IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 27 July 2022, accepted 1 November 2022, date of publication 14 November 2022,
date of current version 6 December 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3222354

THEORY

A Note on the Maximum Size of a Prefix Code

VILIAM HROMADA AND OTOKAR GROŠEK
Faculty of Electrical Engineering and Information Technology, Slovak University of Technology in Bratislava, 812 19 Bratislava, Slovakia
Corresponding author: Viliam Hromada (viliam.hromada@stuba.sk)

This work was supported by Slovenská sporiteľňa, a.s., under Contract 0824/2019//CE.

ABSTRACT In the presented paper, we investigate the problem of finding the maximum possible cardinality of a dictionary of a prefix code for a string of a given length. Namely, we present a sharp proof of the cardinality of such a dictionary using results from the number theory. What is more, the presented formula is for the general case of a string over any, not just binary, alphabet. Furthermore, we give conditions on the existence of the so-called canonical dictionary for such a string, where the codewords of the dictionary have at most two different lengths, differing by one. Our approach is based on reformulating the problem of finding the maximum possible cardinality of a dictionary for a string of a given length as the problem of finding the maximum possible number of summands in the Kraft-Szillard partition of the number representing the length of the string, by solving a Diophantine equation related to the canonical partition of the number. One of the areas of applications of presented results is the security-estimate of ciphers based on prefix codes.

INDEX TERMS Prefix codes, maximum minimal dictionary, partitions of a natural number, Kraft-Szillard partition, cipher based on prefix codes.

I. INTRODUCTION

A prefix code (a prefix-free code) over some q-ary alphabet, or P-code for short, is a code where no codeword is a prefix of another codeword [1], i.e. the codewords are prefix-free. Due to this specific property, the code can be efficiently decoded [5], which has led to the wide-spread adoption of P-codes in many applications. For example, the well-known Huffman code [7], commonly used for loss-less data compression, is a particular type of a binary P-code.

Historically, prefix codes emerged from the area of cryptography. The oldest known example of a P-code is the Argenti code [8] from the 16th century. P-codes were also used for encryption purposes by Peter the Great, where the plaintext was the Cyrillic alphabet [8]. Furthermore, the Soviet cipher known as VIC [9], used P-codes as one of the rounds during the encryption. Nowadays, one area of cryptography that employs prefix codes is the DNA cryptography [15], in which many cryptosystems utilize binary prefix codes as the plaintext space [11], [13], [14].

Example 1: To illustrate the idea of P-codes, let us consider Table 1, where each lowercase letter is encoded (mapped) to a codeword of a binary prefix code; each codeword is of length four or five. The set of such codewords is usually called a dictionary. It is readily seen that no codeword is a prefix of any other codeword. The source text, written as a sequence of lowercase letters, is encoded by substituting each letter with the corresponding binary string and concatenating the strings into the resulting encoded message.

TABLE 1. An example of a binary P-code.

a	b	c	d	e	f	g
00000	00001	00010	00011	1101	00101	00110
h	i	j	k	l	m	n
00111	1110	01001	01010	01011	01100	01101
o	p	q	r	s	t	u
1111	01111	10000	10001	10010	10011	10100
v	w	x	y	z		
10101	10110	10111	11000	11001		

Let the source text be `pcode`. *Substituting each letter by its corresponding binary codeword, the encoded form is* 01111000101111000111101. *Due to the prefix property of codewords, the decoding of* 01111000101111000111101 *is unique, i.e. the message* `pcode`, *and can be done efficiently.*

In [4], the authors propose a modern symmetric encryption algorithm based on binary prefix codes, whose goal is to supplement the set of ciphers deployable in the IoT setting, see e.g. [2]. The key of this cipher is the P-code itself, e.g. both the set of prefix-free codewords and the mapping of some source

The associate editor coordinating the review of this manuscript and approving it for publication was Shuai Liu.

alphabet onto this set. The encryption is done by encoding a given plaintext with the secret P-code, along with the inclusion of so-called null-ciphers, and the resulting ciphertext is then the resulting concatenation of the codewords. For example, if the P-code presented in Example 1 was the secret key, then the plaintext `pcode` would be encrypted as the ciphertext 011110001011110000111101. Since the P-code is known only to legitimate users, an attacker wishing to decrypt (decode) the ciphertext must first try to determine the used secret prefix code. If we assume the attacker is in the possession of the ciphertext, he/she may try to determine the secret P-code by examining all possible ways in which the ciphertext can be split into different potential prefix-free binary words, since they form the candidates for a part of the dictionary of the P-code used for the encryption/decryption, i.e. the key of the cipher. This directly leads to the following problem, which was first studied in the paper [4]:

Given the length n of a binary string x, $|x| = n$, where x is a concatenation of words of an unknown P-code, what is the maximum number $k(n)$ of distinct prefix-free words to which string x can be split?

Example 2: Consider the string $x = 01111000101111000111101$ *from Example 1. This string can be split into distinct prefix-free words in a number of ways, e.g.:*

- $01111-00010-111-1000111101$*, i.e. into four prefix-free words,*
- $011-1100-010-111-10-0011-1101$*, i.e. into seven prefix-free words,*
- $01111-00010-1111-00011-1101$*, i.e. into five prefix-free words (this case corresponds with the dictionary used in Example 1),*
- ...
- 011110001011110000111101*, i.e. we can consider it as one word.*

The string x is of length $|x| = 23$*. It can be shown that binary words of length 23 can be split at most into 7 distinct prefix-free binary words, as seen in the second case, i.e. in the binary case* $k(23) = 7$*. Therefore, if the string x was a ciphertext, the attacker could try to find the dictionary of the used P-code by examining all the possible ways how to split the string x into prefix-free words, which in this case means to try all possible splits of x into sets of* $1, 2, 3, \ldots, 7$ *prefix-free words.*

For $n \leq 26$, the values of $k(n)$ were found by exhaustive search in the paper [4]. They are listed in column k of the cited paper, Table 2. As indicated above, the value $k(n)$ plays an important part in the cryptanalysis of cryptosystems based on prefix codes proposed in paper [4], since the value $k(n)$ provides an upper-bound on the complexity of a possible attack on the cipher in which the attacker would try to enumerate all possible P-codes used in the encryption.

The above-mentioned problem concerns only the binary prefix codes. In this paper, we study the generalization of the problem to q-ary prefix codes, i.e.:

Problem 1: Given the length n of a q-ary string x, $|x| = n$, where x is a concatenation of words of an unknown q-ary P-code, what is the maximum number $k(n)$ of distinct q-ary prefix-free words to which string x can be split?

Furthermore, this problem can be reformulated into a number theory problem of finding an integer partition of $n = n_1 + n_2 + \ldots + n_k$, where the numbers n_i satisfy the so-called Kraft-Szillard inequality [1], [10]. The numbers n_i then represent the lengths of distinct prefix-free words into which the string x can be split.

To our knowledge, the only known results on this problem are the previously mentioned paper [4], which contains experimentally determined values of $k(n)$ for $n \leq 26$ in the binary case and the paper [6], which contains the exact formula for $k(n)$, again in the binary case, determined by geometrical assumptions. In the same paper, the authors prove that for any integer partition of $n = n_1 + n_2 + \ldots + n_k$ there exists a partition of n consisting of k elements where $n_i = a$ or $a + 1$ for a suitable number a. Such partition is called the $(a, a+1)$ canonical partition. Thus, finding the maximal value of k for a canonical partition solves the problem in general.

In this paper, we present a sharp formula for $k(n)$ using results from the number theory. What is more, our formula deals with the general case of a q-ary alphabet. This improves the results of [4] and [6], which focus on the binary case only. Furthermore, we determine the conditions on the existence of two specific types of canonical partitions. Namely, we prove step-by-step in Theorems 3-6 the following result:

Theorem 1: Let q, t be positive integers and n be such an integer that $tq^t \leq n < (t+1)q^{t+1}$. Then the theoretical maximum number of q-ary prefix-free words, into which a q-ary string x of length n can be split, is $k(n) = \lfloor \frac{(q-1)n+q^{t+1}}{(q-1)t+q} \rfloor$*. Moreover, let $n > q$. Then*

1) *If* $\frac{n}{t+1} > \lfloor \frac{(q-1)n+q^{t+1}}{(q-1)t+q} \rfloor$*, a q-ary string of length n might be splittable into q-ary prefix-free words of two lengths $(t+1, t+2)$, such that there are* $\alpha = -n + (t+2)\lfloor \frac{(q-1)n+q^{t+1}}{(q-1)t+q} \rfloor$ *words of the length $(t+1)$ and* $\beta = n - (t+1)\lfloor \frac{(q-1)n+q^{t+1}}{(q-1)t+q} \rfloor$ *words of the length $(t+2)$.*
2) *Otherwise, a q-ary string of length n might be splittable into q-ary prefix-free words of two lengths $(t, t+1)$, such that there are* $\alpha = -n + (t+1)\lfloor \frac{(q-1)n+q^{t+1}}{(q-1)t+q} \rfloor$ *words of the length t and* $\beta = n - t\lfloor \frac{(q-1)n+q^{t+1}}{(q-1)t+q} \rfloor$ *words of the length $(t+1)$.*
3) *There exist exactly* $(q-1)\binom{t+1}{2}$ *values of n, such that strings of length n cannot be split into q-ary prefix-free words of lengths $(t, t+1)$.*

II. PRELIMINARIES

We first recall the definition of a q-ary prefix code. For a set of symbols (or words) $\mathcal{Q}$, $\mathcal{Q}^+$ denotes the set of all non-empty finite concatenations of elements of $\mathcal{Q}$.

Definition 1: Let $\mathcal{A}$ be an alphabet, $\mathcal{Q}$ be a q-ary alphabet and V be a set, $V \subset \mathcal{Q}^+$. Then a q-ary code is a bijection

TABLE 2. Ternary prefix-free code mapping κ from Example 3.

a	b	c	d	e
0	10	11	12	2

$\kappa : \mathcal{A} \rightarrow V$; *$\mathcal{A}$ is called the source alphabet, elements of V are called codewords and V is also called the dictionary of the code. Specifically, a q-ary prefix code, for short a q-ary P-code, is a q-ary code where no codeword is a prefix of another codeword. A message x is a concatenation of finitely many words from the dictionary V.*

Problem 1, presented in the Introduction, can be reformulated as determining the maximum possible cardinality $k(n)$ of the so-called minimal dictionary with respect to some q-ary string x of length n. The definition of the minimal dictionary follows.

Definition 2: Let x be a q-ary string. Then a dictionary V of a P-code with $x \in V^+$ is called minimal with respect to x, if for any $w \in V, x \notin (V \setminus w)^+$.

Example 3: Example 1 provides an example of a binary P-code, where the source alphabet $\mathcal{A} = \{a, b, \ldots, z\}$. For a binary string $x = 01111000101111000111101$, Example 2 provides examples of four different minimal dictionaries with respect to the string x, e.g. $V = \{01111, 00010, 111, 1000111101\}$ or $V = \{011, 1100, 010, 111, 10, 0011, 1101\}$ are both such dictionaries, where $x \in V^+$ and $x \notin (V \setminus w)^+$ for any $w \in V$. What is more, the minimal dictionary $V = \{011, 1100, 010, 111, 10, 0011, 1101\}$ is of the maximum possible cardinality, $|V| = 7$, for a binary string x of length 23.

As an example of a non-binary code, let $\mathcal{A} = \{a, b, c, d, e\}$, and $\mathcal{Q} = \{0, 1, 2\}$ be a ternary ($q = 3$) alphabet. Let $V = \{0, 10, 11, 12, 2\}$ be a set of ternary prefix-free words. Then we can construct a ternary P-code, e.g. by the P-code mapping κ presented in Table 2.

Then, for example the string `bead` *would be encoded as 102012. Examples of minimal dictionaries with respect to string 102012 include dictionaries $V = \{102012\}$, $V = \{1, 02, 012\}$, $V = \{10, 2, 0, 12\}$, but not $\{1, 020, 12\}$, since such a set of words does not fulfil the prefix-free property or $\{10, 2, 0, 12, 11\}$, since such a set is not minimal. From all minimal dictionaries with respect to string 102012, the dictionary $V = \{10, 2, 0, 12\}$ is of the maximum possible cardinality, i.e. $|V| = 4$. The value 4 is also the maximum possible cardinality of a minimal dictionary with respect to any ternary string of length 6, i.e. for $q = 3$, $k(6) = 4$.*

A q-ary prefix code can be represented by a q-ary code tree, in which each internal node has at most q children and the leaves, i.e. the external nodes, represent the codewords and can be labeled with the corresponding characters from the source alphabet.

Example 4: The ternary prefix code from Example 3 can be visualized with the following ternary code tree, where the internal nodes are represented by circles and the external nodes are represented by squares.

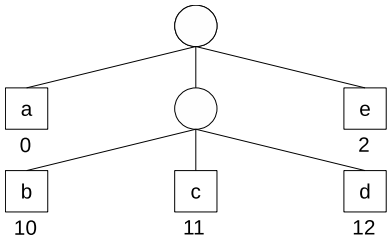

FIGURE 1. Ternary code tree of the ternary prefix code from Example 3.

We now recall the Kraft-Szillard inequality, which states the necessary and sufficient conditions on the existence of prefix codes.

Theorem 2: Let integers $n_1, n_2, \ldots, n_k$ satisfy inequality

$$\sum_{i=1}^{k} q^{-n_i} \leq 1. \tag{1}$$

Then there exists a q-ary P-code with codewords of lengths $n_i, i = 1, 2, \ldots, k$. Conversely, if such numbers do not exist, then there is no unambiguously decodable code with given lengths of codewords.

In order to determine the formula for the maximum possible cardinality $k(n)$ of a minimal dictionary with respect to some q-ary string x of length n, we follow the approach used in [6] and view the problem of finding the maximum possible cardinality of the minimal dictionary as the problem of finding an integer partition of $n = n_1 + n_2 + \ldots + n_k$ with the maximum number of summands k, where the numbers n_i satisfy the Kraft-Szillard inequality [1], [10].

Definition 3: A k-partition of n is a sequence of k natural numbers $n_1, \ldots, n_k$ such that

$$n = n_1 + \ldots + n_k.$$

A partition $n = n_1 + \ldots + n_k$ is a K-S k-partition (Kraft-Szilard) with respect to q, if the numbers n_i satisfy (1). In addition, this partition will be called canonical if $|n_i - n_j| \leq 1$ for all $1 \leq i, j \leq k$. We will denote such canonical partition $(a, a + 1)$, where $a = \min\{n_i\}$.

It can be easily seen that determining the maximum possible cardinality $k(n)$ of a minimal dictionary with respect to some q-ary string x of length n is equal to determining the maximum number k of distinct q-ary substrings $x_1, x_2, \ldots, x_k$ into which a string x of length n can be split, so that each substring x_i is not a prefix of another substring x_j. This in turn means that the lengths of these substrings, $|x_1| = n_1, |x_2| = n_2, \ldots, |x_k| = n_k$ satisfy the Kraft-Szillard inequality (1) and therefore form the K-S k-partition of the number n.

Example 5: Let $q = 3$ and let $x = 0121110$. We are again interested in the maximum possible cardinality of the minimal dictionary with respect to string x. One approach is to try to split the string x into substrings x_i which would be prefix-free and count the number of distinct substrings, e.g.:

- *(the trivial case): 0121110 (no splitting) leads to the dictionary $V = \{0121110\}$ with one word of length 7,*

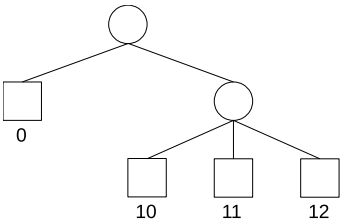

FIGURE 2. Ternary code tree with the maximum number of external nodes w.r.t. $x = 0121110$.

- $01-21110$ *leads to the dictionary* $V = \{01, 21110\}$ *with 2 words of lengths 2 and 5 respectively,*
- $0-12-1110$ *leads to the dictionary* $V = \{0, 12, 1110\}$ *with 3 words of lengths 1, 2, and 4 respectively,*
- $0 - 12 - 11 - 10$ *leads to the dictionary* $V = \{0, 10, 11, 12\}$ *with 4 words, one of length 1 and three of length 2.*
- *Since any such dictionary is prefix-free, the words satisfy the Kraft-Szillard inequality.*
- *The string x cannot be split into 5 words, which would be prefix-free, therefore the maximum cardinality of the corresponding minimal dictionary would be 4, as seen for example in the case* $V = \{0, 10, 11, 12\}$.

Equivalently, if we use the tree-representation of a prefix code, we are looking for a ternary tree with the maximum number of external nodes, where the given string 0121110 *is formed by a concatenation of codewords corresponding to all external nodes, in some order. Such a ternary tree, with 4 external nodes, is presented in Figure 2.*

Generally, for any ternary string of length $n = 7$*, we are looking for such a partition into substrings of lengths* $n_1, \ldots, n_k$ *so that* $7 = n_1 + \ldots + n_k$*, which would satisfy the Kraft-Szillard inequality where* $q = 3$*. There is for example the K-S 4-partition of the number 7 into* $7 = 1 + 2 + 2 + 2$*, i.e.* $n_1 = 1, n_2 = n_3 = n_4 = 2$*, since* $3^{-1} + 3^{-2} + 3^{-2} + 3^{-2} \leq 1$*, i.e. the potential lengths of substrings are 1, 2, 2 and 2, as seen in the partition of* 0121110 *into* $0 - 12 - 11 - 10$*. Note that this partition is also the canonical partition, since* $|n_i - n_j| \leq 1$ *and since the values are* 1 *and* 2*, it is denoted the* $(1, 2)$ *canonical partition.*

Furthermore, there does not exist a K-S 5-partition of 7, since there does not exist any way how to split a ternary string of length 7 into 5 substrings whose lengths would satisfy the Kraft-Szillard inequality. Therefore, in the ternary case, $k(7) = 4$.

Note that the value 4 of the maximum possible cardinality for minimal dictionaries of ternary strings of length 7 is the theoretical maximum. Some ternary strings of length 7 may have minimal dictionaries of cardinalities strictly smaller, e.g. the string 0000000 *has only minimal dictionaries* $V = \{0\}$ *and* $V = \{0000000\}$*, both of cardinality one.*

In the next section, we derive the formula for the maximum value of k such that the K-S k-partition of the number n exists, for the general q-ary case. This value k then provides the maximum possible cardinality $k(n)$ of a minimal dictionary with respect to some q-ary string of length n, or in another terms, the maximum number $k(n)$ of distinct q-ary prefix-free words to which a q-ary string of length n could be split.

Furthermore, we will determine the condition for the existence of two special canonical K-S k-partitions, $(t, t+1)$ and $(t+1, t+2)$ partitions.

III. CANONICAL *K-S k*-PARTITIONS OF *N*

In [6], the authors prove, for the binary case ($q = 2$), that for every n there exists a maximum k such that a canonical K-S k-partition of n exists. We will prove a generalized version of the same result, for an arbitrary q, using the number theory. Hereafter, we denote $\mathbb{N} = \{1, 2, \ldots\}$ and $\mathbb{N}_0 = \{0, 1, 2, \ldots\}$. For $n \in \mathbb{N}$, $k(n)$ will stand for the largest integer k such that there exists a K-S k-partition of n.

Throughout this paper, we will use the following fact: For a given integer q and any integer n, there exists a unique integer t such that

$$t\,q^t \leq n < (t+1)\,q^{t+1}. \tag{2}$$

Let $N_t = \{n > 2 : t\,q^t \leq n < (t+1)\,q^{t+1}\}$. Then $|N_t| = q^t((q-1)t + q)$. For further purposes, we derive: If $n < (a+1)\,q^{a+1}$, then $n(q-1) < (q-1)(a+1)\,q^{a+1}$ and $n(q-1) + q^{a+1} < (q-1)(a+1)q^{a+1} + q^{a+1}$ which yields

$$\frac{n(q-1) + q^{a+1}}{(q-1)a + q} < q^{a+1}. \tag{3}$$

On the other hand, if $aq^a \leq n$, then $(q-1)aq^a \leq n(q-1)$ and $(q-1)aq^a + q^{a+1} \leq n(q-1) + q^{a+1}$ which yields

$$q^a \leq \frac{n(q-1) + q^{a+1}}{(q-1)a + q}. \tag{4}$$

Thus

$$q^a \leq \frac{n(q-1) + q^{a+1}}{(q-1)a + q} < q^{a+1}. \tag{5}$$

Suppose now that for all n_i, $n_i = a$ or $n_i = a + 1$. Hence

$$n = \alpha a + \beta(a+1), \tag{6}$$

i.e. we have α members of the partition of value a, and β of value $a+1$, where $\alpha + \beta = k$ is the number of elements of the partition, $\alpha, \beta \in \mathbb{N}_0$. In our special case, all solutions α, β of this equation can be found as follows [12]:

1) Since $\gcd(a, a+1) = 1$, one can easily find integers u and v such that $au + (a+1)v = 1$. Moreover, in this case we can set $u = -1$, $v = 1$.
2) Thus we have one solution of (6), namely $\alpha_0 = nu = -n$, $\beta_0 = nv = n$.
3) Therefore every integer solution (α, β) can be written as $(\alpha_0 + (a+1)j, \beta_0 - aj), j = 0, \pm1, \pm2, \ldots$, or $(-n + (a+1)j, n - aj), j = 0, \pm1, \pm2, \ldots$.
4) In our case $\alpha, \beta \in \mathbb{N}_0$. Thus j must satisfy

$$\frac{n}{a+1} \leq j \leq \frac{n}{a} \tag{7}$$

and the number of non-negative solutions is not greater than

$$\frac{n}{a} - \frac{n}{a+1} + 1 = \frac{n}{a(a+1)} + 1.$$

Since our partition must be a K-S k-partition, we have

$$\frac{\alpha}{q^a} + \frac{\beta}{q^{a+1}} \leq 1, \quad \text{or} \quad q\alpha + \beta \leq q^{a+1}. \tag{8}$$

This yields

$$\begin{aligned} &q(-n + (a+1)j) + (n - aj) \\ &\quad = -n(q-1) + (q(a+1) - a)j \leq q^{a+1}, \end{aligned} \tag{9}$$

or

$$j \leq \frac{n(q-1) + q^{a+1}}{(q-1)a + q}, \tag{10}$$

and it is not difficult to prove that for $n \in N_a$, it follows from $aq^a \leq n$ that

$$\frac{n(q-1) + q^{a+1}}{(q-1)a + q} \leq \frac{n}{a}. \tag{11}$$

On the other hand, it follows from $n < (a+1)q^{a+1}$ that

$$\frac{n}{a+1} < \frac{n(q-1) + q^{a+1}}{(q-1)a + q}. \tag{12}$$

Since $\alpha + \beta = -n + (a+1)j + n - aj = j$, to have $k(n) = \max\{\alpha + \beta\}$, we must also have j maximal. At the same time, assuming (2), the smallest $a = t$ and we have obtained the proof of the following theorem, which states the existence of the $(t, t+1)$ canonical K-S k-partition of n.

Theorem 3: Let $n \in N_t$ and $\lfloor \frac{n(q-1)+q^{t+1}}{(q-1)t+q} \rfloor \in [\frac{n}{t+1}, \frac{n}{t}]$. Then there exists a solution of the equation $n = \alpha t + \beta(t+1)$ satisfying conditions $q\alpha + \beta \leq q^{t+1}$ and $k(n) = \max\{\alpha + \beta\} = \max\{j\}$ where α, β are non-negative integer solutions. This solution is attained for $j = \lfloor \frac{n(q-1)+q^{t+1}}{(q-1)t+q} \rfloor = k(n)$.

Now, we illustrate our steps with two Examples.

Example 6: Let $q = 2, n = 21$. Then $2 \times 2^2 \leq 21 < 3 \times 2^3$, i.e. $t = 2$ and we put $a = t = 2$. Thus we have

$$2\alpha + 3\beta = 21.$$

Since $\gcd(2, 3) = 1$, there exist solutions of our equation of the form $(-21 + 3j, 21 - 2j)$. Considering only non-negative solutions, $\frac{21}{3} \leq j \leq \frac{21}{2}$, yields $j = 7, 8, 9, 10$. Thus all non-negative solutions $(\alpha, \beta) = (9, 1), (6, 3), (3, 5), (0, 7)$.

At the same time, $j \leq \frac{n(q-1)+q^{a+1}}{a(q-1)+q} = \frac{21+8}{4} = 7.25$. Thus there exists $j \in \mathbb{N}$ satisfying both conditions, namely $j = 7$ and thus there exists one canonical K-S 7-partition $(2, 3)$ of the number 21, i.e. $21 = 2 \times 0 + 3 \times 7$.

Moreover, we may try to find a $(t+1, t+2)$ partition. Then

$$\begin{aligned} &3\alpha + 4\beta = 21, \\ &(\alpha, \beta) = (-21 + 4j, 21 - 3j) = (3, 3), (7, 0), \\ &\frac{21}{4} \leq j \leq \frac{21}{3}, \end{aligned}$$

i.e. $j = 6, 7$. Both values of j satisfy (10), and thus $k = \alpha + \beta = 6$ or 7, and $k(n) = \max j = 7$, the same as in the case of $(t, t+1)$ partition, i.e. in this case the canonical K-S 7-partition $(3, 4)$ of the number 21 is $21 = 3 \times 7 + 4 \times 0$.

Example 7: Let $q = 3, n = 70$. Then $2 \times 3^2 \leq 70 < 3 \times 3^3$, i.e. $t = 2$. Thus we have

$$2\alpha + 3\beta = 70.$$

All solutions are of the form $(-70 + 3j, 70 - 2j)$. Considering only non-negative solutions, $\frac{70}{3} \leq j \leq \frac{70}{2}, j = 24, 25, \ldots, 35$. This yields all non-negative solutions $(\alpha, \beta) = (2, 22), (5, 20), (8, 18), (11, 16), (14, 14), (17, 12), (20, 10), (23, 8), (26, 6), (29, 4), (32, 2), (35, 0)$.

At the same time $j \leq \frac{n(q-1)+q^{a+1}}{a(q-1)+q} = \frac{140+27}{7} \doteq 23.86$. Thus, since $23 \notin \{24, 25, \ldots, 35\}$ there does not exist j satisfying both conditions.

But if we let $a = t + 1 = 3$, i.e.

$$3\alpha + 4\beta = 70,$$

then $\frac{70}{4} \leq j \leq \frac{70}{3}, j = 18, 19, \ldots, 23$. This yields all non-negative solutions $(\alpha, \beta) = (2, 16), (6, 13), (10, 10), (14, 7), (18, 4), (22, 1)$.

Moreover, $j \leq \frac{n(q-1)+q^{a+1}}{a(q-1)+q} = \frac{140+81}{9} \doteq 24.56$. Thus there exist 6 canonical K-S k-partitions $(3, 4)$ of the number 70 with $k = \alpha + \beta = 18, 19, \ldots, 23$, $\max k = 23$. This example also shows that to have a maximal $k = k(n)$, it is not sufficient just to find some canonical K-S k-partition, since e.g. the canonical partition $(3, 4)$ with $(\alpha, \beta) = (2, 16)$ has $k = 18$, but the canonical partition $(3, 4)$ with $(\alpha, \beta) = (22, 1)$ has $k = 23$, the maximum one. Therefore, in this case, the canonical partition $(3, 4)$ of the number 70 with maximum $k = 23$ is $70 = 3 \times 22 + 4 \times 1$.

We emphasize that if $j = \lfloor \frac{n(q-1)+q^{t+1}}{(q-1)t+q} \rfloor \notin [\frac{n}{t+1}, \frac{n}{t}]$, i.e. the value of j is smaller than the minimal integer on the left end of the interval $[\frac{n}{t+1}, \frac{n}{t}]$, then it must be the first integer on the right end of the interval $[\frac{n}{t+2}, \frac{n}{t+1}]$. This suggests to let $a = t + 1$ in (6), which leads to the following theorem on the existence of the $(t+1, t+2)$ canonical K-S k-partition of n.

Theorem 4: Let $n \in N_t$ and $\lfloor \frac{n(q-1)+q^{t+1}}{(q-1)t+q} \rfloor \in [\frac{n}{t+2}, \frac{n}{t+1}]$. Then there exists a solution of the equation $n = \alpha(t+1) + \beta(t+2)$ satisfying conditions $q\alpha + \beta \leq q^{t+2}$ and $k(n) = \max\{\alpha + \beta\} = \max\{j\}$ where α, β are non-negative integer solutions. This solution is attained for $j = \lfloor \frac{n(q-1)+q^{t+1}}{(q-1)t+q} \rfloor = k(n)$.

Proof: All values of j leading to non-negative solutions α, β are in the interval $[\frac{n}{t+2}, \frac{n}{t+1}]$. Under our supposition $\lfloor \frac{n(q-1)+q^{t+1}}{(q-1)t+q} \rfloor$ belongs to this interval, and thus

$$\begin{aligned} &(\alpha, \beta) = (-n + (t+2)j, n - (t+1)j), \\ &j = \left\lceil \frac{n}{t+2} \right\rceil, \ldots, \left\lfloor \frac{n(q-1) + q^{t+1}}{(q-1)t + q} \right\rfloor. \end{aligned}$$

Moreover,

$$\max\{\alpha+\beta\} = \max\{j\} = \left\lfloor \frac{n(q-1)+q^{t+1}}{(q-1)t+q} \right\rfloor.$$

Now, we must prove that this solution also satisfies (9), i.e.

$$\begin{aligned}
&q\alpha+\beta \\
&= -n(q-1) + ((t+1)(q-1)+q)\left\lfloor \frac{n(q-1)+q^{t+1}}{(q-1)t+q} \right\rfloor \\
&= -((q-1)t+q)\frac{n(q-1)+q^{t+1}}{(q-1)t+q} + q^{t+1} \\
&\quad + ((t+1)(q-1)+q)\left\lfloor \frac{n(q-1)+q^{t+1}}{(q-1)t+q} \right\rfloor \\
&= ((q-1)t+q)\left(-\frac{n(q-1)+q^{t+1}}{(q-1)t+q} \right. \\
&\quad \left. + \left\lfloor \frac{n(q-1)+q^{t+1}}{(q-1)t+q} \right\rfloor\right) \\
&\quad + q^{t+1} + (q-1)\left\lfloor \frac{n(q-1)+q^{t+1}}{(q-1)t+q} \right\rfloor.
\end{aligned}$$

Since $-1 \le \lfloor x \rfloor - x \le 0$, and by (3) we have

$$\begin{aligned}
q\alpha+\beta &\le q^{t+1} + (q-1)\left\lfloor \frac{n(q-1)+q^{t+1}}{(q-1)t+q} \right\rfloor \\
&\le q \times q^{t+1} = q^{t+2}.
\end{aligned}$$

This finishes the proof. □

Now, we will find a condition under which we can distinguish to which interval, $[\frac{n}{t+2}, \frac{n}{t+1}]$ or $[\frac{n}{t+1}, \frac{n}{t}]$, $\left\lfloor \frac{n(q-1)+q^{t+1}}{(q-1)t+q} \right\rfloor$ belongs to. This is determined by values on the left side of the interval $[\frac{n}{t+1}, \frac{n}{t}]$. If $\frac{n}{t+1} > \left\lfloor \frac{n(q-1)+q^{t+1}}{(q-1)t+q} \right\rfloor$, then necessarily $\left\lfloor \frac{n(q-1)+q^{t+1}}{(q-1)t+q} \right\rfloor \in [\frac{n}{t+2}, \frac{n}{t+1}]$ which yields that there is a $(t+1, t+2)$ partition only. Thus we proved the following theorem:

Theorem 5: Let $n \in N_t$. If $\frac{n}{t+1} > \left\lfloor \frac{n(q-1)+q^{t+1}}{(q-1)t+q} \right\rfloor$, then there exists a $(t+1, t+2)$ K-S partition of n, only. Otherwise, there exists a $(t, t+1)$ partition. In both cases, $k(n) = \lfloor \frac{n(q-1)+q^{t+1}}{(q-1)t+q} \rfloor$.

Example 8: Let $q = 2, t = 2$. Then $|N_t| = 2^2(2+2) = 16, N_t = \{8, 9, \ldots, 23\}$. There are precisely 3 integers n such that there exists a $(t+1, t+2)$ partition only. Namely for $n = 19, 22, 23$. For $n = 19$, the value $k(n) = k(19) = 6$ and for $n = 22$ and $n = 23$, the value $k(22) = k(23) = 7$. This in turn means w.r.t. our original problem that the maximum cardinality of the minimal dictionary of a binary string of length 19 can be 6, and the maximum cardinality of the minimal dictionary of a binary string of length 22 or 23 can be 7.

Let $q = 3, t = 2$. Then $|N_t| = 3^2(2 \times 2 + 3) = 63, N_t = \{18, 19, \ldots, 80\}$. There are precisely 6 integers n such that there exists a $(t+1, t+2)$ partition only, namely for $n = 70, 73, 76, 77, 79, 80$. For $n = 70$, we have $k(70) = 23$, for $n = 73$ we have $k(73) = 24$, for $n = 76$ and $n = 77$ we have $k(76) = k(77) = 25$, and for $n = 79$ and $n = 80$ we have $k(79) = k(80) = 26$, which in turn means w.r.t. our original problem that a ternary string of length 70 might be splittable up to 23 ternary prefix-free substrings, a ternary string of length 73 might be splittable up to 24 ternary prefix-free substrings, etc.

Consider now the numbers $n \in N_t$ which do not have a K-S $(t, t+1)$ partition, i.e. for these numbers $\frac{n}{t+1} > \lfloor \frac{(q-1)n+q^{t+1}}{(q-1)t+q} \rfloor$. Let us denote the remainder of $(q-1)n + q^{t+1}$ after division by $(q-1)t+q$ as r_n, i.e.

$$r_n = \left((q-1)n + q^{t+1}\right) \bmod (q-1)t+q. \tag{13}$$

Since $\lfloor \frac{(q-1)n+q^{t+1}}{(q-1)t+q} \rfloor = \frac{(q-1)n+q^{t+1}-r_n}{(q-1)t+q}$, the inequality $\frac{n}{t+1} > \lfloor \frac{(q-1)n+q^{t+1}}{(q-1)t+q} \rfloor$ can be rewritten as

$$q^{t+1} - \frac{n}{t+1} < r_n. \tag{14}$$

Therefore, a number n does not have a K-S $(t, t+1)$ partition if its corresponding value r_n satisfies (14). Now, we will find all integers $n \in N_t$ satisfying (14). For simplicity, we will denote the modulus $(q-1)t+q$ as m, i.e. $m = (q-1)t+q = (q-1)(t+1)+1$. We will use the following notation: we can arrange all integers from the set N_t, $|N_t| = q^t m$, into a matrix of the size $q^t \times m$ as follows

$$\mathbf{N}_t = \begin{pmatrix} n_0^{(0)} & n_1^{(0)} & \ldots & n_{m-1}^{(0)} \\ n_0^{(1)} & n_1^{(1)} & \ldots & n_{m-1}^{(1)} \\ \vdots & \vdots & \ldots & \vdots \\ n_0^{(q^t-1)} & n_1^{(q^t-1)} & \ldots & n_{m-1}^{(q^t-1)} \end{pmatrix}.$$

Here $N^{(i)}$ are rows and M_j are columns, where $i = 0, 1, \ldots, q^t - 1$, and $j = 0, 1, \ldots, m-1$. Thus $n_j^{(i)} = tq^t + im + j$.

Now we must identify all numbers $n = n_j^{(i)} \in N_t$ which satisfy (14). This will be done by proving the next 3 lemmas.

Lemma 1: Let $n_j^{(i)}$ and $r_{n_j^{(i)}}$, where $0 \le j \le (q-1)(t+1)$ be as above. Then, for each $n \in N_t$ there exist unique integers u, k such that $j = (t+1)u + k$, $0 \le u \le q-1$, $0 \le k \le t$, and

$$r_{n_j^{(i)}} = \begin{cases} (q-1)k - u; & \text{if } (q-1)k - u \ge 0 \\ (q-1)k - u + m; & \text{otherwise.} \end{cases} \tag{15}$$

Proof: We start with the calculation of

$$\begin{aligned}
r_{n_j^{(i)}} &= \left((q-1)n_j^{(i)} + q^{t+1}\right) \bmod m \\
&= q^t m + (q-1)im + (q-1)j \bmod m \\
&= (q-1)j \bmod m.
\end{aligned}$$

Since $\gcd(q-1, m) = 1$, $r_{n_j^{(i)}}$ is uniquely determined for $0 \le j < m$, i.e. its value remains the same within each column M_j. Further, for $j = (t+1)u + k$ this yields

$$(q-1)((t+1)u + k) \bmod m = (q-1)k - u \bmod m.$$

□

Now we will calculate the left and right sides of (14) for a special case of $i = q^t - 1 - t, j = t + 1$. On the left side we have

$$\begin{aligned} & q^{t+1} - \frac{n_{t+1}^{(q^t-1-t)}}{t+1} \\ &= q^{t+1} - \frac{tq^t + (q^t - 1 - t)((q-1)t + q) + (t+1)}{t+1} \\ &= \frac{q(t^2+2t+1) - (t^2+2t+1)}{t+1} = \frac{(q-1)(t+1)(t+1)}{t+1} \\ &= (q-1)(t+1). \end{aligned}$$

On the right side if $j = t + 1$, then for all numbers in the column M_{t+1} we have

$$r_{n_{t+1}^{(i)}} = -1 + (q-1)(t+1) + 1 = (q-1)(t+1).$$

Thus in this special case left and right sides of (14) are the same. Moreover, since the left hand side sequence is decreasing and $r_n \leq (q-1)(t+1)$, it follows that $n_{t+1}^{(q^t-1-t)}$ is the first number in the matrix where the left and right sides of (14) are the same. Thus we proved

Lemma 2: The first $n \in N_t$, where $q^{t+1} - \frac{n}{t+1} = r_n$, is $n_{t+1}^{(q^t-1-t)}$.

The next lemma states some periodical properties of entries in the matrix $\mathbf{N}_t$.

Lemma 3: The following is valid for numbers $n_j^{(i)}$:

1) *Let $h = t + 1, j + h < m$. If $q^{t+1} - \frac{n_j^{(i)}}{t+1} = r_{n_j^{(i)}}$, then $q^{t+1} - \frac{n_{j+h}^{(i)}}{t+1} = r_{n_{j+h}^{(i)}}$.*
2) *Let $h = t + 1, j + h \geq m$. If $q^{t+1} - \frac{n_j^{(i)}}{t+1} = r_{n_j^{(i)}}$, then $q^{t+1} - \frac{n_{j+h-m}^{(i+1)}}{t+1} = r_{n_{j+h-m}^{(i+1)}}$.*
3) *The number of places where $q^{t+1} - \frac{n}{t+1} = r_n$ in each row, beginning with the row $N_j^{(q^t-1-t)}$, is $q - 1$.*

Proof:

1) Under our supposition it follows that
$$\frac{n_{j+h}^{(i)} - n_j^{(i)}}{t+1} + r_{n_{j+h}^{(i)}} - r_{n_j^{(i)}} = 1 + h(q-1) \bmod m = 0.$$
2) Again direct calculation yields
$$\begin{aligned} & \frac{n_{j+h-m}^{(i+1)} - n_j^{(i)}}{t+1} + r_{n_{j+h-m}^{(i)}} - r_{n_j^{(i)}} \\ &= 1 + (h-m)(q-1) \bmod m = 0. \end{aligned}$$
3) Since we start at the position $t + 1$, and the next such situation occurs after each $h = t + 1$ entries, we have $(q-1)(t+1)$ such positions altogether. In the row $N^{(q^t-1-t)}$, the last position where this happens is the last column. Therefore, due to 2., in the next row, such position starts at the position t, etc. Thus in the last row $N^{(q^t-1)}$, first such position is in the column M_1 and the last one is in column $M_{1+(q-2)(t+1)}$.

□

From Lemma 1 it follows that $r_j^{(i)}$ does not depend on i, i.e. is the same for a fixed j. Thus, if both sides of (14) are equal for $n_j^{(i)}$, then all $n_j^{(i+k)}, k = 1, 2, \ldots$ satisfy (14). By Lemma 2, the first number with both sides of (14) equal is $n_{t+1}^{(q^t-1-t)}$ in row $N^{(q^t-1-t)}$ and column M_{t+1}. The column M_{t+1} therefore contains t numbers that satisfy (14). And by Lemma 3, also columns $M_{(t+1)+k(t+1)}, k = 1, 2, \ldots, (q-2)$ contain t numbers that satisfy (14), so this argument leads to $(q-1)t$ numbers that satisfy (14).

Further, it follows from Lemma 3 that the first number in row $N^{(q^t-t)}$ with both sides of (14) equal is $n_t^{(q^t-t)}$, therefore the column M_t contains $t - 1$ numbers satisfying (14). And again by Lemma 3, there are altogether $q - 1$ columns, $M_{t+k(t+1)}, k = 1, 2, \ldots, (q-2)$, which contain $(t-1)$ numbers that satisfy (14), i.e. this argument leads to another $(q-1)(t-1)$ numbers that satisfy (14), and so on.

The resulting number of $n_j^{(i)}$, which satisfy (14) is therefore

$$\begin{aligned} & (q-1)(t + (t-1) + (t-2) + \ldots + 2 + 1) \\ &= (q-1)\frac{t(t+1)}{2} = (q-1)\binom{t+1}{2}, \end{aligned}$$

which proves the following theorem on the non-existence of $(t, t+1)$ partition for some numbers n.

Theorem 6: Let $n \in N_t$. Then the number of integers n for which there does not exist a $(t, t+1)$ K-S partition is $(q-1)\binom{t+1}{2}$.

Combining the results in Theorems 3-6, we obtain the following theorem, which summarizes the results presented in this paper and is a reformulation of Theorem 1, in terms of K-S k-partitions of n with respect to q.

Theorem 7: Let q, t be positive integers and let n be such an integer that $tq^t \leq n < (t+1)q^{t+1}$. Then the maximum k, for which there exists a K-S k-partition of n is $k = \lfloor \frac{(q-1)n+q^{t+1}}{(q-1)t+q} \rfloor$, with respect to q. Moreover, let $n > q$. Then

1) *If $\frac{n}{t+1} > \lfloor \frac{(q-1)n+q^{t+1}}{(q-1)t+q} \rfloor$, there exists a canonical $(t+1, t+2)$ K-S k-partition of n, with respect to q, $n = \alpha(t+1) + \beta(t+2)$, such that $\alpha = -n + (t+2)\lfloor \frac{(q-1)n+q^{t+1}}{(q-1)t+q} \rfloor$ and $\beta = n - (t+1)\lfloor \frac{(q-1)n+q^{t+1}}{(q-1)t+q} \rfloor$.*
2) *Otherwise, there exists a canonical $(t, t+1)$ K-S k-partition of n, with respect to q, $n = \alpha(t) + \beta(t+1)$, such that $\alpha = -n + (t+1)\lfloor \frac{(q-1)n+q^{t+1}}{(q-1)t+q} \rfloor$ and $\beta = n - t\lfloor \frac{(q-1)n+q^{t+1}}{(q-1)t+q} \rfloor$.*
3) *There exist exactly $(q-1)\binom{t+1}{2}$ values of n, such that there does not exist the $(t, t+1)$ canonical K-S k-partition of n, with respect to q.*

IV. CONCLUSION

In this paper, we investigated the maximum cardinality of minimal dictionaries of strings of q-ary prefix codes. We have reformulated the problem of finding such cardinalities, as the problem of finding the maximum number of summands k in the q-ary K-S k-partition of the number n.

The main result of our paper, a sharp formula on the maximum k such that the K-S k-partition of n exists, is presented in Theorem 7 and proven per-partes in Theorems 3-6. Its practical implications include the security-estimate of encryption systems based on prefix codes, e.g. [4], since the value $k(n)$ provides an upper-bound on the complexity of a possible attack on the cipher, in which the attacker would try to exhaustively search through all possible P-codes potentially used to create a given q-ary string of length n. For example, if the attacker obtains a binary ($q = 2$) ciphertext of length 1024, it can be split into $k(1024) = 142$ prefix-free words, i.e. the attacker can potentially create minimal dictionaries of cardinalities 1, 2, 3, up to 142, which could all have been a part of the used key of the cipher. Obviously, the larger the number $k(n)$, the longer it takes the attacker to create all the possible dictionaries.

Furthermore, we have also determined the conditions of the existence of the canonical K-S k-partition of n of type $(t, t+1)$, where $tq^t \leq n < (t+1)q^{t+1}$.

ACKNOWLEDGMENT

The authors would like to thank the anonymous reviewers whose insightful comments have helped to improve the quality of this article.

REFERENCES

[1] J. Adamek, *Foundations of Coding: Theory and Applications of Error-Correcting Codes With an Introduction to Cryptography and Information Theory*. Hoboken, NJ, USA: Wiley, 1991.
[2] S. Balogh, O. Gallo, R. Ploszek, P. Špaček, and P. Zajac, "IoT security challenges: Cloud and blockchain, postquantum cryptography, and evolutionary techniques," *Electronics*, vol. 10, no. 21, p. 2647, Oct. 2021, doi: 10.3390/electronics10212647.
[3] R. G. Gallager, *Information Theory and Reliable Communication*. New York, NY, USA: Wiley, 1968.
[4] O. Grošek, V. Hromada, and P. Horák, "A cipher based on prefix codes," *Sensors*, vol. 21, no. 18, p. 6236, Sep. 2021, doi: 10.3390/s21186236.
[5] D. S. Hirschberg and D. A. Lelewer, "Efficient decoding of prefix codes," *Commun. ACM*, vol. 33, no. 4, pp. 449–459, Apr. 1990, doi: 10.1145/77556.77566.
[6] P. Horák, V. Hromada, and O. Grošek, "On the maximum size of a prefix code," *TechRxiv*, vol. 19866205, pp. 1–15, Jun. 2022, doi: 10.36227/techrxiv.19866205.v1.
[7] D. A. Huffman, "A method for the construction of minimum-redundancy codes," *Proc. Inst. Radio Eng.*, vol. 40, no. 9, pp. 1098–1101, Sep. 1952, doi: 10.1109/JRPROC.1952.273898.
[8] D. Kahn, *The Codebreakers*. London, U.K.: Weidenfeld and Nicolson, 1967.
[9] D. Kahn, *Kahn on Codes: Secrets of the New Cryptology*. New York, NY, USA: Macmillan, 1984.
[10] B. Mandelbrot, "Leo Szilard and unique decipherability (Corresp.)," *IEEE Trans. Inf. Theory*, vol. IT-11, no. 3, pp. 455–456, Jul. 1965, doi: 10.1109/TIT.1965.1053782.
[11] M. Meftah, A. A. Pacha, and N. Hadj-Said, "DNA encryption algorithm based on Huffman coding," *J. Discrete Math. Sci. Cryptogr.*, vol. 25, pp. 1–14, Dec. 2020, doi: 10.1080/09720529.2020.1818450.
[12] I. Niven, H. S. Zuckerman, and H. L. Montgomery, *An Introduction to Theory of Numbers*, 5th ed. New York, NY, USA: Wiley, 1991.
[13] A. Sen, R. Roy, and S. R. Dash, "Implementation of public key cryptography in DNA cryptography," in *Proc. Adv. DNA Comput. Cryptogr.*, S. Namasudra and G. C. Deka, Eds. Boca Raton, FL, USA: Chapman & Hall, 2018, pp. 20–36.
[14] H. Shaw, "A cryptographic system based upon the principles of gene expression," *Cryptography*, vol. 1, no. 3, p. 21, Nov. 2017, doi: 10.3390/cryptography1030021.
[15] G. Xiao, M. Lu, L. Qin, and X. Lai, "New field of cryptography: DNA cryptography," *Chin. Sci. Bull.*, vol. 51, no. 12, pp. 1413–1420, Jun. 2006, doi: 10.1007/s11434-006-2012-5.

VILIAM HROMADA was born in Piešany, Slovakia, in 1987. He received the bachelor's, master's, and Ph.D. degrees in applied informatics from the Slovak University of Technology in Bratislava, Slovakia, in 2009, 2011, and 2014, respectively.

Since 2014, he has been a Lecturer/Researcher with the Faculty of Electrical Engineering and Information Technology, Institute of Computer Science and Mathematics, Slovak University of Technology in Bratislava. His research interests include post-quantum cryptography and side-channel cryptanalysis.

OTOKAR GROŠEK received the M.Sc. degree in applied mathematics from Comenius University in Bratislava, in 1973, and the Ph.D. degree in applied mathematics from the Slovak Technical University in Bratislava, in 1978.

From 2004 to 2010, he was the Chairperson with the Department of Mathematics, Slovak University of Technology in Bratislava. From 2011 to 2019, he was the Director of the Institute of Computer Science and Mathematics, Slovak University of Technology in Bratislava. From 2014 to 2017, he was the NATO Project Director of the team working on the NATO SPS Project G4520 "Secure Implementation of Post-Quantum Cryptography." The project was recognized as the outstanding NATO SPS multi-year project of the last decade in the category "Cyber Defense" and was awarded the NATO Science Partnership Prize 2018. He is currently the NATO Project Director of the team working on the NATO SPS Project G5448 "Secure Communication in the Quantum Era."