PROLIFIC CODES WITH THE IDENTIFIABLE PARENT PROPERTY[*]

SIMON R. BLACKBURN[†], TUVI ETZION[‡], AND SIAW-LYNN NG[§]

Abstract. Let $\mathcal{C}$ be a code of length n over an alphabet of size q. A word $\mathbf{d}$ is a *descendant* of a pair of codewords $\mathbf{x}, \mathbf{y} \in \mathcal{C}$ if $d_i \in \{x_i, y_i\}$ for $1 \leq i \leq n$. A code $\mathcal{C}$ is an *identifiable parent property* (IPP) code if the following property holds. Whenever we are given $\mathcal{C}$ and a descendant $\mathbf{d}$ of a pair of codewords in $\mathcal{C}$, it is possible to determine at least one of these codewords.

The paper introduces the notion of a prolific IPP code. An IPP code is *prolific* if all q^n words are descendants. It is shown that linear prolific IPP codes fall into three infinite ('trivial') families, together with a single sporadic example which is ternary of length 4. There are no known examples of prolific IPP codes which are not equivalent to a linear example: the paper shows that for most parameters there are no prolific IPP codes, leaving a relatively small number of parameters unsolved. In the process the paper obtains upper bounds on the size of a (not necessarily prolific) IPP code which are better than previously known bounds.

Key words. error-correcting codes, identifying parent property, linear codes, MDS codes, orthogonal arrays, copyright protection.

AMS subject classifications. 94B60, 94A60, 94B65

1. Introduction.

Codes with the Identifiable Parent Property (IPP codes) were first introduced by Hollmann, van Lint, Linnartz and Tolhiuzen [9] in 1998, motivated by an application to prevent software piracy. IPP codes and various generalisations have since been intensively studied: see, for example, papers of Alon, Cohen, Krivelevich and Litsyn [2], Alon, Fischer and Szegedy [3], Alon and Stav [4], Barg, Cohen, Encheva, Kabatiansky and Zémor [5], Barg and Kabatiansky [6], Blackburn [7], Lindkvist, Löfvenberg and Svanström [11], Löfvenberg [10], Staddon, Stinson and Wei [12], Tô and Safavi-Naini [13], van Trung and Martirosyan [14] and Yemane [15].

To define IPP codes we need the notion of a descendant, which is defined as follows. Let F be an alphabet of size q. Let $\mathbf{x} = x_1x_2 \dots x_n \in F^n$ and $\mathbf{y} = y_1y_2 \dots y_n \in F^n$ be q-ary words of length n. The set of *descendants* $\mathrm{desc}(\mathbf{x}, \mathbf{y})$ of $\mathbf{x}$ and $\mathbf{y}$ is defined to be

$$\mathrm{desc}(\mathbf{x}, \mathbf{y}) = \{d_1d_2 \cdots d_n \in F^n : d_i \in \{x_i, y_i\} \text{ for } i = 1, 2, \dots, n\}.$$

If $\mathbf{d} \in \mathrm{desc}(\mathbf{x}, \mathbf{y})$, we say that $\mathbf{d}$ is a *descendant* of $\mathbf{x}$ and $\mathbf{y}$, and we say that $\{\mathbf{x}, \mathbf{y}\}$ is a set of *parents* of $\mathbf{d}$. We say that the parent $\mathbf{x}$ *contributes to the ith component of* $\mathbf{d}$ if $x_i = d_i$. Clearly $|\mathrm{desc}(\mathbf{x}, \mathbf{y})| = 2^{d(\mathbf{x},\mathbf{y})}$, where $d(\mathbf{x}, \mathbf{y})$ is the Hamming distance between $\mathbf{x}$ and $\mathbf{y}$.

Let $\mathcal{C}$ be an (n, q, M)-code (so $\mathcal{C}$ is a q-ary code of length n, containing M codewords). Informally, $\mathcal{C}$ has the *Identifiable Parent Property* (we say $\mathcal{C}$ is an *IPP code*, or an (n, q, M)*-IPP code*) if whenever we are given a descendant $\mathbf{d}$ of two codewords, we are able to identify one of the parents. More formally, $\mathcal{C}$ is an IPP code if the

[*] This research was supported in part by E.P.S.R.C. grant EP/E034632/1.

[†] Department of Mathematics, Royal Holloway, University of London, Egham, Surrey TW20 0EX, United Kingdom, `s.blackburn@rhul.ac.uk`.

[‡] Technion — Israel Institute of Technology, Department of Computer Science, Technion City, Haifa 32000, Israel, `etzion@cs.technion.ac.il`.

[§] Department of Mathematics, Royal Holloway, University of London, Egham, Surrey TW20 0EX, United Kingdom, `s.ng@rhul.ac.uk`.

following holds. For $\mathbf{d} \in F^n$, define

$$P_{\mathbf{d}} = \{\{\mathbf{x}, \mathbf{y}\} \subseteq \mathcal{C} : \mathbf{d} \in \mathrm{desc}(\mathbf{x}, \mathbf{y})\}.$$

Then $\mathcal{C}$ is an IPP code if for all $\mathbf{d} \in F$ which are descendants of one or more pairs of codewords

$$\bigcap_{\{\mathbf{x},\mathbf{y}\} \in P_{\mathbf{d}}} \{\mathbf{x}, \mathbf{y}\} \neq \emptyset.$$

The following lemma, due to Hollmann *et al.* [9], gives simple criteria for a code to have the Identifiable Parent Property.

LEMMA 1.1. *An (n, q, M) code $\mathcal{C}$ is an IPP code if and only if*

IPP1 *For any three distinct codewords $\mathbf{x}, \mathbf{y}, \mathbf{z} \in \mathcal{C}$ there exists $i \in \{1, 2, \ldots, n\}$ such that x_i, y_i and z_i are distinct.*

IPP2 *For any four codewords $\mathbf{x}, \mathbf{y}, \mathbf{z}, \mathbf{v} \in \mathcal{C}$ such that $\{\mathbf{x}, \mathbf{y}\} \cap \{\mathbf{z}, \mathbf{v}\} = \emptyset$, there exists $i \in \{1, 2, \ldots, n\}$ such that $\{x_i, y_i\} \cap \{z_i, v_i\} = \emptyset$.*

Hollmann *et al.* [9] observed that the ternary Hamming code of length 4 is an example of a $(4, 3, 9)$-IPP code:

$$\mathcal{C} = \{0000, 0111, 0222, 1012, 1120, 1201, 2021, 2102, 2210\}.$$

To see why $\mathcal{C}$ is an IPP code, note that since all codewords are at distance 3 a descendant $\mathbf{d} \in \mathrm{desc}(\mathbf{x}, \mathbf{y})$ is at distance at most 1 from exactly one of its parents $\mathbf{x}, \mathbf{y}$. But the minimum distance of the code shows that $\mathbf{d}$ cannot be of distance at most 1 from two distinct codewords. Thus $\mathcal{C}$ is an IPP code, with the identified parent of a descendant $\mathbf{d}$ being the unique codeword at distance at most 1 from $\mathbf{d}$.

This example has the beautiful property that every possible word is a descendant. We say that a code is *prolific* if every word is a descendant of some pair of codewords. The main question this paper asks is: what other examples of prolific IPP codes are there? This question is motivated by an attempt to draw parallels between error correcting codes and IPP codes. There are clear connections between the two areas: at a most basic level, we observed above that the size of the set of descendants is related to Hamming distance; moreover, error correcting codes of high minimum distance provide good explicit constructions of IPP codes (see Hollmann *et al.* [9, Theorem 4], for example). From this perspective, prolific IPP codes may be thought of as analogues of perfect error correcting codes.

There are three *trivial* families of prolific IPP codes. Firstly the set F of all words of length 1 is a prolific $(1, q, q)$-IPP code. Secondly, the repetition code of length 2 (with codewords of the form aa where $a \in F$) is a prolific $(2, q, q)$-IPP code. Thirdly, any binary word and its complement form a prolific $(n, 2, 2)$-IPP code. It is easy to see that a prolific IPP code which has length 1 or 2, or which is a binary code, must be equivalent to a member of one of these three families, and so from now on we assume that $n \geq 3$ and $q \geq 3$.

All the examples above are linear codes. One main aim of the paper is to show that the $(4, 3, 9)$-code above is the only non-trivial example of a linear prolific IPP code (up to equivalence). In general, we conjecture that there are no more examples of prolific IPP codes (linear or not). We do not know how to show this, but we are able to prove that there are no more examples when $n \leq 5$, and we rule out many other parameters. As a side-benefit of our investigations, we are able to provide new upper bounds on the size of a (not necessarily prolific) IPP code.

For the rest of this paper, $\mathcal{C}$ will be a q-ary code of length n with M codewords. We write words and subwords in a bold font, to distinguish them from components of codewords. We write $\ell(\mathbf{x})$ for the length of the (sub)word $\mathbf{x}$.

The paper is structured as follows. In Section 2, we provide some simple upper and lower bounds for the size of a prolific IPP code. Section 3 shows that there are no non-trivial prolific IPP codes which are MDS codes (or even orthogonal arrays) other than the $(4, 3, 9)$-code above. Section 4 then extends this result to general linear codes. We then turn our attention away from the linear case. We provide new upper bounds for (not necessarily prolific) IPP codes in Section 5. Sections 6, 7 and 8 show that there are no non-trivial examples of prolific IPP codes for lengths 3, 4 and 5 respectively, other than the $(4, 3, 9)$-example. Finally, Section 9 summarises the parameters where it is unknown whether a prolific IPP code exists, and comments on possibilities for future work.

2. General bounds for prolific IPP codes. Since prolific IPP codes have many descendants, it seems intuitively plausible that they must be fairly large. This is indeed the case, and this section makes this precise by establishing lower bounds on the size of a prolific IPP code. For many parameters these lower bounds conflict with known upper bounds on the size of an IPP code, and so the bounds rule out the existence of a prolific IPP code for these parameters.

The simplest lower bound on the size of a prolific IPP code is stated in the following theorem.

THEOREM 2.1. *If $\mathcal{C}$ is an (n, q, M) prolific IPP code then $\binom{M}{2}2^n \geq q^n$.*

Proof. There are at most $\binom{M}{2}$ pairs of codewords from $\mathcal{C}$. Each pair of codewords can produce at most 2^n descendants, since there are at most two possibilities for each component of a descendant once the pair of parents is fixed. So $\mathcal{C}$ has at most $\binom{M}{2}2^n$ descendants. The bound follows once we observe that all q^n words must be descendants since $\mathcal{C}$ is prolific. □

The counting argument used above will tend to significantly overcount descendants which are close to a codeword (in terms of Hamming distance). We can overcome this problem by counting the descendants of the code in another way, giving us the following improvement on Theorem 2.1.

THEOREM 2.2. *Let $\mathcal{C}$ be an (n, q, M) prolific IPP code and let k be a positive integer. Then*

$$M\left(\sum_{i=0}^{k}\binom{n}{i}(q-1)^i + \frac{M-1}{2}\sum_{i=k+1}^{n-k-1}\binom{n}{i}\right) \geq q^n.$$

Proof. We count the descendants of $\mathcal{C}$ as follows. A sphere of radius k contains $\sum_{i=0}^{k}\binom{n}{i}(q-1)^i$ words, and so there are at most $M\left(\sum_{i=0}^{k}\binom{n}{i}(q-1)^i\right)$ descendants of $\mathcal{C}$ at distance at most k from the code. A descendant of a pair $\{\mathbf{c}_1, \mathbf{c}_2\} \subseteq \mathcal{C}$ of codewords is formed by choosing i components from $\mathbf{c}_1$ and the remaining $n - i$ components from $\mathbf{c}_2$ for some $i \in \{0, 1, \ldots, n\}$. But when $0 \leq i \leq k$ or $n - k \leq i \leq n$ the resulting descendant is within distance k of the code. So each of the $\binom{M}{2}$ pairs of codewords gives rise to at most $\sum_{i=k+1}^{n-k-1}\binom{n}{i}$ descendants of distance greater than k from the code. Thus $\mathcal{C}$ has at most $M\left(\sum_{i=0}^{k}\binom{n}{i}(q-1)^i + \frac{M-1}{2}\sum_{i=k+1}^{n-k-1}\binom{n}{i}\right)$ descendants, and the theorem follows by the same argument as in Theorem 2.1. □

The *covering radius* of an (n, q, M) code $\mathcal{C}$ is the smallest integer R such that any word lies in the union of the spheres of radius R about the codewords. A prolific code cannot have a large covering radius, as the following theorem shows. We make use of the key idea in the proof of Theorem 2.3 in several places in the sections that follow.

THEOREM 2.3. *The covering radius of an (n, q, M) prolific IPP code $\mathcal{C}$ is at most $\lfloor \frac{n}{2} \rfloor$.*

Proof. A descendant $\mathbf{d}$ has at least $\lceil \frac{n}{2} \rceil$ positions in common with one of its parents. Since $\mathcal{C}$ is prolific, every word is a descendant and so the covering radius of $\mathcal{C}$ is at most $\lfloor \frac{n}{2} \rfloor$. □

3. MDS codes. The aim of this section is to prove Theorem 3.1, that there are no non-trivial prolific IPP codes which are (n, q, q^k)-codes of minimum distance $n - k + 1$ other than the $(4, 3, 9)$-code from the introduction. We will use this result twice. Firstly, in the next section, we use the result in our proof that the examples in the introduction are the only linear prolific IPP codes. (Recall that an *MDS code* is a linear-(n, q, q^k) code of minimum distance $n - k + 1$: it is this case that we use in the next section.) Secondly, in Section 8, we use the result to prove that there exist no non-trivial examples of prolific IPP codes of length 5.

Throughout this section, we assume that the all-zero word is a codeword. This is no loss of generality, as we may replace a code with an equivalent code containing the all-zero word if necessary.

Note that an (n, q, q^k)-code of minimum distance $n - k + 1$ meets the Singleton bound. In particular, whenever we restrict all codewords to a set of k positions we find that every word of length k appears exactly once as a restriction. We abuse notation slightly and refer to this property as *the MDS property* of the code, even though we are not assuming our code is linear.

We use the well known result that an (n, q, q^k)-code $\mathcal{C}$ with minimum distance $n - k + 1$ can only exist when $q \geq n - k + 1$. To see this, note that we can construct an $(n - k + 2, q, q^2)$ code $\mathcal{C}'$ of minimum distance $n - k + 1$ by taking all codewords in $\mathcal{C}$ ending in $k - 2$ zeros, and then removing these zeros to produce words of length $n - k + 2$. But then $\mathcal{C}'$ implies the existence of a set of $n - k$ mutually orthogonal Latin squares of order q (see Hill [8, Theorem 10.20]), and such a set can have size at most $q - 1$ (see Hill [8, Theorem 10.18]).

THEOREM 3.1. *Let $\mathcal{C}$ be an (n, q, q^k)-code of minimum distance $n - k + 1$. Let $n \geq 3$ and $q \geq 3$. If $\mathcal{C}$ is a prolific IPP code, then $n = 4$, $q = 3$, and $k = 2$. In particular, no MDS code of length strictly greater than 4 is a prolific IPP code.*

Proof. Let $\mathcal{C}$ be an (n, q, q^k) code which has minimum distance $n - k + 1$. Assume that we are not in the case when $n = 4$, $q = 3$ and $k = 2$. We need to show that $\mathcal{C}$ is not a prolific IPP code. We deal with the length 3 and 4 cases first, and then go on to consider the remaining cases.

Suppose that $n = 3$. When $k = 0$ or $k = 1$, we see that $\mathcal{C}$ is too small to be a prolific code, by Theorem 2.1. When $k = 2$ or $k = 3$, we see that $\mathcal{C}$ is too large to be an IPP code: a bound of Hollmann *et al.* [9, Theorem 1] states that an IPP code of length 3 has at most $3q - 1$ codewords. So we get a contradiction when $n = 3$, as required.

Suppose that $n = 4$. Theorem 2.1 implies that $k \geq 2$, and Theorem 4.3 implies that $k \leq 2$. So we may assume that $k = 2$, and thus $q > 3$ and $\mathcal{C}$ is a $(4, q, q^2)$ code of minimum distance 3. The union of spheres of radius 1 about codewords contains $q^2(1 + 4(q - 1))$ words, and since $q > 3$ we have that $q^2(1 + 4(q - 1)) < q^4$. So there exists a word $\mathbf{d} = d_1 d_2 d_3 d_4$ of distance at least 2 from any codeword. By the MDS property of $\mathcal{C}$, there exist codewords $\mathbf{c}_1, \mathbf{c}_2, \mathbf{c}_3, \mathbf{c}_4 \in \mathcal{C}$ of the form $\mathbf{c}_1 = d_1 d_2 {*}{*}$, $\mathbf{c}_2 = {*}{*} d_3 d_4$, $\mathbf{c}_3 = d_1 {*}{*} d_4$ and $\mathbf{c}_4 = {*} d_2 d_3 {*}$. These codewords are distinct, since $\mathbf{d}$

has distance 2 from $\mathcal{C}$. But then the sets $\{\mathbf{c}_1, \mathbf{c}_2\}$ and $\{\mathbf{c}_3, \mathbf{c}_4\}$ violate IPP2. (Recall the definition of IPP2 from Lemma 1.1.) So we have a contradiction in this case.

It remains to consider the situation when $n > 4$. We distinguish between six cases. In the first two cases we show that $\mathcal{C}$ is not an IPP code; in the remaining cases we show that $\mathcal{C}$ cannot be prolific.

Case 1: $n \leq 3k-3$. Define integers n_1, n_2 and n_3 by $n_1 = \lceil \frac{n}{3} \rceil$, $n_2 = \lfloor \frac{n}{3} \rfloor$ and $n_3 = n - \lceil \frac{n}{3} \rceil - \lfloor \frac{n}{3} \rfloor$. We can write each codeword in the form $\mathbf{x}_1\mathbf{x}_2\mathbf{x}_3$, where $\ell(\mathbf{x}_i) = n_i$. We are assuming that the all-zero word $\mathbf{c}_0 = \mathbf{000}$ lies in $\mathcal{C}$. Since $n_1 \leq k$, there are exactly q^{k-n_1} codewords of the form $\mathbf{0}{*}{*}$. Since $n_1 < k$, we have that $q^{k-n_1} > 1$ and so there exists a codeword $\mathbf{c}_1 \in \mathcal{C} \setminus \{\mathbf{c}_0\}$ of the form $\mathbf{c}_1 = \mathbf{0}{*}\mathbf{y}$ for some word $\mathbf{y}$ of length n_3. Similarly, since $n_2 < k$ there exists a codeword $\mathbf{c}_2$ distinct from $\mathbf{c}_0$ of the form $\mathbf{c}_2 = {*}\mathbf{0}{*}$, and since $n_3 < k$ there is a codeword $\mathbf{c}_3$ distinct from $\mathbf{c}_1$ of the form ${*}{*}\mathbf{y}$. If $\mathbf{y} = \mathbf{0}$, then the codewords $\mathbf{c}_0, \mathbf{c}_1$ and $\mathbf{c}_2$ violate IPP1; if $\mathbf{y} \neq \mathbf{0}$ then the sets $\{\mathbf{c}_0, \mathbf{c}_3\}$ and $\{\mathbf{c}_1, \mathbf{c}_2\}$ violate IPP2. So $\mathcal{C}$ is not an IPP code.

Case 2: $n = 3k-2$. Note that since $n > 4$, we have that $k \geq 3$. Define $n_1 = n_2 = k-1$ and $n_3 = k$. As before, we can write any codeword in the form $\mathbf{x}_1\mathbf{x}_2\mathbf{x}_3$, where $\ell(\mathbf{x}_i) = n_i$. Let $\mathbf{c}_0 \in \mathcal{C}$ be the all-zero word. Since $n_1 < k$, there exists a codeword $c_1 \in \mathcal{C} \setminus \{\mathbf{c}_0\}$ of the form $\mathbf{0}{*}\mathbf{y}$ for some word $\mathbf{y}$ of length n_3. Since $n_2 = k-1$, there are $q-1$ non-zero codewords of the form ${*}\mathbf{0}{*}$; moreover no two distinct words of this form can agree anywhere in their last n_3 positions as this would contradict the MDS property of the code. This implies (since $q \geq 3$ and $n_3 = k \geq 3$) that we may choose a codeword $\mathbf{c}_2$ distinct from $\mathbf{c}_0$ of the form $\mathbf{c}_2 = {*}\mathbf{0}\mathbf{z}$ where $d(\mathbf{y}, \mathbf{z}) \geq 2$. Let $\mathbf{w}$ be a word of length n_3 such that $\mathbf{w} \in \mathrm{desc}(\mathbf{y}, \mathbf{z}) \setminus \{\mathbf{y}, \mathbf{z}\}$. Such a word exists since $d(\mathbf{y}, \mathbf{z}) \geq 2$. Let $\mathbf{c}_3$ be the (unique) codeword of the form ${*}{*}\mathbf{w}$. The sets $\{\mathbf{c}_0, \mathbf{c}_3\}$, $\{\mathbf{c}_1, \mathbf{c}_2\}$ violate IPP2. Hence the code is not an IPP code.

Case 3: $n = 3k-1$ and $k \geq 4$. We can write any word in the form $\mathbf{xy}$ where $\ell(\mathbf{x}) = 2k-2$ and $\ell(\mathbf{y}) = k+1$. Consider the set $\mathcal{D}$ of all words of the form $\mathbf{0y}$ where $\mathbf{y}$ has length $k+1$, all entries in $\mathbf{y}$ are non-zero and $\mathbf{y}$ does not occur as a suffix of a codeword. There are exactly $(q-1)^k$ codewords that end in k non-zero symbols, and so $|\mathcal{D}| \geq (q-1)^{k+1} - (q-1)^k = (q-1)^k(q-2)$. We aim to show that $\mathcal{C}$ cannot be prolific since it cannot have all the words in $\mathcal{D}$ as descendants.

Note that the all-zero word cannot be a parent of any word $\mathbf{d} \in \mathcal{D}$. To see this, note that the all-zero word cannot contribute to any of the last $k+1$ components of $\mathbf{d}$, and so these $k+1$ components must come from the other parent. But this would mean that the last $k+1$ entries of $\mathbf{d}$ would be a suffix of a codeword, contradicting the definition of $\mathcal{D}$.

The MDS property of the code shows that any non-zero codeword in $\mathcal{C}$ has at most $k-1$ zero entries (for otherwise the codeword would be too close to the all-zero codeword). Since any word $\mathbf{d} \in \mathcal{D}$ begins with $2(k-1)$ zeroes, any pair of parents $\mathbf{c}_1$ and $\mathbf{c}_2$ for $\mathbf{d}$ must each have $k-1$ zeroes in their first $2(k-1)$ positions, and the positions where the zeroes of $\mathbf{c}_1$ occur must be disjoint from the positions where the zeroes of $\mathbf{c}_2$ occur. Without loss of generality, we may assume that $\mathbf{c}_1$ has a zero in its first position. There are $\frac{1}{2}\binom{2k-2}{k-1}$ choices for the positions where $\mathbf{c}_1$ is zero; the positions where $\mathbf{c}_2$ is zero are determined by this choice. By the MDS property, there are exactly $q-1$ choices for a non-zero codeword $\mathbf{c}_1$ with zeroes in the specified positions; similarly, there are $q-1$ choices for $\mathbf{c}_2$. Each pair of codewords $\{\mathbf{c}_1, \mathbf{c}_2\}$ gives rise to at most 2^{k+1} descendants which start with $2k-2$ zeroes. So the pair $\{\mathbf{c}_1, \mathbf{c}_2\}$ can have at most $2^{k+1}-2$ descendants in $\mathcal{D}$, since no element of $\mathcal{D}$ ends with the suffix of a codeword. Moreover, the MDS property shows that for every choice of

$\mathbf{c}_1$, there is a unique choice for $\mathbf{c}_2$ that agrees with $\mathbf{c}_1$ in its last position. When $\mathbf{c}_2$ is of this form, the pair gives rise to at most $2^k - 2$ descendants in $\mathcal{D}$. Thus $\mathcal{C}$ can have at most

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

descendants in $\mathcal{D}$. For $\mathcal{C}$ to be prolific, all words in $\mathcal{D}$ must be descendants, and so

$$\binom{2k-2}{k-1}(q-1)\left((q-2)(2^k-1)+(2^{k-1}-1)\right) \geq |\mathcal{D}| \geq (q-1)^k(q-2).$$

Using the fact that $q \geq n-k+1 = 2k$, we find that there are no solutions k and q to this inequality, since we are assuming that $k \geq 4$. So there are no prolific codes in this case, as required.

Case 4: $n = 3k-1$ with $k = 2$. Note that $(5, q, q^2)$-codes of minimum distance $5-1=4$ do not exist when $q = 3$, and so we may assume that $q \geq 4$.

A descendant of $\mathcal{C}$ must agree with one of its parents in at least 3 positions, and so is at distance 2 from this parent. So the set of descendants is contained in the union of the spheres of radius 2 about codewords. We show that these spheres cannot cover all words, and so $\mathcal{C}$ cannot be prolific.

We begin by counting the number of words $\mathbf{d}$ that are in spheres of radius 2 about two codewords $\mathbf{c}_1$ and $\mathbf{c}_2$. Note that since all codewords are at distance at least 4, the word $\mathbf{d}$ has distance exactly 2 from both $\mathbf{c}_1$ and $\mathbf{c}_2$, and the distance from $\mathbf{c}_1$ to $\mathbf{c}_2$ is exactly 4. This implies that the positions where $\mathbf{c}_1$ and $\mathbf{d}$ differ must be disjoint from the positions where $\mathbf{c}_2$ and $\mathbf{d}$ differ, and so no codeword $\mathbf{d}$ lies in more than two spheres of radius 2 about codewords, since we cannot have three pairwise disjoint 2-subsets of a 5-set.

There are $5(q-1)$ codewords at distance 4 from a fixed codeword and so the number of pairs of codewords at distance 4 is $5q^2(q-1)/2$. Each such pair gives rise to exactly $\binom{4}{2} = 6$ words that lie in spheres of radius 2 about both codewords. So the number of words that are in two spheres of radius 2 is exactly $15q^2(q-1)$ (and no words lie in 3 or more spheres). Hence the number of words in spheres of radius 2 about $\mathcal{C}$ is

$$q^2(1+5(q-1)+10(q-1)^2) - 15q^2(q-1) = q^2(1-10(q-1)+10(q-1)^2).$$

Since $q \geq 4$, this expression is less than q^5 and so there are words that are not descendants of the code. Thus $\mathcal{C}$ is not a prolific IPP code as required.

Case 5: $n = 3k-1$ and $k = 3$. We may assume that $q \geq n-k+1 = 6$. Indeed, since there do not exist 5 mutually orthogonal Latin squares of order 6, no $(8, 6, 6^3)$-code with minimum distance 6 exists and so we may assume that $q \geq 7$. We show the code cannot be prolific by showing that the number of descendants of the code is less than q^8.

There are $q^3(1+8(q-1)+\binom{8}{2}(q-1)^2)$ words within spheres of radius 2 about codewords. All these words are descendants, by the MDS property of the code. It remains to count descendants of distance at least 3 from every codeword.

Let p_i be the number of (unordered) pairs of codewords at distance i. So $p_i = 0$ when $1 \leq i \leq 5$. An upper bound for the number of descendants at distance at least 3 from the code is

$$p_6\binom{6}{3} + p_7\left(\binom{7}{3}+\binom{7}{4}\right) + p_8\left(\binom{8}{3}+\binom{8}{4}+\binom{8}{5}\right). \qquad (3.1)$$

We count the number s_6 of codewords $\mathbf{x}$ at distance 6 from a fixed codeword $\mathbf{c}$ as follows. There are exactly $\binom{8}{2}q$ pairs $(\mathbf{x}, I)$, where $\mathbf{x}$ is a codeword, I is a 2-subset of $\{1, 2, \ldots, 8\}$, and $c_i = x_i$ for $i \in I$. The codeword $\mathbf{c}$ appears $\binom{8}{2}$ times in a pair and every codeword $\mathbf{x}$ at distance 6 appears exactly once, so

$$s_6 + \binom{8}{2} = \binom{8}{2}q.$$

Write s_7 for the number of codewords at distance 7 from $\mathbf{c}$. Counting pairs as above, but with $|I| = 1$ shows that

$$s_7 + 2s_6 + \binom{8}{1} = \binom{8}{1}q^2.$$

Finally, writing s_8 for the number of codewords at distance 8 from $\mathbf{c}$ we see that $s_8 = q^3 - s_6 - s_7 - 1$. Solving these equations shows that

$$\begin{aligned} s_6 &= 28q - 28 \\ s_7 &= 8q^2 - 56q + 48 \\ s_8 &= q^3 - 8q^2 + 28q - 21. \end{aligned}$$

Now $p_i = q^3 s_i/2$. Substituting these values into Equation 3.1 above, and adding the term which counts descendants at distance at most 2 from the code, we can easily check that the number of descendants is less than q^8 whenever $q \geq 7$ and so $\mathcal{C}$ cannot be prolific in this case.

Case 6: $n \geq 3k$. Consider the set $\mathcal{D}$ of words $\mathbf{0y}$, where $\ell(\mathbf{y}) = k + 1$ and $\mathbf{y}$ is not a suffix of a codeword and contains no zero entries. Note that $|\mathcal{D}| \geq (q-1)^{k+1} - (q-1)^k = (q-1)^k(q-2) > 0$, and so $\mathcal{D}$ is non-empty. We show that no word in $\mathcal{D}$ can be a descendant of $\mathcal{C}$, and so $\mathcal{C}$ cannot be prolific.

Suppose, for a contradiction, that $\mathbf{d} \in \mathcal{D}$ is a descendant of $\mathcal{C}$. A descendant must agree with one of its parents in at least k of its first $2k - 1$ positions. Since $\mathbf{d}$ begins with $2k - 1$ zeroes, and so one of its parents must have k zero entries and so must be the all-zero codeword. But the final $k + 1$ entries of $\mathbf{d}$ are non-zero, and so the other parent must have contributed them. But this implies that the last $k + 1$ entries of $\mathbf{d}$ is a suffix of this parent, contradicting the definition of $\mathcal{D}$ as required.

From all these six cases we deduce that there are no prolific IPP (n, q, q^k)-codes of minimum distance $n - k + 1$ when $n \geq 3$, unless $n = 4$, $q = 3$ and $k = 2$. So the theorem is established. □

4. Linear IPP codes. Linear IPP codes were considered first in [5]. (An $[n, k]$ *linear IPP code* is an IPP code which is a k-dimensional subspace of the vector space $(\mathbb{F}_q)^n$.) We begin this section with a classification of prolific linear IPP codes (Theorem 4.1). We then prove a new upper bound on the size of a linear IPP code (Theorem 4.4) which emerged from our investigations of the prolific case.

THEOREM 4.1. *The only linear non-binary prolific IPP code of length* 3 *or more is the* $(4, 3, 9)$ *IPP code.*

Proof. Let $\mathcal{C}$ be a prolific $[n, k]$ linear IPP code, where $n \geq 3$. By Theorem 3.1 there are no prolific MDS IPP codes other than the $(4, 3, 9)$ example, so we may assume that $\mathcal{C}$ is not an MDS code. The theorem follows if we can derive a contradiction from this assumption.

Since $\mathcal{C}$ is not MDS, we may permute the columns of the code so that the last k columns of the generator matrix $\mathcal{G}$ form a $k \times k$ matrix with rank $k-1$.

We can write each codeword in the form $\mathbf{xy}$ where $\ell(\mathbf{x}) = n-k$ and $\ell(\mathbf{y}) = k$. Consider a word $\mathbf{0y}$ where $\mathbf{y}$ has no zeroes and is not a suffix of any codeword. A choice for $\mathbf{y}$ certainly exists, since there are at most $(q-1)^{k-1}$ suffixes of length k of codewords with no zeroes, by our condition on the generator matrix $\mathcal{G}$, and so there are at least $(q-1)^k - (q-1)^{k-1}$ choices for $\mathbf{y}$.

Since $\mathcal{C}$ is prolific it follows that $\mathbf{0y}$ is a descendant, so there exist codewords $\mathbf{x}_1\mathbf{y}_1, \mathbf{x}_2\mathbf{y}_2 \in \mathcal{C}$ such that $\mathbf{0y} \in \mathrm{desc}(\mathbf{x}_1\mathbf{y}_1, \mathbf{x}_2\mathbf{y}_2)$. Since $\mathbf{y}$ is not a suffix of a codeword these parents are distinct; moreover, neither parent can be the all zero codeword. Clearly, $\mathbf{0y}_1 \in \mathrm{desc}(\mathbf{x}_1\mathbf{y}_1, \mathbf{x}_2\mathbf{y}_2)$. The suffix $\mathbf{y}_1$ appears in q codewords of $\mathcal{C}$ by our condition on $\mathcal{G}$, and hence there is a codeword $\mathbf{x}_3\mathbf{y}_1$ where $\mathbf{x}_3 \notin \{\mathbf{x}_1, \mathbf{x}_2\}$. But then we have that $\mathbf{0y}_1 \in \mathrm{desc}(\mathbf{00}, \mathbf{x}_3\mathbf{y}_1)$ which implies that $\mathcal{C}$ is not an IPP code. This contradiction establishes the theorem, as required. □

Hollmann *et al.* [9] proved several upper bounds on the size of IPP codes. One of their results can be stated as follows:

THEOREM 4.2. *Let $\mathcal{C}$ be an IPP code of length* 3*, where position i, $1 \le i \le 3$, of a codeword is taken from an alphabet Q_i. Then*

$$|\mathcal{C}| \le |Q_1| + |Q_2| + |Q_3| - 1 \ . \tag{4.1}$$

Hollmann *et al.* used Equation (4.1) to obtain the following bounds on the size of IPP codes:

THEOREM 4.3. *Let $\mathcal{C}$ be an (n, q, M) IPP code.*

(i) *If $n = 3\ell - 2$ then $M \le q^\ell + 2q^{\ell-1} - 1$.*

(ii) *If $n = 3\ell - 1$ then $M \le 2q^\ell + q^{\ell-1} - 1$.*

(iii) *If $n = 3\ell$ then $M \le 3q^\ell - 1$.*

Proof. Write $n = k_1 + k_2 + k_3$, where $k_i \in \{\ell - 1, \ell\}$. Define $Q_i = F^{k_i}$. We may write any codeword in $\mathcal{C}$ in the form $\mathbf{x}_1\mathbf{x}_2\mathbf{x}_3$ where $\ell(\mathbf{x}_i) = k_i$. So $\mathcal{C}$ can be regarded as a length 3 code, with symbols in position i taken from the alphabet Q_i. It is easy to verify that $\mathcal{C}$ is still an IPP code when thought of in this way, and so the bound follows by Theorem 4.2. □

Theorem 4.3 implies the following bounds on the dimension k of an $[n, k]$ linear IPP code when $q > 3$:

(i) if $n \equiv 0 \bmod 3$ then $k \le \frac{n}{3}$.

(ii) if $n \equiv 1 \bmod 3$ then $k \le \frac{n+2}{3}$.

(iii) if $n \equiv 2 \bmod 3$ then $k \le \frac{n+1}{3}$.

The theorem below improves Theorem 4.3 when $n \equiv 1 \bmod 3$.

THEOREM 4.4. *Let $\mathcal{C}$ be an $[n, k]$ linear IPP code, and suppose that $k \ge 3$. Then $k \le \lfloor \frac{n+1}{3} \rfloor$.*

The main part of the proof of Theorem 4.4 is contained in the following lemma:

LEMMA 4.5. *Let $\mathcal{C}$ be a non-binary $[n, k]$ code, and let $\mathcal{G}$ be a generator matrix for $\mathcal{C}$. Let k_1, k_2 and k_3 be positive integers such that $n = k_1 + k_2 + k_3$, and suppose that the following properties hold:*

1. *The first k_1 columns of $\mathcal{G}$ have rank less than k;*
2. *The next k_2 columns of $\mathcal{G}$ have rank less than k;*
3. $1 \le k_3 \le k$*;*
4. *When the final k_3 columns of $\mathcal{G}$ have rank k, we have that $k \ge 3$.*

Then $\mathcal{C}$ is not an IPP code.

Proof. Suppose that the final k_3 columns of $\mathcal{G}$ have rank less than k. We show that $\mathcal{G}$ is not an IPP code in this 'small rank' case.

We may write any codeword $\mathbf{c} \in \mathcal{C}$ in the form $\mathbf{c} = \mathbf{xyz}$ where $\ell(\mathbf{x}) = k_1$, $\ell(\mathbf{y}) = k_2$, and $\ell(\mathbf{z}) = k_3$.

Let $\mathbf{c}_1$ be the all-zero codeword. By Property 1 there exists a non-zero codeword $\mathbf{c}_2$ of the form $\mathbf{c}_2 = \mathbf{0}{*}\mathbf{z}_2$. Similarly, by Property 2 there exists a non-zero codeword $\mathbf{c}_3$ of the form $\mathbf{c}_3 = {*}\mathbf{0}\mathbf{z}_3$. Indeed, there are at least $q-1$ choices for $\mathbf{c}_3$. Since $q-1 \geq 3-1 = 2$, we may therefore choose $\mathbf{c}_3$ so that $\mathbf{c}_3 \neq \mathbf{c}_2$.

If $|\{\mathbf{0}, \mathbf{z}_2, \mathbf{z}_3\}| < 3$ then $\{\mathbf{c}_1, \mathbf{c}_2, \mathbf{c}_3\}$ violates IPP1. Otherwise, let $\mathbf{c}_4 \in \mathcal{C}$ be of the form $\mathbf{c}_4 = {*}{*}\mathbf{z}_2$ where $\mathbf{c}_4 \neq \mathbf{c}_2$. Such a codeword exists, since we are assuming the rank of the last k_3 columns of $\mathcal{C}$ is less than k. Then the pairs $\{\mathbf{c}_1, \mathbf{c}_4\}$ and $\{\mathbf{c}_2, \mathbf{c}_3\}$ violate IPP2. So $\mathcal{C}$ is not an IPP code in this 'small rank' case.

Assume that we are not in the small rank case, so the final k_3 columns of $\mathcal{G}$ have rank k. This implies in particular that $k_3 = k$. Moreover, Property 4 implies that $k \geq 3$.

Choose distinct codewords $\mathbf{c}_1 = \mathbf{000}$, $\mathbf{c}_2 = \mathbf{0}{*}\mathbf{z}_2$ and $\mathbf{c}_3 = {*}\mathbf{0}\mathbf{z}_3$ as above. If $|\{\mathbf{0}, \mathbf{z}_2, \mathbf{z}_3\}| < 3$ then $\{\mathbf{c}_1, \mathbf{c}_2, \mathbf{c}_3\}$ violates IPP1. So it remains to deal with the case when $\mathbf{z}_2 \neq \mathbf{z}_3$ (and both $\mathbf{z}_2$ and $\mathbf{z}_3$ are non-zero).

We claim that we may assume $d(\mathbf{z}_2, \mathbf{z}_3) > 1$. For suppose that $d(\mathbf{z}_2, \mathbf{z}_3) = 1$. By Property 1, there are at least q choices for a codeword starting with k_1 zeroes, and so there exists a codeword $\mathbf{c}_4$ of the form $\mathbf{c}_4 = \mathbf{0}{*}\mathbf{z}_4$ where $\mathbf{c}_4 \neq \mathbf{c}_1$ and $\mathbf{c}_4 \neq \mathbf{c}_2$. Suppose that $\mathbf{z}_2$ and $\mathbf{z}_4$ agree in some component (the ith say). Now, $\mathbf{c}_2$ and $\mathbf{c}_4$ are distinct codewords which agree in their first k_1 positions and their $(k_1 + k_2 + i)$-th position, and so the first k_1 columns of $\mathcal{G}$ together with the (k_1+k_2+i)-th column has rank less than k. We may replace $\mathcal{C}$ by an equivalent code, moving the $(k_1 + k_2 + i)$-th column of $\mathcal{G}$ to the left hand side of the generator matrix. This new code is not an IPP code, by the small rank case using the partition $k_1' = k_1 + 1$, $k_2' = k_2$ and $k_3' = k_3 - 1$ of the columns of the generator matrix. So $\mathcal{C}$ is not an IPP code in this case. So we may assume that $d(\mathbf{z}_2, \mathbf{z}_4) = k$. This implies, since $d(\mathbf{z}_2, \mathbf{z}_3) = 1$ and $k \geq 3$, that $d(\mathbf{z}_4, \mathbf{z}_3) \geq k-1 > 1$. So our claim follows, since we may replace $\mathbf{c}_2$ by $\mathbf{c}_4$ if necessary.

By the previous paragraph, we may assume (without loss of generality) that $d(\mathbf{z}_2, \mathbf{z}_3) > 1$. Let $\mathbf{z}_5 \in \mathrm{desc}(\mathbf{z}_2, \mathbf{z}_3\} \setminus \{\mathbf{z}_2, \mathbf{z}_3\}$; we may choose $\mathbf{z}_5$ of this form since $d(\mathbf{z}_2, \mathbf{z}_3) > 1$. Since $\ell(\mathbf{z}_3) = k_3$ and the last k_3 columns of $\mathcal{C}$ are linearly independent there exists a codeword $\mathbf{c}_5$ of the form $\mathbf{c}_5 = {*}{*}\mathbf{z}_5$. But the two pairs $\{\mathbf{c}_1, \mathbf{c}_5\}$, $\{\mathbf{c}_2, \mathbf{c}_3\}$ violate IPP2. Hence $\mathcal{C}$ is not an IPP code, as required. □

Proof of Theorem 4.4. Suppose that $n \leq 3k-2$. We can write $n = k_1 + k_2 + k_3$ where $k_1 < k$, $k_2 < k$ and $k_3 = k \geq 3$. But then $\mathcal{C}$ contradicts Lemma 4.5. Therefore $n \geq 3k-1$ and so $k \leq \frac{n+1}{3}$. Since k is an integer, the theorem follows. □

A further improvement for non-MDS codes is given in the following corollary to Lemma 4.5.

Corollary 4.6. *Let $\mathcal{C}$ be an $[n,k]$ linear IPP code which is not an MDS code. If $k \geq 3$ then $k \leq \lfloor \frac{n}{3} \rfloor$.*

Proof. By Theorem 4.4, no $[n,k]$ linear IPP code exists when $n \leq 3k-2$. Suppose, for a contradiction, that $n = 3k-1$. As $\mathcal{C}$ is not an MDS code, we may (by permuting the columns of $\mathcal{C}$ if necessary) assume that the first k columns of the generator matrix for $\mathcal{C}$ are linearly dependent. But now Lemma 4.5 (in the case when $k_1 = k$, $k_2 = k-1$ and $k_3 = k$) shows that $\mathcal{C}$ is not an IPP code and we have a contradiction. Therefore, $n \geq 3k$ and so $k \leq \frac{n}{3}$. The corollary now follows since k is an integer. □

5. New upper bound on the size of IPP codes. This section establishes a new upper bound on the size of an IPP code, which improves the leading coefficient in the bound of Theorem 4.3. When the code has length 5, our techniques yield especially good results (and we will need a good bound in Section 8). We begin by considering this special case.

LEMMA 5.1. *Let $\mathcal{C}$ be a $(5, q, M)$ IPP code, where $M > q^2$. Then the minimum distance $d(\mathcal{C})$ of $\mathcal{C}$ is at least* 3.

Proof. Assume, for a contradiction, that $d(\mathcal{C}) \leq 2$. Suppose $\mathbf{c}_1$ and $\mathbf{c}_2$ are codewords at distance 1 or 2. Without loss of generality, assume that $\mathbf{c}_1$ and $\mathbf{c}_2$ agree in their first 3 positions. Since $M > q^2$, there exist distinct codewords $\mathbf{c}_3$ and $\mathbf{c}_4$ that agree in their final two positions. If the sets $\{\mathbf{c}_1, \mathbf{c}_3\}$ and $\{\mathbf{c}_2, \mathbf{c}_4\}$ are disjoint, then IPP2 is violated. Otherwise, the set $\{\mathbf{c}_1, \mathbf{c}_2, \mathbf{c}_3, \mathbf{c}_4\}$ has size 3 and IPP1 is violated. In either case, we have a contradiction, as required. □

Theorem 4.3 implies that for a $(5, q, M)$ IPP code we have $M \leq 2q^2 + q - 1$. The theorem below significantly improves this bound.

THEOREM 5.2. *If $\mathcal{C}$ is a $(5, q, M)$ IPP code then $M < \frac{5}{4}q^2 + 5q$.*

Proof. If there is a symbol $x \in F$ and a position $i \in \{1, 2, 3, 4, 5\}$ such that x occurs just once as the ith position of a codeword, we remove this codeword to produce a smaller code. Repeating this process as often as is necessary, we eventually obtain a code $\mathcal{C}'$ in which no symbol appears exactly once in any fixed position. Note that we have removed at most $5q$ codewords to obtain $\mathcal{C}'$, and so $\mathcal{C}'$ has M' codewords where $M - M' \leq 5q$. To prove the theorem, it suffices to show that $M' < \frac{5}{4}q^2$. The theorem follows trivially when $M' \leq q^2$, and so we may assume that $M' = q^2 + \mu$ for some positive integer μ.

For integers i and j such that $1 \leq i < j \leq 5$, define the subset $S_{ij} \subseteq \mathcal{C}'$ by

$$S_{ij} = \{\mathbf{x} \in \mathcal{C}' : \exists \mathbf{y} \in \mathcal{C}' \setminus \{\mathbf{x}\} \text{ such that } x_i = y_i \text{ and } x_j = y_j\}.$$

Note that $|S_{ij}| > \mu$.

We claim that $S_{ij} \cap S_{i'j'} = \emptyset$ whenever $\{i, j\}$ and $\{i', j'\}$ are disjoint pairs of positions. Without loss of generality, it is sufficient to show that $S_{12} \cap S_{34} = \emptyset$. Suppose, for a contradiction, that $\mathbf{c}_1 \in S_{12} \cap S_{34}$. Writing $\mathbf{c}_1 = x_1x_2x_3x_4x_5$, there exist codewords $\mathbf{c}_2, \mathbf{c}_3 \in \mathcal{C}' \setminus \{\mathbf{c}_1\}$ of the form $\mathbf{c}_2 = x_1x_2{*}{*}y$ and $\mathbf{c}_3 = {*}{*}x_3x_4z$. Note that $\mathbf{c}_2 \neq \mathbf{c}_3$, by Lemma 5.1. If $|\{x_5, y, z\}| < 3$ then $\{\mathbf{c}_1, \mathbf{c}_2, \mathbf{c}_3\}$ violates IPP1. If x_5, y and z are distinct, let $\mathbf{c}_4 \in \mathcal{C}'$ be another codeword ending in y (which exists by our choice of $\mathcal{C}'$). Then the sets $\{\mathbf{c}_1, \mathbf{c}_4\}$, $\{\mathbf{c}_2, \mathbf{c}_3\}$ violate IPP2. This contradiction establishes our claim.

For any given disjoint pairs $\{i_1, j_1\}, \{i_2, j_2\} \subseteq \{1, 2, 3, 4, 5\}$ define $Q_{i_1j_1i_2j_2}$ to be the set of symbols that occur in the ℓth positions of codewords in $S_{i_1j_1}$, where ℓ is the unique position not equal to any of i_1, j_1, i_2, j_2.

We claim that $Q_{i_1j_1i_2j_2} \cap Q_{i_2j_2i_1j_1} = \emptyset$ for any disjoint pairs $\{i_1, j_1\}$ and $\{i_2, j_2\}$. (In particular, since there are q symbols in total, this claim implies that $|Q_{i_1j_1i_2j_2}| + |Q_{i_2j_2i_1j_1}| \leq q$.) To see why our claim holds, we show (without loss of generality) that $Q_{1234} \cap Q_{3412} = \emptyset$. Assume, for a contradiction, that $x \in Q_{1234} \cap Q_{3412}$. Then the following 4 codewords lie in $\mathcal{C}'$: $\mathbf{c}_1 = x_1x_2{*}{*}x$, $\mathbf{c}_2 = x_1x_2{*}{*}{*}$, $\mathbf{c}_3 = {*}{*}x_3x_4x$ and $\mathbf{c}_4 = {*}{*}x_3x_4{*}$. (These codewords are distinct, since $S_{12} \cap S_{34} = \emptyset$.) But then the pairs $\{\mathbf{c}_1, \mathbf{c}_4\}$ and $\{\mathbf{c}_2, \mathbf{c}_3\}$ violate IPP2, and so our claim follows.

There are $\binom{5}{2}\binom{3}{2} = 30$ subsets $Q_{i_1j_1i_2j_2} \subseteq F$, and the previous paragraph shows that at least half of them are 'small' in the sense of satisfying $|Q_{i_1j_1i_2j_2}| \leq \frac{1}{2}q$.

The map from S_{12} to $Q_{1235} \times Q_{1234}$ where $x_1x_2x_3x_4x_5 \mapsto x_4x_5$ is injective, since $S_{12} \cap S_{45} = \emptyset$. Thus $|S_{12}| \leq |Q_{1235}||Q_{1234}|$. Arguing similarly, we find that

$$|S_{12}| \leq \min\{|Q_{1234}||Q_{1235}|, |Q_{1234}||Q_{1245}|, |Q_{1235}||Q_{1245}|\}.$$

A similar inequality exists for any subset S_{ij}. There are in total $\binom{5}{2} = 10$ such inequalities, each involving 3 subsets $Q_{iji'j'}$. Averaging over all pairs $\{i,j\}$, the expected number the subsets $Q_{iji'j'}$ in such an inequality which satisfy $|Q_{iji'j'}| \leq \frac{1}{2}q$ is at least $\frac{3}{2}$. So we may find a pair $\{i,j\}$ such that the inequality involves at least two sets $Q_{iji'j'}$ of size at most $\frac{1}{2}q$. But then the inequality implies that $|S_{ij}| \leq (\frac{1}{2}q)^2 = \frac{1}{4}q^2$. Since $\frac{1}{4}q^2 \geq |S_{ij}| > \mu$, we find that $M' = q^2 + \mu < \frac{5}{4}q^2$, as required. □

We comment that, arguing more carefully, it is possible to reduce the term $5q$ in the bound above. Indeed, we have the outline of a proof (with many special cases) that the term can be eliminated. For the sake of simplicity, we content ourselves with proving a bound that eliminates this term in the case of prolific IPP codes.

Theorem 5.3. *If $\mathcal{C}$ is a $(5,q,M)$ prolific IPP code, then $M < \frac{5}{4}q^2$.*

Proof. Suppose that there is no symbol x that appears just once as ith position of a codeword. The argument of Theorem 5.2 (where $M = M'$ in our situation) now shows that $M < \frac{5}{4}q^2$. So we may assume that there is a symbol x and a position i such that x appears exactly once as the ith position of a codeword. Replacing $\mathcal{C}$ by an equivalent code if necessary, we may assume (without loss of generality) that $x = 0$, $i = 1$ and $\mathcal{C}$ contains the all-zero word $\mathbf{0}$. So no codeword starts with 0, other than the all-zero codeword.

The word 01111 is a descendant, since $\mathcal{C}$ is prolific. Now, $\mathbf{0}$ is a parent, since no other codeword can contribute to the first position of the descendant. But $\mathbf{0}$ cannot contribute to any of the remaining positions, and so there exists a codeword $\mathbf{c}_1$ of the form $\mathbf{c}_1 = *1111$. Similarly, considering the descendant 01112, there exists a codeword of the form $\mathbf{c}_2 = *1112$. But then $d(\mathbf{c}_1, \mathbf{c}_2) \leq 2$, and so Lemma 5.1 implies that $M \leq q^2 < \frac{5}{4}q^2$, as required. □

We do not see how to generalise the bound of Theorem 5.2, as it is not clear what the analogue of the final paragraph of the proof should be. However, we are able to establish the following theorem.

Theorem 5.4. *Let $\mathcal{C}$ be an (n,q,M) IPP code where $n = 3k-1$. Then $M < \frac{3}{2}q^k + 3q^{k-1}$.*

Proof. We begin by proving the weaker bound

$$M < \frac{3}{2}q^k + \binom{n}{k-1}q^{k-1}, \tag{5.1}$$

and we will then show how our argument can be modified to give the bound of the theorem.

If there are any codewords $\mathbf{c}$ that are uniquely defined by a set of $k-1$ positions (so there exists a $(k-1)$-set X of positions such that $\{\mathbf{u} \in \mathcal{C} : c_i = u_i \,\forall i \in X\} = \{\mathbf{u}\}$) we remove them. Repeating this process as often as is necessary, we obtain a code $\mathcal{C}'$ with the property that for all $\mathbf{c} \in \mathcal{C}'$ and for all $(k-1)$-sets $X \subseteq \{1,2,\ldots,n\}$ we have that

$$|\{\mathbf{u} \in \mathcal{C}' : c_i = u_i \,\forall i \in X\}| \geq 2.$$

Note that $\mathcal{C}'$ is an (n,q,M')-code with $M - M' < \binom{n}{k-1}q^{k-1}$. If $M' \leq q^k$ then bound (5.1) holds trivially, and so we may assume that $M' > q^k$. Let μ be the positive integer such that $M' = q^k + \mu$. To show (5.1) holds, it suffices to show that $\mu \leq \frac{1}{2}q^2$.

For a subset $T \subseteq \{1, 2, \ldots, n\}$ of k positions, define a subset $S_T \subseteq \mathcal{C}'$ by

$$S_T = \{\mathbf{x} \in \mathcal{C}' : \exists \mathbf{y} \in \mathcal{C}' \setminus \{\mathbf{x}\} \text{ such that } x_i = y_i \text{ for } i \in T\}.$$

Note that $|S_T| > \mu$. We claim that $S_{T_1} \cap S_{T_2} = \emptyset$ whenever T_1 and T_2 are disjoint. To see this, assume (without loss of generality) that $T_1 = \{1, 2, \ldots, k\}$ and $T_2 = \{k+1, k+2, \ldots, 2k\}$ and suppose (for a contradiction) that $\mathbf{c} \in S_{T_1} \cap S_{T_2}$. We may write $\mathbf{c} = \mathbf{xyz}$ where $\ell(\mathbf{x}) = \ell(\mathbf{y}) = k$ and $\ell(\mathbf{z}) = k-1$. Since $\mathbf{c} \in S_{T_1} \cap S_{T_2}$, there exist codewords $\mathbf{c}_2, \mathbf{c}_3 \in \mathcal{C}' \setminus \{\mathbf{c}_1\}$ of the form $\mathbf{c}_2 = \mathbf{x}{*}\mathbf{w}$ and $\mathbf{c}_3 = {*}\mathbf{yu}$. Suppose that $|\{\mathbf{z}, \mathbf{w}, \mathbf{u}\}| < 3$. Then we find that IPP1 is violated. For if $\mathbf{c}_2 \neq \mathbf{c}_3$ then $\{\mathbf{c}_1, \mathbf{c}_2, \mathbf{c}_3\}$ violates IPP1. Moreover, if $\mathbf{c}_2 = \mathbf{c}_3$ (which implies $\mathbf{z} \neq \mathbf{w}$), then $\{\mathbf{c}_1, \mathbf{c}_2, \mathbf{c}_4\}$ violates IPP1, where $\mathbf{c}_4 \in \mathcal{C}$ has $\mathbf{w}$ as a suffix and is distinct from $\mathbf{c}_2$. (Our construction of $\mathcal{C}'$ guarantees that $\mathbf{c}_4$ exists.) Now suppose that $|\{\mathbf{z}, \mathbf{w}, \mathbf{u}\}| = 3$. We find that IPP2 is violated by the sets $\{\mathbf{c}_1, \mathbf{c}_4\}$ and $\{\mathbf{c}_2, \mathbf{c}_3\}$, where $\mathbf{c}_4$ is defined as before. So we have a contradiction, and therefore our claim follows.

Define $S_1 = S_{\{1,2,\ldots,k\}}$ and $S_2 = S_{\{k+1,k+2,\ldots,2k\}}$. For $i = 1, 2$, let P_i be the set of length $k-1$ suffixes of codewords in S_i. We claim that $P_1 \cap P_2 = \emptyset$. To see this, suppose (for a contradiction) that there exists a suffix $\mathbf{z} \in P_1 \cap P_2$. Since $\mathbf{z} \in P_1$, there exist distinct codewords of the form $\mathbf{c}_1 = \mathbf{x}{*}\mathbf{z}$ and $\mathbf{c}_2 = \mathbf{x}{*}{*}$. Since $\mathbf{z} \in P_2$, there exist distinct codewords of the form $\mathbf{c}_3 = {*}\mathbf{yz}$ and $\mathbf{c}_4 = {*}\mathbf{y}{*}$. Since $\{\mathbf{c}_1, \mathbf{c}_2\} \subseteq S_1$ and $\{\mathbf{c}_3, \mathbf{c}_4\} \subseteq S_2$, and since $S_1 \cap S_2 = \emptyset$, we find that the codewords $\mathbf{c}_i$ are pairwise distinct. But then the sets $\{\mathbf{c}_1, \mathbf{c}_4\}$ and $\{\mathbf{c}_2, \mathbf{c}_3\}$ violate IPP2. This contradiction shows that $P_1 \cap P_2 = \emptyset$, as required.

Since P_1 and P_2 are disjoint subsets of a set of the q^{k-1} possible suffixes of length $k-1$, we find that $|P_i| \leq \frac{1}{2}q^{k-1}$ for some i. Suppose that $|P_1| \leq \frac{1}{2}q^{k-1}$, so the number of length $k-1$ suffixes of codewords in S_1 is at most $\frac{1}{2}q^{k-1}$. The number of suffixes of length k of codewords in S_1 is therefore at most $\frac{1}{2}q^k$, and the length k suffixes of any two codewords in S_1 are distinct, since $S_1 \cap S_{\{2k,2k+1,\ldots,3k-1\}} = \emptyset$. Thus $\frac{1}{2}q^k \geq |S_1| > \mu$, and so (5.1) holds in this case. In the case when $|P_2| \leq \frac{1}{2}q^{k-1}$, a similar argument establishes (5.1): instead of suffixes, we consider sub-words consisting of the first component and the last $k-1$ components of a word and we use the fact that $S_2 \cap S_{\{1,2k+1,2k+2,\ldots,3k-1\}} = \emptyset$.

It remains to show that the above argument can be tightened in order to establish the theorem.

The argument above uses the fact that $S_{T_1} \cap S_{T_2} = \emptyset$ for a limited range of sets T_1 and T_2. (Indeed, it uses this equality when $T_1 = \{1, 2, \ldots, k\}$ and $T_2 = \{k+1, k+2, \ldots, \ldots, 2k\}$, when $T_1 = \{1, 2, \ldots, k\}$ and $T_2 = \{2k, 2k+1, \ldots, 3k-1\}$ and when $T_1 = \{k+1, k+2, \ldots, 2k\}$ and $T_2 = \{1, 2k+1, 2k+2, \ldots, 3k-1\}$.) Because of this, we see that the argument still works when $\mathcal{C}'$ is defined to be a larger subcode, where less than $3q^{k-1}$ codewords have been removed: we remove codewords that are uniquely defined by their positions in X, where $X = \{2k+1, 2k+2, \ldots, 3k-1\}$, $X = \{k+1, k+2, \ldots, 2k-1\}$ or $X = \{2, 3, \ldots, k\}$. This modification establishes the theorem, as required. □

6. Prolific IPP codes of length 3. The aim of this section is to prove Theorem 6.7, which states that there are no non-trivial prolific IPP codes of length 3.

LEMMA 6.1. *Let $\mathcal{C}$ be a non-binary prolific IPP code of length* 3. *Then $|\mathcal{C}| > q$, and the minimum distance $d(\mathcal{C})$ of $\mathcal{C}$ is at least* 2.

Proof. If $\mathcal{C}$ contains q or fewer codewords, then the bound of Theorem 2.1 is violated. So $|\mathcal{C}| > q$.

Suppose $\mathcal{C}$ contains codewords $\mathbf{x}$ and $\mathbf{y}$ at distance 1. There exist distinct codewords $\mathbf{u}, \mathbf{v} \in \mathcal{C}$ that agree at the position where $\mathbf{x}$ and $\mathbf{y}$ disagree, since $|\mathcal{C}| > q$. If the codewords $\mathbf{x}, \mathbf{y}, \mathbf{u}, \mathbf{v}$ are not distinct (and so they form a set of size 3) then IPP1 is violated; if the codewords are distinct then IPP2 is violated. This contradiction shows that $d(\mathcal{C}) \geq 2$. □

Lemma 6.2. *There are no non-trivial prolific IPP codes of length* 3 *when* $q \geq 6$.

Proof. Let $\mathcal{C}$ be a $(3, q, M)$ prolific IPP code. Every symbol must occur at least once at the start of a codeword, since if there are no codewords of the form $x{*}{*}$ then there are no descendants of the form $x{*}{*}$, contradicting the fact that $\mathcal{C}$ is prolific. Since $M \leq 3q - 1$ by Theorem 4.3, there is a symbol that occurs (A) exactly once or (B) exactly twice as the first position of a codeword. Without loss of generality, suppose this value is 0.

Case (A): Without loss of generality, we may assume that $000 \in \mathcal{C}$ and that no other codeword starts with 0. There are $(q-1)^2$ words $\mathbf{d}$ of the form $0ab$ where $a, b \neq 0$, and all of these must occur as descendants. Now 000 must be a parent of $\mathbf{d}$ (since no other codeword starts with 0), but cannot contribute to the remaining two positions of $\mathbf{d}$ (since $a, b \neq 0$). So the other parent must be a codeword of the form ${*}ab$. Thus there are at least $(q-1)^2$ codewords not equal to 000. But $|\mathcal{C}| \geq (q-1)^2 + 1$ contradicts Theorem 4.3 since $q \notin \{3, 4\}$.

Case (B): We may assume that $000, 0xy \in \mathcal{C}$ for some $x, y \in \{0, 1, \ldots, q-1\}$ that are not both zero. A similar argument to (A), with the codewords d of the form $0ab$ where $a \notin \{0, x\}$ and $b \notin \{0, y\}$, shows that there are at least $(q-2)^2$ other codewords, and so $|\mathcal{C}| \geq (q-2)^2 + 2$ in this case. This contradicts Theorem 4.3 when $q \geq 6$.

Thus, no $(3, q, M)$ prolific IPP codes exist if $q \geq 6$. □

The following lemma is a special case of a result of Tô and Safavi-Naini [13, Theorem 34].

Lemma 6.3. *A* 3*-ary IPP code* $\mathcal{C}$ *of length* 3 *must have* $|\mathcal{C}| \leq 4$.

Proof. Consider codewords of the form $0{*}{*}$. Suppose (for a contradiction) there are 3 such codewords. Since $d(\mathcal{C}) \geq 2$ (by Lemma 6.1) we may assume, without loss of generality, that the codewords are 000, 011, 022. There exists another codeword, and it cannot start with 0 (again since $d(\mathcal{C}) \geq 2$), so without loss of generality it is of the form $1ab$ for some $a, b \in \{0, 1, 2\}$. The minimum distance of the code implies that $a \neq b$, and so we may assume, without loss of generality, that $a = 0$ and $b = 1$ and so 101 is a codeword. But this gives a contradiction, since the set $\{000, 011, 101\}$ violates IPP1. So we may assume that any symbol is the first component of 0, 1 or 2 codewords.

Suppose (again for a contradiction) that there are two symbols, 0 and 1 say, each being the first component of exactly 2 codewords. Without loss of generality, we may assume that $000, 011 \in \mathcal{C}$. Let $1ab$, $1cd$ be the two codewords starting with 1. Now $\{a, b\} \subseteq \{0, 1\}$ leads to a contradiction, since then $\{000, 011, 1ab\}$ violates IPP1. So $ab \in \{02, 20, 12, 21, 22\}$; similarly for cd. Since $d(C) \geq 2$ we cannot have $ab = 22$ or $cd = 22$ (since otherwise $1ab$ and $1cd$ would be too close). Indeed, without loss of generality we must have $ab = 02$ and $cd = 20$, or $ab = 02$ and $cd = 21$. But in the first case, IPP1 is violated and in the second case IPP2 is violated.

So at most one symbol starts 2 codewords, the remaining symbols start at most one codeword each. Since there are three possibilities for the first symbol of a codeword, there are at most $2+1+1=4$ codewords in total. □

Corollary 6.4. *There is no* 3*-ary prolific IPP code of length* 3.

Proof. Suppose a 3-ary prolific IPP code $\mathcal{C}$ of length 3 exists. All symbols must occur as the start of a codeword, since $\mathcal{C}$ is prolific. Since $|\mathcal{C}| \leq 4$, there is a symbol that starts a unique codeword. So we are in Case (A) of Lemma 6.2. The argument there shows that $|\mathcal{C}| \geq 1+(q-1)^2 = 5$, and this contradicts Lemma 6.3, as required. □

Lemma 6.5. *There is no* 4*-ary prolific IPP code of length* 3.

Proof. Suppose, for a contradiction, that a symbol occurs exactly once as the start of a codeword. Without loss of generality $000 \in \mathcal{C}$ and no other codeword starts with 0. Considering parents of the 9 descendants of the form $0xy$ where $x, y \neq 0$, we see that there are codewords of the form $*xy$ for all x, y; moreover these codewords cannot start with 0, since 0 starts a unique codeword. But these 9 codewords now form a 3-ary IPP code: this contradicts Lemma 6.3.

Using the argument above on the second and third positions of $\mathcal{C}$, we may assume all symbols occur at least twice in every position in the code.

Choose a codeword $\mathbf{x}$. Choose a codeword $\mathbf{y} \neq \mathbf{x}$ such that $x_1 = y_1$. Choose a codeword $\mathbf{z} \neq \mathbf{y}$ such that $y_2 = z_2$. Choose a codeword $\mathbf{w} \neq \mathbf{z}$ such that $z_3 = w_3$. Since $\mathcal{C}$ has minimum distance 2, we find that $\mathbf{x} \neq \mathbf{z}$ and $\mathbf{y} \neq \mathbf{w}$. There are two cases: if $\mathbf{x} \neq \mathbf{w}$, then the pairs $\{\mathbf{x}, \mathbf{z}\}$, and $\{\mathbf{y}, \mathbf{w}\}$ violate IPP2. If $\mathbf{x} = \mathbf{w}$ then $\{\mathbf{x}, \mathbf{y}, \mathbf{z}\}$ violates IPP1. We have produced a contradiction, and so the lemma follows. □

Lemma 6.6. *There is no* 5*-ary prolific IPP code of length* 3.

Proof. If such code $\mathcal{C}$ exists then we cannot be in Case (A) in Lemma 6.2 (as $1+(q-1)^2 \geq 3q$, contradicting Theorem 4.3. So we may assume that for all positions i and symbols a there are at least 2 codewords equal to a in position i. But now the argument in the final paragraph of Lemma 6.5 shows that $\mathcal{C}$ cannot be an IPP code. This contradiction establishes the lemma. □

The above results together show:

Theorem 6.7. *There are no non-binary prolific codes of length* 3.

7. Prolific IPP codes of length 4. This section aims to prove Theorem 7.5, which states that there are no non-trivial examples of prolific codes of length 4.

Lemma 7.1. *Let $\mathcal{C}$ be a non-binary prolific IPP code of length* 4. *Then the minimum distance $d(\mathcal{C})$ of $\mathcal{C}$ is at least* 3.

Proof. Suppose $\mathcal{C}$ has M codewords. Theorem 2.1 shows that $\binom{M}{2}2^4 \geq q^4$. If $M \leq q$ we have that $q^2 2^3 \geq q^4$, which implies that $q \leq 2^{3/2} < 3$, a contradiction. So we may assume that $M > q$.

Suppose, for a contradiction, that $d(\mathcal{C}) = 1$. Without loss of generality, we may assume that $0000, 0001 \in \mathcal{C}$. If there is another codeword whose final symbol is 0 or 1, then IPP1 is violated. If this is not the case, choose distinct codewords $\mathbf{c}_1$ and $\mathbf{c}_2$ that agree in their final position. (Such codewords exist since $|\mathcal{C}| > q$.) Then $\{0000, \mathbf{c}_1\}$ and $\{0001, \mathbf{c}_2\}$ violate IPP2. This contradiction shows that $d(\mathcal{C}) \geq 2$.

We claim that every symbol must occur at least twice in any position of the code. (The prolific property shows that every symbol must occur at least once.) For assume that a symbol, 0 say, occurs exactly once as the start of a codeword. Without loss of generality, we may assume that $0000 \in \mathcal{C}$. For $x, y, z \in F \setminus \{0\}$, the descendant $0xyz$ must have 0000 and $*xyz$ as parents. So there are at least $(q-1)^3$ codewords of the form $*xyz$ where x, y and z are non-zero. None of these codewords can start with 0, and so we have a collection $\mathcal{C}'$ of $(q-1)^3$ codewords over an alphabet of size $q-1$. Since $(q-1)^3 > (q-1)^2$, we can find distinct codewords $\mathbf{c}_1, \mathbf{c}_2 \in \mathcal{C}'$ that agree in their first two positions. There are $q-2$ codewords in $\mathcal{C}'\setminus\{\mathbf{c}_2\}$ that agree with $\mathbf{c}_2$ in its last two positions: pick $\mathbf{c}_3$ of this form. Then $\mathbf{c}_1$ and $\mathbf{c}_3$ have $\mathbf{c}_2$ as their descendant. This contradiction establishes our claim.

Now suppose, for a contradiction, that we can find two distinct codewords $\mathbf{c}_1$, $\mathbf{c}_2$ that are at distance 2. Without loss of generality, we may take $\mathbf{c}_1 = 0000$, $\mathbf{c}_2 = 0011$. Let $\mathbf{c}_3 \in \mathcal{C}\setminus\{\mathbf{c}_1\}$ be of the form $**0*$. Similarly let $\mathbf{c}_4 \in \mathcal{C}\setminus\{\mathbf{c}_2\}$ be a codeword of the form $***1$. If $\mathbf{c}_3 = \mathbf{c}_4$ then the set $\{\mathbf{c}_1,\mathbf{c}_2,\mathbf{c}_3\}$ violates IPP1. But if $\mathbf{c}_3 \neq \mathbf{c}_4$ then $\{\mathbf{c}_1,\mathbf{c}_4\}$ and $\{\mathbf{c}_2,\mathbf{c}_3\}$ violates IPP2. This contradiction shows that there are no pairs of codewords at distance less than 3, and so the lemma follows. □

Lemma 7.2. *Let $\mathcal{C}$ be a q-ary prolific IPP code of length 4, and let $i \in \{1,2,3,4\}$. Then every symbol occurs in the ith position of either $q-1$ or q codewords.*

Proof. Without loss of generality, we may assume that $i = 1$.

Since $d(\mathcal{C}) \geq 3$, the first two positions of a codeword uniquely determine that codeword. So each symbol occurs at most q times as the start of a codeword (as there are q pairs starting with this symbol).

Suppose a symbol, 0 say, occurs less than q times as the start of a codeword. Then there exist symbols $x, y, z \in F$ such that there are no codewords of the form $0x**$, $0*y*$ or $0**z$. Without loss of generality, assume that $x = y = z = 0$.

Consider the descendant 0000. One parent must start with 0, but this codeword cannot contribute to any of the remaining positions in the descendant. So there is a codeword of the form $w000$ for some w (and w is clearly non-zero). Since $d(C) = 3$, we see that $w000$ is the unique codeword of the form $*00*$.

Now consider the descendant $000a$ where $a \neq 0$. One parent starts with 0 and this parent cannot contribute to the middle two positions. Hence $w000$ is the other parent. But $w000$ cannot contribute to the last position, and so there must be a codeword of the form $0**a$. Since we have $q-1$ choices for a, the symbol 0 occurs at least $q-1$ times as the start of a codeword. This proves the lemma. □

Lemma 7.3. *There is no q-ary prolific IPP code of length 4 when $q > 4$.*

Proof. Let C be a q-ary prolific IPP code of length 4. Consider the words starting with 0. By Lemma 7.2 there are at least $q-1$ such words and no two of these words can agree in any position other than the first (since $d(C) = 3$). So, without loss of generality, we may assume the codewords starting with 0 are $0111, 0222, \ldots, 0(q-1)(q-1)(q-1)$, and possibly 0000.

Fix a symbol $\ell \in F$. Choose $m \in F$ such that $\ell \neq m$ and there exists a codeword of the form $*\ell m*$. Such a choice for m exists: indeed, by Lemma 7.2 there are at least $(q-1)-1$ choices for m. Choose $n \in F$ such that $\ell \neq n$, $m \neq n$, there exists a codeword of the form $**mn$ but there does not exist a codeword of the form $*\ell mn$. There are at least $q-1$ choices for n such that there is a codeword of the form $**mn$, and at most 3 of these choices are ruled out by the other conditions we place on n. Since $q-1-3 \geq 1$, there exists at least one choice for n.

Consider the descendant $0\ell mn$. This is a descendant of $0\ell\ell\ell$ and the codeword $**mn$, and of $0nnn$ and $*\ell m*$. Our choice of ℓ, m, n means that these sets of parents are disjoint. So we get a contradiction to IPP2, as required. □

Lemma 7.4. *There is no 4-ary prolific IPP code of length 4.*

Proof. The proof of Lemma 7.3 works (with $q-1$ replaced by q at various points) when every symbol occurs 4 times in every position of a codeword. So we know that if a 4-ary prolific IPP code $\mathcal{C}$ of length 4 exists then $12 = q(q-1) \leq |C| < q^2 = 16$. We now argue that no such code can exist.

Let $M = |\mathcal{C}|$. There are $M(1+4\times 3) = 13M$ words at distance at most 1 from $\mathcal{C}$.

All the remaining words must be descendants, and indeed they must be descendants of a unique pair of codewords at distance 4. (Codewords at distance 3 can never produce descendants of distance more than 1 from the code.)

A pair of codewords at distance 4 produce exactly 6 descendants at distance 2 from the code. So $4^4 = 6M_4 + 13M$, where M_4 is the number of pairs of codewords at distance 4. Therefore

$$M_4 = \frac{256 - 13M}{6},$$

and so $256 - 13M \equiv 0 \bmod 6$ and so $M \equiv 4 \bmod 6$. But this cannot happen, as no number M such that $12 \leq M \leq 15$ is such that $M \equiv 4 \bmod 6$. This proves the theorem. □

THEOREM 7.5. *If $\mathcal{C}$ is a non-binary prolific IPP code of length* 4*, then $\mathcal{C}$ is equivalent to the* $(4, 3, 9)$*-code given in the introduction.*

Proof. Let $\mathcal{C}$ be a q-ary prolific IPP code of length 4, where $q > 2$. Lemmas 7.3 and 7.4 show that we must, in fact, have that $q = 3$.

Lemma 7.1 shows that $d(\mathcal{C}) \geq 3$ and so the Singleton bound shows that $|\mathcal{C}| \leq 9$. Moreover, we know from Lemma 7.2 that every symbol occurs at least twice in each position of the code.

We claim that $|\mathcal{C}| = 9$. Suppose, for a contradiction, that $|\mathcal{C}| \leq 8$ and so some symbol occurs twice at the start of a codeword. Without loss of generality, we may assume that 0 occurs just twice at the start of a codeword, and that $0000, 0111 \in \mathcal{C}$. Since 0222 is a descendant of the code, we must have a codeword of the form $x222$ where $x \in \{1, 2\}$.

The descendant 0012 must either have parents 0000 and $**12$, or 0111 and $*0*2$. Since $d(\mathcal{C}) \geq 3$, any word of the form $**12$ or $*0*2$ must in fact be of the form $*012$ (for otherwise there would be a codeword too close to 0000, 0111 or $x222$), and so we may deduce that $\mathcal{C}$ contains a codeword of the form $*012$. Indeed, if $a, b, c \in F$ are distinct we may argue in the same way (using the descendant $0abc$) that there exists a codeword of the form $*abc$. These 6 codewords, together with the codewords 0000, 0111 and $x222$ show that $|\mathcal{C}| \geq 9$, contradicting our assumption that $|\mathcal{C}| < 9$.

So we may assume that $|\mathcal{C}| = 9$. But this implies that $\mathcal{C}$ is an MDS code, and so is equivalent to the $(4, 3, 9)$-code from the introduction. □

8. Prolific IPP codes of length 5. This section establishes the following theorem, which follows from Lemmas 8.6, 8.13 and 8.15 below.

THEOREM 8.1. *There are no non-binary prolific IPP codes of length* 5.

LEMMA 8.2. *Let $\mathcal{C}$ be a non-binary prolific IPP code of length* 5*. Then each symbol appears at least twice in every co-ordinate.*

Proof. Every symbol occurs at least once in every position, since $\mathcal{C}$ is prolific. Suppose, for a contradiction, that a symbol (0, say) occurs exactly once in some position. Without loss of generality, we may suppose that 0 occurs once in the first position of a codeword, and that $00000 \in \mathcal{C}$.

Any descendant $0abcd$ where a, b, c, d are non-zero must have 00000 as one of its parents. But 00000 cannot contribute to any of the last four positions of this descendant, and so there is a codeword of the form $*abcd$. So the number of codewords is at least $(q - 1)^4 + 1$. But this contradicts the upper bound of Theorem 5.3, and so the lemma follows. □

LEMMA 8.3. *Let $\mathcal{C}$ be a non-binary prolific IPP code of length* 5*. Then $d(\mathcal{C}) = 3$.*

Proof. Suppose, for a contradiction, that $d(\mathcal{C}) = 1$. Without loss of generality, we may assume that $00000, 00001 \in \mathcal{C}$. By Lemma 8.2, there exists another codeword $\mathbf{c} \in \mathcal{C} \setminus \{00001\}$ that ends with a 1. Then $\{00000, 00001, \mathbf{c}\}$ violates IPP1, and so we have a contradiction.

Suppose, for a contradiction, that $d(\mathcal{C}) = 2$. Without loss of generality, we may assume that $00000, 00011 \in \mathcal{C}$. By Lemma 8.2, there exist codewords $\mathbf{c}_1 \in \mathcal{C} \setminus \{00000\}$ of the form $\mathbf{c}_1 = {*}{*}{*}0{*}$ and a codeword $\mathbf{c}_2 \in \mathcal{C} \setminus \{00001\}$ ending in a 1. If $\mathbf{c}_1 = \mathbf{c}_2$ we find that IPP1 is violated; otherwise we find that IPP2 is violated by these codewords. So again we have a contradiction.

Suppose that $d(\mathcal{C}) = 5$. We may assume, without loss of generality, that every codeword is of the form $aaaaa$ for some symbol $a \in F$. But then no word containing 3 or more distinct symbols can be a descendant of $\mathcal{C}$ and so $\mathcal{C}$ is not prolific. Thus $d(\mathcal{C}) \neq 5$.

Suppose that $d(\mathcal{C}) = 4$. The Singleton bound shows that $|\mathcal{C}| \leq q^2$. We aim to show that $\mathcal{C}$ is a $(5, q, q^2)$-code.

Let $\mathbf{c} \in \mathcal{C}$ have the form $\mathbf{c} = x_1x_2x_3x_4x_5$. Let $a, b \in F$ be such that $a \neq x_1$ and $b \neq x_2$. Consider the descendant $abx_3x_4x_5$. A parent must contribute two or more of the last 3 components, and so must be equal to $\mathbf{c}$ since $d(\mathcal{C}) = 4$. But $\mathbf{c}$ cannot contribute to the first two positions, and so there must be a codeword of the form $ab{*}{*}{*}$.

Assume, without loss of generality, that $00000 \in C$. The previous paragraph shows that there are codewords of the form $ab{*}{*}{*}$ for any non-zero $a, b \in F$. Applying the previous paragraph to the codeword of the form $11{*}{*}{*}$, and then to the codeword of the form $22{*}{*}{*}$ shows that all possible prefixes of length 2 occur in codewords. Thus $\mathcal{C}$ is a $(5, q, q^2)$-code. But then $\mathcal{C}$ cannot be a prolific IPP code, by Theorem 3.1. This contradiction shows that $d(\mathcal{C}) \neq 4$, as required. □

LEMMA 8.4. *Let $\mathcal{C}$ be a non-binary prolific IPP code of length 5. Let $\mathbf{c} \in \mathcal{C}$ agree in positions i and j with another codeword. Let $k, \ell \in \{1, 2, 3, 4, 5\} \setminus \{i, j\}$ be distinct positions. Then $\mathbf{c}$ does not agree with another codeword in positions k and ℓ.*

Proof. Without loss of generality, we may assume that $i = 1$, $j = 2$, $k = 3$ and $\ell = 4$, and that $\mathbf{c} = 00000$. So there is a codeword $\mathbf{c}_1$ that agrees with $\mathbf{c}$ in its first two positions. Since $d(\mathcal{C}) = 3$ by Lemma 8.3, we may assume without loss of generality that $\mathbf{c}_1 = 00111$. Suppose, for a contradiction, that there exists codeword $\mathbf{c}_2 \in \mathcal{C} \setminus \{\mathbf{c}\}$ of the form $\mathbf{c}_2 = {*}{*}00{*}$. If $\mathbf{c}_2$ ends with a 1, IPP1 is violated. Otherwise, by Lemma 8.2 there exists a codeword $\mathbf{c}_3 \in \mathcal{C} \setminus \{\mathbf{c}_1\}$ ending in a 1 and the sets $\{\mathbf{c}, \mathbf{c}_3\}$, $\{\mathbf{c}_1, \mathbf{c}_2\}$ violate IPP2. This contradiction establishes the lemma. □

LEMMA 8.5. *Let $\mathcal{C}$ be a non-binary prolific IPP code of length 5. Suppose that there exist codewords $\mathbf{c}_1, \mathbf{c}_2, \mathbf{c}_3, \mathbf{c}_4$ of the form $\mathbf{c}_1 = x_1x_2a{*}{*}$, $\mathbf{c}_2 = x_1x_2{*}{*}{*}$, $\mathbf{c}_3 = {*}{*}bx_4x_5$ and $\mathbf{c}_4 = {*}{*}{*}x_4x_5$. Suppose that $\mathbf{c}_1 \neq \mathbf{c}_2$ and $\mathbf{c}_3 \neq \mathbf{c}_4$. Then $a \neq b$.*

Proof. Note that the codewords $\mathbf{c}_1, \mathbf{c}_2, \mathbf{c}_3$ and $\mathbf{c}_4$ are distinct, by Lemma 8.4. If $a = b$, it is easy to see that IPP2 is violated, and so $a \neq b$. □

LEMMA 8.6. *No 3-ary prolific IPP code of length 5 exists.*

Proof. Suppose that $\mathcal{C}$ is a 3-ary prolific IPP code of length 5. Since $d(\mathcal{C}) = 3$, we may assume (without loss of generality) that there exist codewords $\mathbf{c}_1, \mathbf{c}_2 \in \mathcal{C}$ where $\mathbf{c}_1 = 00000$ and $\mathbf{c}_2 = 00111$. If there exist two codewords $\mathbf{c}_3$ and $\mathbf{c}_4$ that agree in their last two positions, we claim that we have a contradiction. To see this, firstly note that $\mathbf{c}_3$ and $\mathbf{c}_4$ cannot agree in their third position, since $d(\mathcal{C}) = 3$. Since $q = 3$, we cannot have two disjoint pairs of symbols in the 3rd position, and so (swapping over $\mathbf{c}_1$ and $\mathbf{c}_2$, or $\mathbf{c}_3$ and $\mathbf{c}_4$, if necessary) we have a contradiction to Lemma 8.5.

So our claim follows, and we may assume that no codewords agree in their last two positions. In particular, we find that $|\mathcal{C}| \leq 3^2 = 9$.

The argument above shows that no codewords agree in positions 4 and 5. The argument equally works (using the appropriate analogue of Lemma 8.5) to show that no codewords agree in any pair of positions in the set $\{3, 4, 5\}$. Because of this, a descendant of the form $ab000$ must have 00000 as a parent. When $a \neq 0$ and $b \neq 0$, this means that there is a codeword of the form $ab{*}{*}{*}$, since 00000 cannot contribute to the first two positions of the descendant. Let $\mathbf{c} \in \mathcal{C}$ have the form $11cde$. The above argument, using the descendant $abcde$, shows that codewords exist of the form $ab{*}{*}{*}$ whenever $a \neq 1$ and $b \neq 1$. Repeating the argument with a codeword of the form $22cde$, we see that all 9 words of length 2 are prefixes of codewords. Since there are two codewords of the form $00{*}{*}{*}$, this implies that $\mathcal{C}$ contains at least 10 codewords. But we have already shown that $|\mathcal{C}| \leq 9$, and so we have a contradiction, as required. □

The argument in the last paragraph of the proof above is also useful when $q > 3$, as the following two lemmas show.

Lemma 8.7. *Let $\mathcal{C}$ be a non-binary prolific IPP code of length 5. Suppose (without loss of generality) that $00000, 00111 \in \mathcal{C}$. Then there exists a pair of codewords which agree in positions i and j, where $\{i, j\} \subseteq \{3, 4, 5\}$.*

Proof. Assume, for a contradiction, that no pair of codewords agree in two of their final three positions. In particular, since the final two positions of any two codewords are distinct, we have that $|\mathcal{C}| \leq q^2$.

The argument in the final paragraph of the proof of Lemma 8.6 just uses the fact that $q \geq 3$ and that any two of the final three positions of a codeword determines that codeword uniquely. So the argument implies that there is a codeword of the form $ab{*}{*}{*}$ for any symbols $a, b \in F$. But since there are two codewords of the form $00{*}{*}{*}$, we see that $|\mathcal{C}| \geq q^2 + 1$. This contradiction establishes the lemma. □

Lemma 8.8. *Let $\mathcal{C}$ be a non-binary prolific IPP code of length 5. Suppose there are two codewords $\mathbf{c}, \mathbf{c}' \in \mathcal{C}$ that agree in positions i and j. Then for all $a, b \in F$ with $a \neq c_i$, $b \neq c_j$ there exists a codeword whose ith and jth positions are equal to a and b respectively.*

Proof. Without loss of generality, assume that $i = 1$, $j = 2$, $\mathbf{c} = 00000$ and $\mathbf{c}' = 00111$. Let $a, b \in F$ be non-zero: we must show there exists a codeword of the form $ab{*}{*}{*}$.

By Lemma 8.4, no non-zero codeword has two or more zeroes in its final three positions. But then the descendant $ab000$ must have 00000 as a parent, and so the other parent is a codeword of the form $ab{*}{*}{*}$. □

Lemma 8.9. *Let $\mathcal{C}$ be a q-ary prolific IPP code of length 5, where $q \neq 2$. No set of $q - 1$ codewords can pairwise agree in any fixed set of two positions.*

Proof. Suppose a set of $q - 1$ codewords agree in a pair of positions, say (without loss of generality) their first two positions. The third positions of these codewords must all be distinct (by Lemma 8.3). Thus if there exist two codewords that agree in their final two positions, we obtain a contradiction by Lemma 8.5. A similar argument shows that no two codewords can agree in their ith and jth positions, where $\{i, j\} \subseteq \{3, 4, 5\}$. But this contradicts Lemma 8.7. □

Lemma 8.10. *Let $\mathcal{C}$ be a q-ary prolific IPP code of length 5, where $q \neq 2$. Then every symbol occurs at least q times in each co-ordinate of the code.*

Proof. Assume, for a contradiction, that there is a symbol that occurs fewer than q times in some position. Without loss of generality, we may assume that at most

$q-1$ codewords are of the form $0{*}{*}{*}{*}$.

Let $a, b, c, d \in F$ be symbols such that there are no codewords of the form $0a{*}{*}{*}$, $0{*}b{*}{*}$, $0{*}{*}c{*}$ or $0{*}{*}{*}d$. By considering the descendant $0abcd$ we see that there is a codeword $\mathbf{c}$ of the form $\mathbf{c} = {*}abcd$. Let $z \in F \setminus \{d\}$. The word $0abcz$ has parents $\mathbf{c}_1$ and $\mathbf{c}_2$ where $\mathbf{c}_1 = 0{*}{*}{*}{*}$ and $\mathbf{c}_2 = {*}abc{*}$. But since $d(\mathcal{C}) = 3$ we find that $\mathbf{c}_2 = \mathbf{c}$ and so $\mathbf{c}_1$ must contribute to the final component of the descendant $0abcz$. Thus $\mathbf{c}_1 = 0{*}{*}{*}z$. Since there are at most $q-1$ codewords starting with 0, and since there are $q-1$ choices for z, we find that there are exactly $q-1$ codewords starting with 0, and all their last co-ordinates differ.

By Lemma 8.9, there are at most $q-2$ codewords of the form ${*}ab{*}{*}$, and so we may find distinct symbols $u, v \in F$ such that there are no codewords of the form ${*}abu{*}$ or ${*}abv{*}$. Let $w \in F$ be such that there are no codewords of the form ${*}ab{*}w$. Considering the parents of $0abuw$ and $0abvw$, we see that there must be codewords of the form $0{*}{*}uw$ and $0{*}{*}vw$. But these are distinct codewords starting with 0, and their last co-ordinates are equal. This contradicts the previous paragraph, and so the lemma follows. □

Lemma 8.11. *Let $\mathcal{C}$ be a non-binary prolific IPP code of length 5. Assume that $00000, 00111 \in \mathcal{C}$ and that there are two codewords that agree in their last two positions. Then the non-zero codewords of the form ${*}{*}0{*}{*}$ must be of the form ${*}{*}01a$ or ${*}{*}0a1$ for some $a \in F \setminus \{0,1\}$.*

Proof. Note that a non-zero codeword of the form ${*}{*}0xy$ cannot have $x, y \in \{0,1\}$, for otherwise IPP1 is violated. Moreover, a non-zero codeword of the form ${*}{*}00a$ or ${*}{*}0a0$ would contradict Lemma 8.4. So to prove the lemma, it suffices to show that there are no codewords of the form ${*}{*}0xy$ where $x \notin \{0,1\}$ and $y \notin \{0,1\}$.

Suppose, for a contradiction, that there exists a codeword $\mathbf{c} \in \mathcal{C}$ of the form ${*}{*}0xy$, where $x \notin \{0,1\}$ and $y \notin \{0,1\}$. Without loss of generality, we may assume that $x = y = 2$. By Lemma 8.5, $\mathbf{c}$ is the unique codeword of the form ${*}{*}{*}22$. By Lemma 8.4, the only codeword of the form ${*}{*}{*}11$ is the codeword 00111. Since we are assuming that there exist two codewords that agree in their final two positions, Lemma 8.8 implies that there exists $\mathbf{c}' \in \mathcal{C}$ of the form $\mathbf{c}' = {*}{*}{*}12$ or $\mathbf{c}' = {*}{*}{*}21$. In either case, the sets $\{00111, \mathbf{c}\}$ and $\{00000, \mathbf{c}'\}$ violate IPP2. This contradiction establishes the lemma, as required. □

Lemma 8.12. *Let $\mathcal{C}$ be a non-binary prolific IPP code of length 5. Let $i, j \in \{1, 2, \ldots, 5\}$ be a pair of positions. Then there cannot exist a set of three codewords that pairwise agree in position i and position j.*

Proof. Suppose, for a contradiction, that such a set of three codewords exists. Without loss of generality, we may assume $i = 1$ and $j = 2$ and the positions where the three codewords agree are both 0. Since $d(\mathcal{C}) = 3$, we may assume that $00000, 00111, 00222 \in \mathcal{C}$. By Lemma 8.7, we may assume (without loss of generality) that there is a pair of codewords that agrees in their final two positions.

By Lemma 8.11, the non-zero codewords of the form ${*}{*}0{*}{*}$ must have the form ${*}{*}01a$ or ${*}{*}0a1$ for some $a \in F \setminus \{0,1\}$. But replacing the symbol 1 by the symbol 2 throughout Lemma 8.11, since $00222 \in \mathcal{C}$ we deduce that these codewords must have the form ${*}{*}02b$ or ${*}{*}0b2$ where $b \in F \setminus \{0,2\}$. So a codeword of the form ${*}{*}0{*}{*}$ must have one of the forms 00000, ${*}{*}012$ or ${*}{*}021$. Since $d(\mathcal{C}) = 3$, there is at most one codeword of each of the forms ${*}{*}012$ and ${*}{*}021$, and so 0 occurs at most 3 times in the third position of a codeword. By Lemma 8.6 we have that $q > 3$, and so we have a contradiction by Lemma 8.10. □

Lemma 8.13. *No q-ary length 5 prolific IPP code exists when $q \geq 5$.*

Proof. Lemmas 8.11 and 8.12 imply that there are at most 5 codewords which are zero in their third position. Indeed, without loss of generality, we may assume these codewords have one of the forms 00000, $**012$, $**013$, $**021$ and $**031$.

In order to obtain descendants of the form $**044$, we must have a codeword of the form $***44$. By Lemma 8.12, there are at most two codewords of the form $***44$, and so there are at least $(q-2)^2$ choices for symbols $a, b \in F$ such that there are no codewords of the form $a**44$ or $*b*44$. Since $q \geq 5$, we have that $(q-2)^2 > 5$ and so we may in addition choose a and b to have the property that there are no codewords of the form $ab0**$. But then $ab044$ is not a descendant of $\mathcal{C}$, and so we have a contradiction as required. □

Lemma 8.14. *A* 4*-ary prolific IPP code* $\mathcal{C}$ *of length* 5 *has exactly* 16 *codewords.*

Proof. By Lemma 8.10, each symbol must occur at least 4 times at the start of a codeword, and so $|\mathcal{C}| \geq 16$. So it suffices to show that $|\mathcal{C}| \leq 16$.

If no pair of codewords agree on some fixed pair of positions $\{i, j\}$, we have that $|\mathcal{C}| \leq 16$ and the lemma follows trivially. So we may assume that for all positions i and j we may find distinct codewords $\mathbf{c}, \mathbf{c}' \in \mathcal{C}$ such that $\mathbf{c}_i = \mathbf{c}'_i$ and $\mathbf{c}_j = \mathbf{c}'_j$.

Without loss of generality, assume that $00000, 00111 \in \mathcal{C}$.

Suppose another pair of symbols is the prefix of distinct codewords, so $aby_3y_4y_5 \in \mathcal{C}$ and $abz_3z_4z_5 \in \mathcal{C}$ where a and b are not both zero. If $y_3, y_4, y_5 \in \{0, 1\}$ then $\{00000, 00111, aby_3y_4y_5\}$ violates IPP1. So, without loss of generality, we may assume that $y_3 = 2$. Note that $y_3 \neq z_3$, since $d(\mathcal{C}) = 3$. Let $\mathbf{c}_3$ and $\mathbf{c}_4$ agree in their final two positions. Since $q = 4$, we find that Lemma 8.5 is violated, in the case when $\mathbf{c}_1 = 00000$, $\mathbf{c}_2 = 00111$ or in the case when $\mathbf{c}_1 = aby_3y_4y_5$ and $\mathbf{c}_2 = abz_3z_4z_5$. This contradiction shows that at most one pair of symbols is the prefix of two or more codewords. Indeed, more generally, we find that for any pair $\{i, j\}$ of positions at most one pair of symbols occurs twice as the ith and jth positions of a codeword. Since, by Lemma 8.12, no pair of symbols occurs more than twice in any two positions of the code, we see that $|\mathcal{C}| \leq 17$, and if $|\mathcal{C}| = 17$ then every pair of symbols occurs in any two positions of the code, with exactly one pair occurring twice.

Suppose that $|\mathcal{C}| = 17$. Then the previous paragraph shows that there are codewords of the form $**0*2$ and $**0*3$. By Lemma 8.11, these codewords must be of the form $**012$ and $**013$. Similarly, there exist codewords of the form $***02$ and $***03$, and Lemma 8.11 implies that these codewords are actually of the form $**102$ and $**103$. But then both 01 and 10 repeat in the third and forth positions of codewords. This contradicts the previous paragraph. Hence $|\mathcal{C}| \leq 16$, as required. □

Lemma 8.15. *No* 4*-ary prolific IPP code of length* 5 *exists.*

Proof. Let $\mathcal{C}$ be a 4-ary prolific IPP code of length 5. By Lemma 8.14, we have that $|\mathcal{C}| = 16$ and so, by Lemma 8.10, every symbol occurs exactly 4 times in any position.

Since $d(\mathcal{C}) = 3$, we may assume that $00000, 00111 \in \mathcal{C}$. Since 0 occurs exactly 4 times in the third position of the code, Lemma 8.11 implies that (without loss of generality) there exist codewords of the form $**012$, $**021$ and $**013$. In particular, two codewords agree in their third and fourth positions. So we may apply Lemma 8.11 to deduce that the non-zero codewords ending in 0 are, without loss of generality, of the form $**210$, $**120$ and $**130$. (Note that we cannot have both a codeword of the form $**210$ and a codeword of the form $**310$, for then 1 would occur too many times in the fourth position of the code.) Arguing similarly, we find that (without loss of generality) the non-zero codewords with 0 in their fourth position are of the form $**102$, $**201$ and $**301$. So to summarise, we may assume that $\mathcal{C}$ consists of 16

$$\begin{bmatrix}
0 & 0 & 0 & 0 & 0 \\
0 & 0 & 1 & 1 & 1 \\
* & * & 0 & 1 & 2 \\
* & * & 0 & 2 & 1 \\
* & * & 0 & 1 & 3 \\
* & * & 2 & 1 & 0 \\
* & * & 1 & 2 & 0 \\
* & * & 1 & 3 & 0 \\
* & * & 1 & 0 & 2 \\
* & * & 2 & 0 & 1 \\
* & * & 3 & 0 & 1 \\
* & * & * & * & * \\
* & * & * & * & * \\
* & * & * & * & * \\
* & * & * & * & * \\
* & * & * & * & *
\end{bmatrix}$$

TABLE 8.1
Structure of the 4*-ary code*

codewords of the forms given in Table 8.1, where the last three positions of the final 5 codewords do not involve 0 or 1.

Since there are two codewords of the form $00{*}{*}{*}$, and 0 occurs exactly 4 times in any position, there are symbols x and y such that there are no codewords of the form $0x{*}{*}{*}$ or $y0{*}{*}{*}$. Without loss of generality, we may assume $x = y = 3$, and so there are no codewords of the form $30{*}{*}{*}$ or $03{*}{*}{*}$. There are at most two codewords of the form $33{*}{*}{*}$ by Lemma 8.12. Let $\mathbf{c}$ be a codeword of one of the forms ${*}{*}021$, ${*}{*}210$, ${*}{*}102$ which is not of the form $33{*}{*}{*}$. Lemma 8.12, together with the fact that $00000, 00111 \in \mathcal{C}$, implies that $\mathbf{c}$ is not of the form $00{*}{*}{*}$. Write $\mathbf{c} = c_1c_2c_3c_4c_5$. Note that there is no codeword $\mathbf{c}' \in \mathcal{C} \setminus \{\mathbf{c}\}$ that agrees with $\mathbf{c}$ in two of its final three positions, and so any descendant of the form ${*}{*}c_3c_4c_5$ must have $\mathbf{c}$ as a parent. Since $\mathbf{c}$ is not of the form $33{*}{*}{*}$ or $00{*}{*}{*}$, one of the words $03c_3c_4c_5$ and $30c_3c_4c_5$ cannot be a descendant and so we have a contradiction as required. □

9. Conclusion and open problems. As we stated in the introduction, we conjecture that there are no examples of prolific IPP codes other than those listed in the introduction. We begin this section by discussing how close we are to proving this conjecture.

When $n \geq 3$ but $n \neq 4$, the lower bound of Theorem 2.2 and the upper bound of due to Hollmann *et al.* (Theorem 4.3) together imply that for any fixed length n, there are no q-ary prolific IPP codes of length n provided that q is sufficiently large. When $n = 4$, the same is true if an improved, but less explicit, bound due to Alon, Fischer and Szegedy [3] is used; or we may deduce this from Theorem 7.5. Sadly, we are not able to bring the number of open parameters n and q (where it is not known whether a prolific q-ary IPP code of length n exists) down to a finite number. In particular, when q is fixed and $3 \leq q \leq 8$, all the parameters n and q are open for $n \geq 6$. Table 9.1 lists the parameters for which the existence of a prolific (n, q, M) IPP code is as yet undetermined. For the values of n listed in the table, a

n	q_{max}	n	q_{max}	n	q_{max}
6	11	7	34	8	13
9	10	10	18	11	12
12	10	13	14	14	11
15	10	16	13	17	10
18	9	19	12	20	10
21	9	22	11	23	10
24	9	25	10	26	9
27	9	28	10	29	9
30	9	31	10	32	9
33	9	34	10	35	9
36	9	37	9	38	9
39	8	40	9	41	9
42	8	43	9	44	9
45	8	46	9	47	9
48	8	49	9	50	8
51	8	52	9	53	8
54	8	55	9	56	8
57	8	58	9	59	8
60	8	61	9	62	8
63	8	64	9	65	8
66	8	67	8	68	8

Table 9.1
Open parameters for prolific IPP codes

non-binary prolific (n, q, M)-IPP code might exist for $3 \leq q \leq q_{max}$, and for $n > 68$, a non-binary prolific (n, q, M)-IPP code might exist for $3 \leq q \leq 8$. We used the lower bound of Theorem 2.2 with $k = \lceil \frac{n}{3} \rceil$. We used the upper bound of Theorem 5.4 when $n \equiv 2 \bmod 3$, and the upper bound of Theorem 4.3 otherwise.

As a step towards proving the conjecture, is it possible to show that no q-ary prolific IPP code of length n exists for all sufficiently large n, where q is a fixed integer with $3 \leq q \leq 8$?

We remark that the notion of a prolific code makes sense for any class of code which has a natural notion of a descendant. In particular, are there non-trivial examples of prolific k-IPP codes where $k > 2$? (See Staddon, Stinson and Wei [12] or Blackburn [7] for the definition of a k-IPP code).

Is it possible to improve our upper bounds on IPP codes (Theorems 5.2 and 5.4) significantly? We believe that the constants in the leading terms of these upper bounds can be reduced. Indeed, it might be possible to prove more than this. Let n be fixed, and suppose that 3 does not divide n. Let ϵ be a positive constant. Is it the case that a q-ary IPP code $\mathcal{C}$ of length n must satisfy $|\mathcal{C}| \leq \epsilon q^{\lceil n/3 \rceil}$ when q is sufficiently large? This is true when $n = 4$, by a bound of Alon, Fischer and Szegedy [3, Theorem 2.5].

Is it possible to generalise the techniques of Theorems 5.2 and 5.4, to provide better upper bounds for k-IPP codes when $k > 2$? We have established new bounds on 3-IPP codes using these techniques. We hope that these bounds will form the subject of a future paper; but we also hope that our techniques can be stretched further.

Acknowledgments. The authors would like to thank Noga Alon for some helpful remarks.

REFERENCES

[1] R. Ahlswede and L. H. Khachatrian, *The complete intersection theorem for systems of finite sets*, European Journal Combinatorics, 18 (1997), pp. 125–136.

[2] N. Alon, G. Cohen, M. Krivelevich, and S. Litsyn, *Generalised hashing and parent identifying codes*, J. Combinatorial Theory, Series A, 104 (2003), pp. 207–115.

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

[4] N. Alon and U. Stav, *New bounds on parent-identifying codes: the case of multiple parents*, Combinatorics, Probability and Computing, 13 (2004), pp. 795–807.

[5] 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 Journal on Discrete Mathematics, 14 (2001), pp. 423–431.

[6] A. Barg and G. Kabatiansky, *A class of IPP codes with efficient identification*, J. Complexity, 20 (2004), pp. 137–147.

[7] S. R. Blackburn, *An upper bound on the size of a code with the k-identifiable parent property*, Journal of Combinatorial Theory, Series A, 102 (2003), pp. 179–185.

[8] R. Hill, *A First Course in Coding Theory* Oxford University Press, Oxford, 1986.

[9] 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, Series A, 82 (1998), pp. 121–133.

[10] J. Löfvenberg, *Binary fingerprinting codes*, Designs, Codes and Cryptography, 36 (2005), pp. 69-81.

[11] T. Lindkvist, J. Löfvenberg and M. Svanström, *A class of traceability codes*, IEEE Trans. Information Theory, 48 (2002), pp. 2094–2096.

[12] J. N. Staddon, D. R. Stinson, and R. Wei, *Combinatorial properties of frameproof and traceability codes*, IEEE Transactions on Information Theory, 47 (2001), pp. 1042–1049.

[13] V. D. Tô and R. Safavi-Naini, *On the maximal codes of length 3 with the 2-identifiable parent property*, SIAM J. Discrete Mathematics, 17 (2004), pp. 548–570.

[14] T. van Trung and S. Martirosyan, *New constructions for IPP codes*, Designs, Codes and Cryptography, 35 (2005), pp. 227–239.

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