# Hash Families and Covering Arrays

(Extended Abstract)

Charles J. Colbourn
Arizona State University

## 1 Introduction

Let $N$, $k$, $t$, and $v$ be positive integers. Let $C$ be an $N \times k$ array with entries from an alphabet $\Sigma$ of size $v$; we typically take $\Sigma = \{0, \ldots, v-1\}$. When $(\nu_1, \ldots, \nu_t)$ is a $t$-tuple with $\nu_i \in \Sigma$ for $1 \leq i \leq t$, $(c_1, \ldots, c_t)$ is a tuple of $t$ column indices ($c_i \in \{1, \ldots, k\}$), and $c_i \neq c_j$ whenever $\nu_i \neq \nu_j$, the $t$-tuple $\{(c_i, \nu_i) : 1 \leq i \leq t\}$ is a *$t$-way interaction*. The array *covers* the $t$-way interaction $\{(c_i, \nu_i) : 1 \leq i \leq t\}$ if, in at least one row $\rho$ of $\mathsf{C}$, the entry in row $\rho$ and column $c_i$ is $\nu_i$ for $1 \leq i \leq t$. Array $\mathsf{C}$ is a *covering array* $\mathsf{CA}(N; t, k, v)$ of *strength* $t$ when every $t$-way interaction is covered. Figure 1 gives an example of a covering array with $N = 12$ rows, ten factors having two symbols, and strength three. Consider, for example, the 3-way interaction $\{(2, 0), (5, 1), (6, 1)\}$; it is covered in the fifth and eighth rows.

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| 1 | 1 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 1 |
| 1 | 0 | 1 | 1 | 0 | 1 | 0 | 1 | 0 | 0 |
| 1 | 0 | 0 | 0 | 1 | 1 | 1 | 0 | 0 | 0 |
| 0 | 1 | 1 | 0 | 0 | 1 | 0 | 0 | 1 | 0 |
| 0 | 0 | 1 | 0 | 1 | 0 | 1 | 1 | 1 | 0 |
| 1 | 1 | 0 | 1 | 0 | 0 | 1 | 0 | 1 | 0 |
| 0 | 0 | 0 | 1 | 1 | 1 | 0 | 0 | 1 | 1 |
| 0 | 0 | 1 | 1 | 0 | 0 | 1 | 0 | 0 | 1 |
| 0 | 1 | 0 | 1 | 1 | 0 | 0 | 1 | 0 | 0 |
| 1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 1 |
| 0 | 1 | 0 | 0 | 0 | 1 | 1 | 1 | 0 | 1 |

Figure 1: CA(12;3,10,2)

We denote by $\mathsf{CAN}(t, k, v)$ the minimum $N$ for which a $\mathsf{CA}(N; t, k, v)$ exists; fewer rows is what we are after. Because $\mathsf{CAN}(1, k, v) = v$, $\mathsf{CAN}(t, k, v) = v^t$ when $k < t$, and $\mathsf{CAN}(t, k, 1) = 1$, we generally assume that $k \geq t \geq 2$ and $v \geq 2$.

Applications to interaction testing, in particular to testing component-based software, have driven much recent research. In applications in testing, columns of the array correspond to experimental *factors*, and the symbols in the column form *values* or *levels* for the factor. Each row specifies the values to which to set the factors for an experimental *run*. The array is 'covering' in the sense that every $t$-way interaction appears in at least one run. Covering arrays are employed in numerous testing applications in which experimental factors interact to detect the presence of faults, to detect the location of faults, to detect interactions in biological networks, to generate representative multiple sequence alignments of genomic data, and to learn an unknown function by nonadaptive tests. They have also arisen in numerous other disguises, as $t$-universal sets, existentially closed graphs, $t$-qualitatively independent sets of partitions, $t$-surjective codes, face transversals of the $n$-cube, and others. For these reasons, the construction of covering arrays has been a topic of much research.

Our primary concern is with recursive constructions that make larger covering arrays from smaller ones by a technique of 'column replacement'. A *perfect hash family* $\mathsf{PHF}(N; k, w, t)$ is an $N \times k$ array on $w$ symbols, in which in every $N \times t$ subarray, at least one row consists of distinct symbols. The smallest $N$ for which a $\mathsf{PHF}(N; k, v, t)$ exists is the *perfect hash family number*, denoted $\mathsf{PHFN}(k, v, t)$. Figure 2 shows a $\mathsf{PHF}(6; 12, 3, 3)$. For instance, in columns 1, 3, and 5, the first row contains `0 2 1`.

$$\begin{bmatrix} 0 & 1 & 2 & 2 & 1 & 2 & 2 & 0 & 1 & 1 & 0 & 0 \\ 0 & 2 & 1 & 0 & 2 & 2 & 2 & 1 & 0 & 1 & 2 & 1 \\ 1 & 0 & 0 & 2 & 2 & 2 & 1 & 1 & 2 & 1 & 0 & 2 \\ 2 & 0 & 1 & 1 & 2 & 0 & 2 & 0 & 1 & 1 & 2 & 1 \\ 2 & 0 & 2 & 1 & 2 & 1 & 0 & 2 & 2 & 1 & 1 & 0 \\ 2 & 0 & 1 & 2 & 1 & 1 & 2 & 2 & 0 & 1 & 2 & 1 \end{bmatrix}$$

Figure 2: A $\mathsf{PHF}(6; 12, 3, 3)$

Perfect hash families were introduced an efficient tool for compact storage and fast retrieval of frequently used information. In this setting, each row defines a *hash function* from a domain of size $k$ to a range of size $v$; we employ the array formulation instead. Perfect hash families can be used to construct separating systems, key distribution patterns, group testing algorithms, cover-free families, secure frameproof codes, and broadcast encryption schemes. Of particular concern here is that perfect hash families arise as ingredients in some recursive constructions for covering arrays.

# 2 Column Replacement

Our reason for interest in PHFs is that they are used to construct covering arrays:

**Theorem 2.1** *If a* $\mathsf{PHF}(s; k, m, t)$ *and a* $\mathsf{CA}(N; t, m, v)$ *both exist then a* $\mathsf{CA}(sN; t, k, v)$ *exists.*

*Proof.* Let $\mathsf{B} = (b_{ij})$ be an $s \times k$ array on $m$ symbols forming a $\mathsf{PHF}(s; k, m, t)$. Let $\mathsf{A} = (a_{ij})$ be an $N \times m$ array on $v$ symbols forming a $\mathsf{CA}(N; t, m, v)$. We produce an $sN \times k$ array $\mathsf{C} = (c_{ij})$ as follows. For each $1 \le i \le s$, $1 \le j \le N$, and $1 \le \ell \le k$, set $c_{(i-1)N+j,\ell} = a_{j,b_{i,\ell}}$. The verification that $\mathsf{C}$ is a $\mathsf{CA}(sN; t, k, v)$ is straightforward; one needs only check that every $t$-way interaction is covered. Consider the $t$-way interaction $\{(\gamma_1, \nu_1), \ldots, (\gamma_t, \nu_t)\}$. Because $\mathsf{B}$ is a perfect hash family of strength $t$, it is also a perfect hash family of strength $t'$ for all $t' \le t$. Therefore there is a row $\rho$ of $\mathsf{B}$ in which $b_{\rho,\gamma_i} \neq b_{\rho,\gamma_j}$ whenever $\nu_i \neq \nu_j$ (and hence $\gamma_i \neq \gamma_j$). Set $d_i = b_{\gamma_i}$ for $1 \le i \le t$. In columns $(d_1, \ldots, d_t)$ of $\mathsf{A}$, there is a row $\tau$ in which $a_{\tau,d_i} = \nu_i$, because $\mathsf{A}$ is a covering array of strength $t$ and $d_i \neq d_j$ when $\nu_i \neq \nu_j$. But then $c_{(\rho-1)N+\tau,\gamma_i} = \nu_i$ for $1 \le i \le t$, and the $t$-way interaction is covered in $\mathsf{C}$. ■

Less formally, the perfect hash family $\mathsf{B}$ is used as a 'pattern' to select columns from the covering array $\mathsf{A}$, so that every symbol $\sigma$ of $\mathsf{B}$ is replaced by the entire column of $\mathsf{A}$ that is indexed by $\sigma$.

Many improvements on Theorem 2.1 have been developed recently. We introduce the most general one next.

The standard definition of covering array asks for all $t$-way interactions to be covered. We consider restrictions in which all sets of $t$ columns are treated similarly, but not all $t$-way interactions need to be covered.

The *species* of a $t$-way interaction $\mathcal{S} = \{(c_i, \nu_i) : 1 \le i \le t\}$ is the multiset $\{\nu_i : 1 \le i \le t\}$; hence a species in general encompasses a number of specific $t$-way interactions. (A species can be represented as a weak composition of $t$ with $v$ parts, and there are $\binom{t+v-1}{v-1}$ species.) Often we are not concerned with the specific symbols used in defining the species. Then the *family* of a species is its orbit under the action of the symmetric group on $v$ letters, and hence a family consists of a set of species, and by inheritance, a set of $t$-way interactions. (A family can be represented as a partition of $t$ into at most $v$ parts.)

Let $\mathbb{S}$ be a set of species for $t$ and $v$. An $N \times k$ array with $v$ symbols is an $\mathbb{S}$-*quilting array* if every interaction whose species is in $\mathbb{S}$ is covered. The notation $\mathbb{S}\text{-}\mathsf{QA}(N; t, k, v)$ is used for such an array when $\mathbb{S}$ contains interactions of strength at most $t$, and $\mathbb{S}\text{-}\mathsf{QAN}(t, k, v)$ is the smallest $N$ for which an $\mathbb{S}\text{-}\mathsf{QA}(N; t, k, v)$ exists. An $\mathbb{S}\text{-}\mathsf{QA}(N; t, k, v)$ is equivalent to a $\mathsf{CA}(N; t, k, v)$ when $\mathbb{S}$ contains all possible species of $t$-way interactions.

We also employ variants of perfect hash families. An $N \times t$ array $A$ on $w$ symbols (with columns $C = \{1, \ldots, t\}$) is $(t, v)$-*distributing* if, for every partition $\{C_1, \ldots, C_v\}$ of $C$ into $v$ parts, there is at least one row of $A$, $(a_1, \ldots, a_t)$, in which $a_i = a_j$ only if $i$ and $j$ belong to the same class of the partition. An $N \times k$ array is $(t, v)$-*distributing* if every $N \times t$ subarray is $(t, v)$-distributing; such an array is called a *distributing hash family*, and is denoted by $\mathsf{DHF}(N; k, w, t, v)$. An $(N; k, v, \{w_1, w_2, \ldots, w_t\})$-*separating hash family*, or $\mathsf{SHF}(N; k, v, \{w_1, w_2, \ldots, w_t\})$, is an $(N; k, v)$-hash family $\mathcal{H}$ that satisfies the property: For any $C_1, C_2, \ldots, C_t \subseteq \{1, 2, \ldots, k\}$ such that $|C_1| = w_1$, $|C_2| = w_2$, …, $|C_t| = w_t$, and $C_i \cap C_j = \emptyset$ for every $i \neq j$, there exists at least one function $h \in \mathcal{H}$ such that $\{f(y) : y \in C_i\} \cap \{f(y) : y \in C_j\} = \emptyset$. A *heterogeneous hash family*, denoted $\mathsf{HHF}(N; k, (v_1, \ldots, v_N))$, is an $N \times k$ array in which the $i$th row contains (at most) $v_i$ symbols for $1 \le i \le N$. Often we write $(v_1, \ldots, v_N)$ in exponential notation: $v_1^{u_1} \cdots v_c^{u_c}$ means that the $N = \sum_{i=1}^{c} u_i$ rows can be partitioned into classes, so that in the $i$th class there are $u_i$

rows each employing (at most) $v_i$ symbols. These notions combine in the obvious manner to form heterogeneous DHFs and SHFs; we denote these using DHHF and SHHF.

Let $\mathbb{S}$ be the set of all multisets $\{\nu_1, \ldots, \nu_t\}$ with $\nu_i \in \{1, \ldots, v\}$ for $1 \leq i \leq t$. Let $A$ be an $M \times k$ array with $v$ symbols. Define a function $\Phi$ with $\Phi : \mathbb{S} \mapsto 2^{\{1,\ldots,M\}}$. Then $A$ is a $\Phi$-*separating hash family* if for every $S = \{\nu_1, \ldots, \nu_t\} \in \mathbb{S}$ and for every choice of $t$ distinct columns $(c_1, \ldots, c_t)$, there is at least one row $\rho \in \Phi(S)$ in which, for $1 \leq i < j \leq t$, row $\rho$ has different symbols in columns $c_i$ and $c_j$ if $\nu_i \neq \nu_j$. Such an array is denoted by $\Phi$-SHF$(M; k, v, t)$. Again we generalize to the heterogeneous case: A $\Phi$–*separating heterogeneous hash family* $\Phi$-SHHF$(M; k, v_1 \cdots v_M, t)$ contains at most $v_i$ symbols in the $i$th row for $1 \leq i \leq M$, and satisfies the same separation condition.

When $\Phi : \mathbb{S} \mapsto 2^{\{1,\ldots,M\}}$ is specified, we define a vector $(\Psi_1, \ldots, \Psi_M)$ so that $\Psi_i = \{S : i \in \Phi(S)\}$. In words, $\Phi$ associates each $t$-way interaction with a set of rows of the array, while $\Psi_i$ contains the $t$-way interactions thereby associated with the $i$th row.

**Theorem 2.2** *Let $t$ be a positive integer. Suppose that a $\Phi$-SHHF$(M; k, k_1 \cdots k_M, t)$ exists, and that a $\Psi_i$-QA$(R_i; t, k_i, v)$ exists for each $1 \leq i \leq M$. Then a CA$(\sum_{i=1}^{M} R_i; t, k, v)$ exists.*

*Proof.* Let $D$ be a $\Phi$-SHHF$(M; k, k_1 \cdots k_M, t)$. Form $E$ by replacing each entry $j$ in row $i$ of $D$ by the $j$th column of the $\Psi_i$-QA$(R_i; t, k_i, v)$. It suffices to prove that $E$ is a covering array of strength $t$. Fix a tuple $C = (c_1, \ldots, c_t)$ of $t$ columns in $E$ (equivalently, in $D$), and fix a $t$-way interaction $T$ by selecting value $\nu_j$ for column $c_j$ for $1 \leq j \leq t$. We must show that $T$ is covered in $E$. Let $W = \Phi(T)$, the set of rows that (together) separate $T$ in $D$; then for some $w \in W$, $T$ is separated in row $w$. Because $T \in \Psi_w$, it is covered in the $\Psi_w$-QA$(R_w; t, k_w, v)$ and therefore also covered in $E$. ■

Naturally the concern with such general theorems is finding the ingredients with which to apply them! In the presentation, we describe methods that are appropriate for making such ingredients, and discuss consequences for the existence of covering arrays.

(This research is reported in papers by the author with Alan Ling, Pepe Torres-Jiménez, and Junling Zhou. Copies are available from the author.)