**Des. Codes. Cryptogr. (2016) 80:11–28**
Author's accepted manuscript. The final publication is available at
http://link.springer.com/article/10.1007/s10623-015-0060-z

# Almost separating and almost secure frameproof codes over $q$-ary alphabets

**José Moreira · Marcel Fernández · Grigory Kabatiansky**

Received: date / Accepted: date

**Abstract** In this paper we discuss some variations of the notion of separating code for alphabets of arbitrary size. We show how the original definition can be relaxed in two different ways, namely almost separating and almost secure frameproof codes, yielding two different concepts. The new definitions enable us to obtain codes of higher rate, at the expense of satisfying the separating property partially. These new definitions become useful when complete separation is only required with high probability, rather than unconditionally. We

J. Moreira and M. Fernández have been supported in part by the Spanish Government through projects Consolider-Ingenio 2010 CSD2007-00004 "ARES" and TEC2011-26491 "COPPI", and by the Catalan Government through grant 2014 SGR-1504. G. Kabatiansky has been supported in part by the Russian Foundation for Basic Research through grants RFBR 13-01-12458, RFBR 13-07-00978, and RFBR 14-01-93108. The material in this work was presented in part at the 2011 IEEE International Symposium on Information Theory [10].

J. Moreira
Department of Network Engineering, Universitat Politècnica de Catalunya (UPC), C. Jordi Girona 1-3, Building C3, 08034 Barcelona, Spain
E-mail: jose.moreira@entel.upc.edu
*and*
SCYTL Secure Electronic Voting, Pl. Gal.la Placídia 1-3, 1st floor, 08006 Barcelona, Spain
E-mail: jose.moreira@scytl.com

M. Fernández
Department of Network Engineering, Universitat Politècnica de Catalunya (UPC), C. Jordi Girona 1-3, Building C3, 08034 Barcelona, Spain
E-mail: marcel@entel.upc.edu

G. Kabatiansky
National Research University Higher School of Economics (HSE), Myasnitskaya ul. 20, Moscow 101000, Russia
*and*
Institute for Information Transmission Problems, Russian Academy of Sciences, Bol'shoi Karetnyi per. 19, Moscow 127051, Russia
E-mail: kaba@iitp.ru

also show how the codes proposed can be used to improve the rate of existing constructions of families of fingerprinting codes.

**Keywords** Separating code · Secure frameproof code · Fingerprinting · Traitor tracing

**Mathematics Subject Classification (2000)** 94B60 · 94B65

## 1 Introduction

In this paper we shall proceed to relax the notion of separating code to derive two new families of codes, which we will prove to be useful in the field of digital fingerprinting.

Separating codes have been known for some decades now, being discussed for the first time by Friedman et al. [12] in the late 1960s. In those days, one of the original applications that motivated the study of separating codes was the need of an encoding scheme for the inner stable states of asynchronous automata. These states can be stored in $n$ binary memory cells, and due to their asynchronous nature, the content of those cells are allowed to change nonuniformly in the transition between two stable states. Let $a$, $a'$, $b$ and $b'$ be the binary representations of four different stable states. A critical race occurs if in the transition from $a$ to $a'$, and in the transition from $b$ to $b'$ the same intermediate state appears. Critical races are undesirable, since from that common intermediate state it is not possible to determine if the final stable state of the transition is $a'$ or $b'$. Hence, one would like an encoding for the stable states so that any pair of "disjoint" transitions (i.e., having different initial and final stable states) shares a common intermediate state. It has been shown that separating codes provide an appropriate encoding that avoid the rise of these common intermediate states.

Codes with the separating property have also proved to be useful in many other areas, such as technical diagnosis, construction of hash functions, Internet routing, and even in genetics. Such codes have been subsequently investigated by many authors, obtaining lower and upper bounds on the code rate, and establishing connections with similar concepts. See for instance the overviews [18] and [7].

With the advent of digital fingerprinting [3, 4], the interest in separating codes reemerged once again. The concept of a digital fingerprinting scheme is that of traitor tracing [5, 6] applied to the distribution of digital contents. In this new area of application, separating codes are also known under the names of frameproof and secure frameproof codes [19, 20].

A fingerprinting scheme is a cryptographic technique that enables the identification of the source of leaked information. In a fingerprinting scheme, a distributor delivers copies of a given content to a set of authorized users. If there are dishonest members (traitors) among them, the distributor can deter plain redistribution of the content by delivering a marked copy to each user. The set of all user marks is known as a fingerprinting code. There is, however, another

threat. If several traitors collude to create a combination of their copies, then the pirated copy that they generate will contain a corrupted mark, which may obstruct the identification of the traitors.

As stated above, the main result of this paper is the fact that relaxing the notion of separating code shall lead us to two different families of codes, coined as almost separating and almost secure frameproof codes. When absolute separation is not strictly necessary, then these two relaxations of a separating code allows us to construct codes with better rate. We will derive existence bounds for such new codes in the case of $q$-ary alphabets. Moreover, we will prove them to be useful to construct families of binary fingerprinting codes with efficient identification algorithms and exponentially small identification error.

The structure of the paper is as follows. In Section 2 we introduce the topic and present some previous results. In Section 3 and Section 4, we obtain existence bounds on the rate of the relaxed versions of separating codes that we are introducing. Next, in Section 5 we compare the obtained results with the current known state of the art. Our motivation for studying almost separating and almost secure frameproof codes is their application to fingerprinting schemes. In Section 6, we construct a family of fingerprinting codes with small error using almost separating and almost secure frameproof codes. Finally, the conclusions are drawn in Section 7.

## 2 Definitions and previous results

We begin our discussion by recalling some coding-theoretic definitions and some useful bounds from probability theory.

Let $Q$ be an alphabet of size $q$, i.e., a nonempty set of size $q$. When $Q$ is the finite field of $q$ elements, we denote it by $\mathbb{F}_q$. Also, let $Q^n$ be the set of all possible $n$-tuples over an alphabet $Q$. We denote the elements of $Q^n$ in boldface, e.g., $\mathbf{u} = (u_1, \ldots, u_n)$. The *(Hamming) distance* between two elements $\mathbf{u}, \mathbf{v} \in Q^n$ is denoted $d(\mathbf{u}, \mathbf{v})$.

A $q$-ary $(n, M)$-*code* is a subset $C \subseteq Q^n$ of size $M$, where $|Q| = q$. A code $C$ is a *$q$-ary linear $[n, k]$-code* if $C$ is a vector subspace of $\mathbb{F}_q^n$ of dimension $k$. The elements of a code $C$ are called *codewords*. A code $C$ has *minimum distance* $d$ if $d$ is the smallest distance between any two of its codewords. The *rate* of a $q$-ary $(n, M)$-code $C$, denoted $R(C)$, is defined as $R(C) = n^{-1} \log_q M$.

Informally in our discussion, we refer to a *random $(n, M)$-code* $C$ over a $q$-ary alphabet $Q$ as the result of the experiment consisting in choosing $M$ vectors uniformly and independently from $Q^n$. That is, we generate $M$ codewords $(u_1, \ldots, u_n) \subseteq Q^n$ where each $u_i$ is chosen from $Q$ independently with probability $1/q$.

*Remark 1* A clarification is probably in order at this point. Observe that in the definition of a random code that we have just introduced, the codewords are chosen with replacement, i.e., the resulting code should be regarded as a

multiset, rather than a set. Apparently, this may seem unsuitable, since the repeated codewords of a multiset should not be taken into account in the computation of the code rate. However, in our standard probabilistic analysis below, we will be interested in the occurrence of certain significant events in a random code. Namely, we will use the probability of occurrence of such significant events to show that, if they are subject to certain restrictions, then a code with some desired properties exists.

By convention in probability theory, we shall use the abbreviations "r.v." and "pmf" to denote *random variable* and *probability mass function*, respectively. Moreover, we will have several occasions to use the following well-known results. Let $X_1, \ldots, X_n$ be $n$ independent indicator r.v.'s, i.e., taking on values in $\{0, 1\}$. Also, let $X = \sum_1^n X_i$, and $p = EX/n$, where $EX$ denotes the expected value of $X$. In other words, $X$ counts the number of successes in $n$ trials with average probability of success $p$. Then, the probabilities of the tails can be bounded as

$$\Pr\{X/n \geq p + \delta\} \overset{\text{(a)}}{\leq} 2^{-nD(p+\delta\|p)} \overset{\text{(b)}}{\leq} e^{-2n\delta^2}, \qquad \text{for } \delta > 0, \tag{1}$$

and

$$\Pr\{X/n \leq p - \delta\} \overset{\text{(a)}}{\leq} 2^{-nD(p-\delta\|p)} \overset{\text{(b)}}{\leq} e^{-2n\delta^2}, \qquad \text{for } 0 < \delta < p. \tag{2}$$

Here, $D(x\|y)$ denotes the Kullback-Leibler divergence between two Bernoulli distributed r.v.'s of parameters $x$ and $y$, respectively,

$$D(x\|y) \overset{\text{def}}{=} x \log_2(x/y) + (1-x) \log_2((1-x)/(1-y)).$$

Inequalities (a) in (1) and (2) are known as the *Chernoff bounds*, and inequalities (b) are a special case of the *Hoeffding bounds* [14]. Observe that $D(x\|y) \geq 0$, and $D(x\|y) = 0$ if and only if $x = y$. Moreover,

$$2^{-nD(p+\delta\|p')} \leq 2^{-nD(p+\delta\|p)}, \qquad \text{for } p' \leq p, \tag{3}$$

since $D(x\|y)$ is strictly decreasing in the interval $0 \leq y \leq x$. Finally, note that bounds (1) and (2) hold when $X$ is a binomial r.v. of parameters $n$ and $p$.

### 2.1 Separating and secure frameproof codes

Let $C$ be a code over a $q$-ary alphabet $Q$. We call a subset of $c$ codewords $U = \{\mathbf{u}^1, \ldots, \mathbf{u}^c\} \subseteq C$ a *c-subset* or a *c-coalition*. For a $c$-subset $U$, denote by $P_i(U)$ the *projection* of $U$ on the $i$th position, i.e., the set of elements of the code alphabet at the $i$th position,

$$P_i(U) \overset{\text{def}}{=} \{u_i^1, \ldots, u_i^c\}. \tag{4}$$

For a pair of (disjoint) subsets $U, V \subseteq C$, a position $1 \leq i \leq n$ is *separating* if $P_i(U)$ and $P_i(V)$ have empty intersection. The pair of subsets $U, V$ are *separated* if there exists a separating position for them.

**Definition 1** A code $C$ is $(c, c')$-*separating* if every pair of disjoint subsets $U, V \subseteq C$, with $|U| = c$ and $|V| = c'$, are separated.

After its introduction in the work by Friedman et al. in [12], the separating property has been investigated by many authors, e.g. in [7,8,15,16,17,18].

In connection with digital fingerprinting, we say that a position $i$ is *undetectable* for a $c$-coalition $U$ if the codewords of $U$ match in their $i$th position, that is, $u_i^1 = \cdots = u_i^c$, or equivalently, $|P_i(U)| = 1$. A position that does not satisfy this property is called *detectable*.

In the field of digital fingerprinting, the *marking assumption* [3,4] states that when a $c$-coalition $U$ generates a forged copy of the content, the undetectable positions remain unchanged in the pirated word. For the detectable positions, the traitors are allowed to alter them in some way, possibly making them unreadable.

We will restrict our study to the so-called *narrow-sense envelope model* [1]. That is, we consider that the set of pirated words that a $c$-coalition $U$ can generate, denoted $\mathrm{desc}(U)$, is

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

For the binary case, and from the distributor's perspective, the study of fingerprinting codes under the narrow-sense envelope model yields the same results as the study under other envelope models, as it was shown in [1].

A pirated word $\mathbf{z} \in \mathrm{desc}(U)$ is also known as a *descendant*, and the codewords from $U$ are also called the *parents* of $\mathbf{z}$. Additionally, for a code $C$, we define the *$c$-descendant code* under the narrow-sense envelope model, denoted $\mathrm{desc}_c(C)$, as

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

Thus, $\mathrm{desc}_c(C)$ contains all vectors that can be generated by any $c$-coalition.

**Definition 2** A code $C$ is *$c$-secure frameproof* if for any $U, V \subseteq C$ with $|U| \leq c$, $|V| \leq c$ and $U \cap V = \emptyset$, we have $\mathrm{desc}(U) \cap \mathrm{desc}(V) = \emptyset$.

The concept of frameproof code was introduced in [19,20]. It is easy to see that a $c$-secure frameproof code is the same as a $(c, c)$-separating code. Also, a code $C$ such that for any $c$-subset $U \subseteq C$ satisfies $\mathrm{desc}(U) \cap C = U$ is called a *$c$-frameproof code*, which is the same as a $(c, 1)$-separating code.

Let $R_q^{\mathrm{sep}}(n, c, c')$ denote the rate of a maximal $q$-ary $(c, c')$-separating code of length $n$, i.e.,

$$R_q^{\mathrm{sep}}(n, c, c') \stackrel{\text{def}}{=} \max_{\substack{C \subseteq Q^n \text{ s.t. } C \text{ is} \\ (c, c')\text{-separating}}} R(C), \qquad \text{where } |Q| = q.$$

Also, consider the corresponding asymptotical rates

$$\underline{R}_q^{\mathrm{sep}}(c, c') \stackrel{\text{def}}{=} \liminf_{n \to \infty} R_q(n, c, c'), \qquad \overline{R}_q^{\mathrm{sep}}(c, c') \stackrel{\text{def}}{=} \limsup_{n \to \infty} R_q(n, c, c').$$

Lower bounds on the rate of $(2,2)$-separating codes were studied in [12,16]. Some important, well-known results for binary separating codes are worth mentioning. For example, from [16,18], we have $\underline{R}_2^{\mathrm{sep}}(2,2) \geq 1 - \log_2(7/8) = 0.0642$, which also holds for linear codes [16]. Also, for the general case, it was shown in [1] that

$$\underline{R}_2^{\mathrm{sep}}(c,c') \geq -\frac{\log_2(1-2^{-c-c'+1})}{c+c'-1}. \tag{5}$$

Regarding the upper bounds, it is known that $\overline{R}_2^{\mathrm{sep}}(2,2) < 0.2835$ for arbitrary codes [15,18], and $\overline{R}_2^{\mathrm{sep}}(2,2) < 0.108$ for linear codes [18].

## 3 Separating and almost separating codes over $q$-ary alphabets

We start the study of separating and almost separating codes by obtaining lower bounds for separating codes over arbitrary alphabets. This will allow us to compare these results with the concepts of almost separating and almost secure frameproof codes that we will introduce below. We will use a standard probabilistic argument to obtain a generalization of (5).

### 3.1 Lower bounds on the rate of separating codes

Let us begin our discussion with the following two lemmas.

**Lemma 1** *Let $v(j;q,c)$ be the pmf at $j$ of an r.v. that counts the number of different symbols of a $q$-ary vector of length $c$ chosen uniformly at random. Then,*

$$v(j;q,c) \stackrel{\mathrm{def}}{=} \frac{q^{\underline{j}}}{q^c} \left\{ {c \atop j} \right\}, \qquad 1 \leq j \leq \min\{q,c\}, \tag{6}$$

*where $q^{\underline{j}} \stackrel{\mathrm{def}}{=} q(q-1)\cdots(q-j+1)$ denotes the falling factorial and $\left\{ {c \atop j} \right\}$ denotes the Stirling number of the second kind.*

*Proof* A set of size $c$ can be partitioned into $j$ nonempty subsets in $\left\{ {c \atop j} \right\}$ different ways. For each such partition there are exactly $q(q-1)\cdots(q-j+1) = q^{\underline{j}}$ possible assignments using $j$ different elements from a $q$-ary alphabet. The product of these two terms gives the number of $q$-ary vectors of length $c$ that contain exactly $j$ different symbols. The proof follows after dividing by the total number of vectors of length $c$. □

For convenience, we shall say that two $q$-ary vectors are *disjoint* if they have no common element.

**Lemma 2** *Let $p^{\text{disj.}}_{q,c,c'}$ be the probability that two $q$-ary vectors of lengths $c$ and $c'$, respectively, chosen uniformly and independently at random are disjoint. Then, we have*

$$p^{\text{disj.}}_{q,c,c'} = \sum_{j=1}^{\min\{q,c\}} (1-j/q)^{c'}\, v(j;q,c).$$

*Proof* Let $\mathbf{a} = (a_1,\ldots,a_c)$ and $\mathbf{b} = (b_1,\ldots,b_{c'})$ be two random vectors, of length $c$ and $c'$, respectively, and let $X$ be the r.v. that counts the number of different symbols in $\mathbf{a}$. The probability that $\mathbf{a}$ and $\mathbf{b}$ are disjoint, i.e., $\{a_1,\ldots,a_c\} \cap \{b_1,\ldots,b_{c'}\} = \emptyset$, can be computed as

$$p^{\text{disj.}}_{q,c,c'} = \sum_j \Pr\{\mathbf{a} \text{ and } \mathbf{b} \text{ disjoint} \,|\, X = j\} \Pr\{X = j\}.$$

Clearly, $\Pr\{X = j\} = v(j;q,c)$. Also, since $\mathbf{b}$ has $c'$ elements, independently chosen from $\mathbf{a}$, we have $\Pr\{\mathbf{a} \text{ and } \mathbf{b} \text{ disjoint} \,|\, X = j\} = (1-j/q)^{c'}$. □

Now, consider two disjoint subsets $U, V$ of a random $(n, M)$-code $C$ over a $q$-ary alphabet $Q$, with $|U| = c$ and $|V| = c'$. According to our definition of a random code, these subsets can be regarded as choosing uniformly and independently (with replacement) $c$ and $c'$ codewords from $Q^n$. Hence, the probability that a position $1 \le i \le n$ is separating, i.e., $P_i(U) \cap P_i(V) = \emptyset$ is precisely $p^{\text{disj.}}_{q,c,c'}$. Using this fact, combined with the probabilistic argument borrowed from [1, Proposition 3.4], it is easy to see the following result. We provide the proof below for completeness.

**Corollary 1** *There exist $q$-ary $(c,c')$-separating codes of asymptotical rate satisfying*

$$\underline{R}^{\text{sep}}_q(c,c') \ge -\frac{\log_q(1-p^{\text{disj.}}_{q,c,c'})}{c+c'-1}.$$

*Proof* Let $C$ be a random $q$-ary $(n, M)$-code, and let $E$ be the expected number of "bad" pairs $U, V$ of subsets with $|U| = c$ and $|V| = c'$, i.e., pairs that are not separated. If $E < M/2$, then a $q$-ary $(n, M/2)$-code with the $(c,c')$-separating property exists, since by removing one codeword from each bad pair, the remaining codewords yield a $(c,c')$-separating code. The probability that a pair $U, V$ of such subsets is not separated is $(1-p^{\text{disj.}}_{q,c,c'})^n$. Hence, we have

$$E \le \binom{M}{c}\binom{M-c}{c'}(1-p^{\text{disj.}}_{q,c,c'})^n < \frac{M^{c+c'}}{c!\,c'!}(1-p^{\text{disj.}}_{q,c,c'})^n.$$

Observe that taking $M = \left(\frac{c!c'!}{2}(1-p^{\text{disj.}}_{q,c,c'})^{-n}\right)^{1/(c+c'-1)}$, we have $E < M/2$. Finally, since $(\frac{c!c'!}{2})^{1/(c+c'-1)} \ge 1$, we can disregard the logarithm of this term in the lower bound on the code rate. □

### 3.2 Lower bounds on the rate of almost separating codes

The separating property imposes a very strict combinatorial restriction to the code, namely that every pair of $c$-subsets is separated. One could obtain codes with better rates by relaxing this condition, and asking for codes where it is satisfied with high probability, rather than in all cases. In this section we elaborate our first proposal to relax the separating condition.

Let us introduce some useful definitions. For a code $C$, we say that a $c$-subset $U \subseteq C$ is *separated* if $U$ is separated from any other disjoint $c$-subset $V \subseteq C$.

**Definition 3** A code $C \subseteq Q^n$ is $\varepsilon$-*almost* $(c, c)$-*separating* if the proportion of $c$-subsets that are separated is at least $1 - \varepsilon$.

A sequence of codes $\mathcal{C} = (C_i)_{i \geq 1}$ of growing length $n_i$ is an *asymptotically almost* $(c, c)$-*separating family* if every code $C_i$ is $\varepsilon_i$-almost $(c, c)$-separating and $\lim_{i \to \infty} \varepsilon_i = 0$.

We also define the asymptotical rate of a sequence $\mathcal{C} = (C_i)_{i \geq 1}$ as

$$R(\mathcal{C}) = \liminf_{i \to \infty} R(C_i). \tag{7}$$

We are interested in estimating the maximal possible asymptotical rate, denoted $R_q^{\text{sep}^*}(c)$, among all asymptotically almost $(c, c)$-separating families of codes.

To derive lower bounds, we make use of a restricted version of strongly typical subsets of codewords [9]. That is, subsets of codewords that appear with high probability in a random code.

Let $C$ be a $q$-ary $(n, M)$-code, and let $U \subseteq C$ be a $c$-subset. We say that a position $i$ is $j$-*valued* if its projection $P_i(U)$ contains exactly $j$ different symbols from the code alphabet. We denote $N(j; U)$, for $1 \leq j \leq \min\{q, c\}$, the number of positions $1 \leq i \leq n$ that are $j$-valued. For example, if $Q = \{0, 1, 2\}$ and

$$\begin{aligned} U = \{\ & (2,1,0,0,2,0,0,1,0,0,0,2,1,0,2), \\ & (1,1,1,1,1,0,0,2,1,0,0,0,0,0,2), \\ & (1,2,2,0,1,0,1,0,0,0,1,1,0,2,0), \\ & (2,1,1,1,1,1,0,2,0,0,2,1,1,1,2)\ \}, \end{aligned}$$

then $N(1; U) = 1$, $N(2; U) = 9$, $N(3; U) = 5$ and $N(j; U) = 0$ otherwise. Note that for a $c$-subset $U$, uniformly chosen from a random $q$-ary $(n, M)$-code, the empirical distribution $N(j; U)/n$ satisfies

$$E\, N(j; U)/n = v(j; q, c), \qquad 1 \leq j \leq \min\{q, c\}.$$

**Definition 4** Let $0 < \delta \leq 1$. For a $q$-ary $(n, M)$-code $C$, we say that a given $c$-subset $U \subseteq C$ is $\delta$-*typical* if the empirical distribution of the number of $j$-valued positions, i.e., $N(j; U)/n$, is "close" to the expected value $v(j; q, c)$ of a $c$-subset in a random code. Namely,

$$|N(j; U)/n - v(j; q, c)| < \delta, \qquad 1 \leq j \leq \min\{q, c\}. \tag{8}$$

The set of all $\delta$-typical $c$-subsets of a random code is denoted by $A_\delta^{(n)}(q,c)$,

$$A_\delta^{(n)}(q,c) \stackrel{\text{def}}{=} \{U \subseteq C : |U| = c \text{ and } U \text{ is } \delta\text{-typical}\}.$$

Note that if $U$ is a uniformly chosen $c$-subset from a random code, each $N(j;U)$ can be regarded as a binomial r.v. of parameters $n$ and $v(j) = v(j;q,c)$. Then, combining the union bound with (1) and (2), it is not difficult to see that the probability that $U$ is not contained in the typical set satisfies

$$\Pr\{U \notin A_\delta^{(n)}(q,c)\} \leq \sum_{j=1}^{\min\{q,c\}} 2^{-nD(v(j)-\delta\|v(j))} + 2^{-nD(v(j)+\delta\|v(j))} \leq 2q\, e^{-2n\delta^2}. \quad (9)$$

With these results in mind, we are ready to derive a lower bound for almost separating $q$-ary codes. The key idea in our probabilistic approach below is to bound the probability that a given $c$-subset is separated (from any other disjoint $c$-subset) in a random code, and derive the conditions on the code rate for which this probability vanishes as the code length increases.

**Theorem 1** *For the maximal possible asymptotical rate $R_q^{\text{sep}^*}(c)$ among all asymptotically almost $(c,c)$-separating families of $q$-ary codes we have*

$$R_q^{\text{sep}^*}(c) \geq -\frac{1}{c} \sum_{j=1}^{\min\{q,c\}} \log_q(1-(1-j/q)^c)\, v(j;q,c).$$

*Proof* Consider a random $(n,M)$-code $C$ over a $q$-ary alphabet. For a given $c$-subset $U \subseteq C$ there are exactly $N(j;U)$ $j$-valued positions. For each such position $i$, the probability that another random disjoint $c$-subset $V$ satisfies $P_i(U) \cap P_i(V) = \emptyset$ equals $(1-j/q)^c$. Thus,

$$\Pr\{U \text{ and } V \text{ are not separated}\} = \prod_j (1-(1-j/q)^c)^{N(j;U)}.$$

Now, using (8) and (9), the probability $\varepsilon$ that $U$ is not separated can be upper bounded as follows:

$$\begin{aligned}
\varepsilon &= \Pr\{U \text{ is not separated} \,|\, U \text{ is typical}\} \Pr\{U \text{ is typical}\} + \\
&\qquad \Pr\{U \text{ is not separated} \,|\, U \text{ is not typical}\} \Pr\{U \text{ is not typical}\} \\
&\leq \Pr\{U \text{ is not separated} \,|\, U \text{ is typical}\} + \Pr\{U \text{ is not typical}\} \\
&\leq \binom{M-c}{c} \prod_j (1-(1-j/q)^c)^{n(v(j;q,c)-\delta)} + 2q\, e^{-2n\delta^2}.
\end{aligned}$$

Observe that taking $\delta = \delta(n) = \ln n/\sqrt{n}$, then $\lim_{n\to\infty} \varepsilon \leq q^{nA}$, where

$$A = c\, R(C) + \sum_j \log_q(1-(1-j/q)^c)\, v(j;q,c).$$

Now take a sequence of codes $\mathcal{C} = (C_i)_{i\geq 1}$ of growing length, such that each $(n_i, M_i)$-code $C_i$ is a random code. The probabilistic argument above shows that, taking an appropriate value for $\delta_i$, for example $\delta_i = \delta_i(n_i) = \ln n_i/\sqrt{n_i}$, there exists a sequence of codes with $\lim_{i\to\infty} \varepsilon_i = 0$ for any rate

$$R < -\frac{1}{c} \sum_j \log_q(1 - (1 - j/q)^c)\, v(j; q, c),$$

which completes the proof. □

## 4 Almost secure frameproof codes

In this section we relax the definition of secure frameproof (i.e., separating) code, again, in order to obtain codes with better rates. The notion that we introduce here allows us to separate the concepts of almost separating and almost secure frameproof codes.

Let us call a vector $\mathbf{z} \in \mathrm{desc}_c(C)$ *c-uniquely decodable* if $\mathbf{z} \in \mathrm{desc}(U)$ for some $c$-subset $U \subseteq C$ and $\mathbf{z} \notin \mathrm{desc}(V)$ for any $c$-subset $V \subseteq C$ such that $U \cap V = \emptyset$. Note that the $c$-secure frameproof codes from Definition 2 can be regarded as codes where all vectors $\mathbf{z} \in \mathrm{desc}_c(C)$ are $c$-uniquely decodable. This alternate definition allows us to introduce the following concept.

**Definition 5** A code $C \subseteq Q^n$ is *$\varepsilon$-almost c-secure frameproof* if the proportion of $c$-uniquely decodable vectors among all $\mathbf{z} \in \mathrm{desc}_c(C)$ is at least $1 - \varepsilon$.

A sequence of codes $\mathcal{C} = (C_i)_{i\geq 1}$ of growing length $n_i$ is an *asymptotically almost c-secure frameproof family* if every code $C_i$ is $\varepsilon_i$-almost $c$-secure frameproof and $\lim_{i\to\infty} \varepsilon_i = 0$.

Consider again the asymptotical rate of a sequence of codes (7). As above, we are interested in estimating the maximal possible asymptotical rate, denoted $R_q^{\mathrm{sf}^*}(c)$, among all asymptotically almost $c$-secure frameproof families. Similarly, as in Theorem 1, our approach will be to bound the probability that a given descendant is $c$-uniquely decodable in a random code, and derive conditions on the code rate for which this probability goes to 0 as the code length goes to infinity.

**Theorem 2** *For the maximal possible asymptotical rate $R_q^{\mathrm{sf}^*}(c)$ among all asymptotically almost c-secure frameproof families of q-ary codes we have*

$$R_q^{\mathrm{sf}^*}(c) \geq -\frac{1}{c} \log_q(1 - (1 - 1/q)^c).$$

*Proof* Consider a random $(n, M)$-code $C$ over a $q$-ary alphabet. Also, consider a vector $\mathbf{z} = (z_1, \ldots, z_n)$ which is generated by a $c$-coalition $U \subseteq C$. For a random $c$-coalition $V \subseteq C$ such that $U \cap V = \emptyset$, using Lemma 2, we have

$$\Pr\{\mathbf{z} \in \mathrm{desc}(V)\} = (1 - p_{q,c,1}^{\mathrm{disj.}})^n = (1 - (1 - 1/q)^c)^n.$$

Certainly, there are $n$ positions, and the probability that each position $i$ satisfies $y_i \notin P_i(V)$ equals $p^{\text{disj.}}_{q,c,1}$, because $V$ is a random, independent coalition from $U$. Therefore, the probability that a given vector $\mathbf{z} \in \text{desc}_c(C)$ is not $c$-uniquely decodable is at most

$$\varepsilon \le \binom{M-c}{c}(1-p^{\text{disj.}}_{q,c,1})^n \le M^c(1-p^{\text{disj.}}_{q,c,1})^n.$$

Hence, there is a sequence $\mathcal{C} = (C_i)_{i\ge 1}$ of growing length $n_i$ such that for each $(n_i, M_i)$-code $C_i$ the proportion of $c$-uniquely decodable vectors among all $\mathbf{z} \in \text{desc}_c(C_i)$ is at least $1-\varepsilon_i \ge 1 - M_i^c(1-p^{\text{disj.}}_{q,c,1})^{n_i}$. Taking $M_i = o((1-p^{\text{disj.}}_{q,c,1})^{-n_i/c})$, i.e.,

$$R(C_i) < -\frac{1}{c}\log_q(1-(1-1/q)^c),$$

we have $\lim_{i\to\infty}\varepsilon_i = 0$, and the proof follows. □

*Remark 2* If $C \subseteq Q^n$ is an $\varepsilon$-almost $c$-secure frameproof code, then for the family of codes $\varphi(C)$, where $\varphi$ runs over the group $G$ of all isometries of the Hamming space $Q^n$, the probability that any given vector $\mathbf{z}$ can be generated by two disjoint coalitions is at most $\varepsilon$ (since the group $G$ is twice transitive). This property allows us to replace the $(c,c)$-separating codes in the main construction of fingerprinting codes from [1] with asymptotically almost $c$-secure frameproof families, what will result in larger code rate with the same polynomial complexity identification algorithm. See Section 6 below.

*Remark 3* For the case of a family of codes (instead of a single code) we can say "probability" instead of "proportion." Namely, for *every* "received" vector $\mathbf{z}$ the probability (i.e., the "proportion" of codes) that there exist at least two different $c$-coalitions $U, V$ of codewords which can generate $\mathbf{z}$, is at most $\varepsilon$. Then, of course, for $c = 2$ the lower bound on the code rate is the same and it also follows from [2].

### 4.1 Geometric interpretation

For an $(n, M)$-code $C$ over a $q$-ary alphabet $Q$, consider the set of convex combinations between two vectors $\mathbf{u}, \mathbf{v} \in C$ as

$$\{\mathbf{z} \in Q^n : d(\mathbf{u},\mathbf{z}) + d(\mathbf{z},\mathbf{v}) = d(\mathbf{u},\mathbf{v})\}. \tag{10}$$

Note that for a $c$-subset $U \subseteq C$, its *convex hull* $[U] \subseteq Q^n$, i.e., the smallest set containing all convex combinations between any two of its elements, is precisely the envelope under the narrow-sense model, $\text{desc}(U)$. Therefore, for the case $c = 2$ and $U = \{\mathbf{u}, \mathbf{v}\} \subseteq C$, equation (10) suggests calling the set $[\{\mathbf{u}, \mathbf{v}\}]$ a *segment* of $C$ with *vertices* $\mathbf{u}$ and $\mathbf{v}$. For $c = 3$ and a 3-coalition $U \subseteq C$, the set $[U]$ could be called a (convex) *polygon*, and so on. For arbitrary $c$, let us call $[U]$ a (convex) *$c$-polytope* with vertices in $C$.

Hence, a $c$-secure frameproof code, or, what is the same, a $(c, c)$-separating code, can be regarded as a set of points $C$ in the $q$-ary Hamming space $Q^n$ with the property that any two $c$-polytopes $[U]$, $[V]$ with vertices in $C$ do not intersect, provided that they do not share a common vertex from $C$.

For a random $q$-ary code $C$, consider the union $C^{[c]}$ of all points generated from $c$-polytopes $[U]$ such that $U \subseteq C$, as in the proof of Theorem 2. In other words, $C^{[c]} = \mathrm{desc}_c(C)$. For a given $\mathbf{z} \in Q^n$ and a random $c$-subset $V \subseteq C$, let us call

$$g(n) = \Pr\{\mathbf{z} \in [V]\} = \Pr\{\mathbf{z} \in \mathrm{desc}(V)\} = (1 - p^{\mathrm{disj.}}_{q,c,1})^n,$$

which follows from the proof of Theorem 2 above. Hence, the size of $C^{[c]}$ can be estimated as

$$|C^{[c]}| = \sum_{\mathbf{z} \in Q^n} \Pr\{\mathbf{z} \in C^{[c]}\} = q^n \Pr\{\mathbf{z} \in C^{[c]}\} = q^n (1 - (1 - g(n))^{\binom{M}{c}}). \quad (11)$$

Now, let us define the *volume* of $C^{[c]}$ by counting every point $\mathbf{z} \in C^{[c]}$ with its multiplicity, i.e., the number of $c$-polytopes that contain $\mathbf{z}$. Using (6), and calling $v(j) = v(j; q, c)$, we have

$$|C^{[c]}| \le \mathrm{vol}(C^{[c]}) = \binom{M}{c} \Big( \sum_j j\, v(j) \Big)^n = \binom{M}{c} q^n\, g(n). \quad (12)$$

This result can be obtained in two different but equivalent ways. Indeed, there are $\binom{M}{c}$ $c$-polytopes, and the probability that each $\mathbf{z} \in Q^n$ is generated by a given $c$-polytope $[U]$ is $g(n)$. Alternatively, the average number of points generated by every $c$-polytope $[U]$ can be computed as

$$\begin{aligned}
&\sum_{j_1=1}^{c} \cdots \sum_{j_n=1}^{c} j_1 \cdots j_n \Pr\{|P_1(U)| = j_1, \ldots, |P_n(U)| = j_n\} \\
&\quad = \sum_{j_1} \cdots \sum_{j_n} j_1 \cdots j_n\, v(j_1) \cdots v(j_n) = \Big( \sum_j j\, v(j) \Big)^n = \Big( q^{-c} \sum_j j\, q^{\underline{j}} \left\{ {c \atop j} \right\} \Big)^n \\
&\quad \overset{(a)}{=} \Big( q^{-c} \sum_j q(q^{\underline{j}} - (q-1)^{\underline{j}}) \left\{ {c \atop j} \right\} \Big)^n \overset{(b)}{=} (q^{-c+1}(q^c - (q-1)^c))^n = q^n\, g(n).
\end{aligned}$$

Here, (a) is obtained by routine algebraic manipulation, and (b) follows from the well-known identity $x^c = \sum_j x^{\underline{j}} \left\{ {c \atop j} \right\}$.

Hence, from (11) and (12) two nontrivial observations can be drawn. First, for $M = o(g(n)^{-1/c})$, i.e., $M = o((1 - p^{\mathrm{disj.}}_{q,c,1})^{-n/c})$, as we took in Theorem 2, we have $\lim_{n\to\infty} \mathrm{vol}(C^{[c]})/|Q^n| = 0$, i.e., the volume of $C^{[c]}$ is relatively small compared to the volume of the whole Hamming space. Second, consider the average asymptotical multiplicity of the points from $C^{[c]}$,

$$\lim_{n\to\infty} \frac{\mathrm{vol}(C^{[c]})}{|C^{[c]}|} = \lim_{n\to\infty} \frac{M^c\, g(n)}{1 - (1 - g(n))^{M^c}} = \lim_{n\to\infty} \frac{M^c\, g(n)}{1 - e^{-M^c\, g(n)}}.$$

The last equality follows from the fact that $\lim_{n\to\infty} g(n) = 0$. Taking again $M = o(g(n)^{-1/c})$, it is easy to see that the main part of points from $C^{[c]}$ have multiplicity 1, i.e., covered only once by code polytopes, which is a stronger statement than Theorem 2.

## 5 Comparison of results

In Table 1 we give some figures for the lower bounds on the asymptotical rate of $q$-ary separating, almost separating and almost secure frameproof codes.

It can be seen that for binary codes and $c = 2$ the ratio $R_q^{\mathrm{sep}^*}(c)/\underline{R}_q^{\mathrm{sep}}(c,c)$ is about 1.6, and it approaches to 2 for $c$ growing. That is, the lower bound on the asymptotical rate of almost separating codes roughly doubles the currently known lower bound on the asymptotical rate of ordinary separating codes. This ratio suffers from minor fluctuations, with a slight decay, for $c$ fixed and $q$ growing.

On the other hand, for the case of almost secure frameproof codes, the ratio $R_q^{\mathrm{sf}^*}(c)/\underline{R}_q^{\mathrm{sep}}(c,c)$ starts at about 3.2 for binary codes and $c = 2$. This ratio increases significantly for $c$ growing, and it decreases, at a much slower speed, for $q$ growing. For example, for $c = 5$ and $q = 3$ the lower bound on the asymptotical rate of almost secure frameproof codes is about 80 times the value of the lower bound for the case of ordinary separating codes.

| $q$ | Code | $c=2$ | 3 | 4 | 5 | 10 | 15 |
|---|---|---|---|---|---|---|---|
| | Separating | 6.422E−2 | 9.161E−3 | 1.616E−3 | 3.134E−4 | 1.448E−7 | 9.266E−11 |
| 2 | Almost sep. | 1.038E−1 | 1.605E−2 | 2.910E−3 | 5.725E−4 | 2.753E−7 | 1.792E−10 |
| | Almost s.f. | 2.075E−1 | 6.422E−2 | 2.328E−2 | 9.161E−3 | 1.410E−4 | 2.935E−6 |
| | Separating | 7.625E−2 | 1.080E−2 | 1.796E−3 | 3.191E−4 | 8.433E−8 | 2.997E−11 |
| 3 | Almost sep. | 1.249E−1 | 1.948E−2 | 3.320E−3 | 5.954E−4 | 1.609E−7 | 5.798E−11 |
| | Almost s.f. | 2.675E−1 | 1.066E−1 | 5.008E−2 | 2.571E−2 | 1.592E−3 | 1.387E−4 |
| | Separating | 9.562E−2 | 1.561E−2 | 2.889E−3 | 5.624E−4 | 2.327E−7 | 1.415E−10 |
| 4 | Almost sep. | 1.524E−1 | 2.772E−2 | 5.288E−3 | 1.040E−3 | 4.428E−7 | 2.735E−10 |
| | Almost s.f. | 2.982E−1 | 1.318E−1 | 6.860E−2 | 3.908E−2 | 4.181E−3 | 6.470E−4 |
| | Separating | 1.114E−1 | 2.091E−2 | 4.307E−3 | 9.053E−4 | 4.067E−7 | 2.158E−10 |
| 5 | Almost sep. | 1.744E−1 | 3.671E−2 | 7.853E−3 | 1.674E−3 | 7.741E−7 | 4.173E−10 |
| | Almost s.f. | 3.174E−1 | 1.486E−1 | 8.185E−2 | 4.934E−2 | 7.058E−3 | 1.484E−3 |
| | Separating | 1.549E−1 | 4.329E−2 | 1.350E−2 | 4.201E−3 | 6.568E−6 | 5.615E−9 |
| 10 | Almost sep. | 2.357E−1 | 7.372E−2 | 2.419E−2 | 7.728E−3 | 1.251E−5 | 1.086E−8 |
| | Almost s.f. | 3.606E−1 | 1.890E−1 | 1.159E−1 | 7.755E−2 | 1.862E−2 | 6.675E−3 |
| | Separating | 1.752E−1 | 5.723E−2 | 2.162E−2 | 8.418E−3 | 4.303E−5 | 7.895E−8 |
| 15 | Almost sep. | 2.649E−1 | 9.653E−2 | 3.840E−2 | 1.539E−2 | 8.202E−5 | 1.527E−7 |
| | Almost s.f. | 3.783E−1 | 2.064E−1 | 1.313E−1 | 9.098E−2 | 2.572E−2 | 1.081E−2 |

**Table 1** Lower bounds on the rate of some families of $q$-ary codes

## 6 Application to fingerprinting codes

In this section we show how almost separating and almost secure frameproof codes can be used to construct a family of binary fingerprinting codes.

The contents of this section are inspired by the work presented in [1], where the authors present a construction of a family of binary fingerprinting codes using code concatenation [11]. Their construction uses ordinary separating codes as inner codes, and the authors propose an efficient identification algorithm when the outer code is a Reed-Solomon or an algebraic-geometric code.

Our main goal is to show that replacing the inner codes used in [1] by almost separating or almost secure frameproof codes can yield a family of binary fingerprinting codes of higher rate, with a small probability of error and with an efficient identification algorithm. We will outline the family construction and derive existence conditions. Of course, in our analysis below, we will also show how to deal with the issues implied by the fact that the inner codes are relaxed versions of ordinary separating codes.

### 6.1 Fingerprinting codes

A collusion attack, described in Section 1, is modeled by the generation of a descendant. In this case the descendant is the word in the pirated copy and the parents are the codewords belonging to the colluders.

As it is shown in [1, 3, 4], for a fingerprinting scheme to achieve an error probability as small as desired a single code is not sufficient, but a *family of codes* is needed. Below we denote a family of fingerprinting codes as $\mathcal{C} = \{C_j\}_{j\in T}$, where $T$ is a finite set, and each $C_j$ is an $(m, M)$-code. A fingerprinting scheme also has the need for randomness in the following sense. The family $\mathcal{C} = \{C_j\}_{j\in T}$ is publicly known, however, the specific code $C_j$ used by the distributor is chosen at random with probability $\pi(j)$, and it is kept secret.

For the family of fingerprinting codes $\mathcal{C} = \{C_j\}_{j\in T}$ we also need an *identification algorithm*, which is a collection of functions $\mathcal{A} = \{A_j\}_{j\in T}$, where each $A_j$ is a map from the set of descendants of $C_j$ to the subsets of codewords of $C_j$ of size at most $c$,

$$A_j \colon \mathrm{desc}_c(C_j) \to \{U \subseteq C_j : |U| \leq c\}.$$

**Definition 6** Let $T$ be a finite set and let $\pi$ be a pmf on $T$. For $c \geq 2$, a family of fingerprinting codes $\mathcal{C} = \{C_j\}_{j\in T}$, is *$c$-secure with $\varepsilon$-error* if there exists an identification algorithm $\mathcal{A} = \{A_j\}_{j\in T}$ that satisfies the following condition: if a coalition $U$ of size at most $c$ creates a descendant $\mathbf{z}$, then $A_j(\mathbf{z})$ is not empty, and

$$\Pr(A_j(\mathbf{z}) \subseteq U) > 1 - \varepsilon,$$

where the probability is taken over the random choices made by the coalition when creating the descendant, and over the pmf $\pi$.

### 6.2 Family construction and existence conditions

Let $C_{\text{out}}$ be an $(n, M)$-code over a $q$-ary alphabet $Q$, and let $C_{\text{in}}$ be a binary $(l, q)$-code. Consider a vector mapping $(\phi_1, \ldots, \phi_n)$, where $\phi_i, 1 \leq i \leq n$ are bijections between $Q$ and $C_{\text{in}}$. It is clear that there are $(q!)^n$ different such vector mappings. If the mappings are arbitrarily numbered from 1 to $(q!)^n$, then

$$\Phi_t \stackrel{\text{def}}{=} (\phi_{t1}, \ldots, \phi_{tn}) \tag{13}$$

denotes the mapping indexed by $t$. Now, for $\mathbf{w} = (w_1, \ldots, w_n) \in C_{\text{out}}$, we denote by $\Phi_t(\mathbf{w})$ the following binary vector of length $m = n\,l$:

$$\Phi_t(\mathbf{w}) \stackrel{\text{def}}{=} (\phi_{t1}(w_1), \ldots, \phi_{tn}(w_n)).$$

**Construction 1 (Family of binary concatenated codes)** *Let $C_{\text{out}}$ be an $(n, M)$-code over a $q$-ary alphabet $Q$ (the outer code), and let $C_{\text{in}}$ be a binary $(l, q)$-code (the inner code). Also, let $\Phi_t$ denote the mapping indexed by $t$ as in (13). Denote by $C_t$ the binary $(m, M)$-code, with $m = n\,l$, constructed in the following way:*

$$C_t \stackrel{\text{def}}{=} \{\Phi_t(\mathbf{w}) : \mathbf{w} \in C_{\text{out}}\}.$$

*The set $\mathcal{C} = \{C_t\}_{t \in T}$, with $T = \{1, \ldots, (q!)^n\}$, constitutes the family of binary concatenated codes.*

Taking into account Definition 6, to use the family from Construction 1, $\mathcal{C} = \{C_t\}_{t \in T}$, first we have to choose a value $t \in T$ according to a pmf $\pi(t)$. This value $t$ must be kept secret. Each user is then assigned a codeword from $C_t$, and the copies of the content are delivered correspondingly marked.

Let $U = \{\mathbf{u}^1, \ldots, \mathbf{u}^c\} \subseteq C_t$ be a $c$-coalition, and $W = \{\mathbf{w}^1, \ldots, \mathbf{w}^c\} \subseteq C_{\text{out}}$ be the subset of the corresponding codewords of the outer code. In other words, $\mathbf{u}^j = \Phi_t(\mathbf{w}^j)$ for $1 \leq j \leq c$. Also, let

$$\mathbf{z} = (\mathbf{z}_1, \ldots, \mathbf{z}_n) = (\underbrace{z_{11}, \ldots, z_{1l}}_{\mathbf{z}_1}, \ \ldots, \ \underbrace{z_{n1}, \ldots, z_{nl}}_{\mathbf{z}_n}) \in \text{desc}(U)$$

be a descendant created by coalition $U$.

In the discussion of the identification algorithm, we will consider that the identification process of each inner block $\mathbf{z}_i$ returns a set $V_i \subseteq C_{\text{in}}$ of at most $c$ codewords, such that $\mathbf{z}_i \in \text{desc}(V_i)$. Observe that if the inner code $C_{\text{in}}$ is an $\varepsilon_{\text{in}}$-almost $(c, c)$-separating or an $\varepsilon_{\text{in}}$-almost $c$-secure frameproof code, then with probability $\geq 1 - \varepsilon_{\text{in}}$ there is a $\mathbf{v} \in V_i$ such that $\mathbf{v}$ agrees with the $i$th block of a traitor codeword, i.e., $\mathbf{v} = \phi_{ti}(w_i)$ for some $\mathbf{w} = (w_1, \ldots, w_n) \in W$.

We now state, in the form of a theorem, the precise parameters of the family of codes from Construction 1 so that it can achieve exponentially small error probability when used in conjunction with Algorithm 1.

**Algorithm 1**

*Input:* A concatenated code $C_t$ from Construction 1, satisfying the conditions from Theorem 3, and a descendant $\mathbf{z} \in \mathrm{desc}_c(C_t)$,

$$\mathbf{z} = (\underbrace{z_{11}, \ldots, z_{1l}}_{\mathbf{z}_1}, \ \ldots, \ \underbrace{z_{n1}, \ldots, z_{nl}}_{\mathbf{z}_n}).$$

*Output:* A subset of codewords of $C_t$.

1. For $1 \leq i \leq n$, decode each block $\mathbf{z}_i = (z_{i1}, \ldots, z_{il})$ of the the descendant $\mathbf{z}$ as follows:
   (a) Find all $c$-subsets $V \subseteq C_{\mathrm{in}}$ such that $\mathbf{z}_i \in \mathrm{desc}(V)$.
   (b) If the intersection of all $c$-subsets $V$ found in Step 1a) is empty, set $\mathcal{Z}_i = \emptyset$.
   (c) Otherwise, pick an arbitrary $c$-subset $V$ from Step 1a) and use the inverse mapping
   $$\phi_{ti}^{-1} \colon C_{\mathrm{in}} \to Q$$
   to obtain a set $\mathcal{Z}_i$ of $c$ symbols from $Q$.
2. Construct the vector of sets
   $$\boldsymbol{\mathcal{Z}} := (\mathcal{Z}_1, \ldots, \mathcal{Z}_n).$$
3. For each $\mathbf{w} \in C_{\mathrm{out}}$, define the similitude $\mathrm{s}(\mathbf{w}, \boldsymbol{\mathcal{Z}})$, as
   $$\mathrm{s}(\mathbf{u}, \boldsymbol{\mathcal{Z}}) \stackrel{\mathrm{def}}{=} |\{i : u_i \in \mathcal{Z}_i, 1 \leq i \leq n\}|.$$
4. Output the set $L := \{\mathbf{u}^1, \ldots, \mathbf{u}^s\}$, consisting of all codewords $\mathbf{u} = \Phi_t(\mathbf{w}) \in C_t$, such that
   $$\mathrm{s}(\mathbf{w}, \boldsymbol{\mathcal{Z}}) \geq n\frac{1-\sigma}{c},$$
   for some codeword $\mathbf{w} \in C_{\mathrm{out}}$. If $L = \emptyset$, declare identification error.

**Theorem 3** *Let $C_{\mathrm{out}}$ be an $(n, M)$-code over a $q$-ary alphabet $Q$, with minimum distance $d$, and let $C_{\mathrm{in}}$ be a binary $\varepsilon_{\mathrm{in}}$-almost $(c, c)$-separating or $\varepsilon_{\mathrm{in}}$-almost $c$-secure frameproof $(l, q)$-code. Let $\mathcal{C} = \{C_t\}_{t \in T}$ be the family of codes from Construction 1 with outer code $C_{\mathrm{out}}$, inner code $C_{\mathrm{in}}$, the mappings $\Phi_t$, the set of keys $T$, and $\pi(t) = |T|^{-1}$. For $q > c^2$, if*

$$d > n - \frac{n(1-\sigma)}{c^2} + \frac{n(c-1)}{c(q-c)}, \qquad \text{with } \varepsilon_{\mathrm{in}} < \sigma < \frac{q-c^2}{q-c}, \tag{14}$$

*the family of codes $\mathcal{C} = \{C_t\}_{t \in T}$ together with Algorithm 1 is a $c$-secure with $\varepsilon$-error family of codes, with exponentially small error,*

$$\varepsilon \leq M 2^{-nD(\rho \| \frac{c-1}{q-c})} + 2^{-nD(\sigma \| \varepsilon_{\mathrm{in}})} = \exp(-\Omega(n)), \tag{15}$$

*where $\rho = \frac{1-\sigma}{c} - c(1 - d/n)$.*

*Proof* Let $U \subseteq C_t$ be a $c$-coalition, and let $W \subseteq C_{\mathrm{out}}$ be the subset of their corresponding outer codewords, as stated above. Also, let $\mathbf{z}$ be

$$\mathbf{z} = (\underbrace{z_{11}, \ldots, z_{1l}}_{\mathbf{z}_1}, \ \ldots, \ \underbrace{z_{n1}, \ldots, z_{nl}}_{\mathbf{z}_n})$$

a descendant created by coalition $U$.

First, note that in Step 1b) of Algorithm 1 we are discarding all blocks $\mathbf{z}_i$ that violate the separating property by setting $\mathcal{Z}_i = \emptyset$, an event that occurs with probability $\leq \varepsilon_{\text{in}}$, due to the properties of the inner code. Hence, $\mathcal{Z}_i \cap P_i(W) \neq \emptyset$, i.e., $\mathcal{Z}_i$ contains at least one element $w_i$ for some $\mathbf{w} = (w_1, \ldots, w_n) \in W$, with probability $\geq 1 - \varepsilon_{\text{in}}$.

Let $X$ be the number of discarded blocks, which can be upper bounded using a binomial r.v. of parameters $n$ and $p' \leq \varepsilon_{\text{in}}$. Since $\varepsilon_{\text{in}} < \sigma$, we can use (1) and (3) to see that

$$\Pr\{X \geq n\sigma\} \leq 2^{-nD(\sigma \| \varepsilon_{\text{in}})}, \tag{16}$$

which decreases exponentially with $n$.

That is, with high probability, there is a coalition codeword $\hat{\mathbf{u}} = \Phi_t(\hat{\mathbf{w}}) \in U$ for some $\hat{\mathbf{w}} \in W$, such that

$$\mathrm{s}(\hat{\mathbf{w}}, \boldsymbol{\mathcal{Z}}) \geq n\frac{1-\sigma}{c}, \tag{17}$$

hence, the output of the algorithm is not empty and a traitor is identified.

On the other hand, if $\mathbf{u} = \Phi_t(\mathbf{w})$ is a codeword of an innocent user, i.e., $\mathbf{w} \notin W$, the element $w_i$ could appear in a nondiscarded set $\mathcal{Z}_i$ when some codeword from $W$ matches this position, i.e., when $w_i \in P_i(W)$. Since any two codewords of $C_{\text{out}}$ can agree in $\leq n - d$ positions, this event can happen in at most $c(n-d)$ positions. Also, whenever $w_i \notin P_i(W)$ the probability that $w_i \in \mathcal{Z}_i$ can be bounded as

$$p_i = \Pr\{w_i \in \mathcal{Z}_i | w_i \notin P_i(W)\} \leq \frac{c-1}{q-c}. \tag{18}$$

For $1 \leq i \leq n$, let $Y_i$ be an r.v. that takes the value 1 with probability $p_i$ and 0 with probability $1 - p_i$. Therefore, for $\mathbf{w} \notin W$,

$$\begin{aligned}
\Pr\Big\{\mathrm{s}(\mathbf{w}, \boldsymbol{\mathcal{Z}}) \geq n\frac{1-\sigma}{c} \Big| \mathbf{w} \notin W\Big\} &\leq \Pr\Big\{c(n-d) + \sum_{i=1}^{n-X-c(n-d)} Y_i \geq n\frac{1-\sigma}{c}\Big\} \\
&\leq \Pr\Big\{c(n-d) + \sum_{i=1}^{n} Y_i \geq n\frac{1-\sigma}{c}\Big\} = \Pr\Big\{\sum_{i=1}^{n} Y_i \geq n\rho\Big\} \\
&\overset{\text{(a)}}{\leq} \Pr\Big\{Y \geq n\rho\Big\} \leq 2^{-nD(\rho \| \frac{c-1}{q-c})}.
\end{aligned}$$

Inequality (a) above follows from (18), by comparing the summation $\sum_{i=1}^{n} Y_i$ with an appropriate binomial r.v. $Y$ of parameters $n$ and $(c-1)/(q-c)$. Also, since $(c-1)/(q-c) < \rho$, which is implied by the condition in the minimum distance of the outer code (14), applying (1) and (3) again gives the last inequality above.

Since there are $M$ codewords, the probability of accusing an innocent user as guilty is upper bounded as

$$\begin{aligned}
&\Pr\Big\{\max_{\mathbf{w}\notin W} \mathrm{s}(\mathbf{w},\boldsymbol{\mathcal{Z}}) \geq n\frac{1-\sigma}{c}\Big\} \\
&\quad\leq M\Pr\Big\{\mathrm{s}(\mathbf{w},\boldsymbol{\mathcal{Z}}) \geq n\frac{1-\sigma}{c}\Big|\mathbf{w}\notin W\Big\} \\
&\quad\leq M2^{-nD(\rho\|\frac{c-1}{q-c})}. \qquad (19)
\end{aligned}$$

Recall that the probability of not accusing a real traitor is (16). Putting this together with (19), we have

$$\varepsilon \leq M2^{-nD(\rho\|\frac{c-1}{q-c})} + 2^{-nD(\sigma\|\varepsilon_{\mathrm{in}})}.$$

Moreover, this shows that with error probability $\varepsilon$ no codeword $\mathbf{w}\notin W$ will lie within the decoding radius (17). □

The existence of a family of fingerprinting codes with error probability decreasing exponentially in the outer code length is guaranteed using similar arguments to those from [1]. Using Reed-Solomon as outer codes we have the following result, which assumes $c$ fixed and $q$ growing.

**Corollary 2** *Let $C_{\mathrm{out}}$ be an extended $[n,k]$-Reed-Solomon code over $\mathbb{F}_q$ of rate $R_{\mathrm{out}} = R(C_{\mathrm{out}})$, and let $C_{\mathrm{in}}$ be a binary $\varepsilon_{\mathrm{in}}$-almost $(c,c)$-separating or $\varepsilon_{\mathrm{in}}$-almost $c$-secure frameproof $(l,q)$-code of rate $R_{\mathrm{in}} = R(C_{\mathrm{in}})$. Let $\mathcal{C} = \{C_t\}_{t\in T}$ be the family of codes from Construction 1 with outer code $C_{\mathrm{out}}$, inner code $C_{\mathrm{in}}$, the mappings $\Phi_t$, the set of keys $T$, and $\pi(t) = |T|^{-1}$. For $q > c^2$, and any*

$$R_{\mathrm{out}} < \frac{1-\sigma}{c(c+1)}, \qquad \text{with } \varepsilon_{\mathrm{in}} < \sigma < \frac{q-c^2}{q-c}, \qquad (20)$$

*the family of codes $\mathcal{C} = \{C_t\}_{t\in T}$ together with Algorithm 1 is a $c$-secure with $\varepsilon$-error family of codes, of length $m = n\,l$, rate*

$$R = R_{\mathrm{out}}\,R_{\mathrm{in}},$$

*and error probability $\varepsilon$ decreasing exponentially as*

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

*Proof* If $C_{\mathrm{out}}$ is an extended Reed-Solomon code with minimum distance $d$, we have $n = q$ and $1 - d/n = R_{\mathrm{out}} - 1/n$. Hence, from Theorem 3,

$$\rho = \frac{1-\sigma}{c} - c\Big(R_{\mathrm{out}} - \frac{1}{n}\Big). \qquad (21)$$

Now, since $C_{\mathrm{out}}$ has size $M = q^k$, the error probability from (15) can be rewritten as

$$\varepsilon \leq 2^{-n\,l R_{\mathrm{in}}((\log_2 q)^{-1}D(\rho\|\frac{c-1}{q-c})-R_{\mathrm{out}})} + 2^{-nD(\sigma\|\varepsilon_{\mathrm{in}})}.$$

The proof follows after substituting (21) into the previous equation, and taking into account that

$$\lim_{q\to\infty} (\log_2 q)^{-1} D\Big(\rho \Big\| \frac{c-1}{q-c}\Big) = \rho$$

for $c$ fixed and $q$ growing. □

Besides Reed-Solomon codes, algebraic-geometric codes are also proposed as outer codes in [1]. As noted in Section 5, replacing ordinary separating codes by almost separating or almost secure frameproof codes enables us to increase the asymptotical rate of the family of fingerprinting codes proposed in [1], while maintaining an exponentially small identification error.

Considering the assumptions of Corollary 2, the decoding process of a single inner code made in Step 1a) of Algorithm 1 takes time $O(l\, n^c)$ in the worst case. That is, the overall decoding time for the whole set of inner codes is $O(l\, n^{c+1})$. Moreover, we would like to stress that the main reason for Construction 1, Theorem 3 and Corollary 2 is to mimic the following strategy from [1]. If the outer code $C_{\text{out}}$ is a Reed-Solomon or an algebraic-geometric code, then Steps 3) and 4) can be efficiently done in $O(\text{poly}(n))$ by using the list decoding algorithms from [13]. Therefore, traitor identification can be efficiently achieved in polynomial time in the code length.

## 7 Conclusions

In this paper we have presented two different relaxed versions of separating codes, namely almost separating and almost secure frameproof codes. The notions introduced allows us to separate two concepts that coincide in the case of absolute separation.

To show existence bounds for almost separating codes we have used the concept of typicality. Our analysis is based in the fact that a typical set of at most $c$ codewords is separated, with very high probability, with all other disjoint sets also of at most $c$ codewords. This analysis shows that there exists almost separating codes that double the asymptotical rate of ordinary separating codes.

For almost secure frameproof codes the probabilistic analysis shows that there exist almost secure frameproof codes with asymptotical rate even higher, with relative difference to the rate of both separating and almost separating codes increasing with the coalition size.

We believe that these two notions are essentially different, in particular, we conjecture that for asymptotical rates

$$R_q^{\text{sf}^*}(c) > R_q^{\text{sep}^*}(c),$$

but it could be a rather difficult question since even for the simplest case $q = c = 2$ the best upper bound for the rate of $(2,2)$-separating codes $\overline{R}_2^{\text{sep}}(2,2) \leq 0.2835$ is very far from being "useful."

Finally, we have presented a construction of a family of fingerprinting codes. The use of almost separating or almost secure frameproof codes as inner codes allows us to obtain better code rates, preserving the exponential decline of the error probability on the outer code length, and it also enables us to obtain a polynomial-time identification algorithm.

**Acknowledgements** We would like to thank the anonymous Reviewers, whose insightful comments and observations helped to improve the contents and presentation of the paper.

## References

1. Barg, A., Blakley, G.R., Kabatiansky, G.: Digital fingerprinting codes: Problem statements, constructions, identification of traitors. IEEE Trans. Inf. Theory **49**(4), 852–865 (2003)
2. Blakley, G.R., Kabatiansky, G.: Random coding technique for digital fingerprinting codes: fighting two pirates revisited. In: Proc. IEEE Int. Symp. Inform. Theory (ISIT), p. 203. Chicago, IL (2004)
3. Boneh, D., Shaw, J.: Collusion-secure fingerprinting for digital data. In: Proc. Int. Cryptol. Conf. (CRYPTO), *Lecture Notes Comput. Sci. (LNCS)*, vol. 963, pp. 452–465. Santa Barbara, CA (1995)
4. Boneh, D., Shaw, J.: Collusion-secure fingerprinting for digital data. IEEE Trans. Inf. Theory **44**(5), 1897–1905 (1998)
5. Chor, B., Fiat, A., Naor, M.: Tracing traitors. In: Proc. Int. Cryptol. Conf. (CRYPTO), *Lecture Notes Comput. Sci. (LNCS)*, vol. 839, pp. 480–491. Santa Barbara, CA (1994)
6. Chor, B., Fiat, A., Naor, M., Pinkas, B.: Tracing traitors. IEEE Trans. Inf. Theory **46**(3), 893–910 (2000)
7. Cohen, G.D., Schaathun, H.G.: Asymptotic overview on separating codes. Tech. Rep. 248, Deptartment of Informatics, University of Bergen, Norway (2003)
8. Cohen, G.D., Schaathun, H.G.: Upper bounds on separating codes. IEEE Trans. Inf. Theory **50**(6), 1291–1294 (2004)
9. Csiszár, I., Körner, J.: Information Theory: Coding Theorems for Discrete Memoryless Systems, second edn. Cambridge Univ. Press, Cambridge, UK (2011)
10. Fernández, M., Kabatiansky, G., Moreira, J.: Almost separating and almost secure frameproof codes. In: Proc. IEEE Int. Symp. Inform. Theory (ISIT), pp. 2696–2700. Saint Petersburg, Russia (2011)
11. Forney, G.D.: Concatenated codes. Tech. Rep. 440, Res. Lab. Electron., MIT, Cambridge, MA (1966)
12. Friedman, A.D., Graham, R.L., Ullman, J.D.: Universal single transition time asynchronous state assignments. IEEE Trans. Comput. **C-18**(6), 541–547 (1969)
13. Guruswami, V., Sudan, M.: Improved decoding of Reed-Solomon and algebraic-geometry codes. IEEE Trans. Inf. Theory **45**(6), 1757–1767 (1999)
14. Hoeffding, W.: Probability inequalities for sums of bounded random variables. J. Amer. Stat. Assoc. **58**(301), 13–30 (1963)
15. Körner, J., Simonyi, G.: Separating partition systems and locally different sequences. SIAM J. Discr. Math. (SIDMA) **1**(3), 355–359 (1988)
16. Pinsker, M.S., Sagalovich, Y.L.: Lower bound on the cardinality of code of automata's states. Probl. Inform. Transm. **8**(3), 59–66 (1972)
17. Sagalovich, Y.L.: Completely separating systems. Probl. Inform. Transm. **18**(2), 140–146 (1982)
18. Sagalovich, Y.L.: Separating systems. Probl. Inform. Transm. **30**(2), 105–123 (1994)
19. Staddon, J.N., Stinson, D.R., Wei, R.: Combinatorial properties of frameproof and traceability codes. IEEE Trans. Inf. Theory **47**(3), 1042–1049 (2001)
20. Stinson, D.R., van Trung, T., Wei, R.: Secure frameproof codes, key distribution patterns, group testing algorithms and related structures. J. Stat. Plan. Infer. **86**(2), 595–617 (2000)