Cross-Bifix-Free Codes Within a Constant Factor of Optimality

Yeow Meng Chee, *Senior Member, IEEE*, Han Mao Kiah, Punarbasu Purkayastha, *Member, IEEE*, and Chengmin Wang

Abstract—A cross-bifix-free code is a set of words in which no prefix of any length of any word is the suffix of any word in the set. Cross-bifix-free codes arise in the study of distributed sequences for frame synchronization. We provide a new construction of cross-bifix-free codes which generalizes the construction by Bajic to longer code lengths and to any alphabet size. The codes are shown to be nearly optimal in size. We also establish new results on Fibonacci sequences, which are used in estimating the size of the cross-bifix-free codes.

Index Terms—Cross-bifix-free code, Fibonacci sequence, synchronization sequence.

I. INTRODUCTION

A crucial requirement to reliably transmit information in a digital communication system is to establish synchronization between the transmitter and the receiver. Synchronization is required not only to determine the start of a symbol, but also to determine the start of a frame of data in the received signals. The initial acquisition of frame synchronization and the maintenance of this synchronization has been a widely studied field of research for several decades. Early works on frame synchronization concentrated on introducing a synchronization word periodically into the data stream [8], [11]. In the receiver, correlation techniques were used to determine the position of the synchronization sequence within the data stream. Massey [8] introduced the notion of *bifix-free* synchronization word in order to achieve fast and reliable synchronization in binary data streams. A bifix-free word denotes a sequence of symbols in which no prefix of any length of the word is identical to any suffix of the word.

The current methods for achieving frame synchronization at the receiver do not look at exact matching of the synchronization word. Instead, the objective is to search for a word that is within a specified Hamming distance of the transmitted synchronization word. This procedure allows for faster synchronization between the transmitter and the receiver [2].

Van Wijngaarden and Willink [14] introduced the notion of a *distributed sequence* where the synchronization word is not a contiguous sequence of symbols but is instead interleaved into the data stream. For instance the binary sequence $110*0*0$ is a distributed sequence, where the symbol $*$ denotes a data symbol that can take either of the values 0 or 1. Van Wijngaarden and Willink [14] provided constructions of such sequences for binary data streams and studied their properties. Bajic *et al.* [1], [2] showed that the distributed sequence entails a simultaneous search for a set of synchronization words. Each word in the set of sequences is required to be bifix-free. In addition, there arises a new requirement that no prefix of any length of any word in the set should be a suffix of any other word in the set. This property of the set of synchronization words was termed as *cross-bifix-free* in [1], [2], and [13]. In the same works, the properties of sets of words that are cross-bifix-free were statistically analyzed. In this paper, we term the set of words which are cross-bifix-free as a *cross-bifix-free code*. In the above example of a distributed sequence, the set of words $\{(1,1,0,0,0,0,0), (1,1,0,0,0,1,0), (1,1,0,1,0,0,0), (1,1,0,1,0,1,0)\}$ forms a cross-bifix-free code.

In a follow up work, Bajic [3] provided a new construction of cross-bifix-free codes over a binary alphabet for word lengths up to eight. This specific construction uncovers interesting connections to the Fibonacci sequence of numbers. In particular, the number $S(n)$ of binary words of length n, for $3 \leq n \leq 8$, which are cross-bifix-free satisfies the Fibonacci recursion

$$S(n) = S(n-1) + S(n-2).$$

It is noted in [3] that although this construction gives larger sets compared to distributed sequences [14] for $n \leq 8$, the sizes of the sets are relatively smaller for lengths greater than eight. In a recent work, Bilotta *et al.* [4] introduced a new construction of binary cross-bifix-free codes based on lattice paths and showed that their construction attains greater cardinality compared to the ones in [3].

In this study, we revisit the construction in [3]. We give a new construction of cross-bifix-free codes that generalizes the construction of [3] in two ways. First, we provide new binary codes that are greater in cardinality compared to the ones in [4] for larger lengths. In the process, we discover interesting connections of the size of the codes obtained to the so-called k-generalized Fibonacci numbers. Second, we generalize the construction to q-ary alphabets for any $q, q \geq 2$. To the best of our knowledge, this is the first construction of cross-bifix-free codes over alphabets of size greater than two. The size of the generalized q-ary constructions are also related to a Fibonacci

sequence, which we call the $(q-1)$-weighted k-generalized Fibonacci sequence (see Section II for the exact definition). Using this relation to the Fibonacci sequences, we analyze the asymptotic size of our construction. In the process of this asymptotic analysis, we generalize a result of Dresden [5] on k-generalized Fibonacci sequence to $(q-1)$-weighted k-generalized Fibonacci sequence. The main asymptotic result on the size of cross-bifix-free codes that we prove is described in the theorem below.

Theorem 1.1: Let $C(n,q)$ denote the maximum size of a cross-bifix-free code of length n over an alphabet of size q. Then

$$\liminf_{n\to\infty} \frac{C(n,q)}{q^n/n} \geq \frac{q-1}{qe} \simeq 0.368\frac{q-1}{q} \qquad (1)$$

$$\limsup_{n\to\infty} \frac{C(n,q)}{q^n/n} \leq \frac{1}{2} = 0.5. \qquad (2)$$

Note that the lower bound is within a constant factor of the best possible construction. The ratio between the lower and the upper bound increases toward $2/e = 0.736$ for larger alphabet sizes. In comparison, a similar ratio of the size of the binary codes constructed by Bilotta *et al.* [4], or the distributed sequences by van Wijngaarden and Willink [14], to the quantity $2^n/n$, asymptotically goes to zero.

The rest of this paper is organized as follows. Our presentation is provided for general alphabet size q, $q \geq 2$, and the results for the binary alphabet are obtained as a special case. In Section III, we provide the construction of the cross-bifix-free code and show that for the binary alphabet, it is optimal for lengths $n \leq 14$, barring an exception at $n = 9$. In Section IV, we study the asymptotic behavior of the size of cross-bifix-free codes obtained from our construction. In particular, we exhibit (1) and (2) in Theorems 4.2 and 4.3, respectively. Results on the behavior of the $(q-1)$-weighted k-generalized Fibonacci sequence are also presented in this section. Lengthy calculations and some proofs are deferred to the Appendix. In the following section, we introduce the basic notations and definitions required.

II. NOTATIONS AND DEFINITIONS

Let $\mathbb{Z}_q = \{0, \ldots, q-1\}$ be an alphabet of q elements. We denote by $\mathbb{Z}_q^*$ all the nonzero elements of the set $\mathbb{Z}_q$, that is, $\mathbb{Z}_q^* = \mathbb{Z}_q \setminus \{0\}$. A consecutive sequence of m elements $b \in \mathbb{Z}_q$ is denoted by the short form b^m. As an example, the word $(0, 0, 1, 1, 1, 0, 1)$ is represented in short as $(0^2, 1^3, 0, 1)$. For convenience, if $m = 0$, then b^m is used to denote the absence of any element.

Definition 2.1: For a word $\mathbf{u} \in \mathbb{Z}_q^n$, a word $\mathbf{v}$ is called a *prefix* of $\mathbf{u}$ if we can write $\mathbf{u}$ as $\mathbf{u} = (\mathbf{v}, \mathbf{w})$, for some word $\mathbf{w}$. The word $\mathbf{w}$ is called a *suffix* of $\mathbf{u} \in \mathbb{Z}_q^n$ if we can write $\mathbf{u}$ as $\mathbf{u} = (\mathbf{v}, \mathbf{w})$, for some word $\mathbf{v}$.

For any word $\mathbf{u}$, we only consider prefixes and suffixes which have length strictly less than the length of $\mathbf{u}$.

Definition 2.2: A word $\mathbf{u} \in \mathbb{Z}_q^n$ is called *bifix-free* if the prefix of any length of the word is not a suffix of the word.

Definition 2.3: A *cross-bifix-free code* is a set of words in $\mathbb{Z}_q^n$ which satisfy the property that the prefix of any length of any word is not the suffix of any word in the set, including itself.

We denote the maximum size of a cross-bifix-free code by the notation $C(n,q)$.

Definition 2.4: The $(q-1)$-*weighted k-generalized Fibonacci sequence* is a sequence of numbers which satisfies the recurrence relation

$$F_{k,q}(n) = (q-1)\sum_{l=1}^{k} F_{k,q}(n-l)$$

for some initial values of $F_{k,q}(0), \ldots, F_{k,q}(k-1)$. For $q = 2$, the sequence obtained is called a k-*generalized Fibonacci sequence*. For $q = 2$, $k = 2$, and the initialization $F_{2,2}(0) = 1$, $F_{2,2}(1) = 2$, we obtain the usual Fibonacci sequence.

The $(q-1)$-weighted k-generalized Fibonacci sequence is a special case of the *weighted k-generalized Fibonacci sequence* which satisfies the recurrence relation [7], [12]

$$F_k(n) = a_1 F_k(n-1) + a_2 F_k(n-2) + \cdots + a_k F_k(n-k)$$

where the weights are given by $a_1, a_2, \ldots, a_k \in \mathbb{Z}$, and $\mathbb{Z}$ denotes the integers. Setting all the weights equal to $q-1$ gives the sequence in the above definition.

The $(q-1)$-weighted k-generalized Fibonacci sequence arises in the study of cross-bifix-free codes as described in the following section.

III. CONSTRUCTION OF CROSS-BIFIX-FREE CODES

In this section, we provide a general construction of cross-bifix-free codes over the q-ary alphabet. Interestingly, the sizes of our construction are related to the $(q-1)$-weighted k-generalized weighted Fibonacci numbers $F_{k,q}(n)$. The initialization on $F_{k,q}(n)$, $0 \leq n \leq k-1$ that we use is given as

$$F_{k,q}(n) = q^n, \quad n = 0, \ldots, k-1. \qquad (3)$$

Below, we describe the family of cross-bifix-free codes in the space $\mathbb{Z}_q^n$. The family is obtained by varying the value of k.

The construction: For any $2 \leq k \leq n-2$, denote by $\mathcal{S}_{k,q}(n)$ the set of all words $(s_1, s_2, \ldots, s_n)$ in $\mathbb{Z}_q^n$ that satisfy the following two properties:
1) $s_1 = s_2 = \cdots = s_k = 0$, $s_{k+1} \neq 0$ and $s_n \neq 0$,
2) the subsequence $(s_{k+2}, s_{k+3}, \ldots, s_{n-1})$ does not contain any string of k consecutive 0's.

This construction implies that $\mathcal{S}_{k,q}(n)$ contains all possible words of length n that start with k zeroes, end with a nonzero element, and have at most $k-1$ consecutive zeroes in the last $n-1$ coordinates. In the remaining part of this section, we show that for every k, $k = 2, \ldots, n-2$, this set of words forms a cross-bifix-free code. We determine its size in terms of the Fibonacci sequence. First, in the theorem below, we show that $\mathcal{S}_{k,q}(n)$ is a cross-bifix-free code. Additionally, we show that the code $\mathcal{S}_{k,q}(n)$ has the property that it cannot be expanded while preserving the property that it is cross-bifix-free. That is, for every word $\mathbf{x} \in \mathbb{Z}_q^n \setminus \mathcal{S}_{k,q}(n)$, the set $\{\mathbf{x}\} \cup \mathcal{S}_{k,q}(n)$ is not cross-bifix-free.

Theorem 3.1: For any k, $2 \leq k \leq n-2$, the set $\mathcal{S}_{k,q}(n)$ is a nonexpandable cross-bifix-free code.

Proof: To see that $\mathcal{S}_{k,q}(n)$ is a cross-bifix-free code, note that the prefix of any word of $\mathcal{S}_{k,q}(n)$ starts with k consecutive zeroes. But in the last $n-1$ coordinates of any word, we have at most $k-1$ consecutive zeroes, and the last coordinate is always

nonzero. Thus, no prefix of any length of any word can match any suffix of itself or of any other word in $\mathcal{S}_{k,q}(n)$.

To show that $\mathcal{S}_{k,q}(n)$ is nonexpandable, we consider all the possible configurations of words that could be appended to the set $\mathcal{S}_{k,q}(n)$. First, we note that we cannot append any word starting with a nonzero element since the nonzero element occurs in the last coordinate of some word in $\mathcal{S}_{k,q}(n)$. Similarly, we cannot append any word ending with a zero element. The other possible configurations of words that we need to consider are as follows.

1) Let $\mathbf{s}$ be a word which contains at least k consecutive zeroes in the last $n-1$ coordinates. We consider the suffix in $\mathbf{s}$ that starts with the last set of k consecutive zeroes and contains at most $k-1$ consecutive zeroes following it, that is, the suffix has the form $(0^k, \alpha, \mathbf{u})$, where α is nonzero and $\mathbf{u}$ is a vector of length m that has at most $k-1$ consecutive zeroes. Then, the word of length n $(0^k, \alpha, \mathbf{u}, 1^{n-m-k-1})$ is a word in $\mathcal{S}_{k,q}(n)$ and has a prefix matching a suffix of $\mathbf{s}$. Thus, $\mathbf{s}$ cannot be appended to $\mathcal{S}_{k,q}(n)$.

2) Let $\mathbf{s}$ be a word which contains a prefix of at most $k-1$ zeroes followed by a nonzero element, that is $\mathbf{s} = (0^l, \alpha, \mathbf{u})$, where $0 < l \leq k-1$, α is nonzero, and $\mathbf{u}$ has length $n-l-1$. It is readily seen that $(0^l, \alpha)$ is also the suffix of the word $(0^k, 1^{n-k-l-1}, 0^l, \alpha)$ in $\mathcal{S}_{k,q}(n)$. Hence, such a word cannot be appended to $\mathcal{S}_{k,q}(n)$.

Thus, no additional word can be appended to the set $\mathcal{S}_{k,q}(n)$, while still preserving the cross-bifix-free property. ∎

The nonexpandability of the construction above parallels the nonexpandability of the cross-bifix-free codes obtained in [3] and [4]. However, note that the nonexpandability does not automatically indicate the optimality of the construction, as is evident from the many values of k for which the nonexpandability holds true. In the following sections, we instead show that the largest sized set obtained by optimizing over the value of k, $k = 2, \ldots, n-2$, differs (in ratio) from the size of the optimal code by only a factor of a constant $2(q-1)/(qe)$.

We first describe a recursive construction of the set $\mathcal{S}_{k,q}(n)$ in terms of the sets $\mathcal{S}_{k,q}(n-l)$, $l = 1, \ldots, k$. This recursive construction immediately establishes the connection to the Fibonacci recurrence and helps us determine the size of the set in terms of the Fibonacci numbers.

Theorem 3.2:

$$\mathcal{S}_{k,q}(n) = \begin{cases} \{(0^k, \alpha, \mathbf{s}, \beta) : \alpha, \beta \in \mathbb{Z}_q^*, \mathbf{s} \in \mathbb{Z}_q^{n-k-2}\}, \\ \qquad\qquad\qquad\qquad\qquad k+2 \leq n \leq 2k+1, \\ \bigcup_{l=1}^{k} \{(\mathbf{s}, 0^{l-1}, \alpha) : \mathbf{s} \in \mathcal{S}_{k,q}(n-l), \alpha \in \mathbb{Z}_q^*\}, \\ \qquad\qquad\qquad\qquad\qquad\qquad\qquad 2k+2 \leq n. \end{cases}$$

Proof: For $n = k+2, \ldots, 2k+1$, the coordinates $s_{k+2}, \ldots, s_{n-1}$ necessarily have at most $n-k-2 < k$ zeroes and hence can contain all the words of length $n-k-2$. This establishes the result for $n = k+2, \ldots, 2k+1$.

Now, let $n \geq 2k+2$. For brevity, denote each set on the right-hand side (RHS) of the equation in Theorem 3.2 by

$$\mathcal{T}_l(n) \triangleq \{(\mathbf{s}, 0^{l-1}, \alpha) : \mathbf{s} \in \mathcal{S}_{k,q}(n-l), \alpha \in \mathbb{Z}_q^*\}. \quad (4)$$

Note that the sets $\mathcal{T}_l(n)$ are mutually disjoint for different l since the last l coordinates have different structure for the different sets. To show that $\mathcal{S}_{k,q}(n) \subseteq \cup_l \mathcal{T}_l(n)$, note that any

element $\mathbf{u} \in \mathcal{S}_{k,q}(n)$ has at most $k-1$ zeroes in the last $n-1$ coordinates, and hence, the word $\mathbf{u}$ must be of the form $\mathbf{u} = (0^k, u_{k+1}, \ldots, u_{n-l-1}, 0^{l-1}, \alpha)$ where $u_{n-l-1}, \alpha \in \mathbb{Z}_q^*$ and $l \in \{1, \ldots, k\}$. Thus, $\mathbf{u} \in \mathcal{T}_l(n)$.

To show the reverse inclusion, let $l \in \{1, \ldots, k\}$ and let $\mathbf{s} \in \mathcal{S}_{k,q}(n-l)$. Note that $\mathbf{s}$ ends with a nonzero element. The word $(\mathbf{s}, 0^{l-1}, \alpha)$, where $\alpha \in \mathbb{Z}_q^*$, starts with a sequence 0^k, ends with a nonzero element, and has at most $k-1$ consecutive zeroes in the last $n-1$ coordinates. Hence, $(\mathbf{s}, 0^{l-1}, \alpha) \in \mathcal{S}_{k,q}(n)$ and the set $\{(\mathbf{s}, 0^{l-1}, \alpha) : \alpha \in \mathbb{Z}_q^*\}$ is a subset of $\mathcal{S}_{k,q}(n)$. Hence, $\mathcal{T}_l(n) \subseteq \mathcal{S}_{k,q}(n)$ for every $l = 1, \ldots, k$. ∎

Corollary 3.1: The cardinality of $\mathcal{S}_{k,q}(n)$ for $n \geq 3$ is given by the equation

$$S_{k,q}(n) = |\mathcal{S}_{k,q}(n)| = (q-1)^2 F_{k,q}(n-k-2).$$

Proof: For $n = k+2, \ldots, 2k+1$, the corollary can be readily verified from the expression in (3) and Theorem 3.2. We use an induction argument for $n \geq 2k+2$. Assume that the corollary is true for $n < N$ where $N \geq 2k+2$. First, note that by using the definition in (4), we get

$$\sum_{l=1}^{k} |\mathcal{T}_l(N)| = \sum_{l=1}^{k} \left| \{(\mathbf{s}, 0^{l-1}, \alpha) : \mathbf{s} \in \mathcal{S}_{k,q}(N-l), \alpha \in \mathbb{Z}_q^*\} \right|$$
$$= \sum_{l=1}^{k} (q-1) S_{k,q}(N-l).$$

Now

$$S_{k,q}(N) = \sum_{l=1}^{k} |\mathcal{T}_l(N)|$$
$$= \sum_{l=1}^{k} (q-1) S_{k,q}(N-l)$$
$$= \sum_{l=1}^{k} (q-1)(q-1)^2 F_{k,q}(N-l-k-2)$$
$$= (q-1)^2 F_{k,q}(N-k-2).$$

We used the induction argument in the second last step. This proves the corollary. ∎

For fixed n and q, the largest size of the set $\mathcal{S}_{k,q}(n)$ can be obtained by optimizing over the choice of k. Let $S(n,q)$ denote this maximum. It is given by the expression

$$S(n,q) = \max\{(q-1)^2 F_{k,q}(n-k-2) : 2 \leq k \leq n-2\}. \quad (5)$$

In particular, the size $S(n,q)$ is upper bounded by the maximum cardinality $C(n,q)$ of a cross-bifix-free code.

A. Sizes of Cross-Bifix-Free Codes for Small Lengths

The size of binary cross-bifix-free codes obtained in [4] is obtained by counting lattice paths, in particular, Dyck paths.

Theorem 3.3 (see [4]): Let $B(n)$ denote the size of a binary cross-bifix-free code of length n constructed by Bilotta *et al.*

TABLE I
COMPARING THE VALUES FROM [4] WITH (5)

n	[4]	Eq. (5)	k	n	[4]	Eq. (5)	k
3	**1**	**1**	-	17	1430	1705	3
4	**1**	**1**	2	18	2529	3136	3
5	**2**	**2**	2	19	4862	5768	3
6	**3**	**3**	2	20	8790	10671	4
7	**5**	**5**	2	21	16796	20569	4
8	**8**	**8**	2	22	30275	39648	4
9	**14**	13	2	23	58786	76424	4
10	23	**24**	3	24	107786	147312	4
11	42	**44**	3	25	208012	283953	4
12	72	**81**	3	26	380162	547337	4
13	132	**149**	3	27	742900	1055026	4
14	227	**274**	3	28	1376424	2033628	4
15	429	**504**	3	29	2674440	3919944	4
16	760	**927**	3	30	4939443	7555935	4

[4]. For $m \geq 1$, let $C_m = \frac{1}{m+1}\binom{2m}{m}$ denote the mth Catalan number. Then

$$
B(n) =
\begin{cases}
C_m, & n = 2m+1, m \geq 1 \\
\sum_{i=0}^{m/2} C_i C_{m-i}, & n = 2m+2, m \text{ odd} \\
\sum_{i=0}^{(m+1)/2} C_i C_{m-i} - C_{(m-1)/2}^2, & n = 2m+2, m \text{ even.}
\end{cases}
$$

For values of $n \leq 16$, it is verified by numerical computations that the sizes obtained by our construction are all optimal, except for the value $n = 9$. In particular, we get Table I of values for $3 \leq n \leq 30$. The first column gives the value of the word length n, the second column shows the sizes of the codes obtained in [4], the third column gives the sizes obtained from our construction after optimizing over different values of k, and finally, the last column gives the values of k for which $S_{k,q}(n)$ achieves the maximal size in the third column. The numbers in bold denote the sizes that are known to be optimal.

The optimality of the values for $n \leq 16$ is proved computationally by setting up a specific program that searches for the largest clique in a graph. The graph consists of vertices which correspond to the set of all words in $\mathbb{Z}_2^n$ that are bifix-free. An edge exists between two vertices, i.e., two words, if they are mutually cross-bifix-free. The algorithm `MaxCliqueDyn`[6] is used to determine the maximum size of the clique in the graph. This algorithm shows that the values denoted by bold in Table I are optimal.

Note that our construction has larger size than the construction in [4] for all values of n, $13 \leq n \leq 30$. This trend is observed asymptotically too, as we describe in the following sections.

IV. NEAR OPTIMALITY OF THE SIZE $S(n, q)$

In this section, we show that the size $S(n, q)$ is close to the maximum size $C(n, q)$. The ratio $S(n, q)/C(n, q)$ measures how close the construction in Section III is to the optimal value. The following theorem gives an asymptotic lower bound on this ratio.

Theorem 4.1: The following limit holds:

$$
\liminf_{n \to \infty} \frac{S(n, q)}{C(n, q)} \geq 2\frac{q-1}{qe}. \tag{6}
$$

This lower bound is proved by showing a lower bound on $S(n, q)$ and an upper bound on $C(n, q)$. The derivation of the lower bound on $S(n, q)$ crucially depends on the properties of the $(q-1)$-weighted k-generalized Fibonacci sequence of numbers. We digress in the next section to first establish these needed properties.

A. Properties of the Fibonacci Sequence $F_{k,q}(n)$

Levesque [7] showed in a very general context that to every weighted k-generalized Fibonacci sequence of numbers we can associate a *characteristic polynomial* (see Theorem A.1 in the Appendix). For the $(q-1)$-weighted k-generalized Fibonacci sequence, this polynomial specializes to the following form:

$$
f(x) = x^k - (q-1)\sum_{i=0}^{k-1} x^i. \tag{7}
$$

Below, we state the properties of this polynomial and of the corresponding Fibonacci numbers. The initialization sequence that we use is the one described in (3). The proofs in this section are omitted for clarity of presentation and are instead provided in the Appendix.

Proposition 4.1: The polynomial $f(x)$ has distinct roots with a unique real root $\alpha \equiv \alpha(k, q)$ outside the unit circle. The root α lies in the interval $(1, q)$.

The value of the root α is in fact close to q. An estimate of this root is given by the following lemma.

Lemma 4.1: There exists a number K_q such that the following holds. For all $k \geq K_q$, there exists a $\beta \equiv \beta(k, q)$ in the interval $(q - \frac{1}{q^{k-1}}, q)$ such that

$$
q - \frac{q-1}{\beta^k} < \alpha < q. \tag{8}
$$

Finally, the Fibonacci numbers can be expressed in terms of this real root α. Let $[x]$ denote the integer closest to the real number x.

Proposition 4.2: Let $q \geq 2$. The nth number in the $(q-1)$-weighted k-generalized Fibonacci sequence is given by the expression

$$
F_{k,q}(n) = \left[\frac{(\alpha - 1)\alpha^{n+1}}{(q + (k+1)(q - \alpha))(q-1)} \right].
$$

We note here that Proposition 4.1 is a generalization to $q \geq 3$ of the result obtained by Miles [9] for $q = 2$. We adopt a technique similar to the one in [10]. Additionally, Proposition 4.2 is a generalization of the result in Dresden [5] to $(q-1)$-weighted k-generalized Fibonacci numbers, for $q \geq 3$. For $q = 2$, the expression above reduces to the expression for the sequence $F_{k,2}(n)$ as obtained in [5].

B. Lower Bound on $S(n, q)$

Using the properties of the Fibonacci numbers from Section IV-A, we establish an asymptotic lower bound on the size $S(n, q)$.

Theorem 4.2: The asymptotic size $S(n, q)$ satisfies the limit

$$\liminf_{n \to \infty} \frac{S(n, q)}{q^n/n} \geq \frac{q-1}{qe}. \tag{9}$$

Proof: Using Corollary 3.1 and Proposition 4.2 in successive steps, we obtain

$$\frac{S(n, q)}{q^n/n} \geq \frac{n(q-1)^2 F_{k,q}(n-k-2)}{q^n}$$

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

$$\geq \left(\frac{q-1}{q} \right) \left(\frac{\alpha-1}{\alpha} \right) \left(\frac{\alpha}{q} \right)^n \left(\frac{n}{\alpha^k} \right) - o(1)$$

where the term $o(1) \to 0$ as $n \to \infty$. To derive the asymptotics, we choose n as an increasing function of k:

$$n \equiv n(k) = \lceil c\alpha^k \rceil$$

where c is a positive constant. Note that α is also a function of k. We obtain

$$\frac{S(n, q)}{q^n/n} \geq \left(\frac{q-1}{q} \right) \left(\frac{\alpha-1}{\alpha} \right) \left(\frac{\alpha}{q} \right)^{\lceil c\alpha^k \rceil} \left(\frac{\lceil c\alpha^k \rceil}{\alpha^k} \right) - o(1)$$

$$\geq \left(\frac{q-1}{q} \right) \left(\frac{\alpha-1}{\alpha} \right) \left(\frac{\alpha}{q} \right)^{(c\alpha^k+1)} \cdot c - o(1)$$

$$= \left(\frac{q-1}{q} \right) \left(\frac{\alpha-1}{q} \right) c \left(\frac{\alpha}{q} \right)^{c\alpha^k} - o(1).$$

The last term in the RHS of the equation above can be further lower bounded by using Lemma 4.1. We assume that there exists a number K_q and $k \geq K_q$, as required by the lemma

$$c \left(\frac{\alpha}{q} \right)^{c\alpha^k} \geq c \left(1 - \frac{q-1}{q\beta^k} \right)^{cq^k}$$

$$= c \left(\left(1 - \frac{q-1}{q\beta^k} \right)^{c\beta^k} \right)^{(q/\beta)^k}.$$

The RHS of the above equation tends to $ce^{-\frac{c(q-1)}{q}}$ as $k \to \infty$ since $\left(1 + \frac{1}{x} \right)^x \to e$ as $x \to \infty$ and $\beta(k, q) \to q$ as $k \to \infty$. The term $ce^{-\frac{c(q-1)}{q}}$ attains a maximum of $\frac{q}{(q-1)e}$ when $c = q/(q-1)$. The theorem follows by substituting this value into the lower bound above. $\blacksquare$

C. Upper Bound on the Maximum Size $C(n, q)$

Let M denote the size of a cross-bifix-free code of length n over an alphabet of size q. An upper bound for the maximum size of a cross-bifix-free code is readily obtained from the study of the statistical properties of such sets in the data stream. The main object of study is the time when the search for any word of the cross-bifix-free code in the data stream returns with a positive match. Bajic *et al.* [1], [2] establish the probability distribution function of this time, the expected time duration for a

match, and the variance of this distribution. The variance σ^2 of the time for the first match is given by the expression [2, eq. (18)]

$$\sigma^2 = (1 - 2n) \frac{q^n}{M} + \frac{q^{2n}}{M^2}. \tag{10}$$

Using the property that the variance is always nonnegative immediately gives us the required upper bound on any cross-bifix-free code. In particular, we have the following theorem.

Theorem 4.3: Let $C(n, q)$ denote the maximum size of a cross-bifix-free code in $\mathbb{Z}_q^n$. Then

$$C(n, q) \leq \frac{q^n}{2n - 1}.$$

We remark that this upper bound, albeit immediate from (10), was not noted in the previous works on the size of the cross-bifix-free codes. Combining Theorems 4.2 and 4.3, we obtain Theorems 4.1 and 1.1.

D. Comparison to Earlier Results

To compare the construction in this study with the new construction of binary cross-bifix-free codes [4] and the construction of distributed sequences [14], we study the asymptotics of their respective constructions for large n. In both cases, we exhibit that the size of the previous constructions is a negligible fraction of $2^n/n$, in contrast to the nearly optimal construction described in the previous section.

The asymptotic behavior of the construction in [4] is obtained from the expressions in Theorem 3.3. We obtain that

$$B(n) \begin{cases} = C_m, & n = 2m + 1 \\ \leq C_{m+1}, & n = 2m + 2 \end{cases}$$

where $C_m = \frac{1}{m+1} \binom{2m}{m}$ is the mth Catalan number and $m \geq 1$. Using Stirling's approximation, we get that the number C_m is approximately

$$C_m \simeq \frac{1}{m+1} \frac{2^{2m}}{\sqrt{\pi m}}.$$

Thus, for n odd

$$\frac{B(n)}{2^n/n} = \frac{(2m+1)C_m}{2^{2m+1}} \simeq \frac{2m+1}{2(m+1)} \frac{1}{\sqrt{\pi m}},$$

which goes to zero as $2m + 1 = n \to \infty$. Similar conditions hold for the case $n = 2m + 2$. Thus, the construction in [4] is a negligible fraction of $2^n/n$.

On the other hand, van Wijngaarden and Willink [14, eq. (4)] showed that for a set of distributed sequences of length n, and with h synchronization positions

$$n \leq \lfloor h^2/4 \rfloor + 1. \tag{11}$$

Let $D(n)$ denote the maximum size of a set of distributed sequences. Then, it follows from (11) that

$$D(n) \leq 2^{n-h} \leq 2^{n-2\sqrt{n-1}}.$$

Hence, the ratio $D(n)/(2^n/n)$ tends to zero with increasing n.

V. CONCLUSION

We provided a new construction of cross-bifix-free codes that are close to the maximum possible size. The construction for the binary codes is shown to be larger than the previously constructed codes for all lengths $n \leq 30$, barring an exception at $n = 9$. We also provided the first construction of q-ary cross-bifix-free codes for $q > 2$. In the process, we established new results on the Fibonacci sequences, generalizing some earlier works on these sequences.

APPENDIX

In this appendix, we provide the proofs of the results on the Fibonacci sequences that are stated in Section IV-A. First, we recall a very general theorem on weighted k-generalized Fibonacci sequences proved in [7].

Theorem A.1 (see [7]): Let $k \geq 2$. Let $F_k(n)$ be defined by the following recurrence relation:

$$F_k(n) = \sum_{i=1}^{k} a_i F_k(n-i), \qquad \text{for } n \geq k$$

for $a_i \in \mathbb{Z}$, $i = 1, \ldots, k$, and with the initial conditions, $F_k(0), F_k(1), \ldots, F_k(k-1)$. Additionally, suppose that the *characteristic polynomial* $h(x)$ associated with the sequence $\{F_k(n)\}_{k=0}^{\infty}$,

$$h(x) = x^k - \sum_{i=0}^{k-1} a_{k-i} x^i,$$

has distinct roots $\gamma_1, \gamma_2, \ldots, \gamma_k$. Then, for $n \geq k$, the values of $F_k(n)$ are given by the expression

$$F_k(n) = \sum_{j=0}^{k-1} p_{n-j} v_j$$

where

$$v_0 = u(0),$$
$$v_j = u(j) - \sum_{i=1}^{j} a_i u(j-i), \quad \text{for } 1 \leq j \leq k-1$$
$$p_j = \sum_{i=1}^{k} \frac{\gamma_i^{k-1+j}}{h'(\gamma_i)}, \quad \text{for } j \geq 1.$$

For $a_i = q - 1$, $i = 1, \ldots, k$, we obtain the corresponding expressions for the $(q-1)$-weighted k-generalized Fibonacci numbers. In particular, the polynomial $h(x)$ reduces to the polynomial $f(x)$ defined in (7).

We proceed with the proofs of the propositions in Section IV-A. In order to prove Proposition 4.1, we first establish two lemmas below. Define a polynomial $g(x)$ as

$$g(x) \overset{\triangle}{=} (x-1)f(x) = x^k(x-q) + (q-1). \qquad (12)$$

Lemma A.1: The polynomial $f(x)$ has a real root in the interval $(1, q)$.

Proof: This follows from the fact that $f(1) = 1 - k(q-1) < 0$ and $f(q) = g(q)/(q-1) = 1 > 0$. ∎

Lemma A.2: Let $\alpha \equiv \alpha(k, q)$ be the real root of $f(x)$ in $(1, q)$. Then, the polynomial $g(x)$, and consequently the polynomial $f(x)$, satisfies the following inequalities:

(i) $g(x) > 0$ for $x \in (\alpha, \infty)$,

(ii) $g(x) < 0$ for $x \in (1, \alpha)$.

Proof: Observe that

$$g'(x) = x^{k-1}((k+1)x - kq)$$

and so $g'(x) < 0$ for $x \in [1, kq/(k+1))$ and $g'(x) > 0$ for $x \in (kq/(k+1)), q]$. Since $g(1) = g(\alpha) = 0$, $g(kq/(k+1)) < 0$ and $g(q) > 0$, the lemma follows. ∎

Next, we establish Proposition 4.1.

Proof of Proposition 4.1: First, we show that the roots of $g(x)$ in (12), and hence of $f(x)$, are distinct. Indeed, $g'(x) = 0$ if and only if $x = 0$ or $x = \frac{kq}{k+1}$. However, $g(0) \neq 0$ and $g\left(\frac{kq}{k+1}\right) \neq 0$. Therefore, the roots are distinct.

Next, let γ be a root of $f(x)$ with $\gamma \neq \alpha$. We prove by contradiction that $|\gamma| < 1$. We consider the two cases $|\gamma| > |\alpha|$ and $|\gamma| < |\alpha|$ separately. Suppose $|\gamma| > |\alpha|$. Since $\gamma^k = (q-1) \sum_{i=1}^{k-1} \gamma^i$, we get

$$|\gamma|^k = |(q-1) \sum_{i=0}^{k-1} \gamma^i| \leq (q-1) \sum_{i=0}^{k-1} |\gamma|^i$$

and so, $f(|\gamma|) \leq 0$, contradicting part (i) of Lemma A.2.

Next, suppose $|\gamma| \in [1, |\alpha|]$. Since γ is also a root of $g(x)$, $q\gamma^k = \gamma^{k+1} + (q-1)$. Then

$$q|\gamma|^k = |\gamma^{k+1} + (q-1)| \leq |\gamma|^{k+1} + (q-1) \qquad (13)$$

which implies that $g(|\gamma|) \geq 0$. Then, by part (ii) of Lemma A.2, $|\gamma| \in \{1, |\alpha|\}$ and equality in (13) holds. Hence, γ^{k+1} and γ^k are real, implying that γ is real. Since the roots of $g(x)$ are distinct, $\gamma \in \{-1, -\alpha\}$. But $g(-1) = (-1)^{k+1}(q+1) - (q-1) \neq 0$ and $g(-\alpha) = 2\alpha^k q + (1 + (-1)^k)(q-1) \neq 0$, contradicting the fact that γ is a root of $g(x)$.

Therefore, α is the only root outside the unit circle. By Lemma A.1, α is in the required interval. ∎

Proof of Proposition 4.2: We apply Lemma A.1 with $a_i = q - 1$, $i = 1, \ldots, k$ and with $h(x) = f(x)$. Observe that $v_0 = v_1 = \cdots = v_{k-1} = 1$. We obtain

$$\begin{aligned}
F_{k,q}(n) &= \sum_{j=0}^{k-1} p_{n-j} \\
&= \sum_{j=0}^{k-1} \sum_{i=1}^{k} \frac{\gamma_i^{k-1+n-j}}{f'(\gamma_i)} \\
&= \sum_{i=1}^{k} \frac{\gamma_i^n}{f'(\gamma_i)} \sum_{j=0}^{k-1} \gamma_i^j \\
&= \sum_{i=1}^{k} \left(\frac{\gamma_i^n}{f'(\gamma_i)}\right) \left(\frac{\gamma_i^k}{q-1}\right) \\
&= \sum_{i=1}^{k} \left(\frac{\gamma_i^n(\gamma_i-1)}{(q+(k+1)(\gamma_i-q))\gamma_i^{k-1}}\right) \left(\frac{\gamma_i^k}{q-1}\right) \\
&= \sum_{i=1}^{k} \frac{\gamma_i^{n+1}(\gamma_i-1)}{(q-1)(q+(k+1)(\gamma_i-q))}.
\end{aligned}$$

To obtain the fourth step, we used the fact that γ_i's are roots of $f(x)$. Without loss of generality, let γ_1 be the $\alpha(k, q)$ defined in Proposition 4.2 and so $|\gamma_i| < 1$ for $i \geq 2$. We note that

$q + (k+1)(\gamma_i - q) = -kq + (k+1)\gamma_i$. We get the following sequence of inequalities for $q \geq 3$:

$$\left| \sum_{i=2}^{k} \frac{\gamma_i^{n+1}(\gamma_i - 1)}{(q-1)(q+(k+1)(\gamma_i - q))} \right|$$

$$\leq \sum_{i=2}^{k} |\gamma_i^{n+1}| \left| \frac{(\gamma_i - 1)}{(q-1)(q+(k+1)(\gamma_i - q))} \right|$$

$$\leq \sum_{i=2}^{k} \frac{|\gamma_i - 1|}{(q-1)|kq - (k+1)\gamma_i|}$$

$$< \sum_{i=2}^{k} \frac{1}{q-1} \frac{2}{k(q-1) - 1}$$

$$< \frac{1}{2}$$

where the second last step is obtained by observing that for $i \geq 2$, we have the inequalities $|\gamma_i - 1| < 2$ and $|kq - (k+1)\gamma_i| > kq - (k+1)$. The last step is obtained by applying the inequality $q \geq 3$. This completes the proof for $q \geq 3$. The proof for $q = 2$ is present in [5]. ∎

Finally, we prove Lemma 4.1. Again, for brevity, we denote $\alpha(k,q)$ and $\beta(k,q)$ by α and β, respectively.

Proof of Lemma 4.1: Observe that $g(q - 1/q^{k-1}) < 0$ if and only if

$$(1 - 1/q^k)^k > (1 - 1/q)$$

where $g(x)$ is the polynomial defined in (12). Since $(1 - 1/q^k)^k \to 1$ as $k \to \infty$, there exists a constant K_q such that $g(q - 1/q^{k-1}) < 0$ for all $k \geq K_q$.

Hence, for all $k \geq K_q$, there exists β in the interval $(q - 1/q^{k-1}, q)$ such that $g(\beta) < 0$. We claim that β satisfies (8) by showing that $g(\beta) < 0$ implies that $g(q - 1/\beta^k) < 0$. Indeed, since

$$g(\beta) = \beta^k(\beta - q) + (q-1) < 0$$

we get

$$\beta < q - \frac{q-1}{\beta^k}$$

$$\Rightarrow \quad g\left(q - \frac{q-1}{\beta^k}\right) = \left(q - \frac{q-1}{\beta^k}\right)^k \left(-\frac{q-1}{\beta^k}\right) + (q-1)$$
$$< 0.$$

Since $g(\alpha) = 0$, we get $q - \frac{q-1}{\beta^k} < \alpha < q$. ∎

ACKNOWLEDGMENT

We are grateful to D. Bajic for providing us with a copy of her work [3], and the anonymous reviewers for their helpful comments. In particular, the comparison with distributed sequences in Section IV-D was prompted by a perspicacious reviewer.

REFERENCES

[1] D. Bajic, J. Stojanovic, and J. Lindner, "Multiple window-sliding search," in *Proc. IEEE Int. Symp. Inf. Theory*, Yokohama, Japan, Jun. 2003, p. 249.

[2] D. Bajic and J. Stojanovic, "Distributed sequences and search process," in *Proc. IEEE Int. Conf. Commun.*, Paris, France, Jun. 2004, pp. 514–518.

[3] D. Bajic, "On construction of cross-bifix-free kernel sets," presented at the 2nd MCM, Lisbon, Portugal, Feb. 2007, COST 2100, TD07237.

[4] S. Bilotta, E. Pergola, and R. Pinzani, "A new approach to cross-bifix-free sets," *IEEE Trans. Inf. Theory*, vol. 58, no. 6, pp. 4058–4063, Jun. 2012.

[5] G. P. B. Dresden, A simplified Binet formula for k-generalized Fibonacci numbers [Online]. Available: arXiv:0905.0304v2.   , 2011

[6] J. Konc and D. Janezic, "An improved branch and bound algorithm for the maximum clique problem," *MATCH Commun. Math. Comput. Chem.*, vol. 58, pp. 569–590, 2007 [Online]. Available: http://www.sicmm.org/konc/maxclique

[7] C. Levesque, "On m-th order linear recurrences," *Fibonacci Quart.*, vol. 23, no. 4, pp. 290–293, 1985.

[8] J. L. Massey, "Optimum frame synchronization," *IEEE Trans. Commun.*, vol. COM-20, no. 2, pp. 115–119, Apr. 1972.

[9] E. P. Miles, "Generalized Fibonacci numbers and associated matrices," *Amer. Mathe. Monthly*, vol. 67, no. 8, pp. 745–752, Oct. 1960.

[10] M. D. Miller, "On generalized Fibonacci numbers," *Amer. Math. Monthly*, vol. 78, no. 10, pp. 1108–1109, Dec. 1971.

[11] P. T. Nielsen, "On the expected duration of a search for a fixed pattern in random data," *IEEE Trans. Inf. Theory*, vol. IT-19, no. 10, pp. 702–704, Sep. 1973.

[12] M. Schork, "The r-generalized Fibonacci numbers and polynomial coefficients," *Int. J. Contemp. Math. Sci.*, vol. 3, no. 24, pp. 1157–1163, Sep. 2008.

[13] C. Stefanovic and D. Bajic, "On the search for a sequence from a predefined set of sequences in random and framed data streams," *IEEE Trans. Commun.*, vol. 60, no. 1, pp. 189–197, Jan. 2012.

[14] A. J. de Lind van Wijngaarden and T. J. Willink, "Frame synchronization using distributed sequences," *IEEE Trans. Commun.*, vol. 48, no. 12, pp. 2127–2138, Dec. 2000.

Yeow Meng Chee (SM'08) received the B.Math. degree in computer science and combinatorics and optimization and the M.Math. and Ph.D. degrees in computer science, from the University of Waterloo, Waterloo, ON, Canada, in 1988, 1989, and 1996, respectively.

Currently, he is an Associate Professor at the Division of Mathematical Sciences, School of Physical and Mathematical Sciences, Nanyang Technological University, Singapore. Prior to this, he was Program Director of Interactive Digital Media R&D in the Media Development Authority of Singapore, Postdoctoral Fellow at the University of Waterloo and IBM's Zürich Research Laboratory, General Manager of the Singapore Computer Emergency Response Team, and Deputy Director of Strategic Programs at the Infocomm Development Authority, Singapore. His research interest lies in the interplay between combinatorics and computer science/engineering, particularly combinatorial design theory, coding theory, extremal set systems, and electronic design automation.

Han Mao Kiah received the B.Sc.(Hon) degree in mathematics from the National University of Singapore, Singapore in 2006. Currently, he is working towards his Ph.D. degree at the Division of Mathematical Sciences, School of Physical and Mathematical Sciences, Nanyang Technological University, Singapore.

His research interest lies in the application of combinatorics to engineering problems in information theory. In particular, his interests include combinatorial design theory, coding theory and power line communications.

Punarbasu Purkayastha (M'10) received the B.Tech. degree in electrical engineering from Indian Institute of Technology, Kanpur, India in 2004, and the Ph.D. degree in electrical engineering from University of Maryland, College Park, U.S.A., in 2010.

Currently, he is a Research Fellow at the Division of Mathematical Sciences, School of Physical and Mathematical Sciences, Nanyang Technological University, Singapore. His research interests include coding theory, combinatorics, information theory and communication theory.

Chengmin Wang received the B.Math. and Ph.D. degrees in mathematics from Suzhou University, China in 2002 and 2007, respectively.

Currently, he is an Associate Professor at the School of Science, Jiangnan University, China. Prior to this, he was a Visiting Scholar at the School of Computing, Informatics and Decision Systems Engineering, Arizona State University, USA, from 2010 to 2011 and was a Research Fellow at the Division of Mathematical Sciences, School of Physical and Mathematical Sciences, Nanyang Technological University, Singapore from 2011 to 2012. His research interests include combinatorial design theory and its applications in coding theory and cryptography.