Generalized hashing and applications to digital fingerprinting

Noga Alon[*], Gérard Cohen[†], Michael Krivelevich[‡] and Simon Litsyn[§]

Abstract

Let C be a code of length n over an alphabet of q letters. An n-word y is called a descendant of a set of t codewords $x^1, \ldots, x^t$ if $y_i \in \{x_i^1, \ldots, x_i^t\}$ for all $i = 1, \ldots, n$. A code is said to have the t-identifying parent property if for any n-word that is a descendant of at most t parents it is possible to identify at least one of them. We study a generalization of hashing, (t, u)-hashing, which ensures identification, and provide tight estimates of the rates.

Keywords: error-correcting codes, identifying parent property, generalized hashing.

1 Background

Consider the distribution of digital content to subscribers over a broadcast channel. Each authorized user is given a decoder (could be a smartcard) with a secret decryption key. The distributor broadcasts an encrypted version of the content, which is decrypted by the authorized users. The scope of applications encompasses pay-per-view television, e-commerce, any broadcasting system to subscribers (see [2]), as well as some watermarking or fingerprinting questions.

We search for codes such that pooling t legal deciphering keys (t codewords) does not allow for creating an illegal hybrid deciphering key whose origin (the t codewords) would not be partially identifiable by the distributor.

Let us illustrate the problem on the binary alphabet $Q = \{0, 1\}$.

Suppose a *Distributor* wishes to create and distribute a large number of copies of a large file Φ of length N. In order to trace illegal copies he will *mark* each copy of Φ. The marking process consists of changing the bits of Φ belonging to some subset of a privileged set $M \subset \{1, \ldots N\}$ of coordinates called *marks*. The subset of marks associated to a copy of Φ is called a *fingerprint* and can be seen as a binary vector of length $m = |M|$. The set of marks M is supposed to be unknown to anyone but the distributor. Furthermore, the set of marks is usually supposed to be a small subset of $\{1, \ldots N\}$, so that modifying a fingerprint by randomly changing bits of a copy of Φ implies changing many bits of the original file and damaging the data significantly.

[*]Dept. of Mathematics, Tel Aviv University, Tel Aviv, Israel.
[†]ENST, 46 rue Barrault, 75013, Paris, France.
[‡]Dept. of Mathematics, Tel Aviv University, Tel Aviv, Israel.
[§]Dept. of Elect. Eng.- Systems, Tel Aviv University, Tel Aviv, Israel.

The problem of *collusion* occurs when a coalition of t pirate users compare their decoders: whenever they differ on some coordinate they will know it is a mark. They can then produce an illegal decoder changing at will bits on the subset of marks they have found out.

2 Introduction

Let Q be an alphabet of size q, and let us call any subset C of Q^n an (n, M)-*code* when $|C| = M$. Elements $x = (x_1, \dots, x_n)$ of C will be called *codewords*.

Let C be an (n, M)-code. Suppose $X \subseteq C$. For any coordinate i define the *projection*

$$P_i(X) = \bigcup_{x \in X} x_i.$$

Define the *envelope* $e(X)$ of X by:

$$e(X) = \{x \in Q^n : \forall i, x_i \in P_i(X)\}.$$

Elements of the envelope $e(X)$ will be called *descendants* of X. Observe that $X \subseteq e(X)$ for all X, and $e(X) = X$ if $|X| = 1$.

Given a word $s \in Q^n$ (a son) which is a descendant of X we would like to identify without ambiguity at least one member of X (a parent). From [1], we have the following definition, a generalization of the case $t = 2$ from [5].

Definition 1 *For any $s \in Q^n$ let $\mathcal{H}_t(s)$ be the set of subsets $X \subset C$ of size at most t such that $s \in e(X)$. We shall say that C has the* identifiable parent property of order t *(or is a t-identifying code, or is t i.p.p. for short) if for any $s \in Q^n$, either $\mathcal{H}_t(s) = \emptyset$ or*

$$\bigcap_{X \in \mathcal{H}_t(s)} X \neq \emptyset.$$

It is convenient to view $\mathcal{H}_t(s)$ as the set of edges of a hypergraph. Its vertices are codewords of C.

The concept of t-identification originates with the work of Chor, Fiat and Naor on broadcast encryption [4]. It is also related to the problem of fingerprinting numerical data [3].

It is not difficult to prove that if the minimum Hamming distance of C is big enough, then C must be t-identifying: we have [4]:

Proposition 1 *If C has minimum Hamming distance d satisfying*

$$d > (1 - 1/t^2)n,$$

then C is a t-identifying code.

As usual, let $R = R(C) = \log_q M/n$ denote the rate of the (n, M)-code C. Let $R_q(t) = \liminf_{n\to\infty} \max R(C_n)$, where the maximum is computed over all t-identifying codes C_n of length n.

In [1], the following is proved:

Theorem 1 *$R_q(t) > 0$ if and only if $t \leq q - 1$.*

Recall that a subset C of Q^n is said to be t-*hashing* (or t-separating, see, e.g. [6]) if any t of its members have t distinct entries in some common coordinate $i \in \{1, \dots, n\}$.

In the next section, we recall an extension of hashing and a few results from [1].

3 Partially hashing families

Definition 2 *Let us say that a subset $C \subset Q^n$ is (t,u)* partially hashing *if for any two subsets T, U of C such that $T \subset U \subset C$, $|T| = t$, $|U| = u$, there is some coordinate $i \in \{1, \dots, n\}$ such that for any $x \in T$ and any $y \in U, y \neq x$, we have $x_i \neq y_i$.*

The concept of (t,u)-hashing is easily seen to generalize the well known notion of hashing. Indeed, when $u = t+1$, a (y,u)-partially hashing family is $(t+1)$-hashing.

Barg et al. proved in [1] that the property of (t,u) partial hashing can be used to ensure the t-IPP property, and obtained a lower bound of the rate of (t,u)-hashing families. Their results are summarized below.

Lemma 1 *Let $u \geq t+1$ and $\varepsilon > 0$: infinite sequences of (t,u) partially hashing codes exist for all rates R such that*

$$R + \varepsilon \leq \frac{1}{u-1} \log_q \frac{(q-t)!q^u}{(q-t)!q^u - q!(q-t)^{u-t}}.$$

Lemma 2 *Let $u = \lfloor (t/2+1)^2) \rfloor$. If C is (t,u) partially hashing then C is a t-identifying code.*

Theorem 2 *Let $u = \lfloor (t/2+1)^2) \rfloor$. We have*

$$R_q(t) \geq \frac{1}{u-1} \log_q \frac{(q-t)!q^u}{(q-t)!q^u - q!(q-t)^{u-t}}.$$

4 New bounds for (t,u)-hashing

In this section we present new bounds on the rate of (t,u) partially hashing families and indicate how they can be proved. For simplicity we consider here only the case of the smallest possible alphabet $q = t+1$. We denote $Q = \{0, \dots, t\}$.

Two families $A \subset B \subseteq Q^n$ are called *separated* if there exists a coordinate i, $1 \leq i \leq n$, so that for every $a \in A$ and every $b \in B - a$ one has $a_i \neq b_i$. Then such a coordinate i is called *separating*.

Theorem 3 *Let $u \geq t+1$, $q = t+1$ and $\varepsilon > 0$. Infinite sequences of (t,u) partially hashing codes exist for all rates R such that*

$$R + \varepsilon \leq \frac{t!(u-t)^{u-t}}{u^u(u-1)\ln(t+1)} .$$

Proof. (Outline) We will apply the probabilistic method with expurgation to (t,u)-hashing codes. Choose $2m$ vectors in Q^n independently with repetitions, where each vector c is generated according to the following distribution: for each coordinate $1 \leq i \leq n$, $Pr[c_i = 0] = (u-t)/t$, and $Pr[c_i = j] = 1/u$ for $j = 1, \dots, t$. The value of m will be chosen later. Denote the obtained random family by C_0. Now estimate the expected number of non-separated pairs $T \subset U \subset C_0$, where $|T| = t$, $|U| = u$. The probability that a coordinate i separates $T = \{a^1, \dots, a^t\}$ and $U = T \cup \{b^1, \dots, b^{u-t}\}$ is at least as large as the probability that all a_i^k

are different and are different from 0, and $b_i^l = 0$, $l = 1, \ldots, u-t$. The latter probability is exactly $t!\left(\frac{1}{u}\right)^t\left(\frac{u-t}{u}\right)^{u-t} = \frac{t!(u-t)^{u-t}}{u^u}$. As all coordinates behave independently we get

$$Pr[T, U \text{ are not separated}] \leq \left(1 - \frac{t!(u-t)^{u-t}}{u^u}\right)^n \ .$$

Hence the expected number of non-separated pairs A, B in C_0 is at most $\binom{2m}{u}\binom{u}{t}$ times the above expression. We obtain that if

$$\binom{2m}{u}\binom{u}{t}\left(1 - \frac{t!(u-t)^{u-t}}{u^u}\right)^n \leq m\,, \tag{1}$$

then there exists a code $C_0 \subset Q^n$ of cardinality $|C_0| = 2m$ with at most m non-separated pairs $T \subset U \subset C_0$, $|T| = t$, $|U| = u$. Fix such a code and for each non-separated pair (T, U) delete one vector from T. Denote the resulting code by C. Then C is (t, u) partially hashing and $|C| \geq m$. We infer that for every m satisfying (1), there exists a (t, u)-separating code $C \subset Q^n$ of cardinality m. Solving (1) for m gives the desired bound. □

Corollary 1 *Let $u = \lfloor (t/2+1)^2 \rfloor$. Then*

$$R_{t+1}(t) \geq \frac{t!(u-t)^{u-t}}{u^u(u-1)\ln(t+1)} \ .$$

Theorem 4 *Let $C \subset \{0, \ldots, t\}^n$ be a (t, u) partially hashing code. Then*

$$\frac{1}{n}\log_{t+1}|C| \leq \frac{\ln 3(t+1)!(u-t-1)^{u-t-1}}{2(u-2)^{u-2}} + o(1) \ .$$

Proof. (Outline) The argument here borrows some ideas from the proof of Nilli [7] for the upper bound for hashing. We first prove the following claim.

Claim 1 *If C contains subsets $T_0 \subset U_0$ of cardinalities $|T_0| = t-1$, $|U_0| = u-2$, respectively, such that (T_0, U_0) has at most μ separating coordinates, then $|C| - u + 2 \leq 3^\mu$.*

Claim proof. Fix such T_0, U_0 and assume to the contrary that $|C| - u + 2 > 3^\mu$. Let $I \subset [n]$ be the set of coordinates separating T_0 and U_0. Then $|I| \leq \mu$. For each $i \in I$ set $Q_i = \{a_i : a \in T_0\}$. Obviously, $|Q_i| = t-1$. By the pigeon hole principle it follows that the set $C \setminus U_0$ contains two vectors c^1, c^2 so that for every $i \in I$, $c_i^1 = c_i^2$ or $c_i^1, c_i^2 \in Q_i$. Define $T = T_0 + c^1$, $U = U_0 + \{c^1, c^2\}$. We claim that the pair (T, U) violates the condition of (t, u)-hashing. Indeed, if a coordinate i separates T and U then it already separates T_0 and U_0 and thus $i \in I$. But then, if $c_i^1 = c_i^2$, then $c^1 \in T$, $C^2 \in U \setminus T$ and therefore i does not separate T and U. In the second case $c_i^1 \in Q_i$, and hence $c^1 \in T$ and c_i^1 coincides with a_i for some $a \in T_0$. The obtained contradiction establishes the result. □

Returning to the theorem proof we now show that there exists a pair (T_0, U_0) as in the above claim with few separating coordinates. To this end, we choose T_0 and U_0 at random and estimate from above the expected number of coordinates separating T_0 and U_0. Fix a coordinate i and for all $0 \leq j \leq t$ denote $p_j = \frac{|\{c \in C : c_i = j\}|}{|C|}$, i.e., p_j is the frequency of symbol j in coordinate i. Then

$$Pr[i \text{ separates } T_0 \text{ and } U_0] = \sum_{I \subset Q, |I| = t-1} (t-1)! \prod_{j \in I} p_j (1 - \sum_{j \in I} p_j)^{u-t-1} \ .$$

One can show that for a fixed $I \subset Q$, $|I| = t-1$, $\prod_{j\in I} p_j (1-\sum_{j\in I} p_j)^{u-t-1} \le \frac{(u-t-1)^{u-t-1}}{(u-2)^{u-2}}$. Hence the probability that i is separating is at most

$$\binom{t+1}{t-1}(t-1)!\frac{(u-t-1)^{u-t-1}}{(u-2)^{u-2}} = \frac{(t+1)!}{2}\frac{(u-t-1)^{u-t-1}}{(u-2)^{u-2}} .$$

By linearity of expectation there exists a pair (T_0, U_0) with $T_0 \subset U_0 \subset C$, $|T_0| = t-1$, $|U_0| = u-2$, and with at most $\mu = \frac{(t+1)!}{2}\frac{(u-t-1)^{u-t-1}}{(u-2)^{u-2}} n$ separating coordinates. Plugging this estimate into Claim 1 gives the required upper bound on C. $\square$

It is instructive to compare the lower and the upper bounds for (t,u)-hashing families given by Theorems 3 and 4. One can easily see that for large t, both bounds on the rate are exponentially small in t, while their ratio is $O(1)tu^3/(u-t)$ and thus is only polynomial in case u is polynomial in t (as happens for example when applying (t,u) partial hashing families for constructing codes with the identifying parent property, see Lemma 2). Thus to a certain extent we can claim that the obtained bounds for (t,u)-hashing match each other.

Comparing the lower bounds of Lemma 1 and Theorem 3, one can easily show that in case u is quadratic in t the bound of Theorem 3 is exponentially better than that of Lemma 1.

References

[1] A. Barg, G. Cohen, S. Encheva, G. Kabatiansky and G. Zémor, "A hypergraph approach to the identifying parent property", *SIAM J. Disc. Math.*, to appear.

[2] D. Boneh and M. Franklin, "An efficient public-key traitor-tracing scheme", *LNCS Crypto'99*

[3] D. Boneh and J. Shaw, "Collusion-secure fingerprinting for digital data", *IEEE Trans. on Inf. Theory*, **44** (1998), pp. 480–491.

[4] B. Chor, A. Fiat and M. Naor, "Tracing traitors", *Crypto'94* LNCS 839 (1994), pp. 257–270.

[5] H. D. L. Hollmann, J. H. van Lint, J.-P. Linnartz and L. M. G. M. Tolhuizen, "On codes with the identifiable parent property", *J. Combinatorial Theory*, Series A, **82** (1998) pp. 121–133.

[6] J. Körner and A. Orlitski, "Zero-error information theory," *IEEE Trans. Information Theory*, **44** (1998), pp. 2207–2229.

[7] A. Nilli, "Perfect hashing and probability", *Combinatorics, Probability and Computing*, **3** 1994, pp. 407–409.