k-IPP code C of length n. We aim to provide good bounds when the alphabet size is large. As a byproduct of the techniques we use, we obtain a shorter proof of one of the bounds given in the paper of Hollmann *et al* [3, Theorem 1].

The paper is organised as follows. In Section 2 we reprove Theorem 1 of the paper of Hollmann *et al.* This provides an introduction to the techniques we use in Section 3, where we prove our main result. Finally, Section 4 discusses the known existence results for k-IPP codes, and relates these to our upper bound. Our bound combines with these results to determine the asymptotic value (as $q \to \infty$ with n and k fixed) of the optimal rate of a q-ary k-IPP code of length n in many cases.

2 2-IPP Codes of Length 3

Hollman *et al* [3] proved that a q-ary 2-IPP code C of length 3 has at most $3q-1$ codewords. This section aims to reprove this bound, as an illustration of some of the techniques we will use to prove a more general bound in the next section. In fact, the proof we give will establish a (very slightly) stronger result:

Theorem 1 *Let C be a q-ary 2-IPP code of length 3. Then $|C| < 3q - 1$.*

Before embarking on the proof, we establish some notation. Let D be a set of q-ary words of length n. We define a graph $\Gamma(D)$, whose edges are labelled by elements of the set $\{1, 2, \ldots, n\}$, as follows. We take the vertex set of $\Gamma(D)$ to be D, and we join distinct vertices $a, b \in D$ by an edge labelled i if and only if $a_i = b_i$. So $\Gamma(D)$ might have several edges between a given pair of vertices, but contains no loops.

For $i \in \{1, 2, \ldots, n\}$, let $\Gamma_i(D)$ be the graph obtained by deleting all the edges in $\Gamma(D)$ other than those labelled i. The definition of $\Gamma(D)$ implies that $\Gamma_i(D)$ is a simple graph and is a disjoint union of at most q cliques. In particular, $\Gamma_i(D)$ has at most q isolated vertices. Indeed, when $|D| \neq q$, the graph $\Gamma_i(D)$ has at most $q-1$ isolated vertices.

The results contained in the following lemma are proved in the paper of Hollmann *et al* [3, Lemmas 2 and 3]. For the sake of completeness, we include a proof here.

Lemma 1 *Let C be a q-ary 2-IPP code of length 3.*

(i) $\Gamma(C)$ *does not contain a triangle whose edges are labelled with three different labels.*

(ii) $\Gamma(C)$ *does not contain a chain a, b, c, d whose edges ab, bc, cd are labelled* 1, 2 *and* 3 *respectively and where a, b, c, d are pairwise distinct.*

(iii) *When $|C| > q$, no two vertices in $\Gamma(C)$ are joined by more than one edge.*

Proof: Suppose that $\Gamma(C)$ contains a triangle $\{a, b, c\}$ whose edges are labelled with three different labels. Then for any $i \in \{1, 2, 3\}$, two (or more) of a_i, b_i and c_i are equal: define x_i to be this repeated value. Let $x = (x_1, x_2, x_3)$. Then x is a descendant of $\{a, b\}$, $\{b, c\}$ and $\{a, c\}$, and these three sets have trivial intersection. This contradicts the fact that C is a 2-IPP code, and so we have proved Part (i) of the lemma.

Now suppose that $\Gamma(C)$ contains a chain a, b, c, d where a, b, c, d are pairwise distinct and where ab, bc and cd are labelled 1, 2 and 3 respectively. Then it is easy to check that $\{a, c\}$ and $\{b, d\}$ are both parent sets of the word (a_1, b_2, c_3). Since these parent sets are disjoint, this contradicts the fact that C is a 2-IPP code, and so we have proved Part (ii) of the lemma.

Suppose that $|C| > q$, and let $a, b \in C$ be vertices joined by two edges. Suppose, without loss of generality, that these edges are labelled 1 and 2, so the codewords a and b agree in their first two positions. Since $|C| > q$, there exist distinct codewords $c, d \in C$ that agree in their 3rd position. By exchanging c and d if necessary, we may assume that $a \neq d$ and $b \neq c$. But then $\{a, c\}$ and $\{b, d\}$ are disjoint parent sets of (a_1, a_2, c_3), contradicting the fact that C is a 2-IPP code. So the lemma is proved. $\Box$

Proof of Theorem 1: Let C be a q-ary 2-IPP code of length 3, and suppose that $|C| \geq 3q - 1$. We derive a contradiction as follows.

Recalling the definitions of $\Gamma(D)$ and $\Gamma_i(D)$ given above, we define subsets D_0, D_1, D_2 and D_3 of C as follows. Let $D_0 = C$, and for $i \in \{1, 2, 3\}$ let

$$D_i = \{c \in D_{i-1} : c \text{ is not an isolated vertex in } \Gamma_i(D_{i-1})\}.$$

Since $|D_0| \geq 3q - 1 > q$, the graph $\Gamma_1(D_0)$ has at most $q - 1$ isolated vertices, and so $|D_1| \geq (3q - 1) - (q - 1) = 2q$. Similarly, $|D_2| \geq q + 1$ and $|D_3| \geq 2$. In particular, D_3 is not empty.

Let $d \in D_3$. By definition of D_3, there exists a vertex $c \in D_2$ such that there is an edge cd labelled 3. By definition of D_2, there is an edge bc labelled 2 for some vertex $b \in D_1$. Note that $b \neq d$, for otherwise there would be edges labelled 2 and 3 between c and d in $\Gamma(C)$, contradicting Lemma 1 (iii). Finally, by definition of D_1, there exists $a \in D_0$ such that there is an edge ab labelled 1. We find that $a \neq c$, for $a = c$ contradicts Lemma 1 (iii) as before. But $a \neq d$, as $a = d$ contradicts Lemma 1 (i). So a, b, c, d is a chain in $\Gamma(C)$ whose edges ab, bc and cd are labelled 1, 2 and 3 respectively and a, b, c and d are distinct. This contradicts Lemma 1 (ii). Hence Theorem 1 is established. $\Box$

3 An upper bound on the Size of k-IPP Codes

This section aims to prove the main result of the paper: an upper bound on the number of codewords of a q-ary k-IPP code of length n. The bulk of the section is concerned with showing that q-ary k-IPP codes of short length can have at most $O(q)$ codewords. To be more precise, we will prove the following theorem.

Theorem 2 *Let C be a q-ary k-IPP code of length n. Let $u = \lfloor (k/2 + 1)^2 \rfloor$. Then whenever $n < u$, we have that $|C| \leq \frac{1}{2}u(u - 1)(q - 1) + 1$.*

Our arguments generalise the techniques we used in Section 2 for the case of 2-IPP codes. We will then use the techniques in the paper of Hollmann *et al* [3] to derive a bound for k-IPP codes of arbitrary length:

Theorem 3 *Let C be a q-ary k-IPP code of length n. Let $u = \lfloor (k/2+1)^2 \rfloor$. Then*

$$|C| \leq \frac{1}{2}u(u-1)q^{\lceil n/(u-1) \rceil}.$$

Recall the notion of the graphs $\Gamma(C)$ and $\Gamma_i(C)$ from the previous section.

Lemma 2 *Let u be a positive integer. Let C be a q-ary code of length n and suppose that $|C| \geq \frac{1}{2}u(u-1)(q-1)+2$. Let T be a tree on u vertices, whose edges are labelled with elements of the set $\{1, 2, \ldots, n\}$. Then $\Gamma(C)$ contains a subgraph isomorphic to T (as a labelled graph).*

Proof: We use induction on u. When $u = 1$, the tree T is a single point, and since we are assuming that $|C| \geq 2$ the lemma is trivially true in this case.

Assume, as an inductive hypothesis, that $u > 1$ and the lemma is true for all smaller values of u. Let a be a vertex of T of degree 1; so there is a unique vertex $b \in T$ and a unique integer $i \in \{1, 2, \ldots, n\}$ such that there is an edge ab in T labelled i. Define

$$D = \{c \in C : c \text{ has degree at least } u-1 \text{ in } \Gamma_i(C)\}.$$

Now, $\Gamma_i(C)$ is the union of disjoint cliques, and so $c \in C \setminus D$ if and only if c is contained in a clique of size at most $u-1$. But $\Gamma_i(C)$ consists of at most q cliques, and one of these must contain more than $u-1$ vertices since

$$|C| \geq \frac{1}{2}u(u-1)(q-1)+2 > (u-1)q.$$

So $\Gamma_i(C)$ contains at most $q-1$ cliques of size at most $u-1$. Hence $|D| \geq |C| - (u-1)(q-1) \geq \frac{1}{2}(u-1)(u-2)(q-1)+2$.

Let $T' = T \setminus \{a\}$. By our inductive hypothesis (applied to the code D and the tree T') we find that $\Gamma(D)$ contains a subgraph L' that is isomorphic to T'. Let d be the vertex corresponding to b in L'. Since $d \in D$, there exist at least $u-1$ vertices in C that are connected to d via an edge with label i. Since are $u-2$ vertices in L' besides d, we find that d is connected to

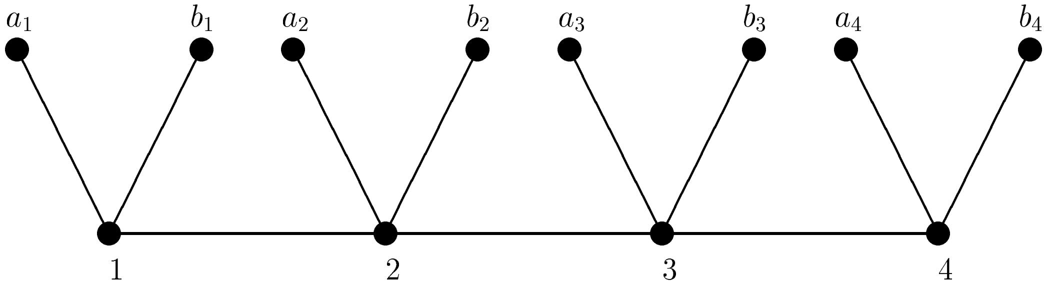

Figure 1: Constructing T when $k = 5$

a vertex e outside L' by an edge labelled with i. Adding the vertex e and the edge de to L', we obtain a subgraph of $\Gamma(C)$ that is isomorphic to T, as required. So the lemma follows by induction on u. $\Box$

Proof of Theorem 2: Suppose that C is a q-ary code of length n, where $n < u$ and $|C| \geq \frac{1}{2}u(u-1)(q-1)+2$. We will show that C is not a k-IPP code.

We define a labelled tree T as follows. Let $r = \lceil k/2 \rceil$ and $s = \lfloor k/2 \rfloor$. Let R be a set of size $r+1$, and let $\{S_x : x \in R\}$ be a collection of sets of size s. We define the vertex set of the tree T to be the disjoint union $R \cup (\cup_{x\in R} S_x)$. Note that T has u vertices, since

$$(r+1)+(r+1)s = (r+1)(s+1) = \lfloor (k/2+1)^2 \rfloor = u.$$

We add edges to R so that R becomes a tree, and then extend the tree to the whole vertex set by joining each $x \in R$ to every vertex in S_x. Figure 1 gives an example of this construction in the case when $k = 5$: here $R = \{1, 2, 3, 4\}$ and $S_x = \{a_x, b_x\}$ for all $x \in R$. Finally, we label the edges of T in an arbitrary manner subject to the condition that every label in the set $\{1, 2, \ldots, n\}$ occurs at least once. Note that we can do this since T has $u-1$ edges and $n < u$.

By Lemma 2, there exists a subgraph L of $\Gamma(C)$ that is isomorphic to T. We identify T and L, so the vertices of T are codewords, and $x, y \in T$ agree in their ith position if they are joined by an edge of T labelled i.

For $x \in R$, define the set $P_x \subseteq C$ by

$$P_x = (R \setminus \{x\}) \cup S_x.$$

Note that $|P_x| = r + s = k$, and P_x contains at least one end point of every edge in T. Moreover, $\cap_{x\in R} P_x = \emptyset$. Define a word w as follows. For each

$i \in \{1, 2, \ldots, n\}$, choose an edge yz of T labelled i and define w_i to be the common value of y_i and z_i. We remark that for any $x \in R$ the set P_x contains at least one of y and z (since yz is an edge) and so w_i agrees with the ith component of some element of P_x. Set $w = (w_1, w_2, \ldots, w_n)$. By our previous remark and the fact that $|P_x| \leq k$, the set P_x is a parent set of w for any $x \in R$. But $\cap_{x \in R} P_x = \emptyset$ and so C is not a k-IPP code, as required. □

Proof of Theorem 3: Let C be a k-IPP code C of length n over an alphabet Q of size q. Let $r = \lceil n/(u-1) \rceil$. It is easy to check that, by regarding r-tuples of elements from Q as elements from a larger alphabet Q^r, the code C may be thought of as a q^r-ary k-IPP code of length $u - 1$. But now Theorem 2 implies that

$$|C| \leq \frac{1}{2}u(u-1)(q^r - 1) + 1 \leq \frac{1}{2}u(u-1)q^{\lceil n/(u-1) \rceil},$$

and so Theorem 3 is proved. □

4 Discussion

This section discusses how far the bounds proved in Section 3 are tight, by relating them to the known existence results for k-IPP codes. Throughout this discussion, the value u will always defined by $u = \lceil (k/2+1)^2 \rceil$.

Recall that the rate of a q-ary code C of length n is defined to be $\frac{1}{n}\log_q |C|$. Theorem 3 implies that the rate of a q-ary k-IPP code of length n can be at most about $\frac{1}{n}\lceil n/(u-1) \rceil$. (Indeed, as q tends to infinity with k and n fixed, we find that an upper bound on the rate tends to $\frac{1}{n}\lceil n/(u-1) \rceil$.)

Barg, Cohen, Encheva, Kabatiansky and Zémor [2] establish probabilistic existence results for k-IPP codes. They are most interested in the case when n tends to infinity with k and q fixed, but their methods also show the following. Let k and n be fixed. Let ϵ be chosen so that $\epsilon > 0$. Then for all sufficiently large integers q there exists a q-ary k-IPP code C of length n such that $|C| \geq q^{(1/(u-1)-\epsilon)n}$. (See Yemane [5] for an alternative approach that also gives this result.)

In particular, the upper and lower bounds for the optimal rate of a k-IPP code match (at $1/(u-1)$) as $q \to \infty$ in the case when n is a multiple of $u-1$.

We conjecture that the upper bound for the rate is the correct one:

Conjecture 1 *The optimal rate for a q-ary k-IPP code of length n tends to $\frac{1}{n}\lceil n/(u-1)\rceil$ as $q \to \infty$ with k and n fixed.*

The conjecture is true when $n < u$ or when n is a multiple of $u - 1$. A result of Alon, Fischer and Szegedy [1] implies that the conjecture holds when $k = 2$ and $n = 4$. It would be very interesting to know whether their construction could work more generally. However, progress in additive number theory might be needed in order to extend their construction to other parameters.

References

[1] N. Alon, E. Fischer and M. Szegedy, 'Parent-identifying codes', *J. Comb. Theory, Series A* **95** (2001), 349-359.

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

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

[4] J.N. Staddon, D.R. Stinson and R. Wei, 'Combinatorial properties of frameproof and traceability codes', *IEEE Trans. Information Theory*, **47** (2001), 1042-1049.

[5] Y. Yemane, *Codes with the k-identifiable parent property*, PhD Thesis, Royal Holloway, University of London, 2002.