Short c-Secure Fingerprinting Codes

Tri Van Le, Mike Burmester, and Jiangyi Hu

Department of Computer Science
Florida State University
Tallahassee, FL 32306-4530, USA

Abstract. In this paper we consider c-secure fingerprinting codes for copyright protection. We construct a probabilistic fingerprint code and show that at least one colluder in a coalition of up to c users can be traced with high probability. We prove that this code is shorter than the Boneh-Shaw code. In addition, we show that it is asymptotically optimal when c is constant.

Keywords: Fingerprinting, copyright protection, collusion secure.

1 Introduction

Digital fingerprinting is a watermarking technique that protects the intellectual property of multimedia data. It consists of uniquely marking each copy of the data in such a way that the positions marked and their values are kept secret. This marking allows a distributor to detect any unauthorized copy and trace it back to the owner (buyer) of the original copy.

Collusions of dishonest buyers are a major threat to digital fingerprinting. Although the collusion cannot detect the marks where their copies agree, they *can* detect the positions of the marks where the copies differ by comparing copies. Thus they can make pirate copies that either cannot be linked to any particular buyer, or that are linked to other buyers, in which case innocent buyers are framed. A fingerprinting scheme that enables the capture of a member of a collusion of c buyers with probability greater than $1-\varepsilon$ is called c-secure with ε error. In [1, 2], D. Boneh and J. Shaw give a general construction for fingerprinting c-secure codes with ε error. For n possible buyers, given $\varepsilon > 0$, a code with n codewords of length $\ell = 32c^4 \log(2n/\varepsilon) \log(16c^2 \log(2n/\varepsilon)/\varepsilon)$ is constructed allowing at least one of the colluders to be identified with probability at least $1-\varepsilon$.

We present in this paper a construction for binary probabilistic c-secure fingerprinting codes which are shorter than those obtained by using the Boneh-Shaw construction. The basic idea is that each bit position of a codeword (each mark), is assigned a value 0 or value 1 with a certain specified probability. That is, the bits in each codeword are scattered randomly with each bit selected independently.

While it is possible to apply the general construction of Boneh-Shaw for the case when c is constant, such a construction is not very efficient due to

the large length $\ell = 32c^4 \log(\frac{n}{\varepsilon}) \log(\frac{1}{\varepsilon})$ when ε is relatively small. In fact, there are several improvements for the code lengths when c is small. In particular, Sebe and Joancomarti [4] have shown that for $c = 2$, collusion security can be obtained by using dual Hamming codes [7]. These codes have lengths that are much shorter than those of the general construction in [1, 2]. Sebe and Domingo-Ferrer [3] further constructed 3-secure codes that are shorter than the codes in [1, 2]. No other results are known that improve the code length when $c > 3$. In this paper, we extend these results and present a construction for c-secure codes whose length is $\ell = \ln(\frac{n}{\varepsilon})/g(c)$ for any $c \geq 2$, where $g(c)$ depends only on c. For constant c, this improves on the construction of [1, 2] by a factor of $O(\ln \frac{1}{\varepsilon})$. For the special case when $c = 3$, our construction gives codes of length $\ell = 9851 \ln(\frac{n}{\varepsilon})$, which is better than [3] when $n > 6000$.

The paper is organized as follows. We define our notation in Section 2. In Section 3 we present our construction of probabilistic c-secure fingerprinting codes. In Section 4 we show that these codes are optimal. We conclude in Section 5.

2 Model and Definitions

Definition 1. *An (ℓ, n)-code over alphabet Σ is a set of n distinct words in Σ^* that have length ℓ. Each codeword, also called watermark, is to be embedded in a data object to be given to a buyer. The information in the watermark allows its creator to identify the user or users who illicitly distribute forged copies of the watermarked object. In this paper we are concerned with the case $\Sigma = \{0, 1\}$.*

Naturally, users may form coalitions in order to break a watermarking scheme. For example two colluders may run a `diff` on their copies and determine the bit positions where these differ, which of course must belong to the watermark. Therefore, they can damage the watermark by changing these bit positions. Of course, if all users collude then there is nothing we can do. So we assume that each coalition has at most c colluders for some fixed $1 < c < n$, where n is the total number of users. As in [1, 2, 4] we shall assume that the colluders can only modify the bit positions of their codewords that differ. This is known as the Marking Assumption.

Definition 2 (Marking Assumption). *Let $W = \{w_1, \ldots, w_c\}$ be the set of codewords of an (ℓ, n)-code given to a coalition C of c users. Then the coalition C can only produce codewords that belong to $\Gamma(C)$, where:*

$$\Gamma(C) = \{z \in \Sigma^\ell \mid \forall\, i \in \{1, \ldots, \ell\} : (w_1[i] = \ldots = w_c[i]) \Rightarrow (z[i] = w_1[i])\}.$$

The positions i where $w_1[i] = \ldots = w_c[i]$ are called hidden positions. The Marking Assumption tells us that values in the hidden positions cannot be changed by the collusion. For a binary code, we have $\Gamma(C) = \{z \in \Sigma^\ell \mid \forall\, i \in \{1, \ldots, \ell\}, \exists\, j \in \{1, \ldots, c\} : z[i] = w_j[i]\}$.

Definition 3. *A code is totally c-secure if, for any coalition C of at most c users we can identify at least one colluder by using information from the codeword z*

forged by C. That is, there exists a deterministic tracing algorithm A such that, for all colluding strategies of C: $A(z) \in C$. Furthermore, a code is c-secure with ε error if there is a probabilistic tracing algorithm A such that for all colluding strategies of C: $\Pr[A(z) \notin C] \leq \varepsilon$, where the probability is taken over the random coin tosses of A and C.

3 Probabilistic c-Secure Fingerprinting Codes

We now construct a c-secure code for n users with ε error over $\Sigma = \{0,1\}$. Let $\{1,\ldots,n\}$ the set of n users and $p = \frac{1}{c} \in (0,1)$. We define our c-secure codes as follows. Let F_c be an (ℓ, n)-code whose codewords are chosen at random and independently in such a way that:

$$\forall x \in F_c, \forall i \in \{1,\ldots,\ell\} : \Pr[x[i] = 1] = p,$$

where $x[i]$ is the i^{th} bit of x. The code F_c is kept secret. As we will show, despite its simplicity, this code is very efficient at detecting coalition users.

Let $C \subset \{1,\ldots,n\}$ be a coalition and z a codeword illegally constructed by C. Let x_u be the codeword given to user $u \in \{1,\ldots,n\}$. We shall consider the asymmetric Hamming distance $H_0(u,z) = \#\{i \mid x_u[i] = 1 \wedge z[i] = 0\}$. Let u^* be a user for whom $H_0(u^*,z)$ is minimal, that is: $H_0(u^*,z) \leq H_0(u,z)$ for all $u \in \{1,\ldots,n\}$. We shall show that:

Lemma 1. $\Pr[u^* \notin C] \leq e^{-O(\ell)+\ln n}$.

Proof. Let $|C| = c$. For $k \in \{0,\ldots,c\}$, let $\mathcal{B}_0^k(C)$ be the set of all bit positions in which the coalition C sees exactly k bits 0 and $c-k$ bits 1, that is $\mathcal{B}_0^k(C) = i \in \{1,\ldots,\ell\} \mid \#\{u \in C \mid x_u[i] = 0\} = k\}$. Define $d(x,y) = 1$ if $x = 1$ and $y = 0$, and $d(x,y) = 0$ otherwise. For $u \in \{1,\ldots,n\}\backslash C$ and $i \in \{1,\ldots,\ell\}$, let:

$$d_i(u,C) = d(x_u[i], z[i]) - \frac{1}{c}\sum_{v \in C} d(x_v[i], z[i])\,.$$

By definition, we have:

$$\sum_{i=1}^{\ell} d_i(u,C) = H_0(u,z) - \frac{1}{c}\sum_{v \in C} H_0(v,z)\,.$$

We now show that:

$$\Pr\left[\sum_{i=1}^{\ell} d_i(u,C) \leq 0\right] \leq \frac{\varepsilon}{n}\,.$$

Assume that $i \in \mathcal{B}_0^k(C)$ and let $p_0^{ik} = \Pr\left[z[i] = 0 \mid i \in \mathcal{B}_0^k(C)\right]$. Since $u \notin C$, x_u is independent of all x_v with $v \in C$. This means that $x_u[i]$ is independent of $z[i]$ and i. Furthermore, $d(x_u[i], z[i])$ equals 1 when $x_u[i] = 1$ and $z[i] = 0$, and

0 otherwise. On the other hand, $\sum_{v \in C} d(x_v[i], z[i])$ equals $c - k$ when $z[i] = 0$, and 0 otherwise. Therefore the distribution of $d_i(u, C)$ given $i \in \mathcal{B}_0^k(C)$ is:

$x_u[i]$	$z[i]$	$d_i(u,C)$	*probability*
0	0	$\frac{k}{c} - 1$	$(1-p)p_0^{ik}$
0	1	0	$(1-p)(1-p_0^{ik})$
1	0	$\frac{k}{c}$	pp_0^{ik}
1	1	0	$p(1-p_0^{ik})$

Since $Pr[i \in \mathcal{B}_0^k(C)] = \binom{c}{k}(1-p)^k p^{c-k}$, we get:

$$d_i(u,C) = \begin{cases} \frac{k}{c} - 1 & \text{with probability } \binom{c}{k}(1-p)^{k+1}p^{c-k}p_0^{ik},\ 1 \le k \le c; \\ \frac{k}{c} & \text{with probability } \binom{c}{k}(1-p)^k p^{c-k+1}p_0^{ik},\ 1 \le k \le c; \\ 0 & \text{otherwise.} \end{cases}$$

Since $p_0^{i0} = 0$, $p_0^{ic} = 1$ (from the Marking Assumption) and $0 \le p_0^{ik} \le 1$ for all $0 < k < c$, we obtain that $d_i(u,C)$ is bounded below by the random variable D_i, whose distribution is:

$$- \ D_i = \begin{cases} \frac{k}{c} - 1 & \text{with probability } \binom{c}{k}(1-p)^{k+1}p^{c-k},\ 1 \le k \le c-1; \\ \frac{k}{c} & \text{with probability } \binom{c}{k}(1-p)^k p^{c-k+1},\ 1 \le k \le c-1; \\ 1 & \text{with probability } (1-p)^c p; \\ 0 & \text{otherwise.} \end{cases}$$

Note that $p = \frac{1}{c}$, so we have:

$$\begin{aligned} E(D_i) &= (1-p)^c p + \sum_{k=1}^{c-1} \binom{c}{k}(1-p)^k p^{c-k} \left((1-p)(\frac{k}{c} - 1) + p\frac{k}{c} \right) \\ &= (1-p)^c p - \sum_{k=1}^{c-1} \binom{c}{k}(1-p)^k p^{c-k} \left(1 - p - \frac{k}{c} \right) \\ &= (1-p)^c p - (1-p) \sum_{k=1}^{c-1} \binom{c}{k}(1-p)^k p^{c-k} + \frac{1}{c} \sum_{k=1}^{c-1} \binom{c}{k}(1-p)^k p^{c-k} k \\ &= (1-p)^c p - (1-p)\left(1 - (1-p)^c - p^c\right) + \frac{1}{c}\left((1-p)c - (1-p)^c c\right) \\ &= (1-p)p^c. \end{aligned}$$

By applying Lemma 2 (Appendix) to the random variables $X_1, \ldots, X_\ell$ with $\delta = E(D_i)/(1 + E(D_i))$ and $a = -1$, $b = 1$, we get:

$$\Pr\left[\sum_{i=1}^{\ell} D_i \le 0 \right] \le e^{-g(c)\ell},$$

where $(1-\delta)E(D_i) + \delta a = 0$, and

$$g(c) = \frac{1}{3}\delta^2 (E(D_i) - a)/(b - a) = \frac{(1-p)^2 p^{2c}}{6(1 + (1-p)p^c)} \ .$$

Since $D_i \leq d_i(u, C)$, and by the definition of u^*,

$$H_0(u^*, z) \leq \frac{1}{c} \sum_{v \in C} H_0(v, z),$$

we get:

$$\Pr[u^* \notin C] \leq n \Pr[H_0(u, z) = H_0(u^*, z)] \leq e^{-g(c)\ell + \ln n}.$$

Hence Lemma 1 is proven.

We can now state our main theorem.

Theorem 1. *Let $c \geq 2$ be constant. Then for all $\varepsilon > 0$, $n \geq 1$, there is a (ℓ, n)-code with $\ell = O(\ln \frac{n}{\varepsilon})$ that is c-secure with ε error.*

Proof. We will catch any user u whose $H_0(u, w)$ is minimal. Then apply Lemma 1 to get $\Pr[u \notin C] \leq \varepsilon$. Here the constant inside the big O is bounded above by $g(c)^{-1}$.

4 Asymptotic Optimality

Boneh and Shaw proved that $O(c \log \frac{1}{\varepsilon})$ is a lower bound for the length of c-secure codes [1, 2]. This implies that for constant c our codes with length $\ell = O(\ln \frac{1}{\varepsilon})$ are asymptotically optimal, provided that $\varepsilon \leq n^{-a}$ for some constant $a > 0$, which is normally the case. In general, for constant c and arbitrary $\varepsilon > 0$ our codes are more efficient than those obtained from the general construction in [1, 2] by a factor $O(\frac{1}{\varepsilon})$.

In the case when $c = 3$ our code has length $\ell = 9851 \ln(\frac{n}{\varepsilon})$, which is better than [3] for relatively small values of n. For example when $\varepsilon = 10^{-10}$, our code lengths for $n = 8000, 16000, 32000$ are $\ell = 315352, 322180, 329008$, whereas the corresponding code lengths in [3] and [1, 2] are $\ell = 439992, 879984, 1759968$ and $\ell = 5619790, 5742669, 5865627$ respectively.

5 Conclusion

We have presented probabilistic fingerprinting codes that are secure against collusions. The codes are shorter than the Boneh-Shaw codes.

References

1. D.Boneh and J. Shaw, Collusion-secure fingerprinting for digital data, in *Advanced in Cryptology-CRYPTO95*, LNCS 963, pp.452-465, 1995.
2. D.Boneh and J. Shaw, Collusion-secure fingerprinting for digital data, *IEEE Trans. Inf. Theory*, vol IT-44, no. (5), pp. 1897-1905, 1998.
3. Francesco Sebe and Josep Domingo-Ferrer, Short 3-secure fingerprinting codes for copyright protection, *ACISP 2002*, LNCS 2384, pp. 316-327, 2002.

4. J. Domingo-Ferrer and J. Herrera-Joancomarti, Short collusion-secure fingerprints based on dual binary Hamming codes, *Electronics Letters*, vol. 36, no.20, pp.1697-1699, 2000.
5. W. Hoeffding, Probability inequalities for some of bounded random variables, *American Statistical Association Journal*, pp. 13-30, 1963.
6. H. Chernoff, A measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations, *Annals of Mathematical Statistics*, 23:493-509, 1952.
7. F.J. MacWilliams and N.J.A. Sloane, *The Theory of Error-Correcting Codes*, Amsterdam, North-Holland, 1977.

Appendix

Lemma 2. *Let $X_1, X_2, \ldots, X_\ell$ be bounded, independent, identically distributed random variables, with (the same) expected value x and range $[a, b]$. Then the following inequality holds for all $0 < \delta < 1$:*

$$\Pr\left[\sum_{i=1}^{t} D_i \leq \ell((1-\delta)x + \delta a)\right] \leq e^{-\frac{1}{3}\delta^2(x-a)(b-a)^{-1}\ell}.$$

Proof. Let $D_i' = \frac{D_i - a}{b-a}$. Applying the Chernoff-Hoeffding bound [5,6] on ℓ independent, identically distributed, random variables $X_1', X_2', \ldots, X_\ell'$, whose range is $[0,1]$ and whose expected value is $x' = \frac{x-a}{b-a}$, gives us:

$$\Pr\left[\sum_{i=1}^{t} D_i' \leq \ell(1-\delta)x'\right] \leq e^{-\frac{1}{3}\delta^2 x'\ell}.$$

By substituting back $D_i' = \frac{D_i - a}{b-a}$, and $x' = \frac{x-a}{b-a}$ we get:

$$\Pr\left[\frac{(\sum_{i=1}^{t} D_i) - \ell a}{b-a} \leq \ell(1-\delta)\frac{x-a}{b-a}\right] \leq e^{-\frac{1}{3}\delta^2(x-a)(b-a)^{-1}\ell}.$$

Simplifying this inequality gives us:

$$\Pr\left[\sum_{i=1}^{t} D_i \leq \ell((1-\delta)x + \delta a)\right] \leq e^{-\frac{1}{3}\delta^2(x-a)(b-a)^{-1}\ell}.$$

Lemma 2 is now proven.