Almost k-wise Independent Sample Spaces and Their Cryptologic Applications

Kaoru Kurosawa[1], Thomas Johansson[2], Douglas Stinson[3]

[1] Dept. of Computer Science
Graduate School of Information Science and Engineering
Tokyo Institute of Technology
2-12-1 O-okayama, Meguro-ku, Tokyo 152, Japan
kurosawa@ss.titech.ac.jp

[2] Dept. of Information Technology, Lund University,
PO Box 118, S-22100 Lund, Sweden
thomas@it.lth.se

[3] Dept. of Computer Science and Engineering
University of Nebraska
Lincoln NE 68588, USA
stinson@bibd.unl.edu

Abstract. An almost k-wise independent sample space is a small subset of m bit sequences in which any k bits are "almost independent". We show that this idea has close relationships with useful cryptologic notions such as multiple authentication codes (multiple A-codes), almost strongly universal hash families and almost k-resilient functions.

We use almost k-wise independent sample spaces to construct new efficient multiple A-codes such that the number of key bits grows linearly as a function of k (here k is the number of messages to be authenticated with a single key). This improves on the construction of Atici and Stinson [2], in which the number of key bits is $\Omega(k^2)$.

We also introduce the concept of ϵ-almost k-resilient functions and give a construction that has parameters superior to k-resilient functions.

Finally, new bounds (necessary conditions) are derived for almost k-wise independent sample spaces, multiple A-codes and balanced ϵ-almost k-resilient functions.

1 Introduction

An *almost k-wise independent sample space* is a probability space on m-bit sequences such that any k bits are almost independent. A *ϵ-biased sample space* is a space in which any (boolean) linear combination of the m bits has the value 1 with probability close to 1/2. These notions were introduced by Naor and Naor [17] and further studied in [1] due to their applications to algorithms and complexity theory. However, there are also cryptographic applications: Krawczyk applied ϵ-biased sample spaces to the construction of authentication codes [13].

In this paper, we investigate several new relationships between almost k-wise independent sample spaces and useful cryptologic notions such as multiple

W. Fumy (Ed.): Advances in Cryptology - EUROCRYPT '97, LNCS 1233, pp. 409-421, 1997.

authentication codes (multiple A-codes) [2] and k-resilient functions [10, 3, 11, 24, 4].

In a multiple A-code, $k \geq 2$ messages are authenticated with the same key. (In "usual" A-codes, just one message is authenticated with a given key.) Recently, Atici and Stinson [2] defined some new classes of almost strongly universal hash families which allowed the construction of multiple A-codes. Here, we prove that almost k-wise independent sample spaces are equivalent to multiple A-codes. This allows us to obtain a more efficient construction of multiple A-codes from the almost k-wise independent sample spaces of [1].

Next, we present a lower bound on the size of the keyspace in a multiple A-code. Numerical examples show that the multiple A-codes we construct are quite close to this bound. Further, from the above equivalence, a lower bound on the size of almost k-wise independent sample spaces is obtained for free. (While a lower bound on the size of ϵ-biased sample spaces was given in [1], no lower bound was known for the size of almost k-wise independent sample spaces.)

Finally, we generalize the idea of resilient functions. A function $\phi : \{0,1\}^m \to \{0,1\}^l$ is called *k-resilient* if every possible output l-tuple is equally likely to occur when the values of k arbitrary inputs are fixed by an opponent and the remaining $m - k$ input bits are chosen at random. This is a useful tool for achieving key renewal: an m-bit secret key $(x_1, \cdots, x_m)$ can be renewed to a new l-bit secret key $\phi(x_1, \cdots, x_m)$ about which an opponent has no information if the opponent knows at most k bits of $(x_1, \cdots, x_m)$.

We show that k can be made larger if the definition of resilient function is slightly relaxed. Thus, we define an ϵ-almost k-resilient function as a function ϕ such that every possible output l-tuple is almost equally likely to occur when the values of k arbitrary inputs are fixed by an opponent. (The statistical difference between the output distribution of a k-resilient function and an ϵ-almost k-resilient function is ϵ.) We prove that a large set of almost k-wise independent sample spaces is equivalent to a balanced ϵ-almost k-resilient function, generalizing a result of [24]. From this equivalence, we are able to obtain both efficient constructions and bounds for balanced ϵ-almost k-resilient functions.

2 Almost k-wise independent sample spaces

Let $S_m \subseteq \{0,1\}^m$, and let $X = x_1 \cdots x_m$ be chosen uniformly from S_m.

Definition 1. [1] We say that S_m is an *(ϵ, k)-independent sample space* if for any k positions $i_1 < i_2 < \cdots < i_k$ and any k-bit string α, we have

$$|\Pr[x_{i_1} x_{i_2} \cdots x_{i_k} = \alpha] - 2^{-k}| \leq \epsilon. \tag{1}$$

If $\epsilon = 0$, then S_m is equivalent to an *orthogonal array* $\mathrm{OA}_\lambda(k, m, 2)$, where $\lambda = |S_m|/2^k$.

The following efficient construction for (ϵ, k)-independent sample spaces is proved in [1].

Proposition 2. *There exists an* (ϵ, k)*-independent sample space* S_m *such that*

$$\log_2 |S_m| = 2(\log_2 \log_2 m - \log_2 \epsilon + \log_2 k - 1).$$

In this section, we prove that almost k-wise independent sample spaces are equivalent to multiple authentication codes (more precisely, almost strongly universal-k hash families, as defined in [2]). This allows us to obtain more efficient multiple A-codes than were previously known.

2.1 Multiple A-codes and ASU-k hash families

We briefly review basic concepts of (multiple) authentication codes. In the usual Simmons model of authentication codes (A-codes) [21, 22], there are three participants, a *transmitter*, a *receiver* and an *opponent*. In an *A-code without secrecy*, the transmitter sends a *message* (s, a) to the receiver, where s is a *source state* (plaintext) and a is an *authenticator*. The authenticator is computed as $a = e(s)$, where e is a secret *key* shared between the transmitter and the receiver. The key e is chosen according to a specified probability distribution.

In a *multiple* A-code, we suppose that an opponent observes $i \geq 2$ messages which are sent using the same key. Then the opponent places a new bogus message (s', a') into the channel, where s' is distinct from the i source states already sent. This attack is called a *spoofing attack of order i*. P_{d_i} denotes the success probability of a spoofing attack of order i, see [15].

Almost strongly universal hash families are a very useful way of constructing practical A-codes. This idea was introduced by Wegman and Carter [26], and further developed and refined in papers such as [23, 5, 13, 12]. Atici and Stinson [2] generalized the definitions so that they could be applied to multiple A-codes. We review these definitions now.

Definition 3. An $(N; m, n)$ *hash family* is a set F of N functions such that $f : A \to B$ for each $f \in F$, where $|A| = m, |B| = n$ and $m > n$.

Definition 4. An $(N; m, n)$ hash family F of functions from A to B is ϵ *almost strongly universal-k* (or ϵ-ASU $(N; m, n, k)$) provided that, for all distinct elements $x_1, x_2, \cdots, x_k \in A$, and for all (not necessary distinct) $y_1, y_2, \cdots, y_k \in B$, we have

$$|\{f \in F : f(x_i) = y_i, 1 \leq i \leq k\}| \leq \epsilon \times |\{f \in F : f(x_i) = y_i, 1 \leq i \leq k-1\}|.$$

The following result gives the connection between ϵ-ASU $(N; m, n, k)$ hash families and multiple A-codes.

Proposition 5. *[2] There exists an A-code without secrecy for m source states, having n authenticators and N equiprobable authentication rules and such that $P_{d_{k-1}} \leq \epsilon$, if and only if there exists an ϵ-ASU $(N; m, n, k)$ hash family F.*

2.2 Equivalence of hash families and sample spaces

We can can rephrase Definition 1 in terms of hash families, and generalize it to the non-binary case, as follows.

Definition 6. An $(N; m, n)$ hash family F of functions from A to B is (ϵ, k)-*independent* if for all distinct elements $x_1, x_2, \cdots, x_k \in A$, and for all (not necessary distinct) $y_1, y_2, \cdots, y_k \in B$, we have

$$|\Pr(f(x_i) = y_i, 1 \leq i \leq k) - n^{-k}| \leq \epsilon, \tag{2}$$

where $f \in F$ is chosen uniformly at random.

The following results are straightforward.

Proposition 7. *An (ϵ, k)-independent sample space S_m is equivalent to an (ϵ, k)-independent $(|S_m|; m, 2)$ hash family.*

Proposition 8. *If there exists an (ϵ, k)-independent sample space S_m, then there exists an $(\epsilon, k/t)$-independent $(|S_m|; m/t, 2^t)$ hash family.*

Now we show the equivalence of (ϵ, k)-independent sample spaces and almost strongly universal-k hash families.

Theorem 9. *If F is an (ϵ, k)-independent $(N; m, n)$ hash family, then F is a δ-ASU $(N; m, n, k)$ hash family, where*

$$\delta = \frac{(n^{-k} + \epsilon)}{n(n^{-k} - \epsilon)}.$$

Proof. Suppose that Eq. (2) holds. Then for any $y_1, \cdots, y_k \in B$, we have

$$\Pr[f(x_i) = y_i, 1 \leq i \leq k] \geq n^{-k} - \epsilon,$$
$$\sum_{y_k \in B} \Pr[f(x_i) = y_i, 1 \leq i \leq k] \geq \sum_{y_k \in B} (n^{-k} - \epsilon), \quad \text{and}$$
$$\Pr[f(x_i) = y_i, 1 \leq i \leq k-1] \geq n(n^{-k} - \epsilon).$$

From the above inequality and Eq. (2), we have

$$\frac{\Pr[f(x_i) = y_i, 1 \leq i \leq k]}{\Pr[f(x_i) = y_i, 1 \leq i \leq k-1]} \leq \frac{n^{-k} + \epsilon}{n(n^{-k} - \epsilon)}.$$

Let $\delta \triangleq (n^{-k} + \epsilon)/(n(n^{-k} - \epsilon))$. Then

$$|\{f \in F : f(x_i) = y_i, 1 \leq i \leq k\}| \leq \delta \times |\{f \in F : f(x_i) = y_i, 1 \leq i \leq k-1\}|.$$

Hence, F is a δ-ASU $(N; m, n, k)$ hash family. □

Definition 10. An $(N;m,n)$ hash family F of functions from A to B is *strongly (ϵ,k)-independent* if for any t such that $1 \le t \le k$ and for all distinct elements $x_1, x_2, \cdots, x_t \in A$, and for all (not necessary distinct) $y_1, y_2, \cdots, y_t \in B$, we have

$$|\Pr(f(x_i) = y_i, 1 \le i \le t) - n^{-t}| \le \epsilon \tag{3}$$

where $f \in F$ is chosen uniformly at random.

Theorem 11. *If an $(N;m,n)$ hash family F is strongly (ϵ,k)-independent, then F is a δ-ASU $(N;m,n,k)$ hash family, where $\delta = (n^{-k} + \epsilon)/(n^{-(k-1)} - \epsilon)$.*

Proof. The proof is similar to the proof of Theorem 9. □

Lemma 12. *[2] Suppose that a hash family F of functions from A to B is ϵ-ASU $(N;m,n,k)$. Then for for all $1 \le j \le k$, for all distinct elements $x_1, x_2, \cdots, x_j \in A$, and for all (not necessary distinct) $y_1, y_2, \cdots, y_j \in B$, we have*

$$|\{f \in F : f(x_i) = y_i, 1 \le i \le j\}| \le \epsilon^j \times N \tag{4}$$

Lemma 13. *[2] If a hash family F is ϵ-ASU $(N;m,n,k)$, then $\epsilon \ge 1/n$.*

Theorem 14. *If a hash family F is ϵ-ASU $(N;m,n,k)$, then F is (δ,k)-independent, where $\delta = (n^k - 1)(\epsilon^k - n^{-k})$.*

Proof. From Lemma 12, we have

$$\Pr[f(x_i) = y_i, 1 \le i \le k] \le \epsilon^k \quad \text{and} \tag{5}$$

$$\Pr[f(x_i) = y_i, 1 \le i \le k] - n^{-k} \le \epsilon^k - n^{-k}. \tag{6}$$

On the other hand, from eq.(5), we have

$$\sum_{(\hat{y}_1,\cdots,\hat{y}_k)\ne(y_1,\cdots,y_k)} \Pr[f(x_i) = \hat{y}_i, 1 \le i \le k] \le (n^k - 1)\epsilon^k.$$

Therefore, we have

$$\begin{aligned}\Pr[f(x_i) = y_i, 1 \le i \le k] &= 1 - \sum_{(\hat{y}_1,\cdots,\hat{y}_k)\ne(y_1,\cdots,y_k)} \Pr[f(x_i) = \hat{y}_i, 1 \le i \le k] \\ &\ge 1 - (n^k - 1)\epsilon^k.\end{aligned}$$

Hence,

$$\begin{aligned}\Pr[f(x_i) = \hat{y}_i, 1 \le i \le k] - n^{-k} &\ge 1 - (n^k - 1)\epsilon^k - n^{-k} \\ &= 1 - \epsilon^k n^k + \epsilon^k - n^{-k} \\ &= -(n^k - 1)(\epsilon^k - n^{-k}).\end{aligned}$$

From Lemma 13, we see that $\epsilon^k - n^{-k} \ge 0$. Hence,

$$-(n^k - 1)(\epsilon^k - n^{-k}) \le \Pr[f(x_i) = \hat{y}_i, 1 \le i \le k] - n^{-k} \le \epsilon^k - n^{-k}$$

Then the family is (δ,k)-independent, where

$$\delta = \max\{|\epsilon^k - n^{-k}|, |-(n^k - 1)(\epsilon^k - n^{-k})|\} = (n^k - 1)(\epsilon^k - n^{-k})$$

□

2.3 New multiple A-codes

By combining Propositions 2 and 8 with Theorem 9 or Theorem 11, we can obtain new multiple A-codes (ASU-k hash families) from an (ϵ, k)-independent sample space. Since the (ϵ, k)-independent sample spaces from [1] mentioned in Proposition 2 can be shown to be strong, we will apply Theorem 11.

Theorem 15. *There exists a δ-ASU $(N; m, n, k)$ hash family where*

$$\log_2 N = 2(\log_2 \log_2(m \log_2 n) + k \log_2 n - \log_2(n\delta - 1) + \log_2(k \log_2 n) - 1). \quad (7)$$

Proof. Define $l = k \log_2 n$, $u = m \log_2 n$, and

$$\epsilon = \frac{n^{-k}(\delta n - 1)}{\delta + 1} \approx n^{-k}(\delta n - 1).$$

Apply Proposition 2 and 8, constructing a strongly (ϵ, k)-independent (N, m, n) hash family, where $\log_2 N = 2(\log_2 \log_2 u - \log_2 \epsilon + \log_2 l - 1)$. Now apply Theorem 11, to obtain a δ-ASU $(N; m, n, k)$ hash family. We compute $\log_2 N$ as

$$\begin{aligned}\log_2 N &= 2(\log_2 \log_2(m \log_2 n) - \log_2(n^{-k}(\delta n - 1)) + \log_2(k \log_2 n) - 1) \\ &= 2(\log_2 \log_2(m \log_2 n) + k \log_2 n - \log_2(\delta n - 1) + \log_2(k \log_2 n) - 1).\end{aligned}$$

□

3 A lower bound

In this section, we present a lower bound on the size of ASU-k hash families and almost k-wise independent sample spaces.

Theorem 16. *If there exists an ϵ-ASU$(N; m, n, k)$ hash family such that*

$$\epsilon^k \leq 1/n, \quad (8)$$

then

$$N \geq \frac{1}{\epsilon^k} \left(\frac{\log\left(\frac{mn}{k-1}\right)}{\log\left(\frac{1-\epsilon^k}{\frac{1}{n}-\epsilon^k}\right)} - 1 \right).$$

Proof. Suppose F is an ϵ-ASU$(N; m, n, k)$ hash family from A to B, where $|A| = m$, $|B| = n$ and $k \geq 2$. Construct an $N \times mn$ binary matrix $G = (g_{ij})$, with rows indexed by the functions in F and columns indexed by $A \times B$, defined by the rule

$$g_{f,(x,y)} = \begin{cases} 1 \text{ if } f(x) = y \\ 0 \text{ if } f(x) \neq y. \end{cases}$$

Interpret the columns of G as incidence vectors of the N-set F. We obtain a set-system $(F, \mathcal{C} = \{C_{x,y} : x \in A, y \in B\})$, where

$$C_{x,y} = \{f \in F : f(x) = y\}$$

for all $x \in A$, $y \in B$. Let

$$t \triangleq \lfloor \epsilon^k N \rfloor + 1. \tag{9}$$

This set-system satisfies the following properties: (A) $|F| = N$, (B) $|\mathcal{C}| = mn$, (C) $\sum_{C \in \mathcal{C}} |C| = Nm$, (D) there does not exist a subset of t points that occurs as a subset of k different blocks (see Lemma 12).

Property (D) says that $(F, \mathcal{C})$ is a *t-packing of index* $\lambda = k - 1$ (i.e., no t-subset of points occurs in more than λ blocks). Hence we obtain the following:

$$\lambda \binom{N}{t} \geq \sum_{C \in \mathcal{C}} \binom{|C|}{t}. \tag{10}$$

Property (C) implies that the average block size is $Nm/mn = N/n$. Define a real-valued function $f(x)$ as

$$f(x) = \begin{cases} 0 & \text{if } x < t \\ x(x-1)\ldots(x-t+1) & \text{otherwise.} \end{cases}$$

Since $f(x)$ is convex, we have

$$\frac{\lambda}{mn}\binom{N}{t} \geq \frac{1}{mn}\sum_{C \in \mathcal{C}} \binom{|C|}{t} \geq \frac{f(N/n)}{t!} \tag{11}$$

from Jensen's inequality. We observe that $N/n \geq t - 1$ follows from Eq. (8) and Eq. (9). Then, we obtain

$$(k-1)\frac{N(N-1)\cdots(N-t+1)}{\frac{N}{n}\left(\frac{N}{n}-1\right)\cdots\left(\frac{N}{n}-t+1\right)} \geq mn, \tag{12}$$

and hence

$$(k-1)\left(\frac{N-t+1}{\frac{N}{n}-t+1}\right)^t \geq mn. \tag{13}$$

From Eq. (9), we have $t \leq \epsilon^k N + 1$. Then Eq. (13) can be simplified as follows.

$$(k-1)\left(\frac{1-\epsilon^k}{\frac{1}{n}-\epsilon^k}\right)^t \geq mn, \quad \text{and hence}$$

$$(\epsilon^k N + 1)\log\left(\frac{1-\epsilon^k}{\frac{1}{n}-\epsilon^k}\right) \geq \log\left(\frac{mn}{k-1}\right),$$

from which our bound is obtained. □

Corollary 17. *Suppose* S_m *is an* (ϵ, k)*-independent sample space. Denote* $\delta = (2^{-k}+\epsilon)/(2(2^{-k}-\epsilon))$. *If* $\delta^k \leq 1/2$, *then*

$$|S_m| \geq \frac{1}{\delta^k}\left(\frac{\log\left(\frac{2m}{k-1}\right)}{\log\left(\frac{1-\delta^k}{\frac{1}{2}-\delta^k}\right)} - 1\right).$$

Proof. This follows from Theorem 9. □

3.1 Some numerical examples of multiple A-codes

We give some numerical examples to compare the multiple A-codes constructed by Atici and Stinson in [2], our new multiple A-codes obtained from Theorem 15, and the lower bound of Theorem 16. Suppose we want an authentication code for $m = 2^{2^{128}}$ source states with deception probability $\delta = 2^{-40}$. We tabulate the number of key bits (i.e., $\log_2 N$) for $k = 3, 4, 10$. Note that we take $n = 2/\delta = 2^{41}$ in Theorem 15 and Theorem 16 (whereas in [2], $n > 2/\delta$).

k	[2]	Theorem 15	Lower bound
3	657	518	243
4	1043	602	283
10	5376	1096	523

A counter-based multiple authentication scheme would (of course) require less key bits than the proposed construction. For example, tabulated values from [2] show that the construction from [5] would for the parameters above and $k = 4$ require 447 key bits. Hence, the $602 - 447 = 155$ additional key bits we use can be thought of as the price payed for having a stateless multiple authentication scheme. An interesting property that can be verified through Theorem 15 is the following. When $k \to \infty$, the number of key bits required per message approaches $\log_2 n$, which is the same as for the counter-based multiple authentication scheme.

4 Almost resilient functions

In what follows, let $m \geq l \geq 1$ be integers and let $\phi : \{0,1\}^m \to \{0,1\}^l$.

Definition 18. ϕ is called an *(m, l, k)-resilient function* if

$$\Pr[\phi(x_1, \ldots, x_m) = (y_1, \ldots, y_l) \mid x_{i_1} x_{i_2} \cdots x_{i_k} = \alpha] = 2^{-l}$$

for any k positions $i_1 < \cdots < i_k$, for any k-bit string α and for any $(y_1, \cdots, y_l) \in \{0,1\}^l$, where the values x_j ($j \notin \{i_1, \ldots, i_k\}$) are chosen independently at random.

Resilient functions have been studied in several papers, e.g., [10, 3, 11, 24, 4]. We now introduce a generalization, which we call ϵ-almost resilient functions, in which the the output distribution may deviate from the uniform distribution by a small amount ϵ.

Definition 19. We say that ϕ is an *ϵ-almost (m, l, k)-resilient function* if

$$|\Pr[\phi(x_1, \ldots, x_m) = (y_1, \ldots, y_l) \mid x_{i_1} x_{i_2} \cdots x_{i_k} = \alpha] - 2^{-l}| \leq \epsilon$$

for any k positions $i_1 < \cdots < i_k$, for any k-bit string α and for any $(y_1, \cdots, y_l) \in \{0,1\}^l$, where the values x_j ($j \notin \{i_1, \ldots, i_k\}$) are chosen independently at random.

4.1 Relation with (ϵ, k)-independent sample space

It is well-known that a resilient function is equivalent to a large set of orthogonal arrays [24]. Here we prove a similar result for almost resilient functions that involves k-wise independent sample spaces.

Definition 20. A *large set of* (ϵ, k, m, t)*-independent sample spaces*, denoted $LS(\epsilon, k, m, t)$, is a set of 2^{m-t} (ϵ, k, m, t)-independent sample spaces, each of size 2^t, such that their union contains all 2^m binary vectors of length m.

Theorem 21. *If there exists an* $LS(\epsilon, k, m, t)$*, then there exists a* δ*-almost* $(m, m-t, k)$*-resilient function, where* $\delta = \epsilon/2^{m-t-k}$.

Proof. There are 2^{m-t} (ϵ, k)-independent sample spaces in the set. Name the (ϵ, k)-independent sample spaces C_γ, $\gamma \in \{0,1\}^{m-t}$. Then define a function $\phi : \{0,1\}^m \to \{0,1\}^{m-t}$ by the rule

$$\phi(x_1, \ldots, x_m) = \gamma \text{ if and only if } (x_1, \ldots, x_m) \in C_\gamma.$$

For any k positions $i_1 < \cdots < i_k$, any k-bit string α and any $\gamma \in \{0,1\}^{m-t}$, let

$$L \triangleq |\{(x_1, \ldots, x_m) : x_{i_1} \cdots x_{i_k} = \alpha, (x_1, \ldots, x_m) \in C_\gamma\}|.$$

Then

$$\Pr[\phi(x_1, \ldots, x_m) = \gamma \mid x_{i_1} x_{i_2} \cdots x_{i_k} = \alpha] = \frac{L}{2^{m-k}}. \tag{14}$$

From Definition 1, we have

$$2^{-k} - \epsilon \le \frac{L}{2^t} \le 2^{-k} + \epsilon. \tag{15}$$

Hence, from (14) and (15), we obtain

$$|\Pr[\phi(x_1, \ldots, x_m) = \gamma \mid x_{i_1} x_{i_2} \cdots x_{i_k} = \alpha] - 2^{-(m-t)}| \le \frac{\epsilon}{2^{m-t-k}}.$$

□

Definition 22. The function $\phi : \{0,1\}^m \to \{0,1\}^l$ is called *balanced* if we have

$$\Pr[\phi(x_1, \ldots, x_m) = (y_1, \ldots, y_l)] = 2^{-l}$$

for all $(y_1, \cdots, y_l) \in \{0,1\}^l$.

For balanced functions, we can prove the converse of Theorem 21.

Theorem 23. *If there exists a balanced* ϵ*-almost* (m, l, k)*-resilient function,* ϕ*, then there exists an* $LS(\delta, k, m, m-l)$*, where* $\delta = \epsilon/2^{k-l}$.

Proof. For $\gamma \in \{0,1\}^l$, let

$$C_\gamma \triangleq \{(x_1,\ldots,x_m) : \phi(x_1,\ldots,x_m) = \gamma\}.$$

Since ϕ is balanced, $|C_\gamma| = 2^{m-l}$. If each C_γ is an (ϵ,k)-independent sample space, then we automatically get a large set. For any k positions $i_1 < \cdots < i_k$, for any k-bit string α for and any $\gamma \in \{0,1\}^l$, let

$$L \triangleq |\{(x_1,\ldots,x_m) : x_{i_1}\cdots x_{i_k} = \alpha, (x_1,\ldots,x_m) \in C_\gamma\}|.$$

Then, within the sample space C_γ, we have

$$\Pr[x_{i_1}x_{i_2}\cdots x_{i_k} = \alpha] = \frac{L}{|C_\gamma|} = \frac{L}{2^{m-l}}. \tag{16}$$

From Definition 19, we get

$$2^{-l} - \epsilon \le \frac{L}{2^{m-k}} \le 2^{-l} + \epsilon. \tag{17}$$

Hence, from (16) and (17), we obtain

$$|\Pr(x_{i_1}x_{i_2}\cdots x_{i_k} = \alpha) - 2^{-k}| \le \frac{\epsilon}{2^{k-l}}.$$

□

4.2 Constructions of ϵ-almost resilient functions

Definition 24. An (ϵ,k)-independent sample space S_m is *t-systematic* if $|S_m| = 2^t$, and there exist t positions $i_1 < \cdots < i_t$ such that each t-bit string occurs in these positions for exactly one m-tuple in S_m.

A t-systematic (ϵ,k)-independent sample space can be transformed into an $LS(\epsilon,k,m,t)$ by using the same technique as [25, Theorem 3]. We have the following result.

Theorem 25. *If there exists a t-systematic (ϵ,k)-independent sample space S_m, then there exists a balanced δ-almost $(m, m-t, k)$-resilient function, where $\delta = \epsilon/2^{m-t-k}$.*

Due to space limitations, we will present only a very brief summary of our construction for t-systematic (ϵ,k)-independent sample spaces. Our approach is similar to [12] (see also [18]), and depends on the Weil-Carlitz-Uchiyama bound. In what follows, let Tr denote the *trace* function from $GF(2^t)$ to $GF(2)$.

Proposition 26 Weil-Carlitz-Uchiyama bound. *[9] Let $f(x) = \sum_{i=1}^{D} f_i x^i \in GF(2^t)[x]$ be a polynomial that is not expressible in the form $f(x) = g(x)^2 - g(x) + \theta$ for any polynomial $g(x) \in GF(2^t)[x]$ and for any $\theta \in F_{2^t}$. Then*

$$\left| \sum_{\alpha \in GF(2^t)} (-1)^{Tr(f(\alpha))} \right| \le (D-1)\sqrt{2^t}.$$

Definition 27. A polynomial $h(x) \in GF(2^t)[x]$ is a $(2^t, D)$*-polynomial* if h has degree at most D and $a_i = 0$ for all even i, where $h = \sum_{i=0}^{D} a_i x^i$. Define $H(2^t, D, k)$ to be a set of $(2^t, D)$-polynomials such that any k polynomials in the set are independent over $GF(2)$.

For $h_{i_1}, h_{i_2}, \ldots, h_{i_k} \in H(2^t, D, k)$ and for any k elements $\alpha_1, \cdots, \alpha_k \in GF(2)$, define

$$N_{\alpha_1,\ldots,\alpha_k}(h_{i_1}, \ldots, h_{i_k}) \triangleq |\{x \in GF(2^t) : Tr(h_{i_1}(x)) = \alpha_1, \cdots, Tr(h_{i_k}(x)) = \alpha_k\}|.$$

Lemma 28. *[12]* $|N_{\alpha_1,\ldots,\alpha_k}(h_{i_1}, \ldots, h_{i_k}) - 2^{t-k}| \leq (D-1)\sqrt{2^t}$.

Proof. The proof is an application of Proposition 26. The case $k = 2$ can be found in [12] and the general case is proved similarly. □

Theorem 29. *Suppose that β is a primitive element of $GF(2^t)$, and $H(2^t, D, k)$ is chosen such that $\{x, \beta x, \beta^2 x, \ldots, \beta^{t-1} x\} \subseteq H(2^t, D, k)$. There exists a t-systematic (ϵ, k)-independent sample space S_m where $m = |H(2^t, D, k)|$ and $\epsilon = (D-1)/\sqrt{2^t}$.*

Proof. Let $H(2^t, D, k) = \{h_1, \cdots, h_m\}$. Construct a sample space S_m as follows: A binary string $X_\gamma = x_1 x_2 \cdots x_m \in S_m$ is specified by any $\gamma \in GF(2^t)$, where the ith bit of X_γ is $x_i = Tr(h_i(\gamma))$. The proof that S_m is (ϵ, k)-independent follows from Lemma 28. Further, S_m can be shown to be systematic using the fact that $\{x, \beta x, \beta^2 x, \ldots, \beta^{t-1} x\} \subseteq H(2^t, D, k)$ (the proof will be given in the final paper). □

4.3 An Application

In our approach, using Theorem 29, we need to construct a set of polynomials $H(2^t, D, k)$ such that any k of them are linearly independent over $GF(2)$. For this we can use linear error-correcting codes (see [14]). For a fixed (odd) degree D, we can express each polynomial as a linear combination of polynomials in the set

$$\{x, \beta x, \ldots, \beta^{t-1} x, x^3, \beta x^3, \ldots, \beta^{t-1} x^3, \ldots, x^D, \beta x^D, \ldots, \beta^{t-1} x^D\}.$$

Indexing the polynomials in $H(2^t, D, k)$ as $h_1, h_2, \ldots, h_m$ we obtain a binary $tD' \times m$ matrix, where $D' = (D+1)/2$, which is a parity check matrix of an $[m, l, d]$ error correcting code in which $m - l = tD'$ and $d = k + 1$. Conversely, given such a code, we obtain a t-systematic sample space, and hence a balanced ϵ-almost $(m, m-t, k)$-resilient function, as follows.

Theorem 30. *Suppose $D = 2D' - 1$ and there is a $[m, m - tD', k+1]$ code. Then there exists a balanced ϵ-almost $(m, m-t, k)$-resilient function such that*

$$\epsilon = \frac{(D-1)\sqrt{2^t}}{2^{m-k}}.$$

A suitable value of ϵ would be 2^{-m+t-1}. We obtain the following corollary of Theorem 30 by taking $D = 3$ and $k = (t/2) - 2$.

Corollary 31. *Suppose there is an $[m, m-4k-8, k+1]$ code. Then there exists a balanced $2^{-m+2k+3}$-almost $(m, m-2k-4, k)$-resilient function.*

As a typical example, suppose we take $m = 160$ and $k = 18$. A $[160, 80, 23]$ code is known to exist see ([6]), so we obtain a balanced 2^{-121}-almost $(160, 120, 18)$-resilient function.

Let's compare the above result to the best-known $(160, 120, k)$-resilient function. The most important construction method for resilient functions [3, 10] uses linear error-correcting codes, as follows: Let G be a generator matrix for an $[m, l, d]$ linear code. Define a function $f : (GF(2))^m \mapsto (GF(2))^l$ by the rule $f(x) = xG^T$. Then f is an $(m, l, d-1)$ linear resilient function. The maximum d for which a $[160, 120, d]$ code is known to exist is $d = 12$ (see [6]). Hence, the maximum k for which we can construct a $(160, 120, k)$-resilient function is $k = 11$.

5 Comments

The techniques of this paper can also be used to construct "almost" versions of other cryptographic tools. These include *correlation-immune functions* (see, for example, [19, 8, 7]) and *locally random pseudo-random number generators* (see [20, 16, 18]). Details will be given in the full version of the paper.

References

1. N. Alon, O. Goldreich, J. Hastad, and R. Peralta. Simple constructions of almost k-wise independent random variables. *Random Structures and Algorithms* **3** (1992), 289–304.
2. M. Atici and D. R. Stinson. Universal hashing and multiple authentication. *Lecture Notes in Computer Science* **1109** (1996), 16–30 (CRYPTO '96).
3. C. H. Bennett, G. Brassard, and J.-M. Robert. Privacy amplification by public discussion. *SIAM Journal on Computing* **17** (1988), 210–229.
4. J. Bierbrauer, K. Gopalakrishnan and D. R. Stinson. Bounds for resilient functions and orthogonal arrays. *Lecture Notes in Computer Science* **839** (1994), 247–257 (CRYPTO '94).
5. J. Bierbrauer, T. Johansson, G. Kabatianskii and B. Smeets. On families of hash functions via geometric codes and concatenation. *Lecture Notes in Computer Science* **773** (1994), 331–342 (CRYPTO '93).
6. A. E. Brouwer. Bounds on the minimum distance of binary linear codes. `http://www.win.tue.nl/win/math/dw/voorlincod.html`
7. P. Camion and A. Canteaut. Generalization of Siegenthaler inequality and Schnorr-Vaudenay multipermutations. *Lecture Notes in Computer Science* **1109** (1996), 372–386 (CRYPTO '96).
8. P. Camion, C. Carlet, P. Charpin and N. Sendrier. On correlation-immune functions. *Lecture Notes in Computer Science* **576** (1992), 86–100 (CRYPTO '91).

9. L. Carlitz and S. Uchiyama. Bounds on exponential sums. *Duke Math. Journal*, (1957), 37–41.
10. B. Chor, O. Goldreich, J. Hastad, J. Friedman, S Rudich and R. Smolensky. The bit extraction problem or t-resilient functions. *26th IEEE symposium on Foundations of Computer Science*, pages 396–407, 1985.
11. J. Friedman. On the bit extraction problem. *33rd IEEE symposium on Foundations of Computer Science*, pages 314–319, 1992.
12. T. Helleseth and T. Johansson. Universal hash functions from exponential sums over finite fields and Galois rings. *Lecture Notes in Computer Science* **1109** (1996), 31–44 (CRYPTO '96).
13. H. Krawczyk. New hash functions for message authentication. *Lecture Notes in Computer Science* **921** (1995), 301–310 (EUROCRYPT '95).
14. F. J. MacWilliams and N. J. A. Sloane. *The Theory of Error-Correcting Codes*. North-Holland, 1977.
15. J. L. Massey. Cryptography – A selective survey. *Digital Communications*, North-Holland (1986), 3–21.
16. U. M. Maurer and J. L. Massey. Perfect local randomness in pseudo-random sequences. *Lecture Notes in Computer Science* **435** (1990), 100–112 (CRYPTO '89).
17. J. Naor and M. Naor. Small bias probability spaces: efficient constructions and applications. *SIAM Journal on Computing* **22** (1993), 838–856.
18. H. Niederreiter and C. P. Schnorr. Local randomness in polynomial random number and random function generators. *SIAM Journal on Computing* **22** (1993), 684–694.
19. T. Siegenthaler. Correlation-immunity of nonlinear combining functions for cryptographic applications. *IEEE Trans. Inform. Theory* **30** (1984), 776–780.
20. C. P. Schnorr. On the construction of random number generators and random function generators. *Lecture Notes in Computer Science* **330** (1988), 225–232 (EUROCRYPT '88).
21. G.J. Simmons. A game theory model of digital message authentication. *Congressus Numeratium* **34** (1982), 413–424.
22. G.J. Simmons. Authentication theory/coding theory, *Lecture Notes in Computer Science.* **196** (1985), 411–431 (CRYPTO '84).
23. D. R. Stinson. Universal hashing and authentication codes. *Lecture Notes in Computer Science* **576** (1992), 74–85 (CRYPTO '91).
24. D. R. Stinson. Resilient functions and large set of orthogonal arrays. *Congressus Numerantium* **92** (1993), 105–110.
25. D .R. Stinson and J. L. Massey. An infinite class of counterexamples to a conjecture concerning nonlinear resilient functions. *Journal of Cryptology* **8** (1995), 167–173.
26. M. N. Wegman and J. L. Carter. New hash functions and their use in authentication and set equality. *Journal of Computer and System Sciences* **22** (1981), 265–279.