TOWARDS OPTIMAL DESIGN OF SIGNAL FINGERPRINTING CODES

Jean-François Jourdas and Pierre Moulin

University of Illinois
Beckman Inst., Coord. Sci. Lab., & ECE Dept.
405 N. Mathews Ave., Urbana, IL 61801, USA
Email: *jourdas2@uiuc.edu, moulin@ifp.uiuc.edu*

ABSTRACT

Digital fingerprinting aims at protecting multimedia contents from illegal redistribution by embedding imperceptible fingerprints identifying the users. We propose two approaches for building fingerprinting codes that accommodate millions of users and resist tens of colluders. These approaches are based on recent information-theoretic analyses of good fingerprinting codes in two regimes: 1) very low rates, and 2) rates near capacity. Good low-rate codes have high minimum distance. Good high-rate codes are short and random-like. Simulation results are presented to assess decoding performance.

Index Terms— Digital Fingerprinting

1. INTRODUCTION

Digital fingerprinting aims at protecting multimedia contents from illegal redistribution by embedding imperceptible fingerprints identifying the users. Some users may collude and construct a pirated copy, or forgery, and illegally redistribute it. To deter them from doing this, one needs to be able to trace the pirated copy back to at least one of the pirates (colluders). The colluders may use various attacks to thwart detection of their fingerprints from the pirated copy. A popular one is averaging of the colluders' copies. Another one is interleaving, where each colluder contributes parts of his marked copy to construct the pirated copy. Following the averaging or interleaving operation, the pirates may add low-level noise to further impair detectability of their fingerprints without degrading content quality too much. While noise-resilient codes exist, it remains a challenging research problem to design tracing algorithms that are robust against collusion attacks and have reasonably low execution time, especially when the number of users is large.

In the literature, two distinct frameworks have been used to build collusion-resistant fingerprinting codes. The first one relies on the Boneh-Shaw *marking assumption*. Examples include Reed-Solomon based codes [1, 2, 3] and anti-collusion codes [4]. For media fingerprinting however, the marking assumption is less natural than a *distortion assumption* which limits the amount of distortion that the fingerprint distributor and the colluders are allowed to introduce.

For real-valued signals, optimal fingerprinting codes built under the distortion framework are known when M, the number of users, does not exceed $N+1$, the fingerprint codelength. The optimal codes are simplex codes and their performance is asymptotically equal to that of orthogonal codes for large N [5]. When $M > N+1$, no such simple solution exists. Most codes developed in fingerprinting literature have large minimum distance [1, 2, 3, 6, 7]. Recent results [8] have established optimality properties of this approach *at low rates*, in the following sense. Let $R = \frac{1}{N}\log_2 M$ be the rate of the fingerprinting code. Denote by C the fingerprinting capacity, i.e., the supremum of achievable rates. If $R > C$, it is impossible to reliably detect even one colluder. For Gaussian hosts and Euclidean distortion metrics, capacity is a function of SNR (the ratio of fingerprint power to colluder noise power) and of the number K of colluders [9]:

$$C = \frac{1}{2K}\log_2\left(1+\frac{\mathrm{SNR}}{K}\right). \qquad (1)$$

It is thus theoretically possible to catch at least one colluder with low error probability if $R < C$ and N is large enough.

When $R/C \to 0$, expurgated spherical codes (which have large minimum distance) are optimal [8]. Motivated by this result, we have designed a family of modulated fingerprinting codes based on maximum-distance separable algebraic codes. These codes can accommodate an exponential number of users, admit efficient decoding algorithms, and their traceability properties can be theoretically estimated. These codes have low rate and thus long fingerprints (for a given M).

However, in media fingerprinting, short fingerprints are valuable because there are relatively few media features in which they can be robustly embedded. This suggests developing *high-rate* codes, where R potentially approaches C. For such codes what matters is not the minimal distance, but the global distribution of the fingerprints. Since nonexpurgated

This work was supported by NSF under grant CCF 06-35137, and by HP Laboratories.

random-like codes are capacity-achieving, this motivated us to develop fingerprinting codes based on this idea.

This paper present both the low-rate and the high-rate codes discussed above and compare their performance with existing codes in the literature.

2. LOW-RATE FINGERPRINTING

Expurgated spherical codes have high minimum distance and are theoretically excellent for low-rate fingerprinting [8]. In order to design a code that can accommodate many users, we recently proposed the following construction [10]. We chose a $[n, k]$ Reed-Solomon outer code over the Galois field GF(q) with dimension $k \ll n$ and maximal length $n = q - 1$. This code is modulated onto an orthonormal q-dimensional constellation. The length of the code is $N = qn$ and the number of users it can accommodate is $M = q^k$.

We hard decoded the inner code and used the state-of-the-art Guruswami-Sudan [11] decoding algorithm which has the capacity to correct as many as $n(1 - \sqrt{k/n})$ errors. As a result, this code is a K-resilient anti-averaging collusion code if K satisfies $n/k \geq K^2$. The level of noise we can handle depends on how tight the inequality is and also on how the inner code is decoded. For a [31,5] Reed-Solomon code over GF(32), we have $M = 32^5 = 33,554,432$ and $N = 31 \times 32 = 992$, and the Guruswami-Sudan algorithm can correct 18.55 errors. When $K = 2$, there are 15.5 errors in average so we can recover the two fingerprints if the noise level is not too high.

We assessed the performance of this code against the averaging attack with $\mathrm{SNR} = 1$. The number of Monte Carlo simulations performed ranged from 350 to 1,040 depending on the experiment. In the table below, $d_{\min}$ is the normalized minimal distance of the code (would be 2 for orthogonal codes), $\hat{P}_{\mathrm{e}}$ represents the probability that no colluder is caught, and $\hat{P}_{\mathrm{e}}^{\mathrm{all}}$ represents the probability that not all colluders are caught.

exp. #	k	n	q	$N = qn$	$M = q^{\mathrm{k}}$	R
(1)	3	63	64	4,032	262,144	0.045
(2)	4	31	32	992	1,048,576	0.020
(3)	4	31	32	992	33,554,432	0.025

exp. #	$d_{\min}^2$	K	SNR	$\hat{P}_{\mathrm{e}}$	$\hat{P}_{\mathrm{e}}^{\mathrm{all}}$
(1)	1.94	3	1	0	0.002
(2)	1.80	2	1	0.001	0.013
(3)	1.75	2	1	0.017	0.04

Table 1. Results obtained for the low-rate fingerprinting code.

The code of length $N = 4032$ bits can accommodate a large amount of users ($M = 262,144$) and resist 3 colluders. This code is shorter than the code given in [4] and supports many more users for the same collusion resistance. The code with the same parameters but 33,554,432 users can resits 2 colluders. Note that [6, 7] also used fingerprinting codes based on a Reed-Solomon inner code but did not use the Guruswami-Sudan decoding algorithm. The decoder of [6] uses brute-force search, which is manageable only if M is small.

In fingerprinting problems, the goal is generally to catch only one of the colluders, and this is what $\hat{P}_{\mathrm{e}}$ measures. Catching all colluders is far more difficult, especially if they do not contribute equally. We notice from the results, that the value of $\hat{P}_{\mathrm{e}}^{\mathrm{all}}$ is only a little higher than $\hat{P}_{\mathrm{e}}$ meaning that in most cases we can identify all the colluders. Our code has a very low rate in comparison to the capacity of the collusion attack channel, justifying the use of codes with high minimum distance as outer and inner code. If the goal is to catch only one of the colluders then C is given by (1).

If we have $q = 2^m$ where $m \in \mathbb{N}$ and if our goal is to catch only one of the colluders, then the code is K-resilient if

$$R \leq \frac{m}{2^{m-1} K \log_2(1 + \mathrm{SNR}/K)} C \,. \qquad (2)$$

We see from (2) that $R \ll C$. The ratio C/R is dominated by 2^{m-1} which is close to $q = 2^m$ when m is large. We need to have large q if we want to accommodate many users or to be resilient to large coalitions. For instance, for $K = 3$, $m = 6$ and $k = 3$, we have a code of length $N = 31 \times 32 = 4032$ bits and rate $R = 0.0045$, accomodating $M = 64^3 = 262,144$ users. From (1) we obtain $C = 0.069 \approx 15R$. The codes with high minimum distance are optimal when $R \ll C$. The numerical results above suggest that our construction is in this regime. It is thus a good idea to use a Reed-Solomon code (which is maximum distance separable) as outer code, and an orthonormal code as inner code.

To summarize, the above construction gives a low-rate code. The decoder is able to identify at least one of the colluders and often all of the colluders against averaging attacks with low error probability. However, when the number of colluders increases, we need very long codewords to resist the collusion which makes it difficult to embed the fingerprint into a multimedia content, and dramatically increases the execution time of the Guruswami-Sudan decoding algorithm. This limitation motivates the approach described next.

3. HIGH-RATE FINGERPRINTING

Perhaps motivated by codes designed under the marking assumption, most authors have focused on the design of codes with high minimum distance. For signal fingerprinting however, this approach is optimal only if the code rate is much lower than capacity. At rates close to the capacity, the overall distribution of the fingerprints is what matters. Therefore, we propose the following design of a random-like fingerprinting code. Preliminary results were reported in [12].

Each user is identified by a bitstring of length $n = \log_2 M$. First, the bitstring is encoded with a recursive systematic convolutional (RSC) code of rate R_1 which gives a binary subcodeword of size n/R_1. Then, it is interleaved and encoded with the same or another RSC code of rate R_2 which yields a different subcodeword of size n/R_2. This last operation is repeated a total of N_i times. At the end, the $N_\mathrm{i}+1$ subcodewords generated are concatenated to form a fingerprint of size $N = n\sum_{i=1}^{N_\mathrm{i}+1}(1/R_i)$. The number of users that can be accommodated is $M = 2^n$, and the rate of the fingerprinting code is $R = 1 \Big/ \sum_{i=1}^{N_\mathrm{i}+1}(1/R_i)$.

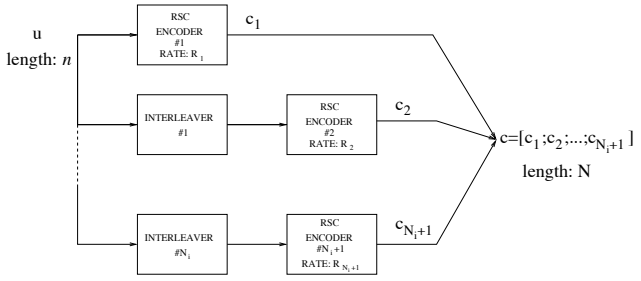

Fig. 1. Encoding Scheme.

This code does not have a high minimum distance but the introduction of random interleavers makes the $N_\mathrm{i}+1$ subcodewords uncorrelated; thus we gain by exchanging information after decoding each subcodeword. Moreover, we can use different polynomial generators g_i in each branch to increase diversity.

To decode a codeword, we first retrieve separately the subcodewords $\{r_\mathrm{i}\}$ and decode them with a list Viterbi decoding algorithm outputting a list L_i of size D. Our simulations showed that, usually, one list contains one or many colluders but their subcodewords are not at the top of the list because the forgery is closer to the subcodeword of an innocent user. However, by comparing the users accused in each list, we gain reliability. Once all subcodewords have been decoded, we take the union of the $N_\mathrm{i}+1$ lists L_i to get a final list L_f. Among all the users in the final list, we run a matched filter to select the fingerprint that has the highest correlation with the forgery. This algorithm aims at catching only one of the pirates.

The binary fingerprint is made of blocks of size n/R_i. To map the binary fingerprint to a vector in $\mathbb{R}^N$, we applied the discrete cosine transform (DCT) to each block and concatenated the resulting DCT coefficients. We chose the DCT because it is fast to compute and suitable for robust embedding.

We assessed the performance of this code against averaging attack followed by additive white Gaussian noise with 100 Monte-Carlo simulations. In our simulations, we targeted 30 millions users and a coalition of about 50 colluders. The parameters of the code are chosen to meet these requirements

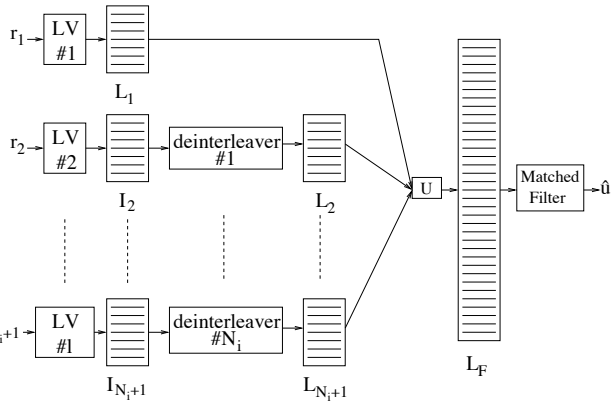

Fig. 2. Decoding Scheme.

and we determine the collusion resistance through simulations. For all the experiments, we chose the same rate R_1 for the polynomial generators g_i. In some cases, we even used the same polynomial generator to encode all the subcodewords. The table below lists the code rate R, the ratio C/R obtained from (1), and the probability $\hat{P}_\mathrm{e}$ that the algorithm accuses an innocent user.

exp. #	g_k	R_1	D	N	N_i	K
(1)	identical	$\frac{1}{1400}$	512	$201,600$	3	50
(2)	identical	$\frac{1}{700}$	512	$196,000$	7	40
(3)	different	$\frac{1}{10}$	256	$1,160$	3	3

exp. #	SNR	R	C/R	$\hat{P}_\mathrm{e}$
(1)	1	1.24×10^{-4}	1.16	0.01
(2)	1	1.28×10^{-4}	1.24	0
(3)	1	0.022	3.14	0

Table 2. Results obtained for the high-rate fingerprinting code. For all experiments, the number of users is $M = 33,554,432$. $\hat{P}_\mathrm{e}$ is the empirical probability that the algorithm accuses an innocent user.

The results are very promising. We are able to accommodate millions of colluders and resist collusion of about 50 colluders with a very low probability of error $\hat{P}_\mathrm{e}$. For comparison, Trappe *et al.* [4] used a code of length $10,000$ accommodating 20 users and resiting 3 colluders against the averaging attack. Here, with a code of length $N = 1,160$, we accommodate $M = 33,554,432$ users and resist up to 3 colluders. He and Wu [6] used a code of length $N = 30,000$ accommodating only 1024 users but resisting up to 25 colluders under interleaving and averaging attacks. Later, they used a joint coding/embedding technique [7] which resists up to 100 colluders against averaging attacks. But the decoding algorithm, which relies on the use of trimming symbols, has complexity $O(qN)$ and N is extremely large ($N = 260$ Mbits).

The complexity of our algorithm is dominated by the

matched filter which is $O(NDN_i + 1))$. In our scheme N is relatively small, making the computations tractable. The list size D is a very important parameter. The larger D is, the larger $L_{\rm f}$ is, and the more accurate the matched filter detection. On the other hand, computational complexity increases. Other than that, using large D is a good way to improve the performance of the code without modifying its length.

For a given $N/n = (N_{\rm i} + 1)/R_1$, better performance is obtained when R_1 and $N_{\rm i}$ are both small.

To summarize, this code is the fingerprinting code with the highest rate ever proposed for a given number of colluders. Because it is short, it can be embedded easily and efficiently into multimedia content. It distinguishes itself from other designs by relying not on minimum distance but on the randomness of the codewords and decoding efficiency.

Reduction of the probability of false alarm. The decoding algorithm above always accuses a user. But, it is very important not to accuse an innocent user. We present a simple modification of our algorithm that reduces the probability of false positives.

First an observation. With $K = 3$, $N = 1,160$, $M = 33,554,432$, $N_i = 3$, $R_1 = \frac{1}{10}$ and $D = 256$, we ran 100 Monte Carlo simulations to estimate the distribution of the correlations of the users's fingerprints in the final list $L_{\rm f}$ with the forgery. After the decoding, we split $L_{\rm f}$ into two lists: one containing only innocent users and the other one containing colluders. We then computed the correlation of these user's fingerprints with the pirated copy and computed histograms to see how the correlation statistics are distributed for guilty and innocent users.

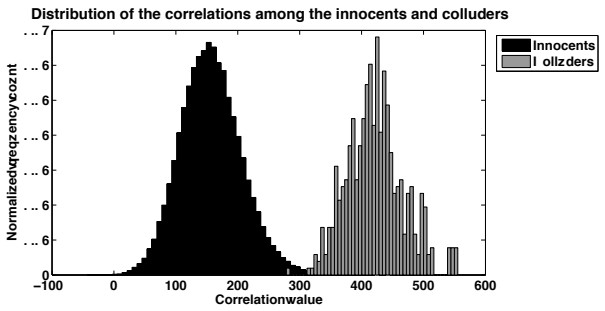

Fig. 3. Distributions of correlation statistics for guilty and innocent users.

The two histograms show that the distributions of the correlation statistic for innocent and guilty users respectively have peaks around 150 and 425. The two distributions are well separated.

This suggests a simple modification of the previous scheme. We choose a threshold and declare a decoding failure if the correlation statistic for the suspect identified by the decoding algorithm falls below the threshold. The threshold can be chosen by fixing the desired probability of false alarm, e.g., 0.1%.

4. CONCLUSION

In this paper, we have presented two fingerprinting codes. One with low rate and the other one with high rate. The low-rate code is easy to analyze but has long codewords even if it is shorter than other existing fingerprinting codes. The high-rate code is short but can accommodate millions of users and tens of colluders. This code operates close to the fundamental capacity limit, so dramatic improvements over this design are unlikely.

5. REFERENCES

[1] A. Barg, G. R. Blakley and G. Kabatiansky, "Digital Fingerprinting Codes: Problem Statements, Constructions, Identification of Traitors," *IEEE Trans. Information Theory*, Vol. 49, No. 4, pp. 852—865, Apr. 2003.

[2] A. Silverberg, J. Staddon and J.L. Walker, "Efficient Traitor Tracing Algorithms Using List Decoding". *IEEE Trans. Information Theory*, Vol. 49, No. 5,pp. 1312—1318, May 2003.

[3] M. Fernandez and M. Soriano, "Identification of Traitors in Algebraic-Geometry Traceability Codes," *IEEE Trans. on Signal Processing* supplement on secure media, Vol. 52, No. 10, pp. 3073-3077, Oct. 2004.

[4] W. Trappe, M. Wu, J. Wang and K.J.R. Liu, "Anti-Collusion Fingerprinting For Multimedia," *IEEE Trans. Signal Processing*, Vol. 51, No. 4, pp. 1069—1087, Apr. 2003.

[5] N. Kiyavash and P. Moulin, "Regular Simplex Fingerprints and Their Optimality Properties," *Proc. Int. Workshop on Digital Watermarking*, Siena, Italy, Sep. 2005.

[6] S. He and M. Wu, "Joint Coding and Embedding Techniques for Multimedia Fingerprinting," *IEEE Trans. on Information Forensics and Security*, Vol. 1, No. 2, pp. 231—247, June 2006.

[7] S. He and M. Wu, "Collusion-Resistant Video Fingerprinting For Large User Group," *2006 IEEE Conference on Image Processing*, Atlanta, GA, pp. 2301—2304, Oct. 2006.

[8] P. Moulin and N. Kiyavash, "Expurgated Gaussian Fingerprinting Codes," *Proc. IEEE Int. Symp. on Information Theory*, Nice, France, June 2007.

[9] Y. Wang and P. Moulin, "Capacity and Optimal Collusion Attack Channels for Gaussian Fingerprinting Games," *Proc. IS&T/SPIE Sym. on Electronic Imaging—Conference on Security, Steganography, and Watermarking of Multimedia Contents IX*, San Jose, Jan. 2007.

[10] J-F. Jourdas and P. Moulin, "A Low-Rate Fingerprinting Code And Its Application To Blind Image Fingerprinting," *to appear in SPIE Proceedings*, San Jose, CA, January 2008.

[11] V. Guruswami and M. Sudan, "Improved Decoding Of Reed-Solomon And Algebraic-Geometry Codes," *IEEE Trans. Information Theory*, Vol. 45, No. 6, pp. 1757—1767, Sept. 1999.

[12] J-F. Jourdas and P. Moulin, "A High-Rate Fingerprinting Code," *to appear in SPIE Proceedings*, San Jose, CA, January 2008.