ASPECTS OF SYNCHRONIZABLE CODING

L.J. Cummings
University of Waterloo

1. Introduction

Unlike classical algebraic coding theory which is concerned with the possibility of error in the transmission of single digits, synchronizable coding is concerned with "misframing"errors in the transmission of block codes. It is the purpose of synchronziable coding to construct codes which permit the correction of this kind of error.

In this survey we will consider only synchronizable codes with length n. The study of synchronizable codes with codewords of varying wordlengths is of separate interest in problems involving data compression, for example, but requires methods different from those discussed here.

To construct codes with synchronization properties, we must find a set of codewords which may initially allow "misframings" to occur but must detect the "misframing" in bounded time. The set of all English words with just three letters does not have this property, for example. Consider the message stream formed by repeating the codeword "EAT". If decoding were to begin with, say, the second letter transmitted, framings of length three would yield the incorrect message "ATE".

2. Synchronization

We will consider only those synchronizable codes which have the property that, for some positive integer d, knowledge of d consecutive digits of any encoded message is sufficient to establish the separation of codewords. These have also been called codes with bounded synchronization delay [17]. The integer d is called the synchronization delay of the code. It is difficult, in general, to determine the synchronization delay of a synchronizable code.

Let $\sum$ be a finite alphabet with σ elements and $\sum^n$ denote the words or strings of length n with letters taken from $\sum$. Given a word $w = a_1 \cdots a_n \epsilon \sum^n$ we obtain $n-1$ other words

$$a_2 \cdots a_n a_1, a_3 \cdots a_n a_1 a_2, \cdots, a_n a_1 \cdots a_{n-1} \tag{2.1}$$

by cyclic permutations of the entries of w. If w coincides with one of the words (2.1), say,

$$w = a_p \cdots a_n a_1 a_2 \cdots a_{p-1} \quad 1 < p \leq n$$

then w is said to be *periodic*. There are several interesting studies of periods in words [19,21], but our interest here is in aperiodic words. If w is not periodic then w is said to be *aperiodic* and the words (2.1) are distinct and form a *complete orbit* of the action of the cyclic group C_n<(12...n)> acting on the set $\sum^n$.

Proposition 2.1. No synchronizable code contains a cyclic permutation of one of its words.

JCMCC 1(1987), pp. 67–84

Proof. If $a_1 \cdots a_n$ is a code word then the message sequence

$$...a_1...a_i a_{i+1} \cdots a_n a_1 \cdots a_{i-1} a_i \cdots a_n a_1 \cdots a_{i-1} a_i \cdots a_n \cdots$$

could be decoded incorrectly if $a_i a_{i+1} \cdots a_{i-1}$ is also a code word. That is, there exists an infinite ambiguous message generated by any code which contains a word together with any of its cyclic permutations.

Proposition 2.2. No synchronizable code contains a periodic word.

Proof. Any periodic word is a cyclic permutation of itself.

Theorem 2.3. The number of $w(n,\sigma)$ of complete orbits under the action of C_n on $\sum^n$ is

$$w(n,\sigma) = \frac{1}{n}\sum_{d\,|n}\mu(n/d)\sigma^d. \tag{2.2}$$

Proof. Consider the set of all σ^n words of length n over the alphabet $\sum$. Each word has some period. (An aperiodic word has period n.) If $P_d(\sigma)$ is the number of words with period d then

$$\sigma^n = \sum_{d\,|n} P_d(\sigma).$$

Therefore,

$$P_n(\sigma) = \sum_{d\,|n}\mu(n/d)\sigma^d,$$

where μ is the Möbius function of elementary number theory. Since each complete orbit of C_n acting on $\sum^n$ has n words we conclude (2.2).

Example 2.4.

C_4 **acts on** $\{0,1\}^4$

0 0 0 0			
0 0 0 1	0 0 1 0	0 1 0 0	1 0 0 0
0 0 1 1	0 1 1 0	1 1 0 0	1 0 0 1
0 1 0 1	1 0 1 0		
0 1 1 1	1 1 1 0	1 1 0 1	1 0 1 1
1 1 1 1			

Each row of the above is an orbit of this action. We see that the orbit of 0101 consists of only two words, each an overlap of the other. Thus, neither of the words in this orbit can occur in a synchronizable code. The code of non-constant weakly increasing sequences of 0's and 1's of length obtained by choosing the first word in each of the aperiodic orbits in Example 2.4 is a synchronizable code. Indeed, an immediate corollary of Theorem 2.3 is:

Corollary 2.5. The maximum number of words in any synchronizable code in $\sum^n$ is $w(n,\sigma)$.

Proof. Since we can choose at most one word for a synchronizable code from each complete orbit of the action of C_n on $\sum^n$, we conclude the maximum number of of codewords in a synchronizable code is at most $w(n,\sigma)$.

Consider the code obtained by choosing from each complete orbit of the action of C_n on $\sum^n$ that word which is least in the lexicographical ordering of $\sum^n$ induced by a fixed ordering of $\sum$. It was first proved by Golomb and Gordon [17] that this code is synchronizable.

It has been shown that the synchronization delay d of a synchronizable code with $w(n,\sigma)$ words is bounded above by $2(n-1)$ when n is odd and $(n/2)\sigma^{n/2}$ when n is even [17].

With the notable exception of the code mentioned in the proof of Corollary 2.5, relatively few constructive techniques are known for maximal synchronizable codes. We turn now to comma free codes, the class of synchronizable codes which has been most widely studied and for which the synchronization delay and some constructive techniques are known.

3. Comma-Free Codes

One solution to the synchronization problem is to construct codes which do not contain "overlaps" of two codewords. Thus, no "misframing" will be a codeword and no confusion can result in a noiseless channel. Such a code is called comma-free.

Definition 3.1. Let $\sum$ be a finite set of σ elements called the alphabet. $\sum^n$ is the set of all words of length n with symbols from $\sum$. A subset $C \subset \sum^n$ is a comma-free code with block length n if

$$a_1 \cdots a_n,\, b_1 \cdots b_n \in C$$

always implies that

$$a_2 \cdots a_n b_1,\, a_3 \cdots a_n b_1 b_2,\, \cdots,\, a_n b_1 \cdots b_{n-1} \notin C. \tag{3.1}$$

The words (3.1) are called *overlaps*. An alternate definition which is applicable to variable word length codes as well is given by:

Definition 3.1'. If u and v are any words then a word $w = u_2 v_1$ is an *overlap* of u and v provided there exist non-empty words $u_i, v_i; i = 1,2$ such that $u = u_1 u_2$ and $v = v_1 v_2$. A code C is *comma-free* if for any two words $u = u_1 u_2$ and $v = v_1 v_2$ with $u_i, v_i; i = 1,2$ nonempty it follows that uv is not in C.

In any comma-free code with block length u the synchronization d delay will be at most 2n-1. That is, after at most d symbols have been received from any message stream, synchronization can be established. This is almost obvious since a complete word of the code must have been received after 2n-1 symbols of any message stream has been received.

Let $CF(n,\sigma)$ denote the set of comma-free codes over an alphabet of σ symbols with block length n. A code in this set is said to be a $CF(n,\sigma)$ code. The maximum number of words in any $CF(n,\sigma)$ code will be denoted by $cf(n,\sigma)$. In this more restricted class of codes we shall see that Corollary 2.5 no longer holds but (2.2) is still an upper bound.

Lemma 3.2.

$$cf(n,\sigma) \leq w(n,\sigma). \tag{3.2}$$

Proof. See Proposition 2.1 and Theorem 2.5.

In 1958 Golomb, Gordon, and Welch [15] conjectured that equality holds in Lemma 3.2 whenever n is odd. In 1965, W. Eastman resolved this issue by giving an construction which produced a comma-free code with $w(n,\sigma)$ words for every odd word length n. We will not discuss Eastman's somewhat awkward construction, but rather an easily implemented algorithm which was published by R.A. Scholtz [28] a year after Eastman's paper appeared.

Scholtz's algorithm is recursive. Define a sequence of sets X_i as follows:

$$X_0 = \sum \tag{3.3}$$
$$X_{i+1} = x_i^*(X_i - x_i)$$

where the words x_i are chosen from X_i with the additional requirement that they have non-decreasing odd word length. (Here the notation x_i^* denotes the set $\{0,x,xx,xxx,...\}$. The empty word is 0. The 'multiplication' indicated in (3.3) is concatenation. This means that the preceding set with x_i deleted is contained in X_{i+1}. Except for X_0, the sets X_i are infinite. To obtain finite codes let $S = \cup X_i$. Then, for odd n, it can be shown that

$$C = S \cap \sum^n$$

is a comma-free code with $w(n,\sigma)$ words [28]. An example of Scholtz's algorithm is given in Example 3.3 with initial alphabet size 3 and the sets X_i being truncated at length 4. The column of words of length 3 is a maximal comma-free code with $w(n,\sigma) = 8$ words.

Example 3.3

length:	1	2	3	4
Y_0:	0 1 2			
		x_0=0		
Y_1:	1 2	01 02	001 002	0001 0002
		x_1=1		
Y_2:	2	12	112 101 102	1112 1101 1102 1001 1002
		x_2=2		
Y_3:			201 202 212	2201 2202 2212 2001 2002 2112 2101 2102

Here $Y_0 = X_0$ and $Y_i = X_i - X_{i+1}$ for $i > 0$. If the algorithm is continued then the next choice, x_4 must be one of the length 3 words.

4. Comma-Free Codes with Even Word Length

In section 3 we saw that for comma-free codes with odd word length there are constructive techniques which produce maximal $CF(n,\sigma)$ codes with $w(n,\sigma)$ words. While $w(n,\sigma)$ remains an upper bound for the number of words in a comma-free code with even word length, we shall see in Theorem 4.2 that, for a sufficiently large alphabet, no maximal comma-free code with even word length has $w(n,\sigma)$ words. There are no general techniques for the construction of maximal comma-free codes with even word length except for $n = 2$.

Golomb, Gordon, and Welch [15] first proved that $cf(2,\sigma) = \left[\frac{\sigma^2}{3}\right]$. With A. Ball we found all inequivalent maximal $CF(2,\sigma)$ comma-free codes. We call two codes equivalent *equivalent* if one can be obtained from the other by a permutation of the alphabet $\sum$ and/or a simultaneous reversal of all code words. The inequivalent $CF(2,\sigma)$ codes are easily described graphically as subgraphs of $\overline{K}_\sigma$, the complete directed graph on σ vertices. Here, the vertex set of $\overline{K}_\sigma$ is taken to be $\sum = \{0,1,...,\sigma-1\}$. The pairs xy with $x,y \in \sum$ will be taken as the edges of the graph $\overline{K}_\sigma$. Thus, $CF(2,\sigma)$ codes correspond to sets of edges in $\overline{K}_\sigma$.

The defining condition of a comma-free code is that no overlaps of codewords may appear in the code. This condition has an easy interpretation in $\overline{K}_\sigma$: A collection of edges in $\overline{K}_\sigma$ represents a $CF(2,\sigma)$ code if and only if it does not contain 3 edges which form a directed path of length 3. (The middle edge would represent an overlap of the first and third codewords.) Of course, a comma-free code corresponds to an asymmetric digraph; i.e., ab and ba cannot both be edges.

Define the subgraph $G_1(\sigma)$ of $\overline{K}_\sigma$ by the edges

$$xy \equiv 01{,}02{,}or\,12 \ (mod 3)$$

where, by an abuse of notation, $xy \equiv ab$ if $x \equiv a$ and $y \equiv b$ $(mod 3)$. It is not difficult to see that $\{01,02,12\}$ is a maximal $CF(2,3)$ code. Similarly, define a subgraph $G_2(\sigma)$ of $\overline{K}_\sigma$ by edges

$$xy \equiv 10{,}02{,}or\,12 \ (mod 3)$$

and a subgraph $G_3(\sigma)$ by edges

$$xy \equiv 01{,}02{,}or\,21 \ (mod 3).$$

Theorem 4.1. An asymmetric digraph without loops or multiple edges which contains the maximal number of edges with respect to the property of containing no directed path of length 3 is isomorphic to

$G_1(\sigma)$, if $\sigma = 0 \ (mod 3)$

$G_1(\sigma)$, $G_2(\sigma)$ or $G_1(\sigma)'$, if $\sigma = 1 \ (mod 3)$

$G_1(\sigma)$, $G_3(\sigma)$ or $G_1(\sigma)'$, if $\sigma = 2 \ (mod 3)$.

Here, $G_1(\sigma)'$ is the reversal of $G_1(\sigma)$.

A proof of this theorem may be found in [4]. Representatives of these codes for $\sigma = 3,4,5$ and 6 are given in Figure 4.1. Figure 4.2 illustrates embeddings between these digraphs and hence embeddings of the corresponding maximal $CF(2,\sigma)$ codes into maximal $CF(2,\tau)$ codes with alphabet size $\tau > \sigma$.

There are always $q+1$ distinct embeddings corresponding to each of the lines. It is known that the digraphs $G_1(3q)$,$G_2(3q+1)$, and $G_3(3q+2)$ are isomorphic to their reversals.

We turn now to $CF(2k,\sigma)$ codes with $k > 1$ where known results are much less comprehensive.

In example 3.3 the 15 words of length 4 form a comma-free code but the value of $w(4,3)$ is 18. Indeed, an early result due to Jewett [20] states that

Theorem 4.2. If $n = 2m$ then $cf(n,\sigma) < w(n,\sigma)$ whenever

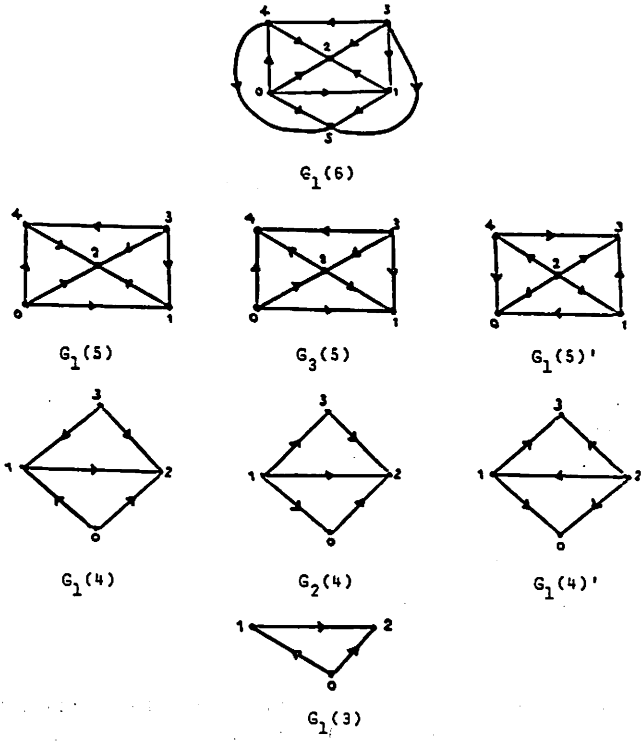
$$\sigma > 2^m + m. \tag{4.1}$$

A backtracking program reported in [26] has shown that a maximum $CF(4,4)$ with $57 < w(4,4)$ words exists.

It has been known for some time that for words of length 2

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

where $[x]$ denotes the integral part of x [14]. All comma-free codes with words of length 2 were first constructively classified in [3].



$G_1(6)$

$G_1(5)$ $G_3(5)$ $G_1(5)'$

$G_1(4)$ $G_2(4)$ $G_1(4)'$

$G_1(3)$

Figure 4.1

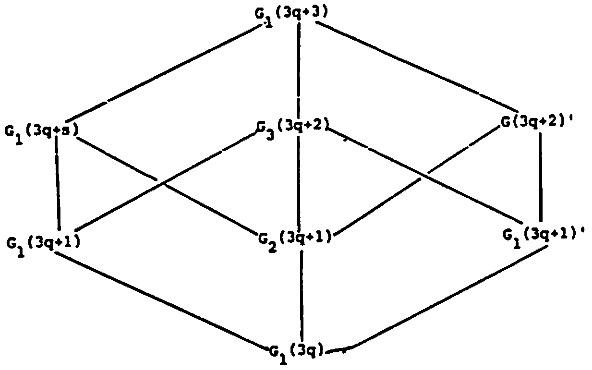


FIGURE 4.2

There are always $q+1$ distinct embeddings corresponding to each of the lines. It is known that the digraphs $G_1(3q)$,$G_2(3q+1)$, and $G_3(3q+2)$ are isomorphic to their reversals.

If we consider only the binary alphabet maximal then $CF(n,2)$ codes have been constructed for $n = 2,4,6,8,10$ [15,26]. Niho [26] constructed a maximal $CF(10,2)$ with 99 words by a program which involved extensive backtracking. It happens that $w(10,2) = 99$. Similarly, the bound (2.2) is met by the constructions for $n = 2,4,6,8$. This naturally suggests

Conjecture 4.2. (Niho) $cf(2n,2) = w(2n,2)$.

Important recent work on the construction of comma-free codes with even word length is due to Golomb, Graham and Tang [18]. They improved (4.1) to

$$\sigma - m > m^{\log m)/0.71}.$$

If we measure the density of codewords in comma-free codes by the limit

$$a_n = \lim_{n \to \infty} \frac{cf(n,\sigma)}{\sigma^n}$$

then, since the right-hand side of (2.2) is asymptotic to σ^n/n , we have that $a_n \leq 1/n$ whenever n is odd. On the other hand, Golomb, Gordon, and Welch proved that $a_n \geq 1/n$ [15]. For n even they proved:

Theorem 4.3. If $n = 2m$ and $2 < m$ then

$$1/ne < a_n < 1/n,$$

where $e = 2.71828$ ... is the base of natural logarithms.

5. Cosets of Linear Codes

In a *linear code* the codewords form a subspace of the finite vector space of n-tuples over $GF(q)$ for some prime power $q = p^m$. Codes with extremely good error-correcting properties, such as the BCH codes, are linear codes. Clearly no synchronizable code can be linear since the n-tuple 0...0 is periodic. This obvious fact does not end the matter, however. Synchronizable codes can be found as cosets (i.e., translates by a single vector) of linear codes. This is particularly convenient because the error-correcting properties of a linear code are inherited by each of its cosets. Finding synchronizable codes as cosets of linear codes provides codes with both error-correcting and synchronization properties. This was realized quite early and word synchronization was established in the 1969 Mariner IV Mars mission with a coset of the (32,6) biorthogonal Reed-Muller code [36].

J.J. Stiffler [29] first studied the error-correcting properties of comma-free codes. He proved that only the trivial (3,1) Hamming code can have a comma-free coset. Figure 5.1 shows why the coset of the Hamming (7,4) code resulting from translation by the vector 0000001 is not comma-free. The indicated matrices, H_3 and G_3, are the parity-check matrix and the generator matrix respectively of the (7,4) code. Hamming further showed that for linear codes which are cyclic of dimension k, there is a comma-free coset if and only if $k \leq (n-1)/2$ [29]. The question arises whether the comma-free codes obtained in this way are maximal. The answer to this question is essentially negative:

HAMMING (7,4) CODE

$$H_3 = \begin{bmatrix} 1&1&0&1&1&0&0\\ 1&0&1&1&0&1&0\\ 0&1&1&1&0&0&1 \end{bmatrix}$$

$$\begin{array}{ll} & \text{code has } 2^4 = 16 \text{ words} \\ G_3 = \begin{bmatrix} 1&1&0&1&0&0&0\\ 1&0&1&0&1&0&0\\ 0&1&1&0&0&1&0\\ 1&1&1&0&0&0&1 \end{bmatrix} & \begin{matrix} 1&1&0&1&0&0&1\\ 1&0&1&0&1&0&1\\ 0&1&1&0&0&1&1\\ 1&1&1&0&0&0&0 \end{matrix} \\ \begin{matrix} 0&1&1&1&1&0&0\\ 1&0&1&1&0&1&0\\ 0&0&1&1&0&0&1\\ 1&1&0&0&1&1&0\\ 0&1&0&0&1&0&1\\ 1&0&0&0&0&1&1\\ 0&0&0&1&1&1&0\\ 0&1&0&1&0&1&1\\ 1&0&0&1&1&0&1\\ 0&0&1&0&1&1&1\\ 1&1&1&1&1&1&1\\ 0&0&0&0&0&0&0 \end{matrix} & \begin{matrix} 0&1&1&1&1&0&1\\ 1&0&1&1&0&1&1\\ 0&0&1&1&0&0&0\\ 1&1&0&0&1&1&1\\ 0&1&0&0&1&0&0\\ 1&0&0&0&0&1&0\\ 0&0&0&1&1&1&1\\ 0&1&0&1&0&1&0\\ 1&0&0&1&1&0&0\\ 0&0&1&0&1&1&0\\ 1&1&1&1&1&1&0\\ 0&0&0&0&0&0&1 \end{matrix} \end{array}$$

$$\text{overlap in coset: } \overbrace{1\,0\,0\,0\,0\,1\,0}\,\overbrace{1\,0\,1\,0\,1\,0\,1}$$

$$\underbrace{0\,1\,0\,1\,0\,1\,0}$$

FIGURE 5.1

Theorem 5.1. Any comma-free code C with word length n over $\sum$ which has

$$w(n,\sigma) = \frac{1}{n}\sum_{d|n}\mu(n/d)\sigma^d$$

words is a coset of a linear code iff

$$(n,\sigma)=(2,2) \text{ and } C=\{01\} \text{ or } C=\{10\}$$

or

$$(n,\sigma)=(3,2) \text{ and } C \text{ is one of}$$

$$\{001,011\},\{010,011\},\{100,101\}$$
$$\{001,101\},\{010,110\},\{100,110\}$$
$$\{001,110\},\{100,011\}.$$

The proof of Theorem 5.1 [8] follows easily from a number theoretic result of independent interest proved by V. Kumar and Ram Murty:

Lemma 5.2. For positive integers σ,n

$$w(n,\sigma)=\frac{1}{n}\sum_{d\,|n}\mu(n/d)\sigma^{d}\,|\,\sigma^{n}$$

if and only if

$$(n,\sigma)=(2,3),(3,2),or(2,2).$$

In practice, large comma-free codes are found as cosets of powerful error-correcting codes, as indicated by Stiffler's result, and are useful in applications in spite of Theorem 5.1.

6. Isomorphism Classes of Comma-Free Codes.

The isomorphism classes of $CF(2,\sigma)$ codes are determined by representing the codes as sets of directed edges in the digraph $\overline{K}_n$ and applying Theorem 4.1. Curiously, the isomorphism classes of $CF(3,\sigma)$ codes were studied first because they figured in early theories of the genetic code. The first work concerning comma-free codes was published in 1957 [6] by a group at the Cavendish laboratory in England. At that time it was thought genetic coding should involve some synchronization device. It was known experimentally that there were 4 bases or nucleotides appearing in the DNA molecule namely, adenine, thymine, guanine, and cytosine. The Cavendish group, which included F.H.C. Crick, proposed the comma-free hypothesis in opposition to the "overlapping" codes notion which had been suggested in Nature by George Gamow [14]. They speculated that amino acids were coded by triplets of the bases. Having defined comma-free codes, they determined by elementary arguments that a maximal such code with words of length 3 using 4 symbols must have exactly 20 words, which was precisely the number of amino acids. They also constructed a family of maximal $CF(3,\sigma)$ codes for each σ as follows:

$$0\ 1\ \begin{matrix}0\\1\end{matrix} \qquad \begin{matrix}0\\1\end{matrix}\ 2\ \begin{matrix}0\\1\\2\end{matrix}\ \ldots\ \begin{matrix}0\\ \cdot\\ \cdot\\ \cdot\\ \cdot\\ \sigma-2\end{matrix}\ \sigma-1\ \begin{matrix}0\\ \cdot\\ \cdot\\ \cdot\\ \cdot\\ \sigma-1\end{matrix} \tag{6.1}$$

where say, the configuration

$$\begin{array}{ccc} & & 0\\ 0 & 2 & 1\\ 1 & & 2\end{array}$$

represents the six codewords

020 021 022 120 121 122.

We now know that the genetic code is "degenerate". It uses all 64 of the triplets on 4 symbols and different triplets yield the same amino acid with one triplet, called the terminator, acting as a comma. This does not end the matter, however, since some biologists argue that at one time in its evolutionary history the genetic code was comma-free. Clearly there are

$$\sum_{i=1}^{\sigma-1} i(i+1) = \frac{\sigma^3-\sigma}{3} = w(3,\sigma)$$

codewords in (6.1) and so it represents a maximal $CF(3,\sigma)$ code provided it can be shown to have the comma-free property. This follows easily because of the structure of suffixes and prefixes of (6.1).

All isomorphism classes of maximal $CF(3,4)$ codes with 20 words were determined by Golomb, Welch, Gordon, and Delbruck [16]. In Figure 6.1, a representative for each of the 5 isomorphism classes of maximal $CF(3,4)$ codes is given with the number of distinct codes of each class noted at the bottom.

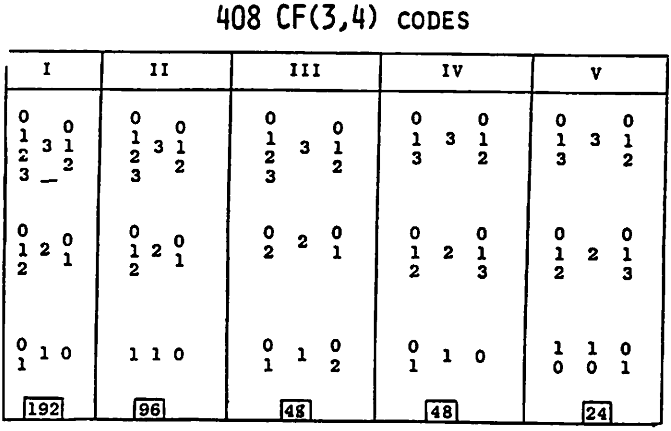


Figure 6.1

The isomorphism clases of $CF(3,3)$ codes in Figure 6.2 were determined by A. Ball [2].

In his thesis at the University of Waterloo [2] A. Ball proved that every maximal $CF(3,\sigma)$ code is either an extension of a maximal $CF(3,\sigma-1)$ or a maximal $CF(3,\sigma-2)$ code. (See also [1].) If T is the number of isomorphism classes of maximal $CF(3,\sigma)$ codes then Ball's proof shows that

$$T_\sigma = T_{\sigma-1} + T_{\sigma-2} \qquad (6.2)$$

with initial values (from Figures 6.1 & 6.2) $T_3 = 3$ and $T_4 = 5$.

Since the recurrence relation (6.2) yields a Fibonacci sequence it follows that:

42 CF(3,3) CODES

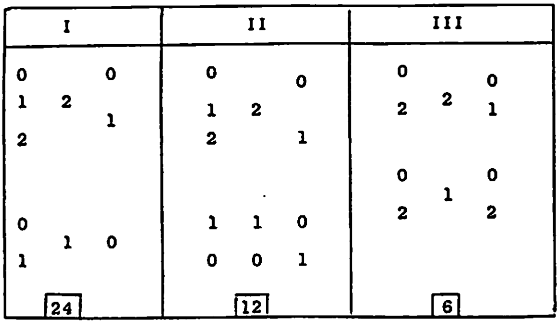


Figure 6.2

Theorem 6.1 (Ball). The number of isomorphism classes of maximum $CF(3,\sigma)$ codes is:

$$T_\sigma = \frac{(-1)^{\sigma+1}}{\sqrt{5}}\left[\left(\frac{-1+\sqrt{5}}{2}\right)^{\sigma+1} - \left(\frac{-1-\sqrt{5}}{2}\right)^{\sigma+1}\right] \quad 3 \leq \sigma.$$

7. Synchronizable Codes in the DeBruijn Graph

The de Bruijn graph, $G_{n,\sigma}$, is the directed graph whose vertices are the σ^n words of length $n > 1$ over the alphabet $\sum = \{0,\ldots,\sigma-1\}$. There is a directed edge from $\bar{a} = a_1 \cdots a_n$ to $\bar{b} = b_1 \cdots b_n$ in $G_{n,\sigma}$, precisely when $a_2 \cdots a_n = b_1 \cdots b_{n-1}$. The edge may be labelled by $a_1 \cdots a_n b_1$. The de Bruijn graph is thus regular with indegree and outdegree equal to σ at every vertex. It contains σ^{n+1} directed edges among which are σ loops at each vertex $a^n = a \cdots a, a \in \sum$. Figure 7.1 shows $G_{3,2}$.

$G_{3,2}$:

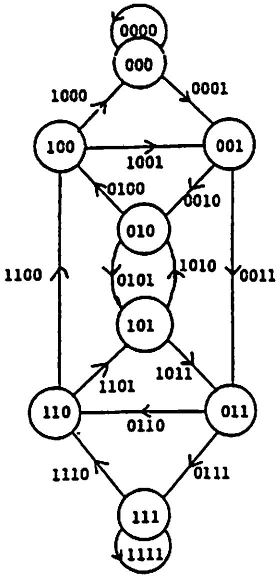


FIGURE 7.1

It was established in [25] that the only planar de Bruijn graphs are:

$$G_{n,1}\ G_{1,2}\ G_{1,3}\ G_{1,4}\ G_{2,2}\ G_{2,3}\ G_{3,2}.$$

For arbitrary σ, T. van Aardenne-Ehrenfest and de Bruijn [10] proved that if $n > 1$ then $G_{n-1,\sigma}$ contains $(\sigma)^{\sigma^{n-1}}\sigma^{-n}$ Euler circuits. Each Euler circuit in $G_{n-1,\sigma}$ determines a circular word of length σ^n over $\sum$ with the property that all the σ^n words of length n appear as subwords precisely once. In $G_{1,2}$ there is only one Euler circuit which determines the circular word 00110. In $G_{2,2}$ there are two Euler circuits corresponding to the circular words 0001011100 and 0001110100. Over the binary alphabet there are $2^{2^{n-1}}2^{-n}$ Euler circuits in $G_{n-1,2}$ and these determine all circular words with the aforementioned property. Although the binary result is usually attributed to de Bruijn, according to de Bruijn's pamphlet [12], R.P. Stanley discovered that the problem of constructing such circular words had been proposed in 1894 by A. de Riviere and in the same year a solution was given by C. Flye Sainte-Marie in the French problem journal "l'Intermédiaire des Mathématiciens".

The existence of circular words over $\sum$ in which all σ^n words of length n appear was first shown by M.H. Martin [24]. Lempel [22,23] has extended this work.

Any code over $\sum$ with block length n may be represented as a set of vertices in $G_{n,\sigma}$ or, equivalently, as a set of edges in $G_{n-1,\sigma}$, whenever $1 < n$. For example, with the binary alphabet, the synchronizable code mentioned after Corollary 2.5 has, for each n, a striking representation as a collection of edges in $G_{n-1,\sigma}$, as we will see in Theorem 7.2 below.

Definition 7.1. The canonical synchronizable code, $\Lambda(n,\sigma)$ is the set of words of length n which are lexicographically least in the orbits containing aperiodic words in $\sum^n$.

Golomb and Gordon [17] have shown that, for every n and σ, $\Lambda(n,\sigma)$ is a synchronizable code with bounded synchronizable delay.

Theorem 7.2. If the binary canonical code, $\Lambda(n,\sigma)$, is considered as a collection of edges in $G_{n,\sigma}$ then it is a union of disjoint paths.

The proof of Theorem 7.2. appears in [10]. Figure 7.2 shows $\Lambda(7,2)$ as a subgraph of $G_{6,2}$ with the vertex labels suppressed.

For alphabets with $\sigma > 2$ the subgraphs of $G_{n-1,\sigma}$ determined by the canonical code $\Lambda(n,\sigma)$ are not easily characterized. Figure 7.3 shows the subgraph of $G_{3,3}$ determined by $\Lambda(4,3)$.

We offer the following theorem as a first step toward understanding the structure of these graphs.

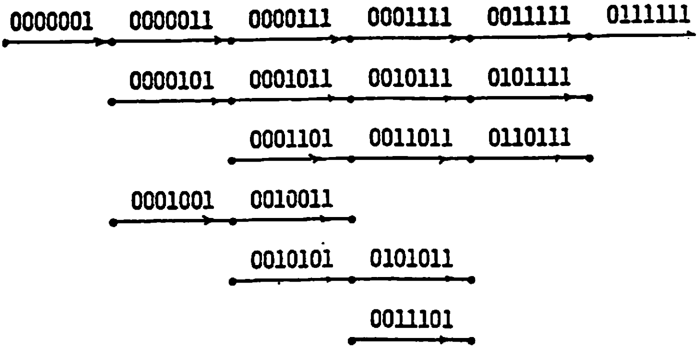


FIGURE 7.2

Theorem 7.3. If $n > 1$ and $\sigma > 2$ then the subgraph $G_{n-1,\sigma}$ determined by edge labels in $\Lambda(n,\sigma)$ contains a unique path, P, of maximal length $(\sigma-1)(n-1)$. For each $n > 1$ the set of vertices $G_{n,\sigma}$ determining P_{n+1} is $P_n \cap \{a^n : a \in \sum\}$.

Proof. P_n' is the path

$$a_1^{n-1}a_2, a_n^{n-2}, \dots, a_1a_2^{n-1}, a_1a_2^{n-1}, a_2^{n-1}a_3, \dots, a_{\sigma-1}a_\sigma^{n-1}.$$

Since the words of P_n with first letter a_i, $i = 1, \dots, \sigma-1$ form a set of $n-1$ words, P_n has $(\sigma-1)(n-1)$ edges. The uniqueness of P_n is established by induction on n.

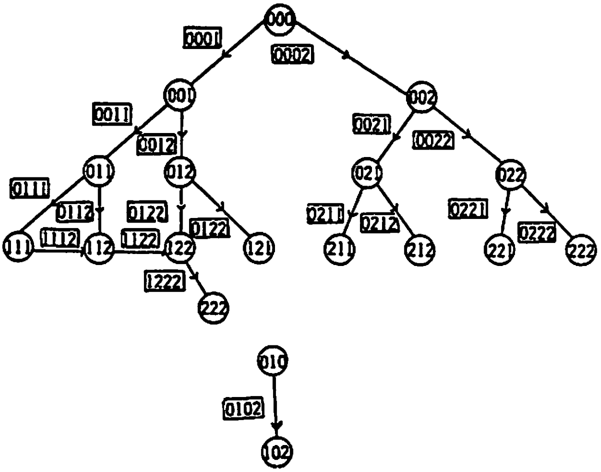


FIGURE 7.3

We give the proof for $n = 2$ and suppress the induction step. Assume

$$b_1b_2, b_2b_3, \dots, b_{\sigma-1}b_\sigma, b_{\sigma-1}b_\sigma, b_\sigma\, b_{\sigma+1}, \dots$$

is a path of length σ or greater in $\Lambda(2,\sigma)$. Since $b_ib_j \in \Lambda(2,\sigma)$ we have

$$b_1 < b_2 < \cdots b_\sigma < b_{\sigma+1}$$

which is impossible since $\sum$ has only σ elements.

Any word of weight 2 in $\Lambda(n,2)$ has the form 0^p10^q1 where $p+q = n-2$ and $p > q \geq 0$.

The number of such words is $\left\lceil\frac{n-2}{2}\right\rceil$ where $\lceil\cdot\rceil$ is the ceiling function. This provided a crude lower bound for the paths determined by $\Lambda(n,2)$.

Problem 7.6. Determine the number of paths and the cardinality of the weight classes in $\Lambda(n,2)$.

The condition of comma-freedom naturally imposes a more restricted structure on the corresponding subgraphs of G_n,σ. We restate here two important theorems about maximal $CF(n,\sigma)$ codes, both due to Golomb and Welch [15] as results about edges in $G_{n-1,\sigma}$.

Definition 7.3. By a bipartite subgraph of G_n,σ we shall mean a collection of edges in $G_{n,\sigma}$ not containing a directed path of length 2.

The example of a maximum $CF(5,2)$ code considered as a set of edges given in Figure 7.4 is a bipartite subgraph of $G_{4,2}$.

Theorem 7.4 (Golomb,Gordon) In $G_{2,\sigma}$ with $2 < \sigma$, a collection of $cf(3,\sigma) = (\sigma^3-\sigma)3$ edges is a maximum $CF(3,\sigma)$ code if and only if it is a bipartite subgraph of $G_{2,\sigma}$.

Proof. If $a = a_1a_2a_3$ and $b = b_1b_2b_3$ are edges of the collection then neither of the edges labelled $a_2a_3b_1$ and $a_3b_1b_2$ can appear in the collection. Since the number of edges in the collection is the upper bound (3.2), the edges form a maximum $CF(3,\sigma)$ code.

We show the converse by contradiction. Suppose vertices abc and bcd are codewords. The code would contain an overlap if it contained vertices with a label ending in a or beginning with d. In particular, a and d must be distinct. From word complete orbit

$$aax \quad axa \quad xaa \qquad x \neq a \tag{7.1}$$

only aax can appear in the code. Similarly only ydd can appear in the code from the complete orbit

$$ddy \quad dyd \quad ydd \qquad y \neq d. \tag{7.2}$$

Since $2 < \sigma$ choose $z \epsilon \sum$ distinct from a and b and consider the complete orbit

$$adz \quad dza \quad zad. \tag{7.3}$$

If $x = d$ in (7.1) and $y = z$ in (7.2) then edge adz is an overlap of aad and zdd. Therefore adz is not a vertex of the code. Similarly, zad is an overlap of aaz and add. Finally, dza cannot be a codeword because its label ends in a. We conclude that the complete orbit (7.3) contains no codeword. Therefore the code cannot be a maximum $CF(3,\sigma)$ code.

For $\sigma = 2$ Theorem 7.5 is not true. All nine choices of a pair of vertices, one from each of complete orbits of 001 and 011, form $CF(3,2)$ codes with the exception of the pair {010,101}. Of the eight remaining pairs, only the pairs {001,011} and {100,110} are not bipartite subgraphs of $G_{2,2}$.

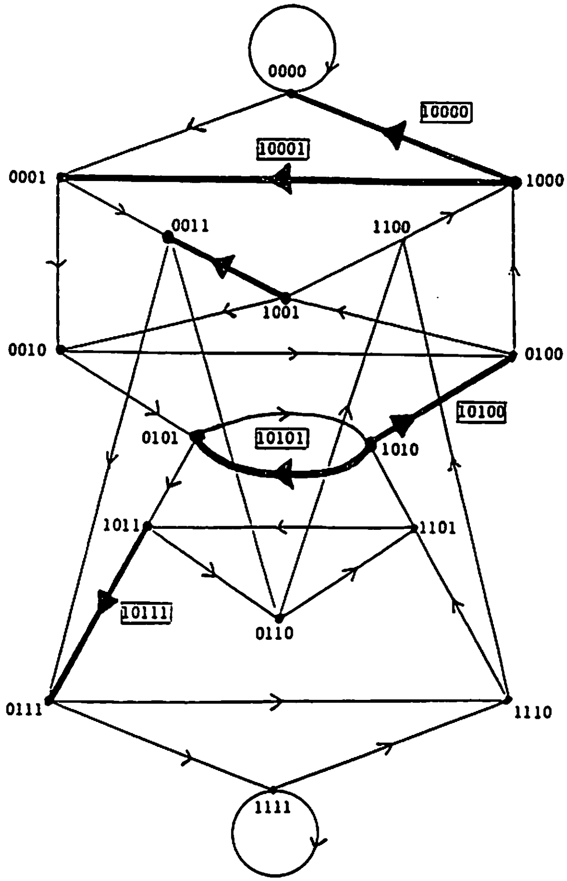


FIGURE 7.4

The theorem is not true in $G_{3,2}$ either since the directed edges in the path of Figure 7.4 are a maximum $CF(4,2)$ code.

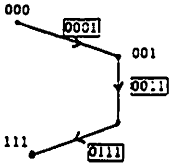


Figure 7.4

Theorem 7.6. (Golomb,Gordon) Let $n \geq 5$ be an odd integer. If a collection of edges C in $G_{n-1,\sigma}$ is a comma-free code with $w(n,\sigma)$ words then C is a bipartite subgraph of $G_{n-1,\sigma}$.

Proof. Suppose two edges of C form a directed path of length 2. Then by definition there are symbols $a,b \in \sum$ such that the edges are aw and wb where w has length $n-1$. Since C is a comma-free code no edge of C can terminate in a nor begin with b. Because C is comma-free with $w(n,\sigma)$ words it must contain precisely one edge from every complete orbit. In particular, only the edges with the following labels can be chosen from the complete orbits which contain them:

$$\begin{array}{c} ab.....b \\ aab...b \\ \cdot \\ \cdot \\ \cdot \\ aa.....ab \end{array}$$

Now consider the complete orbit containing $abab...b$. The only edges of this orbit ending with b are $abab...b$ and $ab...bab$. But C cannot contain $abab...b$ since it is an overlap of $a...ab$ and $ab...b$. The edge $ab...bab$ is excluded similarly. This means that C can contain no edge from the complete orbit containing $abab...b$ and therefore does not have $w(n,\sigma)$ words.

Corollary 7.7. If $5 \leq n$ and $2 < \sigma$ then no maximum $CF(n,\sigma)$ code can contain the directed path of $n-1$ edges between $a...ab$ and $ab...b$ where $a,b \in \sum$ and $a \neq b$.

We have been able to extend Theorem 7.6 to even word length [9]. But in view of Jewett's result in Theorem 4.2 our extension can only apply to finitely many cases for each even n.

Theorem 7.8. If $n > 3$ and $\sigma > 2$ then any collection of $w(n,\sigma)$ edges in $G_{n-1,\sigma}$ which form a comma-free code is a bipartite subgraph.

Figure 7.5 gives an example of a bipartite subgraph of $G_{4,2}$ with $w(5,2)$ edges which does not determine a $CF(n,\sigma)$ code. Obviously, an additional condition is needed to characterize all maximal $CF(n,\sigma)$ codes.

8. Summary

This brief survey is intended to convey a flavour of the subject and some of the directions it has taken. Much work remains for the interested reader which can only be hinted at here. For comma-free codes, finding maximal codes of even word length remains the premier task. Ball [2] has given an algorithm for constructing all isomorphism classes of odd word length maximal comma-free codes, but the algorithm requires sorting partitions and is useful only for relatively short word lengths.

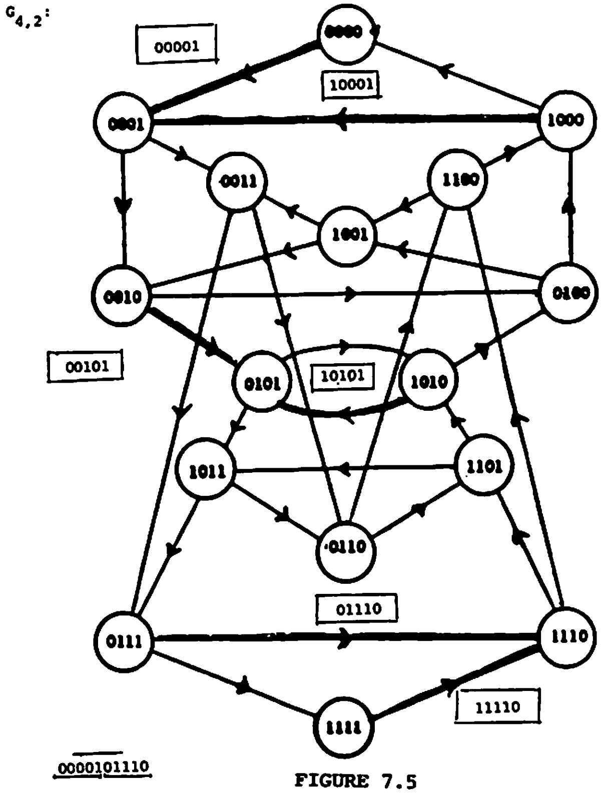


FIGURE 7.5

References

[1] A.H. Ball,; *The comma-free codes with words of length three,* Ars Combinatoria 6(1978), 117-140.

[2] A.H. Ball,; *The construction of comma-free codes with odd word length,* Ph.D. thesis. Department of Combinatorics and Optimization. University of Waterloo, Canada. 1980.

[3] A.H. Ball, and L.J. Cummings,; *The Comma-free codes with words of length 2,* Bull. Austral. Math. Soc., 14(1976), 249-258.

[4] Ball, A.H., and L.J. Cummings,; *Extremal digraphs and comma-free codes,* Ars Combinatoria, 1(1976), 239-251.

[5] A.H. Ball, and L.J. Cummings,; *Lower bounds for* $CF(n,6)$ *codes,* Proceedings of the Seventh Manitoba Conference on Numerical Mathematics, University of Manitoba, Winnipeg, (1977), 177-190.

[6] F.H.C. Crick, J.S. Griffith, and L.E. Ongel, *Codes without commas,* Proc. U.S. Nat. Acad. Sci., 43(1957), 416-421.

[7] L.J. Cummings, *Comma-free codes and incidence algebras,* Combinatorial Mathematics IV, Proceedings of the 4th Australian Conference on Combinatorial Mathematics (invited address), Springer-Verlag Lecture Notes in Mathematics, Vol. 560, Berlin, 1976.

[8] L.J. Cummings; *Maximum comma-free codes as cosets of linear codes,* Ars Combinatoria, 3(1977), 115-122.

[9] L.J. Cummings; *Comma-free codes in the De Bruijn graph.* Caribb. J. Math., 3(1983), 65-68.

[10] L.J. Cummings; *Synchronizable codes in the de Bruijn graph* Ars Combinatoria, 19(1985), 73-80.

[11] N.G. de Bruijn, and T. van Aardenne-Ehrenfest; *Circuits and trees in oriented linear graphs,* Simon Steven 28(1951), 203-217.

[12] N.G., de Bruijn; *Acknowledgement of priority to C. Flye Sainte-Marie on the counting of circular arrangements of 2^n zeros and ones that show each n-letter word exactly once.* T.H. - Report 75-WSK-06 Technische Hogeschool, Eindhoven, 1975.

[13] W.L. Eastman; *On the construction of comma-free codes,* IEEE Trans. Information Theory, 11(1965), 263-67.

[14] G. Gamow; *Possible relation between deoxyrebonucleic acid and protein structures.* Nature 173(1954), 318.

[15] S.W. Golomb, B. Gordon, and L.R. Welch; *Comma-free codes,* Canadian J. Math. 10(1958), 202-209.

[16] S.W. Golomb, M. Delbruck and L.R. Welch; *Construction and properties of comma-free codes,* Biol. Meed. Dan. Vid. Selsk., 23(1958), pp.3-34.

[17] S.W. Golomb, and B. Gordon, *Codes with bounded synchronization delay,* Information and Control, 8(1965), 355-372.

[18] S.W. Golomb, R.L. Graham, and B. Tang, *A new result on comma-free codes of even word length,* (to appear), Candian J. Math.

[19] L.J. Guibas and A.M. Odlyzko, *Periods in strings,* JCT (A), 30(1981), 19-42.

[20] B.H. Jiggs, *Recent results in comma-free codes,* Canadian J. Math., 15(1963), 178-187.

[21] R.B. Kirk and C.E. Langehop, *Periodic chains of beads,* Utilitas Math. 21C(1982), 55-84.

[22] A. Lempel; *M-ary closed sequences,* JCT 10(1971), 253-258.

[23] A. Lempel; *On the extremal factors of the de Bruijn graph,* JCT 11(1971), 17-27.

[24] M.H. Martin, *A problem in arrangements,* Bull. Amer. Math. Soc. 40(1934), 859-864.

[25] N.S. Mendelsohn; *Directed graphs with the unique path property.* Colloquium on Combinatorial theory and its Applications. Balatonfured, Hungary, 1959. Bolyai Janos Mathematikai Tarsulat. Budapest, 1970. 783-799.

[26] Y. Niho; *On maximal comma-free codes,* IEEE Transactions on Information Theory, 19(1973), 580-81.

[27] E.C. Posner, *Combinatorial structures in planatary reconnaissance,* in Error-Correcting Codes, H.B. Mann, ed.,Wiley, New York, 1968, pp. 15-46.

[28] R.A. Scholtz; *Codes with synchronization capability,* IEEE Trans. Information Theory, 12(1966), 135-142.

[29] J.J. Stiffler; *Comma-free error-correcting codes,* IRE Trans. on Information Theory, 11(1965), 107-112.