Generalized hashing and parent-identifying codes

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

Abstract

Let C be a code of length n over an alphabet of q letters. For a pair of integers $2 \leq t < u$, C is (t,u)-hashing if for any two subsets $T, U \subset C$, satisfying $T \subset U$, $|T| = t$, $|U| = u$, there is a coordinate $1 \leq i \leq n$ such that for any $x \in T$, $y \in U - x$, x and y differ in the i-th coordinate. This definition, generalizing the standard notion of a t-hashing family, is motivated by an application in designing the so-called parent identifying codes, used in digital fingerprinting. In this paper we provide lower and upper bounds on the best possible rate of (t,u)-hashing families for fixed t, u and growing n. We also describe an explicit construction of (t,u)-hashing families.

1 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 $x = (x_1, \ldots, x_n)$ of C will be called *codewords*. As usual, let $R = R(C) = \log_q M/n$ denote the rate of C.

For a parameter $t \geq 2$ a code C is called t-*hashing* if for any t distinct codewords $x^1, \ldots, x^t \in C$ there is a coordinate $1 \leq i \leq n$ such that all values x_i^j, $1 \leq j \leq t$ are distinct. The concept of a hashing family is certainly between the most central in Computer Science and Coding Theory, and its numerous applications have been described in the literature, see, e.g., [10] and its references. An obvious necessary condition for the existence of a t-hashing family of positive rate is $q \geq t$, and indeed large hashing codes are known to exist for this range of parameters (see [9], [13], [14], [16] for bounds on the rate of t-hashing families of growing length).

In this note we consider a different notion of hashing.

Definition 1 *Let $2 \leq t < u$ be integers. A subset $C \subset Q^n$ is* (t,u)-hashing *if for any two subsets T, U of C such that $T \subset U$, $|T| = t$, $|U| = u$, there is some coordinate $i \in \{1, \ldots, n\}$ such that for any $x \in T$ and any $y \in U, y \neq x$, we have $x_i \neq y_i$.*

[*] Mathematics and Computer Science, Tel Aviv University, Tel Aviv 69978, Israel. E-mail: nogaa@tau.ac.il. Research supported in part by a USA Israeli BSF grant, by a grant from the Israel Science Foundation, and by the Hermann Minkowski Minerva Center for Geometry at Tel Aviv University.

[†] ENST, 46 rue Barrault, 75013, Paris, France. E-mail: cohen@inf.enst.fr.

[‡] Department of Mathematics, Tel Aviv University, Tel Aviv 69978, Israel. E-mail: krivelev@post.tau.ac.il. Research supported in part by a USA-Israel BSF grant, by a grant from the Israel Science Foundation and by a Bergmann Memorial Award.

[§] Dept. of Elect. Eng.- Systems, Tel Aviv University, Tel Aviv, Israel. E-mail: litsyn@eng.tau.ac.il. Research supported in part by a USA-Israel BSF grant and by a grant from the Israel Science Foundation.

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

As it turns out the somewhat artificially looking class of (t,u)-hashing codes is exactly what is required to show the existence of high rate codes in a relatively well studied class of codes, called parent-identifying codes. This connection together with known bounds on the rate of parent-identifying codes is described in the the next section.

2 Parent identifying codes

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 [4], we have the following definition, a generalization of the case $t = 2$ from [11].

Definition 2 *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 has the t-IPP, for short) if for any $s \in Q^n$, either $\mathcal{H}_t(s) = \emptyset$ or*

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

The study of parent identifying codes is motivated by its connection to digital fingerprinting and schemes against software piracy, see, e.g., [8], [7], [17]. Currently there are already several papers discussing bounds on the size/rate of parent identifying codes. The case of a fixed length and large alphabet size has been considered in [11], [2], [3], [5], while the case of a fixed size alphabet and growing length has been treated in [8], [4]. Here we will be concerned with the latter case.

It is not difficult to prove that if the minimum Hamming distance of C is large enough, then C must be t-identifying: we have [8]:

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

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

then C is a t-identifying code.

In fact, the condition $d/n > 1 - 1/t^2$ guarantees a stronger property: *t-traceability* ([8]), namely, that all closest codewords to the produced descendant are part of the coalition producing it. It thus insures the t-IPP, with the extra feature of a search algorithm linear in $|C|$.

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 [4], the following is proved:

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

Barg et al. discovered in [4] a connection between (t,u)-hashing and t-IPP. Specifically, they proved the following:

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

They also obtained a lower bound on the rate of (t,u)-hashing families:

Lemma 2 *Let $u \geq t+1$ and $\varepsilon > 0$. Infinite sequences of (t,u)-hashing codes exist for all rates R such that*

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

By combining Lemmas 1 and 2 one gets the following lower bound on the rate of t-identifying codes:

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

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

Our main result here is an improvement of the bounds in Lemma 2 and in Theorem 2. We also obtain an explicit construction of high rate (t,u)-hashing families, based on some known explicit constructions of codes.

3 New bounds for (t,u)-hashing

In this section we present new bounds on the rate of (t,u)-hashing families. For simplicity we consider here only the case of the smallest possible alphabet $q = t+1$. We denote $Q = \{0,\ldots,t\}$.

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)-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. 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 $1 \leq i \leq n$, $Pr[c_i = 0] = (u-t)/u$, and $Pr[c_i = j] = 1/u$ for $j = 1,\ldots,t$. 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 $T = \{a^1, \ldots, a^t\}$ and $U = T \cup \{b^1, \ldots, b^{u-t}\}$ 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)-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. It now remains to solve (1) for m. Observe that

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

and thus is order to satisfy (1) it is enough to require:

$$2^u u^t m^u e^{-\frac{t!(u-t)^{u-t}}{u^u}n} \leq m\,,$$

or

$$m \leq \left(\frac{1}{2^u u^t}\right)^{\frac{1}{u-1}} e^{\frac{t!(u-t)^{u-t}}{u^u(u-1)}n} .$$

It follows that there exists a (t, u)-hashing family of rate

$$R = \frac{1}{n}\log_{t+1} m = \frac{1}{n}\frac{\ln m}{\ln(t+1)} = \frac{t!(u-t)^{u-t}}{u^u(u-1)\ln(t+1)} - o(1)\,,$$

as claimed. □

Recalling now Lemma 1 and performing simple asymptotic manipulations we get the following asymptotic lower bound on the rate of t-identifying codes:

Corollary 1 *There exists an absolute constant $c > 0$ such that:*

$$R_{t+1}(t) \geq \frac{ct!2^{2t}}{t^2(et^2)^t} = t^{-t(1+o(1))}.$$

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

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

Proof. The argument here borrows some ideas from the proof of Nilli [16] 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, $|Q_i| = t-1$ and thus $|Q \setminus Q_i| = 2$. By the pigeonhole principle it follows that the set $C \setminus U_0$ contains two distinct vectors c^1, c^2 so that for every $i \in I$, $c^1_i = c^2_i \in Q \setminus Q_i$ or $c^1_i, c^2_i \in Q_i$. Define $T = T_0 \cup \{c^1\}$, $U = U_0 \cup \{c^1, c^2\}$. 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 $c^1_i = c^2_i$, then, as $c^1 \in T$ and $c^2 \in U \setminus T$, i does not separate T and U. In the second case $c^1_i \in Q_i$, and hence $c^1 \in T$ and c^1_i coincides with a_i for some $a \in T_0$. The obtained contradiction establishes the result. □

Returning to the proof of the theorem, 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 (with repetitions) 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

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

By the arithmetic-geometric means inequality, for a fixed $I \subset Q$, $|I| = t-1$,

$$\left(\prod_{j \in I} (u-t-1) p_j) \cdot (1 - \sum_{j \in I} p_j)^{u-t-1} \right)^{\frac{1}{u-2}} \leq \frac{(u-t-1) \sum_{j \in I} p_j + (u-t-1)(1 - \sum_{j \in I} p_j)}{u-2},$$

implying that $\prod_{j \in I} p_j (1 - \sum_{j \in I} p_j)^{u-t-1} \leq \frac{(u-t-1)^{u-t-1}}{(u-2)^{u-2}}$. 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. □

It is instructive to compare the upper and the lower bounds for (t, u)-hashing families given by Theorems 4 and 3, respectively. One can easily see that for large t, both bounds on the rate are exponentially small in t, while their ratio is (up to negligible terms)

$$\begin{aligned} & \frac{\ln 3 (t+1)! (u-t-1)^{u-t-1}}{2 \ln(t+1) (u-2)^{u-2}} \cdot \left(\frac{t! (u-t)^{u-t}}{u^u (u-1) \ln(t+1)} \right)^{-1} \\ = \; & \frac{\ln 3}{2} \cdot (t+1) \cdot \frac{(u-t-1)^{u-t-1}}{(u-t)^{u-t}} \cdot \frac{u^u (u-1)}{(u-2)^{u-2}} \leq O(1) \frac{t}{u-t} \cdot u^3 \left(\frac{u-1}{u-2} \right)^{u-2} \\ = \; & O(1) \frac{t u^3}{u-t}, \end{aligned}$$

and thus is only polynomial in case u is polynomial in t (as happens for example when applying (t,u)-hashing families for constructing codes with the identifying parent property, see Lemma 1). Thus, the obtained bounds for (t,u)-hashing nearly match each other.

Comparing the lower bounds of Lemma 2 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 2. For t-identifying codes over an alphabet of $t+1$ elements we get here that the best possible rate is $t^{-t(1+o(1))}$, whereas the lower bound that follows from Lemma 2 is only $t^{-\Theta(t^2)}$.

4 Explicit constructions

Concatenation, see e.g. [15], is a powerful method to construct infinite families of codes with a required property by combining a "seed" code with the property over a small alphabet, together with an appropriate code over a larger alphabet (whose size is the size of the seed).

Let C_1 be an (N,M) code over $\mathcal{U}$, where $|\mathcal{U}| = 4ut$; let C_2 (the seed) be an $(n, 4ut)$ code over $\mathcal{Q}$, $|\mathcal{Q}| = t+1$. We fix a bijection $\phi : \mathcal{U} \to C_2$.

Denoting by $C_1 \star C_2$ the concatenation of C_1 and C_2, obtained by replacing in codewords of C_1 every occurrence of a symbol $a \in \mathcal{U}$ by its image $\phi(a)$ in C_2, we have the following result:

Proposition 2 *If C_1 is (t,u)-hashing of rate R_1 and C_2 is (t,u)-hashing of rate R_2, then $C_1 \star C_2$ is a (t,u)-hashing (Nn, M) code of rate R_1R_2 over $\mathcal{Q}$.*

Proof. The fact that the rate of the concatenation is the product of the rates of concatenated codes is standard and easy to verify. It thus remains to check that the concatenation $C_1 \star C_2$ is indeed (t,u)-hashing. Let $U \subset C_1 \star C_2$, $|U| = u$, $T \subset U$, $|T| = t$. Let $a^1, \ldots, a^t$ be the codewords of C_1 corresponding to those of T, and let $b^1, \ldots, b^{u-t}$ be the codewords of C_1 corresponding to $U \backslash T$. As C_1 is (t,u)-hashing there exists a coordinate $1 \le i \le N$ in which all symbols $a_i^1, \ldots, a_i^t \in \mathcal{U}$ are distinct and disjoint from the set $\{b_i^1, \ldots, b_i^{u-t}\} \subset \mathcal{U}$ of cardinality at most $u-t$. As C_2 is (t,u)-hashing as well, there is a coordinate $1 \le j \le n$, where all symbols $(\phi(a_i^1))_j, \ldots, (\phi(a_i^t))_j$ are distinct and disjoint from the set $\{(\phi(b_i^1))_j, \ldots, (\phi(b_i^{u-t}))_j\}$. Hence in coordinate $n(i-1)+j$ all codewords of T are pairwise distinct and disjoint from those of $U \setminus T$. □

It is easy to check that if the minimum distance of a code of length N is at least $(1-\frac{1}{ut})N$, then this code is a (t,u)-hashing. Indeed, for any sets $T \subset U$ of codewords, $|T| = t$, $|U| = u$, there are at most

$$\left(\binom{|T|}{2} + |T|(|U| - |T|)\right) \frac{1}{ut} N < N$$

coordinates in which some member of T coincides with another member of U. Thus, there is a separating coordinate, as needed. In [1] the authors describe an explicit construction of codes of length N over an alphabet of size g with minimum distance $(1-\delta)N$ and rate at least

$$Max_{\delta \le \mu \le 1-1/g} c(1 - H_g(\mu))(1 - \frac{\delta}{\mu})$$

where c is an absolute positive constant, and $H_g(x) = -x \log_g x - (1-x) \log_g(1-x) + x \log_g(g-1)$. Taking $g = 4ut$, $\delta = 1 - \frac{4}{g} = 1 - \frac{1}{ut}$ we obtain the following, by substituting $\mu = 1 - \frac{2}{g}$ in the above estimate:

Proposition 3 *There exists an explicit (t, u)-hashing family over an alphabet of size $4ut$ of rate $\Theta(\frac{1}{u^2t^2\log(4ut)})$.*

For the seed C_2 now, we use the following general trivial construction:

Proposition 4 *For every t, there exists a (t, u)-hashing $(\binom{4ut}{t}, 4ut)$ code over Q.*

Proof. Write as columns all binary vectors of weight t and length $4ut$; then, in every column, replace the t ones by all the non-zero elements of Q. The $4ut$ rows thus obtained are the required codewords. □

Putting the above three propositions together we get the following result.

Theorem 5 *There exists an absolute constant $c > 0$ such that for all $2 \leq t < u$ there is an explicit construction of a (t, u)-hashing code over an alphabet of size $t + 1$ of rate $R \geq u^{-ct}$.*

Corollary 2 *For t large enough there is a constructive infinite sequence of t-identifying codes over an alphabet of size $t + 1$ of rate $t^{-O(t)}$.*

Proof. Choose $u = \lfloor (t/2 + 1)^2 \rfloor$ and apply Lemma 1 and Theorem 5. □

5 Concluding remarks

The construction in Proposition 3 can be performed using other known explicit codes, and in particular using the celebrated algebraic geometry codes described in [18], [12] (which supply a similar estimate).

We have mostly considered here the minimal possible alphabet size ($q = t + 1$). On the other hand, for large q, the asymptotic rates are known: the non-constructive approach yields the lower bound $R \geq (1 + o(1))/(u - 1)$ for the rates of both (t, u) and u-hashing families.

This coincides with the upper bound for codes with the t-IPP proved in [5] and in [3].

References

[1] N. Alon, J. Bruck, J. Naor, M. Naor and R. Roth, “Construction of asymptotically good, low-rate error-correcting codes through pseudo-random graphs”, IEEE Transactions on Information Theory, 38 (1992), 509-516.

[2] N. Alon, E. Fischer and M. Szegedy, “Parent-identifying codes”, *J. Combin. Theory Ser. A* **95** 2001, pp. 349–359.

[3] N. Alon and U. Stav, “New bounds on parent-identifying codes: the case of multiple parents”, submitted.

[4] A. Barg, G. Cohen, S. Encheva, G. Kabatiansky and G. Zémor, “A hypergraph approach to the identifying parent property”, *SIAM J. Disc. Math.*, **14** 2001, pp. 423-432.

[5] S. Blackburn, “An upper bound on the size of a code with the k-identifiable parent property”, *J. Combin. Theory Ser. A*, to appear.

[6] D. Boneh and M. Franklin, "An efficient public-key traitor-tracing scheme", *Crypto'99*, LNCS 1666 (1999), pp. 338–353.

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

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

[9] M. Fredman and J. Komlós, "On the size of separating systems and perfect hash functions", *SIAM J. Algebraic and Disc. Meth*, **5** (1983), pp. 61–68.

[10] T. H. Cormen, C. E. Leiserson, and R. L. Rivest, Introduction to Algorithms, MIT Press, 1990, Chapter 12.

[11] 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. Combin. Theory Ser. A*, **82** 1998, pp. 121–133.

[12] G. L. Katsman, M. A. Tsfasman and S. G. Vlăduţ, "Modular curves and codes with a polynomial construction", *IEEE Trans. Inform. Theory,* IT-30 (1984), 353–355.

[13] J. Körner, "Fredman-Komlós bounds and information theory", *SIAM J. Algebraic and Disc. Methods*, **7** 1986, pp. 560–570.

[14] J. Körner and K. Marton, "New bounds for perfect hashing via information theory", *Europ. J. Combinatorics*, **9** 1988, pp. 523–530.

[15] F. J. MacWilliams and N. J. A. Sloane, *The Theory of Error-Correcting Codes*, North-Holland, Amsterdam, (1977).

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

[17] J. N. Staddon, D. R. Stinson and R. Wei, "Combinatorial properties of frameproof and traceability codes", *IEEE Trans. Information Theory*, **47** 2001, pp. 1042–1049.

[18] M. A. Tsfasman, S. G. Vlăduţ and Th. Zink, "Modular curves, Shimura curves, and Goppa codes, better than Varshamov-Gilbert bound", *Math. Nachr.,* 109 (1982), 21–28.