A Dual Watermarking and Fingerprinting System

Darko Kirovski, Henrique Malvar, and Yacov Yacobi
{darkok,malvar,yacov}@microsoft.com

June, 2001

Technical Report
MSR-TR-2001-57

We present a new dual watermarking-fingerprinting (WM/FP) system, where initially all copies of a protected object are identically watermarked using a secret key, but individual detection keys are distinct. By knowing a detection key, an adversary cannot recreate the original content from the watermarked content. However, knowledge of any one detection key is sufficient for modifying the object so that a detector using that key would fail to detect the marks. Detectors using other detection keys would not be fooled, and such a modified object necessarily contains enough information about the broken detector key – the fingerprint. Our dual WM/FP system limits the scope of possible attacks, when compared to classic fingerprinting systems. Under optimal attacks, the size of the collusion necessary to remove the marks without leaving a detectable fingerprint is superlinear in object size, whereas classic fingerprinting has a lower bound on collusion resistance that is approximately fourth root in object size. By using our scheme one can achieve collusion resistance of up to 100,000 users for a two hour high-definition video.

Microsoft Research
Microsoft Corporation
One Microsoft Way
Redmond, WA 98052
http://www.research.microsoft.com

1. Introduction

With the growth of the Internet, unauthorized copying and distribution of digital media has never been easier. As a result, the music industry claims a $5B annual revenue loss due to piracy, which is likely to increase due to Web communities such as Napster and Gnutella. Legal attempts to alleviate the problem have shown limited success so far, in view of the complexity of the issues involved.

One source of hope for copyrighted content distribution on the Internet lies in technological advances that would provide ways of enforcing copyright in server-client scenarios. Traditional data protection methods such as scrambling or encryption cannot be used, since the content must be played back in the original form, at which point it can always be re-recorded and then freely distributed. One approach for this problem is marking the media signal with a secret, robust, and imperceptible watermark. The media player at the client side can detect this mark and thus enforce the corresponding e-commerce policy. Although the effectiveness of such a system requires global adoption of many standards, the industry is determined to carry out this task [SDMI].

1.1. General.

Watermarks (WM) and fingerprints (FP) are marks hidden in an object for two distinct purposes. The former are used to designate an object as protected, and signal to the client machine that some license is needed in order to use the object. The latter is used to trace piracy to its origins. The detection process of a WM is done in real-time even on small devices, while FPs are detected by powerful machines, that can devote significant resources to the forensic process. Watermarks are identical in all the copies, while FPs are individualized. If necessary, the FP detector can have access to the original unmarked object, using it to improve its likelihood of success in detecting the FPs, even from content modified by malicious attacks.

In this paper, we propose a watermarking system, where as usual, all the copies of a protected object are identically watermarked, but where each user has a distinct secret detection key. All such detection keys are different from the secret embedding key. By gaining the knowledge of a small number of detection keys, an adversary cannot remove the marks from the protected content. We assume that the watermarking system is robust against signal-processing attacks on the protected object and concentrate on collusion attacks against the detection keys. We show that an attacker who has access to one detection key can always fool the corresponding WM detector, but not other WM detectors. Also, in that process, the attacker necessarily inserts a fingerprint in the modified content.

The main entities of the WM/FP system and their interactions are illustrated in Figure 1. In the following sections we quantify the security properties of the proposed scheme:

- Construction of distinct detection keys from a secret watermark key,
- The probabilities of false positive and false negative decisions for detectors using a fixed-length fixed key,
- The size of a collusion clique that would fool: (*i*) a single watermark detector, (*ii*) all WM detectors, and (*iii*) the FP detector respectively, and
- The related probabilities of false positive and negative decisions for the three (*i*-*iii*) respective types of collusion.

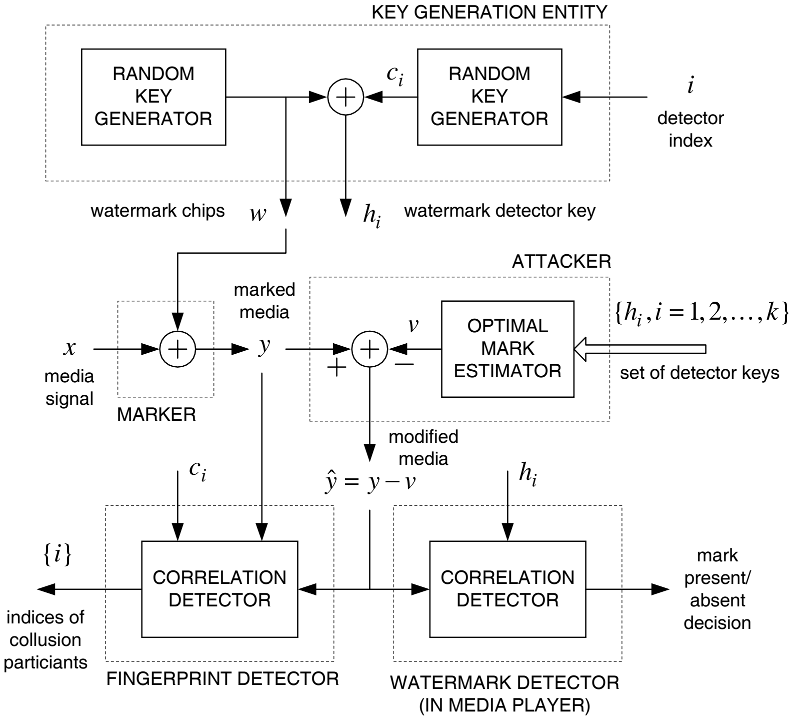

FIGURE 1. General system block diagram for the proposed WM/FP system. Note that each watermark detector i uses a different key h_i. In the attack model, a set of detection keys is colluded to form an estimate v of the watermark w.

A main contribution of this paper is to show that our proposed WM/FP system can achieve a minimum collusion size K that grows linearly with the size N of the marked object. A second contribution is that we can augment our WM/FP system with a segmentation layer. The media content is partitioned into S segments, where a watermark or fingerprint can be reliably detected within each segment. Only detection keys that belong to the same segment can participate in the collusion clique. With segmentation, the minimum collusion size K grows as $O(N \log N)$. Therefore, with or without segmentation, our WM/FP system significantly improves on the best known asymptotic resistance to (fingerprint) collusion attacks of $O(N^{1/4})$ [3]. Since we use a new protection protocol, comparison to classic fingerprinting systems is not completely fair. However, comparison is important from the perspective of building a potential content protection application based upon the two schemes.

1.2. **Previous work.** A survey of watermarking techniques is presented in [8]. We point our reader to review a watermarking technology that succeeds to imperceptibly hide data in audio while being robust with respect to numerous attacks (including sequence desynchronization) specifically designed to prevent detection of spread-spectrum watermarks [9]. In the fingerprinting domain, Ergun et. al. [Erg99] have considered embedding distinct spread-spectrum sequences per copy and have modeled collusion attacks as averaging of copies with additive noise. Boneh and

Shaw [Bon95] have defined a lower bound on the collusion size with a proposal for collusion-secure encoding and an improved "majority attack" model. The previous two works put an upper $O((N/\log(N))^{1/2})$ and lower $O(N^{1/4})$ bound respectively on collusion-secure fingerprinting. Fiat and Tassa [7] introduce a dynamic traitor-tracing mechanism where the set of users is randomly grouped into r subsets, each receiving a distinct symbol. After a subset is identified as the one that includes the pirate(s), the search continues within that subset only. The assumption is that, per round of the tracing process, the pirates simply choose one of the multi-bit symbols available to them. The assumption of [3] is that for the bits where a collusion disagrees, the colluders may choose any value.

2. System Description

In this Section we briefly review the basis of spread-spectrum watermarking, and introduce our WM/FP system.

2.1. **Traditional Spread-Spectrum Watermarking.** The media signal to be watermarked $x \in \mathcal{R}^N$ can be modeled as a random vector, where each element of x is a normal random variable with standard deviation A, i.e. $x_j \sim \mathcal{N}(0, A)$. For example, for audio signals A ranges typically within $A \in \{5, 15\}$, after necessary media preprocessing steps [9]. A *watermark key* w is defined as a spread-spectrum sequence vector $w \in \{\pm 1\}^N$, where each element w_j is usually called a "chip." . The marked signal y is created by vector addition $y = x + w$.

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 $w^2 \equiv w \cdot w$. For example, for w as defined above we have $w^2 = 1$. We assume that the media player contains a watermark (WM) detector that receives a modified version $\hat{y}$ of the watermarked signal y. The WM detector performs a correlation (or matched filter) test $d_W = \hat{y} \cdot w$, and decides that the watermark is present if $d_W > \delta_W$, where δ_W is the detection threshold that controls the tradeoff between the probabilities of false positive and false negative decisions. We recall from modulation and detection theory that such a detector is optimal [12].

Under no malicious attacks or other signal modifications, i.e. $\hat{y} = y$, if the signal y has been marked, then $d_W = 1 + g_W$, where the *detection noise* g_W is a normal zero-mean random variable with variance $\sigma^2_{g_W} = A^2/N$. Otherwise, the correlation test yields $d_W = 0 + g_W$. For equal probabilities of false positives and false negatives, we should set $\delta_W = 1/2$. For robustness against attacks, the signal domain x must be appropriately chosen, and some small modifications on the watermark pattern may be necessary [9]. In this paper we assume that such precautions have been taken care of in the design of the WM detector, so we can disregard media attacks. See [8] for an overview of techniques that use this paradigm for hiding data in audio, images, and video.

2.2. **The Dual WM/FP System.** Traditional spread-spectrum watermarking systems detect watermarks using a key w that is in essence a *secret watermarking key* (SWK). Typically, in many copyright enforcement schemes, the watermark detection is done at the client (the media player), which must then have access to the SWK. An adversary can thus recreate the original content if they succeed in obtaining the SWK, e.g. by breaking into a detector. In our dual WM/FP system, the *watermark detection key* (WDK) is different from the SWK, so breaking into a single detector does not provide enough information to remove the watermark w.

Our WM/FP system is depicted in Figure 1. The media signal x is watermarked in the same way as in traditional spread-spectrum watermarking. However, for each media player i an individualized WM/FP detection key WDK h_i is created from a SWK w in the following way. Let $C = \{c_{ij}\}$ denote an $m \times N$ matrix, where $c_{ij} \in \mathcal{R}$, $c_{ij} \sim \mathcal{N}(0, B)$, i.e. each entry is a zero-mean normal random variable with standard deviation $\sigma_c = B$. Each row i contains a *watermark carrier*, denoted by c_i. The ith WDK is defined as $h_i = w + c_i$. The goal of the watermark carrier c_i is to hide the SWK w in h_i so that knowledge of h_i does not imply knowledge of w, as long as B is large enough. In other words, no player contains the SWK w, but rather a modified version of it. Because the players use a correlation-based WM detection, they should still be capable of detecting the watermark in a marked content y, as long as the number of chips N is large enough to attenuate the noise introduced by the watermark carriers c_i.

The detection process is carried out by correlating the received media file $\hat{y}$ with h_i, generating a detector output $d_W = \hat{y} \cdot h_i$. Similarly to traditional spread-spectrum watermarking, if $\hat{y}$ was marked, then $d_W = 1 + g_W$; otherwise $d_W = 0 + g_W$. The difference is that now g_W is a function of both the media x and the watermark carrier c_i. If there are no attacks, i.e. $\hat{y} = y$, then

$$\begin{aligned} d_W &= y \cdot h_i = (x + w) \cdot (w + c_i) = 1 + g_W, \quad \text{where} \\ g_W &= x \cdot (w + c_i) + w \cdot c_i \end{aligned}$$

from which we compute the detection noise variance as $\sigma^2_{gW} = (A^2 + B^2 + A^2B^2)/N$. We see that the detection noise variance is significantly increased because of the watermark carrier c_i, and so our WM/FP system requires larger N than traditional spread-spectrum, for the same WM detector performance.

2.3. **Copyright Enforcement using WM/FP.** We identify here the main entities in our WM/FP system and describe their roles.

2.3.1. *Watermark Detector (WMD).* As described above, the WMD correlates a potentially marked signal $\hat{y}$ with client WDK h_i, i.e. $d_W = \hat{y} \cdot h_i$. It decides that the content is marked if $d_W > \delta_W$. The probability of false positives (identifying an unmarked content as marked) is denoted as ε_1, which must be very small, e.g. $\varepsilon_1 = 10^{-9}$.

2.3.2. *Attacker.* Breaks K clients and extracts their WDKs $\{h_i, i = 1, \dots K\}$. Creates an attack vector v as an optimal estimate of the SWK w given the collusion key set $\{h_i, i = 1, \dots K\}$, and creates an attacked signal $\hat{y} = y - v$. The closer v estimates w, the more the attacker will clean the watermark in generating $\hat{y}$. We use ε_2 to denote the probability that a watermark chip is incorrectly estimated by the attacker, i.e. $\varepsilon_2 = \Pr[v_j \neq w_j]$. The attacker aims at forcing ε_2 as small as possible, whereas we design the system parameters such that ε_2 is close to 1/2.

2.3.3. *Fingerprint Detector (FPD).* Recovers the attack vector v from an attacked content $\hat{y}$ and the originally marked content y simply by $v = \hat{y} - y$. Unlike the WMDs, the FPD has access to the watermark carrier matrix C. Thus, the FPD correlates v with a suspect watermark carrier c_i, i.e. it computes $d_F = v \cdot c_i$, and decides that the ith client is part of the collusion if $d_F > \delta_F$, i.e. δ_F is the FPD threshold. Compared to the WMD, the FPD has less noise in the correlated vectors, and thus the FPD collusion resistance is much higher than that of the WMD. We use ε_3 to denote the probability of false positives in the FPD, i.e. incriminating a player that was not in the collusion set. Therefore, ε_3 must be very small, just like ε_1). We use η to denote the probability of false negatives at the FPD. We would like it to be small, but do not have to insist that it is as small as ε_1 and ε_3.

3. Attacks Without Collusion

In this section we discuss briefly the kinds of attacks that can be performed on an object with knowledge of at most one WMD key. In the next Section we consider attacks based on a collusion of WMD's, which is the main objective of this paper.

3.1. **Attacks on a Protected Object.** Here we elaborate on a basic assumption for our WM/FP mechanism that we mentioned in the previous section: that there exists a spread-spectrum watermarking mechanism that can be broken only by modifying the marked content beyond the threshold for low fidelity of the attacked copy with respect to the original recording [Kir01]. Typical attacks in this domain range from compression, filtering, resampling, equalization, and various other editing procedures [8], to de-synchronization (or data shifting) techniques that aim at misaligning the embedded spread-spectrum sequence in the content (e.g. the Stirmark attack [2]).

Having a robust watermarking technology is not the only requirement for secure e-commerce of content. Traditional watermarking assumes that the watermarking key (SWK) is hidden at the client side. By breaking a single client, the adversary can create the original content and thus enable all clients to play that content as non-marked. We refer to that as BORE – break once run everywhere. In our WM/FP system, we assume that the attacker will eventually break at least one client and capture that machine's WDK. This can be accomplished by physically breaking the machine (code debugging, reverse engineering) or by using the sensitivity attack [10].

Our scheme is generally BORE-resistant at the protocol level. By breaking a single client, the adversary can play content as non-marked on that broken client, but needs to collude the extracted client WDKs with other clients to finally create content that can play on all players. With our dual WM/FP system, we significantly improve collusion resistance through a fingerprinting mechanism that can identify the members of the clique if its cardinality is smaller than a relatively large lower bound, which is determined in the next section.

3.2. **The Subtraction Attack.** Suppose that an adversary breaks client i and extracts its WDK $h_i = c_i + w$. Then, the adversary can create an attack vector $v = \alpha h_i$ such that the modified media $\hat{y} = y - v$ will produce $E[d_W] = E[\hat{y} \cdot h_i] << \delta_W$, and thus defeating that client's WM detector. To determine α, we note that $d_w = \hat{y} \cdot h_i = [x + w - \alpha(c_i + w)] \cdot (c_i + w) = 1 - \alpha(1 + c_i^2) + x \cdot c_i + x \cdot w + (1 - 2\alpha)c_i \cdot w$, from which we have $\mathrm{E}[d_W] = 1 - \alpha(1 + B^2)$. Thus, by setting $\alpha = (1 + B^2)^{-1}$ we get $\mathrm{E}[d_W] = 0$, that is $d_W = 0 + g_W$. Also, we see that $\sigma_{g_W}^2 \simeq (3 + A^2 + B^2 + A^2B^2)/N$, and that $\sigma_v^2 = \alpha^2(1 + B^2)^{-2} = \alpha << 1$.

Therefore, we see that given knowledge of the client's detection key, the subtraction attack can drive the detector correlation all the way to zero, with just a slight increase in the detector noise $\sigma_{g_W}^2$ and a negligible increase in distortion in the content (since $\sigma_v^2 << w^2 = 1$). If the attacker tries to use a key h_l to break a detector $i \neq l$, it is easy to see that to drive $\mathrm{E}[d_W] = 0$ the attacker would then need to set $\alpha = 1$. However, that would drive $\sigma_v^2 = (1 + B^2) >> 1$, causing too much distortion to the content. Also, it would make σ_{gW}^2 increase by an amount equal to $3B^4/N$, which would make the decisions in the ith WM detector erratic. In other words, even by driving $\mathrm{E}[d_W] = 0$ the ith detector would not be broken with probability much better than 1/2.

3.3. **Resemblance to Public-Key Systems.** We have concluded that if the attacker knows the WDK key h_i of a single detector, that information is not sufficient to break any other detector via the key subtraction attack. Knowing h_i is not enough to infer w, either. In that respect, our dual WM/FP system resembles a public-key cryptosystem, since knowledge of the verification key (in our case h_i) does not imply knowledge of the signing key (in our case w).

4. Collusion Attacks on Detection Keys

Consider a collusion clique of size K that has broken their players and extracted K different WDKs h_i. We now devise the optimal attack based on that set of keys $\{h_i, i = 1, 2, \ldots, K\}$. Without loss of generality, we assume that those extracted WDKs (with indices 1 to K) are the ones in the collusion.

4.1. **The Optimal Attack.** The attacker's job is to estimate the SWK key w by an attack vector v, so that the modified media $\hat{y} = y - v$ will not show significant correlation in any WM detector j, i.e. even for $j > K$. The best job the attacker can possibly perform is given by the following:

Lemma 1. *The optimal attack vector is given by* $v = \operatorname{sign}\left(\sum_{i=1}^{K} h_i\right)$.

Proof. The optimal estimate for each component v_j of the attack vector is clearly given by $v_j = +1$ if $\Pr[w_j = +1|\{h_i\}] \geq 1/2$ and $v_j = -1$ if $\Pr[w_j = +1|\{h_i\}] < 1/2$. That estimate is optimal because it minimizes $\Pr[v_j \neq w_j]$.

Since $h_{ij} = w_j + c_{ij}$, where the c_{ij} are independent and normally distributed, we can write $\Pr[w_j = +1|\{h_i\}] = 1/(1 + \nu_j)$, where $\nu_j = \prod_{i=1}^{K} p_c(h_{ij} + 1)/p_c(h_{ij} - 1)$ and $p_c(\zeta) = (2\pi)^{-1/2} \exp[-\zeta^2/(2B^2)]$. Thus, we can write $\nu_j = \exp(-2\rho_j/B^2)$, where $\rho_j = \sum_{i=1}^{K} h_{ij}$. Thus, $\Pr[w_j = +1|\{h_i\}] \geq 1/2]$ when $s_j \geq 0$ and $\Pr[w_j = +1|\{h_i\}] < 1/2]$ when $s_j < 0$. □

4.2. **WMD Performance.** Given the optimal attack above, we can compute the average estimation error in the attack vector, $\varepsilon_2 = \Pr[v_j \neq w_j]$, by the following:

Lemma 2. *A collusion of size K produces*

$$\varepsilon_2 = \frac{1}{2} \operatorname{erfc}\left(\frac{\sqrt{K}}{B\sqrt{2}}\right) < \frac{1}{2} \exp\left(-\frac{K}{2B^2}\right)$$

Proof. Since the w_j chips are equally likely to be +1 or −1, it is clear that $\varepsilon_2 = (\Pr[s_j \geq 0|w_j = -1] + \Pr[s_j < 0|w_j = +1])/2$. Because of symmetry, this simplifies to $\varepsilon_2 = \Pr[s_j \geq 0|w_j = -1]$. Since for $w_j = -1$ we have $s_j = -K + \bar{c}_j$, where $\bar{c}_j = \sum_{i=1}^{K} c_{ij}$. Therefore, $\varepsilon_2 = \Pr[\bar{c}_j \geq K]$, where $\bar{c}_j$ has a normal distribution with mean zero and variance $\sqrt{K}B$. From the definition of the complementary error function erfc(·) and its relation to the normal probability distribution, the result follows [1]. The inequality follows from the well-known upper bound $\operatorname{erfc}(u/\sqrt{2}) < \exp(-u^2/2)$, for $u > 0$ [1]. □

Given ε_2, we can evaluate the efficiency of the subtraction attack $\hat{y} = y - v$ for the optimal attack vector v. Since $\mathrm{E}[v \cdot w] = \Pr[v_j = w_j] - \Pr[v_j \neq w_j] = 1 - 2\varepsilon_2$, it is easy to see that after attack the expected output of the WM correlation detector drops to $\mathrm{E}[d_W] = 2\varepsilon_2$. The attacker may attempt a stronger subtraction attack, of the form $\hat{y} = y - \beta v$, with $\beta > 1$, because that would bring the WMD output further down to $\mathrm{E}[d_W] = 2\beta\varepsilon_2 - (\beta - 1)$. As long as β is not too large, the attacked content $\hat{y}$ may be acceptable to users.

4.3. **Collusion Size.** For a desired attack efficiency, we can determine the necessary collusion size by the following:

Lemma 3. *In order to reduce the correlation value to* $E[d_W] = \theta$, *the adversary needs to collude* K *WDKs, with*

$$K = 2B^2 \left[\mathrm{erf}^{-1} \left(\frac{1-\theta}{\beta} \right) \right]^2$$

Proof. Follows directly from Lemma 2 and the fact that in order to reduce the expected correlation value to $\mathrm{E}[d_W] = \theta$, the adversary needs to achieve an attack vector error rate of $\varepsilon_2 = (\theta + \beta - 1)/(2\beta)$ through collusion. We see that for fixed θ and β the minimum collusion size grows proportional to B^2. □

Example 1. *For* $B = 10$, $\theta = 0.25$ *and* $\beta = 2$, *the attacker must collude at least* $K = 24$ *keys. For* $\beta = 1$, *the attacker must collude at least* $K = 133$ *keys.*

We note that the attacker needs to set θ much smaller than δ_W, otherwise the probability that a WMD will still detect the watermark is not low enough to justify the attacker's effort. In other words the attack is successful only if it makes $\epsilon_1 \simeq 1$. For that it is not necessary to set θ all the way to zero, because it would require K to be excessively large. By setting $\beta > 1$, though, it is possible to force $\theta = 0$.

To make the attacker's job more difficult, we need to increase the parameter B, the standard deviation of the watermark carrier c, since K grows with B^2. In doing so, however, we increase the detection noise variance $\sigma_{g_W}^2 = (A^2 + B^2 + A^2B^2)/N$, where we recall that A is the standard deviation of the original content x and N is the object size. For a given σ_{g_W}, we can determine that the probability of false positives $\varepsilon_1 = \Pr[d_w > \delta_w \,|\, \text{object is not marked}]$ is given by:

Lemma 4. *An object of size* N *produces*

$$\varepsilon_1 = \frac{1}{2} \,\mathrm{erfc} \left(\frac{\delta_W \sqrt{N}}{\sqrt{2(A^2 + B^2 + A^2B^2)}} \right) < \frac{1}{2} \exp \left(- \frac{\delta_W^2 N}{2(A^2 + B^2 + A^2B^2)} \right)$$

Proof. Follows immediately from the fact that the WMD noise g_W is normal with zero mean and variance $\sigma_{g_W}^2 = (A^2 + B^2 + A^2B^2)/N$. □

We note that if $\delta_W = 1/2$, then ε_1 is also the probability of false negatives, i.e. the probability of a WMD not detecting a marked object that was not attacked.

From the result above we can compute N by:

Corollary 1. *The object size* N *required to achieve a given* ε_1 *is*

$$N = \frac{2[A^2 + B^2(1 + A^2)]}{\delta_W^2} \left[\mathrm{erf}^{-1}(1 - 2\varepsilon_1) \right]^2$$

By combining the result above with that in Lemma 3 we arrive at one of the main results in this paper:

Theorem 1. *Collusion size K grows linearly with object size N, i.e. $K \sim O(N)$.*

Proof. As N grows, for a given ε_1, B also grows, and thus $\sigma_{gW}^2 \to B^2(1+A^2)/N$. Combining this asymptotic expression for σ_{gW} with the results in Corollary 1 and Lemma 3, we get

$$K = N\frac{\delta_W^2}{1+A^2}\left[\frac{\operatorname{erf}^{-1}\left(\frac{1-\theta}{\beta}\right)}{\operatorname{erf}^{-1}(1-2\varepsilon_1)}\right]^2.$$

The equation above allows us to quickly compute the object size N necessary to achieve any desired collusion resistance K. □

It is important to note that the result above is so far determined only by the WMD performance. In the next Section we will confirm the linear relationship between K and N when considering the FPD performance.

5. Fingerprint Detection

As we mentioned in Section 2, the FPD has less noise in its correlation output. Therefore, it should be able to identify the indices i corresponding all the WDKs h_i used in the collusion by the attacker, even if the collusion size K is large enough to fool all clients, as computed above. In this section we evaluate the error probabilities for the FPD.

We recall that the FPD knows the marked content y, the attacked version $\hat{y}$, and the watermark carriers c_i. It computes the correlation $d_F = (\hat{y}-y)\cdot c_i$, and decides that the the ith client participated in the collusion if $d_F > \delta_F$. Assuming the attack model of the previous section, $\hat{y} = y - \beta v$, the FPD output can be written as

$$d_F = \ (\hat{y}-y)\cdot c_i = \beta(v\cdot c_i) = \mathrm{E}[d_F] + g_F$$

where g_F is the zero-mean FPD correlation noise. The most critical error for the FPD is a false positive, i.e. incriminating a WDK i that did not participate in the collusion. The probability ε_3 of that error is given by the following:

Lemma 5. *An object of size N produces*

$$\varepsilon_3 = \frac{1}{2}\operatorname{erfc}\left(\frac{\delta_F\sqrt{N}}{\sqrt{2}\beta B}\right) < \frac{1}{2}\exp\left(-\frac{\delta_F^2 N}{2\beta^2 B^2}\right)$$

Proof. If c_i is not in the collusion, it is independent of the attack vector βv. Therefore, $\sigma_{g_F}^2 = \mathrm{E}[\beta^2 v_{ij}^2 c_{ij}^2]/N = \mathrm{E}[\beta^2 c_{ij}^2]/N = \beta^2 B^2/N$. The result follows from $\varepsilon_3 = \Pr[g_F > \delta_F]$ and the fact that g_F has normal distribution. □

It is clear that, as expected $\varepsilon_3 << \varepsilon_1$ (usually by several orders of magnitude), since the argument in $\operatorname{erfc}(\cdot)$ for ε_3 is approximately $(A\delta_F)/(\beta\delta_W)$ times larger than the argument in $\operatorname{erfc}(\cdot)$ for ε_1. Thus, by choosing B and N for a sufficiently low ε_1, we achieve a negligibly low probability ε_3 of false positives in the FPD.

To compute the detection performance of the FPD we need to determine its expected output when we correlate with a carrier c_i such that h_i was part of the collusion. We see that $\mathrm{E}[d_F] = \beta \mathrm{E}[z_j]$, where

$$z_j = v_j c_{ij} = \mathrm{sign}\,[s_j]\, c_{ij}, \text{ with } s_j = w_j + b_j, \text{ and } b_j = \frac{1}{K}\sum_{m=1}^{K} c_{mj}$$

Lemma 6. *A collusion of size K produces*

$$E[d_F] = \beta \frac{B}{\sqrt{K}} \sqrt{\frac{2}{\pi}} \exp\left(-\frac{K}{2B^2}\right)$$

Proof. It is clear that $\mathrm{E}[z_j] = (\mathrm{E}[z_j|w = +1] + \mathrm{E}[z_j|w = -1])/2$, since the w_j chips are equally likely. Also, because of the symmetry of the problem we see that $\mathrm{E}[z_j|w = +1] = \mathrm{E}[z_j|w = -1]$, and so $\mathrm{E}[z_j] = \mathrm{E}[z_j|w = +1]$.

Assuming $w_j = +1$, $\mathrm{E}[z_j] = \mathrm{E}[z_j|s_j \geq 0]\mathrm{Pr}[s_j \geq 0] + \mathrm{E}[z_j|s_j < 0]\mathrm{Pr}[s_j < 0] = \mathrm{E}[c_{ij}|s_j \geq 0]\mathrm{Pr}[s_j \geq 0] - \mathrm{E}[c_{ij}|s_j < 0]\mathrm{Pr}[s_j < 0]$. Under each of the conditions $s_j \geq 0$ or $s_j < 0$, we see that $s_j = 1 + b_j$ and c_{ij} are all jointly-normal variables, with variances $\sigma_s^2 = \sigma_b^2 = B^2/K$ and $\sigma_c^2 = B^2$. Furthermore, the correlation coefficient between b_j and c_{ij} is equal to one, since c_{ij} is part of the average that defines b_j. Thus, computing the conditional expectations above is just an exercise of computing expectations of a normal random variable, conditioned on minimum or maximum values for that variable. □

Given the expected FPD output, we usually set $\delta_F = \mathrm{E}[d_F]/2$, which determines the probability η of false negatives, i.e. the probability that a key index i in the collusion will not be detected. The result is given by the following

Lemma 7. *An object of size N produces*

$$\eta = \frac{1}{2}\,\mathrm{erfc}\left(\frac{(E[d_F] - \delta_F)\sqrt{N}}{\sqrt{2}\beta B}\right)$$

Proof. Follows immediately from the fact that the FPD output d_F is normal with expected value $\mathrm{E}[d_F]$ and variance $\sigma_{d_F}^2 = \sigma_{g_F}^2 = \beta^2 B^2/N$. □

Example 2. *For $A = 4$, $B = 20$, $\theta = 0.1$ and $\beta = 1$, the collusion size is $K = 592$, with $\varepsilon_2 = 0.11$. With $N = 10^6$, we achieve $\varepsilon_1 = 7 \times 10^{-10}$, $\varepsilon_3 = \eta = 2.6 \times 10^{-15}$.*

From the result above we can compute the object size N necessary to achieve a desired probability η of false negatives in the FPD. For simplicity, let's assume that we set the FPD threshold in the middle, i.e. $\delta_F = \mathrm{E}[d_F]/2$. We recall from the previous Section that the minimum collusion size is $K = 2B^2\mu^2$, where $\mu = \mathrm{erf}[\beta^{-1}(1-\theta)]$ is fixed for a fixed attack efficiency (i.e. a fixed θ). Therefore, as we increase B the attacker has to increase K proportionally to B^2. We can use that into the equation for η above to obtain the following

Lemma 8. *The object size N required to achieve a given η is*

$$N = K\frac{\pi}{2}\left[\mathrm{erf}^{-1}(1-2\eta)\exp(\mu^2)\right]^2$$

Proof. Since $K = 2B^2\mu^2$, it is easy to see that $\delta_F = \beta\mu^{-1}\pi^{-1/2}\exp(-\mu^2)$, and the result follows by simple substitution. □

This result thus confirms Theorem 1, i.e. that collusion size and object size are linearly related. We note that in fixing the WMD performance we obtained one constant of proportionality, whereas in fixing the FPD performance we obtained another. Therefore, in designing a practical system we determine the desired error probabilities, and select N as the largest of the values computed from the WMD and FPD equations.

6. Segmentation

6.1. **General.** In the WM/FP system, watermarks protect the content and fingerprints enable the copyright owner to identify a clique of users that launched an attack to remove the watermark. This unique property of the protection system, enables us to add multiple watermarks in the object and enforce the adversary to create cliques independently for each watermark. More formally, we divide the protected object into S segments $S_s, s = 1...S$, and watermark each of them with a distinct spread spectrum sequence $w_s, s = 1..S$. Per each segment S_s, we use m distinct WDKs $h_i^{[s]}, i = 1..m$, created in accordance to the described dual WM/FP system. Each client gets a single WDK $h_i^{[s]}$ to exactly one segment.

6.2. **Theoretical system.** In both the theoretical and practical cases, object and collusion of any realistic size result in a probability of false positives (ε_3) close to zero such that it can be neglected. Because of this, we conveniently conclude that "a segment can resist K colluders" without mentioning error probabilities.

Definition 1. *A protected object is defeated if watermarks are removed from all segments, while no fingerprints are introduced in the process. The collusion-resistance κ_s of a segmented WM/FP system with S segments equals the* expected *number of users needed to use their WDKs in S collusion cliques (a clique per segment) to defeat the system.*

Lets denote as q the probability that after distributing κ_s keys into segments, no segment contains less than K keys. Lets also adopt the following assumptions: $S \gg 1$, K is a relatively small constant, and $\frac{\kappa_s}{SK} \gg 1$.

Theorem 2. *If $\kappa_s = S[\ln(S) - \ln(2\varepsilon_4)]$ then $q > 1 - \varepsilon_4$.*

A sketch of the proof is presented in Appendix A.

7. Key Compression

7.1. **The problem.** The major drawback of the basic dual system is the requirement for a relatively large storage space for the detection keys. A brief problem overview: it is hard to compress the sum of two independent pseudo-random sequences, such that it is hard to infer the individual sequences. Let $g(s, n)$ denote the output of length n of generator g, given seed s. We need a way to create two generators g_1, g_2 with two seeds s_1, s_2 such that $\exists(g, s) \mid g_1(s_1, n) + g(s, n) = g_2(s_2, n)$ and the sequences $g_1(s_1, n)$ and $g(s, n)$ are mutually independent. This remains as an open problem. The current situation is that we need to create $g_1(s_1, n)$ and $g(s_2, n)$ independently in a secure machine and store their sum on a client. For realistic loads to the system, the length of the key is in the order of 10^5 bytes, which may be too much data for certain embedded devices.

7.2. **Proposed Solution.** Recall that the WDK of user i is created as $h_i = c_i + w$, where c_i and w are mutually independent. Instead, we can generate the key from a short seed using any standard cryptographically secure pseudo-random key generator, and per chosen w do sieving and select only those seeds for which the resulting long sequence (lets denote it as s) has the property that $s \cdot w \geq 1$, thus, inferring $h_i = s$. The deviation of $s \cdot w$ is roughly $\sigma^* = B\sqrt{N_o}$, so the probability for a randomly chosen seed to meet this criteria is $\varepsilon^* = \frac{1}{2}\mathrm{erfc}(N_o/(B\sqrt{2}))$. For example, for $\varepsilon^* < 10^{-6}$ we get $N_o = 2B^2[\mathrm{erfc}^{-1}(2\varepsilon^*)]^2 = 2000$. Since $N = 10^5$, we partition the generation of h_i into N/N_o segments, where for each segment we perform sieving expected $1/\varepsilon^*$ times. For a seed size of $\xi = 100$ bits, we obtain a compression ratio of $N_o/\xi \sim 20$.

8. WM/FP Summarizing Discussion

The dual WM/FP technology aims at building practical secure content protection mechanisms. Although the main underlying theoretical concepts have been presented so far, in this section we focus on their interaction and practical implications. Solid overview of the mutual impact of scheme parameters can be obtained from Table 1. When designing a realistic protection system, several parameters are given as constants: total object size N_o and media variance A^2. All other parameters can be chosen such that the overall detection mechanisms are of desired quality.

WM/FP Parameter	Related Parameter Dependencies
$\varepsilon_1 = \Pr[d_w > \delta_w \mid \text{object is not marked}]$	$\sim \mathrm{erfc}\left(\frac{\sqrt{N}}{AB}\right)$
Segment length: N	$\sim B^2A^2\left[\mathrm{erf}^{-1}(1-2\varepsilon_1)\right]^2$
$\varepsilon_2 = \Pr[v_j \neq w_j]$	$\sim \mathrm{erfc}\left(\frac{\sqrt{K}}{B}\right)$
Collusion resistance per segment: K	$\sim N$
ε_3	$\sim \mathrm{erfc}\left(\frac{\sqrt{N}}{B}\right)$
System collusion resistance: κ_S	$\sim S(\ln(S) + K)$

Table 1. Dependencies among main parameters of the WM/FP system.

The primary decision is to determine the number of segments S per object. Since collusion resistance within a single segment is $K \sim N$, where $N = N_o/S$ is the length of the segment, and collusion resistance achieved over S segments is $\kappa_S \sim S\ln(S)$ for small K, then the objective is to have as short as possible segments in order to: (i) maximize overall collusion resistance κ_S and (ii) reduce the storage space for a single WDK H_i. On the other hand, due to security measures for hiding w within a watermark carrier c_i, there exists a lower bound on the watermark carrier amplitude B, commonly set to $B \geq A$. Selection of B uniquely identifies the segment length N with respect to a desired probability of a false alarm ε_1 under the optimal sign(mean(h)) attack. Such a setup directly impacts the maximal collusion size per segment K and maximal efficacy of the adversary in guessing SWK bits $1 - \varepsilon_2$. It also traces the guidelines for FPD detection performance ε_3 and η.

For realistic loads to the system, such as high-definition television, the number of bits per object ranges in the order of 10^{11} bytes. Assuming, one chip is embedded per 100 pixels, we derive an object size of $N \approx 10^9$ chips. On the other hand, from $B = A \approx 7$ and $\varepsilon_1 = 10^{-9}$, we derive $N \approx 10^5$ chips. This boosts the number of segments to $S \approx 10^4$. The resulting error probabilities are: (i) desired likelihood of an incorrectly guessed w_j bit during the $Sign(Mean())$ attack of $\varepsilon_2 < 0.40$ can be achieved for $K = 3$ and (ii) the detection accuracy of the fingerprint detector is better than $1 - \varepsilon_3 > 1 - 10^{-112}$. Most importantly, the achieved collusion resistance is lower-bounded by $\kappa_S > 10^5$ users. One can hardly expect that, under realistic piracy scenarios, such a clique could be established to oppose the protection of the proposed dual WM/FP system.

9. Conclusion

We have introduced a new dual WM/FP system, where all copies of a protected object are identically watermarked using an SWK, but where individual WDKs are distinct. By knowing a WDK, an adversary cannot recreate the original from the marked content. However, knowledge of any one WDK is sufficient for modifying the object so that a detector using that key does not detect marks. Such a modified object necessarily contains a fingerprint: sufficient information to point at the WDK used to break the detector.

Our dual WM/FP system limits the scope of possible attacks, when compared to classic fingerprinting systems. Under optimal attacks, the size of the collusion necessary to remove the marks without leaving a detectable fingerprint is asymptotically $K \sim O(N)$ without segmentation, and $\kappa_S \sim O(N \ln(N))$ with segmentation, where N denotes object size. Classic fingerprinting has a lower bound on collusion resistance that is roughly $O(N^{1/4})$. Thus, by using the dual WM/FP system one can achieve content protection with collusion resistance of up to 100,000 users for a two-hour high-definition video, for example.

References

[1] N, Alon, J. H. Spencer and P. Erdős. The Probabilistic Method. Wiley-Interscience series in Discrete Mathematics and Optimization. New York: Wiley, 1992.

[2] R. J. Anderson and F. A. P. Petitcolas, "On the limits of steganography," IEEE J. Selected Areas in Communications, vol.16, pp. 474-481, 1998.

[3] D. Boneh, and J. Shaw, "Collusion secure fingerprinting for digital data," IEEE Transactions on Information Theory, Vol 44, pp. 1897–1905, 1998. Extended abstract in Proc. Crypto'95, LNCS Vol. 963, pp. 452–465. New York: Springer-Verlag, 1995.

[4] B. Chor, A. Fiat, and M. Naor, "Traitor tracing," Proc. Crypto'94, LNCS Vol. 839, pp. 257–270. New York: Springer-Verlag, 1994.

[5] F. Ergun, J. Kilian, and R. Kumar, "A Note on the Limits of Collusion-Resistant Watermarks," Proc. Eurocrypt, 1999.

[6] W. Feller. An introduction to probability theory and its applications, 3rd ed. New York: Wiley - Series in Probability and Mathematical Statistics, 1968.

[7] A. Fiat and T. Tassa, "Dynamic traitor tracing," Proc. Crypto'99, LNCS Vol. 1666, pp. 354-371. New York: Springer-Verlag, 1999.

[8] S. Katzenbeisser and F. A. P. Petitcolas, Eds. Information Hiding Techniques for Steganography and Digital Watermarking. Boston, MA: Artech House, 2000.

[9] D. Kirovski and H. S. Malvar, "Robust spread-spectrum audio watermarking," Proc. IEEE Int. Conf. Acoustic, Speech, and Signal Processing, Salt Lake City, UT, May 1999.

[10] J.-P. M. G. Linnartz and M. Van Dijk, "Analysis of the sensitivity attack against electronic watermarks in images." Workshop on Information Hiding, pp. 258–72, 1998.

[11] Secure Digital Music Initiative. See `http://www.sdmi.org`.

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

[13] Y. Yacobi, "Improved Boneh-Shaw fingerprinting," RSA Conf. 2001, to appear. San Francisco, CA, 2001.

Appendix A

We have S cells and m pebbles thrown uniformly at the segments. We want to find m that will guarantee that with high probability each cell has at least c pebbles. Let x denote the number of pebbles in a given cell. $E[x] = m/S$ and $var[x] = \frac{(S-1)m}{S^2}$. We can model this even as throwing a fair dice with S sides. The attacker aims at having no cells contain less than K pebbles. This is a tail probability at $m/S - K$ from the mean. Assuming normal distribution, this event has probability of occurrence $p = \frac{1}{2}[1 - \text{erf}(\frac{(m-SK)}{\sqrt{2(S-1)m}})] < \exp(-\frac{(m-SK)^2}{2(S-1)m})$.

Let $m = SuK$, where $u \gg 1$. Assume that S is large enough so that $S \gg 1$. Then $p = \frac{1}{2}[1 - \text{erf}(\sqrt{uc/2})] < \exp(-uc/2)$.

Since m and S are large and K is relatively small, then we can assume independence between the tail events of the different cells. Let $q = \Pr[$no cell has less than K balls$]$. Then $q > (1-p)^S \approx 1 - Sp$. In order to obtain $q \approx 1$, the attacker needs $p \ll 1/S$.

Theorem 3. *If* $m = Sln(\frac{S}{2\varepsilon_4})$ *then* $q > 1 - \varepsilon_4$.

Proof. $p < \exp(-uc/2)$, and $\frac{uK}{2} = \frac{SuK}{2S} = \frac{m}{2S} = \frac{2S\ln(\frac{S}{2\varepsilon_4})}{2S} = \ln(\frac{S}{2\varepsilon_4})$. So, $\exp(-uc/2) = \frac{\varepsilon_4}{S}$. $q > 1 - Sp > 1 - S\frac{\varepsilon_4}{S} = 1 - \varepsilon_4$. □