A Bound on the Size of Separating Hash Families[*]

Simon R. Blackburn
Department of Mathematics
Royal Holloway, University of London
Egham, Surrey TW20 0EX, United Kingdom
s.blackburn@rhul.ac.uk

Tuvi Etzion
Department of Computer Science
Technion — Israel Institute of Technology
Technion City, Haifa 32000, Israel
etzion@cs.technion.ac.il

Douglas R. Stinson
David R. Cheriton School of Computer Science
University of Waterloo
Waterloo ON, N2L 3G1, Canada
dstinson@uwaterloo.ca

and

Gregory M. Zaverucha
David R. Cheriton School of Computer Science
University of Waterloo
Waterloo ON, N2L 3G1, Canada
gzaveruc@uwaterloo.ca

July 31, 2007

[*] This research was supported in part by E.P.S.R.C. grant EP/E034632 and by NSERC grant 203114-06.

Abstract

The paper provides an upper bound on the size of a (generalised) separating hash family, a notion introduced by Stinson, Wei and Chen. The upper bound generalises and unifies several previously known bounds which apply in special cases, namely bounds on perfect hash families, frameproof codes, secure frameproof codes and separating hash families of small type.

Keywords: Separating hash family, perfect hash family, frameproof code, secure frameproof code.

1 Introduction

This paper is concerned with generalised separating hash families, which are defined as follows.

Definition 1 Let X and Y be sets of cardinalities n and m respectively. We call a set $\mathcal{F}$ of N functions $f : X \to Y$ *an* $(N; n, m)$*-hash family.*

Definition 2 Let $f : X \to Y$ be a function, and let $C_1, C_2, \ldots, C_t \subseteq X$. We say that f *separates* $C_1, C_2, \ldots, C_t$ if $f(C_1), f(C_2), \ldots, f(C_t)$ are pairwise disjoint.

Note that for a function f to separate subsets $C_1, C_2, \ldots, C_t$, the subsets C_i must be pairwise disjoint.

Definition 3 Let X and Y be sets of cardinalities n and m respectively, and let $\mathcal{F}$ be an $(N; n, m)$-hash family of functions from X to Y. We say that $\mathcal{F}$ is *an* $(N; n, m, \{w_1, w_2, \ldots, w_t\})$ *separating hash family* (which we shall also write as an $\mathrm{SHF}(N; n, m, \{w_1, w_2, \ldots, w_t\})$) if it satisfies the following property. For all pairwise disjoint subsets $C_1, C_2, \ldots, C_t \subseteq X$ with $|C_i| = w_i$ for $i \in \{1, 2, \ldots, t\}$, there exists at least one function $f \in \mathcal{F}$ that separates $C_1, C_2, \ldots, C_t$. The multiset $\{w_1, w_2, \ldots, w_t\}$ is the *type* of the separating hash family.

To avoid trivialities, we always assume that $t \geq 2$ and that $u \leq n$, where we define $u = \sum_{i=1}^{t} w_i$. Clearly we must have that $t \leq m$ if an $\mathrm{SHF}(N; n, m, \{w_1, w_2, \ldots, w_t\})$ exists.

This notion was introduced in the special case when $t = 2$ by Stinson, van Trung and Wei [9] and in full generality by Stinson, Wei and Chen [10]. One of the attractions of the concept of separating hash families is its simultaneous generalisation of several well-studied classes of combinatorial objects.

For example, an $(N; n, m, u)$ *perfect hash family* is a separating hash family in the special case when $u = t$ and $w_i = 1$ for all i. An (k, u)-*hashing family* (introduced by Barg, Cohen, Encheva, Kabatiansky and Zémor [3], see also Alon, Cohen, Krivelevich and Litsyn [1]) is a separating hash family in the case when $t = k + 1$, $w_1 = w_2 = \cdots = w_k = 1$ and $w_{k+1} = u - k$. Given an $(N; n, m)$-hash family $\mathcal{F} = \{f_1, f_2, \ldots, f_N : X \to Y\}$, we may define an m-ary code $C \subseteq Y^N$ of length N by

$$C = \{(f_1(x), f_2(x), \ldots, f_N(x)) : x \in X\}.$$

When $\mathcal{F}$ is a separating hash family, it is easy to see that the code C has exactly n codewords. So we are able to rephrase the separating hash family property in the language of coding theory. This rephrasing provides a link with various codes that have been studied for applications in copyright protection and cryptography. Indeed, *w-frameproof codes* are separating hash families of type $\{1, w\}$ and *w-secure frameproof codes* are separating hash families of type $\{w, w\}$. Codes with the identifiable parent property (2-*IPP codes*) are separating hash families which are simultaneously of type $\{1, 1, 1\}$ and $\{2, 2\}$. See Stinson *et al.* [10] for references to some of the extensive literature on these objects.

The paper aims to prove the following bound on the size of a separating hash family.

Theorem 1 *Suppose an* $\mathrm{SHF}(N; n, m, \{w_1, w_2, \ldots, w_t\})$ *exists. Define* $u = \sum_{i=1}^{t} w_i$. *Then*

$$n \leq \gamma m^{\lceil N/(u-1) \rceil},$$

where γ is a constant which depends only on $w_1, w_2, \ldots, w_t$.

We provide three proofs of this theorem. The first proof (in Section 2) shows that we may take $\gamma = \binom{u}{2}$ in the theorem; this proof uses general results and so the argument is quite short. Our second proof (in Section 3) extends techniques of Stinson *et al.* [10] to reduce the value we may take for γ. We show that we may take $\gamma = 2(u - w_1)w_1 - w_1$, where we assume (without loss of generality) that w_1 is the smallest of the integers w_i. Our final proof (in Section 4) obtains a significantly better value for γ: we may take $\gamma = w_1 w_2 + u - w_1 - w_2$, where we assume (without loss of generality) that w_1 and w_2 are the smallest two of the integers w_i. We give an argument to indicate that this value for γ cannot be improved without introducing some significant new ideas.

Theorem 1 generalises bounds due Stinson, Wei and Chen [10] (which apply only when $u \leq 4$), Stinson and Zaverucha [11] (which apply only for separating hash families of types $\{w-1, w\}$ and $\{w, w\}$) and Blackburn [5] (which apply for separating hash families of types $\{1, w\}$).

We claim that the exponent $\lceil N/(u-1)\rceil$ in the bound of Theorem 1 is realistic, since the exponent in a bound of this form cannot be improved to a value less than $N/(u-1)$. To prove this, first note that a probabilistic construction due to Blackburn [4] shows that an $(N; n, m, u)$ perfect hash family exists provided that

$$N > \frac{\log 4\left(\binom{n}{u} - \binom{n-u}{u}\right)}{\log m^u - \log\left(m^u - u!\binom{m}{u}\right)}.$$

In particular, this implies that for any fixed u, and any real number δ such that $\delta < N/(u-1)$, there exists an $(N; \lfloor m^\delta \rfloor, m, u)$ perfect hash family whenever m is sufficiently large. Since an $(N; n, m, u)$ perfect hash family is an $\mathrm{SHF}(N; , n, m, \{w_1, w_2, \ldots, w_t\})$ for any w_i such that $\sum_{i=1}^{t} w_i = u$, we have established our claim.

2 A proof using labelled graphs

Let $\mathcal{F} = \{f_1, f_2, \ldots, f_N : X \to Y\}$ be an $(N; n, m)$-hash family. We define a labelled graph $\Gamma(\mathcal{F})$ as follows. We define the vertex set of $\Gamma(\mathcal{F})$ to be the domain X of the functions in $\mathcal{F}$. We join distinct vertices x and x' by an edge labelled i if and only if $f_i(x) = f_i(x')$. So $\Gamma(\mathcal{F})$ contains no loops, but there might be as many as N edges between a pair of vertices.

We now state a lemma from Blackburn [6] which will be central to the proof of Theorem 1. The lemma was originally stated using coding theory terminology, but we rephrase the lemma in terms of hash families.

Lemma 2 (Blackburn [6, Lemma 2]) *Let u be a positive integer. Let $\mathcal{F}$ be an $(N; n, m)$-hash family, and suppose that $n \geq \binom{u}{2}(m-1) + 2$. Let T be a tree on u vertices, whose edges are labelled with elements of the set $\{1, 2, \ldots, N\}$. Then the graph $\Gamma(\mathcal{F})$ defined above contains a subgraph isomorphic to T (as a labelled graph).*

We now prove the following lemma, which is an important special case of Theorem 1.

Lemma 3 *Let $\mathcal{F}$ be an $\mathrm{SHF}(N; n, m, \{w_1, w_2, \ldots, w_t\})$. Let $u = \sum_{i=1}^{t} w_i$. Suppose that $N < u$. Then*

$$n \leq \binom{u}{2}(m-1) + 1.$$

Proof: Suppose, for a contradiction, that an $\mathrm{SHF}(N; n, m, \{w_1, w_2, \ldots, w_t\})$ exists with $N < u$ and $n \geq \binom{u}{2}(m-1)+2$. Let $\mathcal{F} = \{f_1, f_2, \ldots, f_N\}$ be such a hash family, where $f_i : X \to Y$ for some sets X and Y. We aim to apply Lemma 2, and to this end we define a tree T as follows.

Let the vertices of T be a set of size u. Partition the vertices of T into t parts $T_1, T_2, \ldots, T_t$, where $|T_i| = w_i$. We now choose the edges of T in such a way that no edge lies within a part of the partition. This can always be done: one way of doing this is as follows. Choose 'special' elements $s_1, s_2, \ldots, s_t \in T$ such that $s_i \in T_i$. For $i \in \{1, 2, \ldots, t-1\}$, join s_i to all the elements in T_{i+1}. Finally, join s_t to all the elements in $T_1 \setminus \{s_1\}$. An example of this process is given in Figure 1. Finally, we label the edges of T with elements of $\{1, 2, \ldots, N\}$ in such a way that every possible label occurs at least once. We may do this since T has $u-1$ edges, and $N < u$.

Recall the graph $\Gamma(\mathcal{F})$ defined above. By Lemma 2, the graph $\Gamma(\mathcal{F})$ contains an isomorphic copy L of T. Let $\phi : T \to L$ be an isomorphism, and define pairwise disjoint subsets $C_1, C_2, \ldots, C_t \subset X$ of the vertices of $\Gamma(\mathcal{F})$ by $C_i = \phi(T_i)$. Note that $|C_i| = |T_i| = w_i$, and the subsets C_i are pairwise disjoint.

We claim that the subsets C_i provide a counterexample to the SHF property, which gives us the contradiction we need. To establish this claim, let $j \in \{1, 2, \ldots, N\}$. It suffices to show that the sets $f_j(C_i)$ are not disjoint, and so f_j does not separate the sets $C_1, C_2, \ldots, C_t$. By our construction of the tree T, there exists an edge $ab \in T$ labelled with the symbol j. Moreover, this edge does not lie within a part of our partition, so $a \in T_i$ and $b \in T_{i'}$ where $i \neq i'$. Since ϕ is an isomorphism, there is an edge in $\Gamma(\mathcal{F})$ between $\phi(a) \in C_i$ and $\phi(b) \in C_{i'}$ labelled with j, and so (by definition of the graph $\Gamma(\mathcal{F})$) we have that $f_j(\phi(a)) = f_j(\phi(b))$. Since $f_j(\phi(a)) \in f_j(C_i)$ and $f_j(\phi(b)) \in f_j(C_{i'})$, we see that $f_j(C_i)$ and $f_j(C_{i'})$ are not disjoint. This gives us the contradiction we were seeking. □

Proof of Theorem 1: Suppose that $\mathcal{F} = \{f_1, f_2, \ldots, f_N : X \to Y\}$ is an $\mathrm{SHF}(N; n, m, \{w_1, w_2, \ldots, w_t\})$. Let k and ℓ be positive integers such that $k\ell \geq N$. We construct a hash family $\mathcal{F}' = \{f'_1, f'_2, \ldots, f'_\ell : X \to Y^k\}$ as follows. Cover the set $\{1, 2, \ldots, N\}$ by ℓ sets $I_1, I_2 \ldots, I_\ell$, each of size k. For $j \in \{1, 2, \ldots, \ell\}$, define a function $f'_j : X \to Y^k$ by $f'_j(x) = (f_i(x) : i \in I_j)$. It is easy to check that $\mathcal{F}'$ is an $\mathrm{SHF}(\ell; n, m^k, \{w_1, w_2, \ldots, w_t\})$, the key observation being that whenever sets $f_i(C_j)$ are disjoint and $x \in \{1, 2, \ldots, \ell\}$ is defined by $i \in I_x$, we have that the sets $f'_x(C_j)$ are also disjoint.

Now let $u = \sum_{i=1}^{t} w_i$. Construct a hash family $\mathcal{F}'$ from $\mathcal{F}$ as above, in the case when $\ell = u-1$ and $k = \lceil N/(u-1) \rceil$. Since $\ell < u$, we may apply

Lemma 3 to $\mathcal{F}'$ to obtain the inequality

$$n \leq \binom{u}{2}\left(m^k - 1\right) + 1 \leq \binom{u}{2} m^k = \binom{u}{2} m^{\lceil N/(u-1) \rceil},$$

as required. □

3 A proof using forbidden configurations

In this section we sketch a proof of Theorem 1 using the technique of 'forbidden configurations'. We modify the proof in Section 2 by replacing Lemma 3 by Lemma 4 below. It is clear that the resulting proof gives rise to a better bound than the proof in Section 2: the constant γ may be taken to be much smaller for many parameters. We only provide a sketch proof at a certain point in our argument, since Section 4 shows how to reduce the value of γ still further; full details of the proof we sketch are given in the technical report [12].

Lemma 4 *Let $\mathcal{F}$ be an* SHF$(N; n, m, \{w_1, w_2, \ldots, w_t\})$*. Let $u = \sum_{i=1}^{t} w_i$. Suppose that $N < u$. Then*

$$n \leq \big(2(u - w_1)w_1 - w_1\big)(m - 1) + 1.$$

The matrix representation of a hash family will prove to be the most useful representation for this section. Let $\{f_1, f_2, \ldots, f_n\}$ be an $(N; n, m)$-hash family, where $f_i : X \to Y$. Write $X = \{x_1, x_2, \ldots, x_n\}$ (so we have ordered the elements of X in some way). Then we may define an $N \times n$ matrix A, the *matrix representation* of the hash family, by $A_{i,j} = f_i(x_j)$. Thus the rows of A correspond to functions in the hash family, and the columns of A correspond to elements of X. In this setting, we say that sets C_1, C_2 of columns *cannot be separated* if for all rows $(y_1, y_2, \ldots, y_n)$ of the matrix A, there exist $j_1 \in C_1$ and $j_2 \in C_2$ with $y_{j_1} = y_{j_2}$. In other words, the sets C_1, C_2 of columns cannot be separated exactly when the corresponding subsets of X are not separated by any of the functions f_i.

To obtain an upper bound on n (in terms of N and m and the type of the separating hash family), the general strategy involves showing A contains a submatrix which is impossible in a separating hash family whenever n is sufficiently large. Such a submatrix is referred to as a *forbidden configuration.* For example, the matrix representation of an SHF$(7; n, m, \{5, 3\})$ does not contain a submatrix isomorphic to any matrix of the form

$$\begin{array}{|c|c|c|c|c|c|c|c|}
\hline
* & * & * & * & * & * & g & g \\
\hline
* & * & * & * & * & f & f & * \\
\hline
* & * & * & * & e & e & * & * \\
\hline
* & * & * & d & d & * & * & * \\
\hline\hline
c & * & * & c & * & * & * & * \\
\hline
b & * & b & * & * & * & * & * \\
\hline
a & a & * & * & * & * & * & * \\
\hline
\end{array}\,.$$

(Here starred entries are arbitrary and $a, b, c, d, e, f, g \in Y$. We say that matrices are isomorphic if they are equal up to a permutation of their rows and columns.) To see why this is the case, note that the sets of columns $\{2, 3, 4, 6, 8\}$ and $\{1, 5, 7\}$ cannot be separated, and so the corresponding elements of X form a counterexample to the separating hash family property.

Generalising this example, we have the following lemma.

Lemma 5 *The matrix representation of an* SHF$(w + d - 1; n, m, \{w, d\})$, $w \geq d$, *does not contain a submatrix isomorphic to*

$$\begin{array}{|c|c|c|c|c|c|c|c|c|c|}
\multicolumn{1}{c}{1} & \multicolumn{1}{c}{2} & \multicolumn{1}{c}{3} & \multicolumn{1}{c}{4} & \multicolumn{1}{c}{\ldots} & \multicolumn{1}{c}{N_1} & \multicolumn{1}{c}{\ldots} & \multicolumn{1}{c}{} & \multicolumn{1}{c}{} & \multicolumn{1}{c}{} \\
\hline
* & * & * & * & * & * & * & * & g & g \\
\hline
* & * & * & * & * & * & * & \iddots & \iddots & * \\
\hline
* & * & * & * & * & * & c & c & * & * \\
\hline
* & * & * & * & * & b & b & * & * & * \\
\hline\hline
a & * & * & * & * & a & * & * & * & * \\
\hline
\vdots & & & & \iddots & & * & * & * & * \\
\hline
x & * & * & x & * & * & * & * & * & * \\
\hline
y & * & y & * & * & * & * & * & * & * \\
\hline
z & z & * & * & * & * & * & * & * & * \\
\hline
\end{array}$$

where there are $N_0 = 2d - 2$ *upper rows and* $N_1 = w - d + 1$ *lower rows.*

Proof: Let $N = N_0 + N_1 = w + d - 1$. Note that the partition C_1, C_2 of the columns cannot be separated, where

$$\begin{aligned}
C_1 &= \{1\} \cup \{N_1 + 2k : 0 \leq k \text{ and } N_1 + 2k \leq N\}\,, \\
C_2 &= \{2, \ldots, N_1\} \cup \{N_1 + 2k + 1 : 1 \leq k \text{ and } N_1 + 2k + 1 \leq N\}\,.
\end{aligned}$$

Thus an $(N; n, m)$-hash family which contains a submatrix isomorphic to the one above is not a separating hash family of type $\{w, d\}$.□

We note that the configuration of Lemma 5 generalises the ‘staircase’ configuration used by Stinson and Zaverucha [11] to provide bounds on separating hash families of types $\{w, w\}$ and $\{w-1, w\}$.

The following lemma gives an upper bound on separating hash families of type $\{w, d\}$.

Lemma 6 *If an* $\mathrm{SHF}(w+d-1; n, m, \{w, d\})$ *exists, then*

$$n \leq 1 + (2dw - w)(m-1).$$

Sketch proof: Assume that an $\mathrm{SHF}(w+d-1; n, m, \{w, d\})$ exists where $n = (2dw - w)(m-1) + 2$, and let A be its matrix representation. To prove the lemma, it suffices to derive a contradiction from this assumption. Let N, N_0 and N_1 be defined as in Lemma 5. If we define $K = N_1 + N_0N - N_0(N_0-1)/2$, a short calculation shows that A has $2 + K(m-1)$ columns.

By deleting columns, we create a series of submatrices of A, each of which satisfies one of two properties, as indicated.

$$\underbrace{A_{N-1} \subset \ldots \subset A_{N_0+1}}_{\text{Property (ii)}} \subset \underbrace{A_{N_0} \subset A_{N_0-1} \subset \ldots \subset A_1}_{\text{Property (i)}} \subset A_0 = A$$

Property (i): the elements in row i of A_i, $1 \leq i \leq N_0$, repeat at least $N - (i-2)$ times.

Property (ii): the elements appearing in row $N_0 + i$ of A_{N_0+i}, $1 \leq i \leq N_1$, appear at least twice.

At each stage, we remove as few columns as possible from A_{i-1} to create A_i. To construct A_i from A_{i-1} for $1 \leq i \leq N_0$, we need to remove at most $(N-(i-1))(m-1)$ columns from A_{i-1}, and so A_i has at least $2 + (K - iN - i(i-1)/2)(m-1)$ columns. Similarly, to construct A_{N_0+i} for $1 \leq i \leq N_1$ we need to remove at most $(m-1)$ columns from A_{N_0+i-1} and so A_{N_0+i} has at least $2 + (N_1 - 1 - i)(m-1)$ columns.

Now, A_{N-1} contains at least two columns and so we may choose a column in A_{N-1} as column 1 of our forbidden configuration. We choose column j of our forbidden configuration, where $2 \leq j \leq N_1$, to lie in A_{N-j+1} and to agree with column 1 in row $N + 2 - j$. Property (ii) ensures that we may choose column j to be distinct from column 1. Suppose that the columns $2, 3, \ldots, N_1$ are distinct, so we have a submatrix of A that is isomorphic to the lower part of the forbidden configuration in Lemma 5. We then construct the remainder of the forbidden configuration in a similar fashion:

Property (i) ensures there are enough repeated elements in the upper N_0 rows to guarantee the existence of a new column with the properties we require at each stage.

We have shown that A contains a forbidden configuration, provided the columns $2, 3, \ldots, N_1$ are distinct. So we have a contradiction in this case, as required.

We may derive a contradiction in a similar way if some of the columns $2, 3, \ldots, N_1$ are equal: if there are t fewer distinct columns, then A is not a separating hash family of type $\{w - t, d\}$ and therefore cannot be a separating hash family of type $\{w, d\}$. $\square$

Proof of Lemma 4: Suppose that $\mathcal{F}$ is an $\mathrm{SHF}(N; n, m, \{w_1, w_2, \ldots, w_t\})$ with $N < u$. By adding functions to $\mathcal{F}$ if necessary, we may assume that $N = u - 1$. Since a separating hash family of type $\{w_1, \ldots, w_{t-1}, w_t\}$ is also an separating hash family of type $\{w_1, w_2 + w_3 + \cdots + w_t\}$, we may apply Lemma 6 (in the case when $w = w_1$ and $d = w_2 + w_3 + \cdots + w_t = u - w_1$) to obtain the bound we require. $\square$

It is not difficult to show that the leading coefficient $\binom{u}{2}$ of the bound using Lemma 3 in the previous section is never better than the leading coefficient $w_1(2u - 2w_1 - 1)$ of bound using Lemma 4. So the methods in this section give a better bound. However, the leading coefficient could still be of the order of $u^2/2$ for some parameters: we will see in the next section that this is far from best possible.

4 Improving the leading term

This section provides a third proof of Theorem 1 which gives a better coefficient for the leading term. Indeed, if we replace Lemma 3 from Section 2 by Lemma 7 below, it is easy to see that we obtain the result we require. So it remains to prove Lemma 7.

Lemma 7 *Let $\mathcal{F}$ be an $\mathrm{SHF}(N; n, m, \{w_1, w_2, \ldots, w_t\})$. Let $u = \sum_{i=1}^{t} w_i$. Suppose that $N < u$. Then*

$$n \leq (w_1 w_2 + u - w_1 - w_2)(m - 1) + 1.$$

Proof: Suppose, for a contradiction, that an $\mathrm{SHF}(N; n, m, \{w_1, w_2, \ldots, w_t\})$ $\mathcal{F}$ exists such that $N < u$ and

$$n \geq (w_1 w_2 + u - w_1 - w_2)(m - 1) + 2.$$

By adding additional functions to $\mathcal{F}$ if necessary, we may assume (without loss of generality) that $N = u - 1$. Let X and Y be sets of size n and m respectively, and suppose that $\mathcal{F} = \{f_1, f_2, \ldots, f_{u-1}\}$ where $f_i : X \to Y$.

The first part of our proof aims to show the existence of elements $y, z \in X$ with certain properties that we need. These elements will be then used in the second part of our proof to derive the contradiction we are seeking. For $x \in X$ and $i \in \{1, 2, \ldots, u-1\}$, define the integer $\mu_i(x)$ by

$$\mu_i(x) = \left|\{x' \in X : f_i(x') = f_i(x)\}\right|.$$

Define subsets $X_2, X_3, \ldots, X_{u-1} \subseteq X$ by

$$\begin{aligned}
&X_i = \{x \in X : \mu_i(x) = 1\} \text{ for } 2 \leq i \leq u - w_1,\\
&X_i = \{x \in X : \mu_i(x) \leq w_2\} \text{ for } u - w_1 + 1 \leq i \leq u - 1.
\end{aligned}$$

Note that $|X_i| \leq m - 1$ when $2 \leq i \leq u - w_1$, and $|X_i| \leq w_2(m-1)$ when $u - w_1 + 1 \leq i \leq u - 1$.

Define $X' = X \setminus (X_2 \cup X_3 \cup \cdots X_{u-1})$. Note that

$$\begin{aligned}
|X'| &\geq |X| - \sum_{i=2}^{u-1} |X_i|\\
&\geq n - (u - w_1 - 1)(m-1) - (w_1 - 1)w_2(m-1)\\
&= (m-1) + 2 > m,
\end{aligned}$$

and so we may choose distinct elements $y, z \in X'$ such that $f_1(y) = f_1(z)$. Note, in particular, that

$$\mu_i(y) \geq 2 \text{ for } 2 \leq i \leq u - w_1, \tag{1}$$

since $y \notin X_i$. Moreover,

$$\mu_i(z) \geq w_2 + 1 \text{ for } u - w_1 + 1 \leq i \leq u - 1, \tag{2}$$

since $z \notin X_i$.

We fix the elements $y, z \in X$ above. We use these elements to construct disjoint subsets $C_1, C_2, \ldots, C_t \subseteq X$ with $|C_i| = w_i$ that are not separated by any of the functions $f_1, f_2, \ldots, f_{u-1}$. This produces the contradiction we are seeking. We construct the subsets by the following algorithm, which we justify below:

```
1      Set $C_1 = \{y\}$, $C_2 = \{z\}$ and $C_3 = C_4 = \cdots = C_t = \emptyset$.

2      For $i = 2, 3, \ldots, u - w_1$:
2ia        Choose $k_i \in \{2, 3, \ldots, t\}$ such that $|C_{k_i}| < w_{k_i}$.
2ib        If $f_i$ does not separate $C_1, C_2, \ldots, C_t$:
               choose $a_i \in X \setminus (C_1 \cup C_2 \cup \cdots \cup C_t)$.
2ic        If $f_i$ separates $C_1, C_2, \ldots, C_t$:
               choose $a_i \in X \setminus \{y\}$ such that $f_i(a_i) = f_i(y)$.
2id        Set $C_{k_i} \leftarrow C_{k_i} \cup \{a_i\}$.

3      For $j = 1, 2, \ldots, w_1 - 1$:
3ja        If $f_{u-w_1+j}$ does not separate $C_1, C_2, \ldots, C_t$:
               choose $b_j \in X \setminus (C_1 \cup C_2 \cup \cdots \cup C_t)$.
3jb        If $f_{u-w_1+j}$ separates $C_1, C_2, \ldots, C_t$:
               choose $b_j \in X \setminus C_2$ such that $f_{u-w_1+j}(b_j) = f_{u-w_1+j}(z)$.
3jc        Set $C_1 \leftarrow C_1 \cup \{b_j\}$.
```

We first justify why choices always exist for the index k_i at stage $2i$a, the elements a_i at stage $2i$b or $2i$c and the elements b_j at stage $3j$a or $3j$b. This will show that the algorithm always terminates.

It is clear that at most one element is added to one set at each iteration of the loop at stage 2 of the algorithm. Indeed, at stage $2i$a we have that $|C_1| = 1$ and

$$\sum_{\ell=2}^{t} |C_\ell| \leq i - 1. \tag{3}$$

Our choice of $k_2, k_3, \ldots, k_{i-1}$ shows that $|C_\ell| \leq w_\ell$ for $\ell \in \{2, 3, \ldots, t\}$. We have that

$$i \leq u - w_1 = \sum_{\ell=2}^{t} w_\ell,$$

and so the inequality (3) implies that $|C_\ell| < w_\ell$ for some ℓ. Therefore a choice for k_i exists.

At stage $2i$b of the algorithm, the inequality (3) implies that

$$|C_1 \cup C_2 \cup \cdots \cup C_t| \leq \sum_{\ell=1}^{t} |C_\ell| \leq 1 + i - 1 \leq w_1 + (u - w_1) - 1 < u \leq n,$$

and so a choice for a_i exists. There is a choice for a_i at stage $2i$c since $\mu_i(y) \geq 2$, by (1).

Throughout stage 3, we find (by our choice of $k_2, k_3, \ldots, k_{u-w_1}$) that $|C_\ell| \leq w_\ell$ for $\ell \in \{2, 3, \ldots, t\}$. It is clear that

$$|C_1| \leq j \tag{4}$$

at stage $3j$a and $3j$b. In particular, there is a choice for b_j at stage $3j$a since

$$|C_1 \cup C_2 \cup \cdots \cup C_t| \leq \sum_{\ell=1}^{t} |C_\ell| \leq j + w_2 + w_3 + \cdots < u \leq n.$$

There is a choice for b_j at stage $3j$b, since the inequality (2) implies that $\mu_{u-w_1+j}(y) \geq w_2 + 1 > |C_2|$.

So we find that the algorithm always terminates; moreover it is clear that on termination we have that $|C_\ell| \leq w_\ell$ for $\ell \in \{1, 2, \ldots, t\}$.

We claim that the elements a_i are 'new' when they are chosen, in the sense that $a_i \notin C_1 \cup C_2 \cup \cdots \cup C_t$. This is obvious when a_i is chosen at stage $2i$b. Suppose a_i is chosen at stage $2i$c. We see that $a_i \notin C_1$, since $C_1 = \{y\}$. If $a_i \in C_\ell$ for some $\ell \in \{2, 3, \ldots, t\}$, then

$$f_i(a_i) = f_i(y) \in f_i(C_1) \cap f_i(C_\ell),$$

contradicting our assumption that f_i separates $C_1, C_2, \ldots, C_\ell$. This establishes our claim. Essentially the same argument shows that the elements b_j are new when they are chosen.

The fact that the a_i and b_j are new when they are chosen implies that the subsets $C_1, C_2, \ldots, C_t$ are always disjoint. Moreover, the inequalities (3) and (4) are in fact equalities and so the algorithm terminates with subsets $C_1, C_2, \ldots, C_t$ such that $|C_\ell| = w_\ell$ for $\ell \in \{1, 2, \ldots, t\}$.

Finally, we observe that none of the functions $f_1, f_2, \ldots, f_{u-1}$ separate the sets $C_1, C_2, \ldots, C_t$ we have constructed. To see this, note that

$$f_1(y) = f_1(z) \in f_1(C_1) \cap f_1(C_2)$$

by our choice of y and z and so f_1 does not separate the sets we have constructed. Let $i \in \{2, 3, \ldots, u - w_1\}$. If we chose a_i at stage $2i$b, then f_i fails to separate the sets $C_1, C_2, \ldots, C_t$ constructed at that stage, and adding more elements to the sets C_i subsequently does not change this. If we chose a_i at stage $2i$c then

$$f_i(y) = f_i(a_i) \in f_i(C_1) \cap f_i(C_{k_i}),$$

and so again we find that f_i fails to separate the subsets $C_1, C_2, \ldots, C_t$. Now let $j \in \{1, 2, \ldots, w_1 - 1\}$ and consider the function f_{u-w_1+j}. If we chose b_j

at stage $3j$a, then f_{u-w_1+j} fails to separate the sets $C_1, C_2, \ldots, C_t$ at this stage and so cannot separate the final sets produced by the algorithm. If we chose b_j at stage $3j$b then

$$f_{u-w_1+j}(b_j) = f_{u-w_1+j}(z) \in f_{u-w_1+j}(C_1) \cap f_{u-w_1+j}(C_2),$$

so f_{u-w_1+j} fails to separate the sets $C_1, C_2, \ldots, C_t$.

So the algorithm operates as claimed, and the resulting contradiction establishes the lemma. □

One might ask whether the sets C_i in Lemma 7 could be built up in a different order, leading to a better coefficient γ. In fact this cannot be done, as the following argument indicates.

We can model the problem as follows. Define a sequence of integer vectors $\mathbf{w}^{[2]}, \mathbf{w}^{[3]}, \cdots, \mathbf{w}^{[u]} \in \mathbb{Z}^t$ representing the sizes of the sets C_i as they are gradually built up by an algorithm similar to that in the proof of Lemma 7. So $\mathbf{w}^{[u]} = (w_1, w_2, \ldots, w_t)$, $\mathbf{w}^{[2]} = e_{i_1} + e_{i_2}$ for some $i_1 \neq i_2$ (where e_i is the ith unit vector) and $\mathbf{w}^{[k+1]} = \mathbf{w}^{[k]} + e_{i_k}$ for some $i_k \in \{1, 2, \ldots, t\}$. Define the cost κ_k of the kth step of the algorithm, where we move from $w^{[k]}$ to $w^{[k+1]}$, as the value of the smallest non-zero component that is unchanged during this step. Define the cost of the sequence of vectors as $1+\kappa_2+\kappa_3+\cdots+\kappa_{u-1}$. The argument in the proof of Lemma 7 shows that such an algorithm would lead to a version of Theorem 1 with γ equal to the cost of the sequence of vectors $\mathbf{w}^{[k]}$. (We comment that there are a few minor technical modifications that need to be made to the argument. Most notably, the definition of X' needs to be modified slightly so that the more general algorithm works, but this modification does not significantly affect its size. Moreover, we must often restrict our choice of the elements a_i, b_j to lie in X'.)

We claim that the lowest cost of a sequence of vectors is $w_1w_2 + u - w_1 - w_2$. A sketch proof of this is as follows. Firstly, it is not difficult to show that when $t = 2$ the cost of any sequence of vectors of the right form is w_1w_2. Now consider the general case. The sequence consisting of the pair of smallest non-zero components of each $\mathbf{w}^{[k]}$ starts at $(1, 1)$ and ends at (w_1, w_2), and so (by the special case when $t = 2$) the contribution to the cost of the steps where this pair changes is w_1w_2. But there are $u - w_1 - w_2$ remaining steps, all of which have cost at least 1, giving an overall cost of at least $w_1w_2 + u - w_1 - w_2$. The vectors associated with the algorithm above show this cost can be achieved, and our claim follows.

So we conclude that we cannot improve the coefficient γ further without incorporating new ideas into the proof.

5 Comments

It would be very interesting to know whether the exponent in the bound of Theorem 1 is tight when $u-1$ does not divide N. Let integers N and w_i be fixed. Is it the case that for any positive real number ϵ, an infinite family of $\mathrm{SHF}(N; n, m, \{w_1, w_2, \dots, w_t\})$ exist with $n \geq m^{\lceil N/(u-1)\rceil - \epsilon}$?

Is it possible to improve the coefficient γ in Theorem 1? For work on improving this coefficient in an analogous upper bound for k-IPP codes, see Alon and Stav [2]. The coefficient γ can certainly be reduced when $u-1$ does not divide N (at the expense of introducing a lower order term) using essentially the methods in this paper. (To see how to do this, replace some of the sets I_j defined in the proof of Theorem 1 by sets of size $k-1$. Identify Y^{k-1} with a subset of Y^k in some fashion, so the functions f'_j still map into the set Y^k. The functions f'_j associated with sets I_j of size $k-1$ cannot be surjective, and so the corresponding sets X_i defined in the proof of Lemma 7 are smaller. This leads to a better bound.) It seems more difficult to improve the coefficient γ when $u-1$ divides N: the argument at the end of Section 4 shows that this cannot be done without some new ideas.

References

[1] N. Alon, G. Cohen, M. Krivelevich and S. Litsyn, 'Generalized hashing and parent-identifying codes', *J. Comb. Theory Ser. A* **104** (2003), 207–215.

[2] N. Alon and U. Stav, 'New bounds on parent-identifying codes: The case of multiple parents', *Combinatorics , Probability and Computing* **13** (2004), 795–807.

[3] A. Barg, G. Cohen, S. Encheva, G. Kabatiansky and G. Zémor, 'A hypergraph approach to the identifying parent property: the case of multiple parents', *SIAM J. Disc. Math.* **14** (2001), 423–431.

[4] S.R. Blackburn, 'Perfect hash families: probabilistic methods and explicit constructions', *J. Comb. Theory Ser. A* **92** (2000), 54–60.

[5] S.R. Blackburn, 'Frameproof codes', *SIAM Journal on Discrete Mathematics* **16** (2003), 499–510.

[6] S.R. Blackburn, 'An upper bound on the size of a code with the k-identifiable parent property', *J. Comb. Theory - Series A* **102** (2003) 179–185.

[7] S.R. Blackburn and P.R. Wild, 'Optimal linear perfect hash families', *J. Comb. Theory Ser. A* **83** (1998), 233–250.

[8] J.N. Staddon, D.R. Stinson, and R. Wei, 'Combinatorial properties of frameproof and traceability codes', *IEEE Transactions on Information Theory* **47** (2001), 1042–1049.

[9] D.R. Stinson, T. van Trung and R. Wei, 'Secure frameproof codes, key distribution patterns, group testing algorithms and related structures', *J. Stat. Planning and Inference* **86** (2000), 595–617.

[10] D.R. Stinson, R. Wei and K. Chen, 'On generalised separating hash families', *J. Comb. Theory - Series A*, to appear.

[11] D.R. Stinson and G.M. Zaverucha, 'Some improved bounds for secure frameproof codes and related separating hash families', preprint, February 2007. Available online: `http://www.cacr.math.uwaterloo.ca/~dstinson`.

[12] D.R. Stinson and G.M. Zaverucha, 'New bounds for generalised separating hash families', *CACR Technical Report 2007-21*, University of Waterloo.

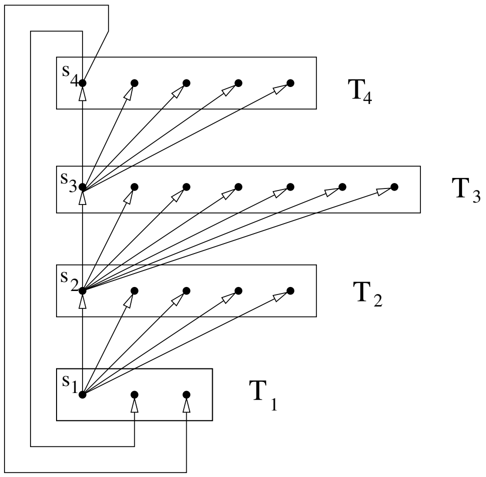

Figure 1: Constructing the tree T