On Single-Deletion-Correcting Codes

N. J. A. Sloane

Information Sciences Research, AT&T Shannon Labs
180 Park Avenue, Florham Park, NJ 07932-0971, U.S.A.

Dedicated to Dijen Ray-Chaudhuri on the occasion of his 65th birthday.

Jan 10, 2000. Revised Jan 20, 2000, Nov 07 2002.

An earlier version of this paper appeared in *Codes and Designs, Ohio State University, May 2000 (Ray-Chaudhuri Festschrift)*, ed. K. T. Arasu and A. Seress, Walter de Gruyter, Berlin, 2002, pp. 273–291.

Abstract

This paper gives a brief survey of binary single-deletion-correcting codes. The Varshamov-Tenengolts codes *appear* to be optimal, but many interesting unsolved problems remain. The connections with shift-register sequences also remain somewhat mysterious.

1 Introduction

The possibility of packet loss on internet transmissions has renewed interest in deletion-correcting codes. The problem is this: you send 12 symbols, but only 11 arrive. One symbol is deleted, but neither you nor the recipient know which of the twelve was lost. We wish to design a code that can correct such errors.

Of course there are many other applications of such codes, including magnetic recording, although in that case there are usually additional conditions that must be satisfied. This paper considers the very simplest family of such codes, binary block codes capable of correcting single deletions. Even for these codes there remain several apparently unsolved problems.

It is surprising, but such codes do not appear to be surveyed in any of the usual references ([MS77], [PH98], etc.). This paper is a first attempt at such a survey.

Proofs are given of a number of results, either because the new proofs are simpler or because the original sources are hard to locate[1].

Definition 1.1. *For a vector $u \in \mathbb{F}_q^n$, let $D_e(u)$ denote the set of e-th order descendants, i.e. the set of vectors $v \in \mathbb{F}_q^{n-e}$ that are obtained if any e components are deleted from u. A subset $C \subseteq \mathbb{F}_q^n$ is said to be an* e-deletion-correcting code *if $D_e(u) \cap D_e(v) = \emptyset$ for all $u, v \in C$, $u \neq v$.*

Our problem is to find the largest such code. In this paper we mostly consider the simplest case, $q = 2$ and $e = 1$. Even here the problem is unsolved.

The *deletion distance* $dd(u, v)$ between vectors $u, v \in \mathbb{F}_q^n$ is defined to be one-half of the smallest number of deletions and insertions needed to change u to v. Then C is e-deletion-correcting if and only if $dd(u, v) \geq e + 1$ for $u, v \in C$, $u \neq v$. (For $dd(u, v) \leq e$ if and only if there is a vector x that can be reached from u by at most e deletions and also from v by at most e deletions, and then C cannot correct e deletions.)

Consider the graph G_n having a node for every vector $u \in \mathbb{F}_q^n$, with an edge joining the nodes corresponding to $u, v \in \mathbb{F}_q^n$, $u \neq v$, if and only if v can be obtained from u by a single deletion and insertion, i.e. if and only if $D_1(u) \cap D_1(v) \neq \emptyset$. The deletion distance $dd(u, v)$ is the length of the shortest path from u to v (this shows that dd is indeed a metric).

In particular, a single-deletion-correcting code corresponds to an independent set in G_n. One can now attempt to calculate the sizes of the largest independent sets by computer. In the binary case we find that the largest single-deletion-correction codes of lengths $1, 2, \ldots, 8$ have sizes

$$1, 2, 2, 4, 6, 10, 16, \geq 30 \ . \tag{1}$$

The last entry in (1) was kindly computed by my colleague David Johnson. Unfortunately G_8 is too large for present computers and 30 is at present only a lower bound on the size of a maximal independent set.[2]

However, (1) turns out to be a useful hint. When one looks up this sequence in [EIS], one finds a unique matching sequence, number A16, whose initial terms N_1, N_2, $N_3, \ldots$ are

$$1, 1, 2, 2, 4, 6, 10, 16, 30, 52, 94, 172, 316, 586, \ldots \tag{2}$$

[1] And when located are sometimes poorly translated or badly photocopied!

[2] Postscript: David Applegate has since used CPLEX's integer programming subroutines (which combine ordinary linear programming with branch-and-bound) to confirm that the largest single-deletion-correcting code of length 8 does indeed have size 30.

and whose nth term is given by

$$N_n = \frac{1}{2n} \sum_{\text{odd } d|n} \phi(d) 2^{n/d}, \ n \geq 1\,, \tag{3}$$

where the sum is over all odd divisors d of n and ϕ is the Euler totient function (sequence A10). The references cited for sequence A16 indicate that it has arisen in connection with the enumeration of shift-register sequences [Go67] and tournaments [Br80]. However there was (at that time) no reference to indicate that this sequence has any connection with codes, nor was there any apparent connection between the shift-register sequences and deletion-correction codes.

More conventional search methods, in particular, consulting some well-known papers of Levenshtein [Lev65], [Lev65a] on codes for correcting deletions, turned up many other relevant references. Some of these will be discussed further in Section 6. The most interesting codes are those of Varshamov and Tenengolts [VT65]. In [VT65] they present a family of codes depending on a certain parameter a. When a is taken to be 0, these codes have size N_{n-1} (see (3)) and thus match (1). These codes are the subject of Section 2.

Sections 3 and 4 will discuss the connection with shift-registers and tournaments, and Section 5 contains some general remarks about the number of descendants of a vector. The final section, Section 6, gives a brief discussion of other papers on deletion-correcting and related codes.

2 The Varshamov-Tenengolts codes

Definition 2.1. *For $0 \leq a \leq n$, the* Varshamov-Tenengolts code $VT_a(n)$ *consists of all binary vectors $(x_1, \ldots, x_n)$ satisfying*

$$\sum_{i=1}^{n} i x_i \equiv a \pmod{n+1}\,, \tag{4}$$

where the sum is evaluated as an ordinary rational integer.

As will appear, the codes with $a = 0$ contain the most codewords. The

first few such codes are

$$\begin{aligned} VT_0(1) &= \{0\} \\ VT_0(2) &= \{00, 11\} \\ VT_0(3) &= \{000, 101\} \\ VT_0(4) &= \{0000, 1001, 0110, 1111\} \\ VT_0(5) &= \{00000, 10001, 01010, 110011, 11100, 00111\}, \end{aligned} \tag{5}$$

of sizes 1,2,2,4,6, matching (1) and (2). These codes were introduced in [VT65] for correcting errors on a Z-channel (or asymmetric channel). Similar constructions have been used in [BR82] and also in [GS80] and [Kl81] to construct constant weight codes.

Levenshtein [Lev65], [Lev65a] observed that the Varshamov-Tenengolts codes could be used for correcting single deletions, proving this by giving the following elegant decoding algorithm.

Decoding algorithm

- Suppose a codeword $x = (x_1, \ldots, x_n) \in VT_a(n)$ is transmitted, the symbol s in position p is deleted, and $x' = (x'_1, \ldots, x'_{n-1})$ is received. Let there be L_0 0's and L_1 1's to the left of s, and R_0 0's and R_1 1's to the right of s (with $p = 1 + L_0 + L_1$).

- We compute the weight $w = L_1 + R_1$ of x' and the new checksum $\sum_{i=1}^{n-1} ix'_i$. If $s = 0$ the new checksum is R_1 ($\leq w$) less than it was before, and if $s = 1$ it is $p + R_1 = 1 + L_0 + L_1 + R_1 = 1 + w + L_0$ ($> w$) less than it was before. (These numbers are less than $n + 1$ so there is no ambiguity.)

- So if the deficiency in the checksum is less than or equal to w we know that a 0 was deleted, and we restore it just to the left of the rightmost R_1 1's. Otherwise a 1 was deleted and we restore it just to the right of the leftmost L_0 0's.

The sizes $|VT_a(n)|$ of the first few codes are shown in Table 1. (This array forms sequence A53633 in [EIS].) These numbers were studied by Varshamov [Var65] and Ginzburg [Gi67], but the following simple formula appears to be new[3].

[3] It appears that Theorem 2.2 had already been discovered by Martirosyan [Mt96].

Table 1: Number of codewords in Varshamov-Tenengolts code $VT_a(n)$.

$n \setminus a$	0	1	2	3	4	5	6	7	8
1	1	1							
2	2	1	1						
3	2	2	2	2					
4	4	3	3	3	3				
5	6	5	5	6	5	5			
6	10	9	9	9	9	9	9		
7	16	16	16	16	16	16	16	16	
8	30	28	28	29	28	28	29	28	28

Theorem 2.2.

$$|VT_a(n)| = \frac{1}{2(n+1)} \sum_{\substack{d|n+1 \\ d \text{ odd}}} \phi(d) \frac{\mu\left(\frac{d}{(d,a)}\right)}{\phi\left(\frac{d}{(d,a)}\right)} 2^{(n+1)/d} , \tag{6}$$

where $\mu(n)$ is the Möbius function (A8683), and $(d,a) = gcd(d,a)$.

Proof. Write $w_a(n) = |VT_a(n)|$. We will calculate $w_a(n-1)$, assuming throughout that $n \geq 1$. It follows from the definition of these codes that the generating function

$$f(z) = \sum_{a=0}^{n-1} w_a(n-1) z^a$$

is equal to

$$\prod_{k=1}^{n-1} (1 + z^k) \mod z^n - 1 .$$

Let $\xi = e^{2\pi i/n}$. Then

$$f(\xi^j) = \sum_{a=0}^{n-1} w_a(n-1)\xi^{ja} = \prod_{k=1}^{n-1} (1 + \xi^{jk}), \ j = 0, \ldots, n-1 .$$

We solve this by taking an inverse discrete Fourier transform (cf. [Ko88], Chap. 97) to obtain

$$w_a(n-1) = \frac{1}{n} \sum_{j=0}^{n-1} f(\xi^j)\xi^{-ja} .$$

Since

$$\prod_{k=0}^{n-1}(z-\xi^k) = z^n - 1\ ,$$

we can calculate $f(\xi^j)$ explicitly. An elementary calculation gives

$$f(\xi^j) = \begin{cases} 2^{g-1} & \text{if} \quad d = n/g \text{ is odd,} \\ 0 & \text{if} \quad d = n/g \text{ is even,} \end{cases}$$

where $g = gcd(n, j)$. Therefore

$$w_a(n-1) = \frac{1}{2n} \sum_{\substack{d|n \\ d \text{ odd}}} 2^{n/d} \sum_{\substack{j=1 \\ gcd(n,j)=n/d}}^{n} \xi^{-ja}$$

which becomes, writing $j = kn/d$,

$$= \frac{1}{2n} \sum_{\substack{d|n \\ d \text{ odd}}} 2^{n/d} \sum_{\substack{k=1 \\ (k,d)=1}}^{d} e^{-2\pi ika/d}\,.$$

The innermost sum is a Ramanujan sum $c_d(a)$ ([Ap76], p. 160), which simplifies to

$$c_d(a) = \phi(d)\frac{\mu\left(\frac{d}{(d,a)}\right)}{\phi\left(\frac{d}{(d,a)}\right)}$$

([Ap76], p. 164). □

Corollary 2.3.

$$\textit{(i)}\ |VT_0(n)| = \frac{1}{2(n+1)} \sum_{\substack{d|n+1 \\ d \text{ odd}}} \phi(d) 2^{(n+1)/d}\ , \tag{7}$$

$$\textit{(ii)}\ |VT_1(n)| = \frac{1}{2(n+1)} \sum_{\substack{d|n+1 \\ d \text{ odd}}} \mu(d) 2^{(n+1)/d}\ , \tag{8}$$

(iii) For any a,

$$|VT_0(n)| \geq |VT_a(n)| \geq |VT_1(a)|\,. \tag{9}$$

Remark 2.4. (i) and the left-hand inequality in (iii) are due to Varshamov [Var65], and (ii) and the right-hand inequality in (iii) to Ginzburg [Gi67].

Proof. (i) and (ii) follow immediately from Theorem 2.2, as does the left-hand side of (iii) using $\mu(k) \leq \phi(k)$ for all k. To establish the right-hand side of (iii), let p be the smallest odd prime dividing both $n+1$ and a (if no such prime exists then $|VT_a(n)| = |VT_1(n)|$). The terms in the expressions for $|VT_a(n)|$ and $|VT_1(n)|$ agree for $d < p$, and at $d = p$ the term in $|VT_a(n)|$ exceeds that in $|VT_1(n)|$ by $p2^{n/p}$. It is easy to check that the remaining terms can never make the sum in $|VT_1(n)|$ catch up with the sum in $|VT_a(n)|$. □

Optimality

It is more difficult to obtain upper bounds for deletion-correcting codes than for conventional error-correcting codes, since the disjoint balls $D_e(u)$ associated with the codewords (see Section 1) do not all have the same size. Furthermore the metric space $(\mathbb{F}_2^n, dd)$ is not an association scheme and so there is no obvious linear programming bound.

The size of $D_1(u)$ is easily seen to be equal to $r(u)$, the number of runs in u. Furthermore the number of vectors in $\mathbb{F}_2^n$ with r runs is $2\binom{n-1}{r-1}$. (We will discuss $|D_e(u)|$ further in Section 5.)

Let $A(n, e)$ denote the size of the largest e-deletion-correcting binary code of length n, and call a code C *optimal* if $|C| = A(n, e)$. The values of $A(n, 1)$ for $n \leq 9$ were given in Section 1, and show that $VT_0(n)$ is optimal for $n \leq 9$.

For large n, the codes $VT_0(n)$ are certainly close to being optimal, since on the one hand we have

$$|VT_0(n)| \geq \frac{2^n}{n+1}\ , \tag{10}$$

from (9), and on the other hand we have the following result of Levenshtein [Lev65]:

Theorem 2.5 ([Lev65]).

$$A(n, 1) \sim \frac{2^n}{n}, \quad \textit{as} \quad n \to \infty\,.$$

Proof. (10) gives a lower bound. Let C be an optimal code. Following Levenshtein, let C_0 denote the subset of C consisting of the vectors $u \in C$ with

$$\frac{n}{2} - \sqrt{n \log n} \leq r(u) \leq \frac{n}{2} + \sqrt{n \log n}$$

and let $C_1 = C \setminus C_0$. Since the sets $D_1(u)$, $u \in C$, must be disjoint,

$$|C_0| \leq \frac{2^{n-1}}{\frac{n}{2} - \sqrt{n \log n}} \lesssim \frac{2^n}{n} \ .$$

Furthermore,

$$|C_1| \leq 2 \sum_{r=1}^{\frac{n}{2}-\sqrt{n \log n}} 2\binom{n-1}{r-1} \ ,$$

which is much smaller than $2^n/n$. □

In a later paper, Levenshtein [Lev92] defines a code C to be *perfect* if the balls $D_e(u)$, $u \in C$, partition the set $\mathbb{F}_2^{n-e}$. In [Lev92] he proves the remarkable fact that *all* the codes $VT_0(n)$, $VT_1(n)$, $VT_2(n), \ldots$ are perfect single-deletion-correcting codes. The argument, not reproduced here, is essentially just a refinement of the decoding algorithm for these codes given above.

It is initially surprising that perfect codes of the same length can have different numbers of codewords, but this is explained by the fact that the balls $D_1(u)$ have different sizes.

In view of this and the result in (9), it is tempting to make the following conjecture.

Conjecture 2.6. *The codes $VT_0(n)$ are optimal for all n.*

This is true for $n \leq 8$, as already mentioned, but for larger n it is possible that other, smaller, perfect codes may exist, or even that smaller, optimal but non-perfect codes may exist.

Indeed, consider the code $\{000, 111\}$. For this code, $\sum_{u \in C} |D_1(u)| = 1 + 1 = 2 < 4$, so this is optimal but not perfect. For length 4, $\{0000, 0011, 1100, 1111\}$ contains as many codewords as $VT_0(4)$ (compare (5)), and again is optimal but not perfect.

At length 6 it is possible to replace two codewords of $VT_0(6)$ by two other vectors without affecting its ability to correct single deletions: 110100 and 001011 can be replaced by 111000 and 000111. The former pair cover eight vectors of length 5, but the latter only cover four vectors of length 5, leaving four vectors uncovered. This suggests the possibility that in some larger code $VT_0(n)$ it may be possible to replace k vectors by $k+1$ vectors, which would prove that these codes are not optimal.

In view of these remarks, Conjecture 2.6 does not seem especially compelling!

Linearity

As can be seen from (5), the codes $VT_0(n)$ are linear for $n \leq 4$. They are never again linear, since, for $n \geq 5$, $VT_0(n)$ contains the vectors $1\,0\,0\,0\ldots0\,0\,1$ and $1\,1\,0\,0\ldots1\,0\,0$ but not their sum.

In particular, even though $|VT_0(7)| = 16$, this code is not linear. One might wonder if it is possible to find a linear code that will do as well, but a computer search has shown that no such code exists.

On the other hand, by adapting a construction of Tenengolts [Ten76], one can modify the Varshamov-Tenengolts construction to obtain linear codes, with only a small increase in the length of the code.

Definition 2.7. *Given $k \geq 1$, let*

$$n = k + \left\lceil \sqrt{2k + 9/4} + 1/2 \right\rceil \ .$$

The linear single-deletion-correcting code $VT_0'(n)$ has dimension k and consists of all vectors $(x_1, \ldots, x_n) \in \mathbb{F}_2^n$, where $x_1, \ldots, x_k$ are information symbols and the $c = n - k$ check symbols $x_{k+1}, \ldots, x_n$ are chosen so that $\sum_{i=1}^{n} i x_i \equiv 0 \pmod{n+1}$.

The construction works because c is just large enough so that $\binom{c+1}{2} \geq n + 1$, and so the sums $\sum_{i=k+1}^{n} i x_i$ cover $n + 1$ consecutive values modulo $n + 1$. We omit the details.

The number of check symbols in these codes is of the order of $\sqrt{2n}$, compared with $O(\log n)$ for the $VT_0(n)$ codes. So we end this section with a final question: What are the optimal linear single-deletion-correcting codes?

3 Shift register sequences

As mentioned in Section 1, the entry for sequence A16 in [EIS] indicates that these numbers also arise in the enumeration of shift register sequences [Go67]. We will show here that indeed this is the same sequence. But whether this is anything more than a coincidence remains an open question. Of course there are well-known connections between shift-register sequences and conventional error-correcting codes (cf. [MS77], Chapter 7), so there should be a deeper explanation.

The context in which sequence A16 appears in Golomb's book [Go67] is the enumeration of the (infinite) output sequences from certain types of n-stage binary shift registers. Following Golomb, we consider four kinds of shift registers: the *pure cycling register* (or PCR), as illustrated in Fig. 1, the

complemented cycling register (or CCR), the *pure summing register* (or PSR) and the *complemented summing register* (or CSR). If the shift register has n cells, initially containing $x_1, x_2, \ldots, x_n$ ($x_i = 0$ or 1), then x_1 is appended to the output stream, symbols $x_2, \ldots, x_n$ move to the left, and the symbol

$$\begin{array}{ll} \text{(PCR)} & x_1 \\ \text{(CCR)} & 1 + x_1 \\ \text{(PCR)} & x_1 + x_2 + \cdots + x_n \qquad \text{or} \\ \text{(CSR)} & 1 + x_1 + x_2 + \cdots + x_n \end{array}$$

is fed back to the right-most cell.

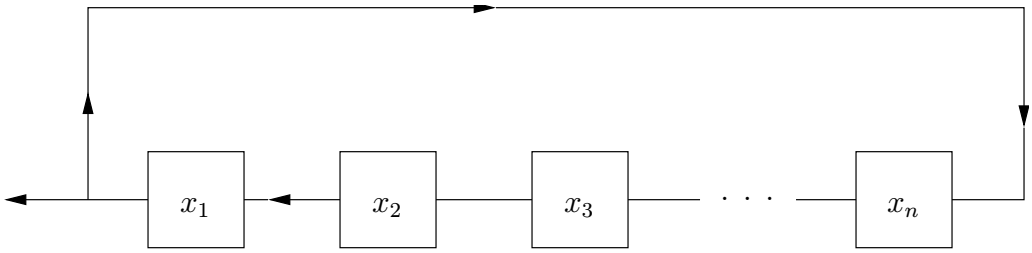

Figure 1: An n-stage pure cycling register.

The problem is to determine the numbers of different possible output sequences from these registers, which we denote by $Z(n)$, $Z^*(n)$, $S(n)$ and $S^*(n)$, respectively. For example $S^*(5) = 6$, corresponding to the sequences

$$\begin{array}{l} \ldots\, 0\,0\,0\,0\,0\,1\,0\,0\,0\,0\,0\,1 \ldots \\ \ldots\, 0\,0\,0\,1\,1\,1\,0\,0\,0\,1\,1\,1 \ldots \\ \ldots\, 0\,0\,1\,0\,1\,1\,0\,0\,1\,0\,1\,1 \ldots \\ \ldots\, 0\,1\,0\,0\,1\,1\,0\,1\,0\,0\,1\,1 \ldots \\ \ldots\, 0\,1\,0\,1\,0\,1\,0\,1\,0\,1\,0\,1 \ldots \\ \ldots\, 0\,1\,1\,1\,1\,1\,0\,1\,1\,1\,1\,1 \ldots\ , \end{array}$$

all having period 6 (or a divisor of 6).

Table 2, based on [Go67, page 172], shows the first few values of these functions, together with the corresponding sequence numbers from [EIS].

Explicit formulas for these functions are given in the next theorem.

Table 2: Number of ouput sequences from n-stage shift registers of types PCR, CCR, PSR, CSR.

	PCR	CCR	PSR	CSR
n	$Z(n)$	$Z^*(n)$	$S(n)$	$S^*(n)$
1	2	1	2	1
2	3	1	2	2
3	4	2	4	2
4	6	2	4	4
5	8	4	8	6
6	14	6	10	10
7	20	10	20	16
8	36	16	30	30
9	60	30	56	52
10	108	52	94	94
...	...	...	...	...
Sequence:	A31	A16	A13	A16

Theorem 3.1. *For $n \geq 1$,*

$$Z(n) = \frac{1}{n}\sum_{d|n} \phi(d) 2^{n/d} , \tag{11}$$

$$Z^*(n) = S^*(n-1) = \frac{1}{2n} \sum_{\substack{d|n \\ d \text{ odd}}} \phi(d) 2^{n/d} , \tag{12}$$

$$S(n) = \frac{1}{2(n+1)} \sum_{d|n+1} \phi(2d) 2^{(n+1)/d} . \tag{13}$$

Remark 3.2. Golomb proves (11) and sketches proofs of the other results. Actually (13) is due to Michael Somos (personal communication), Golomb's version (given in (15) below) being slightly more complicated. The numbers $Z(n)$ (sequence A31) in the first column are also familiar as the number of binary irreducible polynomials of degree dividing n, and the number of n-bead necklaces formed with beads of two colors, when the necklaces may not be turned over (cf. [Be68, Chap. 4], [GR61], [MS77, Chap. 4], [St99, Problem 7.112]). Fredricksen [Fr70] shows that $Z(n) - 1$ is the number of 1's in the truth table defining the lexicographically least de Bruijn cycle.

Proof. Note that sequence A16 appears in two places in the table, for CCR registers of length n and CSR registers of length $n-1$. We begin by explaining this, and thus proving that

$$Z^*(n) = S^*(n-1)\ . \tag{14}$$

Suppose for concreteness that $n = 4$. The output sequences from the four types of register are (omitting plus signs, and writing $1a$ rather than $1 + a$, etc.):

(i)	a	b	c	d	a	b	c	d	a	$\cdots$				
(ii)	a	b	c	d	$1a$	$1b$	$1c$	$1d$	a	b	c	d	$\cdots$	
(iii)	a	b	c	d	$abcd$	a	b	c	d	$abcd$	a	b	c	d	$\cdots$
(iv)	a	b	c	d	$1abcd$	a	b	c	d	$1abcd$	a	b	c	d	$\cdots$

In general these sequences have periods n, $2n$, $n+1$ and $n+1$, respectively. If we replace (ii) by the sums of adjacent pairs we get

$$ab \quad bc \quad cd \quad 1ad \quad ab \quad bc \quad cd \quad 1ad \quad \ldots\ ,$$

a CSR(3) sequence. Conversely, given a CSR(3) sequence, say

$$A \quad B \quad C \quad 1ABC \quad A \quad B \quad C \quad 1ABC \quad \ldots\ ,$$

of period 4, there is a unique CCR(4) sequence of period 8 corresponding to it, namely

$$0 \quad A \quad AB \quad ABC \quad 1 \quad 1A \quad 1AB \quad 1ABC \quad 0 \quad A \quad \ldots\ .$$

Applying this argument in the general case establishes (14).

In the rest of the proof we make use of Burnside's lemma (cf. [St99]), which states that the number of orbits of a finite permutation group G is equal to the average number of points that are fixed by the elements of G.

Let us first prove (11). (This is Golomb's proof [Go67, p. 121].) We take G to be the cyclic group of order n generated by $\pi = (1, 2, \ldots, n)$, acting on $\mathbb{F}_2^n$. The permutation π^i $(1 \le i \le n)$ contains $gcd(n, i)$ cycles, each of length $n/gcd(n, i)$, and has order $n/gcd(n, i)$. There are precisely $2^{gcd(n,i)}$ vectors fixed by π^i, since each cycle must consist of all 0's or all 1's. Hence,

by Burnside's lemma,

$$\begin{aligned} Z(n) &= \frac{1}{n}\sum_{i=1}^{n} 2^{gcd(n,i)} \\ &= \frac{1}{n}\sum_{k|n}\sum_{\substack{i=1 \\ gcd(n,i)=k}}^{n} 2^k \\ &= \frac{1}{n}\sum_{k|n} 2^k \sum_{gcd\left(\frac{n}{k},i\right)=1} 1 \\ &= \frac{1}{n}\sum_{k|n} \phi\left(\frac{n}{k}\right) 2^k \\ &= \frac{1}{n}\sum_{d|n} \phi(d) 2^{n/d} \ . \end{aligned}$$

To establish (12), we note from (iv) that $S^*(n-1)$ is equal to the number of orbits of the same group, but now acting on binary vectors of length n and odd weight. The number of odd weight vectors fixed by π^i is $2^{gcd(n,i)-1}$ if the cycle lengths $n/gcd(n,i)$ are odd, and zero otherwise. Hence

$$\begin{aligned} S^*(n-1) &= \frac{1}{n}\sum_{\substack{i=1 \\ n/gcd(n,i) \text{ odd}}}^{n} 2^{gcd(n,i)-1} \\ &= \frac{1}{n}\sum_{\substack{k|n \\ n/k \text{ odd}}} \phi\left(\frac{n}{k}\right) 2^{k-1} \\ &= \frac{1}{2n}\sum_{\substack{d|n \\ d \text{ odd}}} \phi(d) 2^d \ . \end{aligned}$$

Finally, we prove (13), by determining $S(n-1)$. The group is the same, but now (see (iii)) acting on even weight vectors. If $d = n/gcd(n,i)$ is even there are 2^d fixed vectors, but if d is odd only 2^{d-1} fixed vectors. Hence

$$\begin{aligned} S(n-1) &= \frac{1}{n}\sum_{\substack{d|n \\ d \text{ odd}}} \phi(d) 2^{d-1} + \frac{1}{n}\sum_{\substack{d|n \\ d \text{ even}}} \phi(d) 2^d \qquad (15) \\ &= \frac{1}{2n}\sum_{d|n} \phi(2d) 2^{n/d} \ , \end{aligned}$$

since $\phi(2d) = \phi(d)$ if d odd, $\phi(d) = 2\phi(d)$ if d even. □

But a mystery still remains: is the fact that the number of codewords in $VT_0(n)$ equals $Z(n)$ just a numerical coincidence, or is there a one-to-one correspondence between the codewords and the CCR shift register sequences? (This is essentially equivalent to a research problem stated by Stanley in [St86], Chapter 1, Problem 27(c).)

Furthermore, why is $|VT_1(n)|$ (sequence A48 in [EIS]), equal to the number of $(n+1)$-bead necklaces with beads of two colors and primitive period $n+1$, when the two colors may be interchanged but the necklaces may not be turned over (cf. [Fi58], [GR61])? This is also the number of irreducible polynomials over $\mathbb{F}_2$ of degree $n+1$ in which the coefficient of x^n is 1 [Car52], [CMRSS].

4 Locally transitive tournaments

The entry for A16 in [EIS] also indicates that this sequence arose in Brouwer's enumeration [Br80] of locally transitive tournaments. A tournament is a directed graph with one directed edge between any two nodes. It is transitive if there are no directed cycles.

A locally transitive tournament is a tournament such that the subgraphs on the predecessors of a point and the successors of a point are both transitive.

Brouwer, answering a question raised by P. J. Cameron, determined the number of locally transitive tournaments on n nodes. He began by calculating the first few values by computer. Then he looked up this sequence in [HIS], and found the reference to Golomb's book [Go67]. With this hint alone, and without having access to the book, he established a one-to-one correspondence between these tournaments and output sequences from shift registers of CCR type. From this he obtained the formula

$$\sum_{d|n} \text{odd}\left(\frac{n}{d}\right) \frac{2^{d-1}}{d} \sum_{e|\frac{n}{d}} \frac{\mu(e)}{e} , \tag{16}$$

where odd(i) is 0 or 1 according to whether i is even or odd, and μ is the Möbius function (A8683). Using the identity

$$\phi(n) = \sum_{d|n} \mu(d) \frac{n}{d}$$

([Ap76], p. 26), (16) immediately reduces to (12).

Again we can ask, is there a connection between locally transitive tournaments and the $VT_0(n)$ codes?

5 The number of descendants of a vector

It was already mentioned in Section 2 that $|D_1(u)| = r(u)$, the number of runs in u.

The next theorem was discovered by E. M. Rains and the author. Although this must be well-known, we have not found it in the literature.

The *derivative* $u' \in \mathbb{F}_2^{n-1}$ of $u = (u_1, \ldots, u_n) \in \mathbb{F}_2^n$ is given by

$$u' = (u_1 + u_2, u_2 + u_3, \ldots, u_{n-1} + u_n) \ .$$

Note that $wt(u') = r(u) - 1$.

Theorem 5.1. *The number of grandchildren of a vector:*

$$|D_2(u)| = \binom{r(u)+1}{2} - \delta \ , \tag{17}$$

where $\delta = 2wt(u') - wt(u'')$ *is the* deficiency *of* u.

Sketch of proof. First, suppose u is a "normal" vector, meaning that all runs have length ≥ 2, for example

$$\begin{array}{lllllllllllll} u & = & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 0 & 0 & 0 & \\ u' & = & & 0 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & \\ u'' & = & & & 0 & 0 & 1 & 1 & 0 & 1 & 1 & 0 & \end{array} \tag{18}$$

Then $|D_2(u)| = \binom{r(u)+1}{2}$ is the number of ways of choosing two things out of $r(u)$ with repetitions allowed. Suppose the successive runs in u have lengths $i, j, k, l, \ldots$. Let $v \in D_2(u)$ be any vector obtained by deleting two coordinates from u. Then the lengths of the runs in v are one of the following:

$$\begin{array}{l} i-2,\ j,\ k,\ l,\ \ldots \\ i,\ j-2,\ k,\ l,\ \ldots \\ i,\ j,\ k-2,\ l,\ \ldots \\ \quad \ldots \ \ \ldots \\ i-1,\ j-1,\ k,\ l,\ \ldots \\ i-1,\ j,\ k-1,\ l,\ \ldots \\ \quad \ldots \ \ \ldots \end{array} \tag{19}$$

For a normal vector $wt(u'') = 2wt(u')$ (cf. (18)), $\delta = 0$ and (17) holds.

Next suppose that all runs in u have length ≥ 2 except for a single internal run of length 1, as in

$$\begin{array}{rcccccccccccccccc} u & = & 0 & & 0 & & 0 & & 0 & & 1 & & 0 & & 0 & & 0 \\ u' & = & & 0 & & 0 & & 0 & & 1 & & 1 & & 0 & & 0 & \\ u'' & = & & & 0 & & 0 & & 1 & & 0 & & 1 & & 0 & & \end{array}$$

Then $\delta = 2$, and indeed $|D_2(u)|$ is 2 less than it would be for a normal vector, since one of the possibilities in (19) vanishes and two others coalesce.

The remaining cases, when there are several runs of length 1, possibly including beginning or ending runs, are left to the reader. □

It is not clear how to generalize Theorem 5.1 to k-th order descendants. Certainly $D_3(u)$ is not simply a function of the weights of u, u', u'' and u'''.

Theorem 5.2. *Let*

$$\mu_k(n) = \max_{u \in \mathbb{F}_2^n} |D_k(u)|$$

be the maximal number of k-th order descendants of any binary vector of length n. Then

$$\mu_k(n) = \sum_{i=0}^{k} \binom{n-k}{i} , \tag{20}$$

for $n \geq k+1$. Equality is achieved just by the vectors

$$010101\ldots \quad \textit{and} \quad 101010\ldots . \tag{21}$$

According to Calabi and Hartnett [CH69], (20) is proved in an unpublished 1967 report[4] of Calabi [Cal67]. The first published proof seems to have been given by Levenshtein [Lev96]. It was generalized to the nonbinary case by Hirschberg [Hir99] (see also Levenshtein [Lev01] and Hirschberg and Regnier [HR01]).

It is not difficult to show that the vectors (21) achieve the bound in (20).

Theorem 5.3. *For the two vectors* $010101\ldots$ *and* $101010\ldots$ *we have*

$$|D_k(u)| = \sum_{i=0}^{k} \binom{n-k}{i} . \tag{22}$$

[4] I have been unable to locate a copy of this report. I suspect it was never completed.

Proof. Let $u = 010101\ldots \in \mathbb{F}_2^n$, let $M_{n,k}$ be the set of k-th order descendants of u, and let $m_{n,k} = |M_{n,k}|$. Then

$$\begin{aligned} M_{n,k} &= 0|\bar{M}_{n-1,k} \cup M_{n-1,k-1} \\ &= 0|\bar{M}_{n-1,k} \cup 1|M_{n-2,k-1} \cup M_{n-2,k-2} \ , \end{aligned} \tag{23}$$

where the bars denote binary complementation. However, the last term in (23) can be dropped because it is contained in the union of the other two terms. Since these two terms are disjoint, we have

$$M_{n,k} = M_{n-1,k} + M_{n-2,k-1} \ .$$

This is a disguised version of the recurrence for binomial coefficients, whose solution is given by (22). □

The case $k = 2$ of (20) is a corollary of Theorem 5.1:

Corollary 5.4. *For $n \geq 3$,*

$$\mu_2(n) = \sum_{i=0}^{2} \binom{n-2}{i} = \frac{1}{2}(n^2 - 3n + 4) \ . \tag{24}$$

Proof. Let u achieve $\mu_2(n)$. The result is easily verified if $r(u)$ is 1 or 2, so we assume $r(u) \geq 3$.

Suppose u begins with a string of $k \geq 0$ runs of length 1, followed by a run of length ≥ 2 from position $k+1$. We will show that the vector u^* obtained by complementing u from position $k+2$ onwards satisfies $|D_2(u^*)| \geq |D_2(u)|$. By repeating this operation we eventually arrive at one of the vectors (21).

Since $|D_k(u)| = |D_k(\bar{u})|$, we may assume that the run following the initial k runs of length 1 in u begins $11x\ldots$. In u^* this is replaced by $10\bar{x}\ldots$. Then we find that u^* has $r(u) + 1$ runs, and $wt(u^{*\prime\prime}) = wt(u'') - 2 + 2x$, from which it follows using (17) that $|D_2(u^*)| - |D_2(u)| = r(u) + 2x - 3 \geq 0$, as required. □

6 Related work

The history of deletion-correcting codes is closely tied up with studies of codes for correcting other classes of errors such as:

- erasures, when bits whose positions are known are deleted
- insertions of bits (rather than deletions)

- asymmetric errors, when the only errors that occur are that 1's may be changed to 0's (this is also known as a Z-channel)
- unidirectional errors: 0's may be changed to 1's or 1's to 0's, but only one type of error occurs in any particular transmission
- bit reversals: 0's may be changed to 1's or vice versa — this is the subject of classical coding theory
- transpositions: adjacent bits may be swapped
- any meaningful combination of the above.

Furthermore the alphabet may be changed from $\mathbb{F}_2$ to $\mathbb{F}_q$. This produces an extensive list of families of codes, and of course in each case one can ask for the largest codes.

In this section we give a brief overview of some other relevant papers. First, Levenshtein's papers [Lev65], [Lev65a], [Lev92], [Lev01] should be considered essential reading.

Hartnett [Ha74] (see especially Calabi and Hartnett [CH69]) contains some general investigations of all the above-mentioned codes (both block codes and variable length codes) from a fairly abstract mathematical point of view.

One of the earliest papers to study deletion-correcting codes is Sellers [Se62], which combines a special separating string between blocks with a burst-error correcting code inside the blocks.

Ullman [Ull66] uses a construction similar to that of Varshamov and Tenegolts, but his codes are not as efficient and also use a separating string between blocks. In [Ull67] he gives bounds on the size of codes for correcting synchronization errors.

Tenengolts [Ten84] generalizes the $VT_a(n)$ codes to larger alphabets. Nonbinary codes are also discussed in [Bo94], [Bo95], [Do85], [Ma98].

Other constructions for deletion-correcting and related codes are given by Calabi and Hartnett [CH69a], Iizuka, Kasahara and Namekawa [IKN], Kløve [Kl95] and Tanaka and Kasai [TK76].

The most recent paper on this subject is by Schulman and Zuckerman [SZ99], who present what they describe as "simple, polynomial-time encodable and decodable codes which are asymptotically good for channels allowing insertions, deletions and transpositions". The number of errors that can be corrected is some constant fraction of the block-length n. The constructions are not explicit.

We conclude this section by mentioning some papers on peripherally related codes. Codes for correcting asymmetric and unidirectional errors are discussed in [BR82], [Et91], [EO98], [WVB88] and [WVB89]. Erasure correcting codes are discussed by Alon and Luby [AL96] and Barg [Ba98].

Acknowledgements

I would like to thank Andries Brouwer, Suhas Diggavi, Vladimir Levenshtein, Andrew Odlyzko, Eric Rains and Richard Stanley for conversations about the subject of this paper; and David Applegate, David Johnson and Mauricio Resende for their help in establishing that the $VT_0(n)$ codes are optimal for $n \leq 9$ and for their (so far unsuccessful!) attempts to find better codes. Ulrich Tamm kindly supplied the Martirosyan reference.

References

[AL96] N. Alon and M. Luby, A linear time erasure resilient code with nearly optimal recovery, *IEEE Trans. Inform. Theory*, **42** (1996), 1732–1736.

[Ap76] T. M. Apostol, *Introduction to Analytic Number Theory*, Springer-Verlag, NY, 1976.

[Ba98] A. Barg, Complexity issues in coding theory, pp. 649–754 in *Handbook of Coding Theory*, ed. V. S. Pless and W. C. Huffman, North-Holland, Amsterdam, 1998.

[Be68] E. R. Berlekamp, *Algebraic Coding Theory*, McGraw-Hill, NY, 1968.

[BR82] B. Bose and T. R. N. Rao, Theory of undirectional error correcting/detecting codes, *IEEE Trans.Computers* **31** (1982), 521–530.

[Bo94] P. A. H. Bours, Construction of fixed-length insertion/deletion correcting runlength-limited codes, *IEEE Trans. Inform. Theory*, **40** (1994), 1841–1856.

[Bo95] P. A. H. Bours, On the construction of perfect deletion-correcting codes using design theory, *Designs, Codes Crypt.*, **6** (1995), 5–20.

[Br80] A. E. Brouwer, *The Enumeration of Locally Transitive Tournaments*, Math. Centr. Report ZW138, Amsterdam, April 1980.

[Cal67] L. Calabi, *On the Computation of Levenshtein's Distances*, Report TM-9-0030, Parke Mathematical Laboratories, Inc., Carlisle, MA, 1967. Research supported by Air Force Cambridge Research Laboratories under contracts AF19(628)-3826 and F1962867COO30.

[CH69] L. Calabi and W. E. Hartnett, Some general results of coding theory with applications to the study of codes for the correction of synchronization errors, *Inform. Control*, **15** (1969), 235–249. Reprinted in Hartnett [Ha74].

[CH69a] L. Calabi and W. E. Hartnett, A family of codes for the correction of substitution and synchronization errors, *IEEE Trans. Inform. Theory*, **15** (1969), 102–106.

[Car52] L. Carlitz, A theorem of Dickson on irreducible polynomials, *Proc. Amer. Math. Soc.*, **3** (1952), 693–700.

[CMRSS] K. Cattell, C. R. Miers, F. Ruskey, J. Sawada and M. Serra, The number of irreducible polynomials over $GF(2)$ with given trace and subtrace, preprint, 2000.

[Do85] A. S. Dolgopolov, Nonbinary codes correcting symbol insertions, deletions and substitutions (in Russian), *Problemy Peredachi Informatsii*, **21** (No. 1, 1985), 35–39. English translation in *Problems of Information Transmission*, **21** (No. 1, 1985).

[Et91] T. Etzion, New lower bounds for asymmetric and unidirectional codes, *Trans. Inform. Theory*, **37** (1991), 1696–1704.

[EO98] T. Etzion and P. R. J. Östergård, Greedy and heuristic algorithms for codes and colorings, *IEEE Trans. Inform. Theory*, **44** (1998), 382–388.

[Fi58] N. J. Fine, Classes of periodic sequences, *Illinois J. Math.*, **2** (1958), 285–302.

[Fr70] H. Fredricksen, The lexicographically least de Bruijn cycle, *J. Combin. Theory*, **9** (1970), 1–5.

[GR61] E. N. Gilbert and J. Riordan, Symmetry types of periodic sequences, *Illinois J. Math.*, **5** (1961), 657–665.

[Gi67] B. D. Ginzburg, A number-theoretic function with an application in the theory of coding (in Russian), *Problemy Kibernetiki*, **19**

(1967), 249–252. English translation in *Systems Theory Research*, **19** (1970), 255–259.

[Go67] S. W. Golomb, *Shift Register Sequences*, Holden-Day, San Francisco, 1967.

[GS80] R. L. Graham and N. J. A. Sloane, Lower bounds for constant weight codes, *IEEE Trans. Inform. Theory*, **26** (1980), 37–43.

[Ha74] W. E. Hartnett, editor, *Foundations of Coding Theory*, Reidel, Dordrecht, Holland, 1974.

[Hir99] D. S. Hirschberg, Bounds on the number of string subsequences, pp. 115–122 in *Proc. Symp. on Combinatorial Pattern Matching, Warwick*, *Lecture Notes in Computer Science*, Springer-Verlag, Berlin, 1999.

[HR01] D. S. Hirschberg and M. Regnier, Tight bounds on the number of string subsequences, *J. Discrete Algorithms*, to appear.

[IKN] I. Iizuka, M. Kasahara and T. Namekawa, Block codes capable of correcting both additive and timing errors, *IEEE Trans. Inform. Theory*, **26** (1980), 393–400.

[Kl81] T. Kløve, A lower bound for $A(n, 4, w)$, *IEEE Trans. Inform. Theory*, **27** (1981), 257–258.

[Kl95] T. Kløve, Codes correcting a single insertion/deletion of a zero or a single peak-shift, *IEEE Trans. Inform. Theory*, **41** (1995), 279–283.

[Ko88] T. W. Körner, *Fourier Analysis*, Cambridge Univ. Press, 1988.

[Lev65] V. I. Levenshtein, Binary codes capable of correcting deletions, insertions and reversals (in Russian), *Doklady Akademii Nauk SSSR*, **163** (No. 4, 1965), 845–848. English translation in *Soviet Physics Dokl.*, **10** (No. 8, 1966), 707–710.

[Lev65a] V. I. Levenshtein, Binary codes capable of correcting spurious insertions and deletions of ones (in Russian), *Problemy Peredachi Informatsii*, **1** (No. 1, 1965), 12–25. English translation in *Problems of Information Transmission*, **1** (No. 1, 1965), 8–17.

[Lev92] V. I. Levenshtein, On perfect codes in the deletion/insertion metric (in Russian), *Diskret. Mat.* **3** (No. 1, 1991), 3–20. English translation in *Discrete Mathematics and Applications*, **2** (No. 3, 1992), 241–258.

[Lev96] V. I. Levenshtein, Reconstructing binary sequences by the minimum number of their subsequences or supersequences of a given length, pp. 176–183 in *Proceedings of Fifth Intern. Workshop on Algebr. and Combin. Coding Theory, Sozopol, Bulgaria, June 1-7, 1996.*

[Lev01] V. I. Levenshtein, Efficient reconstruction of sequences, *IEEE Trans. Inform. Theory*, **47** (2001), 2–22.

[MS77] F. J. MacWilliams and N. J. A. Sloane, *The Theory of Error-Correcting Codes*, North-Holland, Amsterdam, 1977.

[Ma98] A. Mahmoodi, Existence of perfect 3-deletion correcting codes, *Designs, Codes Crypt.*, **14** (1998), 81–87.

[Mt96] S. Martirosyan, Single-error correcting close-packed and perfect codes, pp. 90–115 in *Proceedings of the 1st INTAS International Seminar on Coding Theory and Combinatorics*, Thakhadzor, Armenia, 1996.

[PH98] V. S. Pless and W. C. Huffman, *Handbook of Coding Theory*, North-Holland, Amsterdam, 1998.

[SZ99] L. J. Schulman and D. Zuckerman, Asymptotically good codes correcting insertions, deletions and transpositions, *IEEE Trans. Inform. Theory*, **45** (1999), 2552–2557.

[Se62] F. F. Sellers, Jr., Bit loss and gain correction codes, *IEEE Trans. Inform. Theory*, **8** (1962), 35–38.

[HIS] N. J. A. Sloane, *A Handbook of Integer Sequences*, Academic Press, New York 1973.

[SL01] N. J. A. Sloane, http://www.research.att.com/∼njas/ (home page).

[EIS] N. J. A. Sloane, *The On-Line Encyclopedia of Integer Sequences*, http://www.research.att.com/∼njas/sequences/.

[St86] R. P. Stanley, *Enumerative Combinatorics*, Wadsworth, Monterey, CA, Vol. 1, 1986.

[St99] R. P. Stanley, *Enumerative Combinatorics*, Cambridge Univ. Press, Vol. 2, 1999.

[TK76] E. Tanaka and T. Kasai, Synchronization and substitution error-correcting codes for the Levenshtein metric, *IEEE Trans. Inform. Theory*, **22** (1976), 156–162.

[Ten76] G. M. Tenengolts, Class of codes correcting bit loss and errors in the preceding bit (in Russian), *Avtomatika i Telemakhanika*, **5** (1976), 174–179. English translation in *Automation and Remote Control*, **37** (No. 5, 1976), 797–802.

[Ten84] G. Tenengolts, Nonbinary codes correcting single deletion or insertion, *IEEE Trans. Inform. Theory*, **30** (1984), 766–769.

[Ull66] J. D. Ullman, Near-optimal, single-synchronization-error-correcting code, *IEEE Trans. Inform. Theory*, **12** (1966), 418–424.

[Ull67] J. D. Ullman, On the capabilities of codes to correct synchronization errors, *IEEE Trans. Inform. Theory*, **13** (1967), 95–105.

[Var65] R. R. Varshamov, On an arithmetic function with an application in the theory of coding (in Russian), *Dokl. Akad. Nauk SSSR*, **161** (No. 3, 1965), 540–543.

[VT65] R. R. Varshamov and G. M. Tenengolts, Codes which correct single asymmetric errors (in Russian), *Avtomatika i Telemekhanika*, **26** (No. 2, 1965), 288–292. English translation in *Automation and Remote Control*, **26** (No. 2, 1965), 286–290.

[WVB88] J. H. Weber, C. de Vroedt and D. E. Boekee, Bounds and constructions for binary codes of length less than 24 and asymmetric distance less than 6, *IEEE Trans. Inform. Theory*, **34** (1988), 1321–1331.

[WVB89] J. H. Weber, C. de Vroedt and D. E. Boekee, Bounds and constructions for codes correcting unidirectional errors, *IEEE Trans. Inform. Theory*, **35** (1989), 797–810.