Can. J. Math., Vol. XXXIX, No. 3, 1987, pp. 513-526

A NEW RESULT ON COMMA-FREE CODES OF EVEN WORD-LENGTH

BETTY TANG, SOLOMON W. GOLOMB AND RONALD L. GRAHAM

1. Introduction. Comma-free codes were first introduced in [**1**] in 1957 as a possible genetic coding scheme for protein synthesis. The general mathematical setting of such codes was presented in [**3**], and the biochemical and mathematical aspects of the problem were later summarized and extended in [**4**].

Using the notation of [**3**], a set D of k-tuples or k-letter words, $(a_1a_2 \dots a_k)$, where

$$a_i \in \mathbf{Z}_n = \{0, 1, 2, \dots, n - 1\},$$

for fixed positive integers k and n, is said to be a *comma-free dictionary* if and only if, whenever $(a_1a_2 \dots a_k)$ and $(b_1b_2 \dots b_k)$ are in D, the "overlaps"

$$(a_ia_{i+1} \dots a_kb_1 \dots b_{i-1}), \quad 2 \leqq i \leqq k,$$

are not in D. This precludes codewords having a subperiod less than k; and two codewords which are cyclic permutations of one another cannot both be in D. Therefore at most one member from the non-periodic cyclic equivalence class of $(a_1 \dots a_k)$, i.e., from the set

$$\{ (a_j \dots a_ka_1 \dots a_{j-1}) \,|\, 1 \leqq j \leqq k \},$$

can be in D. The maximum number of codewords, $W_k(n)$, in the comma-free dictionary D therefore cannot exceed the number of non-periodic cyclic equivalence classes of sequences of length k formed from an alphabet of n letters. Denoting the latter number by $B_k(n)$, we have formally,

$$W_k(n) \leqq B_k(n)$$

where

$$B_k(n) = \frac{1}{k} \sum_{d|k} \mu(d)n^{k/d}$$

The summation is extended over all divisors d of k, and $\mu(d)$ is the Möbius function.

Received November 21, 1983, and in revised form February 12, 1986. This research was supported in part by the National Security Agency, under Contract No. MDA904-83-H-0004.

Golomb, Gordon and Welch [3] proved that $W_k(n)$ attains the upper bound $B_k(n)$ for arbitrary n if $k = 1, 3, 5, 7, 9, 11, 13, 15$, and conjectured that this is indeed the case for all odd k. The conjecture was proved by Eastman [2], who gave a construction for the maximal comma-free dictionaries. A simpler construction for these dictionaries was found by Scholtz [6].

The results for even integers k were less complete. Golomb, Gordon and Welch [3] were able to prove that $W_k(n)$ cannot attain the bound $B_k(n)$ for $n > 3^{k/2}$; and in particular,

$$W_2(n) = \left[\frac{n^2}{3}\right]$$

where $[x]$ is the integral part of x, whereas

$$B_2(n) = \frac{n^2 - n}{2}.$$

It was also mentioned that for $k = 4$, we in fact have $W_4(n) < B_4(n)$ if $n \geqq 5$, while $W_4(n) = B_4(n)$ if $n = 1, 2, 3$. The case for $n = 4$ was later solved in [5] by exhaustive computer search, which found $W_4(4) = 57 < B_4(4) = 60$.

An improvement on the relation between k and n such that $W_k(n) < B_k(n)$ for even k was given by Jiggs [5]:

$$W_k(n) < B_k(n) \text{ if } n > 2^{k/2} + \frac{k}{2}.$$

We present a further improvement based on Jiggs' proof, which in turn gives rise to a very interesting combinatorial problem. We first present Jiggs' result (attributed by Jiggs to R. I. Jewett) with some modifications of notation.

We consider only the simpler problem of forming a comma-free dictionary D with $\binom{n}{2}$ codewords of length $k = 2l$, with one representative from each cyclic class of the type $(a00\ldots0b00\ldots0)$, with $0 \leqq a < b \leqq n - 1$ and $l - 1$ 0's between a and b. Clearly if these $\binom{n}{2}$ classes cannot be simultaneously represented in a comma-free dictionary, the full set of $B_k(n)$ classes cannot be so represented.

A *half-word* in D is an l-tuple which is either the initial half or final half of some word in D. For each $d \in Z_n$ and $1 \leqq r \leqq k/2$, let $u(d, r)$ denote the half-word with d at the r-th position and 0 everywhere else. We assign a sequence

$$x^d = x_1^d x_2^d \ldots x_l^d$$

to each $d \in \mathbf{Z}_n$ where x_r^d is defined in the following way:

$$x_r^d = \begin{cases} 2 & \text{if } u(d, r) \text{ is both initial and final} \\ 1 & \text{if } u(d, r) \text{ is final only} \\ 0 & \text{if } u(d, r) \text{ is initial only} \\ * & \text{if } u(d, r) \text{ is neither initial nor final.} \end{cases}$$

Jiggs showed that the sequences x^d have the following two properties:

(1) If $d \neq b$, then x_r^d and x_r^b cannot both be 2, for any $1 \leqq r \leqq l$. Thus at most l of the sequences x^d can contain the symbol 2.

(2) Among the sequences in which the symbol 2 does not occur, if $d \neq b$, there exists $1 \leqq r \leqq l$ such that either $x_r^d = 0$ and $x_r^b = 1$, or $x_r^b = 0$ and $x_r^d = 1$. (In particular, distinct letters of the alphabet must have distinct sequences.)

We call two sequences, x^d and x^b, composed of 0, 1, and *, *comparable* if they have property (2). The two properties imply that the maximum number of distinct sequences x^d containing a 2 is l, and the maximum number of distinct sequences x^d containing no 2 is 2^l. Hence if $|D| = B_k(n)$, then $n \leqq 2^{k/2} + k/2$.

Our improvement on Jiggs' result is a consequence of the following observation.

THEOREM 1.1. *If $d \neq b$ and $r \neq s$, we cannot have both $x_r^d = x_s^b = 1$ and $x_r^b = x_s^d = 0$.*

Proof. Suppose there exist $r \neq s$ such that $x_r^d = x_s^b = 1$ and $x_r^b = x_s^d = 0$. Then we will have words of the following form:

$$w_1 = (0 \ldots 0p0 \ldots 0d0 \ldots 0),$$

$$w_2 = (0 \ldots 0b0 \ldots 0q0 \ldots 0),$$

where the non-zero letters appear at positions r and $l + r$, and

$$w_3 = (0 \ldots 0x0 \ldots 0b0 \ldots 0),$$

$$w_4 = (0 \ldots 0d0 \ldots 0y0 \ldots 0),$$

where the non-zero letters appear at positions s and $s + l$. The overlaps of w_1w_2 and w_3w_4 therefore contain all members of the cyclic equivalence class of $(0 \ldots 0b0 \ldots 0d0 \ldots 0)$ and so D cannot contain a representative of this class and still be comma-free.

We will call two sequences x^d and x^b *compatible* if they satisfy the exclusion condition in Theorem 1.1. We will now address the combinatorial problem of determining the maximum size of a set S of sequences of length l, composed of *, 0, and 1 such that the sequences are pairwise comparable and compatible.

2. The minimal array. Let $t = t(l)$ be the maximum number of distinct l-tuples of 0's, 1's, and *'s which are pairwise comparable and compatible.

We will try to determine t indirectly. Suppose we have an array of empty boxes with t rows in the array. We must fill in each empty box with either *, 0 or 1 such that every two rows, taken as sequences, are comparable and compatible. We want to know the minimum number of distinct columns in the array when there are t rows. Let $f(t)$ be that minimum number, and call the array thus obtained the *minimum array* M_t. Obviously, $f(t) \leqq l$.

We define $t(1) = 0$. The value of $f(t)$ for small t can be obtained without much difficulty. (See Table 1).

TABLE 1

$t = 2,\ f(t) = 1$

$$M_2 = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$$

$t = 3,\ f(t) = 2$

$$M_3 = \begin{bmatrix} 0 & 0 \\ 0 & 1 \\ 1 & 1 \end{bmatrix}$$

$t = 4,\ f(t) = 3$

$$M_4 = \begin{bmatrix} 0 & 1 & * \\ * & 0 & 1 \\ 1 & * & 0 \\ 1 & 1 & 1 \end{bmatrix}$$

$t = 5,\ f(t) = 3$

$$M_5 = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 1 & * \\ 1 & * & 0 \\ * & 0 & 1 \\ 1 & 1 & 1 \end{bmatrix}$$

$t = 6,\ f(t) = 4$

$$M_6 = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 1 & * & 1 \\ 1 & * & 0 & 1 \\ * & 0 & 1 & 1 \\ 1 & 1 & 1 & 1 \end{bmatrix}$$

Note that there can be more than one minimal array M_t for each t. Also, t as a function of l is simply the largest number s such that $f(s) = l$. From Table 1 we get the values of $t(l)$ for some l. (See Table 2.)

TABLE 2

l	$t(l)$
1	2
2	3
3	5
4	$\geqq 6$

We can immediately establish a few properties of $f(t)$.

THEOREM 2.1. *$f(t)$ is a monotonically non-decreasing function of t.*

Proof. Let $s > t$. We can remove any $s - t$ rows from the minimal array M_s and the remaining array of t sequences will still be pairwise comparable and compatible. Therefore $f(t) \leqq f(s)$.

THEOREM 2.2. $f(t+1) \leqq f(t)+1$.

Proof. From the minimal array M_t, construct a set of $t+1$ sequences and $f(t)+1$ columns in the following way. A 1 is added to the end of every sequence in M_t, and a sequence x^{t+1} of length $f(t)+1$ containing all 0's is adjoined to the set. The sequences in the new set are still pairwise comparable and compatible, and so

$$f(t+1) \leqq f(t)+1.$$

THEOREM 2.3. $f(t) \leqq t-1$.

Proof. The sequences in the following array of $t-1$ columns are pairwise comparable and compatible, so $f(t) \leqq t-1$.

$$t \text{ rows} \left\{ \overbrace{\begin{array}{cccccc c} 0 & 0 & 0 & \cdots & 0 & 0 & 0 \\ 0 & 0 & 0 & \cdots & 0 & 0 & 1 \\ 0 & 0 & 0 & \cdots & 0 & 1 & * \\ 0 & 0 & 0 & \cdots & 1 & * & * \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & 1 & \cdots & * & * & * \\ 0 & 1 & * & \cdots & * & * & * \\ 1 & * & * & \cdots & * & * & * \end{array}}^{t-1 \text{ columns}} \right.$$

We now require the minimal array M_t to be such that the number of "comparison sites" between every two sequences is as small as possible. In other words, if $x_r^d = x_s^d = 0$ and $x_r^b = x_s^b = 1$ for some $r \neq s$, $1 \leqq r$, $s \leqq f(t)$, we will replace either x_s^d or x_s^b, or both, by * so long as the resulting array is still pairwise comparable and compatible.

LEMMA 2.4. *In a minimal array M_t, there exists some column which contains* *.

Proof. If the first column contains *, we are done. If not, we can assume that $x_1^d = 0$, $d = 1, \ldots, s$, and $x_1^d = 1$, $d = s+1, \ldots, t$. Let $1 < r \leqq f(t)$ and consider the r-th column. If again $x_r^d = 0$, $d = 1, \ldots, s$, and $x_r^d = 1$, $d = s+1, \ldots, t$, we can eliminate the r-th column and the resulting array is still pairwise comparable and compatible, and therefore M_t is not a minimal array. Suppose $x_r^d = 1$ for some $1 \leqq d \leqq s$; then x_r^d must be either * or 1 for all $s+1 \leqq d \leqq t$ or else we will have non-compatible sequences. Since the number of comparison sites between every two sequences has to be minimum, all the x_r^d's, $s+1 \leqq d \leqq t$, in fact have to be * because comparison sites already occur at the first column. The situation is similar if $x_r^d = 0$ for some $s+1 \leqq d \leqq t$.

THEOREM 2.5. $f(2t) > f(t)$.

Proof. We prove this by induction. $f(2) = 1 > 0 = f(1)$. Assume $f(t-1) < f(2t-2)$, but $f(2t) = f(t)$. From Theorems 2.1 and 2.2, we must have

$$f(2t-2) \geqq f(t-1) + 1$$

and

$$f(2t) = f(t) \leqq f(t-1) + 1,$$

and therefore

$$f(s) = f(t-1) + 1, \; t \leqq s \leqq 2t.$$

In particular,

$$f(t+1) = f(t-1) + 1.$$

Consider the minimal array M_{2t}, and suppose the r-th column contains at least one *. The total number of entries which are not * in this column therefore cannot be more than $2t - 1$. Without loss of generality, assume the number of 0's in this column is less than or equal to $t - 1$. If we now remove all rows in M_{2t} with 0 at the r-th position and also remove the r-th column, the resulting array has at least $t + 1$ rows and $f(t) - 1$ columns since $f(2t) = f(t)$. The $t + 1$ rows are still pairwise comparable and compatible, whence $f(t+1) \leqq f(t-1)$, contradicting

$$f(t+1) = f(t-1) + 1.$$

We can now make a rough estimate of $f(t)$. From Table 1 and Theorem 2.5, the best lower bound we can get is

$$f(6 \cdot 2^i) \geqq 4 + i, \quad i = 0, 1, 2, \ldots$$

Using the substitution $t = 6 \cdot 2^i$, we get

$$f(t) \geqq q(t), \quad t \geqq 6,$$

where

$$q(t) = 4 + \frac{\log t - \log 6}{\log 2}$$

which gives

$$t(l) \leqq 3 \cdot 2^{l-3}.$$

3. A graph structure on the minimum array. Given a minimal array M_t, define a graph G_s, for each $1 \leqq s \leqq f(t)$, on the vertex set $V = \{1, 2, \ldots, t\}$ by assigning an edge between vertices b and d, $b \neq d$, if and only if either $x_s^b = 0$ and $x_s^d = 1$, or $x_s^b = 1$ and $x_s^d = 0$. Let

$$A_s = \{b \mid 1 \leqq b \leqq t, \; x_s^b = 0\}$$

and

$$B_s = \{b | 1 \leqq b \leqq t, x_s^b = 1\}.$$

G_s is then a complete bipartite graph on the vertex sets A_s and B_s, and is non-empty by comparability and the minimality of M_t. We have the following observation.

LEMMA 3.1. *There do not exist* s *and* s', $1 \leqq s, s' \leqq f(t)$, *such that both*

$$A_s \cap B_{s'} \neq \emptyset \quad \textit{and} \quad A_{s'} \cap B_s \neq \emptyset.$$

Proof. Suppose there exist b and d such that

$$b \in A_s \cap B_{s'} \quad \text{and} \quad d \in A_{s'} \cap B_s.$$

Then

$$x_s^b = x_{s'}^d = 0 \quad \text{and} \quad x_{s'}^b = x_s^d = 1,$$

which implies x^b and x^d are not compatible sequences.

Now construct a graph G on the vertex set $V = \{1, 2, \ldots, t\}$ by assigning an edge between b and d if and only if x^b and x^d are comparable sequences. Since all the x^b's, $1 \leqq b \leqq t$, are pairwise comparable, G is a complete graph on V. Moreover, the G_s's, $1 \leqq s \leqq f(t)$ are a *minimal cover* of G, that is,

$$G = \bigcup_{s=1}^{f(t)} G_s$$

since every edge in G is also an edge in some G_s, and $f(t)$ is the minimum number of columns in M_t.

Let $\lambda_s = |A_s| \cdot |B_s|$, which gives the number of edges in the graph G_s. Suppose

$$\lambda = \lambda(t) = \max_{1 \leqq s \leqq f(t)} \lambda_s.$$

LEMMA 3.2. $f(t) \geqq \binom{t}{2} / \lambda$, *where* $\binom{t}{2}$ *is the binomial coefficient.*

Proof. Since G is a complete graph on a set of t vertices, there are $\binom{t}{2}$ edges in G. The minimal covering of G by all the G_s's implies

$$\binom{t}{2} \leqq \sum_{s=1}^{f(t)} \lambda_s < \lambda f(t).$$

LEMMA 3.3. *There does not exist* $1 \leqq s \leqq f(t)$ *such that* G_s *has an edge between two vertices in both* $A_{s'}$ *and* $B_{s'}$ *for all* $1 \leqq s' \leqq f(t)$.

Proof. If G_s has an edge in $A_{s'}$ and $B_{s'}$, then there exist b_1, b_2, d_1, d_2 such that $b_1, d_1 \in A_{s'}$ with $b_1 \in A_s$ and $d_1 \in B_s$ and $b_2, d_2 \in B_{s'}$ with $b_2 \in A_s$ and $d_2 \in B_s$. This implies

$$A_{s'} \cap B_s \neq \emptyset \quad \text{and} \quad A_s \cap B_{s'} \neq \emptyset.$$

In particular, let $s' = r$ where $\lambda = \lambda_r$ and assume without loss of generality that $|A_r| \geqq |B_r|$. Lemma 3.3 asserts that in the complete graph G, the edges between vertices in A_r and those in B_r are covered separately. We therefore have

LEMMA 3.4. $f(t) \geqq f(|A_r|) + f(|B_r|)$.

So far f is a function defined on the positive integers only. For convenience sake, extend f to a function $\tilde{f}$ defined on all nonnegative real numbers by the following:

$$\tilde{f}(t) = \begin{cases} f(t) & \text{if } t \text{ is an integer} \\ f(\lceil t \rceil) & \text{if } t \text{ is not an integer} \end{cases}$$

where $\lceil t \rceil$ is the smallest integer larger than or equal to t. Henceforth we will refer to $f(t)$ as a function defined on all $t \in [0, \infty)$ when we really mean $\tilde{f}(t)$.

LEMMA 3.5. $f(t) \geqq f(\sqrt{\lambda}) + f\left(\frac{\lambda}{t}\right)$.

Proof. We have

$$\lambda = \lambda_r = |A_r| \cdot |B_r| \leqq |A_r|^2,$$

or $|A_r| \geqq \sqrt{\lambda}$. Moreover,

$$|B_r| = \frac{\lambda}{|A_r|} \geqq \frac{\lambda}{t}.$$

We then have, from the last lemma and the monotonicity of f,

$$f(t) \geqq f(\sqrt{\lambda}) + f\left(\frac{\lambda}{t}\right).$$

COROLLARY 3.6. $f(t) \geqq \max\left(\frac{t(t-1)}{2}, f(\sqrt{\lambda}) + f\left(\frac{\lambda}{t}\right)\right)$.

This additional property of $f(t)$ helps establish a larger lower bound for it.

THEOREM 3.7. *There exists a constant* $0 < c_0 < 1$ *such that*

$$f(t) \geqq \exp\sqrt{c_0 \log(t)} \quad \textit{for } t \geqq a > 0.$$

Note. We prove the theorem by actually taking $c_0 = 0.71$. It can be shown that

$$q(t) \geqq \exp\sqrt{0.71 \log(t)} \quad \text{for } 6 \leqq t \leqq T_0,$$

where $q(t)$ is the bound in the last section and $T_0 = 208,562$ is the largest integer t such that

$$q(t) \geqq \exp\sqrt{0.71 \log(t)}$$

and hence

$$f(t) \geqq \exp\sqrt{0.71 \log(t)} \quad \text{for } 6 \leqq t \leqq T_0.$$

Proof of Theorem 3.7. We proceed by induction using Corollary 3.6. All we need show is

$$f(t) \geqq \exp\sqrt{0.71 \log(t)} \quad \text{for } t \geqq T_0 + 1.$$

Assume

$$f(s) \geqq \exp\sqrt{0.71 \log(s)}$$

for all $s \leqq t - 1$ where $t \geqq T_0 + 1$. If

$$\frac{t(t-1)}{2\lambda(t)} \geqq \exp\sqrt{0.71 \log(t)},$$

we are done. Otherwise

$$\lambda(t) > \frac{t(t-1)}{2\exp\sqrt{0.71 \log(t)}},$$

and hence

$$f(\sqrt{\lambda(t)}) + f\left(\frac{\lambda(t)}{t}\right) \geqq f\left(\sqrt{\frac{t(t-1)}{2\exp\sqrt{0.71 \log(t)}}}\right) + f\left(\frac{t-1}{2\exp\sqrt{0.71 \log(t)}}\right).$$

For convenience, let $u = \exp\sqrt{c_0 \log(t)}$ where $c_0 = 0.71$ and

$$G(t) = f\left(\sqrt{\frac{t(t-1)}{2u}}\right) + f\left(\frac{t-1}{2u}\right).$$

Also, let

$$g(t) = \frac{t(t-1)}{2u} \quad \text{and} \quad h(t) = \frac{g(t)}{t}.$$

Simple calculus shows that both $g(t)$ and $h(t)$ are increasing functions, in particular for $t \geqq 6$. Moreover, we must have

$$6 < \sqrt{g(t)} < t - 1 \quad \text{and} \quad 6 < h(t) < t - 1.$$

By the induction hypothesis,

$$G(t) \geqq \exp\sqrt{c_0 \log\sqrt{g(t)}} + \exp\sqrt{c_0 \log h(t)},$$
$$= \exp\left(c_0 \log t + \frac{c_0}{2}\beta(t)\right)^{1/2} + \exp(c_0 \log t + c_0\beta(t))^{1/2},$$

where

$$\beta(t) = \log\frac{t-1}{2t} - \log u.$$

Note that $(t-1)/2t$ is an increasing function of t, and larger than $1/e$ for $t \geqq T_0$. Hence

$$\beta(t) > -1 - \log u \quad \text{for } t \geqq T_0 + 1,$$

and therefore

$$G(t) \geqq \exp\left(c_0 \log t - \frac{c_0}{2}(1 + \log u)\right)^{1/2}$$
$$+ \exp(c_0 \log t - c_0(1 + \log u))^{1/2}$$
$$= \exp\left[\log u\left(1 - \frac{c_0}{2(\log u)^2}(1 + \log u)\right)^{1/2}\right]$$
$$+ \exp\left[\log u\left(1 - \frac{c_0}{(\log u)^2}(1 + \log u)\right)^{1/2}\right].$$

Since

$$\frac{c_0}{(\log u)^2}(1 + \log u) < 1,$$

$$\frac{1}{u}G(t) \geqq z^2 + z$$

where

$$z = z(t) = \exp\left[-\frac{c_0}{2}\left(1 + \frac{1}{\log u}\right)\right].$$

Note that for $t \geqq T_0 + 1$,

$$z \geqq \exp\left[-\frac{c_0}{2}\left(1 + \frac{1}{\sqrt{c_0 \log(T_0 + 1)}}\right)\right] > \frac{\sqrt{5} - 1}{2}$$

and therefore $z^2 + z > 1$. Hence $G(t) > u$, or

$$f(t) > \exp\sqrt{0.71 \log(t)}$$

for $t > T_0$ also.

The constant $c_0 = 0.71$ is almost the best possible value, as $T_0(0.72) =$ 132, 284, and in this case

$$z(T_0) < \frac{\sqrt{5} - 1}{2}.$$

With

$$l \geqq f(t) \geqq \exp\sqrt{0.71 \log(t)},$$

we get

$$t(l) \leqq l^{\log l/0.71},$$

and a comma-free dictionary will not have the maximum size $B_k(n)$ if

$$n > \left(\frac{k}{2}\right)^{(\log k/2)/0.71} + \frac{k}{2}, \quad k \geqq 8.$$

Table 3 compares Jiggs' bound and the new bound on n. Asymptotically, the new lower bound for n is significantly smaller. However, we suspect that compatibility is so strong a constraint that the bound on n could be dramatically reduced, probably to a polynomial in k.

TABLE 3

k	Jiggs' bound $2^{k/2} + k/2$	New bound $[(k/2)\exp(\log(k/2)/0.71) + k/2]$
8	20	18
10	37	43
20	1034	1760
30	3.28×10^4	3.06×10^4
40	1.05×10^6	3.09×10^5
80	1.10×10^{12}	2.11×10^8
160	1.21×10^{24}	5.57×10^{11}
320	1.46×10^{48}	5.69×10^{15}

4. A lower bound for $t(l)$. As before, let $t = t(l)$ be the maximum number of l-tuples of 0's, 1's, and *'s which are pairwise comparable and compatible. In the previous section we obtained the upper bound

$$t(l) \leqq l^{\log l/0.71} = e^{c\log^2 l}.$$

The lower bound which we found is

$$t(l) \geqq 15l + 1 \quad \text{for all } l \equiv 0 \pmod 7.$$

The basic construction here is for $l = 7$, with $t(l) = 16$.

0	0	0	0	0	0	0
1	0	0	*	0	*	*
*	1	0	0	*	0	*
*	*	1	0	0	*	0
0	*	*	1	0	0	*
*	0	*	*	1	0	0
0	*	0	*	*	1	0
0	0	*	0	*	*	1
0	*	*	1	*	1	1
1	0	*	*	1	*	1
1	1	0	*	*	1	*
*	1	1	0	*	*	1
1	*	1	1	0	*	*
*	1	*	1	1	0	*
*	*	1	*	1	1	0
1	1	1	1	1	1	1

It is no loss of generality to assume that the array A which achieves $t(l)$ rows with l columns includes an all-0's row, $\vec{0}$, and an all-1's row, $\vec{1}$. Let R denote the reduced $(t(l) - 2) \times l$ array when $\vec{0}$ and $\vec{1}$ and removed from A. Let Z be the $(t(l) - 2) \times l$ matrix of all 0's, and let J be the $(t(l) - 2) \times l$ matrix of all 1's. Then for any multiplicity m, the following array (Table 4), which is $(mt(l) - m + 1) \times (ml)$, clearly consists of rows which are pairwise comparable and compatible. This also yields the general result

$$t(ml) \geqq m(t(l) - 1) + 1,$$

for all $m \geqq 1$, $l \geqq 1$.

TABLE 4

$\vec{0}$	$\vec{0}$	$\vec{0}$	...	$\vec{0}$	$\vec{0}$	$\vec{0}$
Z	Z	Z	...	Z	Z	R
$\vec{0}$	$\vec{0}$	$\vec{0}$	...	$\vec{0}$	$\vec{0}$	$\vec{1}$
Z	Z	Z	...	Z	R	J
$\vec{0}$	$\vec{0}$	$\vec{0}$	...	$\vec{0}$	$\vec{1}$	$\vec{1}$
Z	Z	Z	...	R	J	J
$\vec{0}$	$\vec{0}$	$\vec{0}$	...	$\vec{1}$	$\vec{1}$	$\vec{1}$
...	...	...	...	...	...	...
$\vec{0}$	$\vec{0}$	$\vec{0}$	...	$\vec{1}$	$\vec{1}$	$\vec{1}$
Z	Z	R	...	J	J	J
$\vec{0}$	$\vec{0}$	$\vec{1}$	...	$\vec{1}$	$\vec{1}$	$\vec{1}$
Z	R	J	...	J	J	J
$\vec{0}$	$\vec{1}$	$\vec{1}$	...	$\vec{1}$	$\vec{1}$	$\vec{1}$
R	J	J	...	J	J	J
$\vec{1}$	$\vec{1}$	$\vec{1}$	...	$\vec{1}$	$\vec{1}$	$\vec{1}$

TABLE 5

The construction by Collins, Shor and Stembridge to show that $t(n^2 + n + 1) \geqq n(n^2 + n + 1) + 2$ for all positive integers n. (Use all cyclic shifts for each of the n words of length $l = n^2 + n + 1$, and adjoin the vectors $\vec{0}$ and $\vec{1}$ consisting of l 0's and of l 1's respectively to obtain the dictionary.)

$n = 1, n^2 + n + 1 = 3, t(3) = 5$

1	*	0

$n = 2, n^2 + n + 1 = 7, t(7) = 16$

1	0 0 *	0 * *
1	* 1 1	0 * *

$n = 3, n^2 + n + 1 = 13, t(13) \geqq 41$

1	0 0 0 *	0 0 * *	0 * * *
1	* * 1 1	0 0 * *	0 * * *
1	* * 1 1	* 1 1 1	0 * * *

$n = 4, n^2 + n + 1 = 21, t(21) \geqq 86$

1	0 0 0 0 0 *	0 0 0 0 * *	0 0 0 * * *	0 0 * * * *	0 * * * * *
1	* * * * 1 1	0 0 0 0 * *	0 0 0 * * *	0 0 * * * *	0 * * * * *
1	* * * * 1 1	* * * 1 1 1	0 0 0 * * *	0 0 * * * *	0 * * * * *
1	* * * * 1 1	* * * 1 1 1	* * 1 1 1 1	0 0 * * * *	0 * * * * *
1	* * * * 1 1	* * * 1 1 1	* * 1 1 1 1	* 1 1 1 1 1	0 * * * * *

$n = 5, n^2 + n + 1 = 31, t(31) \geqq 157$

1	0 0 0 0 0 0 *	0 0 0 0 0 * *	0 0 0 0 * * *	0 0 0 * * * *	0 0 * * * * *	0 * * * * * *
1	* * * * * 1 1	0 0 0 0 0 * *	0 0 0 0 * * *	0 0 0 * * * *	0 0 * * * * *	0 * * * * * *
1	* * * * * 1 1	* * * * 1 1 1	0 0 0 0 * * *	0 0 0 * * * *	0 0 * * * * *	0 * * * * * *
1	* * * * * 1 1	* * * * 1 1 1	* * * 1 1 1 1	0 0 0 * * * *	0 0 * * * * *	0 * * * * * *
1	* * * * * 1 1	* * * * 1 1 1	* * * 1 1 1 1	* * 1 1 1 1 1	0 0 * * * * *	0 * * * * * *
1	* * * * * 1 1	* * * * 1 1 1	* * * 1 1 1 1	* * 1 1 1 1 1	* 1 1 1 1 1 1	0 * * * * * *

5. Postscript. The results presented thus far were all obtained in time for inclusion in B. Tang's Ph.D. thesis in May, 1983. Several subsequent results on {0, 1, *}-sequences are presented in [7], and include the following:

i) A simpler proof of the upper bound formula,

$$t(l) < l^{c \log l},$$

attributed to C. L. M. van Pul;

ii) The constructions illustrating $t(1) = 2$, $t(3) = 5$, and $t(7) = 16$ have been generalized. Three students at Eindhoven (F. Abels, W. Janse, and J. Verbakel) found three words of length 13, all of whose cyclic shifts can be used simultaneously in a dictionary, along with the "all 0's" and "all 1's" words, to obtain $t(13) \geqq 41$. Three M.I.T. students (K. Collins, P. Shor, and J. Stembridge) found a general construction which yields

$$t(n^2 + n + 1) \geqq n(n^2 + n + 1) + 2$$

for all positive integers n, from which the lower bound result

$$t(l) > cl^{3/2}$$

clearly follows. This construction is illustrated for $1 \leqq n \leqq 5$ in Table 5.

The large gap which still remains between the upper and lower bound formulas is a clear invitation to further research.

References

1. F. H. C. Crick, J. S. Griffith and L. E. Orgel, *Codes without commas*, Proc. Nat. Acad. Sci. *43* (1957), 416-421.
2. W. L. Eastman, *On the construction of comma-free codes*, IEEE Trans. on Information Theory, *IT-11* (1965), 263-266.
3. S. W. Golomb, B. Gordon and L. R. Welch, *Comma-free codes*, Can. J. Math. *10* (1958), 202-209.
4. S. W. Golomb, L. R. Welch and M. Delbrück, *Construction and properties of comma-free codes*, Biol. Medd. Dan. Vid. Selsk. *23* (1958), 3-34.
5. B. H. Jiggs, *Recent results in comma-free codes*, Can. J. Math. *15* (1963), 178-187.
6. R. A. Scholtz, *Maximal and variable word-length comma-free codes*, IEEE Trans. on Information Theory, *IT-15* (1969), 300-306.
7. J. H. van Lint, {0, 1, *} *distance problems in combinatorics*, Surveys in Combinatorics (1985), London Mathematical Society Lecture Note Series *103* (Cambridge University Press, 1985), 113-135.

University of Southern California,
Los Angeles, California,
AT&T Bell Laboratories,
Murray Hill, New Jersey