A Dual Watermark-Fingerprint System

Darko Kirovski, Henrique Malvar, and Yacov Yacobi
Microsoft Research

A dual-purpose watermarking and fingerprinting system for multimedia screening uses the same secret key to mark all content copies, but different detection keys within each media player. Under optimal attacks, the system's collusion resistance is superlinear in object size.

The Internet's growth has made the unauthorized copying and distribution of digital media easier than ever before. As a consequence, the music industry claims an enormous annual revenue loss due to piracy (see http://www.riaa.org)—a loss allegedly exacerbated by file-sharing Web communities. As Internet bandwidth increases, piracy is likely to affect the movie industry in a similar manner. Legal attempts to alleviate the problem have had limited success so far.

One source of hope for copyrighted content distribution on the Internet lies in technological advances that would allow copyright enforcement in both client-server and peer-to-peer scenarios. Traditional data-protection methods such as scrambling or encryption are minor hurdles to attackers because they can always rerecord the content and freely distribute it. This problem is commonly referred to as the *analog hole*. One proposed solution is to hide within the media signal a secret, robust, and imperceptible watermark. A watermark designates a multimedia clip as protected. Before playing the clip, the client machine searches for watermark existence within the clip. Watermark detection signals to the client machine that playing the clip requires a license. Recent efforts have sought to define standards for protecting music content using watermarks with little success (see http://www.sdmi.org). The main vulnerability of watermark-based content-screening systems is that each client machine must store the same key used to mark the content. By breaking a single client, the adversary inherently breaks the entire system. In addition, watermark detection in this scenario must be performed blindly (that is, without the original recording) and in real time, even on small devices. To date, no technology has been able to robustly detect watermarks under such circumstances.

An alternative to content screening is content fingerprinting. In this scenario, copyright owners distribute a uniquely watermarked content copy to each client. Clients can play or redistribute the content without any barriers. Copyright owners scan public distribution channels for illegally distributed clips. Each identified pirated clip is analyzed for the existence of one or more fingerprints, which are used to trace piracy to its origin. Detecting fingerprints usually requires powerful machines that can devote significant resources to the forensic process. A fingerprint detector can access the original unmarked object and use it to detect fingerprints, even from content modified by malicious attacks. The main vulnerability of this forensic protection scenario is collusion. A clique of malicious users can collude their copies and create a new copy that is statistically clean of any traces that might point to any of the colluders. Typically, the collusion resistance of even the best fingerprint-encoding schemes is low[1] (for example, less than one hundred for a two-hour high-definition video clip).

Dual watermark-fingerprint system

This article proposes a multimedia content protection system in which all copies of a protected object are identically watermarked, but each user has a distinct secret detection key that differs from the secret embedding key. An attacker with access to one detection key can fool the corresponding watermark detector but not other watermark detectors. Surprisingly, analogous to a criminal action, during this attack the attacker necessarily inserts his or her fingerprint into the modified content. Even a collusion clique of relatively large size cannot entirely remove the secret marks from the protected content by colluding their detection keys. More importantly, if the clique is not large enough, traces of the detection keys of all colluders can be detected with relatively high accuracy in the attacked clip. Figure 1 (next page) illustrates the main entities of our dual watermark-fingerprint system and their interactions.

Our proposed watermark-fingerprint system achieves a minimum collusion size K that grows

1070-986X/04/$20.00 © 2004 IEEE

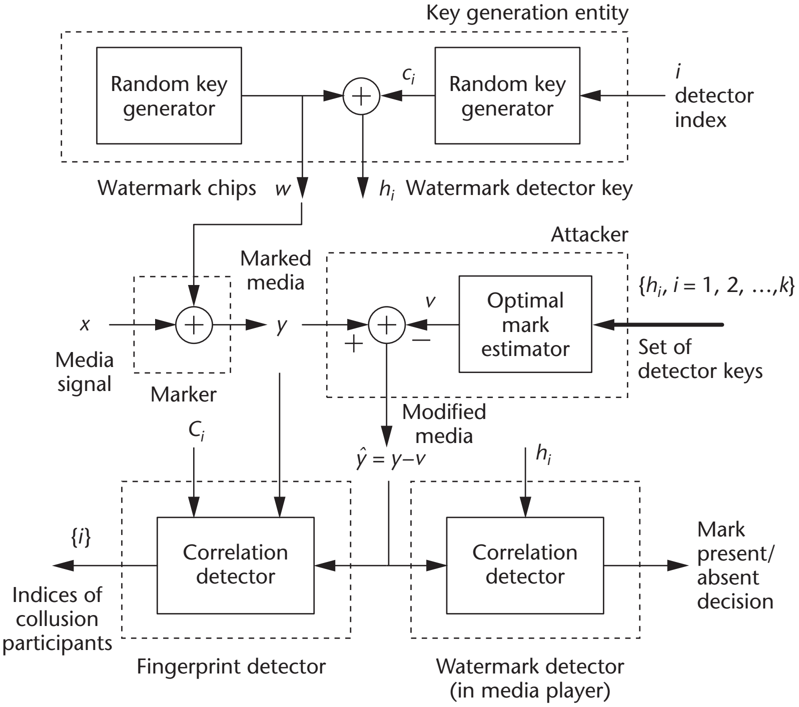


***Figure 1. General system block diagram for the proposed watermark-fingerprint system. Note that each watermark detector i uses a different detection key* h$_i$*. In the attack model, a set of detection keys is colluded to form an estimate* v *of the watermark* w.**

linearly with the size N of the marked object. In addition, we can augment our watermark-fingerprint system with a segmentation layer. The media content is partitioned into S segments, in which media players as well as forensic analyzers can reliably detect a watermark or fingerprint. Only detection keys that belong to the same segment can participate in the collusion clique. With segmentation, the minimum collusion size K grows as $\mathcal{O}(N \log N)$. Therefore, with or without segmentation, our watermark-fingerprint system significantly improves on the best-known asymptotic resistance to (fingerprint) collusion attacks of about $\mathcal{O}(N^{1/4})$.[1] Because we use a new protection protocol, comparing our system to classic fingerprint systems might seem unfair. However, such a comparison is important because the two technologies share a common goal: multimedia copyright enforcement.

Our aim in this article is to characterize the collusion attacks against this system under the assumption that watermark detection is robust against signal-processing attacks on the protected object. To the best of our knowledge, no modern watermark technology demonstrates such robustness. To date, attacks such as Stirmark,[2] the estimation attack,[3] and the swap attack[4] have successfully removed or obfuscated embedded watermarks enough to fool blind detection. When robust watermarking systems become a reality, however, the techniques we describe here might contribute significantly in building an efficient content protection system for multimedia content distribution for a large-scale client base.

In addition to describing the system, we present performance analyses for the low-cost sensitivity attack[5] and an improved attack on the fingerprint detector with additive Gaussian noise. The two efforts are new with respect to our previously published work.[6]

Dual watermarking

Traditional spread-spectrum watermark systems, outlined in the "Spread-Spectrum Watermarking" sidebar, detect watermarks using a key w, which is in essence a *secret watermarking key* (SWK). In copyright enforcement schemes, watermark detection is done at the client (the media player), which must then have access to the SWK. Adversaries can recreate the original content if they succeed in obtaining the SWK from a single player. This can be achieved in several ways: by breaking into a detector—that is, reverse engineering the detection software or hardware—or using the sensitivity attack.[5]

In our dual watermark-fingerprint system, depicted in Figure 1, the *watermark detection key* (WDK) differs from the SWK, so breaking into a single detector does not provide enough information to remove the watermark w. The media signal x is watermarked the same way as in traditional spread-spectrum watermarking. However, for each media player i, an individualized watermark-fingerprint detection key WDK h_i is created from an SWK w in the following way. Let $C = \{c_{ij}\}$ denote an $m \times N$ matrix, where $c_{ij} \in \mathcal{R}$, $c_{ij} = \mathcal{N}(0, B^2)$, that is, each entry is a zero-mean Gaussian random variable with standard deviation $\sigma_c = B$. For any practical purpose of the dual watermark-fingerprint system, we assume $B \approx A$. 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 deterministically 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 correlation-based watermark 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 involves correlating the received media file $\hat{y}$ with h_i, which generates a detector output $d_W = \hat{y} \cdot h_i$. Similarly to traditional spread-spectrum watermarking, if $\hat{y}$ is marked, $d_W = 1 + g_W$; otherwise $d_W = 0 + g_W$. The difference is that now g_W is a function of both the media signal x and the watermark carrier c_i. If there are no attacks—that is, $\hat{y} = y = x + w$—then

$$d_W = y + h_i = (x + w)\ (w + c_i) = 1 + g_W, \text{ where}$$
$$g_W = s\ (w + c_i) + w \cdot c_i$$

is a zero-mean noise component of d_W. For this case, we derive the detection noise variance as $\sigma^2_{gW} \approx (A^2 + B^2 + A^2B^2)/N$.

In the remainder of the article, we assume equality in the expression for detection noise variance σ_{gW}. Because the detection noise variance is significantly increased because of the watermark carrier c_i (if $\hat{y}$ is watermarked, then $g_W = g_T + x \cdot c_i + w \cdot c_i$ or else $g_W = g_T + x \cdot c_i$, where $\mathrm{Var}[x \cdot c_i] >> \mathrm{Var}[g_T + w \cdot c_i]$), our watermark-fingerprint system requires larger N than traditional spread spectrum for the same watermark detector performance. We can improve detector performance by generating watermark carriers such that $(\forall c_i \in C) w \cdot c_i = 0$. Although such carriers do not pose any tangible effect on system security, they reduce the detection noise while screening watermarked clips.

Copyright enforcement

Our dual watermark-fingerprint system comprises a number of entities that combine to enforce copyright protection.

Watermark detector (WMD). The WMD correlates a potentially marked signal $\hat{y}$ with the client's WDK h_i—that is, $d_W = \hat{y} \cdot h_i$. 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 relatively small—for example, $\varepsilon_1 = 10^{-9}$.

Attacker. As part of an optimal attack to the system, the adversary breaks K clients and extracts their WDKs $\{h_i, i = 1 \ldots K\}$. Next, the adversary creates an attack vector v as an optimal estimate of the SWK w given the collusion key set $\{h_i, i = 1, \ldots K\}$, and generates an attacked signal as $\hat{y} = y - v$. The closer v estimates w, the more the attacker removes the watermark while generating $\hat{y}$. We use ε_2 to denote the probability that a watermark chip is incorrectly estimated by the attacker—that is, $\varepsilon_2 = \Pr[v_j \neq w_j]$. The attacker aims at forcing ε_2 to be as small as possible, whereas we design the system parameters such that ε_2 is near 1/2.

Spread-Spectrum Watermarking

In spread-spectrum watermarking, the media signal to be watermarked $x \in R^N$ can be modeled as a random variable, in which each element of x is Gaussain random variable with standard deviation A—that is, $x_j = \mathcal{N}(0, A^2)$. For audio signals, for example, A is typically within $A \in \{5, 15\}$, after necessary media preprocessing steps. A *watermark key* w is a spread-spectrum (SS) sequence vector $w \in \{\pm 1\}^N$, where each element w_j is usually called a *chip*. Vector addition—$y = x + w$—creates the marked signal y.

A small modification of this embedding rule, or *improved spread spectrum* (ISS),[1] leads to lower detection-error probability. To simplify the following analysis, we assume the use of traditional spread spectrum.

Let $w \cdot v$ denote the normalized inner product of vectors w and v—that is, $w \cdot v \equiv N^{-1}\sum w_j v_j$, with $w^2 \equiv w \cdot w$. For example, for w as defined previously, we have $w^2 = 1$. We assume the media player contains a watermark detector that receives a modified version $\hat{y}$ of the watermarked signal y. The watermark detector performs a correlation (or matched filter) test $d_T = \hat{y} \cdot w$, and, using a classical Neyman-Pearson hypothesis test, decides that the watermark is present if $d_T > \delta_T$, where δ_T is the detection threshold controlling the tradeoff between the false positive probabilities and false negative decisions. Modulation and detection theory have shown that such a detector is optimal.[2]

With no malicious attacks or other signal modifications (that is, $\hat{y} = y$), if the signal y is marked, $d_T = 1 + g_T$ where the *detection noise* g_T is a Gaussian zero-mean random variable with variance $\sigma^2_{g_T} = A^2/N$. Otherwise, the correlation test yields $d_T = 0 + g_T$. For equal probabilities of false positives and false negatives, we should set $\delta_T = 1/2$. For robustness against attacks, we must appropriately choose the signal domain x, and might have to make some small modifications on the watermark pattern. In this article, we assume that the designers of the watermark detector have considered these precautions, so we can disregard media attacks.

Finally, modeling the host signal x as a Gaussian random variable is a realistic assumption because most watermarking technologies replicate watermark chips to provide robustness to desynchronization attacks. From the watermark detector's perspective, chip replication is equivalent to averaging samples of the host signal. Consequently, averaged samples of the host signal, due to the central limit theorem, can be relatively accurately approximated using Gaussian distribution, regardless of the distribution of the original host signal samples.

References

1. H.S. Malvar and D. Florêncio, "Improved Spread Spectrum: A New Modulation Technique for Robust Watermarking," *IEEE Trans. Signal Processing*, vol. 51, no. 4, Apr. 2003, pp. 898-905.
2. H.L. van Trees, *Detection, Estimation, and Modulation Theory, Part I*, John Wiley & Sons, 1968.

Fingerprint detector (FPD). The 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—that is, it computes $d_F = v \cdot c_i$ and decides that the ith client is part of the collusion if $d_F > \delta_F$ (in other words, δ_F is the FPD threshold). The FPD has less noise in the correlated vectors than the WMD, and thus the collusion resistance of the FPD is much higher. We use ε_3 to denote the probability of false positives in the FPD—that is, incriminating a player that was not in the collusion set. Therefore, ε_3 must be very small, 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 insist that it is as small as ε_1 and ε_3.

Ultimately, the adversary's goal is to create an optimal attack vector $v \approx w$ based on a collection of K WDKs such that

- v reduces the expected $E[d_W]$ to a level at which the probability of detecting a true positive at the WMD is relatively low ($\approx \varepsilon_1$) and

- v reduces the expected $E[d_F]$ to a level at which the likelihood of a false negative at the FPD is relatively high (for example, $\eta \geq 0.9$).

Attacks without collusion

An adversary with knowledge of at most one WDK can perform several attacks on an object.

Attacks on protected objects

A basic assumption of our watermark-fingerprint mechanism is 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.[3] Typical attacks in this domain range from compression, filtering, resampling, equalization, and various other editing procedures,[7] to desynchronization (or data shifting) techniques that aim to misalign the embedded spread-spectrum sequence in the content (such as the Stirmark attack[2]).

Robustness to desynchronization attacks can be achieved using a certain amount of chip redundancy.[3] However, spread-spectrum chip replication also improves the efficacy of a watermark estimation attack.[3] In addition, the fact that both audio and video are highly repetitive phenomena has spurred a highly successful new generation of *swap attacks*, which replace relatively lengthy watermarked blocks of audio or video with perceptually similar blocks found elsewhere and hence, not marked with the corresponding watermark.[4] It is difficult to achieve protection against such attacks. The basic assumption of this article is that if such improvements to watermaking techniques are ever developed, the attacker would be forced to focus on breaking the player to retrieve the watermark detection keys.

Having a robust watermarking technology is not the only requirement for secure e-commerce transactions of multimedia. Traditional watermarking assumes that the SWK is hidden at the client side. By breaking a single client, the adversary can create the original content and thus allow all clients to play the content as unmarked. We refer to this as BORE: break once, run everywhere. In our system, we assume the attacker will eventually break at least one client and capture that machine's WDK. For the economic viability of the dual watermark-fingerprint system, the effort of breaking a client should be difficult to automate: a trustworthy operating system would bar patching, software debugging, and reverse engineering as simple ways to obtain machines' WDKs. (*Trustworthy* computing is usually defined as storage and computation that cannot be altered or intercepted by any system user without tampering with the computing hardware.) Similarly, watermark detectors should deter trivial implementations of the sensitivity attack as a way to retrieve the detection key, as we discuss later.

Our scheme is generally BORE-resistant at the protocol level. By breaking a single client, the adversary can play content as unmarked on that broken client, but must collude the extracted client WDKs with other clients to finally create content that can play on all players. With our dual watermark-fingerprint system, we significantly improve collusion resistance through a fingerprinting mechanism that can identify the members of the clique if the clique's cardinality is smaller than a relatively large lower bound.

Subtraction attack

Suppose an adversary breaks client i and extracts its WDK $h_i = c_i + w$. The adversary can then create an attack vector $v = \alpha h_i$ such that the modified media $\hat{y} = y - v$ produces $E[d_w] = E[\hat{y} \cdot h_i] \ll \delta w$, thus defeating that client's watermark detector. To determine α, we note that

$$\begin{aligned} d_W &= \hat{y} \cdot h_i \\ &= [x + w - \alpha(c_i + w)] \cdot (c_i + w) \\ &= 1 - \alpha(1 + c_i^2) + x \cdot c_i + x \cdot w + (1 - 2\alpha) c_i \cdot w \end{aligned}$$

Thus, by setting $\alpha = (1 + B^2)^{-1}$, we get $E[d_W] = 0$—that is, $d_W = 0 + g_W$. We also see that σ_{gW}^2 $(3+A^2+B^2+A^2B^2)/N$ and $\sigma_v^2 = \alpha^2(1+B^2) = \alpha \ll 1$.

Therefore, 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 $\sigma_{gW}{}^2$ and a negligible increase in content distortion (because $\sigma_v^2 \ll w^2 = 1$). If the attacker tries to use a key h_l to break a detector $i \neq l$, to drive $E[d_W] = 0$, the attacker would need to set $\alpha = 1$. However, this would drive $\sigma_v^2 = (1+B^2) \gg 1$, causing too much content distortion. It would also make σ_{gW}^2 increase by an amount equal to $3B^4/N$, which would make the decisions in the ith watermark detector erratic. In other words, even by driving $E[d_W] = 0$, the ith detector cannot be broken with probability much better than 1/2.

Resemblance to public-key systems

We have concluded that the attacker's knowledge of a single detector's WDK h_i 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 watermark/fingerprint system resembles a public-key cryptosystem, because knowledge of the verification key (h_i) does not imply knowledge of the signing key (w). However, unlike public-key cryptosystems, our system does not expose the WDK outside an individual player.

Collusion attacks

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

Optimal attack

The attacker's job is to estimate the SWK w by an attack vector v so that the modified media $\hat{y} = y - v$ will not show significant correlation in any watermark detector j, not even for $j > K$. The best job the attacker can perform is given by the sign(mean(·)) attack.

Lemma 1. The optimal attack is performed using the vector $v = sign(\sum_{i=1}^{K} h_i)$.

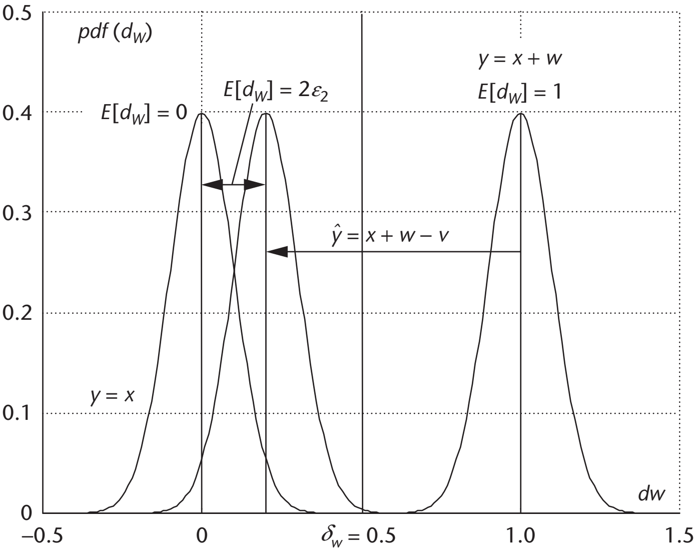


Figure 2. The probability density functions for correlation tests against three signals: a nonmarked $\hat{y}$=x, marked $\hat{y}$=x+w, and attacked using the v = sign(mean(·)) attack $\hat{y}$=x+w−v. All correlation test deviations approximate $\sigma_{gW} \approx AB/\sqrt{N}$. We set exemplary detection threshold to δ_W=1/2.

Proof. The optimal estimate for each element v_j of the attack vector is given by $v_j = +1$ if $\Pr[w_j = +1|\{h_i\}] \geq 1/2$, and $v_j = -1$ if $\Pr[w_j = +1|\{h_i\}] < 1/2$. This estimate is optimal because it minimizes $\Pr[v_j \neq w_j]$. Because $h_{ij} = w_j + c_{ij}$, where c_{ij} is independent and Gaussian, we can write $\Pr[w_j = +1|\{h_i\}] = 1/(1 + \nu_j)$, where $v_j = \prod_{i=1}^{K} p_c(h_{ij} - 1)$ and $p_c(\zeta) = (2\pi)^{-1/2} \exp[-\zeta^2/(2B^2)]$. We can write $v_j = \exp(-2\rho_j/B^2)$, where $\rho_j = \sum_{i=1}^{K} h_{ij}$. Thus, $\Pr[w_j = +1|\{h_i\} \geq 1/2]$ when $\rho_j \geq 0$, and $\Pr[w_j = +1|\{h_i] < 1/2]$ when $\rho_j < 0$. ■

WMD performance

Given the optimal attack, we can compute the average estimation error in the attack vector, $\varepsilon_2 = \Pr[v_j \neq w_j]$, as follows. Because the w_j chips are equally likely to be +1 or –1, due to the symmetry of w, $\varepsilon_2 = \Pr[s_j \geq 0 | w_j = -1]$. Because for $w_j = -1$, we have $s_j = -K + \overline{c}_j$, where $\overline{c}_j = \sum_{i=1}^{K} c_{ij}$. Therefore, $\varepsilon_2 = \Pr[\overline{c}_j \geq K]$, where $\overline{c}_j$ has a Gaussian distribution with zero mean and variance $B\sqrt{K}$.

Corollary 1. A collusion of size K produces

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

Given ε_2, we can evaluate the efficiency of the subtraction attack $\hat{y} = y - v$ for the optimal attack vector v. Because $E[v \cdot w] = \Pr[v_j = w_j] - \Pr[v_j \neq w_j] = 1 - 2\varepsilon_2$, after attack the expected outcome of the watermark correlation detector drops to $E[d_W] = 2\varepsilon_2$. Figure 2 depicts the resulting probability density functions (PDFs) for d_W when computed against an original, marked, and attacked signal. The attacker might attempt a stronger subtrac-

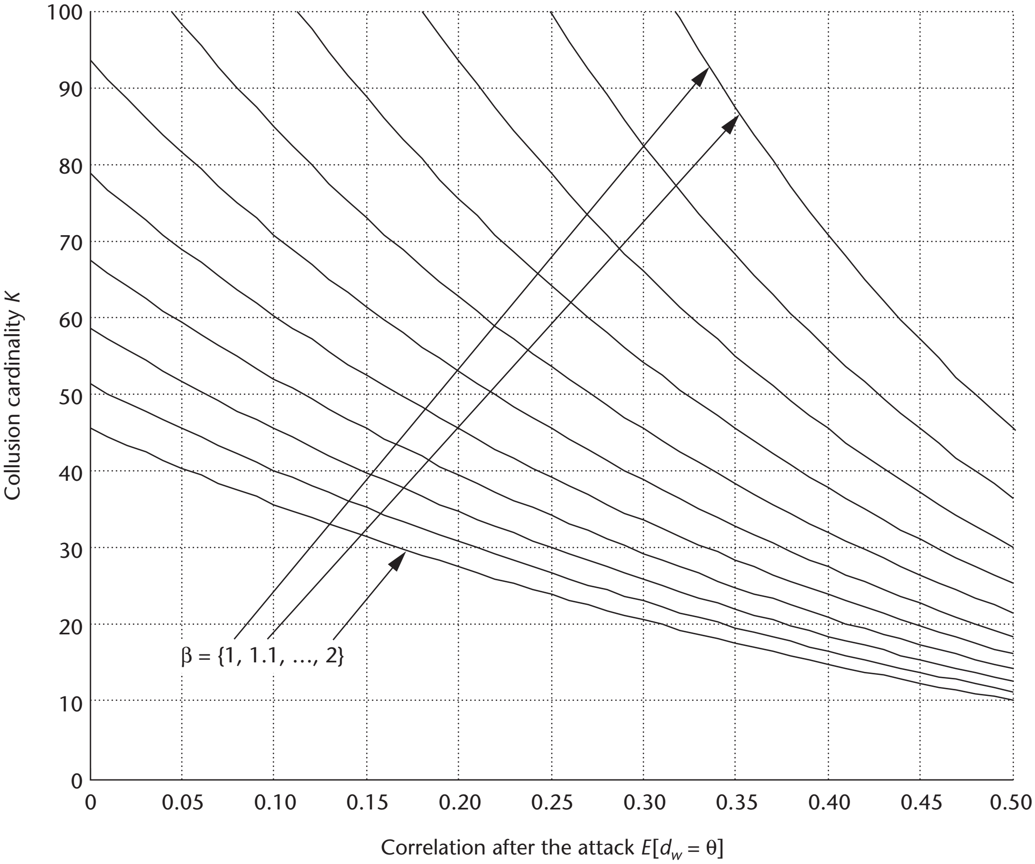


Figure 3. Corollary 2's claim. The figure depicts the collusion size* K *required to reduce the correlation value to $E[d_W] = \theta$.

tion attack, of the form $\hat{y}=y-\beta v$ with $\beta > 1$, to bring the WMD output down further, to $E[d_W] = 2\beta\varepsilon_2 - (\beta - 1)$. As long as β is not too large, the attacked content $\hat{y}$ might be acceptable to users.

Defeating the WMD

To reduce the expected correlation to $E[d_W] = \theta$, the adversary must achieve an attack vector error rate of $\varepsilon_2 = (\theta + \beta - 1)/(2\beta)$ through collusion. From Corollary 1, we see that for fixed θ and β the minimum collusion size grows proportional to B^2.

Corollary 2. To reduce the correlation value to $E[d_w] = \theta$, the adversary must collude K WDKs, with

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

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. Figure 3 illustrates the dependency of K with respect to θ and β for $B = 10$.

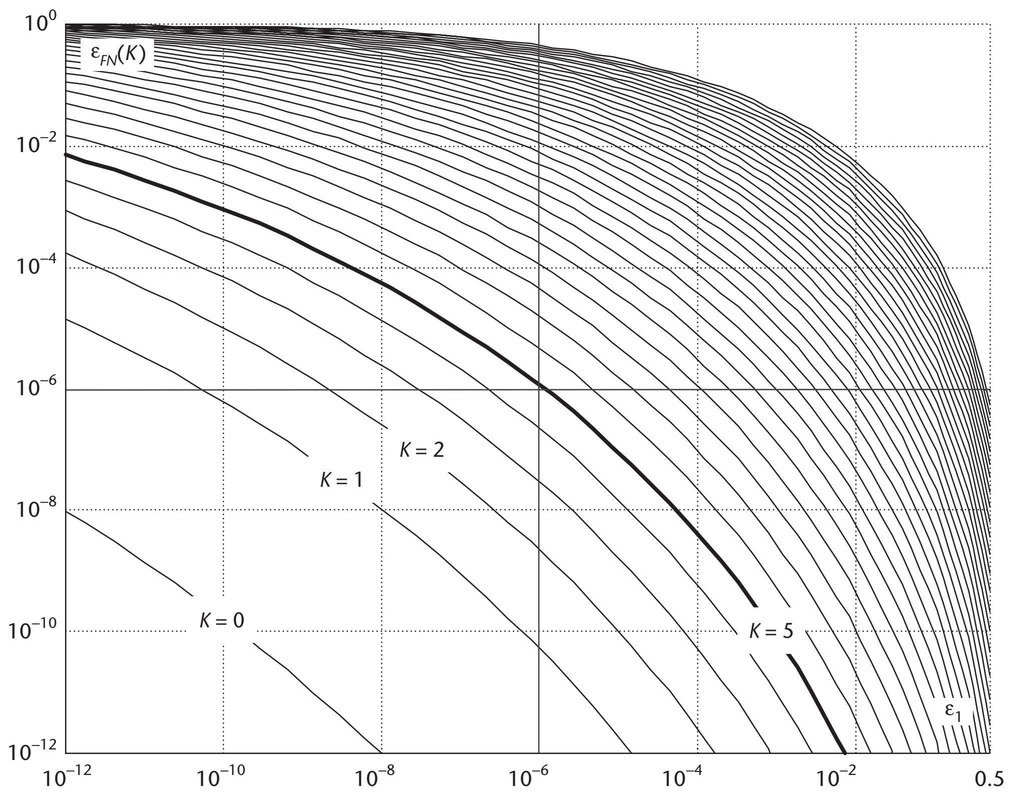
The attacker must set θ much smaller than δ_W, or the probability that a WMD will still detect the watermark will not be low enough to justify the attacker's effort. In other words, the attack is successful only if it makes $\varepsilon_1 \approx 1$. It is not necessary to set θ all the way to zero because doing so would require an excessively large K. By setting $\beta > 1$, however, we can force $\theta = 0$.

To make the attacker's job more difficult, we increase parameter B, the standard deviation of the watermark carrier c, because K grows with B^2. In doing so, however, we increase the detection noise variance $\sigma^2_{gW} = (A^2 + B^2 + A^2B^2)/N$, where A is the standard deviation of the original content x and N is the object size. For a given σ_{gW}, we determine that the probability of false positives $\varepsilon_1 = \Pr[d_w > \delta_w | \text{object is not marked}]$ is given by Corollary 3.

Corollary 3. An object of size N produces

$$\varepsilon_1 = \frac{1}{2}\operatorname{erfc}\left(\frac{\delta_W\sqrt{N}}{\sqrt{2(A^2+B^2+A^2B^2)}}\right)$$

If $\delta_W = 1/2$, e_1 is also the probability of false negatives—that is, the probability that a WMD will not detect a marked object that was not attacked.



$$\varepsilon_1 = \frac{1}{2}\,erfc\left(\frac{\delta_W\sqrt{N}}{\sqrt{2(A^2+B^2+A^2B^2)}}\right) \qquad \varepsilon_{FN} = \frac{1}{2}\,erfc\left(\frac{2(\varepsilon_2-\delta_W)\sqrt{N}}{\sqrt{2(A^2+B^2+A^2B^2)}}\right)$$

Figure 4. WMD receiver operating characteristic graph. The graph considers the false positive probability ε_1 and false negative probability ε_{FN}(K) after an attack with v averaged from K WDKs. In the example, A = B = 7 and N = 4 · 10^5. A possible system design decision is $\varepsilon_1 = \varepsilon_{FN} \leq 10^{-6}$ with K = 5 collusion resistance and $\delta_W = \varepsilon_2$(K = 5). Keeping δ_W = 1/2 reduces the likelihood that the attacked chip will not be detected as marked drops to ε_{FN}(K = 5,δ_W = 1/2) ≈ 10^{-3} at fixed ε_1(K = 5,δ_W = 1/2) ≈ 10^{-10}.

Figure 4 illustrates in more detail the receiver operating characteristic graph of the WMD under the assumption that the marked signal has been attacked by averaging *K* WDKs. From Corollary 1 we compute *N* using Corollary 4.

Corollary 4. The object size *N* required to achieve a given ε_1 is

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

By combining this result with the result in Corollary 2, we arrive at one of the main results in this article.

Theorem 1. A minimal collusion size K_W that achieves a fixed $E[d_W] = \theta$ at a WMD with fixed δ_W, β, and ε_1 grows linearly with object size *N*—that is, $K_W = \mathcal{O}(N)$.

Proof. As *N* grows, for a given ε_1 *B* also grows, and thus $\sigma_{gW}^2 \rightarrow B^2(1+A^2)/N$. Combining this asymptotic expression for σ_{gW} with the results in Corollaries 2 and 4, we get

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

This equation lets us compute the object size *N* necessary to achieve any desired collusion resistance K_W for a given WMD performance. ■

The result is so far determined only by the WMD performance. Next, we confirm the linear relationship between *K* and *N* when considering the FPD performance.

Important suboptimal attack

The computing environment also impacts the security of dual watermark-fingerprint systems. One of the initial assumptions was that both player WDKs and the WMD are trustworthy. For media players implemented entirely in hardware, trustworthiness is achieved with strong tamper-proof design.[5] For software media players, the computing platform must maintain storage and computing processes that system users cannot access using arbitrary software. Hence, because in this case WDKs cannot be extracted using programs, the adversary must tamper with the computing platform. In both cases, reverse engineering a tamper-proof chip is an effort whose cost can easily be kept above the US$100 per player mark.[8,9] Assuming the system collusion resistance analyzed in the "Fingerprint detection" section, such high attack cost per client translates into several tens of millions of dollars as the total cost of breaking the dual watermark-fingerprint system.

We therefore consider alternative attacks that might break the dual watermark-fingerprint system at a lower cost. The premiere candidate is the sensitivity attack, which treats the WMD as a black box and applies series of tests on the WMD to determine the detection key. In a trivial implementation, the adversary first mixes the marked signal with enough noise to bring the detector's correlation with the hidden WDK near the threshold δ_W. The adversary then aims to estimate the value of chip w_j of the watermark by changing the corresponding sample of the marked signal y_j and analyzing the decision of the WMD. The complexity of the attack is in the worst case linear with respect to the length of the detection key. For traditional spread spectrum, the sensitivity attack identifies the actual watermark bit w_j, whereas for the dual watermark-fingerprint system, it identifies only the sign of the WDK hidden in the player: $\text{sign}(h_{ij}) = \text{sign}(c_{ij} + w_j)$. Although such an attack cannot be prevented, using randomized thresholding can make the attack's duration, and ultimately its cost, significant. Because the cost of the sensitivity attack is still estimated to be well below US$100 per player, it is important to analyze its effect on our system collusion resistance.

Lemma 2. Given a set of K vectors $\{u_i = \text{sign}(h_i) = \text{sign}(c_i + w), i = 1 \ldots K\}$, an optimal estimate v of the watermark w is computed as

$$v = \text{sign}\left(\sum_{i=1}^{K} u_i\right)$$

Proof. From Lemma 1 for $K = 1$, it follows that the optimal estimate of w within a single h_i equals $u_i = \text{sign}(h_i)$. The proof of Lemma 1 shows that the optimality of the attack is based on the sign of

$$\rho_j = \sum_{i=1}^{K} h_{ij}$$

Given a set of estimates u_{ij} for each h_{ij}, the best estimate of the $\text{sign}(\rho_j)$ from Lemma 1 is given by

$$\text{sign}\left(\sum_{i=1}^{K} u_{ij}\right)$$

Theorem 2. Let K_o denote the size of a collusion of WDKs: $\{h_i, i = 1 \ldots K_o\}$. Let K_s denote the size of a collusion of WDKs extracted using the sensitivity attack: $\{u_i = \text{sign}(h_i), i = 1 \ldots K_s\}$. The two attacks will have the same efficacy—that is, they will have equivalent probability of error

$$\Pr[\text{sign}(\sum_{i=1}^{K} u_{ij}) \neq w_j] = \Pr[\text{sign}(\sum_{i=1}^{K_O} h_{ij}) \neq w_j]$$

for

$$\frac{K_S}{K_O} = \frac{4\psi(1-\psi)}{(2\psi-1)^2 B^2} = \text{constant}$$

where $\psi = 1 - \text{erfc}(1/B\sqrt{2})/2$.

Proof. From Corollary 1, for $K = 1$ we conclude that the $\Pr[u_{ij} = w_j] = \Pr[|c_{ij}| \leq 1] + \Pr[|c_{ij}| > 1]/2 = \psi$. Hence,

$$E[\sum_{i=1}^{K_s} u_{ij}] = (2\psi - 1)K_s w_j$$

with

$$\text{Var}[\sum_{i=1}^{K+s} u_{ij}] = 4K_s\psi(1-\psi)$$

The probability of an estimation error equals

$$\begin{aligned} \varepsilon_2' &= \Pr[\text{sign}(\sum_{i=1}^{K_s} u_{ij}) \neq w_j] \\ &= \Pr[\sum_{i=1}^{K+s} u_{ij} < 0 \mid w_j = +1] \\ &= \text{erfc}[(2\psi-1)\sqrt{K} / \sqrt{8\psi(1-\psi)}]/2 \end{aligned}$$

Assuming ε_2' equals ε_2 from Corollary 1, we derive the equation in Theorem 2. ■

We can show that $K_s > K_o$ for all positive real values of B, confirming that the

$$\text{sign}\left(\sum_{i=1}^{K} u_i\right)$$

attack is inferior to

$$\text{sign}\left(\sum_{i=1}^{K} h_i\right)$$

Interestingly, within the region of interest for multimedia data ($B \in \{5, 10\}$), the ratio K_s/K_o is approximately constant, with $K_s/K_o \approx 1.56$. Therefore, for brevity and clarity, the remainder of this article analyzes the case in which an attacker obtains the original WDK h_i upon breaking player *i*. If the attacker uses the more realistic sensitivity attack to obtain WDK information, we assume with high accuracy that the attacker would need to collude about 56 percent more WDKs than in the optimal attack to achieve the same goal.

Fingerprint detection

As we mentioned previously, the FPD has less noise in its correlation output. Therefore, it should be able to identify the indices *i* corresponding to all 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 in the previous section.

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 *i*th client participated in the collusion if $d_F > \delta_F$. We assume a realistic modification to the sign(mean(·)) attack model from the previous section, $\hat{y} = y - \beta v + n$, where *n* is a noise the attacker adds to the attack vector $-\beta v$. This noise aims to increase the correlation variance at the FPD and thus reduce its performance. We can model the attack noise *n* as a zero-mean independent identically distributed Gaussian random variable $n_j = \mathcal{N}(0, \sigma_n^2)$ with variance σ_n^2. To preserve the fidelity of the original media clip, the adversary can add only noise of limited variance, usually proportional to the variance of the watermark (for example, $\sigma_n \approx 1$). Now, we can write the FPD output as:

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

where g_F is the zero-mean FPD correlation noise. The most critical error for the FPD is a false positive—that is, incriminating a WDK *i* that did not participate in the collusion. The probability ε_3 of that error is given in Lemma 3.

Lemma 3. An object of size *N* produces

$$\varepsilon_3 = \frac{1}{2} \mathrm{erfc}\left(\frac{\delta_\mathrm{F} \sqrt{N}}{B\sqrt{2(\beta^2 + \sigma_n^2)}} \right)$$

Proof. If c_i is not in the collusion, it is independent of the attack vector $-\beta v + n$. Thus,

$$\begin{aligned}\sigma_{gF}^2 &= \mathrm{E}[(-\beta v_{ij} + n_{oj})^2 c_{ij}^2]/N \\ &= \mathrm{E}[(\beta^2 + n_{ij}^2)c_{ij}^2]/N = (\beta^2 + \sigma_n^2)\beta^2 / N\end{aligned}$$

which follows from $E[v_{ij}n_{ij}] = 0]$, $\varepsilon_3 = \Pr[g_F > \delta_F]$ and the fact that g_F has Gaussian distribution. ■

As expected, $\varepsilon_3 \ll \varepsilon_1$ (usually by several orders of magnitude), because the argument in the complementary error function (erfc(·)) for ε_3 is approximately $(A\delta_F)/(\delta_W\sqrt{\beta^2+\sigma_n^2})$ times larger than the argument in erfc(·) 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 must determine its expected output when we correlate the extracted attack vector $-\beta v + n$ with a carrier c_i to check whether h_i was part of the collusion. The expected output $E[d_F]$ does not depend on the attack noise *n,* assuming $E[n_{ij}v_{ij}] = E[n_{ij}c_{ij}] = 0$. We see that $E[d_F] = \beta E[z_j]$, where $z_j = v_j c_{ij} = \mathrm{sign}[s_j]c_{ij}$, with $s_j = w_j + b_j$, and

$$b_j = \frac{1}{K} \sum\nolimits_{m=1}^{K} c_{mj}$$

Lemma 4. 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. Clearly $E[z_j] = (E[z_j|w = +1] + E[z_j|w = -1])/2$ because the w_j chips are equally likely. Also, because of the symmetry of the problem, $E[z_j|w = +1] = E[z_j|w = -1]$ and so $E[z_j] = E[z_j|w = +1]$. ■

Assuming $w_j = +1$, $E[z_j] = E[z_j|s_j \geq 0]\Pr[s_j \geq 0] + E[z_j|s_j < 0]\Pr[s_j < 0] = E[c_{ij}|s_j \geq 0]\Pr[s_j \geq 0] - E[c_{ij}|s_j < 0]\Pr[s_j < 0]$. Under each of the conditions $s_j \geq 0$ or $s_j < 0$, $s_j = 1 + b_j$ and c_{ij} are jointly Gaussian 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, because c_{ij} is part of the average that defines b_j. Thus, computing these conditional expectations is just an exercise in computing expectations of a Gaussian random variable, conditioned on minimum or maximum values for that variable.

Given the expected FPD output, to concurrently minimize the likelihood of false negatives η (that is, the probability that a key index *i* in the collusion will not be detected) and false positives ε_3, we could set $\delta_F = E[d_F]/2$. However, the FPD does not know *K* at detection time. Because the probability of a false positive does not depend on *K*, we set ε_3 to a constant value $\varepsilon_3 = \tau$ (typically $\tau \leq 10^{-12}$), which determines the detection-bound δ_F.

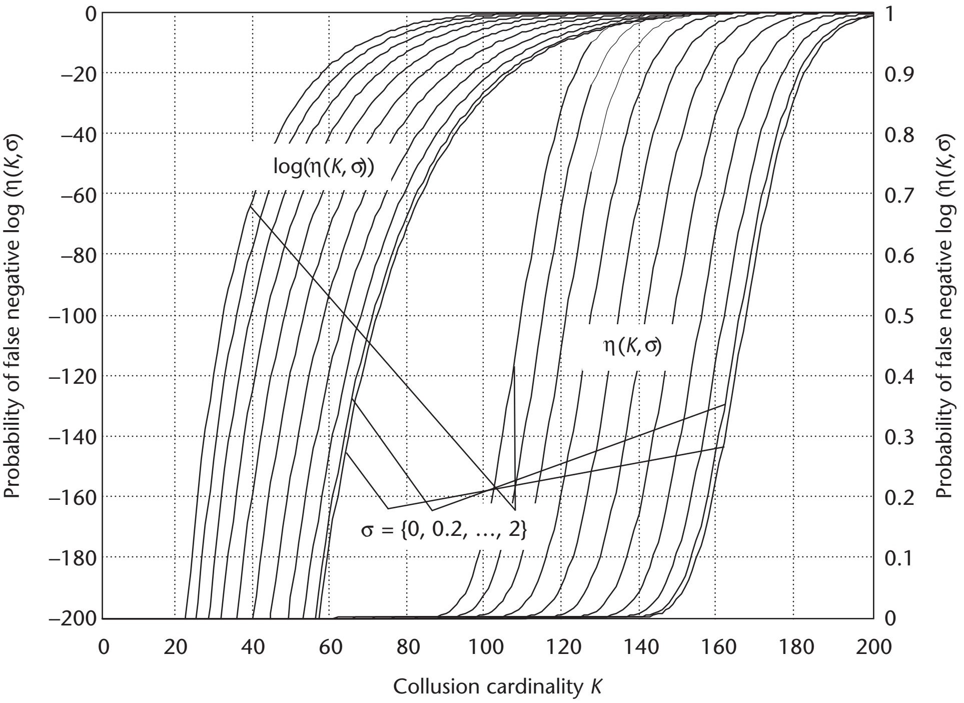


Figure 5. Diagram of η(K,σ) for B = 7, τ = 10^{-12}, β = 1, N = 4 · 10^5. In a realistic scenario where σ = 1, a collection of K ≥ 157 WDKs imposes likelihood of detection $\eta \ge \tau_C$ = 0.9.

Corollary 5. To achieve $\varepsilon_3 = \tau$, FPD must set

$$\delta_F \ge B\sqrt{\frac{2\left(B^2+\sigma_n^2\right)}{N}}\operatorname{erfc}^{-1}(2\tau)$$

The detection threshold uniquely determines the probability of a false negative η. Because the FPD output d_F is Gaussian with expected value $E[d_F]$ and variance $\sigma^2_{d_F} = \sigma^2_{g_F} = (\beta^2 + \sigma_n^2)B^2/N$, we deduce Corollary 6.

Corollary 6. An object of size N produces

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

$$\eta = \frac{1}{2}\operatorname{erfc}\left(\sqrt{\frac{N}{\pi K(1+\frac{\sigma_n^2}{\beta^2}}}\exp\left(-\frac{K}{2B^2}\right)-\operatorname{erfc}^{-1}(2\tau)\right)$$

The collusion's foremost goal is to avoid detection at the FPD. From the second equation, we can compute the minimal size of a collusion clique K_F that would have the probability of individual clique member detection $\eta(K_F)$ above a desired threshold $\eta \ge \tau_C$, where typically $\tau_C \ge 0.9$. Figure 5 shows how $\eta(K, \sigma)$ changes with the increase of K and σ, the only parameters the attacker can change assuming β = 1. For the example system depicted in Figures 4 and 5, cliques of $K \ge K_F = 157$ WDKs at σ = 1 would make only one out of 10 colluders identifiable by our FPD as $\eta(K_F) \ge 0.9$.

We compare this result to the K_W computed in Corollary 2 and Theorem 1. Reducing the expected correlation of the WMD to $E[d_W] = \theta$ requires $K_W = 2B^2\{\operatorname{erf}^{-1}[(1-\theta)\beta^{-1}]^2$ WDKs, where erf(·) is the standard error function. Thus, for the example in Figure 4, the collusion size that drops $E[d_W] = 0.2$ is only $K_W \ge 80$ WDKs (at β = 1). The discrepancy between K_W and K_F is significantly larger with increased β; however, it decreases as stronger attack noise n is superimposed. Hence, the goal of the adversary is to use β = 1 while adding as much noise as possible. We estimate that an adversary can add noise with up to four times stronger variance than the original watermark energy (that is, σ ≤ 2). In this case, for the example in Figure 3, the collusion resistance $K_F|\eta(K_F, \sigma = 2) \ge 0.9$ drops from 157 to 123 WDKs with all other system parameters intact.

Although the adversary can create a signal that can play as unmarked on almost all existing players with colluded K_W WDKs, it would be foolish to expect such a collusion as each colluder could be easily identified. Thus, we assume that the ultimate goal of the collusion clique is to average K_F WDKs such that each WDK is virtually undetectable at FPD time.

Using Corollary 6, we can compute the object size N necessary to achieve a desired probability η of false negatives in the FPD. As noted previously, the minimal collusion size to drive down $E[d_W] = \theta$ can be computed as $K_W = 2B^2\mu^2$, where $\mu = [\mathrm{erf}^{-1}[(1 - \theta)\beta^{-1}]$ is fixed for a fixed attack efficiency (that is, a fixed θ and fixed β). Therefore, as we increase B, the attacker must increase K_F proportionally to B^2, which imposes Corollary 7.

Corollary 7. The object size N required to achieve a given η for fixed τ and μ is:

$$N = K_F \pi (1 + \frac{\sigma_n^2}{\beta^2})$$
$$\left[(\mathrm{erfc}^{-1}(2\eta) + \mathrm{erfc}^{-1}(2\tau)) e^{\mu^2} \right]^2$$

This result confirms Theorem 1: that collusion size and object size are linearly related. 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.

Finally, during the forensic search, the FPD must perform correlation tests individually for each user's WDK h_i to determine whether it is contained in the pirated copy $\hat{y}$. The length of a single watermark $|w| > 0.5 \cdot 10^6$ and the cardinality of the user space $|C| > 10^9$ can make this a time-consuming search. However, the search can be parallelized and effectively distributed over a network of computers, which can significantly reduce the search time. For a large user space, we estimate that the FPD search can be performed in a week using the idle cycles of a typical enterprise network of 1,000+ computers.

Segmentation

In our dual watermark-fingerprint system, watermarks protect the content, and fingerprints let the copyright owner identify a clique of users that launched an attack to remove the watermark. This unique property lets us add multiple watermarks in the object and force the adversary to create cliques independently for each watermark. More formally, we divide the protected object into S segments and watermark each of them with a distinct spread-spectrum sequence. For each segment i, we publish m distinct WDKs $h_{i,j}$, $j = 1 \ldots m$, created in accordance with the described dual watermark-fingerprint system. Each client gets a single WDK $h_{i,j}$ to exactly one segment.

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. For this reason, we conveniently conclude that a segment can resist K colluders without mentioning error probabilities. A protected object is defeated if watermarks are removed from all segments but no fingerprints are introduced in the process. To break the system, the adversary must collect at least K WDKs for each segment. When published, the WDKs are uniformly assigned to random segments. We assume the total number of published WDKs mS significantly surpasses $mS >> KS$. Thus, the adversary's effort to collect a set of WDKs that would break the watermark-fingerprint system can be modeled as *coupon collecting*.[10,11]

Definition 1. The collusion resistance M of a segmented dual watermark-fingerprint system with S segments equals the expected number of WDKs that must be selected from an infinite pool of WDKs, such that each segment has a collusion clique of at least K WDKs.

Let X be a random variable denoting the collusion size in a given segment when we have S segments and overall M broken clients (that is, extracted WDKs). X has a Poisson distribution with mean $\mu = M/S$. Let $p = \Pr[X \leq K]$. Alon et al.[12] show that $\Pr[X \leq \mu(1 - \gamma)] \leq \exp(-\gamma^2\mu/2)$. In our case, $K = \mu(1 - \gamma)$, so $\gamma = 1 - K/\mu$ and $p \leq \exp[-(1 - K/\mu)^2\mu/2]$. Let $q = \Pr$[all segments contain more than K keys]. Then, assuming independence among segments, $q = (1 - p)^S$, which for a small p becomes $q = 1 - pS$. If we try for $q = 1 - \varepsilon_s$, then $\varepsilon_s = pS$. So, $\varepsilon_s/S \leq \exp[-(1 - \mathrm{K}/\mu)^2\mu/2]$. Adding $\mu = M/S$ gives us

$$1\mathrm{n}(S/\varepsilon_S) \geq \frac{M}{2S} - K + \frac{SK^2}{2M}$$

Solving for M, we get Lemma 5.

Lemma 5. If

$$M = S\left(\ln\frac{S}{\varepsilon_S} + K + \sqrt{\ln^2\frac{S}{\varepsilon_S} + 2K\ln\frac{S}{\varepsilon_S}} \right)$$

then $q > 1 - \varepsilon_s$.

Theorem 3. A dual watermark-fingerprint system with segmentation has superlinear collusion resistance.

Proof. For a fixed collusion resistance per segment K, the number of segments S is linear in object size $S = \mathcal{O}(N)$. Therefore, using Lemma 5, we get overall superlinear collusion resistance with respect to object size $M = \mathcal{O}(N \log N)$. ■

In an alternate direction, when the asymptotic case of the coupon collecting problem is analyzed for $S \to \infty$ and minimal K collected WDKs

Table 1. Dependencies among main parameters of the watermark-fingerprint system.

Parameter	Parameter dependencies
$\varepsilon_1 = \Pr[d_w > \delta_w \mid \text{object not marked}]$	$\sim\text{erfc}(\sqrt{N}/AB)$
Segment length, N	$\sim B^2A^2\,[\text{erf}^{-1}(1 - 2\varepsilon_1)]^2$
$\varepsilon_2 = \Pr[v_j \neq w_j]$	$\sim\text{erfc}(\sqrt{K}/B)$
Collusion resistance per segment, K	$\mathcal{O}(N)$
$\varepsilon_3 = \Pr[d_F(c_i) > \delta_F \mid c_i \notin K]$	$\sim\text{erfc}(\sqrt{N}/B)$
System collusion resistance, M	$\mathcal{O}(N(\log(N)))$

per segment, collusion resistance M has a well-known sharp threshold at

$$\lim_{S\to\infty} \Pr\left[M_K > S\left(\ln S + (K-1)\ln\ln S + c\right)\right] = e^{-e^{-c}}$$

for any real number $c \in \mathbf{R}$. This points to the fact that the number of WDKs, M, that the adversary must collect to cover at least K keys in S segments should be centered at $M = S \ln S + (K-1)S \ln\ln S$ with exceptionally small variance at both tails.

Because the solution's variance to the coupon collecting problem is exceptionally small, we expected that for a large number of segments within an object, two distinct attacks to the system would require a similar number of collected WDKs (within S keys) with exceptionally high probability. Thus, although the collection of WDKs during the attack is probabilistic, we can assume with great certainty that the collusion resistance's resulting superlinearity is almost deterministic because of this sharp transition.

Segmentation is impossible in classic fingerprinting systems because they require some form of marking assumption.[1,13]

Key compression

The major drawback of the dual system is its need for a relatively large space in which to store the detection keys. It is difficult to compress the sum of two independent pseudorandom 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 an open problem.

The current situation is that we must 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 approximately 10^5 bytes, which might be too much data for some embedded devices.

Recall that the WDK of user i is created as $h_i = c_i + w$, where c_i and w are mutually independent. Alternatively, we can generate the key from a short seed using any standard cryptographically secure pseudorandom key generator, perform sieving for each chosen w, and select only those seeds whose resulting long sequence (denoted as s) has the property $s \cdot w \geq 1$, thus inferring $h_i = s$. The deviation of $s \cdot w$ is roughly $\sigma^* = B\sqrt{N_O}$, so the probability of a randomly chosen seed meeting these criteria is $\varepsilon^* = 1/2\text{erfc}(N_O/B\sqrt{2})$. For example, for $\varepsilon^* < 10^{-6}$, we get $N_o = 2B^2[\text{erfc}^{-1}(2\varepsilon^*)]^2 = 2000$. Because $N = 10^5$, 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 $N_o/\xi \sim 20$.

Parameter interaction and implications

The dual watermark-fingerprint technology aims to build practical secure content-protection mechanisms. Table 1 presents an overview of the interrelationships among the parameters of our watermark-fingerprint scheme. For example, for a given object size N and variance A^2, the other variables behave as shown in the table.

The designer's primary task is to determine the number of segments S per object. Because collusion resistance within a single segment is $K = \mathcal{O}(N)$, where $N = N_o/S$ is the segment's length, and collusion resistance achieved over S segments is $M = \mathcal{O}(S \ln(S))$, the objective is segments that are as short as possible to

- maximize overall collusion resistance M and
- reduce the storage space for a single WDK.

On the other hand, security measures for hiding w within a watermark carrier c_i make necessary a lower bound on the watermark carrier amplitude B, commonly set to $B \geq A$. Selecting B uniquely identifies the segment length N with respect to a desired probability of a false alarm ε_1 under the optimal sign(mean(·)) 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. Finally, η and N imply the collusion size K (computed from Corollary 6) required to make all colluders invisible at FPD time

Related Work on Content-Screening Technologies

Three main technologies for content protection exist: watermarking, fingerprinting, and traitor tracing (a fourth worth mentioning is digital-rights management, or DRM). Table A briefly compares these technologies to our dual watermark-fingerprint system.

Public-key watermarking

Public-key watermark systems have focused mainly on providing a solution to the *prisoners' problem.*[1] This problem requires two trusting parties (prisoners) to establish a covert communication channel in the presence of a warden. Simmons suggested encrypting the embedded message with the recipient's public key before watermark embedding so a warden could not understand it.[2] Craver extended this protocol to include an active warden.[3] Neither protocol fits the requirements of a content-screening system, which aims to achieve a much harder task—namely, to protect data such that if a server sends one bit to a set of clients, even if an adversary fully controls one client, it cannot interfere with the server's communication with other clients.

Fingerprinting

Ergun et al. consider embedding distinct spread sequences per copy and are among the first to formalize attack metrics (the limits beyond which a copy is considered too corrupt to be useful).[4] They considers one attack: averaging fingerprinted copies with additional noise. This attack is weaker than those Boneh and Shaw consider,[5] and accordingly they show a much higher upper bound on collusion size that can be overcome.[4] Boneh and Shaw construct fingerprint codes, which in the worst case produce collusion resistance of about $K = \mathcal{O}(N^{1/4})$. Pfitzmann and Waidner introduced a fingerprint scheme in which users can buy digital content anonymously but can be identified if they redistribute the fingerprinted content.[6]

Fiat and Tassa's copyright protection approach[7] is less realistic than Boneh and Shaw's[5] because the former assumes that pirates simply choose one of the symbols available to them in each round of the tracing process. Boneh and Shaw assume that they can assign any value to bits about which a collusion disagrees. Symbols are composed of many bits. Thus, the collusion can create symbols not in the original alphabet.

Traitor tracing

Traitor tracing and fingerprint copyright protection systems have important differences. (In particular, one cannot blindly export error correction ideas from traitor tracing to classic digital fingerprint without some form of the Boneh-Shaw marking assumption.[5]) Although they share the same goal, the scenario and means are different.

In traitor tracing, the content is usually broadcast in real time and has little value afterwards. Pirates are assumed to be unable to manipulate and rebroadcast content in (near) real time. The content is encrypted, and legitimate clients have distinct sets of keys that enable decryption when combined. Each legitimate set of keys is uniquely associated with a single client. If a pirate resells his or her keys, law enforcement confiscates the suspect client's box, and uses the set of keys in the confiscated box to

continued on p. 72

For realistic loads to the system, such as high-definition television, the number of bits per object is in the order of 10^{11} bytes. Assuming one chip embedded per 100 pixels, we derive an object size of $NS \approx 10^9$ chips. Alternatively, from $B = A \approx 7$ and $\varepsilon_1 = \varepsilon_{FN} \approx 10^{-10}$, we derive $N \approx 4 \cdot 10^5$ chips. This boosts the number of segments to $S \approx 2.5 \cdot 10^3$. The adversarial collusion clique uses $(\sigma/\beta)^2 = 4$ to create the pirated content by subtracting the optimal watermark estimate amplified by $\beta = 1$ and adding the attack noise as a $n{=}N(0, \sigma_n^2{=}4)$. For a fixed false negative rate of $\varepsilon_3 = 10^{-12}$, the false positive rate follows the diagram in Figure 5, thus yielding a per-segment collusion resistance of $K \geq 123$ for $\eta \geq 0.9$. Most importantly, the achieved overall collusion resistance is lower-bounded by $M > 3 \cdot 10^5$ users. One can hardly expect that, under realistic piracy scenarios, such a clique could be established to oppose the protection of the proposed dual watermark-fingerprint system.

One disadvantage of the dual watermark-fingerprint system is content collusion, in which an adversary uses L media clips marked with an identical watermark w to estimate w using the optimal collusion attack:

$$v = \sum_{i=1}^{L}(x_i + w) + \sum_{j=1}^{K}(h_j + w)$$

For $B \ll A$, this attack can be particularly effective for relatively small L. However, for practical reasons of limited watermark length in the dual watermark-fingerprint system, we use $B \approx A$. To reduce sensitivity to this type of an attack, the set $\{w, C\}$ must be renewed after several tens of movies. The WDKs are distributed to users' players using standard cryptographic tools for authenticated communication.

Conclusion

Our dual watermark-fingerprint system limits the scope of possible attacks, when compared to

continued from p. 71

trace the leak to its source. However, a large enough collusion can create a good set of keys that does not incriminate any of the culprits.

Table A. Comparison of main characteristics of content-screening technologies: traitor tracing, fingerprinting, and the dual watermark-fingerprint system.

Characteristic	Traitor tracing	Fingerprinting	Dual watermarking and fingerprinting
Primary target application	Detection of pirated players	Copyright enforcement	Copyright enforcement
Content replication	Decrypt and capture content	Users collude their content copies	Users collude their keys to remove the protection mark
Content distribution	Real-time broadcast; single encrypted copy of content distributed	Each user receives a unique copy	Single watermarked copy; each user has a distinct detection key
Collusion resistance	Low (hundreds)	Low (tens)	High (millions)
Trace-back mechanism	Player confiscation; player response to a probe with "invalid ciphertext" reveals colluders	Analysis of pirated content; fingerprint detector can compare the pirated content to the original copy and the individual marks	
Advantages	Protocols can be based on provably hard problems	No action required at client side; players remain unchanged	High collusion resistance; copyright is enforced through prevention
Disadvantages	Difficult to enforce player confiscation; low collusion resistance	Exceptionally low collusion resistance; detects fraud, but does not prevent it	Marking key must be replaced after marking 100+ media clips

classic fingerprinting systems. (See the "Related Work on Content-Screening Technologies" sidebar on p. 71.) Under optimal attacks, the size of the collusion necessary to remove the marks without leaving a detectable fingerprint is asymptotically $K = \mathcal{O}(N)$ without segmentation, and $M = \mathcal{O}(N \log(N))$ with segmentation, where N denotes object size. Classic fingerprinting has a lower bound on collusion resistance—roughly $\mathcal{O}(N^{1/4})$. Thus, for example, the dual watermark-fingerprint system can achieve content protection with collusion resistance of up to 300,000 users for a two-hour high-definition video. **MM**

Acknowledgments

We thank M. Kesal, M. Kivanç Mihçak, and R. Venkatesan for providing an analysis of the media collusion attack.

References

1. D. Boneh and J. Shaw, "Collusion Secure Fingerprinting for Digital Data," *IEEE Trans. Information Theory*, vol. 44, no. 5, Sept. 1998, pp. 1897-1905.
2. F.A.P. Petitcolas, R.J. Anderson, and M.G. Kuhn, "Attacks on Copyright Marking Systems," *Proc. Information Hiding Workshop*, LNCS 1525, Springer-Verlag, 1998, pp. 218-238.
3. D. Kirovski and H.S. Malvar, "Spread-Spectrum Watermarking of Audio Signals," *IEEE Trans. Signal Processing*, vol. 51, no. 5, Apr. 2003, pp. 1020-1033.
4. D. Kirovski and F.A.P. Petitcolas, "Replacement Attack on Arbitrary Watermarking Systems," *Proc. ACM Workshop Digital Rights Management*, ACM Press, 2002, pp. 177-189.
5. J.P. Linnartz and M. van Dijk, "Analysis of the Sensitivity Attack Against Electronic Watermarks in Images," *Proc. Information Hiding Workshop,* LNCS 1525, Springer-Verlag, 1998, pp. 258-272.
6. D. Kirovski, H.S. Malvar, and Y. Yacobi, "Multimedia Content Screening Using a Dual Watermarking and Fingerprinting System," *Proc. ACM Multimedia*, ACM Press, 2002, pp. 372-381.
7. S. Katzenbeisser and F.A.P. Petitcolas, eds., *Information Hiding Techniques for Steganography and Digital Watermarking*, Artech House, 2000.
8. Federal Information Processing Standards Publication 140-2, "Security Requirements for Cryptographic Modules," 1994, http://www.itl.nist.gov/fipspubs/by-num.htm.

Boneh and Franklin constructed a public-key encryption scheme with one public encryption key and many private decryption keys.[8] If a broadcaster encrypts once with the public key, each legitimate receiver can decrypt with a different private key. If a coalition of receivers collude to create a new decryption key, an efficient algorithm traces the new key to its creators.

Kiayias and Yung established a black-box traitor-tracing model in which the pirate-decoder employs a self-protection technique.[9] They proved that any system not meeting certain well-defined combinatorial conditions cannot overcome a collusion of size superlogarithmic in object size. To the best of our knowledge the system by Chor et al.[10] is the only traitor-tracing scheme for which the Kiayias-Yung conditions hold.

Fiat and Tassa introduced a dynamic traitor-tracing mechanism in which the set of users is randomly grouped into r subsets, each receiving a distinct symbol.[7] After identifying the subset containing the pirate, the search continues within that subset only. Fiat and Tassa assume that the pirates simply choose one of the multibit symbols available to them during each round of the tracing process.

The main drawback of all traitor-tracing systems (including black-box traitor tracing) is that they require physical confiscation of a suspect client machine to examine it, and they assume that pirates cannot trade the protected content itself. This significantly limits the scenarios in which we can apply traitor tracing.

References

1. G. J. Simmons, “Prisoners’ Problem and the Subliminal Channel,” LNCS 169, Springer-Verlag, 1984, pp. 51-67.
2. R.J. Anderson, “Stretching the Limits of Steganography,” *Proc. Information Hiding Workshop*, LNCS 1174, Springer-Verlag, 1996, pp. 39-48.
3. S. Craver, “On Public-Key Steganography in the Presence of an Activewarden,” *Proc. Information Hiding Workshop*, LNCS 1525, Springer-Verlag, 1998, pp. 355-368.
4. F. Ergun and J. Kilian, “A Note on the Limits of Collusion-Resistant Watermarks,” *Proc. Eurocrypt*, Int’l Assoc. for Crytologic Research (IACR), 1999, pp. 140-149.
5. D. Boneh and J. Shaw, “Collusion Secure Fingerprinting for Digital Data,” *IEEE Trans. Information Theory*, vol. 44, no. 5, Sept. 1998, pp. 1897-1905.
6. B. Pfitzmann and M. Waidner, “Anonymous Fingerprinting,” *Proc. Eurocrypt*, Int’l Assoc. for Crytologic Research (IACR), 1997, pp. 88-102.
7. A. Fiat and T. Tassa, “Dynamic Traitor Tracing,” LNCS 1666, Springer-Verlag, 1994, pp. 354-371.
8. D. Boneh and M. Franklin, “An Efficient Public Key Traitor Tracing Scheme,” LNCS 839, Springer-Verlag, 1999, pp. 338-353.
9. A. Kiayias and M. Yung, “Self-Protecting Pirates and Black-Box Traitor Tracing,” LNCS 2139, Springer-Verlag, 2001, pp. 63-79.
10. B. Chor, A. Fiat, and M. Naor, “Tracing Traitors,” LNCS 839, Springer-Verlag, 1994, pp. 257-270.

9. R.J. Anderson and M. Kuhn, “Tamper Resistance—A Cautionary Note,” *Proc. Usenix Workshop Electronic Commerce*, Usenix, 1996, pp. 1-11.
10. W. Feller, *An Introduction to Probability Theory and Its Applications*, John Wiley & Sons, 1968.
11. R. Motwani and P. Raghavan, *Randomized Algorithms,* Cambridge Univ. Press, 1995.
12. N. Alon, J.H. Spencer, and P. Erdos, *The Probabilistic Method*, John Wiley & Sons, 1992.
13. F. Ergun and J. Kilian, “A Note on the Limits of Collusion-Resistant Watermarks,” *Proc. Eurocrypt, Int. Assoc. for Crytologic Research* (IACR), 1999, pp. 140-149.

Darko Kirovski is a researcher at Microsoft Research. His research interests include system security, multimedia processing, and embedded system design. Kirovski has a PhD in computer science from the University of California, Los Angeles.

Henrique Malvar is a founder and head of the Communication, Collaboration, and Signal Processing group at Microsoft Research. His research interests include signal compression, transforms, wavelets, and cryptography. Malvar has a PhD in electrical engineering from the Massachusetts Institute of Technology.

Yacov Yacobi is a founder and head of the Cryptography and Antipiracy group in Microsoft Research. His current personal research includes economic analysis of antipiracy systems and ID-based encryption. Yacobi has a PhD in electrical engineering from the Technion Israel Institute of Technology.

Contact Darko Kirovski at darkok@microsoft.com.