ROBUST PARENT-IDENTIFYING CODES AND COMBINATORIAL ARRAYS

ALEXANDER BARG* AND GRIGORY KABATIANSKY†

ABSTRACT. An n-word y over a finite alphabet of cardinality q is called a descendant of a set of t words $x^1, \dots, x^t$ if $y_i \in \{x_i^1, \dots, x_i^t\}$ for all $i = 1, \dots, n$. A code $\mathcal{C} = \{x^1, \dots, x^M\}$ is said to have the t-IPP property if for any n-word y that is a descendant of at most t parents belonging to the code it is possible to identify at least one of them. From earlier works it is known that t-IPP codes of positive rate exist if and only if $t \le q - 1$.

We introduce a robust version of IPP codes which allows unconditional identification of parents even if some of the coordinates in y can break away from the descent rule, i.e., can take arbitrary values from the alphabet, or become completely unreadable. We show existence of robust t-IPP codes for all $t \le q - 1$ and some positive proportion of such coordinates. The proofs involve relations between IPP codes and combinatorial arrays with separating properties such as perfect hash functions and hash codes, partially hashing families and separating codes.

For $t = 2$ we find the exact proportion of mutant coordinates (for several error scenarios) that permits unconditional identification of parents.

1. IPP AND FINGERPRINTING CODES

1-A. **Introduction.** Codes with the identifiable parent property (IPP codes) were introduced by Hollmann et al. [23]. They are used in the design of traitor-tracing schemes [16, 29, 9, 14] and digital fingerprinting codes [15, 5, 32, 4, 3]. Some of our terminology is motivated by these applications.

Let $\mathcal{Q}^n$ be the set of all n-words (vectors) over a finite alphabet $\mathcal{Q}$ of size q. A subset $\mathcal{C}$ of $\mathcal{Q}^n$ is called a code of length n. A t-subset $U = \{u^1, \dots, u^t\} \subset \mathcal{C}$ is called a *coalition* of size t. In applications, the elements of $\mathcal{C}$ serve as fingerprints of the users of the system. A *collusion attack* occurs when several uses (pirates) form a coalition U to create a new fingerprint y with the purpose of making it impossible to identify any members of U based on observing y. The vector y is formed as a function of U (the attack map). The general problem considered in the paper is design of codes resilient to collusion attacks. To give a formal definition of the attack, we need to introduce several concepts.

Let $U_i = \{u_i^1, \dots, u_i^t\}$ be the set of the ith coordinates of the elements of U. Coordinate $i, 1 \le i \le n$ is called *undetectable* for U if all vectors in U have the same value in it, i.e., if $|U_i| = 1$, and is called *detectable* otherwise. Denote by $D(U)$ the set of detectable coordinates for the coalition U.

Let $U \subset \mathcal{C}$ be a coalition. Suppose that $y_i \in U_i$ for all $i \in [n]$. Under this restriction the set of possible attack vectors for U forms the subset

$$\langle U \rangle = \{(y_1, \dots, y_n) \in \mathcal{Q}^n : y_i \in U_i,\ i \in [n]\} \tag{1}$$

Date: May 8, 2011.

*Department of ECE and Institute for Systems Research, University of Maryland, College Park, MD 20742, USA and Institute for Information Transmission Problems, Russian Academy of Sciences, Moscow, Russia. Email: abarg@umd.edu. Supported in part by NSF grants CCF0635271, CCF0830699, CCF0916919, and DMS0807411.

†Dobrushin Mathematical Laboratory, Institute for Information Transmission Problems, Russian Academy of Sciences, Moscow, Russia. Email: kaba@iitp.ru. Supported in part by RFFI grants 08-07-92495, 09-01-00536 and 09-01-12171.

called the *narrow-sense envelope* of the coalition and denoted by $\langle U\rangle$. The elements of $\langle U\rangle$ are called *descendants* of U, and for any of the descendants $y \in \langle U\rangle$ the elements of U are called its *parents*. The t-*envelope* of the code $\mathcal{C}$ is defined as follows:

$$\langle \mathcal{C}\rangle_t = \cup_{U\subset\mathcal{C},\,|U|\le t}\langle U\rangle.$$

Definition 1. *(Hollmann et al.* [23]*) The code $\mathcal{C}$ has a* t-identifiable parent property *(is a t-IPP code) if for any $y \in \langle\mathcal{C}\rangle_t$ it is possible to find at least one of its parents, i.e. if*

$$\bigcap_{U\subset\mathcal{C},\,|U|\le t,\,y\in\langle U\rangle} U \neq \emptyset. \tag{2}$$

One of the main problems associated with IPP codes is bounding the cardinality of a t-IPP code $\mathcal{C}$. In this paper we are concerned with the case of large n and fixed q and t. Call the number $R = R(\mathcal{C}) \triangleq \log_q|\mathcal{C}|/n$ the *rate* of the code and let

$$R_q(n,t) = \max\{R(\mathcal{C}) : \mathcal{C}\subset\mathcal{Q}^n \text{ is } t\text{-IPP}\}$$
$$R_q(t) = \liminf_{n\to\infty} R_q(n,t).$$

We are interested in establishing conditions for the existence of sequences of codes of increasing length n and positive rate $R_q(t)$ (called good IPP codes below). It is easily seen that $R_q(t) = 0$ for $t \ge q$. Hollmann et al. [23] proved that $R_q(2) > 0$ for $q \ge 3$. More generally, [6] showed that $R_q(t) > 0$ for all $q \ge t+1$. These papers also provided characterizations of 2- and 3-IPP codes, respectively. An improved lower bound on $R_{t+1}(t)$ was given by Alon et al. in [1] while upper bounds on the cardinality of t-IPP codes were derived in [2, 10].

In a related independent work, Boneh and Shaw [15] introduced a broader class of attack maps. Define the *wide-sense envelope* of the coalition U as

$$\{(y_1,\dots,y_n)\in\mathcal{Q}^n\cup\{*\} :\ y_i = u_i, i\notin D(U)\} \tag{3}$$

(any symbols of $\mathcal{Q}$ or erased symbols * are allowed in the detectable coordinates). Codes that support reliable recovery of the pirates for this problem are called *collusion-secure* or *fingerprinting*. As shown in [15, 5], under this definition, unconditional recovery of pirates is impossible. Moreover, a single code cannot guarantee low error probability of identification, and it is necessary to use a family of codes parametrized by a random key. Such code families are somewhat loosely called fingerprinting codes. Constructions and bounds for fingerprinting codes were studied in [5, 32, 20, 4, 3, 19, 24].

A common feature of the two definitions is the following restriction:

(4) *Marking assumption:* In forming a collusion attack y the coalition cannot change the values of its undetectable coordinates.

The object of this paper is a study of an intermediate concept between the IPP and fingerprinting codes, namely, permitting a limited number of coordinates (mutations) in y that do not follow their parents in that they deviate from the descent rule (1) by breaking the marking assumption or using the wide-sense attack (3) or both.

1-B. **Robust IPP codes.** Call a coordinate i of y a *mutation* if $y_i \notin U_i$. Assume that the coalition U forms y following the IPP attack rule (1) except for εn coordinates that can deviate from this rule. Below we consider mutations of two types: erasures, when the value y_i is replaced by an unreadable mark, and arbitrary symbol $y_i \in \mathcal{Q}\backslash U_i$. We also assume that mutations occur either only in detectable coordinates or in all coordinates of y. Altogether this accounts for the following 4 types of attacks:

(I) Only detectable coordinates can mutate (the marking assumption is followed);
(II) Only detectable coordinates can mutate, and the mutant coordinates always become unreadable marks (erasures);
(III) Any coordinate can mutate to any letter of the alphabet.
(IV) Any coordinate can mutate, and the mutant coordinates always become erasures.

Let $U \subset \mathcal{Q}^n, |U| \leq t$ be a coalition. Denote by $\langle U \rangle_\varepsilon$ the set of all vectors y formed from the vectors in U so that $y_i \in U_i$ for $n(1-\varepsilon)$ coordinates i and y_i is a mutation in at most εn coordinates, formed using one of the rules (I)-(IV) above.

Definition 2. *$\mathcal{C} \subset \mathcal{Q}^n$ is a (t, ε)-IPP code (robust t-IPP code) if*

$$\bigcap_{U \subset \mathcal{C}, |U| \leq t, y \in \langle U \rangle_\varepsilon} U \neq \emptyset.$$

In words: the code $\mathcal{C}$ guarantees exact identification of at least one member of the pirate coalition of size at most t for any collusion attack with at most εn mutations

Below when the attack rule is not explicitly mentioned, we mean that the code has the IPP property for all the rules (I)-(IV).

Introduce the following quantities:

$$R_q(n, t, \varepsilon) = \max\{R(\mathcal{C}) : \mathcal{C} \subset \mathcal{Q}^n \text{ is } (t, \varepsilon)\text{-IPP}\}$$
$$R_q(t, \varepsilon) = \liminf_{n \to \infty} R_q(n, t, \varepsilon).$$

Let

$$\varepsilon_{\text{crit}} = \varepsilon_{\text{crit}}(q, t) := \sup(\varepsilon : R_q(t, \varepsilon) > 0) \tag{5}$$

be the critical value of ε. We use the notation $\varepsilon^D_{\text{crit}}, \varepsilon^{*,D}_{\text{crit}}, \varepsilon_{\text{crit}}, \varepsilon^*_{\text{crit}}$ to refer to the critical values for the attacks of type (I)-(IV), respectively. In the hierarchy of attacks that emerges, the second one is the weakest. Generally, the following inequalities are obvious:

$$\varepsilon_{\text{crit}}(q, t) \leq \varepsilon^*_{\text{crit}}(q, t) \leq \varepsilon^{*,D}_{\text{crit}}(q, t) \tag{6}$$

$$\varepsilon_{\text{crit}}(q, t) \leq \varepsilon^D_{\text{crit}}(q, t) \leq \varepsilon^{*,D}_{\text{crit}}(q, t). \tag{7}$$

The problem addressed in this paper is to determine or bound the critical value of ε. We show that $\varepsilon_{\text{crit}}(q, t) > 0$ if and only if $q \geq t + 1$, thereby establishing existence of good robust t-IPP codes. We also find the exact critical values of $\varepsilon_{\text{crit}}$ for $t = 2$ in all the four cases defined above. Note that permitting an unlimited number of erasures in detectable positions rules out the existence of good IPP codes. Namely it is easy to show that for any code of cardinality at least $2t - 1$, the error probability of identification is close to $1/2$; see Prop. 2.6 in [5]. For the wide-sense attack (3), identification with zero error is also impossible [5].

In establishing our results we rely on properties of combinatorial arrays such as separating families [28], perfect hash families [35, 12] and hash codes [7], and partially hashing families of [6], all of which are defined formally in the next section. Each of these concepts enforces some kind of separation properties between groups of rows of the array. These properties were previously used in establishing existence results of IPP codes [6, 5, 1]. To account for the presence of mutant coordinates, we require that the separation properties hold for a certain number of entries of the rows. This leads to the notion of separating distance for an array, which in some particular cases was studied earlier in [26, 28, 7].

1-C. **Prior work on fingerprinting codes.** Fingerprinting codes were first described by Wagner [34] and Blakley et al. [13] and brought to their modern form by Boneh and Shaw [15]. Some of the earlier works on traitor tracing and collusion-secure codes considered the possibility of unreadable marks or of relinquishing the marking assumption or both. In particular, unreadable marks were added to the wide-envelope definition in [15], permitting some of the detectable coordinates to be erased. However, the present authors observed [5] that this gains no advantage for the pirates. The possibility of adding mutations at a fixed rate ε (similarly to random errors in information transmission) was considered by Guth and Pfitzmann [22]. Billet and Phan [9] considered Tardos's coding scheme [32] that permits up to εn mutations (erasures or bit flips), and estimated the rate of fingerprinting codes that support reliable identification of pirates for that scheme (with small, but positive failure probability), and Sirvent [29] and Boneh and Naor [14] did the same for the Boneh-Shaw scheme [15]. They also suggested, for codes that can handle mutations, the term robust, adopted in the present work.

2. Separating systems, partially hashing families, hash codes

In this section we collect results on set systems that satisfy a range of conditions similar to perfect hashing. We study *separating codes* (separating hash families), *hash codes* (extensions of perfect hash families), and (t,u)-*hashing codes* (partially hashing families). We begin with a general notion of separating codes.

Definition 3. *A code $\mathcal{C}\subset\mathcal{Q}^n$ is separating of type $(\tau_1,\tau_2,\dots,\tau_m)$ if for any m-tuple of pairwise disjoint subsets $U_k\subset\mathcal{C}, |U_k|=\tau_k$ for all k, there exists a coordinate that separates them: for any $1\le k<l\le m$ and for some $i\in[n]$,*

$$(U_k)_i\cap(U_l)_i=\emptyset.$$

If there are at least d coordinates with this property for each choice of the subsets $U_k, k=1,\dots,m$, we say that the code has separating distance d.

The first part of this definition appeared already in [17, 31].

Alternatively, a separating code is a family $\mathcal{F}$ of n functions $f: Y\to\mathcal{Q}$, where Y is a finite set of cardinality $|Y|=|\mathcal{C}|$, such that for any pairwise disjoint subsets $U_k\subset Y, |U_k|=\tau_k, k=1,\dots,m$, there exists at least one function $f\in\mathcal{F}$ such that

$$\{f(y): y\in U_i\}\cap\{f(y): y\in U_j\}=\emptyset$$

for all $i\ne j$. The condition on the distance is equivalent to the existence of at least d such functions for each choice of the subsets U_k.

We will use three special cases of this definition. A code of type (t_1,t_2) is called (t_1,t_2)-separating, A code of type (1^t) is called a perfect t-hash family, and a code of type $(1^t,u-t)$ is called a (t,u) hashing family (partially hashing family). The corresponding distances will be denoted by $d_{\mathrm{s}}^{(t_1,t_2)}$, d_{h}^{t}, and $d_{\mathrm{ph}}^{(t,u)}$ respectively.

Given a separating property $\mathscr{P}\in\{\mathrm{s},\mathrm{h},\mathrm{ph}\}$, define

$$R_{q,\mathscr{P}}(n,\delta)=\max\{R(\mathcal{C}): \mathcal{C}\subset\mathcal{Q}^n, d_{\mathscr{P}}(\mathcal{C})\ge\delta n\}$$

$$R_{q,\mathscr{P}}(\delta)=\liminf_{n\to\infty} R_{q,\mathscr{P}}(n,\delta). \tag{8}$$

A more detailed version of this notation also includes the size of the groups being separated. Thus, for (t_1,t_2)-separating codes we write $R_{q,\mathrm{s}}^{(t_1,t_2)}(\delta)$ and so on.

2-A. **Separating codes.** Separating codes have been studied for a number of years under different names. They were introduced in computer science [21], studied using methods of coding theory and combinatorics (see overviews [28, 17]) and more recently introduced in cryptography under the name of frameproof and secure frameproof codes [30] (the former correspond to $(t, 1)$ separation and the latter to (t, t) separation).

We focus on the case of $t_1 = t_2 = 2$. The maximum value of the relative $(2, 2)$-separating distance for which there exist infinite sequences of codes of positive rate is given in the following proposition.

Proposition 2.1. *(a) Let $\delta < \delta_{\text{crit}}^{(2,2)}(q)$, where*

$$\delta_{\text{crit}}^{(2,2)}(q) = (1 - q^{-1})(1 - 3q^{-1} + 3q^{-2}).$$

Then $R_{q,\text{s}}^{(2,2)}(\delta) \geq \frac{2}{3 \ln q}(\delta - \delta_{\text{crit}}^{(2,2)}(q))^2 > 0$. Moreover, if q is a prime power, this claim is also satisfied by sequences of linear $(2, 2)$-separating codes.

(b) Any code $\mathcal{C} \subset \mathcal{Q}^n$ with $d_{\text{s}}^{(2,2)} = d > n\delta_{\text{crit}}^{(2,2)}(q)$ satisfies

$$|\mathcal{C}| \leq \frac{6d}{d - \delta_{\text{crit}}^{(2,2)} n}. \tag{9}$$

In particular, $R_{q,\text{s}}^{(2,2)}(\delta) = 0$ for $\delta > \delta_{\text{crit}}^{(2,2)}(q)$.

Part (a) was essentially established in [26], with a small refinement in [27]. Regarding part (b), [28] establishes a weaker result, namely that $R_{q,\text{s}}^{(2,2)}(\delta) = 0$ for $\delta > ((q-1)/q)^3$. Bound (9) is proved in the Appendix.

We will also need one extension of the separating property. A code $\mathcal{C} \subset \mathcal{Q}^n$ has the *restricted* (t_1, t_2) *separation* property if for any two its disjoint subsets $U, V, |U| = t_1, |V| = t_2$, there exists a coordinate i such that $|U_i| = 1, |V_i| = 1$, and $U_i \neq V_i$ (all the vectors in U and in V have the same value in the separating coordinate i, and these values are different). This version of the separation property was already considered in [26, 28]. A straightforward application of the probabilistic method gives the following result.

Proposition 2.2. *Let $\alpha > 0$. Infinite sequences of $(2, 2)$ restricted separating codes exist for all rates R such that*

$$R + \alpha \leq -1/3 \log_q(1 - (q-1)/q^3).$$

A proof is given in the Appendix.

2-B. **Perfect hash families and hash distances of codes.** A perfect t-hash family is a set $\mathcal{C} \subset \mathcal{Q}^n$ such that for every t vectors $x^{i_1}, \ldots, x^{i_t} \in \mathcal{C}$ there exists $j \in [n]$ such that $|\{x_j^{i_1}, \ldots, x_j^{i_t}\}| = t$. Below we call such j a *hash coordinate*. Using the language of functions, a perfect hash family is a set $\mathcal{F}$ of n functions $f : Y \to \mathcal{Q}$, where Y is a finite set of cardinality $|Y| = |\mathcal{C}|$, such that for any t-subset $X \subset Y$ there exists $f \in \mathcal{F}$ that is one-to-one on X. We call $\mathcal{C}$ a t-hash code, often omitting the reference to t. The problem of constructing short hash codes of a given cardinality (the most economical collections of functions) has been extensively studied for the last few decades, see [12, 35, 31].

A generalization of this concept introduced in [7] studies hash codes with a given value of t-hash distance $d_{\text{h}}^{(t)}$. Note the $d_{\text{h}}^{(2)}$ is the usual Hamming distance of the code. We need several results for t-hash codes with a given value of $d_{\text{h}}^{(t)}$. These results were announced in [7], but their proofs were never published. Since we rely on them, and to make this paper self-contained, we give them below.

Let

$$\pi_{t,q} \triangleq \prod_{i=1}^{t-1}(1 - iq^{-1}) \tag{10}$$

The following bound is an analog for the hash distance of the Plotkin bound of coding theory [33, p.66].

Proposition 2.3. [7] *Let $\mathcal{C} \subset \mathcal{Q}^n$ be a code such that $d_{\mathrm{h}}^{(t)}(\mathcal{C}) = d$. If $d > n\pi_{t,q}$ then*

$$|\mathcal{C}| \le \binom{t}{2} \frac{d}{d - n\pi_{t,q}}. \tag{11}$$

A proof is given in the Appendix.

In particular, any code $\mathcal{C}$ of length n with $d_{\mathrm{h}}^{(t)}(\mathcal{C}) = d$ where $d/n > \pi_{t,q}$ satisfies $|\mathcal{C}| \le \binom{t}{2} d$. Bearing in mind our definition (8), we obtain the following proposition.

Proposition 2.4. *For all $\pi_{t,q} < \delta < 1$ we have $R_{q,\mathrm{h}}^t(\delta) = 0$.*

At the same time, a random choice argument shows that if $0 \le \delta < \pi_{t,q}$ then $R_{q,\mathrm{h}}^t(\delta) > 0$.

Proposition 2.5. [7] *For any $\delta < \pi_{t,q}$ we have*

$$R_{q,\mathrm{h}}^t(\delta) \ge 2((t-1)\ln q)^{-1}(\delta - \pi_{t,q})^2 > 0. \tag{12}$$

A proof is given in the Appendix. Because of the last two propositions we call the value $\pi_{t,q}$ the critical value of the relative t-hash distance of codes.

Corollary 2.6. *Let $0 < \delta < \pi_{t,q}, \alpha > 0$. There exist infinite sequences of codes with the restricted $(2,2)$-separating property, t-hash distance δn and any rate R that satisfies the inequality*

$$R + \alpha \le \min\{-1/3 \log_q(1 - (q-1)/q^3), 2((t-1)\ln q)^{-1}(\delta - \pi_{t,q})^2\}.$$

This corollary is proved in the Appendix.

Since most existence results in this section rely on a particular application of the probabilistic method, we can similarly claim existence of codes that are simultaneously $(2,2)$-separating and hash, or have a certain value of separating distance and of hash distance at the same time. These claims, whose proofs are analogous to the above corollary, will be freely used in what follows.

It is of interest to show that there exist linear IPP codes. Toward this end, we prove that there exist linear 3-hash codes of positive rate.

Proposition 2.7. [7] *Let $\alpha > 0, \delta < \pi_{3,q}$ and let q be a power of a prime. There exist infinite sequences of q-ary linear codes of the rate $(\ln q)^{-1}(\delta - \pi_{3,q})^2 - \alpha$ that have 3-hash distance at least δn.*

Proof : See Appendix.

We note that for larger t linear hash codes do not exist unless q is sufficiently large [12, 7].

2-C. (t,u)**-hashing families.** A subset $\mathcal{C} \subset \mathcal{Q}^n$ is called (t,u)-*hashing* [6] if for any two subsets T, U of $\mathcal{C}$ such that $T \subset U \subset \mathcal{C}, |T| = t, |U| = u$, there is some coordinate $i \in \{1, \dots, n\}$ such that for any $x \in T$ and any $y \in U, y \ne x$, we have $x_i \ne y_i$. The coordinates whose existence is guaranteed by this definition will be again called *hash coordinates* for given $T, U \subset \mathcal{C}$.

As shown in [6], for any $u \ge t+1$ there exist sequences of good (t,u)-hashing codes. Here we establish a generalization of this result. As usual, we say that the code $\mathcal{C}$ has (t,u)-hash distance $d_{\mathrm{ph}}^{(t,u)}(\mathcal{C}) = d$ if every pair of subsets T, U has at least d hash coordinates. We have the following proposition whose proof is given in the Appendix.

Proposition 2.8. *Let $u \geq t+1$ and let $\delta < P_q(t,u)$, where*

$$P_q(t,u) = 1 - \pi_{t,q}(1 - tq^{-1})^{u-t}. \tag{13}$$

We have

$$R^{(t,u)}_{q,\mathrm{ph}}(\delta) \geq ((u-1)\ln q)^{-1}(\delta - P_q(t,u))^2 > 0.$$

Remark. In our arguments we have applied the probabilistic method in its simplest form. It is possible to improve some of the results stated above relying on more refined arguments. For instance, Alon et al. [1] observed that under biased selection of code symbols, the probability $P_{t+1}(t,u)$ can be shown to be $1 - t!(u-t)^{u-t}u^{-u}$. This number is smaller than the right-hand side of (13), and so the values of δ for which $R^{(t,u)}_{t+1,\mathrm{ph}}(\delta) > 0$ can be larger than in the above proposition. Better rates of some binary (t_1, t_2)-separating codes for $t_1 \neq t_2$, with no consideration of the separating distance, were found in [17] (again, biased selection helps).

Several nonasymptotic improvements of the application of the probabilistic method for perfect hash families and other combinatorial arrays were considered by Blackburn and Wilde [12] and Deng-Stinson-Wei [18].

3. Existence of robust IPP codes

3-A. **Robust IPP codes with traceability property.** Denote by $d_H(x,y) = |\{i = 1,\dots,n : x_i \neq y_i\}|$ the Hamming distance between vectors $x, y \in \mathcal{Q}^n$. The minimum distance $d_H(c,c')$ between distinct codewords $c, c' \in \mathcal{C}$ will be called the (Hammng) distance of the code $\mathcal{C} \subset \mathcal{Q}^n$ and denoted $d_H(\mathcal{C})$. The value $\delta = d_H(\mathcal{C})/n$ is called the relative code distance. We also use the notation $s_H(x,y) \triangleq n - d_H(x,y)$ for the number of equal coordinates in the vectors. Finally for $x \in \mathcal{Q}^n, Y \subset \mathcal{Q}^n$ we write $d_H(x,Y) \triangleq \min_{y\in Y} d_H(x,y)$. For instance, a descendant y of a coalition U with e mutant coordinates is any vector y that satisfies $d_H(y, \langle U\rangle) = e$.

Some of the known results for t-IPP codes can be easily generalized to the new problem. For instance, [16] showed that any code with Hamming distance $d > (1 - t^{-2})n$ is a t-IPP code. A generalization is as follows.

Proposition 3.1. *Let $\mathcal{C} \subset \mathcal{Q}^n$ be a q-ary code with distance $d_H(\mathcal{C}) = \delta n$ such that*

$$\delta > 1 - t^{-2} + \varepsilon(t^{-1} + t^{-2}) \tag{14}$$

where $0 < \varepsilon < (t+1)^{-1}$. Then for any t-coalition $U \subset \mathcal{C}$ and any vector y such that $d(y, \langle U\rangle) \leq \varepsilon n$ there exists at least one vector $u \in U$ such that

$$s_H(y,u) \geq (1-\varepsilon)n/t$$

and for all $c \in \mathcal{C} \setminus U$,

$$s_H(y,c) < (1-\varepsilon)n/t.$$

Proof : Let $U = \{u^1,\dots,u^t\}$ be a coalition and let y be a descendant of U with $e \leq \varepsilon n$ mutant coordinates. For any non-mutant coordinate i there exists at least one $u \in U$ such that $u_i = y_i$. Therefore

$$\sum_{u\in U} s_H(y,u) \geq (1-\varepsilon)n,$$

and hence there exists a vector $u_0 \in U$ such that $s_H(u_0, y) \geq n(1-\varepsilon)t^{-1}$.

On the other hand, let $c \in \mathcal{C}\setminus U$. Then for any non-mutant coordinate i the equality $y_i = c_i$ implies that $c_i = u_i$ for some $u \in U$. Therefore the number of non-mutant positions j such that

$y_j = c_j$ does not exceed $\sum_{u\in U} s_H(u,c)$. Hence,

$$\begin{aligned} s_H(y,c) &\le n\varepsilon + \sum_{u\in U} s_H(u,c) \le n\varepsilon + t(n - d_H(\mathcal{C})) \\ &< (1-\varepsilon)nt^{-1}. \end{aligned}$$

∎

Recall the notion of traceability codes (t-TA codes) [16, 30, 11]: a t-IPP code $\mathcal{C}$ has the t-TA property if for any $y \in \langle \mathcal{C}\rangle_t$ the vector $c \in \mathcal{C}$ closest to y by the Hamming distance is always one of the parents of y, i.e.,

$$c \in \bigcap_{U\subset\mathcal{C},\ |U|\le t,\ y\in\langle U\rangle} U.$$

This implies that a pirate can be provably identified by finding any vector $c \in \mathcal{C}$ such that $c = \arg\min_{x\in\mathcal{C}} d_H(x,y)$. Note that for t-IPP codes, identification of pirates is substantially more complex, requiring a search over all t-subsets of $\mathcal{C}$ as opposed to just finding the closest codevector to y.

Generalizing this definition to include mutations, call a code $\mathcal{C}$ a (t,ε)-TA code (*robust TA code*) if the above property holds true even in the presence of εn mutations. We can rephrase the previous proposition as follows.

Theorem 3.2. *For $q > t^2/(1-\varepsilon(t+1))$ there exist (t,ε)-TA codes with nonvanishing code rate.*

Proof: By the Gilbert-Varshamov bound [33, p.66], for any $0 \le \delta < (q-1)/q$ there exist sequences of q-ary codes of growing length n whose relative Hamming distance converges to δ and rate R converges to a positive number (a function of δ). Using $\delta < (q-1)/q$ in (14) we obtain the inequality $\frac{t^2}{q} < 1 - \varepsilon(t+1)$. ∎

Returning to our main problem, we obtain

Corollary 3.3. *For any $q > t^2$ we have*

$$\varepsilon_{\text{crit}}(q,t) \ge \frac{1}{t+1} - \frac{t^2}{q(t+1)}.$$

3-B. **Existence of robust IPP codes for $q \ge t+1$.** Traceability is a more restrictive property than IPP: it is shown in [6] there that t-IPP codes with positive rate exist over smaller-sized alphabets than t-TA codes. More specifically, [6] proves that good t-IPP codes exist for all $q \ge t+1$, and that this bound is exact. The following generalization of this result to (t,ε)-IPP codes holds true.

Theorem 3.4. *For any $q \ge t+1$ we have $\varepsilon_{\text{crit}}(q,t) > 0$.*

For the proof we need the following lemma which is close to Lemma 3.7 from [6].

Lemma 3.5. *Let m be an integer. If $\mathcal{C}$ has the $(t, m(t-m+2))$-hash property for all $m = 2,3,\dots,t+1$ then it is t-IPP.*

Proof: Let $\mathcal{X} = (X_1,\dots,X_m)$ be a collection of subsets of codewords of a code $\mathcal{C}$ with $|X_i| \le t, i = 1,\dots,m$. Call $\mathcal{X}$ a configuration if $\cap_{i=1}^m X_i = \emptyset$, and call $\mathcal{X}$ a minimal configuration if it is minimal under inclusion.

Suppose the contrary, i.e., that for some $y \in \mathcal{Q}^n$ the set of all subsets $X \subset \mathcal{C}, |X| \le t$ such that $y \in \langle X\rangle$ forms a configuration $\mathcal{X}$. Every configuration contains a minimal configuration, so we can assume that $\mathcal{X} = (X_1,\dots,X_m)$ is minimal. By minimality, for all $j = 1,\dots,m$ there exists an element $b^j \in \cap_{i\ne j} X_i$, and for distinct indices j the elements b^j are different. Let

$B(\mathcal{X}) = \{b^1, \dots, b^m\}$. We have $(B(\mathcal{X})\backslash b^j) \subset X_j$ for any $j = 1, \dots, m$, and hence $m \leq t+1$. Let $U = \cup_{X\in\mathcal{X}} X$, then

$$\begin{aligned} |U| &\leq \sum_i (|X_i\backslash(B(\mathcal{X})\backslash b^i)|) + m \\ &= \sum_{i=1}^{m} (|X_i| - (m-1)) + m \\ &\leq m(t-m+2). \end{aligned}$$

Let $T \in \mathcal{X}$. By assumption, for T, U thus chosen, there is a (t,u)-hash coordinate i, which means that $x_i \neq x'_i$ for all distinct $x \in T, x' \in U$. Moreover, since $y \in \langle T\rangle$, there exists an $x \in T$ such that $x_i = y_i$, so x is a parent of y. Because of the (t,u)-hash property for all other points $x' \in U$, we have $y_i \neq x'_i$. Therefore, this vector x is in fact unique, and every $T \in \mathcal{X}$ contains x, a contradiction. ∎

Remark : Of course, if a code has the (t,u) hashing property for some $u \geq t+1$, it also has the (t,u') hashing property for all $t+1 \leq u' < u$. Thus the statement of Lemma 3.5 can be reduced to one value of m, the one that gives the maximum of $u = m(t-m+2)$. This value $u_0 = \lfloor (t+2)^2/4 \rfloor$ was used in [6] and subsequent works. For our present purposes we need codes that satisfy a set of conditions for all values of $m = 2, \dots, t+1$, which requires more detailed considerations.

Proof of Theorem 3.4 : Let $\mathcal{C}$ be a code. Let $X \subset \mathcal{C}, |X| \leq t$ and let $y \in \langle X\rangle_\varepsilon$. Let $\mathcal{X} = (X_1, X_2, \dots, X_m)$ be the set of all coalitions that could generate y with εn mutations. Let $U = \cup_{X\in\mathcal{X}} X, u = |U|$. By the previous lemma, it suffices to show that for any $i = 1, \dots, m$ the pair X_i, U contains at least one (t,u)-hash coordinate for all $m = 2, \dots, t+1$.

Let us fix m, u such that $2 \leq m \leq t+1 \leq u \leq m(t-m+2)$. Assume that

$$0 < \varepsilon n \leq \frac{d-1}{m} \tag{15}$$

where $d = d_{\text{ph}}^{(t,u)}(\mathcal{C})$ is the (t,u)-hash distance of $\mathcal{C}$. Then the total number of mutations that can be introduced by the coalitions in $\mathcal{X}$ is at most $d-1$ and therefore there exists at least one (t,u)-hash coordinate. Then the previous lemma implies the t-IPP property.

Proposition 2.8 implies that as long as $\delta < P_q(t,u)$, there exist sequences of codes of rate R approaching $R_{q,\text{ph}}^{(t,u)}(\delta) > 0$ and (t,u)-hash distance δn. For a given m it suffices to consider the value $u = m(t-m+2)$ because $d_{\text{ph}}^{(t,u_1)}(\mathcal{C}) \geq d_{\text{ph}}^{(t,u_2)}(\mathcal{C})$ if $u_1 \leq u_2$. Let $\delta_m < P_q(t, m(t-m+2)), m = 2, \dots, t+1$ be a set of real numbers. By a remark made after Corollary 2.6 there exist codes of positive rate and $(t, m(t-m+2))$-hash distance $\geq \delta_m n$ simultaneously for all $m = 2, \dots, m+1$.

To conclude, for any set of t positive real numbers $\delta_m, m = 2, \dots, t+1$ such that $\delta_m < P_q(t, m(t-m+2))$ there exist sequences of codes of positive rate and hash distance $\delta_m n$. These codes have the (t,ε) IPP property for all ε that satisfy

$$0 < \varepsilon < \min_{2\leq m\leq t+1} \frac{\delta_m}{m},$$

which proves our claim. ∎

4. Hash distances and upper bounds for robust IPP codes

In this section we derive upper bounds on $\varepsilon_{\text{crit}}$ under the marking assumption. For that, we show that for $\varepsilon > \pi_{t+1,q}$ in the case of erasures and for $\varepsilon > \pi_{t+1,q}/(t+1)$ in the case of arbitrary mutations, exact identification is impossible, thus rendering $R_q(t,\varepsilon) = 0$ for these values of ε.

Theorem 4.1.

$$\varepsilon^{*,D}_{\text{crit}}(q,t) < \pi_{t+1,q},$$
$$\varepsilon^{D}_{\text{crit}}(q,t) < \pi_{t+1,q}/(t+1).$$

Proof : Consider an arbitrary code $\mathcal{C}$ and let $d = d_{\text{h}}^{(t+1)}(\mathcal{C})$ be its $(t+1)$st hash distance. Take a subset $V = \{v^1, \dots, v^{t+1}\} \subset \mathcal{C}$ and let $A \subset [n], |A| = d$ be the set of its hash coordinates (the coordinates in which all the vectors of V are different). Form a vector $(y_i,\, i \in [n])$ such that

$$y_i = \begin{cases} * & \text{if } i \in A \\ \alpha_i & \text{if } i \notin A \end{cases},$$

where α_i is the most frequent symbol among $v_i^1, \dots, v_i^{t+1}$. Since $\alpha_i = v_i^k$ at least for two values of $k = 1, \dots, t+1$, we have $y \in \langle U \rangle_\delta$ for any t-subset $U \subset V, \delta = d/n$. Thus, it is impossible to identify a parent of y with certainty.

Turning to mutant coordinates of arbitrary value (but still following the marking assumption), let us partition A into $t+1$ disjoint, (almost) equal parts:

$$A = \bigcup_{j=1}^{t+1} A_j.$$

Consider the vector $(y_i,\, i \in [n])$ such that

$$y_i = \begin{cases} v_i^j & \text{if } i \in A_j \\ \alpha & \text{if } i \notin A \end{cases},$$

where α has the same meaning as above. Clearly, y can be generated by any t-subset of V using at most

$$\max_{1 \le j \le t+1} |A_j| \le \lceil (t+1)^{-1} d \rceil$$

mutations.

By Prop. 2.4, any code sequence with positive rate must have $d/n < \pi_{t+1,q}$. Together with (5) this implies our claims. ∎

5. Robust 2-IPP codes

Existence of good $(2, \varepsilon)$-IPP codes for some positive ε has been already established in Theorem 3.4. However its proof uses only a sufficient condition for the robust IPP property, not resulting in an optimal value of ε. In this section we strengthen this result, finding the exact value of $\varepsilon_{\text{crit}}(q, 2)$ for all the four versions of the problem considered.

Hollmann et al. [23] provided a characterization of 2-IPP codes in terms of their separating properties. We start with extending this analysis to the case of robust IPP codes. We write (a, b) to refer to a pair of vectors from $\mathcal{Q}^n$. We also use an abbreviated notation $\langle a, b \rangle^W, W =$ (I)-(IV) (instead of $\langle (a, b) \rangle_\varepsilon^W$) to refer to the set of vectors that can be generated by the vectors a, b following (1) in non-mutant coordinates and following one of the rules (I)-(IV) of Sect. 1 in creating mutations.

For a code $\mathcal{C} \subset \mathcal{Q}^n$ we write

$$\langle \mathcal{C} \rangle^W = \bigcup_{\substack{(a,b) \in \mathcal{C} \times \mathcal{C} \\ a \ne b}} \langle a, b \rangle^W,$$

omitting the subscript ε. For a vector $y \in \langle \mathcal{C} \rangle^W$ consider a graph $G_y(V, E)$, where $V = \mathcal{C}$ and $(c_1, c_2) \in E$ if $y \in \langle c_1, c_2 \rangle^W$. By definition, the code $\mathcal{C}$ can identify at least one of the pirates if and only if for any y the graph G_y is a star (all the edges in E intersect on a point).

Proposition 5.1. *A code $\mathcal{C}$ is a robust 2-IPP code if and only if for any vector $y \in \langle\mathcal{C}\rangle^W$ the graph G_y has no triangles and any two edges have a common vertex.*

This proposition is an easy generalization of Lemma 1 from [23]. According to it, $\mathcal{C}$ is t-IPP if it is simultaneously 3-hash and (2,2)-separating. Switching to the language of graphs, this means that for any $y \in \langle\mathcal{C}\rangle^W$ the graph G_y contains no triangles and no parallel edges.

First we study the case when mutations occur only in detectable coordinates.

Theorem 5.2. *Assume that the pirates follow the attack rule (I) or (II). For all $q \geq 3$,*

$$\varepsilon^{*,D}_{crit}(q,2) = \Big(1-\frac{1}{q}\Big)\Big(1-\frac{2}{q}\Big)$$

$$\varepsilon^{D}_{crit}(q,2) = \frac{1}{3}\Big(1-\frac{1}{q}\Big)\Big(1-\frac{2}{q}\Big).$$

Proof : Owing to Theorem 4.1 we only need to prove existence of sequences of the corresponding $(2,\varepsilon)$-IPP codes. The proof relies on the sufficient conditions for the IPP property of Prop. 5.1. Let $\mathcal{C}$ be a code with the restricted $(2,2)$ separation property of Prop. 2.2. For two pairs (a,b) and (c,d) of distinct codevectors let i be a separating coordinate. Since any y generated by (a,b) obeys the marking assumption (4), clearly $y \notin \langle c,d\rangle^W$. Thus the graph G_y has no parallel edges for either $W =$ (I) or (II).

Suppose in addition that $\mathcal{C}$ has 3-hash distance d. If $d \geq n\varepsilon + 1$, then for any $a,b,c \in \mathcal{C}$

$$\langle a,b\rangle^{(\mathrm{II})} \cap \langle b,c\rangle^{(\mathrm{II})} \cap \langle a,c\rangle^{(\mathrm{II})} = \emptyset.$$

Indeed, if at most εn erasures have occurred, there will be at least one non-mutant hash coordinate for (a,b,c). This implies that for any $y \in \langle\mathcal{C}\rangle^{(\mathrm{II})}$ the graph G_y is triangle-free. In the case of εn arbitrary mutations (Case (I)), the absence of triangles is guaranteed by the condition $d^3_{\mathrm{h}}(\mathcal{C}) \geq 3\varepsilon n+1$ because the pairs (a,b), (b,c) and (a,c) together can create at most $3\varepsilon n$ mutant coordinates, so there will be at least one non-mutant hash coordinate, say i. In this coordinate, if $y_i \in \{a_i,b_i\}$ and $y_i \in \{a_i,c_i\}$, then $y_i \notin \{b_i,c_i\}$, so the vertices a,b,c do not form a triangle in G_y.

By the remark after Corollary 2.6 it is possible to construct a sequence of codes of increasing length n with positive rate that have both the restricted separating property and 3-hash distance δn for all $\delta < \pi_{3,q}$. These codes will have the $(2,\varepsilon)$-IPP property for all $\varepsilon < \delta$ for mutations of type (II) and $\varepsilon < (1/3)\delta$ for type (I). ∎

Now suppose that mutations can stray away from the detectable coordinates.

Theorem 5.3. *Assume that the pirates follow the attack rule (III) or (IV). For all $q \geq 3$,*

$$\begin{aligned}\varepsilon^{*}_{\mathrm{crit}}(q,2) &= \Big(1-\frac{1}{q}\Big)\Big(1-\frac{3}{q}+\frac{3}{q^2}\Big)\\ \varepsilon_{\mathrm{crit}}(q,2) &= \frac{1}{3}\Big(1-\frac{1}{q}\Big)\Big(1-\frac{2}{q}\Big).\end{aligned}$$

Proof : The proof is similar to the proof of the previous theorem. We begin with the case of erasures. Let $\mathcal{C} \subset \mathcal{Q}^n$ be a code with $(2,2)$-separating distance δn, let y be an attack vector, and let G_y be the corresponding graph. Let $(a,b),(c,d)$ be two pairs of distinct codewords, $(a,b)\cap(c,d)=\emptyset$. Any vectors $y \in \langle a,b\rangle$, $y' \in \langle c,d\rangle$ formed according to (1) differ in at least δn coordinates, so even if εn coordinates of are erased, (c,d) is not an edge of G_y. Assuming in addition that $\mathcal{C}$ has the 3-hash distance $\geq \varepsilon n + 1$, we argue as in the previous proof that G_y is triangle-free. By Prop. 2.1 if $\delta < \delta^{(2,2)}_{\mathrm{crit}}$ then there exist infinite sequences of codes of positive rate that have (2,2) separating

distance at least δn. Similarly, by Prop. 2.5 for any $\delta < \pi_{3,q}$ there exist sequences of codes of positive rate and 3-hash distance δn. Together with the remark after Cor. 2.6 this implies that

$$\varepsilon^*_{\text{crit}}(q,2) \geq \min(\pi_{3,q}, \delta^{(2,2)}_{\text{crit}}) = \Big(1-\frac{1}{q}\Big)\Big(1-\frac{3}{q}+\frac{3}{q^2}\Big) \qquad (q\geq 3). \tag{16}$$

Turning to upper bounds on ε, we have from Theorem 4.1 and (6) that $\varepsilon^*_{\text{crit}}(q,2) \leq \pi_{3,q}$. At the same time, if two pairs of codewords $(a,b),(c,d)$ have (2,2) separating distance δn, and y is an attack vector in which the δn separating coordinates are erased, then the graph G_y has both edges (a,b) and (c,d). so exact identification is impossible. However, by (9), the rate of any sequence of codes with $d^{(2,2)}_{\text{s}} > \delta^{(2,2)}_{\text{crit}} n$ approaches 0. This shows that (16) holds with equality, proving the first part of the claim.

Now let us consider arbitrarily valued mutations of Case (III). The arguments are analogous to the first part and lead to the equality

$$\varepsilon_{\text{crit}}(q,2) = \min\Big\{\frac{\pi_{3,q}}{3}, \frac{\delta^{(2,2)}_{\text{crit}}}{2}\Big\} = \frac{1}{3}\Big(1-\frac{1}{q}\Big)\Big(1-\frac{2}{q}\Big) \qquad (q\geq 3).$$

For instance, assuming that each of the two pairs $(a,b),(c,d)$ can alter at most εn positions, the condition $d^{(2,2)}_{\text{s}}(\mathcal{C}) \geq 2\varepsilon n + 1$ suffices to rule out parallel edges, so $\varepsilon_{\text{crit}}(q,2) \geq \delta^{(2,2)}_{\text{crit}}/2$, etc. ∎

Finally, notice that restricting our choice to linear codes (for field-sized alphabets $\mathcal{Q}$) does not change the values of $\varepsilon_{\text{crit}}$ found in the last theorem. This is because standard ensembles of random linear codes with high probability have both $(2,2)$-separating and 3-hash properties as long as the separating and hash distances are less than their critical values $\delta^{(2,2)}_{\text{crit}}$ and $\pi_{3,q}$, respectively.

In particular, from the last two theorems we get the following values for the critical rate of mutations for $q=3$ in $(2,\varepsilon)$-IPP codes:

$$\varepsilon^{*,D}_{\text{crit}}(3,2) = \varepsilon^*_{\text{crit}}(3,2) = 2/9, \quad \varepsilon^D_{\text{crit}}(3/2) = \varepsilon_{\text{crit}}(3,2) = 2/27.$$

These critical values can be attained by sequences of ternary linear codes.

APPENDIX

Proof of Prop. 2.1(b) (outline) : Let $|\mathcal{C}| = M$. Count the sum S of $(2,2)$-separating distances for all choices of 4 distinct codewords of the code $\mathcal{C}$. Let $\lambda_{i,\alpha} = |\{x\in\mathcal{C} : x_i = \alpha\}|$. The contribution of the ith coordinate to S equals

$$S_i = \sum_{\alpha\in\mathcal{Q}} \Big\{\lambda^2_{i,\alpha}(M-\lambda_{i,\alpha})^2 + \sum_{\beta\neq\alpha}\lambda_{i,\alpha}\lambda_{i,\beta}(M-\lambda_{i,\alpha}-\lambda_{i,\beta})^2\Big\}.$$

One checks that the maximum of the form S_i under the condition $\sum_{\alpha\in\mathcal{Q}}\lambda_{i,\alpha} = M$ is attained for $\lambda_{i,\alpha} = M/q$. Therefore,

$$\begin{aligned} S = \sum_{i=1}^n S_i &\leq n\Big(M\frac{M}{q}\Big(M-\frac{M}{q}\Big)^2 + M\Big(M-\frac{M}{q}\Big)\Big(M-\frac{2M}{q}\Big)^2\Big) \\ &= nM^4\delta^{(2,2)}_{\text{crit}}(q). \end{aligned}$$

At the same time,

$$S \geq M(M-1)(M-2)(M-3)d \geq (M^4-6M^3)d.$$

Combining the last two equations yields the result.

Proof of Prop. 2.2 : Suppose that the coordinates of codewords of a code $\mathcal{C}$ of size M are chosen from $\mathcal{Q}$ uniformly with replacement. The probability that a given coordinate in two given pairs

(x^1, x^2) and (x^3, x^4) of codewords fails the restricted separation condition equals $\lambda = 1 - (q-1)/q^3$. Therefore, the expected number of quadruples (more precisely, pairs of unordered pairs) of codewords D that violate the condition in a given coordinate is at most $\binom{M}{2}\binom{M-2}{2}\lambda^n/4 \le M^4\lambda^n/4$. Thus for a random code $\mathcal{C}$

$$\Pr(\sharp \text{ bad quadruples of codewords} \ge n^{-1}M) \le \frac{nM^3\lambda^n}{4}.$$

Taking $M = (4n^{-2}\lambda^{-n})^{1/3}$, we observe that this probability is bounded above by $1/n$. Thus with probability $(n-1)/n$ the number of bad quadruples of codewords in a random code of size M does not exceed $n^{-1}M$. Deleting one element out of each bad quadruple leaves us with a $(2,2)$ restricted separating code of size $q^{nR} = (4n^{-2}\lambda^{-n})^{1/3}(1-n^{-1})$. This concludes the proof.

Proof of Prop. 2.3 : The proof is similar to the proof of Prop. 2.1. Let $|\mathcal{C}| = M$ and let $\lambda_{i,\alpha} = |\{x \in \mathcal{C} : x_i = \alpha\}|$. Clearly,

$$\sum_{\alpha \in \mathcal{Q}} \lambda_{i,\alpha} = M.$$

Denote

$$S_i = \sum_{U \subset \mathcal{C}:\ |U|=t} \chi_i(U)$$

where $\chi_i(U) = 1$ if i is a hash coordinate for the subset U and $\chi_i(U) = 0$ otherwise. We have

$$S_i = \sum_{\{\alpha_1,\dots,\alpha_t\} \subset \mathcal{Q}} \prod_{j=1}^{t} \lambda_{i,\alpha_j} \tag{17}$$

where the sum extends to all the t-tuples of distinct symbols from the alphabet. Indeed, choosing one vector out of each of the t sets λ_{i,α_j} for a fixed t-tuple $\{\alpha_1, \dots, \alpha_t\}$ accounts for an t-tuple of codevectors U for which i is a hash coordinate. The right-hand side of (17) is maximized if $|\lambda_{i,\beta} - \lambda_{i,\gamma}| \le 1$ for all $\beta, \gamma \in \mathcal{Q}, \beta \ne \gamma$. Indeed, suppose the contrary, i.e., that $\lambda_{i,\beta} - \lambda_{i,\gamma} \ge 2$ for some β, γ. Then decrease $\lambda_{i,\beta}$ by one and increase $\lambda_{i,\gamma}$ by one. As a result, every product term in the sum in (17) that involves both β and γ will increase. At the same time, for every term that involves only β the decrease will be compensated by the equal increase of the corresponding term that involves only α. Therefore,

$$S_i \le (M/q)^t \binom{q}{t} = \frac{M^t \pi_{t,q}}{t!}.$$

Then

$$\binom{M}{t} d \le \sum_{i=1}^{n} S_i \le \frac{nM^t\pi_{t,q}}{t!}.$$

Since $M \ge t$ by definition, $\prod_{i=0}^{t-1}(M-i) \ge M^t - \binom{t}{2}M^{t-1}$, and we obtain

$$dM - \binom{t}{2} d \le nM\pi_{t,q}.$$

Solving for M concludes the proof. ∎

Below we use the following standard estimate. Let $Y_i, i = 1, \dots, n$ be i.i.d. Bernoulli random variables with $\Pr(Y_i = 1) = p$ for all i, and let $0 < \alpha \le p$. Then

$$\Pr[\sum_{i=1}^{n} Y_i \le \alpha n] \le e^{-2n(\alpha-p)^2}. \tag{18}$$

This follows from the Pinsker inequality or alternatively, from the Hoeffding bound (see, e.g., [25]). Slightly better estimates are possible (the Chernov bound [25] or the estimate $e^{-\frac{n(\alpha-p)^2}{2p(1-p)}}$ [8]), but we will opt for the above computationally simple bound.

Proof of Prop. 2.5: Consider a random code $\mathcal{C}$ of cardinality M whose codeword coordinates are chosen uniformly and independently from $\mathcal{Q}$. The probability that a given coordinate i is hash for an t-subset $U \subset \mathcal{C}$ equals $\pi_{t,q}$. Since the events $\{\text{coordinate } i \text{ is } (T,U)\text{-hash}\}$ are independent for different i, the probability that U is bad (contains fewer than δn hash coordinates) can be bounded by (18) as follows:

$$P_b = \sum_{j=0}^{\delta n-1} \binom{n}{j} \pi_{t,q}^j (1-\pi_{t,q})^{n-j} \le e^{-2n(\delta-\pi_{t,q})^2}$$

The expected number of bad t-tuples equals $P_b\binom{M}{t}$. From this point on we proceed as in the proof of Proposition 2.2. As a result, we claim that with probability $(1-1/n)$, a random code of size $M = (t!n^{-2}e^{2n(\delta-\pi_{t,q})^2})^{\frac{1}{t-1}}$ will contain not more than M/n bad t-tuples. Deleting one codeword from each of them leaves a code of cardinality $q^{Rn} = M(1-1/n)$ with no bad t-tuples, i.e., with t-hash distance at least δn.

Proof of Corollary 2.6: Consider a random code $\mathcal{C}$ of cardinality

$$M = \min\left\{(4n^{-2}(1-(q-1)/q^3)^{-n})^{1/3}, (t!n^{-2}q^{2n(\ln q)^{-1}(\delta-\pi_{t,q})^2})^{\frac{1}{t-1}}\right\}.$$

From the proofs of Prop. 2.2 and Prop. 2.5 with probability $\ge (n-2)/n$ the code $\mathcal{C}$ contains fewer than M/n quadruples of codewords that fail the restricted separation property and fewer than M/n t-tuples of codewords that contain fewer than δn hash coordinates. Deleting at most $M/2n$ codewords from the code $\mathcal{C}$, we obtain a code $\mathcal{C}'$ with the claimed hash and separating properties. For given α and R we clearly can find n large enough so that the code $\mathcal{C}'$ has rate R.

Proof of Prop. 2.7: Construct a random linear $k \times n$ matrix G whose elements are chosen from $\mathbb{F}_q$ independently and uniformly, and consider the $\mathbb{F}_q$-linear space $\mathcal{C} = \{xG : \ x \in \mathbb{F}_q^k\}$ of cardinality q^k. Denote by $g^1, \dots, g^n$ the columns of G. Then a codevector $c = xG$ can be written as $(c_1, \dots, c_n)$, where $c_i = (x, g^i)$. Consider any three different codevectors c_1, c_2, c_3. Since being 3-hash is a translation invariant property, we assume w.l.o.g. that $c_3 = 0$. Let $c_1 = aG$ and $c_2 = bG$, where $a, b \in \mathbb{F}_q^k$.

Case 1: If a and b are not collinear, then the probability that a given coordinate is 3-hash equals $(1-1/q)(1-2/q) = \pi_{3,q}$. Hence, the probability P_1 that the triple c_1, c_2, c_3 contains fewer than δn hash coordinates equals

$$P_1 = \sum_{j=0}^{d-1} \binom{n}{j} \pi_{3,q}^j (1-\pi_{3,q})^{n-j} \le e^{-2n(\delta-\pi_{3,q})^2}.$$

Case 2: If $b = \lambda a$, then the probability that a given i-th coordinate is 3-hash equals $1-1/q$ and the probability P_2 that the triple c_1, c_2, c_3 contains fewer than δn hash coordinates satisfies $P_2 \le e^{-2n(\delta-(q-1)/q)^2}$.

Using the union bound, there exists a linear code with 3-hash distance δn if $q^{2k}\max(P_1, P_2) < 1$, and at least a $1-1/n$ proportion of linear codes have this property if

$$q^{2k}\max(P_1, P_2) < 1/n.$$

This yields the following condition on the code rate:

$$R + \varepsilon < (\ln q)^{-1}\min((\delta-\pi_{3,q})^2, (\delta-(q-1)/q)^2) = (\ln q)^{-1}(\delta-\pi_{3,q})^2.$$

Proof of Prop. 2.8 The proof proceeds analogously to Prop.2.5 if one observes that, under the uniform distribution for the selection of the code symbols, the probability that a given coordinate is not hash for a given pair of subsets T, U depends only on their cardinalities t, u and equals $P_q(t,u)$. The expected number of bad choices of the subsets equals

$$E_{u,t} = \binom{M}{u}\binom{u}{t} P_q(t,u) \leq \frac{M^{u\,t}}{u}\, u!t! P_q(t,u).$$

The probability that a random code contains more than M/n such choices is not more than $nE_{u,t}/M$ which equals n^{-1} for $M = (u!t!u^{-t}P_q(t,u)^{-n})^{\frac{1}{u-1}}$. Deleting one vector from each pair of subsets $T \subset U$, we obtain a (t,u)-hashing code $\mathcal{C}$ with $d_{\mathrm{ph}}^{(t,u)} \geq \delta n$ and cardinality $q^{nR} \geq M(1-n^{-1})$.

REFERENCES

1. N. Alon, G. Cohen, M. Krivelevich, and S. Litsyn, *Generalized hashing and parent-identifying codes*, J. Combinatorial Theory Ser. A **104** (2003), 207–215.
2. N. Alon and U. Stav, *New bounds on parent-identifying codes: the case of multiple parents*, Combinatorics, Probability and Computing **13** (2004), no. 6, 795–807.
3. E. Amiri and G. Tardos, *High rate fingerprinting codes and fingerprinting capacity*, Proc. 20th ACM-SIAM Sympos. Discrete Algorithms (SODA 2009), 2009, pp. 336–345.
4. N. P. Anthapadmanabhan, A. Barg, and I. Dumer, *Fingerprinting capacity under the marking assumption*, IEEE Trans. Inform. Theory **54** (2008), no. 6, 2678–2689.
5. A. Barg, G. R. Blakley, and G. Kabatiansky, *Digital fingerprinting codes: Problem statements, constructions, identification of traitors*, IEEE Trans. Inform. Theory **49** (2003), no. 4, 852–865.
6. A. Barg, G. Cohen, S. Encheva, G. Kabatiansky, and G. Zémor, *A hypergraph approach to IPP codes: the case of multiple parents*, SIAM J. Discrete Math. **14** (2001), no. 3, 423–431.
7. L. A. Bassalygo, M. Burmester, A. G. Dyachkov, and G. A. Kabatiansky, *Hash codes*, Proc. 1997 IEEE Int. Sympos. Information Theory, Ulm, Germany, 1997, p. 174.
8. L. A. Bassalygo and M. S. Pinsker, *Restricted asynchronous multiple access*, Problems of Information Transmission **19** (1983), no. 4, 92–96.
9. O. Billet and D. Phan, *Efficient traitor tracing from collusion secure codes*, 3rd Internat. Conf. Information Theoretic Security (ICITS 2008) (R. Safavi-Naini, ed.), 2008, pp. 171–182.
10. S. Blackburn, *An upper bound on the size of a code with the k-identifiable property*, J. Combinatorial Theory Ser. A **102** (2003), 179–185.
11. S. R. Blackburn, T. Etzion, and S.-L. Ng, *Traceability codes*, J. Combinatorial Theory Ser. A **117** (2010), 1049–1057.
12. S. R. Blackburn and P. R. Wild, *Optimal linear perfect hash families*, J. Combin. Theory Ser. A **83** (1998), no. 2, 233–250.
13. G. R. Blakley, C. Meadows, and G. Purdy, *Fingerprinting long forgiving messages*, Proc. CRYPTO'85, 1985, pp. 180–189.
14. D. Boneh and M. Naor, *Traitor tracing with constant size cihpertext*, Proceedings of the 15th ACM conference on Computer and Communications Security, CCS'08, ALexandria, VA, 2008, pp. 501–510.
15. D. Boneh and J. Shaw, *Collusion-secure fingerprinting for digital data*, IEEE Trans. Inform. Theory **44** (1998), no. 5, 1897–1905.
16. B. Chor, A. Fiat, and M. Naor, *Tracing traitors*, CRYPTO '94, Lect. Notes Comput. Science, Springer-Verlag, New York e. a., 1994, pp. 257–270.
17. G. D. Cohen and H. G. Schaathun, *Asymptotic overview on separating codes*, Tech. Report 248, Department of Informatics, University of Bergen, Bergen, Norway, 2003.
18. D. Deng, D. R. Stinson, and R. Wei, *The Lovász local lemma and its application to some combinatorial arrays*, Des. Codes Cryptogr. (2004), 121–134.
19. I. Dumer, *Equal-weight fingerprinting codes*, Coding and Cryptology, Lecture Notes in Computer Science, vol. 5557, 2009, pp. 43–51.
20. M. Fernandez and M. Soriano, *Identification of traitors in algebraic-geometric traceability codes*, IEEE Trans. Signal Processing **52** (2004), no. 10, 3073–3077.
21. A. D. Friedman, R. L. Graham, and J. D. Ullman, *Universal single transition time asynchronous state assignments*, IEEE Trans. Comput. **18** (1969), no. 6, 541–547.

22. H.-J. Guth and B. Pfitzmann, *Error- and collusion-secure fingerprinting for digital data*, Proc. Information Hiding Workshop (IH'99) (A. Pfitzmann, ed.), 2000, pp. 134–145.
23. 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 Ser. A **82** (1998), no. 2, 121–133.
24. Y. Huang and P. Moulin, *Saddle-point solution of the fingerpinting capacity game*, Proc. IEEE International Symposium on Information Theory (ISIT2009), Seoul, Korea, 2009, pp. 2256–2260.
25. P. Massart, *Concentration inequalities and model selection*, Lecture Notes in Mathematics, vol. 1896, Springer-Verlag, Berlin, 2007.
26. M. S. Pinsker and Yu. L. Sagalovich, *Lower bound on the cardinality of code of automata's states*, Problems of Information Transmission **8** (1972), no. 3, 59–66.
27. Yu. L. Sagalovich, *Completely separating systems*, Problems of Information Transmission **18** (1982), no. 2, 74–82.
28. ———, *Separating systems*, Problems of Information Transmission **30** (1994), no. 2, 14–35.
29. T. Sirvent, *Traitor tracing scheme with constant ciphertext rate against powerful pirates*, Workshop on Coding and Cryptography, WCC 2007, http://eprint.iacr.org/2006/383.pdf.
30. J. N. Staddon, D. R. Stinson, and R. Wei, *Combinatorial properties of frameproof and traceability codes*, IEEE Trans. Inform. Theory **47** (2001), 1042–1049.
31. D. R. Stinson, R. Wei, and K. Chen, *On generalized separating hash families*, J. Combin. Theory Ser. A **115** (2008), no. 1, 105–120.
32. G. Tardos, *Optimal probabilistic fingerprint codes*, Journal of the ACM **55** (2008), no. 2, Art. 10, 24pp.
33. J. H. van Lint, *Introduction to coding theory*, 3 ed., Springer-Verlag, Berlin e. a., 1999.
34. N. Wagner, *Fingerprinting*, Proc. 1983 IEEE Symposium on Security and Privacy, 1983, pp. 18–23.
35. R. A. Walker, II and C. J. Colbourn, *Perfect hash families: constructions and existence*, J. Math. Cryptol. **1** (2007), no. 2, 125–150.