Separating codes: constructions and bounds

Gérard Cohen[1]
and Hans Georg Schaathun[2]

[1] Ecole Nationale Supérieure des Télécommunications
46 rue Barrault
F-75634 Paris Cedex, France
cohen@enst.fr
[2] Department of Informatics
University of Bergen
Høyteknologisenteret
N-5020 Bergen, Norway
georg@ii.uib.no

Abstract Separating codes, initially introduced to test automaton, have revived lately in the study of fingerprinting codes, which are used for copyright protection. Separating codes play their role in making the fingerprinting scheme secure agains coalitions of pirates. We provide here better bounds, constructions and generalizations for these codes.

1 Introduction

Separating codes were introduced in 1969 and have been the topic of several papers with various motivations. Many initial results are due to Sagalovich; see [3] for a survey, and also [2,5]. New applications of separating codes have appeared during the last decade, namely *traitor tracing* and *fingerprinting*.

Fingerprinting is a proposed technique for copyright protection. The vendor has some copyrighted work of which he wants to sell copies to customers. If he is not able to prevent the customer from duplicating his copy, he may individually mark every copy sold with a unique fingerprint. If an illegal copy (for which the vendor has not been paid) subsequently appears, it may be traced back to one legal copy and one pirate via the fingerprint. A pirate is here any customer guilty of illegal copying of the copyrighted work.

Traitor tracing is the same idea applied to broadcast encryption keys. E.g. the vendor broadcasts encrypted pay-TV, and each customer buys or leases a decoder box to be able to decrypt the programmes. If the vendor is not able to make the decoder completely tamperproof, he may fingerprint the decryption keys which are stored in the box.

The set of fingerprints in use, is called the fingerprinting code. Separating codes are used in the study of *collusion secure* fingerprinting codes. If several pirates collude, they posess several copies with different fingerprints. By comparing their copies, they will find differences which must be part of the fingerprint.

These identified "marks" may be changed to produce a false fingerprint. A collusion secure code should aim to identify at least one of the pirates from this false fingerprint.

We shall introduce two useful concepts regarding collusion secure codes. If the code is t-frameproof, it is impossible for any collusion of at most t pirates to produce a false fingerprint which is also a valid fingerprint of an innocent user. In other words, no user may be framed by a coalition of t pirates or less. A t-frameproof code is the same as a $(t, 1)$-separating code, which will be defined formally in the next section.

If the code is t-identifying, the vendor is always able to identify at least one pirate from any coalition of size at most t, given a false fingerprint created by the coalition. A first step towards identification is (t, t)-separation (see, e.g. [4]), which we study and generalize here.

2 Definitions

For any positive real number x we denote by $\lceil x \rceil$ the smallest integer at least equal to x. Let A be an alphabet of q elements, and A^n the set of sequences of length n over it. A subset $C \subseteq A^n$ is called an $(n, M)_q$ or (n, M)-*code* if $|C| = M$. Its *rate* is defined by $R = (\log_q M)/n$. For any $\mathbf{x} \in A^n$, we write x_i for the i-th component, so that $\mathbf{x} = (x_1, x_2, \ldots, x_n)$. The minimum Hamming distance between two elements (codewords) of C is denoted by $d(C)$ or d, and the normalised quantity d/n by δ.

Consider a subset $\mathcal{C} \subseteq C$. For any position i, we define the *projection* $P_i(\mathcal{C}) = \bigcup_{\mathbf{a} \in \mathcal{C}} \{a_i\}$. The *feasible set* of $\mathcal{C}$ is

$$F(\mathcal{C}) = \{\mathbf{x} \in A^n : \forall i, x_i \in P_i(\mathcal{C})\}.$$

If $\mathcal{C}$ is the fingerprints held by some pirate coalition, then $F(\mathcal{C})$ is the set of fingerprints they may produce. If two non-intersecting coalitions can produce the same descendant, i.e., if their feasible sets intersect, it will be impossible to trace with certainty even one pirate. This motivates the following definition.

Definition 1. *A code C is (t, t')-separating if, for any pair (T, T') of disjoint subsets of C where $|T| = t$ and $|T'| = t'$, the feasible sets are disjoint, i.e. $F(T) \cap F(T') = \emptyset$.*

Such codes are also called *separating systems*, abbreviated by SS.

Since the separation property is preserved by translation, we shall always assume that $\mathbf{0} \in C$. The separation property can be rephrased as follows when $q = 2$: For any ordered $t + t'$-tuple of codewords, there is a coordinate where the $t + t'$-tuple $(1..10..0)$ of weight t or its complement occurs.

Given a (t, t')-configuration (T, T') we define the separating set $\Theta(T, T')$ to be the set of coordinate positions where (T, T') is separated. Let $\theta(T, T') := \#\Theta(T, T')$ be the separating weight. Clearly $\theta(T, T') \geq 1$ is equivalent with (T, T') being separated. The minimum (t, t')-separating weight $\theta_{t,t'}(C)$ is the

least separating weight of any (t,t')-configuration of C. We abbreviate $\theta_{i,i}(C)$ to $\theta_i(C)$ or θ_i. Clearly $\theta_1(C) = d(C)$. The minimum separating weights have previously been studied by Sagalovich [3].

3 Bounds on $(t,1)$ separating codes

The case $t' = 1$ corresponds to "frameproof" codes introduced in [1]. Körner (personal communication) has a simplified proof of $R \leq 1/2$ for (1,2)-separation in the binary case. We generalize it to any t and q, and for bounded separating weight $n\tau$.

A (t,τ)-*coverfree* code is a code with $(t,1)$-separating weight at least equal to τn. Their study in [11] and [9] is motivated by broadcast encryption.

Partition $\{1,2,..n\}$ into t almost equal parts $P_1,\ldots,P_t$ of size approximately n/t. Say a codeword c is *isolated* on P_i if no other codeword projects on P_i on a vector located at distance less than $(n/t)\tau$ from c. Denote by U_i the subset of codewords isolated on P_i.

Lemma 1. *If C is (t,τ)-coverfree, then every codeword c of C is isolated on at least one P_i.*

Proof: Suppose for a contradiction that there is a codeword $\mathbf{c}_0$ which is not isolated. Let $\mathbf{c}_i$ be a codeword which is at distance less than $(n/t)\tau$ when projected onto P_i, for $i = 1,\ldots,t$. Now $\mathbf{c}_0$ is separated from $\{\mathbf{c}_1,\ldots,\mathbf{c}_t\}$ on less than $(n/t)\tau$ coordinates per block, or at most $n\tau - t$ coordinate positions total. This contradicts the assumption on the separating weight τ.

If we let τ tend to zero, we get an upper bound on the size of $(t,1)$-separating codes, which was found independently in [13] and [12]. The proofs are essentially the same as the one presented here.

Theorem 1. *If C is (t,τ)-coverfree, then $|C| \leq tq^{\lceil (1-\tau)n/t \rceil}$.*

For constant t, this asymptotically gives a rate $R \leq (1-\tau)/t$ when n increases. A lower bound on the rate can now be obtained by invoking a sufficient condition for C to be (t,τ)-coverfree, based on its minimum distance d: $td \geq (t-1+\tau)n$. This is proved in a more general form in Proposition 1. Using algebraic-geometric (AG) codes [7] with $\delta > t^{-1}(1-\tau)$ and $R \approx 1-\delta-1/(q^{1/2}-1)$ gives the following asymptotically tight (in q):

Theorem 2. *For fixed t and large enough q, the largest possible rate of a q-ary family of (t,τ)-coverfree codes satisfies $R = t^{-1}(1-\tau)(1+o(1))$.*

4 Large separation

Definition 2. *A code C of length n is (t,t',τ)-separating if, for any pair (T,T') of disjoint subsets of C where $|T| = t$ and $|T'| = t'$, $\theta(T,T') \geq \tau n$.*

Proposition 1. *A code with minimum distance d is (t, t', τ)-separating if*

$$tt'd \geq (tt' - 1 + \tau)n.$$

Proof: Consider two disjoints sets T and T' of sizes t and t' respectively and count the sum Σ of pairwise distances between them: on one hand, $\Sigma \geq tt'd \geq (tt' - 1 + \tau)n$. Computing Σ coordinatewise now, we get that the contribution to Σ of at least τn coordinates must be greater than $tt' - 1$, i.e. tt'. Thus, these coordinates separate T and T'.

To construct infinite families of separating codes over small alphabets, we can resort to the classical notion of *concatenation*.

Definition 3 (Concatenation). *Let C_1 be a $(n_1, Q)_q$ and let C_2 be an $(n_2, M)_Q$ code. Then the concatenated code $C_1 \circ C_2$ is the $(n_1 n_2, M)_q$ code obtained by taking the words of C_2 and mapping every symbol on a word from C_1.*

The following result is an easy consequence of the definition.

Proposition 2. *Let Γ_1 be a $(n_1, M)_{M'}$ code with minimum separating weight $\theta^{(1)}_{t,t'}$, and let Γ_2 be a $(n_2, M')_q$ code with minimum separating weight $\theta^{(2)}_{t,t'}$. Then the concatenated code $\Gamma := \Gamma_2 \circ \Gamma_1$ has minimum separating weight $\theta_{t,t'} = \theta^{(1)}_{t,t'} \cdot \theta^{(2)}_{t,t'}$.*

We shall illustate the concatenation method with $q = 2, t = 2, t' = 1$ in the next section.

5 The binary case

5.1 (2, 1)-separation

In [8], it was pointed out that shortened Kerdock codes $K'(m)$ for $m \geq 4$ are $(2, 1)$-separating. Take an arbitary subcode of size 11^2 in $K'(4)$ which is a $(15, 2^7)$ $(2, 1)$-SS. Concatenate it with an infinite family of algebraic-geometry codes over $\mathsf{GF}(11^2)$ (the finite field with 11^2 elements) with $\delta > 1/2$ (hence $(2, 1)$-separating by Proposition 1) and $R \approx 1/2 - 1/11$ [7]. After some easy computations, this gives:

Theorem 3. *There is a constructive asymptotic family of binary $(2, 1)$-separating codes with rate $R = 0.1845$.*

This can even be refined if we concatenate with the codes contained in the following proposition from [10].

Proposition 3. *Suppose that $q = p^{2r}$ with p prime, and that t is an integer such that $2 \leq t \leq \sqrt{q} - 1$. Then there is an asymptotic family of $(t, 1)$-separating codes with rate*

$$R = \frac{1}{t} - \frac{1}{\sqrt{q} - 1} + \frac{1 - 2\log_q t}{t(\sqrt{q} - 1)}.$$

Remark 1. If we use the Xing's codes ([10]), we get an improved rate of $R \approx 0.2033$, but at the expense of constructivity.

5.2 A stronger property

Definition 4 (Completely Separating Code). *A binary code is said to be (t, t')-completely separating ((t, t')-CSS) if for any set ordered set of $t + t'$ codewords, there is at least one column with 1 in the t upper positions, and 0 elsewhere, and one column with 0 in the t upper positions and 1 in the t' lower ones.*

We define $R_{SS}(t, t')$ as the largest possible asymptotical rate of a family of (t, t')-SS, and similarly $R_{CSS}(t, t')$ for (t, t')-CSS. We clearly have

$$R_{SS}(t, t') \geq R_{CSS}(t, t') \geq \frac{1}{2} R_{SS}(t, t'). \tag{1}$$

5.3 Improved upper bounds on (t, t)-separating codes

Theorem 4. *A (t, t)-separating (θ_0, M, θ_1) code with separating weights $(\theta_1, \ldots, \theta_t)$ gives rise to a (i, i)-CSS $(\theta_{t-i}, M - 2t + 2i, 2\theta_{t+1-i})$ with complete-separating weight θ_i, for any $i < t$.*

Proof: Consider a pair of $(t-i)$-tuples of vectors which are separated on θ_{t-i} positions. Pick any vector $\mathbf{c}$ from the first $(t - i)$-tuple and replace the code C by its translation $C - \mathbf{c}$. Thus all the columns which separates the two tuples have the form $(0 \ldots 01 \ldots 1)$.

Now consider any two i-tuples of vectors. Coupling each i-tuple with a $(t-i)$-tuple, we get two t-tuples which must be separated on θ_t positions, i.e. the two i-tuples must have at least θ_t columns of the form $(0 \ldots 01 \ldots 1)$. Now, observe that we can swap the two $(t - i)$-tuples, and the two resulting t-tuples are still separated. This guarantees at least θ_t columns of the form $(1 \ldots 10 \ldots 0)$.

Deleting all the columns where the two $(t - i)$-tuples are not separated, and the words of these two tuples must this leave us with an (i, i)-CSS with complete-separating weight θ_i and parameters $(\theta_{t-i}, M - 2t + 2i, 2\theta_{t+1-i})$, as required.

Theorem 5. *Any completely (t, t)-separating $(\theta_0, M, 2\theta_1)$ code with complete-separating weights $(\theta_1, \ldots, \theta_t)$ gives rise to a completely (i, i)-separating $(\theta_{t-i}, M - 2t + 2i, 2\theta_{t+1-i})$ code with complete-separating weight θ_i, for any $i < t$.*

This is proved in the same way as the previous theorem.

Theorem 6. *For any (t, t)-CSS, the rate R_t satisfies*

$$R_t \leq \bar{R}(2R_t / \bar{R}_{t-1}),$$

where $\bar{R}(\delta)$ is any upper bound on the rate of error-correcting codes in terms of the normalised minimum distance, and $\bar{R}_{t-1}$ is the upper bound on the rate of any $(t - 1, t - 1)$-CSS.

Proof: Let C_{t-1} be the $(t-1,t-1)$-CSS which exists by Theorem 5, and let R_{t-1} be its rate. We have that

$$\delta_t = 2\frac{\theta_1}{\theta_0} = 2\frac{\log M}{\theta_0}\frac{\theta_1}{\log M} = 2R_t/R_{t-1}.$$

Now, obviously $R_t \leq \bar{R}(\delta_t)$, which is decreasing in δ_t, and this gives the result.

With a completely analogous proof, we also get the following.

Theorem 7. *For any (t,t)-SS, the rate R satisfies*

$$R \leq \bar{R}(R/\bar{R}_{t-1}),$$

where $\bar{R}(\delta)$ is any upper bound on the rate of error-correcting codes in terms of the normalised minimum distance, and $\bar{R}_{t-1}$ is the upper bound on the rate of any $(t-1,t-1)$-CSS.

	Bound 1		D'yachkov et al.	Bound 2	
(t,t)	CSS rate	SS rate	CSS rate	CSS rate	SS rate
(1,1)	1	1	1	1	1
(2,2)	0.1712	0.2835	0.161	–	–
(3,3)	0.03742	0.06998	0.0445	0.0354	0.0663
(4,4)	0.008843	0.01721	0.0123	0.00837	0.0163
(5,5)	0.002156	0.004261	0.00333	0.00204	0.00404

Table 1. Rate bounds on CSS and SS.

Setting equality in the bounds and solving, we get the upper bounds given as 'Bound 1' in Table 1. Comparing with the CSS bounds of [6] shows an improvement from $(3,3)$-CSS onwards. However, [6] has a good bound on $(2,2)$-CSS, used as a seed for the recursive bounds of our theorems to obtain 'Bound 2' in the table.

Example 1. Let C_1 be an asymptotic class of $(\theta_0, 2^k, \theta_1)$ $(3,3)$-SS. Then there is an asymptotic class C_2 of $(\theta_1, 2^k, \theta_2)$ $(2,2)$-CSS. We have that $R_2 = k/\theta_1 \leq 0.161$, and

$$R_1 = k/\theta_0 = R_2\delta_1 \leq 0.161\delta_1,$$

which is equivalent to $\delta_1 \geq R_1/0.161$. We can use any upper bound $\bar{R}(\delta)$ on R_1, and get

$$R_1 \leq \bar{R}(\delta_1) \leq \bar{R}(R_1/0.161),$$

and $R_1 \leq 0.0663$ by the linear programming bound.

References

1. D. Boneh and J. Shaw, "Collusion-secure fingerprinting for digital data", *IEEE Trans. on Inf. Theory*, **44** (1998), pp. 480–491.
2. J. Körner and G. Simonyi, "Separating partition systems and locally different sequences," *SIAM J. Discrete Math.*, **1** No 3 (1988) pp. 355–359.
3. Yu. L. Sagalovich, "Separating systems", *Probl. Inform. Trans.* **30** No 2 (1994) pp. 105–123.
4. A. Barg, G. Cohen, S. Encheva, G. Kabatiansky, and G. Zémor, "A hypergraph approach to the identifying parent property", *SIAM J. Disc. Math.*, vol. 14, 3 (2001) pp. 423-431.
5. G. Cohen, S. Encheva, and H.G. Schaathun, "More on $(2,2)$-separating systems", *IEEE Trans. Inform. Theory*, vol. 48, 9 (2002) pp. 2606-2609.
6. A. D'yachkov, P. Vilenkin, A. Macula, and D. Torney, "Families of finite sets in which no intersection of ℓ sets is covered by the union of s others", *J. Combinatorial Theory*, vol. 99, 195-208 (2002).
7. M.A. Tsfasmann, "Algebraic-geometric codes and asymptotic problems", *Discrete Applied Math.*, vol. 33 (1991) pp. 241-256.
8. A. Krasnopeev and Yu. Sagalovitch, "The Kerdock codes and separating systems", *Eighth International Workshop on Algebraic and Combinatorial Theory*, 8-14 Sept. 2002, pp. 165-167.
9. R. Kumar, S. Rajagopalan and A. Sahai, "Coding constructions for blacklisting problems without computational assumptions", Crypto'99 *LNCS* 1666 (1999) 609-623.
10. Chaoping Xing, "Asymptotic bounds on frameproof codes", *IEEE Trans. Inform. Th.* 40 (2002) 2991-2995.
11. J. Garay, J. Staddon and A. Wool, "Long-lived broadcast encryption", *Springer LNCS* 1880 (2000) pp. 333–352.
12. S. R. Blackburn, "Frameproof codes", *SIAM J. Discrete Math.*, vol. 16 (2003) 499-510.
13. G. Cohen and H.- G. Schaathun, "New upper bounds on separating codes", *2003 International Conference on Telecommunications*, February 2003.