# Hybrid Fingerprints

- an $(\ell, n, q)$-*code* is a subset $\mathcal{C} \subseteq Q^\ell$ such that $|Q| = q$ and $|\mathcal{C}| = n$
- let $\mathcal{C}_0 \subseteq \mathcal{C}$
- define $\mathsf{desc}(\mathcal{C}_0)$ to consist of all $\ell$-tuples $\mathbf{f} = (f_1, \ldots, f_\ell)$ such that, for all $1 \le i \le \ell$, there exists a codeword $\mathbf{c} = (c_1, \ldots, c_\ell) \in \mathcal{C}_0$ such that $f_i = c_i$
- for example,

$$\mathsf{desc}(\{(1,1,2),(2,3,2)\}) = \{(1,1,2),(2,3,2),(1,3,2),(2,1,2)\}$$

- $\mathsf{desc}(\mathcal{C}_0)$ consists of all the hybrid fingerprints that can be constructed from the fingerprints in $\mathcal{C}_0$
- the codewords in $\mathcal{C}_0$ are called the *parents* of the codewords in $\mathsf{desc}(\mathcal{C}_0)\backslash\mathcal{C}_0$

## Descendant Codes and Suspect Coalitions

- for an integer $w \geq 2$, the *$w$-descendant code* of $\mathcal{C}$, denoted $\mathsf{desc}_w(\mathcal{C})$, consists of the following set of $\ell$-tuples:

$$\mathsf{desc}_w(\mathcal{C}) = \bigcup_{\mathcal{C}_0 \subseteq \mathcal{C}, |\mathcal{C}_0| \leq w} \mathsf{desc}(\mathcal{C}_0)$$

- the $w$-descendant code consists of all hybrid fingerprints that could be produced by a coalition of size at most $w$
- suppose that $\mathbf{f} \in \mathsf{desc}_w(\mathcal{C})$
- define the set of *suspect coalitions* as follows:

$$\mathsf{susp}_w(\mathbf{f}) = \{\mathcal{C}_0 \subseteq \mathcal{C} : |\mathcal{C}_0| \leq w, \mathbf{f} \in \mathsf{desc}(\mathcal{C}_0)\}$$

- $\mathcal{C}$ is a *$w$-identifiable parent property* code ($w$-IPP code) provided that, for all $\mathbf{f} \in \mathsf{desc}_w(\mathcal{C})$, there exists a codeword $\mathbf{c} \in \mathcal{C}$ such that $\mathbf{c} \in \mathcal{C}_0$ for all $\mathcal{C}_0 \in \mathsf{susp}_w(\mathbf{f})$

# An Example

We present a $(3,6,3)$ code, and consider coalitions of size at most 2:

$$\mathbf{c}_1 = (0,1,1), \quad \mathbf{c}_2 = (1,0,1), \quad \mathbf{c}_3 = (1,1,0),$$
$$\mathbf{c}_4 = (2,0,2), \quad \mathbf{c}_5 = (1,0,2), \quad \mathbf{c}_6 = (2,1,0)$$

- consider $\mathbf{f}_1 = (1,1,1)$
- $\mathsf{susp}_2(\mathbf{f}_1) = \{\{1,2\},\{1,3\},\{2,3\},\{1,5\},\{2,6\}\}$, which violates the 2-IPP property
- consider $\mathbf{f}_2 = (2,1,2)$
- $\mathsf{susp}_2(\mathbf{f}_2) = \{\{1,4\},\{3,4\},\{4,6\},\{5,6\}\}$, which also violates the 2-IPP property
- the code is not a 2-IPP code

## Definition of IPP Code

- an $(\ell, n, q)$-code is a $w$-IPP code if and only if

$$\bigcap_{\mathcal{C}_0 \in \mathsf{susp}_w(\mathbf{f})} \mathcal{C}_0 \neq \emptyset$$

  for all $\mathbf{f} \in \mathsf{desc}_w(\mathcal{C})$

- given any hybrid fingerprint $\mathbf{f}$ which was created by a coalition of size at most $w$ in a $w$-IPP code, any codeword in the above-defined intersection is an *identifiable parent*
- the $w$-IPP property means that at least one parent is identifiable

## An Example

We present a $(3, 7, 5)$ 2-IPP code:

$$\mathbf{c}_1 = (0,0,0), \quad \mathbf{c}_2 = (0,1,1), \quad \mathbf{c}_3 = (0,2,2), \quad \mathbf{c}_4 = (1,0,3),$$
$$\mathbf{c}_5 = (2,0,4), \quad \mathbf{c}_6 = (3,3,0), \quad \mathbf{c}_7 = (4,4,0)$$

- suppose that $\mathbf{f} = (f_1, f_2, f_3)$ is a hybrid fingerprint
- if any co-ordinate of $\mathbf{f}$ is non-zero, then at least one parent of $\mathbf{f}$ can be identified:

$$f_1 = 1 \Rightarrow \mathbf{c}_4; \quad f_1 = 2 \Rightarrow \mathbf{c}_5; \quad f_1 = 3 \Rightarrow \mathbf{c}_6; \quad f_1 = 4 \Rightarrow \mathbf{c}_7$$
$$f_2 = 1 \Rightarrow \mathbf{c}_2; \quad f_2 = 2 \Rightarrow \mathbf{c}_3; \quad f_2 = 3 \Rightarrow \mathbf{c}_6; \quad f_2 = 4 \Rightarrow \mathbf{c}_7$$
$$f_3 = 1 \Rightarrow \mathbf{c}_2; \quad f_3 = 2 \Rightarrow \mathbf{c}_3; \quad f_3 = 3 \Rightarrow \mathbf{c}_4; \quad f_3 = 4 \Rightarrow \mathbf{c}_5$$

- finally, if $\mathbf{f} = (0,0,0)$, then $\mathbf{c}_1$ must be a parent

## IPP Codes

- in general, it is not an easy task
  - to construct a $w$-IPP code;
  - to verify whether a given code is a $w$-IPP code; or
  - to find an efficient algorithm to identify a parent, given an $\ell$-tuple in the $w$-descendant code of a $w$-IPP code
- it is of interest to design $w$-IPP codes for which efficient parent-identifying algorithms can be constructed
- first, we will pursue these questions in the easiest case, $w = 2$

## Perfect Hash Families

- an $(n, m, w)$-*perfect hash family* is a set of functions, say $\mathcal{F}$, such that $|X| = n$, $|Y| = m$, $f : X \rightarrow Y$ for each $f \in \mathcal{F}$, and for any $X_1 \subseteq X$ such that $|X_1| = w$, there exists at least one $f \in \mathcal{F}$ such that $f|_{X_1}$ is one-to-one
- when $|\mathcal{F}| = N$, an $(n, m, w)$-perfect hash family will be denoted by PHF$(N; n, m, w)$.
- a PHF$(N; n, m, w)$ can be depicted as an $n \times N$ matrix with entries from $Y$, having the property that in any $w$ rows there exists at least one column such that the $w$ entries in the given $w$ rows are distinct
- perfect hash families have been widely studied in the context of information retrieval algorithms

## Separating Hash Families

- an $(n, m, \{w_1, w_2\})$-*separating hash family* is a set of functions, say $\mathcal{F}$, such that $|X| = n$, $|Y| = m$, $f : X \to Y$ for each $f \in \mathcal{F}$, and for any $X_1, X_2 \subseteq X$ such that $|X_1| = w_1$, $|X_2| = w_2$ and $X_1 \cap X_2 = \emptyset$, there exists at least one $f \in \mathcal{F}$ such that

$$\{f(x) : x \in X_1\} \cap \{f(x) : x \in X_2\} = \emptyset$$

- the notation $\mathrm{SHF}(N; n, m, \{w_1, w_2\})$ will be used to denote an $(n, m, \{w_1, w_2\})$-separating hash family with $|\mathcal{F}| = N$.
- an $\mathrm{SHF}(N; n, m, \{w_1, w_2\})$ can be depicted as an $n \times N$ matrix with entries from the set $Y$, having the property that in any $w_1$ rows and any $w_2$ disjoint rows there exists at least one column such that the entries in the given $w_1$ rows are distinct from the entries in the given $w_2$ rows

# An Example

Consider the following 7 by 3 array

| | | |
|---|---|---|
| 0 | 0 | 0 |
| 0 | 1 | 1 |
| 0 | 2 | 2 |
| 1 | 0 | 3 |
| 2 | 0 | 4 |
| 3 | 3 | 0 |
| 4 | 4 | 0 |

- the above array is a $\mathrm{PHF}(3; 7, 5, 3)$
- it is also an $\mathrm{SHF}(3; 7, 5, \{2, 2\})$
- it is *not* a $\mathrm{PHF}(3; 7, 5, 4)$ (consider rows $1, 2, 4$ and $6$)

## 2-IPP Codes

- we will derive an efficient algorithm to determine if a given $(\ell, n, q)$ code, $\mathcal{C}$, is a 2-IPP code
- suppose the codewords are written in the form of an $n$ by $\ell$ array, say $A(\mathcal{C})$
- suppose that $A(\mathcal{C})$ is *not* a PHF$(\ell; n, q, 3)$
  - then there exist three rows, $r_1, r_2, r_3$ of $A$ that violate the PHF property
  - for every column $c$, let $f_c$ be an element that is repeated (i.e., it occurs in at least two of the three given rows in column $c$)
  - define $\mathbf{f} = (f_1, \dots, f_\ell)$
  - clearly $\{r_1, r_2\}, \{r_1, r_3\}, \{r_2, r_3\} \in \mathsf{susp}_w(\mathbf{f})$
  - therefore, $\mathcal{C}$ is not a 2-IPP code

## 2-IPP Codes (cont.)

- suppose that $A(\mathcal{C})$ is *not* an $\mathrm{SHF}(\ell; n, q, \{2,2\})$
  - then there exist two sets of two rows of $A(\mathcal{C})$, $\{r_1, r_2\}$ and $\{r_3, r_4\}$, that violate the SHF property
  - for every column $c$, let $f_c$ be an element such that it occurs in one of rows $r_1$ and $r_2$, and in one of rows $r_3$ and $r_4$, in column $c$
  - define $\mathbf{f} = (f_1, \ldots, f_\ell)$
  - clearly $\{r_1, r_2\}, \{r_3, r_4\} \in \mathsf{susp}_w(\mathbf{f})$
  - therefore, $\mathcal{C}$ is not a 2-IPP code
- hence, a *necessary* condition for $\mathcal{C}$ to be a 2-IPP code is that $A(\mathcal{C})$ is simultaneously a $\mathrm{PHF}(\ell; n, q, 3)$ and an $\mathrm{SHF}(\ell; n, q, \{2,2\})$

## 2-IPP Codes (cont.)

- the converse is also true (proof omitted)
- therefore $\mathcal{C}$ is a 2-IPP code *if and only if* $A(\mathcal{C})$ is simultaneously a PHF$(\ell; n, q, 3)$ and an SHF$(\ell; n, q, \{2, 2\})$
- as a corollary, an $(\ell, n, 2)$ code cannot be a 2-IPP code if $n \geq 3$
- an $(\ell, n, q)$ code, $\mathcal{C}$, can be tested to see if it is a 2-IPP code in time $O(n^4 \ell)$
- if $\mathcal{C}$ is a 2-IPP code and $\mathbf{f} \in \mathsf{desc}_2(\mathcal{C}) \backslash \mathcal{C}$, then we can determine $\mathsf{susp}_2(\mathbf{f})$ in time $O(n^2 \ell)$
- there are two possible cases: either $\mathsf{susp}_2(\mathbf{f})$ consists of a single set of two codewords, both of which are identifiable parents; or $\mathsf{susp}_2(\mathbf{f})$ consists of a two or more sets of two codewords, all of which contain a fixed codeword (i.e., we have one identifiable parent)

## A Construction For 2-IPP Codes with $\ell = 3$

- suppose that $q = r^2 + 2r$
- define

$$\begin{aligned}
S &= \{1, \ldots, r\} \quad (|S| = r)\\
M &= \{r+1, \ldots, 2r\} \quad (|M| = r)\\
L &= \{2r+1, \ldots, q\} \quad (|L| = r^2)\\
\mathcal{C}_1 &= \{(s_1, s_2, rs_1 + s_2 + r) : s_1, s_2 \in S\} \subseteq S \times S \times L\\
\mathcal{C}_2 &= \{(m, sr + m, s) : m \in M, s \in S\} \subseteq M \times L \times S\\
\mathcal{C}_3 &= \{(rm_1 + m_2 - r^2, m_1, m_2) : m_1, m_2 \in M\} \subseteq L \times M \times M
\end{aligned}$$

- $\mathcal{C}_1 \cup \mathcal{C}_2 \cup \mathcal{C}_3$ is a 2-IPP code with $n = 3r^2$
- this code has an (efficient) $O(1)$ time algorithm to find an identifiable parent

## An Example

We present a $(3, 27, 15)$ 2-IPP code:

$$\begin{array}{lll}
\mathbf{c}_1 = (1,1,7), & \mathbf{c}_2 = (1,2,8), & \mathbf{c}_3 = (1,3,9), \\
\mathbf{c}_4 = (2,1,10), & \mathbf{c}_5 = (2,2,11), & \mathbf{c}_6 = (2,2,12), \\
\mathbf{c}_7 = (3,1,13), & \mathbf{c}_8 = (3,2,14), & \mathbf{c}_9 = (3,3,15), \\
\hline
\mathbf{c}_{10} = (4,7,1), & \mathbf{c}_{11} = (5,8,1), & \mathbf{c}_{12} = (6,9,1), \\
\mathbf{c}_{13} = (4,10,2), & \mathbf{c}_{14} = (5,11,2), & \mathbf{c}_{15} = (6,12,2), \\
\mathbf{c}_{16} = (4,13,3), & \mathbf{c}_{17} = (5,14,3), & \mathbf{c}_{18} = (6,15,3), \\
\hline
\mathbf{c}_{19} = (7,4,4), & \mathbf{c}_{20} = (8,4,5), & \mathbf{c}_{21} = (9,4,6), \\
\mathbf{c}_{22} = (10,5,4), & \mathbf{c}_{23} = (11,5,5), & \mathbf{c}_{24} = (12,5,6), \\
\mathbf{c}_{25} = (13,6,4), & \mathbf{c}_{26} = (14,6,5), & \mathbf{c}_{27} = (15,6,6)
\end{array}$$

## Finding an Identifiable Parent

- if $\mathbf{f} = (f_1, f_2, f_3)$ has a co-ordinate in $L$, then a parent is easily identified
  - for example, suppose that $r = 3$ (so $q = 15$) and $f_2 = 13$
  - then $3s + m = 13$, where $s, m \in \{1, 2, 3\}$, so $s = 3$, $m = 4$ and hence $(4, 13, 3)$ is an identifiable parent
- if $\mathbf{f}$ has no co-ordinate in $L$, then it is possible to compute $i \neq j$ such that the two parents of $\mathbf{f}$ are in $\mathcal{C}_i$ and $\mathcal{C}_j$
- the parent that contributed two co-ordinates to $\mathbf{f}$ can then be identified
  - for example, suppose that $\mathbf{f} = (1, 3, 2)$
  - the parents of $\mathbf{f}$ are from $\mathcal{C}_1$ and $\mathcal{C}_2$
  - the parent from $\mathcal{C}_1$ contributes $f_1$ and $f_2$
  - hence $(1, 3, 9)$ is an identifiable parent