CAPACITY OF DIGITAL WATERMARKS SUBJECTED TO AN OPTIMAL COLLUSION ATTACK

Jonathan K. Su *Joachim J. Eggers*
Telecommunications Laboratory
Univ. Erlangen-Nuremberg, Erlangen, Germany
$\{$*su,eggers*$\}$*@LNT.de*

Bernd Girod
Information Systems Laboratory
Stanford Univ., Stanford, CA, USA
girod@isl.stanford.edu

ABSTRACT

One envisioned application of digital watermarking is fingerprinting, in which different information is embedded into several copies of the same original signal. Several attackers may collude by combining their copies to produce an attacked signal. In the case of independent watermarks, a collusion-attack model is presented and shown to be analogous to the Gaussian multiple-access channel. The attack parameters are optimized to minimize the information rate under a constraint on the distortion of the attacked signal. Another fingerprinting method, collusion-secure codes, is then related to the attack. Finally, independent and collusion-secure watermarking are compared for the same attacked-signal distortion and probability of false identification.

1 Introduction

Digital watermarking is the secure, imperceptible, robust transmission of information by embedding it directly into digital signals (e.g., digital audio, images, or video) for later retrieval. Envisioned applications of digital watermarking include tracking of distribution paths, access control, and copyright protection. *Security* means only authorized parties can properly decode the embedded information and requires proper cryptographic methods; it is not treated in this paper. *Imperceptibility* means the original and watermarked signals are perceptually equivalent. An *attack* is any processing of the watermarked signal that might impair the watermark; *robustness* means attacks cannot prevent communication without also making the resulting signal useless.

This paper takes a theoretical approach toward watermarking. It addresses the *fingerprinting scenario*, in which an owner publishes several copies of an original signal[1] with a different watermark or *fingerprint* in each copy. In a *collusion attack*, several attackers, each with a different copy, form a *coalition* and combine their copies to create an attacked signal.

Watermarking is treated as a communications problem, in which the owner attempts to communicate over a hostile channel, where the collusion attack forms the channel. Given the attacked signal, the owner attempts to identify the members of the coalition. The owner is successful if at least one attacker is identified. A *false-identification* (false-ID) error occurs if the owner mistakenly identifies an innocent user who did not participate in the attack.

We introduce an optimal collusion attack and then show how it can be used to compare two possible watermarking strategies: embedding information into each copy using *independent watermarks*, and embedding *collusion-secure fingerprinting codes* (CS codes) [1, 4] using the exactly the same watermarking method for each copy.

2 Preliminaries and Notation

The original signal is modeled as an M-dimensional discrete-time/discrete-space random process $\mathbf{x}[\vec{n}]$ whose elements are independent identically distributed (IID) random variables (RVs) $\sim \mathcal{N}(0, \sigma_x^2)$. Information embedding is accomplished by adding an appropriate watermark signal, which we treat as another random process $\mathbf{w}[\vec{n}]$ with IID elements drawn $\mathcal{N}(0, \sigma_w^2)$. $\mathbf{x}[\vec{n}]$ and $\mathbf{w}[\vec{n}]$ are independent of one another. A watermarked copy is denoted by $\mathbf{y}[\vec{n}] = \mathbf{x}[\vec{n}] + \mathbf{w}[\vec{n}]$. A subscript k or ℓ indicates a particular copy or watermark (e.g., $\mathbf{y}_k[\vec{n}]$). It will be clear from context whether independent or collusion-secure watermarking is being discussed.

The copies are indexed from 1 to K; let $\mathcal{K} = \{1, 2, \ldots, K\}$. The attackers are assumed to have the copies in $\mathcal{L} \subseteq \mathcal{K}$. "The copies in $\mathcal{L}$" means the set of copies $\mathbf{y}_\ell[\vec{n}]$ such that $\ell \in \mathcal{L}$. Let $L = |\mathcal{L}|$, where the cardinality of a set $\mathcal{A}$ is $|\mathcal{A}|$.

As a distortion metric, we adopt the mean-squared error (MSE) between a signal $\hat{\mathbf{x}}[\vec{n}]$ and the original $\mathbf{x}[\vec{n}]$ by $D(\hat{\mathbf{x}}, \mathbf{x}) = \mathrm{E}[(\hat{\mathbf{x}}[\vec{n}] - \mathbf{x}[\vec{n}])^2]$. The distortion of a watermarked signal $\mathbf{y}_k[\vec{n}]$ is $D(\mathbf{y}_k, \mathbf{x}) = \sigma_w^2$. To ensure watermark imperceptibility, $\sigma_x^2 \gg \sigma_w^2$.

In some applications (e.g., wide distribution of an audio file, image, or video), the fact that K may be very large may hinder practical fingerprinting schemes. However, it may be reasonable to assume that $L \ll K$. Also, in other applications (e.g., distribution of a sensitive or classified image to a small number of recipients), both K and L may be small.

3 Independent Watermarking

Here we consider *independent* watermarks. The kth copy is $\mathbf{y}_k[\vec{n}] = \mathbf{x}[\vec{n}] + \mathbf{w}_k[\vec{n}]$, where $\mathbf{w}_k[\vec{n}]$ is the kth watermark.

[1] The original signal is sometimes called the "host signal."

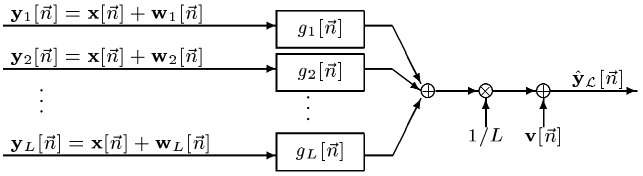
$\mathbf{y}_1[\vec{n}] = \mathbf{x}[\vec{n}] + \mathbf{w}_1[\vec{n}]$
$\mathbf{y}_2[\vec{n}] = \mathbf{x}[\vec{n}] + \mathbf{w}_2[\vec{n}]$
$\mathbf{y}_L[\vec{n}] = \mathbf{x}[\vec{n}] + \mathbf{w}_L[\vec{n}]$
$g_1[\vec{n}]$
$g_2[\vec{n}]$
$g_L[\vec{n}]$
$1/L$
$\mathbf{v}[\vec{n}]$
$\hat{\mathbf{y}}_{\mathcal{L}}[\vec{n}]$

Fig. 1. Collusion attack by LSI filtering and additive noise. In the diagram, it is assumed that $\mathcal{L} = \{1, 2, \ldots, L\}$.

$\mathbf{w}_k[\vec{n}]$ conveys a *message* m_k, $k \in \mathcal{K}$. Each message is a string of B_{ind} bits. The watermarks and messages are assumed to be mutually independent.

Let $\hat{\mathbf{y}}_{\mathcal{L}}[\vec{n}]$ denote the attacked signal generated by the coalition. The owner acquires $\hat{\mathbf{y}}_{\mathcal{L}}[\vec{n}]$ and attempts to identify the colluders. To do so, the owner decodes the messages $\hat{m}_k$, $k \in \mathcal{K}$. For each k, if $\hat{m}_k = m_k$, the owner decides that the coalition used the kth copy.

If copy $\mathbf{y}_k[\vec{n}]$ was *not* used during collusion, the decoded message $\hat{m}_k$ is a string of random, equiprobable bits. The probability of false-ID error is $P_{FI} = (K-L)2^{-B_{\text{ind}}}$. If P_{FI} is given and the coalition may have up to $L' \le K$ copies, then $P_{FI} \ge (K-L)2^{-B_{\text{ind}}}$, $1 \le L \le L'$. The right-hand side is maximum for $L = 1$, so $B_{\text{ind}} \ge \log_2\left(\frac{K-1}{P_{FI}}\right)$.

3.1 Attack Model and Optimization

The attackers wish to generate an *attacked signal* $\hat{\mathbf{y}}_{\mathcal{L}}[\vec{n}]$ such that it is difficult to decode the messages m_ℓ, $\ell \in \mathcal{L}$, and the attacked-signal distortion $D(\hat{\mathbf{y}}_{\mathcal{L}}, \mathbf{x})$ remains acceptably small. To facilitate analysis, we assume that the attackers are limited to multi-input, single-output (MISO) linear shift-invariant (LSI) filtering and additive Gaussian noise; a diagram appears in Fig. 1. The attacked signal is

$$\hat{\mathbf{y}}_{\mathcal{L}}[\vec{n}] = \frac{1}{L}\sum_{\ell\in\mathcal{L}} g_\ell[\vec{n}] * \underbrace{(\mathbf{x}[\vec{n}] + \mathbf{w}_\ell[\vec{n}])}_{\mathbf{y}_\ell[\vec{n}]} + \mathbf{v}[\vec{n}], \qquad (1)$$

where $g_\ell[\vec{n}]$, $\ell \in \mathcal{L}$, are the impulse responses of individual single-input, single-output (SISO) filters, and the noise $\mathbf{v}[\vec{n}]$ has mean zero and power spectrum $\Phi_{vv}(\vec{\omega})$ and is independent of $\mathbf{x}[\vec{n}]$ and $\mathbf{w}_k[\vec{n}]$, $k \in \mathcal{K}$.

The attackers adopt Kerckhoff's principle and assume the owner has knowledge of the filters $g_\ell[\vec{n}]$ and the statistical properties $\mathbf{v}[\vec{n}]$. Depending on the application, the owner may also be able to use the original signal $\mathbf{x}[\vec{n}]$ to assist decoding. To maintain generality, we assume that the owner performs decoding on the signal $\mathbf{z}_{\mathcal{L}}[\vec{n}] = \hat{\mathbf{y}}_{\mathcal{L}}[\vec{n}] - \frac{a}{L}\sum_{\ell\in\mathcal{L}} g_\ell[\vec{n}] * \mathbf{x}[\vec{n}]$, $0 \le a \le 1$. Then $a = 1$ means all original-signal interference can be eliminated, and $a = 0$ means no original-signal interference can be eliminated.[2]

3.1.1 Analogy with Gaussian Multiple-Access Channel

The above arrangement is analogous to the *Gaussian multiple-access channel* (GMAC) with L independent users, each with power P, and additive white Gaussian noise with power N_0. The transmitters then have equal rates R, and the total achievable rate LR. The dominating bound on the rate is $LR \le \frac{1}{2}\log_2\left(1 + \frac{LP}{N_0}\right)$ [3]. When the signals are M-dimensional, the noise is colored and has power spectrum $\Phi_{nn}(\vec{\omega})$, and the transmitted signals each have power spectra $\Phi_{ss}(\vec{\omega})$, the white-noise GMAC result becomes

$$R \le \frac{1}{(2\pi)^M}\int_\Omega \frac{1}{2L}\log_2\left(1 + \frac{L\Phi_{ss}(\vec{\omega})}{\Phi_{zz}(\vec{\omega})}\right) d\vec{\omega}. \qquad (2)$$

where Ω is the M-dimensional hypercube centered at the origin with side length 2π. For the attack (1), each filtered watermark $\frac{1}{L}g_\ell[\vec{n}] * \mathbf{w}_\ell[\vec{n}]$ corresponds to a transmitted signal, and $\frac{1-a}{L}\sum_{\ell\in\mathcal{L}} g_\ell[\vec{n}] * \mathbf{x}[\vec{n}] + \mathbf{v}[\vec{n}]$ corresponds to the noise.

3.1.2 Optimum Attack on Independent Watermarks

Mathematically, the attackers must solve a constrained optimization problem: Given a bound D on the attack distortion, select $g_\ell[\vec{n}]$, $\ell \in \mathcal{L}$ and $\Phi_{vv}(\vec{\omega})$ to minimize R in (2) such that $D(\hat{\mathbf{y}}_{\mathcal{L}}, \mathbf{x}) \le D$. From the symmetry of the problem, the filters should be the same: $g_\ell[\vec{n}] = g[\vec{n}]$, $\ell \in \mathcal{L}$.

Let $G(\vec{\omega})$ denote the transfer function corresponding to $g[\vec{n}]$. The coalition seeks $G(\vec{\omega})$ and $\Phi_{vv}(\vec{\omega})$ to minimize R such that $D(\hat{\mathbf{y}}_{\mathcal{L}}, \mathbf{x}) = D$. The calculus of variations can be used to solve this problem We find that $G(\vec{\omega})$ and $\Phi_{vv}(\vec{\omega})$ are constants: $G(\vec{\omega}) = 1 - D/\sigma_x^2 = G^*$, and $\Phi_{vv}(\vec{\omega}) = G^*(D - G^*\sigma_w^2/L) = \sigma_v^{*2}$. Hence the optimum attack is memoryless. The rate R of each watermark is bounded by

$$R \le \frac{1}{2L}\log_2\left(1 + \frac{(\sigma_x^2 - D)\sigma_w^2/L}{\sigma_x^4 - (\sigma_x^2 - D)\left(a(2-a)\sigma_x^2 + \sigma_w^2/L\right)}\right). \qquad (3)$$

Eq. (3) is a convenient generalization of the single-copy case ($K = L = 1$) considered in [7, Eq. 21]. Note the following: **(A)** If $\sigma_x^2 = 0$, then $R = 0$ for $D \ge 0$. It is not possible to watermark a "flat" original. **(B)** $R = 0$ for $D \ge \sigma_x^2$; the attackers set $\hat{\mathbf{y}}_{\mathcal{L}}[\vec{n}] = 0$. However, $\hat{\mathbf{y}}_{\mathcal{L}}[\vec{n}]$ is unlikely to be useful to the attackers. **(C)** The total rate is LR, and it is not surprising to see that $LR \to 0$ as $L \to \infty$.

3.2 Estimation Attack on Independent Watermarks

A suboptimal but nonetheless potentially effective attack on independent watermarks is an *estimation attack*. The coalition uses MISO filtering to compute the estimate $\hat{\mathbf{x}}_{\mathcal{L}}[\vec{n}] = \frac{1}{L}\sum_{\ell\in\mathcal{L}} h_\ell[\vec{n}] * \mathbf{y}_\ell[\vec{n}]$. The filters are selected to minimize $D(\mathbf{x}, \hat{\mathbf{x}}_{\mathcal{L}})$.

By symmetry, the filters are identical, and the problem becomes equivalent to estimating $\mathbf{x}[\vec{n}]$ from $\mathbf{x}[\vec{n}] + \bar{\mathbf{w}}_{\mathcal{L}}[\vec{n}]$, where $\bar{\mathbf{w}}_{\mathcal{L}}[\vec{n}] = \frac{1}{L}\sum_{\ell\in\mathcal{L}} \mathbf{w}_\ell[\vec{n}]$. The solution is a Wiener filter with impulse response $h[\vec{n}] = \frac{\sigma_x^2}{\sigma_x^2 + \sigma_w^2/L}\delta[\vec{n}]$. It can be shown (e.g., see [6]) that $D(\mathbf{x}, \hat{\mathbf{x}}_{\mathcal{L}}) = \frac{\sigma_x^2\sigma_w^2}{L\sigma_x^2 + \sigma_w^2} = D_{\min}(L)$. $D_{\min}(L)$ indicates that this distortion is also the minimum that the coalition can achieve; $D_{\min}(L) \to 0$ as $L \to \infty$.

Consider image watermarking, where the original signal takes on integer values in $[0, 255]$. If we neglect quantization of the watermarked copies $\mathbf{y}_k[\vec{n}]$, the estimation error is

[2] In the single-copy case ($K = L = 1$), it is theoretically possible to construct a watermarking scheme such that a decoder without access to $\mathbf{x}[\vec{n}]$ can perform as if $a = 1$ [2]. No such result is currently known for the multiple-access channel.

distributed $\mathcal{N}\left(0, \frac{\sigma_x^2 \sigma_w^2}{L\sigma_x^2+\sigma_w^2}\right)$. Suppose the attackers perform estimation and then quantize $\hat{\mathbf{x}}_{\mathcal{L}}[\vec{n}]$ to integer values. When $L = (4 - \sigma_x^{-2})\sigma_w^2$, over 68% of the estimated samples in $\hat{\mathbf{x}}_{\mathcal{L}}[\vec{n}]$ will be quantized to the original values in $\mathbf{x}[\vec{n}]$. When $L = (16 - \sigma_x^{-2})\sigma_w^2$, this percentage exceeds 95%.[3] These results show that the maximum number of independent watermarks may be severely limited in practice.

4 Collusion-Secure Watermarking

The preceding section considered *independent* watermarks. A different way to identify colluders uses *collusion-secure* (CS) *codes* [1, 4]. **CS codes assume the existence of a reliable watermarking method. They describe the information to be embedded rather than the watermarking method itself and are independent of the mechanisms for information embedding and retrieval**.

CS codes were developed independently of the type of data (e.g., ASCII text, audio, or images) to be watermarked. As such, CS codes have not been linked to the distortion of the attacked data. Here, we review CS codes and then tie CS codes to the distortion of $\hat{\mathbf{y}}_{\mathcal{L}}[\vec{n}]$ in our model.

4.1 Review of CS Codes

A CS code consists of K *codewords* or messages, which are length-B_{cs} bit strings b_k, $k \in \mathcal{K}$. Codeword b_k is embedded in copy $\mathbf{y}_k[\vec{n}]$, and $b_k(j)$ denotes the jth bit in b_k. One assumption behind CS codes is that identical bits at the same position in different codewords cause identical embedding modifications in the respective copies. That is, if $b_k(j) = b_\ell(j)$, then identical modifications are made to $\mathbf{y}_k[\vec{n}]$ and $\mathbf{y}_\ell[\vec{n}]$ to embed this bit. A second assumption is that the modifications associated with each bit position j are unknown to the attackers.

A *K-secure code* Γ with ϵ-error [1] ensures that the owner can always identify at least one attacker while keeping $P_{FI} < \epsilon$. We present the code in [4]. The bit positions are partitioned into $K+1$ consecutive *blocks* $\mathcal{B}_s$, $s \in \{0, 1, \ldots, K\}$, and each block contains B bits.

The total length B_{cs} of each codeword is $B_{\text{cs}} = (K+1)B$ bits. Codeword construction is simple: For codeword b_k, the bits in blocks $\mathcal{B}_0$ through $\mathcal{B}_{k-1}$ are all 0, and the bits in blocks $\mathcal{B}_k$ through $\mathcal{B}_K$ are all 1.

By comparing the copies in $\mathcal{L}$, the attackers can *detect* the bits in blocks where two or more of their copies differ. However, because the attackers do not know which modifications correspond to which bit positions, they cannot decode the bits. Let $\mathcal{B}_{\text{det}}(\mathcal{L})$ denote the blocks of bit positions that the attackers can detect. In a worst-case attack, the attackers can produce an attacked signal in which the probability of bit error $P_E = 0.5$ for bits in $\mathcal{B}_{\text{det}}(\mathcal{L})$.

Some of the bits (e.g., those in $\mathcal{B}_0$ and $\mathcal{B}_K$) are identical in every copy in $\mathcal{L}$. The modifications associated with these bits are the same in all of the attackers' copies, so these bits cannot be detected. Let $\mathcal{B}_{\text{undet}}(\mathcal{L})$ denote the blocks that the attackers cannot detect. A third assumption is the *marking assumption* [1], which states that the coalition cannot alter or erase bits in $\mathcal{B}_{\text{undet}}(\mathcal{L})$.[4]

Given $\hat{\mathbf{y}}_{\mathcal{L}}[\vec{n}]$, the owner decodes the bit string $\hat{b}$. Let $\hat{b}(\mathcal{B}_s)$ denote the decoded bits in block $\mathcal{B}_s$, and $\text{wt}(\hat{b}(\mathcal{B}_s))$ be the Hamming weight of $\hat{b}(\mathcal{B}_s)$. If $\text{wt}(\hat{b}(\mathcal{B}_k)) \neq \text{wt}(\hat{b}(\mathcal{B}_{k-1}))$, the owner decides that copy $\mathbf{y}_k[\vec{n}]$ was used during collusion; otherwise, not.

Let P_{FI} be given. It is shown in [4] that the block length B must satisfy $B > 2(K+1)^2 \ln\left(\frac{2}{P_{FI}}\right)$, so $B_{\text{cs}} > (K+1)B = O(K^3)$. The cubic increase in B_{cs} makes K-secure codes impractical for large K. However, in some scenarios the coalition cannot acquire more than L' copies, where $L' \ll K$. In this case, a new CS code Γ_2 can be constructed. Each codeword of Γ_2 is formed by randomly concatenating S codewords from a K_1-secure "inner code" Γ with $L' < K_1 \ll K$. Under the marking assumption,

$$S \geq \frac{L' \ln \frac{2}{P_{FI}}}{\frac{L'}{K} - 1 + \ln \frac{K}{L'}}, \tag{4}$$

and the inner code Γ must have $P_{FI,\Gamma} \leq P_{FI}/2S(K_1 - 1)$. With B_{cs} denoting the codeword length for Γ, the codeword length for the new CS code Γ_2 is SB_{cs}.

4.2 Lower Bound on Distortion

Clearly, the independent watermarking model of the previous section does not apply here. However, for the bits in $\mathcal{B}_{\text{undet}}(\mathcal{L})$, it is as if the attackers had a single copy. Let $D_{\text{undet}} \geq 0$ denote the distortion of $\hat{\mathbf{y}}_{\mathcal{L}}[\vec{n}]$ associated with the undetectable bits. Apply (3) with $L = 1$ and $D = D_{\text{undet}}$ to find the upper bound on the rate of the undetectable bits—the highest rate for which the marking assumption can hold. Also, let $D_{\text{det}} \geq 0$ be the distortion associated with the detectable bits. Note that D_{undet} and D_{det} are independent of the specific set $\mathcal{L}$.

Next, assume the overall distortion can be written as $D(\hat{\mathbf{y}}_{\mathcal{L}}, \mathbf{x}) = \frac{|\mathcal{B}_{\text{undet}}(\mathcal{L})|}{B_{\text{cs}}} D_{\text{undet}} + \frac{|\mathcal{B}_{\text{det}}(\mathcal{L})|}{B_{\text{cs}}} D_{\text{det}}$. This assumption holds for direct-sequence spread-spectrum and related (e.g., frequency-domain or wavelet-based) watermarking methods that apply a unitary transformation to the original signal and operate in the transformed-signal domain. In a worst-case attack, besides $P_E = 0.5$ for detectable bits, D_{undet} is zero.

Typically, the cardinality $|\mathcal{L}|$, not $\mathcal{L}$ itself, is restricted. Let $\mathcal{K}(L') = \{\mathcal{L} : \mathcal{L} \subseteq \mathcal{K}, |\mathcal{L}| = L'\}$ for L' given and $2 \leq L' \leq K$. A lower bound on the distortion when $|\mathcal{L}| = L'$ is then $D(\hat{\mathbf{y}}_{L'}, \mathbf{x}) \geq D_{\text{undet}} \sum_{\mathcal{L} \in \mathcal{K}(L')} \Pr(\mathcal{L}) \frac{|\mathcal{B}_{\text{undet}}(\mathcal{L})|}{B_{\text{cs}}}$, where $\hat{\mathbf{y}}_{L'}$ indicates that the attackers have some combination of L' copies in $\mathcal{K}$. For $\mathcal{L} \subseteq \mathcal{K}$, $|\mathcal{B}_{\text{undet}}(\mathcal{L})| = V_{\text{cs}} - (\max \mathcal{L} - \min \mathcal{L})B$. A counting exercise then shows that the above summation simplifies to $\frac{2}{L'+1}$, which is also sensible when $L' = 1$ and all bits are undetectable.

[3] For $\sigma_x^2 \gg \sigma_w^2$, the estimation attack is only slightly more effective than simply averaging the copies together [5]. In the latter case, the 68%- and 95%-values of L are, respectively, $4\sigma_w^2$ and $16\sigma_w^2$, which are negligibly greater than L for the estimation-attack if σ_x^2 is large.

[4] CS codes that can tolerate a certain probability of undetectable bit error may be found in [4].

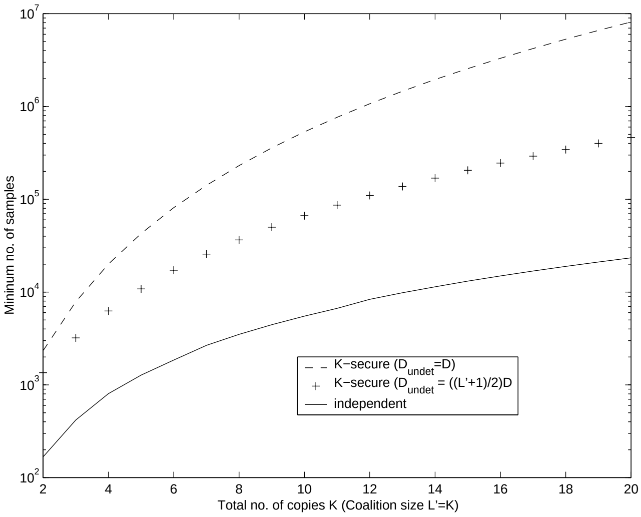

Fig. 2. Comparison of independent and CS watermarks for small K when coalition may have all K copies.

Thus, $D(\hat{\mathbf{y}}_{L'}, \mathbf{x}) \geq \frac{2}{L'+1} D_{\text{undet}}$, $1 \leq L' \leq K$. Now we relate the distortion D to the upper bound on the rate R_{undet} of the undetectable bits. With K, L', and D given, set $L = 1$ and $D_{\text{undet}} = \frac{L'+1}{2} D$ in (3) to get

$$R_{\text{undet}} \leq \frac{1}{2} \log_2 \left(1 + \frac{(\sigma_x^2 - \frac{L'+1}{2} D)\sigma_w^2}{\sigma_x^4 - (\sigma_x^2 - \frac{L'+1}{2} D)\left(a(2-a)\sigma_x^2 + \sigma_w^2\right)} \right). \tag{5}$$

In (5), even if D is small, the coalition can make $D_{\text{undet}} = \frac{L'+1}{2} D$ large if L' is large. In some cases, the coalition requires $D_{\text{undet}} \leq D$; in these cases, we just set $L' = 1$ in (5). When $L' \ll K$ and a new CS code Γ_2 is used, Eq. (5) still applies because the codewords in Γ_2 consist of concatenated codewords from the K_1-secure code Γ with $L' < K_1$.

5 Comparison of Independent and CS Watermarking

With σ_x^2, σ_w^2, D, P_{FI}, K, and L' given, we can compare independent and CS watermarking by looking at the ratios B_{ind}/R and $B_{\text{cs}}/R_{\text{undet}}$ (or $SB_{\text{cs}}/R_{\text{undet}}$). These ratios give the minimum number of samples such that the marking assumption can hold and P_{FI} can be achieved. In the graphs that follow, $\sigma_x^2 = 3000$, $\sigma_w^2 = 30$, $D = 60$, and $P_{FI} = 10^{-5}$. Hence $10 \log_{10} \sigma_x^2/\sigma_w^2 = 20$ dB, and $10 \log_{10} \sigma_x^2/D = 17$ dB.

Fig. 2 considers the case when K is small and the attackers can acquire all copies. Fig. 3 considers large K when the coalition has $L' = \lceil \ln K \rceil$ copies. (The staircase appearance of the curves in Fig. 3 results from the ceiling operation on $\ln K$.) Independent watermarks require far fewer samples than CS codes. In practice, however, the complexity of implementing independent watermarks may be much higher than that of CS codes. The shortened CS codes only become useful for K larger than about 100. Unfortunately, it appears that CS codes may require a prohibitively large number of samples. This requirement may be relaxed by abandoning the marking assumption and allowing for some bit errors among the undetectable bits [4].

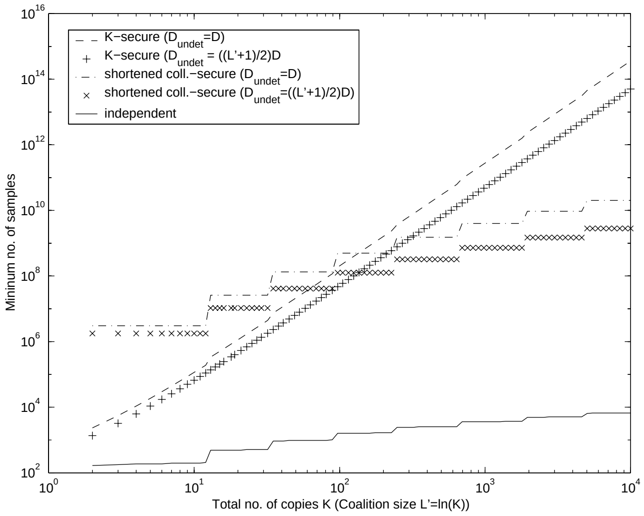

Fig. 3. Comparison of independent and CS watermarks for large K when coalition has up to $\lceil L' = \ln K \rceil$ copies.

6 Conclusions

An optimum collusion attack by a coalition with multiple, independently watermarked copies of a signal was presented. The attack minimizes the amount of information that can be recovered from the coalition's attacked signal for a given attacked-signal distortion. With some additional assumptions, the attack was also applied to watermarking with collusion-secure codes.

References

[1] D. Boneh and J. Shaw. Collusion-secure fingerprinting for digital data. In *Advances in Cryptology – Proc. CRYPTO '95*, vol. 963, pp. 452–465. Lecture Notes in Computer Science, Springer, Aug. 1995.

[2] B. Chen and G. W. Wornell. Provably robust digital watermarking. In *Proc. SPIE Multimedia Systems and Applications II*, vol. 3845, 1999.

[3] T. M. Cover and J. A. Thomas. *Elements of Information Theory*. John Wiley & Sons, New York, 1991.

[4] H.-J. Guth and B. Pfitzmann. Error- and collusion-secure fingerprinting for digital data. In *Prelim. Proc. 3rd Intl. Information Hiding Workshop*, Dresden, Germany, Sep.–Oct. 1999.

[5] H. S. Stone. Analysis of attacks of image watermarks with randomized coefficients. Technical report, NEC Research Inst., May 1996.

[6] J. K. Su and B. Girod. On the imperceptibility and robustness of digital fingerprints. In *Proc. IEEE Intl. Conf. Multimedia Comp. Sys.*, vol. 2, pp. 530–535, Jun. 1999.

[7] J. K. Su and B. Girod. Fundamental performance limits of power-spectrum condition-compliant watermarks. In *Proc. SPIE Security & Watermarking Multimedia Contents*, vol. 3971, Jan. 2000.