Códigos con propiedades de localización basados en matrices de bajo sesgo

José Moreira
Departament d'Enginyeria Telemàtica
Universitat Politècnica de Catalunya
Email: jose.moreira@entel.upc.edu

Marcel Fernández
Departament d'Enginyeria Telemàtica
Universitat Politècnica de Catalunya
Email: marcel@entel.upc.edu

Grigory Kabatiansky
Institute for
Information Transmission Problems
Russian Academy of Sciences
Email: kaba@iitp.ru

***Resumen*—En este artículo presentamos una construcción explícita de un código con propiedades de identificación de traidores, aplicable a entornos de fingerprinting. Nuestro trabajo parte del estudio de una familia de códigos conocidos como *códigos separables*, que en el campo del fingerprinting también se conocen como *códigos seguros contra incriminaciones*. A partir de estos códigos, nos centramos en una versión menos estricta de ellos, en los que no se requiere que la propiedad de separación se cumpla en todos los casos, sino con alta probabilidad. Este tipo de códigos se conocen como *códigos cuasi seguros contra incriminaciones*. En este trabajo mostramos como construir explícitamente estos códigos, basando nuestras construcciones en estructuras conocidas como *matrices de bajo sesgo*. Además, mostramos cómo es posible utilizar dichos códigos para construir de forma explícita una familia de códigos binarios con propiedades de identificación, baja tasa de error y decodificación eficiente.**

***Palabras clave*—Fingerprinting, código seguro contra incriminaciones (*secure frameproof code*), código separable (*separating code*), identificación de traidores (*traitor tracing*)**

I. Introducción

Los códigos con propiedades de localización, también conocidos como códigos de fingerprinting, se utilizan para luchar contra la redistribución ilegítima de contenidos, llevada a cabo por usuarios deshonestos. Un distribuidor que desee proteger un determinado contenido entregará copias marcadas de éste a los usuarios destinatarios. Cada marca introducida identificará a un único usuario. Esto los disuadirá de realizar una redistribución "ingenua" de su copia del contenido. No obstante, puede suceder que diversos usuarios (traidores) confabulen y generen una copia pirata, que no es más que una mezcla, de acuerdo a unas determinadas reglas, de sus propias copias. La copia pirata, por tanto, contendrá una marca corrupta. Por tanto, el objetivo del distribuidor consistirá en determinar un conjunto de marcas tales que sea posible identificar, al menos, a uno de los traidores.

El término de *código seguro contra incriminaciones (secure frameproof code, código SFP)* [1], [2], [3], [4] es el nombre que se dio a los *códigos separables* [5], [6], [7], [8], [9], [10], [11] cuando fueron redescubiertos dentro de los campos del fingerprinting y de la identificación de traidores. Este artículo versa sobre la construcción de lo que denominamos *códigos cuasi seguros contra incriminaciones (almost secure frameproof code, código cuasi SFP)* y su aplicación a la construcción explícita de códigos de fingerprinting. Estos códigos, una versión menos restrictiva de los códigos SFP, fueron introducidos en [12]. En ese trabajo, se mostró su aplicación a la construcción de códigos de fingerprinting, al estilo de [13], mejorando las cotas de existencia previas de dichos códigos. A efectos prácticos, la idea principal consiste en que el reemplazo de códigos SFP por códigos cuasi SFP en ese tipo de construcciones permite al distribuidor utilizar códigos de fingerprinting más cortos, reduciendo así tanto el coste de inserción de las marcas como el coste de identificación de traidores [12], [14].

Sea C un código. Informalmente, diremos que dos conjuntos disjuntos de palabras código $U, V \subseteq C$ son *separados* si existe una posición en la que el conjunto de valores de las palabras de U es disjunto al conjunto de valores de las palabras de V en esa posición. El código C se denomina (c, c)-separable [5], [6], [7], [8], [9], [10], [11] si cada par de conjuntos disjuntos $U, V \subseteq C$ de tamaño c son separados.

Supongamos que, dado un conjunto $U \subseteq C$ de $\leq c$ palabras código, generamos una nueva palabra en la que el valor en cada posición pertenece a alguna de las palabras de U en esa posición. Una palabra generada de esta forma se denomina *descendiente* del conjunto U. Dado que las palabras código corresponden a las marcas de usuario, la generación de un descendiente modela la generación de la marca corrupta de la copia pirata. Un descendiente de U es *unívocamente c-decodificable* si no es descendiente de cualquier otro conjunto disjunto a U de $\leq c$ palabras código. Un código c-SFP es aquel en el que todos los descendientes de conjuntos de $\leq c$ palabras son unívocamente c-decodificables [1], [2], [3], [4]. No es difícil ver que esto es equivalente a la condición descrita para los códigos (c, c)-separables. Es decir, un código (c, c)-separable y un código c-SFP son el mismo concepto.

Consideremos ahora una versión menos estricta de ambas definiciones, en el sentido de no exigir separabilidad completa ni decodificación unívoca completa. Esto nos lleva a considerar dos nociones diferentes, como se expuso en [12]. Un código *cuasi (c, c)-separable* es un código en el que un subconjunto de $\leq c$ palabras código está separado del resto de subconjuntos disjuntos de tamaño $\leq c$ con alta probabilidad. Por otra parte, un código *cuasi c-SFP* es un código en el que cada descendiente es unívocamente c-decodificable con alta probabilidad.

En este artículo conectaremos los conceptos definidos anteriormente con el concepto de *matrices de bajo sesgo* [15], que está estrechamente relacionado con el concepto de espacios

ISBN: 978-84-9717-323-0

probabilísticos de bajo sesgo [16], [17].

Una *matriz binaria de bajo sesgo* es una matriz definida sobre el cuerpo finito de dos elementos, $\mathbb{F}_2 = \{0,1\}$, tal que cualquier combinación lineal de sus columnas tiene, aproximadamente, el mismo número de ceros que de unos.

Bajo unas determinadas condiciones, una matriz binaria de bajo sesgo $A \in \mathbb{F}_2^{n\times M}$ exhibirá la siguiente propiedad: para cualquier subconjunto S de $\leq t$ columnas y cada posible vector $\mathbf{a} \in \mathbb{F}_2^t$, existirá una fila tal que su proyección en las columnas de S coincidirá con $\mathbf{a}$. Una matriz con esta propiedad genera inmediatamente lo que se conoce como un *conjunto (M,t)-universal*. Esta observación será clave para nuestros propósitos, ya que un conjunto $(M,2c)$-universal genera inmediatamente un código (c,c)-separable, es decir, c-SFP.

Como se ha comentado, es fácil ver que un código (c,c)-separable y un código c-SFP son el mismo concepto. No obstante, cuando se consideran sus versiones relajadas ambas nociones difieren. Intuitivamente, parece claro que un código cuasi separable es más restrictivo que un código cuasi SFP. Concretamente, se ha mostrado la existencia de códigos cuasi SFP de tasa mucho mayor que códigos cuasi separables [12]. La estrategia utilizada para establecer dichas cotas de existencia está basada en métodos probabilísticos que, desafortunadamente, no son métodos constructivos que nos lleven a la construcción práctica de estos códigos.

Introducidos estos conceptos, podemos dar una visión general de la estructura de este artículo. En la Sección II mostramos las definiciones formales que necesitaremos, así como una breve revisión de resultados anteriores. Nuestra contribución la presentaremos en la Sección III. Mostraremos como, partiendo de construcciones existentes de matrices de bajo sesgo, podemos obtener construcciones de lo que denominaremos conjuntos cuasi universales. Finalmente, mostramos como esas construcciones pueden emplearse para la construcción explícita de códigos cuasi SFP, lo que culminará con la construcción explícita de un código de fingerprinting en la Sección IV.

II. Definiciones y resultados previos

Dado un alfabeto Q de tamaño $|Q| = q$, denotamos por Q^n el conjunto de todos los vectores q-arios de longitud n. Por ejemplo, $\mathbf{u} = (u_1,\ldots,u_n) \in Q^n$. Un subconjunto $C \subseteq Q^n$ de tamaño M se denomina un *(n,M)-código q-ario*. Los elementos de C se denominan *palabras código*. Si Q es un cuerpo finito de q elementos, lo denotaremos por $\mathbb{F}_q$.

II-A. Códigos cuasi separables y cuasi SFP

Dado un (n,M)-código C, un subconjunto $U = \{\mathbf{u}^1,\ldots,\mathbf{u}^c\} \subseteq C$ de tamaño c se denomina *c-coalición*. Denotaremos por $P_i(U)$ la *proyección* de U en la posición i-ésima, es decir, el conjunto de elementos del alfabeto del código en dicha posición,

$$P_i(U) \stackrel{\text{def}}{=} \{u_i^1,\ldots,u_i^c\}.$$

Dadas dos c-coaliciones $U,V \subseteq C$, diremos que U y V están *separadas* si $P_i(U)\cap P_i(V) = \emptyset$ para alguna posición i. Llamaremos a esa posición una *posición separadora*. También diremos que una c-coalición U es *separada* si está separada de cualquier otra c-coalición del código.

Definición 1: Un código C es *(c,c)-separable* si cualquier par de c-coaliciones $U,V \subseteq C$ tienen una posición separadora. Equivalentemente, si todas las c-coaliciones $U \subseteq C$ son separadas.

Los códigos separables fueron introducidos por Friedman *et al.* en [5] hace más de 40 años. Un código separable es una estructura combinatoria con multitud de aplicaciones, como por ejemplo en la construcción de funciones de hash, testeo de circuitos combinacionales y síntesis de autómatas. Estos códigos han sido posteriormente estudiados por numerosos autores, por ejemplo en [6], [7], [8], [9], [10], [11]. Se han investigado cotas superiores e inferiores sobre su tasa, y se ha mostrado su relación con conceptos matemáticos similares. Véase, por ejemplo, [6] y [10].

Con la aparición del fingerprinting digital, los códigos separables han vuelto a suscitar interés de nuevo. Consideremos una c-coalicón $U = \{\mathbf{u}^1,\ldots,\mathbf{u}^c\} \subseteq C$ de un (n,M)-código $C \subseteq Q^n$. En un ataque de confabulación, las *reglas de marcado (marking assumption)* [18] establecen que las posiciones i tales que todas las palabras de U tienen el mismo símbolo deben permanecer inalteradas en cualquier palabra pirata $\mathbf{z}$ que generen. En concreto, el modelo de generación de palabras pirata que adoptaremos será el conocido como *narrow-sense envelope model* [13], donde para cada posición i tenemos que $z_i \in P_i(U)$. El conjunto de todas las palabras piratas que U puede generar lo denotaremos por $\mathrm{desc}(U)$, y bajo las presuposiciones mencionadas, tenemos que

$$\mathrm{desc}(U) \stackrel{\text{def}}{=} \{\mathbf{z} = (z_1,\ldots,z_n) \in Q^n : z_i \in P_i(U),\ 1 \leq i \leq n\}.$$

Normalmente, a las palabras pirata de $\mathrm{desc}(U)$ se les denomina *descendientes*. También se define el *código c-descendiente* de C, denotado por $\mathrm{desc}_c(C)$, como

$$\mathrm{desc}_c(C) \stackrel{\text{def}}{=} \bigcup_{U\subseteq C, |U|\leq c} \mathrm{desc}(U).$$

Un descendiente $\mathbf{z} \in \mathrm{desc}_c(C)$ es *unívocamente c-decodificable* si $\mathbf{z} \in \mathrm{desc}(U)$ para alguna c-coalición $U \subseteq C$, y $\mathbf{z} \notin \mathrm{desc}(V)$ para cualquier c-coalición V disjunta a U.

Definición 2: Un código C es *c-seguro contra incriminaciones* (c-SFP) si para cualesquiera $U,V \subseteq C$ tales que $|U| \leq c$, $|V| \leq c$ y $U \cap V = \emptyset$, entonces $\mathrm{desc}(U) \cap \mathrm{desc}(V) = \emptyset$. Equivalentemente, si todos los descendientes $\mathbf{z} \in \mathrm{desc}_c(C)$ son unívocamente c-decodificables.

El concepto de código c-SFP fue introducido en [18], [1], [2]. No es difícil ver que, en efecto, se trata de códigos (c,c)-separables.

Sea $R = R(C) \stackrel{\text{def}}{=} n^{-1}\log_q|C|$ la *tasa* de un (n,M)-código sobre un alfabeto q-ario Q. Denotaremos por $R_q(n,c)$ a la tasa máxima que puede alcanzar un código q-ario (c,c)-separable (equivalentemente, c-SFP) de longitud n. Es decir,

$$R_q(n,c) \stackrel{\text{def}}{=} \max_{\substack{C\subseteq Q^n:\, C \text{ es}\\ (c,c)\text{-separable}}} R(C).$$

Consideraremos también los límites asintóticos de dicha tasa

$$\underline{R}_q(c) = \liminf_{n\to\infty} R_q(n,c), \quad \overline{R}_q(c) = \limsup_{n\to\infty} R_q(n,c).$$

En este artículo nos centraremos en códigos sobre el alfabeto binario, es decir, $Q = \{0,1\}$. Para códigos binarios $(2,2)$-separables, se sabe que $\underline{R}_2(2) \geq 0,0642$ [7], [6] (cota también válida para el caso de códigos lineales [7]) y $\overline{R}_2(2) < 0{,}2835$ [6], [9]. Para valores arbitrarios de c, en [13] se obtiene que

$$\underline{R}_2(c) \geq -\frac{\log_2(1-2^{-2c+1})}{2c-1}.$$

Como puede observarse, las cotas de existencias de códigos separables muestran que éstos poseen una tasa muy baja. Con el objetivo de obtener códigos de mejor tasa, en [12] se proponen dos versiones menos estrictas de estos códigos.

Definición 3: Un código $C \subseteq Q^n$ es *ε-cuasi (c,c)-separable* si la proporción de coaliciones separadas de tamaño c, entre todas las posibles coaliciones de tamaño c, es $\geq 1-\varepsilon$.

Una secuencia de códigos $(C_i)_{i\geq 1}$ de longitud n_i creciente es una *familia asintóticamente cuasi (c,c)-separable* si cada código C_i es un código ε_i-cuasi (c,c)-separable y $\lim_{i\to\infty} \varepsilon_i = 0$.

Definición 4: Un código $C \subseteq Q^n$ es *ε-cuasi c-SFP* si la proporción de descendientes $\mathbf{z} \in \mathrm{desc}_c(C)$ unívocamente c-decodificables es $\geq 1-\varepsilon$.

Una secuencia de códigos $(C_i)_{i\geq 1}$ de longitud n_i creciente es una *familia asintóticamente cuasi c-SFP* si cada código C_i es un código ε_i-cuasi c-SFP y $\lim_{i\to\infty} \varepsilon_i = 0$.

Cabe destacar que las definiciones anteriores permiten separar los conceptos de "separación" y "decodificación unívoca", que coincidían en el caso de códigos completamente separables y SFP. Además, estas nuevas definiciones permiten obtener códigos con mayor tasa.

Para una familia de códigos $\mathcal{C} = (C_i)_{i\geq 1}$ definimos su tasa asintótica como

$$R(\mathcal{C}) \stackrel{\text{def}}{=} \liminf_{i\to\infty} R(C_i).$$

Nuestro interés reside en estimar el valor máximo de dicha tasa entre todas las familias de códigos asintóticamente cuasi (c,c)-separables y asintóticamente c-SFP. Denotaremos estas tasas asintóticas por $R_q^{\text{sep}*}(c)$ y $R_q^{\text{SFP}*}(c)$, respectivamente.

Por ejemplo, para el caso binario y coaliciones de tamaño $c=2$ tenemos que $R_2^{\text{sep}*}(2) \geq 0{,}1142$, de [14], y $R_2^{\text{SFP}*}(2) \geq 0{,}2075$, de [12].

II-B. Matrices de bajo sesgo

En esta sección presentamos los conceptos sobre matrices de bajo sesgo que utilizaremos más adelante en nuestras construcciones. Para una explicación más detallada, remitimos al lector a las referencias [17], [16], [15].

Como se ha comentado, nos centraremos en el caso binario, ya que nuestro objetivo final será la construcción de códigos binarios. Por tanto, de aquí en adelante trabajaremos con el alfabeto $\mathbb{F}_2$.

Una *(n,M)-matriz binaria* A es una matriz de tamaño $n \times M$ donde sus elementos pertenecen a $\mathbb{F}_2$. Dada una (n,M)-matriz binaria A y un subconjunto de posiciones $S \subseteq \{1,\ldots,M\}$ de tamaño s, Denotamos por $\nu_S(\mathbf{a};A)$ al número de filas de A cuyas proyecciones en las posiciones de S coincide con el vector $\mathbf{a} \in \mathbb{F}_2^s$. Obviamente, un vector $\mathbf{u} \in \mathbb{F}_2^n$ puede verse como una $(n,1)$-matriz binaria. En este caso, $\nu_{\{1\}}(0;\mathbf{u})$ y $\nu_{\{1\}}(1;\mathbf{u})$ denotan el número de ceros y el número de unos de $\mathbf{u}$, respectivamente.

Definición 5: Sea $\mathbf{u} = (u_1,\ldots,u_n) \in \mathbb{F}_2^n$. El *sesgo* del vector $\mathbf{u}$ se define como

$$n^{-1}|\nu_{\{1\}}(0;\mathbf{u}) - \nu_{\{1\}}(1;\mathbf{u})|.$$

Es decir, un vector $\mathbf{u}$ con, aproximadamente, el mismo número de ceros y de unos tendrá bajo sesgo.

Definición 6: Sea $0 \leq \varepsilon < 1$. Una (n,M)-matriz binaria A es *(ε,t)-sesgada* si cualquier combinación lineal no trivial de $\leq t$ de sus columnas tiene sesgo $\leq \varepsilon$. Si $t = M$ diremos simplemente que A es ε-sesgada (o que tiene sesgo ε).

Definición 7: Sea $0 \leq \varepsilon < 1$. Una (n,M)-matriz binaria A es *ε-cuasi t-independiente* si para cualquier subconjunto $S \subseteq \{1,\ldots,M\}$ de $s \leq t$ columnas satisface

$$\sum_{\mathbf{a}\in\mathbb{F}_2^s} |n^{-1}\nu_S(\mathbf{a};A) - 2^{-s}| \leq \varepsilon.$$

Para nuestros propósitos, el concepto más importante que necesitaremos será el de conjunto (M,t)-universal.

Definición 8: Un *conjunto (M,t)-universal* B es un subconjunto de $\mathbb{F}_2^M$ tal que para cualquier subconjunto $S \subseteq \{1,\ldots,M\}$ de t posiciones, el conjunto de las proyecciones de los elementos de B en las posiciones S contiene todos los vectores $\mathbf{a} \in \mathbb{F}_2^t$.

Dada una (n,M)-matriz binaria A, si para cualquier subconjunto $S \subseteq \{1,\ldots,M\}$ de t columnas y cualquier vector $\mathbf{a} \in \mathbb{F}_2^s$, se satisface $\nu_S(\mathbf{a};A) > 0$, entonces las filas de A forman un conjunto (M,t)-universal. Nos interesan conjuntos universales del mínimo tamaño posible.

En [16] se establece la conexión entre conjuntos universales y matrices cuasi independientes.

Proposición 9: Sea A una (n,M)-matriz binaria. Para $\varepsilon \leq 2^{-t}$, si A es ε-cuasi t-independiente, entonces las filas de A forman un conjunto (M,t)-universal de tamaño n.

Además, en [19], [20], [16] también se relacionan estos conceptos con matrices de bajo sesgo.

Corolario 10: Sea A una (n,M)-matriz binaria. Si A es ε-sesgada, entonces A es $2^{t/2}\varepsilon$-cuasi t-independiente.

Por tanto, la construcción de conjuntos universales se reduce a la construcción de matrices cuasi independientes mediante la Proposición 9, que a su vez se reduce a la construcción de matrices de bajo sesgo, mediante el Corolario 10. Más adelante, presentaremos una construcción de matrices cuasi independientes aún más eficiente que la aplicación directa del Corolario 10.

III. Construcciones

En esta sección presentamos nuestras construcciones de códigos cuasi SFP. Antes de entrar en detalles explícitos, daremos un razonamiento intuitivo de nuestra propuesta.

No es difícil ver que, en un (n, M)-código aleatorio binario, la probabilidad de que dos coaliciones de tamaño c esten separadas se maximiza cuando se generan las palabras código según un vector de probabilidad $\mathbf{p} = (p_1, \ldots, p_n)$ con $p_1 = \cdots = p_n = 1/2$. Es decir, generamos al azar M palabras código $(u_1, \ldots, u_n)$ tales que $\Pr\{u_i = 1\} = p_i = 1/2$. Pero ya que estamos interesados en códigos ε-cuasi c-SFP, se permitirá un pequeño sesgo en estas probabilidades y, por tanto, consideraremos matrices de bajo sesgo.

Existen construcciones explícitas de (n, M)-matrices ε-sesgadas con $n = 2^{O(\log M + \log \varepsilon^{-1})}$ [16], que conducen a obtener conjuntos (M, t)-universales, de tamaño $2^{O(t)} \log M$. Si disponemos los vectores de este conjunto universal como las filas de una matriz, las columnas de esta matriz formarán un código c-SFP para $t = 2c$. Este código tendrá tamaño M, longitud $2^{O(2c)} \log M$ y tasa $2^{-O(2c)}$. Nuestra idea consist en relajar la propiedad de conjunto universal y permitir que un determinado número de vectores $\mathbf{a} \in \mathbb{F}_2^t$ no aparezcan en cada proyección de t posiciones del conjunto. Esto da lugar a conjuntos "cuasi universales". Finalmente demostramos que los conjuntos "cuasi universales" se pueden utilizar para generar códigos ε-cuasi c-SFP.

III-A. Conjuntos universales y cuasi universales

Los conjuntos universales se han descrito en la Definición 8, y mostrado que la construcción de conjuntos universales se puede reducir a la construcción de matrices ε-sesgadas.

Es fácil ver que un conjunto $(M, 2c)$-universal de tamaño n genera un (n, M)-código (c, c)-separable, es decir, c-SFP [21]. Para ello, sea A una (n, M)-matriz cuyas filas forman un conjunto $(M, 2c)$-universal, y tomemos las columnas de A como las palabras de un código C. Consideremos dos c-coaliciones disjuntas $U, V \subseteq C$, es decir, $2c$ columnas de A. Debido a que las filas de A forman un conjunto $(M, 2c)$-universal, entonces para las $2c$ columnas seleccionadas aparezcan todos los posibles vectores $\mathbf{a} \in \mathbb{F}_2^{2c}$. En particular, hay una fila i donde todas las columnas correspondientes a U contienen el símbolo 0 y todas las columnas correspondientes a V contienen el símbolo 1. Por tanto, i es una posición separadora para las coaliciones U, V. Es decir, $P_i(U) \cap P_i(V) = \emptyset$, como se deseaba.

Construcciones eficientes de conjuntos $(M, 2c)$-universales usando matrices ε-cuasi t-independiente se presentan en [16], en virtud de la Proposición 9 y el Corolario 10. Estas construcciones dan lugar a códigos c-SFP de longitud $2^{O(c)} \log M$. Utilizando esta idea, nuestro objetivo es relajar la restricción que la $(M, 2c)$-universalidad impone para asi obtener una matriz más corta, es decir, un código de mejor tasa. De hecho, no es necesario que todos los posibles vectores de $\mathbb{F}_2^{2c}$ aparezcan en el código. Por lo tanto, nos proponemos relajar la Definición 8 permitiendo que un número máximo de vectores $\mathbf{a} \in \mathbb{F}_2^{2c}$, por ejemplo z, no aparezcan en la proyección de cualquier subconjunto $S \subseteq \{1, \ldots, M\}$ de $2c$ posiciones. Esto se formaliza en la siguiente definición.

Definición 11: Un conjunto (M, t, z)-universal B es un subconjunto de $\mathbb{F}_2^M$ tal que para cada subconjunto $S \subseteq \{1, \ldots, M\}$ de t posiciones el conjunto de proyecciones de los elementos de B sobre los índices de S contiene todos los vectores $\mathbf{a} \in \mathbb{F}_2^t$ excepto, como máximo, z.

De nuevo, si A es una (n, M)-matriz, las filas de A generan un conjunto (M, t, z)-universal siempre que existan al menos $2^t - z$ vectores $\mathbf{a} \in \mathbb{F}_2^t$ tales que $\nu_S(\mathbf{a}; A) > 0$, para todos los subconjuntos $S \subseteq \{1, \ldots, M\}$ de t columnas.

De manera similar a la Proposición 9, el siguiente resultado muestra la conexión entre conjuntos (M, t, z)-universales y matrices ε-cuasi t-independientes.

Proposición 12: Sea A una matriz binaria (n, M). Para $\varepsilon \leq (z+1)2^{-t}$, si A es ε-cuasi t-independiente, entonces las filas de A generan un conjunto (M, t, z)-universal de tamaño n.

Demostración: Por contradicción, supondremos que las filas de A no generan generan un conjunto (M, t, z)-universal. En otras palabras, existe un subconjunto $S \subseteq \{1, \ldots, M\}$ de t columnas tales que hay más de z vectores $\mathbf{a} \in \mathbb{F}_2^t$ tal que $\nu_S(\mathbf{a}; A) = 0$. Para este subconjunto particular S se tiene que

$$\sum_{\mathbf{a} \in \mathbb{F}_2^t} |n^{-1} \nu_S(\mathbf{a}; A) - 2^{-t}| \geq (z+1)2^{-t} + \sum_{\substack{\mathbf{a} \in \mathbb{F}_2^t \text{ t.q.} \\ \nu_S(\mathbf{a}; A) > 0}} |n^{-1} \nu_S(\mathbf{a}; A) - 2^{-t}| \geq (z+1)2^{-t+1} > \varepsilon,$$

que contradice el hecho que la matriz A sea ε-cuasi t-independiente. ■

III-B. Conjuntos (M, t, z)-universales

Según la Proposición 12, la construcción de conjuntos (M, t, z)-universales se reduce a construir una matriz binaria $(z+1)2^{-t}$-cuasi t-independiente, y según el Corolario 10, esto se reduce a la construcción de una matriz ε-sesgada. En realidad, la matriz A del Corolario 10 se puede tomar como una matriz (ε, t)-sesgada, que es una condición menos restrictiva que una matriz ε-sesgada.

En [16] se puede encontrar una construcción estándar para matrices binarias (ε, t)-sesgadas.

Teorema 13: Sea A una (n, M')-matriz binaria ε-sesgada, y sea H la matriz de paridad de un código binario lineal de longitud M, dimensión $M - M'$ y distancia mínima $t + 1$. Entonces, el producto matricial $A \times H$ es una(n, M)-matriz binaria (ε, t)-sesgada.

Habitualmente, la matriz H utilizada en el Teorema 13 es la matriz de paridad de un código BCH binario. En este caso, la matriz H tendrá M columnas y $M' = t \log M$ filas. Entonces, empleando el Teorema 13 en el Corolario 10, el número de filas de una (n, M)-matriz binaria ε-cuasi t-independiente se puede reducir de $n = 2^{O(t + \log M + \log \varepsilon^{-1})}$ a $n = 2^{O(t + \log\log M + \log \varepsilon^{-1})}$ [16].

El problema ahora se reduce a obtener (n, M')-matrices binarias ε-sesgadas con el menor número posible de filas. En [17] se presentan construcciones explicitas de dichas matrices tales que su número de filas es

$$n \leq 2^{2(\log_2 M' + \log_2 \varepsilon^{-1})}.$$

Tabla I
TASAS DE CÓDIGO ALCANZABLES PARA CONSTRUCCIONES EXPLÍCITAS DE CÓDIGOS ε-CUASI c-SFP DE TAMAÑOS ENTRE 10^3 Y 10^7

			Tamaño del código				
c	z	$\log_2 \varepsilon$	$M = 10^3$	$M = 10^4$	$M = 10^5$	$M = 10^6$	$M = 10^7$
2	0	n/a	$1{,}531 \times 10^{-6}$	$1{,}148 \times 10^{-6}$	$9{,}187 \times 10^{-7}$	$7{,}656 \times 10^{-7}$	$6{,}562 \times 10^{-7}$
2	1	n/a	$6{,}124 \times 10^{-6}$	$4{,}593 \times 10^{-6}$	$3{,}675 \times 10^{-6}$	$3{,}062 \times 10^{-6}$	$2{,}625 \times 10^{-6}$
2	2	$-5{,}336 \times 10^{5}$	$1{,}378 \times 10^{-5}$	$1{,}034 \times 10^{-5}$	$8{,}268 \times 10^{-6}$	$6{,}890 \times 10^{-6}$	$5{,}906 \times 10^{-6}$
2	3	$-3{,}001 \times 10^{5}$	$2{,}450 \times 10^{-5}$	$1{,}837 \times 10^{-5}$	$1{,}470 \times 10^{-5}$	$1{,}225 \times 10^{-5}$	$1{,}050 \times 10^{-5}$
3	0	n/a	$1{,}063 \times 10^{-8}$	$7{,}975 \times 10^{-9}$	$6{,}380 \times 10^{-9}$	$5{,}316 \times 10^{-9}$	$4{,}557 \times 10^{-9}$
3	1	n/a	$4{,}253 \times 10^{-8}$	$3{,}190 \times 10^{-8}$	$2{,}552 \times 10^{-8}$	$2{,}127 \times 10^{-8}$	$1{,}823 \times 10^{-8}$
3	2	$-3{,}567 \times 10^{7}$	$9{,}569 \times 10^{-8}$	$7{,}177 \times 10^{-8}$	$5{,}742 \times 10^{-8}$	$4{,}785 \times 10^{-8}$	$4{,}101 \times 10^{-8}$
3	3	$-2{,}006 \times 10^{7}$	$1{,}701 \times 10^{-7}$	$1{,}276 \times 10^{-7}$	$1{,}021 \times 10^{-7}$	$8{,}506 \times 10^{-8}$	$7{,}291 \times 10^{-8}$
3	5	$-8{,}917 \times 10^{6}$	$3{,}828 \times 10^{-7}$	$2{,}871 \times 10^{-7}$	$2{,}297 \times 10^{-7}$	$1{,}914 \times 10^{-7}$	$1{,}640 \times 10^{-7}$
3	6	$-6{,}551 \times 10^{6}$	$5{,}210 \times 10^{-7}$	$3{,}908 \times 10^{-7}$	$3{,}126 \times 10^{-7}$	$2{,}605 \times 10^{-7}$	$2{,}233 \times 10^{-7}$

En [15, Teorema 10] se presenta una mejor construcción explícita con un valor inferior de n. Lamentablemente, las condiciones requeridas en esa construcción hacen que no sea aplicable a nuestro caso, por lo que tenemos que recurrir a la construcción de [17].

A continuación, resumimos los pasos para la construcción explícita de un conjunto (M, t, z)-universal.

1. Tomar $\varepsilon = (z+1)2^{-3t/2}$.
2. Construir una (n, M')-matriz A' ε-sesgada, donde $M' = t \log M$.
3. Construir la matriz de paridad H de un código BCH binario de longitud M, codimensión $M' = t \log M$ y distancia mínima $t+1$.
4. El producto matricial $A = A' \times H$ genera una (n, M)-matriz binaria (ε, t)-sesgada.
5. La matriz A es también ε'-cuasi t-independiente, con $\varepsilon' = 2^{t/2}\varepsilon = (z+1)2^{-t}$. Por tanto, las filas de A generan un conjunto (M, t, z)-universal.

Empleando la construcción de (n, M')-matrices binarias ε-sesgadas proporcionada en [17], el conjunto (M, t, z)-universal resultante tendrá tamaño

$$n = 2^{2(3\,t/2 + \log_2 t + \log_2 \log_2 M - \log_2(z+1))}.$$

III-C. *Códigos cuasi SFP*

Se ha visto que un conjunto $(M, 2c)$-universal de tamaño n genera un (n, M)-código c-SFP. Consideremos una (n, M)-matriz binaria A cuyas filas generan un conjunto $(M, 2c, z)$-universal B, y tomemos las columnas de A como las palabras de un código ε-cuasi c-SFP C, y por tanto tendrá tasa $R = \log M/n$. Para $z < 2^c$ el conjunto $(M, 2c, z)$-universal B es, de hecho, un conjunto (M, c)-universal. Para ver esto, nótese que si un vector de $\mathbb{F}_2^c$ no apareciese en una proyección de c posiciones de B, significaría que faltan $\geq 2^c$ vectores de $\mathbb{F}_2^{2c}$ en alguna proyección de $2c$ columnas. Esto contradice la definición de conjunto $(M, 2c, z)$-universal con $z < 2^c$.

Dado un código C construido usando un conjunto $(M, 2c, z)$-universal, para facilitar el análisis, supondremos que por cada c-coalición $U \subseteq C$, cada posible vector de $\mathbb{F}_2^c$ aparece aproximadamente con probabilidad uniforme (ya que el conjunto $(M, 2c, z)$-universal ha sido generado a partir de una matriz cuasi independiente).

El siguiente corolario formaliza la relación entre códigos ε-cuasi c-SFP y conjuntos (M, t, z)-universales.

Corolario 14: Sean $M > 0$, $c \geq 2$, $z < 2^c$, y $\varepsilon \geq p(M, c, z)$, donde

$$p(M, c, z) \stackrel{\text{def}}{=} M^c (1 - 2^{-c})^n.$$

Entonces, un conjunto $(M, 2c, z)$-universal de tamaño n genera un código ε-cuasi c-SFP de tasa $R = \log M/n$.

Demostración: Considérese un código C generado a partir de un conjunto $(M, 2c, z)$-universal. Sea $\mathbf{z}$ un descendiente generado por una c-coalición del código, $\mathbf{z} \subseteq \text{desc}_c(C)$. Por las suposiciones anteriores, la probabilidad que $\mathbf{z}$ pertenezca a otra c-coalición V es $(1 - 2^{-c})^n$. Por tanto, usando la desigualdad de Boole, se puede acotar la probabilidad de que $\mathbf{z}$ sea generada por otra coalición del código como

$$p(M, c, z) = M^c (1 - 2^{-c})^n.$$

El cociente (probabilidad) de descendientes que no son unívocamente c-decodificables en $\text{desc}_c(C)$ es por tanto $\leq p(M, c, z)$, lo que significa que C es un código ε-cuasi c-SFP. ■

III-D. *Resultados*

En la Tabla I se muestran las tasas obtenidas de códigos cuasi c-SFP para el caso de coaliciones de tamaños $c = 2$ y 3. En número máximo de configuraciones $\{0, 1\}^{2c}$ que faltan se denota como z, y la probabilidad que un descendiente no sea unívocamente c-decodificable se denota mediante ε. Obsérvese que cuando $z < 2$ el código es c-SFP, es decir $\varepsilon = 0$. El valor de ε dado para una fila corresponde al peor caso. Los valores de tasa que se obtienen son del orden de, aproximadamente, 10 veces el valor de la tasa que se obtendría para construcciones explícitas de códigos SFP ordinarios.

IV. APLICACIÓN A CÓDIGOS DE FINGERPRINTING

En esta sección se muestra cómo un código binario ε-cuasi c-SFP se puede utilizar para construir una familia de códigos de fingerprinting equipados con un algoritmo de decodificación eficiente.

Para que un esquema de fingerprinting tenga una probabilidad de error baja, un solo código no es suficiente y se necesita una *familia de códigos* $\{C_j\}_{j \in T}$, siendo T un

conjunto finito. La familia $\{C_j\}_{j\in T}$ es pública. El distribuidor elige un código C_j con probabilidad $\pi(j)$. Esta elección se mantiene en secreto.

En [14, Corolario 1] se proponen condiciones de existencia de una familia de códigos de fingerprinting concatenados, que usan un código cuasi separable como código interno. Obsérvese que el código cuasi separable puede ser sustituido por un código cuasi SFP. Combinando este hecho con los resultados presentados en este trabajo, obtenemos una construcción explícita de un código de fingerprinting.

Corolario 15: Sea $C_{\text{out}} \subseteq \mathbb{F}_q^n$ un código de Reed-Solomon extendido de tasa

$$R_o = R(C_{\text{out}}) < \frac{1-\sigma}{c(c+1)},$$

y sea C_{in} un (l, q)-código ε-cuasi c-SFP de tasa $R_i = R(C_{\text{in}})$, con $\varepsilon < \sigma$. Entonces, existe una construcción explícita de una família de códigos binarios de fingerprinting $\{C_j\}_{j\in T}$ con código externo C_{out} y código interno C_{in}, con un algoritmo de identificación en tiempo polinómico, tasa $R = R_i R_o$ y probabilidad de error decreciendo exponencialmente como

$$p_e \leq 2^{-n\, l(\frac{1-\sigma}{c} R_i - (c+1)R + o(1))} + 2^{-nD(\sigma\|\varepsilon)}.$$

Por último, vale la pena señalar aquí que, como se muestra en [12], [14], el uso de códigos ε-cuasi c-SFP en lugar de códigos SFP ordinarios introduce un término de error adicional en el proceso de identificación, como se indica en el Corolario 15. Afortunadamente, este término de error disminuye exponencialmente con la longitud del código exterior.

V. Conclusiones

Los códigos cuasi separables y cuasi SFP son dos versiones menos restrictivas de los códigos separables. En este trabajo, hemos presentado las primeras construcciones explícitas de códigos cuasi SFP.

Nuestro trabajo parte del estudio de la conexión entre matrices de bajo sesgo y conjuntos universales, y la posterior conexión entre conjuntos universales y códigos separables.

A partir de esta idea, hemos introducido una relajación en la definición de conjunto universal. Se demuestra que un conjunto cuasi universal se puede utilizar para construir un código cuasi SFP. Esta observación nos ha llevado a las construcciones explícitas de códigos cuasi SFP.

También hemos demostrado cómo las construcciones propuestas pueden ser usadas para construir de forma explícita una familia de códigos concatenados de fingerprinting. La construcción presentada se basa en los resultados teóricos de existencia de un trabajo anterior, que presupone la existencia de códigos cuasi SFP. Por lo tanto, una de las principales aportaciones de este trabajo ha sido la de proporcionar una implementación "verdadera" de de dicha existencia teórica de un esquema de fingerprinting.

Por último, cabe señalar que a pesar de que un conjunto universal genera un código separable, la relación entre un conjunto cuasi universal y un código cuasi separable no es en absoluto evidente y será objeto de investigación futura.

Agradecimientos

J. Moreira y M. Fernández han sido financiados por el Gobierno de España mediante los proyectos CONSOLIDER INGENIO 2010 CSD2007-00004 "ARES" y TEC2011-26491 "COPPI", y por la Generalitat de Catalunya mediante la ayuda 2009 SGR-1362.

G. Kabatiansky ha sido financiado por la Russian Foundation for Basic Research mediante las ayudas RFBR 13-07-00978 y RFBR 13-01-12458.

Referencias

[1] D. R. Stinson, T. van Trung, and R. Wei, "Secure frameproof codes, key distribution patterns, group testing algorithms and related structures," *J. Stat. Plan. Infer.*, vol. 86, no. 2, pp. 595–617, May 2000.

[2] J. N. Staddon, D. R. Stinson, and R. Wei, "Combinatorial properties of frameproof and traceability codes," *IEEE Trans. Inf. Theory*, vol. 47, no. 3, pp. 1042–1049, Mar. 2001.

[3] D. Tonien and R. Safavi-Naini, "Explicit construction of secure frameproof codes," *Int. J. Pure Appl. Math.*, vol. 6, no. 3, pp. 343–360, 2003.

[4] D. R. Stinson and G. M. Zaverucha, "Some improved bounds for secure frameproof codes and related separating hash families," *IEEE Trans. Inf. Theory*, vol. 54, no. 6, pp. 2508–2514, June 2008.

[5] A. D. Friedman, R. L. Graham, and J. D. Ullman, "Universal single transition time asynchronous state assignments," *IEEE Trans. Comput.*, vol. C-18, no. 6, pp. 541–547, June 1969.

[6] Y. L. Sagalovich, "Separating systems," *Probl. Inform. Transm.*, vol. 30, no. 2, pp. 105–123, 1994.

[7] M. S. Pinsker and Y. L. Sagalovich, "Lower bound on the cardinality of code of automata's states," *Probl. Inform. Transm.*, vol. 8, no. 3, pp. 59–66, 1972.

[8] Y. L. Sagalovich, "Completely separating systems," *Probl. Inform. Transm.*, vol. 18, no. 2, pp. 140–146, 1982.

[9] J. Körner and G. Simonyi, "Separating partition systems and locally different sequences," *SIAM J. Discr. Math. (SIDMA)*, vol. 1, no. 3, pp. 355–359, Aug. 1988.

[10] G. D. Cohen and H. G. Schaathun, "Asymptotic overview on separating codes," Deptartment of Informatics, University of Bergen, Norway, Tech. Rep. 248, Aug. 2003.

[11] G. D. Cohen and H. G. Schaathun, "Upper bounds on separating codes," *IEEE Trans. Inf. Theory*, vol. 50, no. 6, pp. 1291–1294, June 2004.

[12] M. Fernández, G. Kabatiansky, and J. Moreira, "Almost separating and almost secure frameproof codes," in *Proc. IEEE Int. Symp. Inform. Theory (ISIT)*, Saint Petersburg, Russia, Aug. 2011, pp. 2696–2700.

[13] A. Barg, G. R. Blakley, and G. Kabatiansky, "Digital fingerprinting codes: Problem statements, constructions, identification of traitors," *IEEE Trans. Inf. Theory*, vol. 49, no. 4, pp. 852–865, Apr. 2003.

[14] J. Moreira, G. Kabatiansky, and M. Fernández, "Lower bounds on almost-separating binary codes," in *Proc. IEEE Int. Workshop Inform. Forensics, Security (WIFS)*, Foz do Iguaçu, Brazil, Nov. 2011, pp. 1–6.

[15] J. Bierbrauer and H. Schellwat, "Almost independent and weakly biased arrays: Efficient constructions and cryptologic applications," in *Proc. Int. Cryptol. Conf. (CRYPTO)*, ser. Lecture Notes Comput. Sci. (LNCS), vol. 1880, Santa Barbara, CA, Aug. 2000, pp. 533–544.

[16] J. Naor and M. Naor, "Small-bias probability spaces: Efficient constructions and applications," *SIAM J. Comput. (SICOMP)*, vol. 22, no. 4, pp. 838–856, Aug. 1993.

[17] N. Alon, O. Goldreich, J. Håstad, and R. Peralta, "Simple constructions of almost k-wise independent random variables," *Random Struct. Alg.*, vol. 3, no. 3, pp. 289–304, 1992.

[18] D. Boneh and J. Shaw, "Collusion-secure fingerprinting for digital data," *IEEE Trans. Inf. Theory*, vol. 44, no. 5, pp. 1897–1905, Sept. 1998.

[19] U. V. Vazirani, "Randomness, adversaries and computation," Ph.D. dissertation, Dept. Elect. Eng. Comp. Sci., Univ. California, Berkeley, 1986.

[20] P. Diaconis, *Group Representations in Probability and Statistics*. Beachwood, OH: Inst. Math. Stat., 1988.

[21] N. Alon, V. Guruswami, T. Kaufman, and M. Sudan, "Guessing secrets efficiently via list decoding," *ACM Trans. Alg.*, vol. 3, no. 4, pp. 1–16, Nov. 2007.