SIAM J. DISC. MATH.
Vol. 1, No. 3, August 1988

008

SEPARATING PARTITION SYSTEMS AND LOCALLY DIFFERENT SEQUENCES*

JÁNOS KÖRNER† AND GÁBOR SIMONYI†

Abstract. The problem of perfect hashing is generalized and some initial results are obtained. As a corollary, an improvement on earlier results for (i, j)-separating systems of partitions is provided.

Key words. separating partition systems, code distance

AMS(MOS) subject classifications. primary O5B40; secondary O5A15, 94A10

1. Separating partition systems. Separating partition systems have been studied under different names by several authors. Our attention was drawn to this subject by a very stimulating paper by Fredman and Komlós [1] who used information theory to derive nonexistence bounds for separating partition systems in two special cases: systems of perfect hash functions and (i, j)-separating systems. For the latter model, stronger results exist in Djackov and Rykov [2], Erdös, Frankl, and Füredi [3], and Hwang and Sós [4] in the case $i = 1$. However, if $i > 1, j > 1$, no appropriate method seems to be available. By establishing an interesting connection between this problem and a new generalization of perfect hashing, we will improve upon earlier results for (2. 2)-separating systems; the special case is singled out in the pioneering paper of Friedman, Graham, and Ullman [5].

DEFINITION [5]. An (i, j)-*separating* system for the set S is a family of bipartitions $P_1, \cdots, P_t$ of S such that for every pair of disjoint subsets A, B of S such that $|A| = i$, $|B| = j$, there is at least one partition P_v in the given family for which A and B are contained in (the two) different classes of P_v.

Write $|S| = n$ and denote by $M(i, j, n)$ the minimum number t of partitions in any (i, j)-separating partition system (SPS) for S. In the papers [1]–[5] asymptotic bounds have been obtained for $M(i, j, n)$ if n tends to infinity, while i and j remain fixed. The case $i = 1$ seems to have raised the most interest since it has applications in conflict resolution in multi-access communication. The case $i > 1, j > 1$ is treated only in [1] and [5]. We will return to it at the end of this paper in order to improve on the asymptotic bounds for $M(2, 2, n)$.

Our starting point is, however, a different separating partition system, known under the name of perfect hash functions.

DEFINITION [1]. A (b, k)-*system of perfect hash functions* for the set S is a system of partitions $P_1, \cdots, P_t$ of S into at most b classes such that for every k-element subset A of S there is at least one partition P_v in the system for which every element of A falls into a different class of P_v. (Clearly, such a system exists only if $k \leqq b$.)

Following Fredman and Komlós [1], we denote by $Y(b, k, n)$ the smallest t for which a (b, k)-system of perfect hash functions exists for $|S| = n$. The best asymptotic bounds for $Y(b, k, n)$ in the case of $n \to \infty$, arbitrary finite b and k can be found in Körner and Marton [6] (cf. also [1] and Körner [7]). The problem seems to be hopelessly difficult and no exact asymptotic value of $Y(b, k, n)$ is known except for the trivial case $b = 2$.

* Received by the editors February 24, 1988; accepted for publication March 9, 1988.

† Department Informatique, Ecole Nationale Superieure des Telecommunications, 75634 Paris CEDEX 13, France. On leave from the Mathematical Institute of HAS, POB 127, 1364 Budapest, Hungary.

Clearly, the nonbinary case has a different flavor, and in particular, has no natural reformulation using the familiar language of extremal set theory. In the hope that problems for binary sequences are appealing to a larger set of combinatorialists, we are now introducing a generalization of perfect hash functions. Our new problem is very nontrivial already in the binary case and will lead to a new treatment of $M(2, 2, n)$.

Log's and exp's are binary.

2. Locally different sequences. Let B be a set of b elements represented by $\{0, 1, \cdots, b-1\}$, which we will consider as our alphabet. Let B^t denote the set of sequences of length t of elements from B. Given a set $A \subset B^t$ we say that the coordinates $i_1, i_2, \cdots, i_l$ form a *separating domain* for A if the subsequences

$$x_{i1}, x_{i2} \cdots x_{il} \quad \text{of } x = x_1 \cdots x_t, \quad x \in A$$

are all different. Let $L(A)$ denote the smallest number l for which a separation domain with l coordinates exists for A. In this language, the existence of a (b, k)-system of perfect hash functions for an n-element set with $t = Y(b, k, n)$ partitions is equivalent to the existence of an n-element set $E \subset B^t$ such that for every $A \subset E$ with $|A| = k$ we have $L(A) = 1$. In other words, $Y(b, k, n)$ is the smallest length for which n sequences of length t can be constructed from a b-ary alphabet under the condition that for any k sequences (out of these n) their separation domain has just one element.

Intuitively, we shall imagine that a set of sequences is locally different if they have a small separation domain. Systems of perfect hash functions in the sense of Fredman and Komlós [1] are an example of this concept. To be more precise, we shall give the following definition.

DEFINITION. The set $A \subset B^t$ is l-different if its smallest separation domain has at most l coordinates. Let $Z(b, k, n, l)$ be the smallest t for which there exists an n-element set $E \subset B^t$ such that every k-element subset of E is l-different.

Then $Z(b, k, n, 1) = Y(b, k, n)$. Clearly, if l is large relative to k, the problem of determining $Z(b, k, n, l)$ becomes trivial. In fact, we have Proposition 1.

PROPOSITION 1. *If* $l \geq k - 1$, *then* $Z(b, k, l, n) = \log n / \log b$.

Proof. We shall prove that every k-element subset of B^t is $(k-1)$-different, the rest being trivial. We will use induction on k. Notice that no other parameter is relevant.

The statement is obvious for $k = 2$. Therefore set $k > 2$ and suppose that the statement is true for $k' < k$. Let $A \subset B^t$ have k elements. As all its elements differ, consider an arbitrary coordinate in which at least two of them differ. Let A_c be the set of those sequences in A that have $c \in B$ in this coordinate. Then, by definition for some $c \in B$ both A_c and $A \backslash A_c$ are nonempty. The induction hypothesis implies

$$L(A_c) \leq |A_c| - 1 \quad \text{and} \quad L(A \backslash A_c) \leq |A \backslash A_c| - 1.$$

Therefore,

$$L(A) \leq 1 + L(A_c) + L(A \backslash A_c) \leq |A| - 1. \qquad \square$$

It seems to us that the information-theoretic technique used in [6] is quite inefficient in dealing with the problem of locally different sequences if $l > 1$. Unfortunately, we cannot suggest any alternative in the general case. Rather, we have attempted to work on the first nontrivial examples. In fact, from now on, we restrict ourselves to the binary alphabet, i.e., $b = 2$. We know that for $k = 2$ the problem is trivial. The case $k = 3$ is settled by Proposition 1. We note that in order for $Z(b, k, n, l)$ to be finite, we must have

$$l \geq \left\lceil \frac{\log k}{\log b} \right\rceil.$$

The first k for which l can be smaller than $k-1$ is $k=4$. Unfortunately, we do not know the full answer here. We have no reason to believe that any of our bounds is tight. Their main interest is to provide nontrivial estimates in an area characterized by a lack of methods.

3. Two-different quadruples and (2, 2)-separation. Our key result is Theorem 1.

THEOREM 1. *For every sufficiently large n*

$$3.53 \leqq \frac{Z(2,4,2,n)}{\log n} \leqq \frac{3}{\log \frac{4}{3}} \sim 7.23.$$

Proof. Let us fix some n and let us write $t = Z(2,4,2,n)$. Thus we can construct n binary sequences of length t such that for any quadruple of these sequences there are two distinct coordinates where the corresponding four binary pairs are

00
01
10
11

in an arbitrary order.

We claim that this property of our n sequences is equivalent to the following: for any two disjoint pairs of sequences there exists a coordinate in which the two members of both differ simultaneously. In fact, suppose first that our original condition is satisfied. In other words, for every quadruple of our n elements there exist two different binary equipartitions (BEP), generated by the coordinates of the respective sequences. Let us now consider any two disjoint pairs of sequences and their prescribed configurations. Because the two coordinates involved represent BEPs, the two pairs are either both different or both equal, simultaneously. But they can both be equal in at most one of the two coordinates. In the other direction, look at an arbitrary quadruple and form two disjoint pairs in it arbitrarily. There is a coordinate in which both differ simultaneously. This gives one BEP in which the zeros are one class, and the ones are the other class. Now consider these two classes as the two disjoint pairs. Their simultaneous separation gives a new BEP that differs from the previous one.

Let us now proceed to prove

$$3.53 \log n \leqq Z(2,4,2,n).$$

Consider the fixed optimal configuration of n sequences. For $x \in \{0,1\}^t$, $y \in \{0,1\}^t$ let the *Hamming distance* $d(x,y)$ be the number of coordinates in which they differ. Let d be the minimum Hamming distance between sequences in our optimal configuration and let (x^*, y^*) be a pair of sequences achieving it. Let the set of the remaining $(n-2)$ sequences be denoted by C_t. Then the minimum Hamming distance between different elements of C_t is at least d. On the other hand, any pair of sequences in C_t must differ in at least one of those coordinates in which x^* and y^* disagree, and hence C_t cannot have more elements as there are binary sequences of length d. This gives

$$n-2 \leqq 2^d.$$

This can be viewed as a relation between the cardinality of C_t and its minimum Hamming distance, for every n. Denoting

$$R_t = \frac{\log (n-2)}{t},$$

we obtain

$$R_t \leqq \frac{d}{t} \tag{1}$$

for the sequence $t = t(n)$, where $n \to \infty$. Let us denote

$$R(D) = \frac{1}{t}\log |D| \quad \text{and} \quad \delta(D) = \frac{1}{t}\min \{d(x, y); x \in D, y \in D, x \neq y\}$$

for any $D \subset \{0, 1\}^t$. We are interested in how large a set with a given minimum Hamming distance can be. We write

$$R_t(\delta) = \max \{R(D); \delta(D) \leqq \delta, D \subset (0, 1)^t\},$$

and $R(\delta) = \lim \sup_{t \to \infty} R_t(\delta)$. Our bound on $Z(2, 4, 2, n)$ will follow from (1) and an analysis of the function $R(\delta)$. In fact, let us suppose that we have a sequence of constructions satisfying (1) with

$$\limsup_{n \to \infty} R_{t(n)} = R_0.$$

Clearly, for this R_0 we must have

$$R_0 \leqq \max \{\delta; R(\delta) \leqq \delta\}.$$

Obviously, $R(\delta)$ is a monotonically decreasing function of δ, and hence it has a unique point δ^* for which

$$R(\delta^*) = \delta^*.$$

Thus, by the previous inequality,

$$R_0 \leqq R(\delta^*).$$

Although the value of $R(\delta^*)$ is unknown, coding theory provides us with interesting upper bounds on the function $R(\delta)$ that we can use to evaluate our last inequality. Using the celebrated linear programming bound on $R(\delta)$ by McEliece et al. [8], we get

$$R_0 \leqq \max \{\delta, R''(\delta) \leqq \delta\},$$

where $R''(\delta)$ is the rate-distance bound in [8]. As in the previous argument,

$$R_0 \leqq R''(\delta^{**})$$

where δ^{**} is the unique point in which $R''(\delta^{**}) = \delta^{**}$. Looking up the values of $R''(\delta)$ we complete the proof.

The upper bound can be obtained by randomly selecting the binary sequences, independently and equiprobably among all possible sequences of the stated length. We omit the calculations. □

As an immediate consequence of the lower bound in the above theorem, we obtain a lower bound on $M(2, 2, n)$ that is stronger than the one in [1]. We have Theorem 2.

THEOREM 2. *For every sufficiently large* n,

$$3.53 \leqq \frac{M(2, 2, n)}{\log n} \leqq \frac{3}{\log \frac{8}{7}} \sim 15.57.$$

Proof. The upper bound can be obtained by random selection, as in the previous proof. Similarly nonconstructive bounds for this problem can be found in [5]. For this

case those authors obtain $4/\log\frac{8}{7}$. A more careful evaluation gives the above. Once again, we omit the details.

The lower bound will follow upon noticing that

$$Z(2, 4, 2, n) \leqq M(2, 2, n). \tag{2}$$

In order to prove (2), it is enough to observe that a (2, 2)-separating partition system can be described equivalently as a set of binary sequences such that for every quadruple of these binary sequences there exist three coordinates in which they represent three different BEPs. By the first part of the proof of Theorem 1 we also know that in any set of binary sequences satisfying the conditions of that theorem, for every quadruple of sequences there are two coordinates where we have two different BEPs. Thus, our present conditions are stronger. This proves (2); whence the original statement follows by Theorem 1. □

It seems unfortunate to reduce this problem to a weaker one and we would like to see a better argument.

Related problems are discussed in [9].

Acknowledgment. It is a pleasure to acknowledge our fruitful discussions with Vera T. Sós.

REFERENCES

[1] M. FREDMAN AND J. KOMLÓS, *On the size of separating systems and perfect hash functions*, SIAM J. Algebraic Discrete Methods, 5 (1984), pp. 61–68.

[2] A. G. DJACKOV AND V. V. RYKOV, *Bounds on the length of disjunctive codes*, Problems Inform. Transmission, 18 (1982), pp. 7–13.

[3] P. ERDÖS, P. FRANKL, AND Z. FÜREDI, *Families of finite sets in which no set is covered by the union of two others*, J. Combin. Theory Ser. A, 33 (1982), pp. 158–166.

[4] F. K. HWANG AND V. T. SÓS, *Non-adaptive hypergeometric group testing*, Combinatorica, to appear.

[5] A. D. FRIEDMAN, R. L. GRAHAM, AND J. D. ULLMAN, *Universal single transition time asynchronous state assignments*, IEEE Trans. Comput., 18 (1969), pp. 541–547.

[6] J. KÖRNER AND K. MARTON, *New bounds for perfect hashing via information theory*, European J. Combin., to appear.

[7] J. KÖRNER, *Fredman–Komlós bounds and information theory*, SIAM J. Algebraic Discrete Methods, 7 (1986), pp. 560–570.

[8] R. J. MCELIECE, E. R. RODEMICH, H. RUMSEY JR., AND L. R. WELCH, *New upper bounds on the rate of a code via the Delsarte–MacWilliams inequalities*, IEEE Trans. Inform. Theory, 23 (1977), pp. 157–166.

[9] P. FRANKL AND Z. FÜREDI, *Union-free hypergraphs and probability theory*, European J. Combin., 5 (1984), pp. 127–131.