BOUNDED GAUSSIAN FINGERPRINTS AND THE GRADIENT COLLUSION ATTACK

Darko Kirovski and M. Kivanç Mihçak

Microsoft Research, One Microsoft Way, Redmond, WA, USA

ABSTRACT

The difficulty of building an effective digital rights management system stems from the fact that traditional cryptographic primitives such as encryption or scrambling do not protect audio or video signals once they are played in plaintext. This fact, commonly referred to as "the analog hole," has been responsible for the popularity of multimedia file sharing which cannot be controlled, at least technically, by content's copyright owners. In this paper, we explore a specific issue in multimedia fingerprinting as an answer to "the analog hole" problem. We analyze the collusion resistance of three large classes of spread-spectrum fingerprints using a recently introduced collusion procedure, the gradient attack. Surprisingly, we show that the collusion resistance of direct-sequence and uniformly distributed spread spectrum fingerprints is a small constant that does not depend on the object size, whereas bounded Gaussian fingerprints demonstrate significantly better robustness to the gradient attack.

1. INTRODUCTION

Significantly increased levels of multimedia piracy over the last decade have put the movie and music industry under pressure to deploy a standardized anti-piracy technology. Initiatives, such as the Secure Digital Music Initiative [1], have been established to develop open technology specifications that protect the playing, storing, and distributing of digital music and video. The problem of ensuring copyright of multimedia at the client side lies in the fact that traditional data protection technologies such as encryption or scrambling cannot be applied exclusively as they are prone to digital copying or analog re-recording. The moment the adversary obtains a plain-text digital copy of the multimedia clip, its copyright owners, at least technically, lose control over content's distribution. Thus, almost all modern copyright protection mechanisms tend to rely to a certain extent on watermarks: imperceptive and secret marks hidden in host signals. Two different types of protection systems have evolved over the past decade: content screening [2] and fingerprinting, which is the central focus of our work.

In a typical scenario that uses multimedia marking for forensic purposes, studios create a uniquely marked content copy for each individual user request. User-specific distinct watermarks are commonly denoted as fingerprints. The fingerprinted copy is securely distributed to the user who plays the content using a media player which is unmodified compared to modern media players. Certain users may chose to illegally distribute this content. To address this problem, the media studios deploy search robots in order to find content copies on the Internet. Illegally distributed content is retrieved and based upon the known user database as well as the original clip, media studios use forensic analysis tools to identify the pirates.

Imperceptiveness, robustness, and reliability are the key requirements for fingerprints. One major difference with respect to content screening is that the robustness requirement is significantly easier to satisfy - fingerprint detection is done in the presence of the original clip, not "blindly". Major problem for fingerprinting systems is the collusion attack. To launch such an attack, an adversarial clique of malicious users colludes their copies in order to create a copy which is statistically clean of any fingerprint traces (e.g., the original) or a copy that incriminates another innocent user. Collusion resistance for multimedia content is typically low [3]. Because of this deficiency, fingerprinting systems are commonly restricted to small distribution lists. Finally, one of the most devastating problems for fingerprinting systems is surprisingly, successful identity theft. An adversary with a stolen identity can purchase a multimedia clip and then illegally distribute it, leaving multimedia studios without a target for legal action.

2. SPREAD-SPECTRUM FINGERPRINTS

The media signal to be fingerprinted, $x \in \mathcal{R}^N$, can be modeled as a vector, where each element of x is an i.i.d. Gaussian random variable with standard deviation A, i.e., $x_j = \mathcal{N}(0, A^2)$. We discuss three classes of fingerprints:

- A *class-I fingerprint* $w^{(i)}$, uniquely generated for a specific user i, is defined as a spread-spectrum sequence of N independent identically and uniformly distributed random samples $w^{(i)} \in U[-\delta, +\delta]^N$.
- A *class-II fingerprint* $w^{(i)}$, uniquely generated for a specific user i, is defined as a spread-spectrum sequence of N i.i.d. random samples $w^{(i)} \in \{\pm\delta\}^N$.
- A *class-III fingerprint* $w^{(i)}$, uniquely generated for a specific user i, is defined as a spread-spectrum se-

quence of N i.i.d. random samples $w^{(i)} \in \bar{\mathcal{N}}\{\sigma^2, \delta\}^N$, where $\bar{\mathcal{N}}\{\sigma^2, \delta\}$ denotes a random bounded Gaussian variable with zero-mean, δ as a maximum amplitude, and variance equal to σ^2.

Each element $w_j^{(i)}$ is usually called a "chip." The fingerprinted copy $y^{(i)}$ is created by vector addition: $y^{(i)} = x + w^{(i)}$. Maximum fingerprint amplitude δ is selected as large as possible with two constraints: the fingerprint must be imperceptive and robust with respect to the estimation attack [4].

The forensic detector obtains a modified version z of one or more colluded fingerprinted signals $z = a(\{y^{(i)}, i \in \mathcal{K}\})$. We denote the set of users in the collusion clique as $\mathcal{K}$ with cardinality K and an attack function $a(\cdot)$. Next, let $w \cdot v$ denote the normalized inner product of vectors w and v, i.e., $w \cdot v \equiv N^{-1} \sum w_j v_j$, with $w^2 \equiv w \cdot w$. The fingerprint detector performs a normalized correlation (or matched filter) test:

$$d_T^{(i)} = c(f(z,x) - x, w^{(i)}) = \frac{[f(z,x) - x] \cdot w^{(i)}}{\left(w^{(i)}\right)^2}, \qquad (1)$$

against each user i in the user database $\mathcal{U}$. Function $f(\cdot)$ denotes a pirate-to-original alignment function. As the adversary clique may choose to apply a non-linear geometric bending transform such as the StirMark [6] in addition to the collusion attack, the goal of this function is to perform the realignment of the pirated copy with respect to the original. An example of such a function is presented for both audio and video in [8].

Using a classical Neyman-Pearson hypothesis test, the detector decides that a certain user i has participated in $\mathcal{K}$ if her fingerprint $w^{(i)}$ yields $d_T^{(i)} > \Delta_T$. The detection threshold Δ_T controls the trade-off between the probabilities of false positive and false negative decisions. For example, if $z = y^{(i)}$, then $\mathrm{E}[d_T^{(i)}] = 1$. Also, for $z = y^{(i)}$, we have $\mathrm{E}[d_T^{(j)}] = 0, j \neq i$. Since the noise in the detector is Gaussian due to the Central Limit Theorem for all fingerprint classes, the error probabilities of false negatives and positives are computed by integrating the tail of a corresponding Gaussian probability density function. We recall from modulation and detection theory that the correlation detector is optimal in the class of linear detectors in the presence of i.i.d. noise [5].

3. COLLUSION RESISTANCE – $|\mathcal{K}| \geq 2$

Collusion is usually the most effective effort to defeat fingerprinting schemes. While the estimation attack typically produces a pirated copy of inferior quality, the result of collusion is of equal or even better quality than the distributed content. The adversary can have two types of goals: (*i*) removal of their fingerprints from the pirated copy and (*ii*) framing an innocent user. The latter attack is particularly dangerous because it limits the number of copies the studios can distribute. Once innocent users can be framed, the entire system is rendered dysfunctional. Related work by Boneh and Shaw [3] establishes fingerprint encoding schemes that aim at improving collusion resistance w.r.t. type-(*i*) attacks while reducing robustness to type-(*ii*) attacks.

Since spread-spectrum fingerprints are not prone to type-(*ii*) attacks, we analyze the collusion resistance of the three fingerprint classes with respect to the gradient attack [8]. This is a type-(*i*) attack that drastically reduces the collusion resistance of fingerprinted content for certain fingerprint classes. Although the attack can be generalized to several fingerprint modulation schemes, we focus only on spread-spectrum fingerprints. We first revisit two collusion schemes that have been introduced before [7].

Averaging. $z'_j = \frac{1}{K} \sum_{i=1}^{K} y_j^{(i)}, i \in \mathcal{K}$.

Max-Min. $2z''_j = \max\{y_j^{(i)}\} + \min\{y_j^{(i)}\}, i \in \mathcal{K}$.

The max-min attack is equivalent to the majority attack [3] for *class-II* fingerprints. First, we review the effect of the attacks on *class-I* and *II* fingerprints. We denote as d'_T and d''_T the expected forensic correlation $\mathrm{E}[c(z - x, w^{(i)})]$ for the averaging and max-min attack respectively, where $i \in \mathcal{K}$. We derive the following theorems:

Theorem 1 ***Averaging vs. class-I fingerprints.*** $d'_T = E[c(z' - x, w^{(i)})] = \frac{\delta^2}{3K}$.

Theorem 2 ***Max-min vs. class-I fingerprints.*** $d''_T = E[c(z'' - x, w^{(i)})] = \frac{2\delta^2}{(K+1)(K+2)}$.

Proof: We identify two cases. In the first case, we denote as Y the subset of all positions in the media vector x, where a fingerprint under test, $w^{(t)}$, has the largest value. We can compute the expected correlation at these positions:

$$\begin{aligned} c' &= \mathrm{E}\left[\sum_{j \in Y} v_j w_j^{(t)}\right] = \int_{-\delta}^{\delta} \left(\frac{x+\delta}{2\delta}\right)^{K-1} K x^2 \frac{dx}{2\delta} \\ &= \delta^2 K \frac{K^2 - K + 2}{K(K+1)(K+2)}, \end{aligned} \qquad (2)$$

where $v = z'' - x$ is the attack vector extracted from the attacked clip. In the second case, we identify all other positions in the media vector and denote them as $\bar{Y}$. Similarly, we can compute the correlation of $w^{(t)}$ and the attack vector v at these positions only as:

$$\begin{aligned} c'' &= \int_{-\delta}^{\delta} Kx \left(\frac{x+\delta}{2\delta}\right)^{K-1} \int_{-\delta}^{x} y \frac{dy}{x+\delta} \frac{dx}{2\delta} \\ &= \delta^2 K \frac{2 - K}{K(K+1)(K+2)}. \end{aligned} \qquad (3)$$

From c' and c'', we derive the main claim:

$$d''_T = c' \frac{1}{K} + c'' \frac{K-1}{K} = \frac{2\delta^2}{(K+1)(K+2)}. \qquad (4)$$

■

Theorem 3 ***Averaging vs. class-II fingerprints.*** $d'_T = E[c(z' - x, w^{(i)})] = \frac{\delta^2}{K}$.

Theorem 4 ***Max-min vs. class-II fingerprints.*** $d''_T = E[c(z'' - x, w^{(i)})] = \frac{\delta^2}{2^{K-1}}$.

Theorem 5 ***Max-Min*** $<$ ***Averaging.*** *For any bounded distribution of the fingerprint signal w, two pirated signals z' and z'' created by a collusion clique $\mathcal{K}$ of $K > 2$ users using the averaging and max-min attack respectively, result in corresponding expected correlations $d''_T < d'_T$.*

Proof: (sketch) By approximating the distribution of w as a superposition of infinitesimally narrow uniform distributions and using Theorems 1 and 2, one can derive the main claim in this theorem. ■

Based on Theorem 5, we derive the gradient attack.

Definition 1 ***Gradient attack:*** *$z = z''_j - \beta(z'_j - z''_j)$, where β is such that $E[c(z - x, w^{(i)})] \approx 0$ for each user i in the collusion clique $\mathcal{K}$ and the pirated vector is perceptually close to the original as $||z - x|| \leq \delta\sqrt{N}dB$.*

The rationale behind the attack is simple. For bounded distributions of the fingerprint the max-min attack yields always a better **expected** estimate of the original content than averaging. Thus, we can conclude that we have two points in the N dimensional space, z' and z'', which define a direction $z' - z''$ in the space which is opposite to the direction of all fingerprints in the collusion clique. The adversary can move the result of the max-min attack z'' along this direction as far as imperceptiveness with respect to the original allows. However, the goal of the clique is to achieve invisibility for the forensic analyzer $\mathrm{E}[c(z - x, w^{(i)})] \approx 0$ for each participant.

Theorem 6 ***Efficacy of the gradient attack.*** *In order to set $E[c(z - x, w^{(i)})] = 0$ for fingerprints of class-I and II, the adversary has to chose β' and β'' respectively, equal to:*

$$\beta' = \frac{6K}{(K-1)(K-2)} \text{ and } \beta'' = \frac{K}{2^{K-1} - K}. \quad (5)$$

Figure 1 illustrates the efficacy of the gradient attack. For several collusion clique $\mathcal{K}$ cardinalities $3 \geq K \leq 10$, we apply the gradient attack with β chosen according to Theorem 6 which results in $(\forall i \in \mathcal{K})\mathrm{E}[c(z - x, w^{(i)})] = 0$. The right ordinate quantifies the expected correlation for the averaging and max-min attack for fingerprints of *class-I* and *II*. The left ordinate quantifies the resulting relative noise $\rho(z, y, x) = \mathrm{E}[||z - x||/||y - x||]$ introduced due to the gradient attack for fingerprints of *class-I* and *II*. One can observe that even for $K = 3$, the expected output of the gradient attack is actually of better quality than the distributed marked copy as $\rho(\hat{y}, y, x) < 1$.

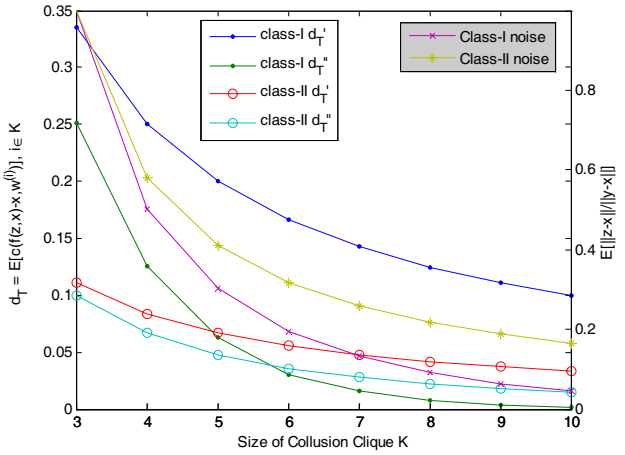

Fig. 1. Efficacy of the gradient attack.

Clearly, the introduced noise is within the imperceptive boundary - hence, the gradient attack is declared successful even for $K = 3$. We conclude that the collusion resistance for the two classes of fingerprints presented in this manuscript, is constant $K = \mathcal{O}(1)$, i.e., invariant of object size. This is significant improvement with respect to averaging and the max-min attack which both enable the forensic analyzer to seek for colluders' traces due to the design of their attack vectors. By using a fingerprint which is sufficiently long, the forensic analyzer can detect all colluders in case of such an attack. The gradient attack removes all traces of the adversarial clique from the perspective of traditional detectors (see Eqn. 1). Finally, the fingerprinting system can be slightly improved by randomizing and hiding certain details (e.g., secret and varying δ) of the fingerprint embedding algorithm.

4. ANALYSIS OF CLASS-III FINGERPRINTS

While fingerprints of class-I and II can be easily cancelled out by the gradient attack, class-III fingerprints approximate the true Gaussian distribution with infinite tails. A Gaussian fingerprint $w^{(i)} = \mathcal{N}(0, \sigma^2)$ results in expected correlations $d'_T = d''_T = \sigma^2/K$ for the averaging and max-min attacks respectively. Hence, the gradient vector $z' - z''$ has no expected correlation with the fingerprint of any participating colluder $E[c(z' - z'', w^{(i)})] = 0, i \in \mathcal{K}$.

Truly Gaussian fingerprints are not practical due to the noise peaks they introduce. Thus, we adopt a bounded Gaussian $\bar{\mathcal{N}}(\sigma^2, \delta)$ for *class-III* fingerprints with the following probability distribution function:

$$p(x) \equiv \begin{cases} 0, & |x| > \delta \\ \frac{\left[1 - \mathrm{erfc}(\delta/\sigma\sqrt{2})\right]^{-1}}{\sigma\sqrt{2\pi}} e^{-\frac{x^2}{2\sigma^2}}, & |x| \leq \delta \end{cases}, \quad (6)$$

which bounds the random variable $w_j^{(i)}$ within $|w_j^{(i)}| \leq \delta$. We denote the resulting variance of this variable as σ_r^2.

In order to evaluate the effect of the gradient attack on *class-III* fingerprints, we adopt several realistic assumptions. First, due to the application of the realignment function $f(\cdot)$, we assume that the extracted signal $f(z,x) = z - x + n$, where n is a noise signal modeled as a normal i.i.d. random variable $n_j = \mathcal{N}(0, \sigma_n^2)$ with $\sigma_n \approx \delta$ due to hi-fidelity piracy. Note that the noise in the fingerprint detector is greater than σ_n. Second, attacker's goal is to set the correlation value of the pirated content $E[c(z-x+n, w^{(i)})], i \in \mathcal{K}$ to τ, where the forensic detector cannot identify the user i as a participant in the collusion clique due to false positives $\varepsilon > \frac{1}{2}\text{erfc}(\tau\sqrt{N}/\sigma_n\sqrt{2})$. Typically, we bound $\varepsilon > 10^{-5}$, so to derive $\tau < \text{erfc}^{-1}(2\varepsilon)\sqrt{2}/\sqrt{N}$. For a realistic range $N \in [10^5, 10^9]$, we conclude $\tau \in [10^{-2}, 10^{-4}]$.

Definition 2 ***Collusion resistance*** *is defined as the size K of the collusion clique $\mathcal{K}$ sufficient to create $z = z'' - \beta(z' - z'')$ such that $E[c(z-x+n, w^{(i)})] \leq \tau, i \in \mathcal{K}$ and the total noise of the attack vector is $||z - x|| \leq \delta\sqrt{N}dB$.*

For presentation simplicity, we evaluate the collusion resistance of *class-III* fingerprints using simulation. In MATLAB, we considered the case when $\sigma \in \{0.1, 0.2, \ldots, 1\}$, $\delta = 1$, and $N = 10^5$. Figure 2 illustrates the simulated results for two extreme cases $\tau = 10^{-2}, N \approx 10^5$ and $\tau = 10^{-4}, N \approx 10^9$. It is important to note that the collusion resistance peaks at a certain σ_r. Higher σ_r makes the bounded Gaussian appear more like uniform distribution resulting in lower collusion resistance, i.e., greater efficacy of the gradient attack. On the other hand, lower σ_r decreases the expected correlation $d_T = \sigma_r^2/K$ at the forensic analyzer, which translates to higher probability of a false positive due to the noise $n = \mathcal{N}(0, \sigma_n)$ in the detector. Note that the peak collusion resistance for long media clips, $N \approx 10^9$, is substantial, $K \approx 180$ for $\sigma_r \approx \sigma = 0.2$. The accuracy of the results obtained for $\tau = 10^{-4}$ in the experiments is subject to certain variance due to relatively short $N = 10^5$ fingerprints used.

5. CONCLUSION

In this paper, we have evaluated three classes of spread-spectrum fingerprints with respect to the recently introduced gradient collusion attack [8]. The advantage of randomly generated spread spectrum fingerprints is their robustness to framing attacks [3]. We demonstrate that certain classes of such fingerprints can provide an efficient collusion resistance. We show that the adversary can cancel out simple direct-sequence (*class-II*) and uniformly distributed (*class-I*) spread spectrum fingerprints using only several content copies. On the other hand, we also show that for bulky multimedia such as hi-quality video, bounded Gaussians are an effective fingerprint choice with carefully selected distribution parameters. For a 10^9-sample fingerprint, considering

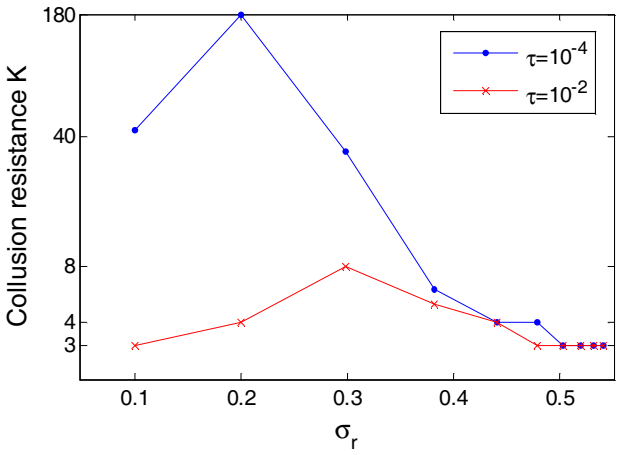

Fig. 2. Collusion resistance of *class-III* fingerprints.

the gradient attack and the amount of noise the adversary can add to the content and still achieve a hi-fi pirated copy, system's collusion resistance of *class-III* fingerprints is on par or better with respect to the resistance achieved using the Boneh-Shaw fingerprint codes [3].

We stress several remaining questions as open problems: (*i*) can the gradient attack be improved by leveraging on the fact that the media signal is quantized using a mid- or high-grain quantizer? (*ii*) is the gradient attack optimal considering a given attack noise level? (*iii*) can we derive fingerprint codes with equivalent estimates regardless of the estimator? and (*iv*) can randomization of the fingerprint generation procedure aid the robustness against the gradient attack?

6. REFERENCES

[1] The Secure Digital Music Initiative. On-line presence at: `http://www.sdmi.org`.

[2] D. Kirovski, H. Malvar, and Y. Yacobi. A dual watermarking and fingerprinting system. *ACM Multimedia*, 2002.

[3] D. Boneh and J. Shaw. Collusion-secure fingerprinting for digital data. *IEEE Trans. on Information Theory*, vol.44, no.5, pp.1897–1905, 1998.

[4] D. Kirovski and H.S. Malvar. Spread-spectrum audio watermarking. *IEEE Transactions on Signal Processing*, 2003.

[5] H.L. Van Trees. Detection, Estimation, and Modulation Theory. Part I, New York: John Wiley and Sons, 1968.

[6] M. Kutter and F.A.P. Petitcolas. A fair benchmark for image watermarking systems. *Security and Watermarking of Multimedia Contents, SPIE*, vol.3657, pp.226–39, 1999.

[7] H. Zhao, M. Wu, Z.J. Wang, and K.J.R. Liu. Nonlinear collusion attacks on independent fingerprints for multimedia. *IEEE ICASSP*, 2003.

[8] D. Schonberg and D. Kirovski. Fingerprinting and Forensic Analysis of Multimedia. ACM *Multimedia*, to appear, 2004.