New bounds on parent-identifying codes: The case of multiple parents

Noga Alon * Uri Stav †

Abstract

Let C be a code of length n over an alphabet of q letters. A codeword y is called a descendant of a set of t codewords $\{x^1, \ldots, x^t\}$ if $y_i \in \{x_i^1, \ldots, x_i^t\}$ for all $i = 1, \ldots, n$. A code is said to have the Identifiable Parent Property of order t if for any word of length n that is a descendant of at most t codewords (parents), it is possible to identify at least one of them. Let $f_t(n, q)$ be the maximum possible cardinality of such a code. We prove that for any t, n, q, $(c_1(t)q)^{\frac{n}{s(t)}} < f_t(n, q) < c_2(t)q^{\lceil \frac{n}{s(t)} \rceil}$ where $s(t) = \lfloor (\frac{t}{2} + 1)^2 \rfloor - 1$ and $c_1(t), c_2(t)$ are some functions of t. We also show some bounds and constructions for $f_3(5, q)$, $f_3(6, q)$, and $f_t(n, q)$ when $n < s(t)$.

1 Introduction

Let Q be an alphabet, $|Q| = q$ and suppose $C \subseteq Q^n$. C is called a code, and the elements of C are called codewords.

Let P be a set of t codewords $P = \{p^1, \ldots, p^t\} \subseteq C$. We define the set of its descendants, $D(P)$, by:

$$D(P) = \{y \in Q^n | y_i \in \{p_i^1, \ldots, p_i^t\}, i = 1, \ldots, n\}$$

A code is said to have the identifiable parent property of order t (or said to have t-IPP for short) if for any $s \in Q^n$ (a son), either it is not a descendant of any set of t codewords, or there exists a codeword p (a parent) that can be identified from s, that is:

$$\forall P \subseteq C, |P| \leq t : (s \in D(P) \Rightarrow p \in P).$$

Note that identifying more than one parent is impossible, since any codeword is a descendant of itself and any other $(t-1)$ codewords.
Define:

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

*Schools of Mathematics and Computer Science, Raymond and Beverly Sackler Faculty of Exact Sciences, Tel Aviv University, Tel Aviv 69978, Israel. Email: noga@math.tau.ac.il. Research supported in part by a USA-Israeli BSF grant, by the Israel Science Foundation and by the Hermann Minkowski Minerva Center for Geometry at Tel Aviv University.

†School of Computer Science, Raymond and Beverly Sackler Faculty of Exact Sciences, Tel Aviv University, Tel Aviv 69978, Israel. Email: suri@tau.ac.il.

The study of $f_t(n,q)$ is motivated by questions about schemes that protect against piracy of software, see, e.g. [5], [7].

The following definition will be helpful later: For a code $C \subseteq Q^n$, we say that a codeword $y \in C$ is **unique** in coordinate i, if

$$\forall x \in C, x \neq y : x_i \neq y_i$$

2 Bounding the growth of $f_t(n,q)$

Our main interest is to explore the growth of $f_t(n,q)$ for some values of n and t, as a function of q. The main result in this section is that $f_t(n,q)$ grows polynomially with q, and the degree depends on n and a function of t, as follows. Denote:

$$s(t) \stackrel{def}{=} \begin{cases} \frac{t^2}{4} + t & \text{when } t \text{ is even} \\ \frac{t^2}{4} + t - \frac{1}{4} & \text{when } t \text{ is odd} \end{cases}$$

The results of this section are summarized in the following theorem:

Theorem 2.1 *There exist two functions $c_1(t)$ and $c_2(t)$, such that for every n,q:*

$$(c_1(t)q)^{\frac{n}{s(t)}} < f_t(n,q) < c_2(t)q^{\lceil \frac{n}{s(t)} \rceil}$$

The lower bound is shown using a probabilistic construction of such codes following the method of [3], while the upper bound is achieved by combining ideas from [6] with some new techniques. The upper bound, with a somewhat worse value of $c_2(t)$, has been obtained, independently, by Blackburn [4].

2.1 The upper bound

Lemma 2.2 *For every t,q,n,a: $f_t(n \cdot a, q) \leq f_t(n, q^a)$*

Proof. Suppose $C \subseteq Q^{n \cdot a}$ has t-IPP. We split the codewords in C into n blocks of a coordinates each. We view the codewords as words of length n over an alphabet of size q^a. It is easy to see that this code has t-IPP, proving the lemma. □

Lemma 2.3 *For every t:*

$$f_t(s(t), q) \leq s(t) \cdot q$$

Proof. Let $C \subseteq Q_1 \times \ldots \times Q_{s(t)}$ have t-IPP, and suppose $|C| \geq s(t) \cdot q + 1$. We show there is some descendant whose parents cannot be identified.

First, we construct a code $\hat{C} \subset C$ as follows: Whenever there is still a codeword containing a unique coordinate (unique among the words that have not been deleted so far), delete it. Any symbol among the q and any coordinate $i = 1, \ldots, s(t)$ can be responsible for deleting at most one codeword: If some symbol was unique in some place, after deleting that word it will never

appear there again. Hence we delete at most $s(t)q$ codewords, and we are thus left with at least one codeword in $\hat{C}$ (this means that in fact we deleted at most $s(t)(q-1)$ codewords, but since we are intersted in fixed t and large q, we ignore the low order terms). Note that by construction, in $\hat{C}$, no codeword will have a unique symbol (and we will thus have, in fact, several codewords in $\hat{C}$).

Suppose t is even, hence $s(t) = \frac{t^2}{4} + t = (\frac{t}{2}+1)^2 - 1$.

We choose the set of parents $X \subset \hat{C}$ as follows: Start by picking some codeword, $x^1 \in \hat{C}$. Next, we pick a codeword $x^2 \in \hat{C}$ whose $(\frac{t}{2}+1)$'th coordinate equals to that coordinate in x^1. The construction of $\hat{C}$ assures us such a codeword exists in $\hat{C}$. Also, denote $m_1 = (\frac{t}{2}+1)$.

To choose x^3, we consider the symbol in place $2(\frac{t}{2}+1) = t+2$ of x^2. This symbol appears in some other codeword in $\hat{C}$. If that other codeword is x^1, we move to the symbol in place $t+3$ and check it. We do so until we find the first symbol that appears in some codeword that is distinct from x^1, x^2. Call this coordinate m_2, and that other codeword x^3.

Later on, the $(k+1)$'th codeword is chosen as follows: let m_k be the first $m_k \geq m_{k-1} + (\frac{t}{2}+1)$ such that x^k 's symbol in the m_k'th coordinate equals to that coordinate in some codeword y which we have not picked yet, i.e. $y_{m_k} = x^k_{m_k}$ and $y \notin \{x^1, x^2, \ldots, x^k\}$. Denote this codeword $x^{k+1} = y$. If no such m_k exists, say that m_k is undefined.

We stop when the next m_k is undefined. Note that at most $(\frac{t}{2}+1)$ codewords were chosen this way. At the end of this process we have a set of codewords $X = \{x^i\}$ and indices $\{m_i\}$, such that the m_k'th coordinate of x^k equals to that coordinate in x^{k+1} for $k = 1, ..., |X| - 1$.

The descendant s takes its first $m_1 = (\frac{t}{2}+1)$ coordinates from x^1, the following coordinates until m_2 from x^2, and so on. The last parent contributes at most $\frac{t}{2}$ coordinates that do not belong to the other members of X.

Obviously, $s \in D(X)$. Yet, $|X| \leq (\frac{t}{2}+1)$ and any $x^i \in X$ can be replaced by at most $\frac{t}{2}$ codewords that give the (at most) $\frac{t}{2}$ coordinates it contributed to s, and no other $x^j \in X$ did. This gives a set of parents of size $\leq t$, that does not include x^i.

Thus, none of the elements in X is a parent, and certainly no other codeword in $\hat{C}$ is. Therefore, no parent of s can be identified and the code does not have t-IPP.

If t is odd, we do exactly the same, only taking $m_{k+1} \geq m_k + (\frac{t+1}{2})$, which gives $|X| \leq (\frac{t+3}{2})$, and all the sets of parents we use are, again, of size at most t.

□

Combining Lemma 2.2 and Lemma 2.3 we obtain the upper bound of Theorem 2.1.

2.2 The lower bound

We use the techniques of [3] to establish the lower bound. Recall the following definition:

Definition 2.4 *A code $C \subset Q^n$ is* (t,u)-partially hashing *if for any two subsets T, U of C such that $T \subset U \subset C$, $|T| = t$, $|U| = u$, there is some coordinate $i \in [1 \ldots n]$ such that*

$$\forall x \in T, y \in U, y \neq x : x_i \neq y_i$$

To prove the lower bound we need the following result:

Lemma 2.5 ([3], Corollary 3) *Let $u = s(t) + 1$. If a code C is (t, u)-partially hashing, then C has t-IPP.*

We prove the lower bound of Theorem 2.1 using the probabilistic method. We show there is a large enough code that is (t, u)-partially hashing, which therefore has t-IPP.

Pick at random, with uniform distribution, a set $C \subset Q^n$ of $|C| = M$ codewords. For any pair of sets $T \subset U \subset C$ that violate the partially hashing property, we delete some codeword in U. This leaves a code that is (t, u)-partially hashing. We choose M such that the expected number of such "bad" couples T, U is at most $\frac{M}{2}$. This assures existence of a t-IPP code of size at least $M - \frac{M}{2} = \frac{M}{2}$.

We now find the probability PR_{bad} of some fixed two sets $|T| = t, |U| = u, T \subset U$ to violate the partially hashing property:

$$\begin{aligned} PR_{bad} &= \left(1 - \frac{q(q-1)\cdots(q-t+1)(q-t)^{u-t}}{q^u}\right)^n \\ &= \left(\frac{\sum_{i=0}^{u-1} a_i(t) q^i}{q^u}\right)^n \\ &\leq \left(\frac{b(t)}{q}\right)^n \end{aligned}$$

for some functions $a_i(t), b(t)$.

The expectation of the number of bad couples therefore satisfies:

$$\begin{aligned} E_{bad} &= \left(\begin{array}{c} M \\ u \end{array}\right)\left(\begin{array}{c} u \\ t \end{array}\right) PR_{bad} \\ &\leq c(t) M^u \left(\frac{b(t)}{q}\right)^n \end{aligned}$$

Choosing $M \leq (\frac{q}{d(t)})^{\frac{n}{u-1}}$ for an appropriate d(t) assures $E_{bad} < \frac{M}{2}$.

This construction gives t-IPP codes of size at least $(c_1(t)q)^{\frac{n}{s(t)}}$, and completes the proof of Theorem 2.1.

Remark 2.6 *The correct power of q for lengths $n \not\equiv 0 \pmod{s(t)}$ remains open. The case $t = 2$, studied in [1] , and the result on $f_3(6, q)$ which we prove in the next section, show there is still much to learn on these cases.*

3 The case t=3

Note that for $t = 2$, $s(2) = 3$. In [6] it is proved that $f_2(3, q) \geq (3 - o(1))q$ (Example 4). It is also shown there that this code is essentially optimal since $f_2(3, q) \leq 3q - 1$ (a similar bound follows from Lemma 2.3). We prove an analogous result for $t = 3$: In this case, $s(3) = 5$ and hence by Lemma 2.3 $f_3(5, q) \leq 5q$.

Theorem 3.1 $f_3(5,q) \geq (5-o(1))q$ *(and hence* $f_3(5,q) = (5-o(1))q$*.)*

Proof. Split the alphabet Q into 5 pairwise disjoint sets as follows: $Q_1, \ldots, Q_4$ consist of $c \cdot q^{\frac{3}{4}}$ letters each for some constant c to be determined later. Q_5 consists of the rest of the alphabet.

Construct the code C as follows: C consists of five sets of codewords, i.e.- $C = \bigcup_{i=1}^{5} C_i$. Each set $C_i \subset Q_i \times Q_{i+1} \times \ldots \times Q_5 \times Q_1 \times \ldots Q_{i-1}$.

First, using the results shown in section 2.2, we construct a code $\hat{C} \subset Q_1 \times Q_2 \times Q_3 \times Q_4$ with the following properties:

(i) $\hat{C}$ has 2-IPP

(ii) No two codewords of $\hat{C}$ share more than one coordinate

(iii) $|\hat{C}| = |Q_5|$

This consruction is done first by picking the code $\hat{C}$ at random. Property (i) is acheived by removing one codeword from any set of codewords that violate the $(2,4)$- partially hashing property of $\hat{C}$. Then, we remove one codeword from any pair of codewords that share more than one coordinate, in order to achieve property (ii). A calculation similar to the one shown in the proof of Theorem 2.1 shows that a sufficiently large q, and a proper choice of c and the initial size of $\hat{C}$ ensure the existence of such a code. The existence of $\hat{C}$ also follows (with room to spare) from the results in [1].

To each of the symbols in Q_5 we match one codeword from $\hat{C}$. Each such couple forms a single codeword in C_1, and with an appropriate cyclic shift, forms a codeword in each of the other C_is. Hence, each C_i consists of $|Q_5|$ codewords. The size of the code C is therefore $5|Q_5| = 5(q - 4cq^{\frac{3}{4}}) = (5 - 4cq^{-\frac{1}{4}})q = (5 - o(1))q$. We prove that C has 3-IPP by showing how to find a parent of a given descendant $s \in Q^5$. If s is not a descendant of any 3 codewords, the search for a parent will fail. Note that in any coordinate, the 5 sets of symbols used by each C_i, are pairwise disjoint. Thus, for each one of s's coordinates we can determine from which C_i it came.

We handle the following cases separately:

Case 1 - All symbols in s are from a single C_i : One of s's coordinates is taken from Q_5. In this case, there is only one codeword with this symbol at that coordinate, and this codeword has to be in any set of parents.

Case 2 - Symbols from 2 different C_i's appear in s: First, suppose s contains 4 coordinates of a single C_i and one coordinate of some other C_j. Then the 4 coordinates are taken from at most 2 codewords in C_i, since we still need another parent from C_j. If one of the 4 coordinates is a symbol from Q_5, then we can identify the parent immediately. Otherwise, we have the 4 coordinates that form a 2-IPP code, hence again we can identify a parent.

Now, assume s contains 3 coordinates from C_i, and the other 2 coordinates are taken from C_j $(i \neq j)$. If a codeword $x \in C_i$ equals s in 3 coordinates, it must be among its parents: Otherwise, since no 2 distinct codewords in C_i share more than 1 coordinate, s should have

at least 3 parents from C_i, but this would leave no room for the essential parent from C_j. If no such x exists, than two of s's parents are members of C_i, and there is only one parent from C_j. Yet s has 2 of that parent's coordinates, and so there is only (at most) one possible parent in C_j.

Case 3 - Symbols from 3 different C_i's appear in s: In this case, the set of parents must consist of exactly one member from each of the 3 C_i sets. Yet, at least one of the parents contributed 2 coordinates or more to s, which allows us to identify it.

Case 4 - s contains symbols from more than 3 C_i's: In this case, s is surely not a descendant of any 3 codewords of C.

This covers all possibilities, hence C indeed has 3-IPP. □

By Theorem 2.1 we know that $\Omega(q^{\frac{6}{5}}) \leq f_3(6,q) \leq O(q^2)$. Using the techniques of [1] modified appropriately, we can prove the following theorem:

Theorem 3.2 $f_3(6,q) = o(q^2)$

To prove this upper bound, we need the following result, proved in [2] by applying the regularity lemma of Szemerédi [8].

Lemma 3.3 ([2], Proposition 4.4) *For every $\gamma > 0$ and every integer k there exists a $\delta = \delta(k,\gamma) > 0$ such that every simple graph G on n vertices containing less than δn^k copies of the complete graph K_k on k vertices, contains a set of less than γn^2 edges whose deletion destroys all copies of K_k in G.*

Suppose that for some $\epsilon > 0$ and every q, $f_3(6,q) \geq \epsilon q^2$. Let C be such a code, for a large enough q. We will show that C doesn't have 3-IPP.

Lemma 3.4 *No two codewords in C share two coordinates.*

Proof. By the result of [1], we have $f_2(q,4) = o(q^2)$. Thus for a large enough q, $f_2(4,q) < (\epsilon q^2 - 2)$. Therefore, considering the code induced by C without any two codewords on any 4 coordinates gives a code that does not have 2-IPP.

Suppose we have some 2 codewords $x, y \in C$ such that x equals y in 2 coordinates i_1, i_2. Inducing the code $C \setminus \{x,y\}$ on the other 4 coordinates, gives a non-2-IPP code. Suppose z is some descendant with 2 or 3 possible sets of parents $\{P_j\}_j$ in that code, that violate the 2-IPP. That is, for every j: $z \in D(P_j)$, and $\cap_j P_j = \emptyset$.

Consider the following codeword: We take the symbols in i_1, i_2 from x (and y), and the other 4 coordinates from z. Call this codeword w. For every j, w is a descendant of both $P_j \cup \{x\}$ and $P_j \cup \{y\}$. All these sets are of size ≤ 3, and surely have an empty intersection. Therefore, a parent of w cannot be identified, contradicting the fact that C has 3-IPP.

□

Proof of Theorem 3.2 Construct a 6-partite graph $G = (V, E)$ as follows: Each vertex class consists of q vertices, i.e. $V = Q_1 \cup Q_2 \cup ... \cup Q_6$, $|Q_i| = q$. We relate Q_i to coordinate i. For every codeword $q_1q_2q_3q_4q_5q_6 \in C$ we add a copy of K_6 on the vertices $q_1 \in Q_1, ..., q_6 \in Q_6$.

By Lemma 3.4, G is simple, and hence contains $15\epsilon q^2$ edges. As it is the edge-disjoint union of ϵq^2 copies of K_6, one has to delete at least ϵq^2 of its edges to destroy all copies of K_6 contained in G. By Lemma 3.3, the graph G contains at least δq^6 copies of K_6, for a constant $\delta = \delta(\epsilon) > 0$.

Among these copies of K_6, the number of K_6 copies that contain at least two edges arising from the same $x \in C$ is at most $O(q^5)$: There are at most q^2 ways to choose x, and $\frac{15 \cdot 14}{2}$ ways to choose two of its edges. This determines already at least three vertices of the K_6, leaving at most q^3 options for the remaining vertices.

It follows that G contains a copy of K_6 in which every edge comes from a different $x \in C$. Suppose $q_1, \ldots, q_6$ are the vertices of this K_6. Then, the codeword $x = q_1 \ldots q_6$ is a descendant of both

(i) The three codewords giving the edges (q_1, q_2), (q_3, q_4), (q_5, q_6)

(ii) The three codewords giving the edges (q_2, q_3), (q_4, q_5), (q_6, q_1)

Since all these codewords are different, a parent of x cannot be identified, hence C does not have 3-IPP. □

4 The case $n \leq s(t)$

In this section we explore some values of $f_t(n, q)$ in those cases where the size of the largest possible code is linear in q. Our main interest is in the constant multiplying q. The sublinear additive error terms will usually be disregarded.

4.1 $n \leq t$

The results on this case are summarized in the following simple lemma:

Lemma 4.1 *For any t, and $n \leq t$: $f_t(n, q) = q$*

Proof. Suppose we have a code of $q + 1$ codewords. In this case, in each coordinate, there is some symbol that appears in two different codewords. Take such a symbol in each coordinate, to generate $s \in Q^n$. Clearly s is a descendant of at most $n \leq t$ codewords; simply take, for each coordinate i, a codeword whose i-th coordinate is s_i. It is also not difficult to see that no parent of s can be identified. This is because we can choose the above set of at most n codewords that generate s even if we have to avoid any single codeword (simply because we have at least two choices in each coordinate). Hence, this code does not have t-IPP. This shows that $f_t(n, q) \leq q$.

Furthermore, the repetition code of any length over a set of q symbols has t-IPP. This code achieves the upper bound, completing the proof.

□

4.2 $n = t + 1$

We have an asymptotically tight result for this case too:

Theorem 4.2 *For any $t \geq 2$,*

$$f_t(t+1, q) = \left(1 + \frac{1}{t - \frac{3}{2}} - o(1)\right) q.$$

We first note the following simple fact.

Fact 4.3 *Assume $C \subseteq Q^{t+1}$ has t-IPP, and $|C| > q+2$. Then there are no two distinct codewords, $x, y \in C$, such that x is not unique in some two coordinates i_1, i_2, and y is not unique in two coordinates i_3, i_4, where all four coordinates i_1, i_2, i_3, i_4 are distinct.*

Proof. Suppose such codewords $x, y \in C$ exist. By taking x's symbols in i_1, i_2 and y's symbols in i_3, i_4, and any symbol that appears in at least two codewords besides x, y in each of the rest of the coordinates, we obtain a descendant whose parents cannot be identified (contradicting our assumption on C). □

Proof of Theorem 4.2 We prove the upper bound first: Assume we have some code $C \subseteq Q^{t+1}$, $|C| > (1 + \frac{1}{t-\frac{3}{2}})q$, that has t-IPP.

Suppose C does not contain any codeword that is not unique in two coordinates. In this case, every codeword of C is unique in at least t coordinates, and since at most $q - 1$ codewords can be unique at each of the $t + 1$ coordinates we get

$$t|C| \leq (q-1)(t+1).$$

Therefore

$$|C| < (1 + \frac{1}{t})q < (1 + \frac{1}{t - \frac{3}{2}})q$$

which contradicts our assumption on the size of C.

Hence C must contain a codeword that is not unique in at least 2 coordinates. We may assume, without loss of generality, that $x \in C$ is not unique in coordinates $\{1, 2\}$. By Fact 4.3, there is no codeword $y \in C$ that is not unique in 2 coordinates among $\{3, \ldots, t+1\}$.

First, assume that there is some codeword $y \in C \setminus \{x\}$ that is not unique in coordinates $\{1, 3\}$ and that no other codeword of C is not unique in coordinates $\{2, 3\}$. In this case, if a codeword of C is not unique in some two coordinates, then at least one of them has to be coordinate 1 (otherwise we would get two codewords that are not unique in two disjoint pairs of coordinates, contradicting

Fact 4.3). Thus any codeword in C (including x and y) has at most one unique coordinate among $\{2,\ldots,t+1\}$. Hence, for the last t coordinates we get

$$(t-1)|C| \leq (q-1)t$$

that is

$$|C| < (1+\frac{1}{t-1})q < (1+\frac{1}{t-\frac{3}{2}})q$$

and again a contradiction is obtained.

We are left with the last option in which there are both $y \in C$ that is not unique in coordinates $\{1,3\}$, and $z \in C$ which is not unique in coordinates $\{2,3\}$. If a codeword $w \in C$ is not unique in coordinate i, for some $i \in \{4,\ldots,t+1\}$, then this must be the only coordinate in which w is not unique (otherwise, a contradiction to Fact 4.3 is obtained). We split the code C into 2 disjoint sets: C_1 contains all the codewords that are not unique in exactly one of the coordinates $\{4,\ldots,t+1\}$. C_2 contains all the codewords of C that are unique in all the coordinates $\{4,\ldots,t+1\}$ (but may not be unique in some of the coordinates $\{1,2,3\}$). We also split the code C_1 into $t-2$ pairwise disjoint sets as follows: for $i \in \{4,\ldots,t+1\}$, C_1^i consists of the codewords that are not unique in coordinate i.

We first claim that the induced code of C_2 on coordinates $\{1,2,3\}$ (which we denote by $C_2|_{\{1,2,3\}}$) must have 2-IPP. To prove this claim, suppose we have some descendant $s \in Q^3$ whose parent cannot be identified, i.e. there are sets of parents $\{P_j\}$ such that $|P_j| \leq 2$, $P_j \subset C_2|_{\{1,2,3\}}$, $s \in D(P_j)$ and $\bigcap_j P_j = \emptyset$. In this case, C itself does not have t-IPP: we construct a descendant by taking the first 3 coordinates from s, and the i'th coordinate from some representative of C_1^i for $i \in \{4,\ldots,t+1\}$ (C_1^i cannot be empty: Otherwise, all the codewords of C would be unique in coordinate i, hence the cardinality of C would not exceed q). This codeword is a descendant of any of the following sets of parents: Taking some P_j and adding one of at least two possible codewords from each C_1^i. It is easy to see that the intersection of these sets is empty, proving the claim.

In the first three coordinates, the codewords of C_2 use different symbols then the ones used by C_1, since the codewords of C_1 are always unique in coordinates $\{1,2,3\}$. Thus, the size of the alphabet left for C_2 in each of these coordinates is $q-|C_1|$. Yet, C_2 has 2-IPP in these coordinates, and since $f_2(3,\hat{q}) \leq 3\hat{q}$ (by Lemma 2.3), we get:

$$|C_2| \leq 3(q-|C_1|) \tag{1}$$

Moreover, for the i'th coordinate ($i \in \{4,\ldots,t+1\}$), there are $|C_1|+|C_2|-|C_1^i|$ codewords that are unique in coordinate i. Therefore, for $i \in \{4,\ldots,t+1\}$, we have:

$$|C_1|+|C_2|-|C_1^i| \leq q \tag{2}$$

Adding the inequalities (2) for $i=4,\ldots,t+1$, and dividing by $t-2$ we obtain:

$$(1-\frac{1}{t-2})|C_1|+|C_2| \leq q \tag{3}$$

Inequality (1) can also be written as

$$3|C_1| + |C_2| \leq 3q \tag{4}$$

Multiplying it by $\frac{1}{2(t-2)}$, and adding it to inequality (3), we get

$$\left(1 + \frac{1}{2(t-2)}\right)|C_1| + \left(1 + \frac{1}{2(t-2)}\right)|C_2| \leq \left(1 + \frac{3}{2(t-2)}\right)q \tag{5}$$

Hence

$$|C_1| + |C_2| \leq \left(\frac{1 + \frac{3}{2(t-2)}}{1 + \frac{1}{2(t-2)}}\right)q = \left(1 + \frac{1}{t - \frac{3}{2}}\right)q$$

which again contradicts our assumption on the size of C and completes the proof of the upper bound.

We now construct a proper code. The construction of the code C simply follows the last part of the proof of the upper bound. To simplify the presentation, we omit all floor and ceiling signs in what follows. Since t is fixed and q is large, this clearly does not affect the asymptotic result.

We choose some special symbol $q_0 \in Q$. First, we construct $t-2$ sets of codewords, C_1^i, as follows: The i'th set ($i \in \{4, \ldots, t+1\}$) consists of $(\frac{1}{t-\frac{3}{2}})q$ codewords with unique symbols in all coordinates except for coordinate i, in which the symbol q_0 appears.

This way we get $C_1 = \bigcup_{i=4}^{t+1} C_1^i$. The size of C_1 is:

$$|C_1| = (t-2)\left(\frac{1}{t - \frac{3}{2}}\right)q = \left(1 - \frac{1}{2t-3}\right)q$$

We add another set of codewords, C_2: For each of the first three coordinates, we find the set of symbols that have not been used yet (by codewords in C_1). Using the construction in [6] (Example 4) we obtain a 2-IPP code of length 3. The size of the alphabet in each coordinate is

$$\hat{q} = q - |C_1| = q - (1 - \frac{1}{2t-3})q = \frac{q}{2t-3}$$

Hence, the size of this code is

$$(3 - o(1))\hat{q} = \left(\frac{3}{2t-3} - o(1)\right)q.$$

To each such codeword we add $t-2$ unique symbols to create codewords of length $t+1$. There are enough symbols, since in each coordinate (among $\{4, \ldots, t+1\}$) we have

$$q - 1 - q(t-3)(\frac{1}{t - \frac{3}{2}}) = q\left(1 - \frac{t-3}{t-\frac{3}{2}}\right) - 1 = q(\frac{3}{2t-3} - o(1))$$

unused symbols after creating C_1.

The code C is the union $C = C_1 \cup C_2$. Its size is

$$|C| = |C_1| + |C_2| = \left(1 - \frac{1}{2t-3}\right)q + \left(\frac{3}{2t-3} - o(1)\right)q = \left(1 + \frac{1}{t-\frac{3}{2}} - o(1)\right)q$$

To complete the proof, we show that C has t-IPP: If a descendant s has some symbol that is unique (i.e. appears only once) in some coordinate, we can identify the parent from which it came. Assume s does not have any unique symbol. In this case, any set $P \subset C$ such that $s \in D(P)$ and $|P| \leq t$, must contain at least one codeword from each C_1^i ($i \in \{4, \ldots, t+1\}$). Yet $|P| \leq t$, hence P contains at most 2 codewords from C_2, that give s its first 3 coordinates (if s inherits one of its first 3 symbols from a codeword in C_1, then this symbol is unique). Since $C_2|_{\{1,2,3\}}$ has 2-IPP, a parent of s from C_2 can be identified.

□

4.3 $t+1 < n < s(t)$

For these cases we do not have a general tight bound. The following lemma, however, provides good upper bounds in some cases:

Lemma 4.4 *Suppose $b + m - 1 \leq t$, then*

$$f_t(b \cdot m, q) \leq (b + o(1))q$$

Proof. We actually show that $f_t(b \cdot m, q) < b(q-1) + m$. Assume we have such a code of size at least $b(q-1) + m$. Split the coordinates into m sets (blocks) of b coordinates each. In each block, there are at most $b(q-1)$ codewords that are unique in some coordinate in that block. Therefore, there are at least m codewords whose symbols in each coordinate of that block are not unique.

This enables us to find a set X of m representatives: x^i, the representative of the i'th block, has no unique symbol in the i'th block. We choose a different representative for each block.

A descendant s that inherits its coordinates in each block from that block's representative is a descendant of $X = \{x^1, \ldots, x^m\}$. Yet every x^i can be replaced by some b codewords, and since $b + m - 1 \leq t$, we still have a set of at most t parents of s that does not contain x^i. This shows that identifying a parent for s is impossible, and the code does not have t-IPP.

□

Remark 4.5 *In the case $n = s(t)$, we only have the upper bound of Lemma 2.3. By [6] (Example 4) and by Theorem 3.1 here this bound is asymptotically optimal for $t = 2, 3$. We conjecture that constructions, similar to the one appearing in the proof of Theorem 3.1, may show this bound is tight also for all other values of t.*

References

[1] N. Alon, E. Fischer and M. Szegedy, Parent-identifying codes, *Journal of Combinatorial Theory Ser. A* 95 (2001), 349-359.

[2] N. Alon, R. A. Duke, H. Lefmann, V. Rödl and R. Yuster, The algorithmic aspects of the Regularity Lemma, *Proceedings of the* 33^{rd} *IEEE FOCS at Pittsburgh* (1992), 473–481. Also: *Journal of Algorithms* 16 (1994), 80–109.

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

[4] S. R. Blackburn, An upper bound on the size of a code with the k-identifiable parent property, Preprint, 2002.

[5] D. Boneh and J. Shaw, Collusion-secure fingerprinting for digital data, IEEE Transactions on Information Theory 44 (1998), 1897-1905.

[6] H. D. L. Hollmann, J. H. Van Lint, J-P. Linnartz and L. M. G. M. Tolhuizen, On codes with the identifiable parent property, *Journal of Combinatorial Theory Ser. A* 82 (1998), 121–133.

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

[8] E. Szemerédi, Regular partitions of graphs, In: *Proc. Colloque Inter. CNRS* No. 260 (J. C. Bermond, J. C. Fournier, M. Las Vergnas and D. Sotteau eds.), 1978, 399–401.

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