Resistance of Digital Watermarks to Collusive Attacks

Joe Kilian[1]

F. Thomson Leighton[2, 6]

Lesley R. Matheson[3, 6]

Talal G. Shamoon[3, 6]

Robert E. Tarjan [3, 4, 6]

Francis Zane [3, 5, 6]

July 27, 1998

Abstract

In many schemes for digital watermarking, an attacker in possession of a small fixed number of differently marked copies of a single document can easily produce an unmarked copy. In contrast, we show in this paper that if the watermarks have a component-wise normal distribution, then $\Omega\left(\sqrt{n/\ln m}\right)$ differently marked copies are required to successfully erase any mark with non-negligible probability. Here n is the dimensionality of the watermarks, and m is the number of differently marked copies distributed. This result holds in a specific but natural statistical model of watermarking.

[1]NEC Research Institute, 4 Independence Way, Princeton, NJ 08540.

[2]Dept. of Mathematics, Massachusetts Institute of Technology, Cambridge, MA 02139.

[3]STAR Lab, InterTrust Technologies, 460 Oakmead Parkway, Sunnyvale, CA 94026.

[4]Department of Computer Science, Princeton University, Princeton, NJ 08544. Research at Princeton University partially supported by the National Science Foundation, Grant CCR 9626862.

[5]Bell Laboratories, Lucent Technologies, Murray Hill, NJ 07974.

[6]Some work done at the NEC Research Institute, 4 Independence Way, Princeton, NJ 08540.

Resistance of Digital Watermarks to Collusive Attacks

1 Introduction

With digital representation of multimedia data becoming more-and-more widespread, the need to protect such data is becoming vastly more important. Cryptography offers one form of protection, but once data is decrypted, it is no longer protected. *Digital watermarking* (also called *digital fingerprinting*) offers a supplemental form of protection that extends into clear space. By watermarking, we mean inserting some information (the watermark) into a document, such as an audio track, still image, or video stream, in such a way that (1) the marked document differs imperceptibly from the original, (2) the information contained in the mark can be recovered by authorized parties, and (3) removal of the mark by modifying the marked document results in rendering the document useless.

There are many ways to watermark documents (see the survey in [5]). Most of these methods, however, are easy to defeat, especially if the attacker has access to a small constant number of differently marked copies of the same document. On the other hand, the paper [5] proposes the use of normally distributed random values as watermarks, and sketches an argument that such watermarks are highly resistant to collusive attacks. The purpose of our paper is to fill in the mathematical justification for this claim. Specifically, we show that, in a natural statistical model, an attacker who does not possess the original document or the watermarks will require $\Omega\big(\sqrt{n/\ln m}\,\big)$differently marked copies of the document in order to have a non-negligible chance of producing an unmarked version. Here n is the dimension of the watermark in a single document, and m is the number of differently marked copies that exist. Our analysis relies on basic properties of normal, chi and chi-square distributions. We argue from first principles.

We are aware of almost no research on the collusion resistance of watermarking schemes. One interesting result was obtained by Boneh and Shaw [3]. In their model, a "fingerprint" consists of a collection of embedded bits. They asume that an attacker possessing differently marked copies can erase a bit if and only if both values of the bit are present among the marked copies. They propose a coding scheme that uses ideas of of Chor, Fiat, and Naor [4] to show that $O(k^4 \log k)$ bits protect against colluders that obtain k different copies. They do not address the question of how to provide individual robust marking bits, nor do they consider the possibility that a bit might be robust to attack even if both its values are present in a collection of marked copies. This makes their approach orthogonal to ours. Work related to that of Boneh and Shaw but in a much weaker model was done by Blakely, Meadows, and Purdy [2]. The Boneh-Shaw scheme has been used as a building block for schemes that achieve more sophisticated cryptographic properties [10, 11, 12].

Work more closely related to ours is that of Karakos and Papamarcou [9], which was done contemporaneously with, and independent of, our work. They consider essentially the same statistical model as we do, but they consider only one kind of attack, namely convex combination of differently marked documents plus the addition of Gaussian noise. They show that maximum-power watermarks can carry up to 0.5 bits of information per dimension while being secure against single-copy attacks, and up to 0.146 bits per dimension while being secure against two-copy attacks. They do not consider k-way collusive attacks for $k > 2$.

Our paper is organized as follows. In Section 2, we describe our statistical model of watermarking. In Section 3, we present an overview of our analysis. In Section 4, we present the remaining details of the analysis. Section 5 contains some final remarks.

2 Statistical Model

We begin by describing a general watermarking scheme. This scheme assumes that the original document is available for watermark extraction. The base document to be protected is simply a vector $V = (v_1, v_2, \ldots, v_n)$ of real numbers. In [5] this vector consists of the most significant DCT components of an image or signal, but our analysis applies to any such vector.

The owner of the document prepares up to m marked copies, denoted by $W^{(1)}, W^{(2)}, \ldots, W^{(m)}$, for distribution. To obtain the j^{th} marked copy $W^{(j)} = (w_1^{(j)}, w_2^{(j)}, \ldots, w_n^{(j)})$, the owner computes a watermark $X^{(j)} = (x_1^{(j)}, x_2^{(j)}, \ldots, x_n^{(j)})$ and then computes $W^{(j)}$ by adding the watermark to the document: $W^{(j)} = V + X^{(j)}$. The owner can then distribute each $W^{(j)}$ to a different individual or group.

The owner looks for a watermark X in a document copy V' by first computing an *implied watermark* $X' = V' - V$, then computing a *similarity measure* $s(X', X)$, and finally reporting that the watermark is present if $s(X', X) > t$, where t is some fixed *similarity threshold*. (We say X' *tests positive* against X'). Otherwise ($s(X', X) \leq t$), the owner reports that the watermark is absent. (We say X' *tests negative* against X.).

The simplest similarity measure is one such that X' tests positive only if it is exactly equal to X in all components. But such a measure is completely useless in practice, since any change at all in the marked document will erase the watermark. We want the watermark to be erasable only by *perceptibly* changing the document. Thus we need a similarity measure that provides robustness against small changes in the document. There are many such measures, including the dot product $X' \cdot X$ and the correlation coefficient $(X' - \overline{X}') \cdot (X - \overline{X}) / (\|X' - \overline{X}'\| \cdot \|X - \overline{X}\|)$. Here $\|\cdot\|$ denotes the Euclidean norm, $\|X\| = \sqrt{X \cdot X}$, and $\overline{X}$ denotes the mean of the components of X; $\overline{X}'$ is defined analogously.

As a similarity measure, we shall use $s(X', X) = X' \cdot X / \|X'\|$, as proposed in [5]. As we shall see, this measure is relatively straightforward to analyze and has nice properties in our setting. Our analysis

applies with minor changes to the dot product itself, to the correlation coefficient, and to related measures.

We need to quantify what it means to "perceptibly" change a document. We adopt the standard approach of calling a copy V' of V *good* if $\|V' - V\| < \delta\sqrt{n}$ for some appropriate constant δ and *bad* otherwise. Whether goodness adequately models the perception threshold is an experimental question; we assume that any copy of the document not sufficiently close to the original in the root-mean-square error sense is sufficiently degraded in fidelity as to be worthless. *This is the first major assumption of our model.*

Our model of attack is as follows. An adversary, having somehow obtained k of the m marked copies, say $W^{(1)}, W^{(2)}, \ldots, W^{(k)}$, prepares an attack document $V^* = (v_1^*, v_2^*, \ldots, v_n^*)$ intended to resemble V but *not* contain *any* of the watermarks. This covers the case that $k = 1$ and the "adversary" is some standard method of document processing, such as lossy compression followed by decompression, or digital-to-analog followed by analog-to-digital conversion. The adversary *succeeds* if V^* is a good copy and in addition either $s(X^*, X^{(j)}) > t$ for some $j > k$ (a *false positive*) or $s(X^*, X^{(j)}) \leq t$ for all $j, 1 \leq j \leq k$ (a *false negative*). Otherwise, the adversary *fails*.

We assume that the watermark detector is under the control of the owner. Thus the attacker gets one try, and is exposed upon failure. *This is the second major assumption of our model.*

Our analysis is statistical; it relies on the base document V having some inherent variability. We assume specifically that each component v_i of V is independent and normally distributed, without loss of generality selected from $N(0, 1)$, the normal (Gaussian) distribution with mean zero and variance one. *This is the third major assumption of our model.* Whether this is a realistic assumption is a question that must be resolved experimentally. We shall see, however, that for a natural class of attacks no such assumption on the variability of V is necessary. We shall discuss the assumptions of our model further in Section 5.

We have not yet specified the nature of the watermarks $X^{(1)}, X^{(2)}, \ldots, X^{(m)}$. This is something that is entirely under the control of the document owner. The particular scheme that we shall study, proposed in [5], selects each $x_i^{(j)}$ independently from $N(0, \alpha^2)$, the normal distribution with mean 0 and variance α^2. We shall require $\alpha < \delta < 1$; if $\delta \geq 1$, a significent fraction of the possible documents are good copies of each other, and if $\alpha \geq \delta$, a marked copy has a significant chance of being bad.

3 Overview of the Analysis

Within the model presented in Section 2, we shall study the adversary's (*any* adversary's) chance of success, as a function of k, the number of copies obtained by the adversary; m, the number of copies distributed; and n, the dimensionality of the watermarks. Our goal is to determine how many copies the adversary needs in order to have a non-negligible chance of success.

From the document owner's point of view, three bad events could occur. First, one or more of the distributed marked copies $W^{(j)}$ could be bad. Second, a copy *not* actually containing a watermark $X^{(j)}$ could pass the watermark detection test for $X^{(j)}$; we call this a *false positive*. A false positive could occur for a marked copy other than $W^{(j)}$ or for the attack document V^*; any false positive is bad. Third, the attack document could be good but fail the watermark detection test for every $X^{(j)}, 1 \leq j \leq k$. We call this a *false negative*.

The owner is in control of the detection threshold t. The owner must set t high enough so that false positives are sufficently unlikely, but no higher, to avoid increasing the likelihood of false negatives.

Analyzing the watermarking scheme requires estimating the probability that each of the three bad events will occur. In this section we analyze the probability of bad copies and of false positives (in the process determining good choices for α and t). We also give an overview of the analysis of false negatives, which is the hardest part of the proof. We fill in the details of the false negative analysis in Section 4.

3.1 Properties of Distributions

Our analysis relies not only on properties of $N(0,1)$, but also on properties of the chi and chi-square distributions. We denote by χ^2_n a random variable having a chi-square distribution with n degrees of freedom. One way to obtain χ^2_n is to sum the squares of n independent $N(0,1)$ random variables. We denote by χ_n a random variable having a chi distribution with n degrees of freedom; χ_n is obtained as the square root of χ^2_n. In our analysis, χ^2_n arises from the dot product with itself of a vector whose components are $N(0,1)$; χ_n arises as the Euclidean norm of such a vector. For large n, χ_n is approximately $N(\sqrt{n}, \frac{1}{2})$ [8]; χ_n tends to this limiting distribution rapidly. For large n, χ^2_n is approximately $N(n,\ 2n)$ [8]; χ^2_n tends to this limiting distribution slowly, but we are interested in large values of n. Furthermore, χ^2_n is skewed so that the right tail is long but the left tail is short, and the left tail is the tail of interest to us.

We shall need bounds on the tails of the normal, chi, and chi-square distributions. We use the following standard bound on the tails of the normal distribution:

Lemma 1 ([7])

$$P(N(0,1) > y) \leq \frac{1}{y\sqrt{2\pi}}\ e^{-y^2/2} \text{ for any } \ y > 0$$

One can approximate the two tails of the chi distribution by applying Lemma 1 to the corresponding normal-distribution approximation. This results in low-order errors. To avoid these errors, we provide below rigorous bounds on the chi-distribution tail probabilities that follow from Lemma 1. Proofs are in the appendix. For simplicity in deriving the left tail bound we assume that n is even and sufficiently large, but this assumption is not inherent.

Lemma 2

$$P(\chi_n < \sqrt{n} - y) \leq \frac{1}{2y\sqrt{\pi}}\ e^{-y^2} \ \textit{for}\ 2 \leq y < \sqrt{n},\ n \geq 12\ \textit{even}$$

Lemma 3

$$P(\chi_n > \sqrt{n} + y) \leq \frac{1}{2y\sqrt{\pi(1 - y/(3\sqrt{n}))}}\ e^{-y^2(1-y/(3\sqrt{n}))}$$
$$\textit{for}\ 0 \leq y < 3\sqrt{n},\ n \geq 1$$

We can obtain probability bounds on the tails of χ_n^2 directly from Lemmas 2 and 3.

3.2 Good Copies

We begin our analysis by studying the chance that all copies are good. The owner can make good copies very likely by choosing α sufficiently small compared to δ. On the other hand, to obtain the most durable watermarks, the owner wants to maximize α. Thus the goodness requirement imposes an upper bound on α; knowing δ and an acceptable probability p of generating a bad copy, the owner wants to maximize α while keeping the probability of a bad copy below p.

For a particular copy $W^{(j)}$ to be good, it must be the case that $\|X^{(j)}\| \leq \delta\sqrt{n}$. Since each component $x_i^{(j)}$ of $X^{(j)}$ is $N(0, \alpha^2) = \alpha N(0,1), \|X^{(j)}\|$ is $\alpha\,\chi_n$. Thus we want

$$\chi_n \leq \frac{\delta}{\alpha}\sqrt{n} \qquad \text{with high probability}$$

For *all* m copies to be good, we require

$$\alpha \leq \frac{\delta\sqrt{n}}{\max_m \chi_n} \qquad \text{with high probability}$$

where $\max_m \chi_n$ denotes the maximum of m independent χ_n random variables. Observe that the bound on α is linear in δ. Table 1 gives allowed ratios of α/δ to guarantee with a given failure probability that all copies are good, based on this formula. We see that, for realistic values of n and m, the owner can make α quite close to δ while guaranteeing goodness with probability close to one.

n	m	p	α/δ
10	1	10^{-2}	.656
10^2	10	10^{-3}	.787
10^3	10^6	10^{-4}	.862
10^4	10^8	10^{-5}	.949

Table 1: Bounds on α/δ to guarantee goodness with specified failure probability.

To study the asymptotic behavior of the goodness constraint, we apply Lemma 3. Let us set $(\delta/\alpha)\sqrt{n} = \sqrt{n} + a$, which is equivalent to $\alpha = \delta/(1 + a/\sqrt{n})$. Then, for $0 \leq a < 3\sqrt{n}$, $n \geq 1$, Lemma 3 gives

$$P\big(\chi_n > (\delta/\alpha)\sqrt{n}\big) \;\leq\; \frac{1}{2a\sqrt{\pi(1 - a/(3\sqrt{n}))}}\, e^{-a^2(1-a/(3\sqrt{n}))}$$

Thus the probability that *any* of the m copies is bad is at most

$$\frac{m}{2a\sqrt{\pi(1 - a/(3\sqrt{n}))}}\, e^{-a^2(1-a/(3\sqrt{n}))}$$

To guarantee a bound of p on the probability of a bad copy, we need

$$p \geq \frac{m}{2a\sqrt{\pi(1 - a/(3\sqrt{n}))}}\, e^{-a^2(1-a/(3\sqrt{n}))}$$

This is equivalent to

$$a \;\geq\; \left(\frac{\ln(m/p) \;-\; \ln\Big(2a\sqrt{\pi(1 - a/(3\sqrt{n}))}\Big)}{1 \;-\; a/(3\sqrt{n})}\right)^{\frac{1}{2}}$$

The choice $a = \sqrt{2\ln(m/p)}$ satisfies this inequality if $a/(3\sqrt{n}) \leq 1/2$, i.e. $\ln(m/p) \leq 9\,n/8$, and $2a\sqrt{\pi/2} \geq 1$, i.e. $\ln(m/p) \geq 1/(4\pi)$. The corresponding choice $\alpha = \delta/(1 + \sqrt{2\ln(m/p)/n})$ guarantees that all copies are good with probability at least $1 - p$, if $1/(4\pi) \leq \ln(m/p) \leq 9n/8$, which will be true for all reasonable choices of m, n, and p. We conclude from this analysis that the owner can distribute a number of copies m polynomial in n, make the probability of any bad copy exponentially small in n, and still have $\alpha/\delta \to 1$ as $n \to \infty$.

3.3 False Positives

Next we determine how high to set the detection threshold t so that the chance of a false positive is small. Specificially, we want any two distinct watermarks $X^{(i)}$, $X^{(j)}$ to satisfy $s\big(X^{(i)}, X^{(j)}\big) \leq t$ with high probability; and, for any attack document V^* constructed with no knowledge of $X^{(j)}$, we want $s\big(X^*, X^{(j)}\big) \leq t$ with high probability. Assuming a single attack document V^*, this gives a total of $\binom{m+1}{2}$ pairs of watermarks that must be dissimilar.

To analyze the rate of false positives, consider the value of the similarity measure $s(X', X)$, where X is one of the original watermarks, and X' is another watermark constructed with no knowledge of

the value of X. Then X' is *independent* of X. That is, X' can be generated before or after X, without affecting the distribution of $s(X', X)$. This allows us to treat X' as a fixed constant and $s(X', X)$ as a random variable whose distribution is a function of the distribution of X. This simplifies the analysis.

Since x_i for any i is $\alpha N(0,1)$, $x_i' x_i$ is $\alpha N\big(0, (x_i')^2\big)$, which implies that $X' \cdot X$ is

$$\alpha\, N(0, X' \cdot X') \;=\; \|X'\|\; N(0,1).$$

But this means that $s(X', X) = X \cdot X' / \|X'\| \;=\; \alpha N(0,1)$. This observation was made in [5].

The fact that this particular similarity measure is normally distributed for independent watermarks is the main reason for its choice; this makes it especially easy to compute the chance of false positives. In particular, applying Lemma 1 we have

$$\begin{aligned} P(s(X', X) > \alpha y) &= P(\alpha(0,1) > \alpha y) \\ &= P(N(0,1) > y) \\ &\leq \frac{1}{y\sqrt{2\pi}}\, e^{-y^2/2} \quad \text{for } y > 0 \end{aligned}$$

Since this bound is valid for any *fixed* value of X', by the laws of conditional probability it is valid *globally*, independent of X'.

Applying the bound, we find that if we choose a threshold $t = \alpha b$ such that

$$\binom{m+1}{2} \frac{1}{b\sqrt{2\pi}}\, e^{-b^2/2} \;\leq\; p,$$

then the probability of *any* false positive will be at most p. It suffices to choose $b = \sqrt{2\ln(m^2/p)} = 2\sqrt{\ln(m/\sqrt{p}\,)}$ to guarantee a probability of at most p of obtaining a false positive. (This is a crude approximation that ignores the "b" term in the denominator.) We also observe that if we set $b = 2\sqrt{\ln m}$, then the probability of any false positive will tend to zero as m tends to infinity.

Table 2 shows the probability of a false positive for various values of m and various lower bounds on the normalized threshold $b = t/\alpha$. The table illustrates that to make it highly unlikely that any watermark will test positive against any different watermark (including the attacker's implied watermark), the threshold simply needs to be set at a small constant number of standard deviations of the distribution of the owner's watermarks. For example, to guarantee a one-in-ten-thousand chance of any false positive, the threshold needs to be (at least) 6.26 α, if the watermarks are drawn from $N(0, \alpha^2)$ and there are 1000 of them.

We can conclude from this analysis that an appropriate choice of the threshold t will make it highly unlikely that two different watermarks will test positive for similarity. This result is powerful because

m	p	t/α
1	10^{-2}	2.33
1	10^{-3}	3.10
10^3	10^{-4}	6.26
10^6	10^{-5}	8.42
10^8	10^{-6}	9.67

Table 2: Lower bounds on t/α to avoid false positives with high probability.

it implies that even a collusive attack has a small chance of producing a false positive. Specifically, an attacker in possession of k different marked copies will have a small chance of generating a copy that tests positive against any of the other $m - k$ watermarks, if he has no information about the values of these other watermarks.

3.4 False Negatives

We have shown that the probability of a false positive (an implied watermark tests positive but does not in fact match an owner's watermark), is extremely small, if the owner chooses the similarity threshold appropriately. We also need to show, however, that with this same threshold the probability of a false negative (an implied watermark, one of the owner's or one derived from some of the owner's, tests negative) is small. Here and in the next section we show that an adversary who wants to remove the owner's watermarks cannot do so with fewer than $\Omega(\sqrt{n/\ln m})$ differently marked copies, except by preparing a bad copy or getting extremely lucky.

Since there are some conceptual subtleties in the argument, and some technical details, we begin with an overview. Let us review the attack framework. The owner creates the base document V and then creates m marked copies, $W^{(1)}, W^{(2)}, \ldots, W^{(m)}$. An adversary obtains $W^{(1)}, W^{(2)}, \ldots, W^{(k)}$, but not V itself. The adversary prepares an attack document V^*. The owner then tests V^* against $W^{(1)}, W^{(2)}, \ldots, W^{(m)}$ for similarity. By the analysis in Section 3.3, V^* will with high probability test negative against $W^{(j)}$ for $j > k$. We want to show that V^*, if it is a good copy of V, will test positive against at least one of $W^{(1)}, W^{(2)}, \ldots, W^{(k)}$ with high probability. This will allow the owner to identify at least one colluder in the attack.

What leverage does the owner have against successful attack? The answer is that the attacker, even knowing $W^{(1)}, W^{(2)}, \ldots, W^{(k)}$, does not know V exactly. Indeed, the attacker's knowledge of V is captured by a posterior probability distribution, namely the conditional distribution of V given the values of $W^{(1)}, W^{(2)}, \ldots, W^{(k)}$ and the known prior distributions of V and $X^{(1)}, X^{(2)}, \ldots X^{(k)}$. This conditional

distribution is something we can compute and use to estimate the attacker's chance of success.

In the analysis to follow, we treat V^* as a fixed constant independent of (the posterior distribution of) V. We can do this because both the marked copies $W^{(1)}, W^{(2)}, \ldots, W^{(k)}$ and the attack document V^* can be constructed before V is completely specified. Specifically, for the purpose of our analysis the following attack framework is entirely equivalent to the original one. The owner could generate $W^{(1)}, W^{(2)}, \ldots, W^{(k)}$ using the prior distributions of V and $X^{(1)}, X^{(2)}, \cdots, X^{(k)}$. Then the adversary could be given $W^{(1)}, W^{(2)}, \ldots, W^{(k)}$ and prepare V^*. Finally, the owner could generate V according to the posterior distribution of V given $W^{(1)}, W^{(2)}, \ldots, W^{(k)}$.

The idea is similar to the independence argument used in Section 3.3. The attack document can be treated as independent of the original because the original, V, could be constructed *after* the attack document V^*, if V is constructed according to its posterior distribution. This allows us to treat V^* as a constant, again if V is generated according to its posterior distribution. Any probability bounds derived in this setting will hold for any fixed value of V^*, and hence will hold globally, independent of the value of V^*. The difference between the situation here and that in Section 3.3 is that there the attacker has no information about any of the watermarks $X^{(k+1)}, \ldots, X^{(m)}$, except for their distributions, and hence their values are truly independent of the value of the attacker's implied watermark.

The attacker must avoid similarity to any of the k watermarks $X^{(1)}, X^{(2)}, \ldots, X^{(k)}$. Thus the natural value for us to study is the maximum similarity, $\max_{1 \le j \le k} s\big(X^*, X^{(j)}\big)$. We would like to show that this exceeds the similarity threshold t with high probability. But it is (much) easier to estimate the average similarity, $\sum_{j=1}^{k} s\big(X^*, X^{(j)}\big)$, and this is what we shall do. Specifically, we shall show that if k is not too large, and the attack document V^* is good, then the average similarity will exceed the threshold t with high probability. This implies that $s\big(X, X^{(j)}\big) > t$ for at least one value of j $(1 \le j \le k)$, and hence that the attack fails. Our analysis applies for *any* attack whatsoever, deterministic or randomized, neutral (such as standard processing or corruption by noise) or malign.

Now we sketch the analysis, details of which appear in the next section. The first step is to compute the posterior distribution of V. This is straightforward: because the prior distributions of V and $X^{(1)}, X^{(2)}, \ldots, X^{(k)}$ are normal, the posterior distribution of V is normal, with a variance of approximately α^2/k. The second step is to do some simple algebra on the average similiarity, breaking it into three terms and estimating each one separately. We discover that the dominant term is n times the (posterior) variance of V divided by $\|X^*\|$, which is approximately $\alpha^2 \chi_n^2 / (k\|X^*\|)$. If the attack document is good, $\|X^*\| \le \delta\sqrt{n}$. Furthermore, χ_n^2 is near n with high probability. We are thus able to conclude that the average similarity is near $\frac{\alpha^2\sqrt{n}}{\delta\, k}$ with high probability. If the threshold t is chosen equal to $c\,\alpha\sqrt{\ln m}$ as proposed in Section 3.3, we discover that as long as

$$k \leq \frac{1}{c}\left(\frac{\alpha}{\delta}\right)\sqrt{n/\ln m} \qquad \text{(approximately)}$$

then the average similarity will exceed the threshold with high probability. Here c depends on the desired bound p on the probability of a false positive, but only linearly in $\sqrt{\ln(1/p)}$, by the analysis in Section 3.3. By the analysis in Section 3.2, if e.g. $\alpha \leq \delta/2$, all copies will be good with high probability. We thus obtain our main theorem:

Theorem 1 *With appropriate choices of α and t by the owner, an attacker needs $\Omega\left(\sqrt{n/\ln m}\right)$ differently marked copies of a document to have a non-negligible chance of creating a good but unmarked copy.*

Corollary 1 *With appropriate choices of α and t, the owner can distribute $2^{O(n/k^2)}$ differently marked copies secure against k-way collusional attacks.*

In the next section, we proceed with the details of the analysis of false negatives, finishing the proof of Theorem 1.

4 Details of the False-Negative Analysis

As outlined in the previous section, first we compute the posterior distribution of V, then we break the average similarity into three terms and estimate them separately, and finally we combine our estimates to give the final result.

4.1 The Posterior Distribution of V

To compute the posterior distribution of $V = (v_1, v_2, \ldots, v_n)$, we consider the distribution of each individual component v_i given $w_i^{(1)}, w_i^{(2)}, \ldots, w_i^{(k)}$. To simplify notation in what follows, we shall deonote the vectors $\left(w_i^{(1)}, w_i^{(2)}, \ldots, w_i^{(k)}\right)$ and $\left(x_i^{(1)}, x_i^{(2)}, \cdots x_i^{(k)}\right)$ by W_i and X_i, respectively. The posterior probability density function $P\big(v_i\big|W_i\big)$ of v_i given W_i is, by the definition of conditional probability,

$$P\big(v_i\big|W_i\big) \;=\; P\big(v_i, W_i\big)/P(W_i),$$

where $P\big(v, W_i\big)$ is the joint probability distribution of v_i and W_i.

For any fixed W_i, the denominator in this expression is just a normalizing constant. In computing $P\big(v_i\big|W_i\big)$, we can ignore this and other constant factors; once we know $P\big(v_i\big|W_i\big)$ to within a constant factor, we can obtain the correct normalizing constant since $P\big(v_i\big|W_i\big)$ is a probability density function and

hence must integrate to one. For this observation we thank Steve Omohundro (private communication, 1996). In what follows, we denote "proportional to" by "$\propto$".

It is inconvenient to work directly with $P(v_i, W_i)$. Instead, we work with $P'(v_i, X_i)$, the joint probability distribtion of v_i and X_i. The following simple relationship holds between these two distributions:

$$P(v_i, W_i) \;=\; P'(v_i, W_i - v_i) \;=\; P'(v_i, X_i)\,.$$

Since v_i is $N(0,1)$, each $x_i^{(j)}$ is $\alpha N(0,1)$, and v_i and all the $x_i^{(j)}$ are independent, P' is just a product of normal distributions:

$$P'\big(v_i, X_i\big) \;\propto\; \exp\big(-v_i^2/2\big) \prod_{j=1}^{k} \exp\big(-\big(x_i^{(j)}\big)^2/\big(2\alpha^2\big)\big)$$

Let $\mu_i \;=\; \frac{1}{k+\alpha^2} \sum_{j=1}^{k} w_i^{(j)}$. Then we have

$$\begin{aligned} P\big(v_i, W_i\big) \;&\propto \exp\big(-v_i^2/2\big) \prod_{j=1}^{k} \exp\big(-\big(w_i^{(j)} - v_i\big)^2/\big(2\alpha^2\big)\big) \\ &= \exp\big(-\big(\alpha^2 v_i^2 + \sum_{j=1}^{k} \big(w_i^{(j)} - v_i\big)^2\big)/\big(2\alpha^2\big)\big) \\ &\propto \exp\big(-\big((k+\alpha^2)v_i^2 - 2v_i \sum_{j=1}^{k} w_i^{(j)}\big)/\big(2\alpha^2\big)\big) \\ &= \exp\big(-\big(v_i^2 - 2v_i\mu_i\big)\big(k+\alpha^2\big)\;/\big(2\,\alpha^2\big)\big) \end{aligned}$$

Completing the square, we find that

$$P\big(v_i, W_i\big) \propto \exp\big(-\big(v_i - \mu_i\big)^2\big(k+\alpha^2\big)/\big(2\alpha^2\big)\big)$$

Thus $P\big(v_i, W_i\big)$, and hence $P\big(v_i \big| W_i\big)$, is proportional to a normal density function with mean μ_i and variance $\frac{\alpha^2}{k+\alpha^2}$. Since $P(v_i|W_i)$ is a probability density function, it must (for any fixed W_i) be exactly the normal density function with mean μ_i and variance $\frac{\alpha^2}{k+\alpha^2}$. Summarizing, we have proved the following lemma:

Lemma 4 *For any $i, 1 \le i \le n$, given any fixed $w_i^{(1)}, w_i^{(2)}, \ldots, w_i^{(k)}$, the posterior probability distribution of v_i is $N\big(\mu_i,\, \alpha^2/\big(k+\alpha^2\big)\big) = \mu_i + \big(\alpha/\sqrt{k+\alpha^2}\big)\, N(0,1)$.*

4.2 The Average Similarity

Now that we know the posterior distribution of V, we can use this to estimate the average similarity of the attacker's implied watermark to the k watermarks in the copies available to him. Then we show that this estimate exceeds the detection threshold. We assume in what follows that V^* is good; if not, the attack fails. The average similarity $\overline{s}$ is

$$\begin{aligned}\overline{s} = \tfrac{1}{k}\sum_{j=1}^{k} s\big(X^*, X^{(j)}\big) &= \tfrac{1}{k}\sum_{j=1}^{k} X^* \cdot X^{(j)}/\|X^*\| \\ &= \big(\tfrac{1}{k}\sum_{j=1}^{k} W^{(j)} - V\big)\cdot X^*/\|X^*\|.\end{aligned}$$

Let $\overline{W} = \frac{1}{k}\sum_{j=1}^{k} W^{(j)}$ and $M = \big(\mu_1,\ \mu_2, \ldots, \mu_n\big)$, where μ_i is the mean of the posterior distribution of v_i as calculated in Section 4.1. Then we can write the average similarity as

$$\big(\overline{W} - V\big)\cdot X^*/\|X^*\|$$

Manipulating the dot product $(\overline{W} - V)\cdot X^*$, we have

$$\begin{aligned}\big(\overline{W} - V\big)\cdot X^* &= \big(\big(M - V\big) + \big(\overline{W} - M\big)\big)\cdot X^* \\ &= \big(M - V\big)\cdot X^* + \big(\overline{W} - M\big)\cdot X^* \\ &= \big(M - V\big)\cdot\big(V^* - V\big) + \big(\overline{W} - M\big)\cdot X^* \\ &= \big(M - V\big)\cdot\big(\big(V^* - M\big) + \big(M - V\big)\big) + \big(\overline{W} - M\big)\cdot X^* \\ &= \big(V - M\big)\cdot\big(V - M\big) - \big(V - M\big)\cdot\big(V^* - M\big) + \big(\overline{W} - M\big)\cdot X^* \\ &= a - b + c\end{aligned}$$

where $a = (V - M)\cdot(V - M)$, $b = (V - M)\cdot(V^* - M)$, and $c = (\overline{W} - M)\cdot X^*$. The average similarity is thus $(a - b + c)/\|X^*\|$.

In the sum $a - b + c$, the "a" term has positive mean and is the only significant one; the "b" term has mean zero and small variance (with respect to the posterior distribution of V), and the factor $\overline{W} - M$ in the "c" term has mean zero and small variance (with respect to the prior distributions of $W^{(1)}, W^{(2)}, \ldots W^{(k)}$). We proceed to estimate a, b, and c, verifying these claims.

To estimate the first term, we use the posterior distribution of V. For any fixed $W^{(1)}, W^{(2)}, \ldots, W^{(k)}$ and V^*, v_i for any particular i is distributed according to $\mu_i + \big(\alpha/\sqrt{k+\alpha^2}\big)N(0,1)$, by Lemma 4. Hence

$$a \ = \ (V - M)\cdot(V - M) \ = \ \big(\alpha^2/(k+\alpha^2)\big)\sum_{i=1}^{n} N(0,1)^2 \ = \ \big(\alpha^2/(k+\alpha^2)\big)\ \chi^2_n$$

To estimate the second term, we also use the posterior distribution on V. For any fixed $W^{(1)}, W^{(2)}, \ldots, W^{(k)}$ and V^*, we have that $V^* - M$ is constant and $V - M$ is a random variable with distribution $(\alpha/\sqrt{k+\alpha^2})N(0,1)$ in each component. Thus

$$\begin{aligned} b \ = \ (V - M)\cdot(V^* - M) \quad &= \ (\alpha/\sqrt{k+\alpha^2})\sum_{i=1}^{n} N(0, (v_i^* - \mu_i)^2) \\ &= \ (\alpha/\sqrt{k+\alpha^2})N(0, \|V^* - M\|^2) \\ &= \ \tfrac{\alpha\|V^* - M\|}{\sqrt{k+\alpha^2}}N(0,1)\end{aligned}$$

Thus b is normally distributed with mean zero. To estimate the variance of b, observe that

$$\begin{aligned}\|V^* - M\| &= \|(V^* - V) + (V - M)\| \leq \|V^* - V\| + \|V - M\| \\ &= \|X^*\| + \|V - M\| \leq \delta\sqrt{n} + \|V - M\|\end{aligned}$$

since V^* is good. Also, $\|V - M\| = \sqrt{a} = (\alpha/\sqrt{k+\alpha^2})\ \chi_n$. Thus we can write b as

$$b = \left(\frac{\alpha\delta\sqrt{n}}{\sqrt{k+\alpha^2}} + \frac{\alpha^2}{k+\alpha^2}\ \chi_n\right) N(0,1)$$

In this expression the "χ_n" and "$N(0,1)$" do not denote independent distributions, but we can (later) bound $|b|$ by using tail probability estimates for χ_n and $N(0,1)$, and noting that the probability of the union of two unlikely events is at most the sum of their probabilities.

To estimate the third term, we use the prior distributions of V and $X^{(1)}, X^{(2)}, \ldots, X^{(k)}$. We have

$$|c| = |(\overline{W} - M) \cdot X^*| \leq \|\overline{W} - M\| \cdot \|X^*\| \text{ by Cauchy's inequality (see [1]).}$$

But

$$\begin{aligned}\|\overline{W} - M\|^2 &= (\overline{W} - M) \cdot (\overline{W} - M) \\ &= \sum_{i=1}^{n} \left(\left(\tfrac{1}{k} - \tfrac{1}{k+\alpha^2}\right) \sum_{j=1}^{k} w_i^{(j)}\right)^2 \\ &= \sum_{i=1}^{n} \left(\tfrac{\alpha^2}{k+\alpha^2} \sum_{j=1}^{k} \left(w_i^{(j)}/k\right)\right)^2 \\ &= \sum_{i=1}^{n} \left(\tfrac{\alpha^2}{k+\alpha^2}\left(v_i + \sum_{j=1}^{k} \left(x_i^{(j)}/k\right)\right)\right)^2 \\ &= \sum_{i=1}^{n} \left(\tfrac{\alpha^2}{k+\alpha^2}\left(N\big(0,1\big) + \tfrac{1}{k} \sum_{j=1}^{k} N\big(0,\alpha^2\big)\right)\right)^2 \\ &= \sum_{i=1}^{n} \left(\tfrac{\alpha^2}{k+\alpha^2}\left(N\big(0, 1+\alpha^2/k\big)\right)^2 \\ &= \tfrac{\alpha^4}{k(k+\alpha^2)} \sum_{i=1}^{n} N\big(0,1\big)^2 \\ &= \tfrac{\alpha^4}{k(k+\alpha^2)}\ \chi_n^2.\end{aligned}$$

From this we conclude that $\|\overline{W} - M\| = \big(\alpha^2/\sqrt{k(k+\alpha^2)}\big)\chi_n$. The inequality $|c| \leq \|\overline{W} - M\| \cdot \|X^*\|$ then gives $|c| \leq \frac{\alpha^2}{k}\ \chi_n \|X^*\| \leq \alpha^2\ \delta\sqrt{n}\ \chi_n/k$ since V^* is a good copy.

Combining the estimates for all three terms, dividing by $\|X^*\|$, and using the goodness constraint $\|X^*\| \leq \delta\sqrt{n}$, we obtain the following lower bound on the average similarity, if it is positive:

$$\begin{aligned}\overline{s} &\geq (a - b - c)/\|X^*\| \geq (a - b - c)/(\delta\sqrt{n}) \\ &\geq \tfrac{\alpha^2}{\delta(k+\alpha^2)\sqrt{n}}\ \chi_n^2 - \big(\tfrac{\alpha}{\sqrt{k+\alpha^2}} + \tfrac{\alpha^2}{\delta(k+\alpha^2)\sqrt{n}}\ \chi_n\big)N(0,1) - \tfrac{\alpha^2}{k}\ \chi_n\end{aligned}$$

Again note that in this expression "χ_n^2", "χ_n", "$N(0,1)$", and "χ_n" do not denote independent distributions, but we can bound $\overline{s}$ with high probability by using tail probability estimates for the respective distributions and summing the probabilities of unlikely events.

4.3 Completion of the Analysis

For purposes of the analysis we shall treat α and δ as fixed constants. The first term in the lower bound on $\overline{s}$ can be bounded below with high probability by

$$\frac{\alpha^2\sqrt{n}}{\delta k}\Big(1-\frac{\alpha^2}{\delta k}\Big) - O\Big(n^{1/6}/k\Big)$$

using the properties of the χ_n^2 distribution. Specifically, by applying Lemma 2, we find that the first term is at least

$$\frac{\alpha^2}{\delta(k+\alpha^2)\sqrt{n}}\ \big(n - c\sqrt{n}\big) \geq \frac{\alpha^2\sqrt{n}}{\delta k}\big(1-\frac{\alpha^2}{\delta k}\big)\ -\ \frac{c\,\alpha^2}{\delta k}$$

except for a probability that is exponentially small in c^2. Choosing $c = n^{1/6}$ gives the claimed inequality except for a probability exponentially small in $n^{1/3}$. Similar arguments can be used to bound the second and third terms. Specifically, Lemma 3 implies that the absolute value of the coefficient of $N(0,1)$ is at most

$$\frac{\alpha}{\sqrt{k}}\ +\ \frac{\alpha^2}{\delta k}\ +\ O(\frac{1}{kn^{2/3}})$$

except for a probability exponentially small in $n^{1/3}$. From Lemma 1 we conclude that the absolute value of the entire second term is $O(n^{1/6}/\sqrt{k})$ except for a probability exponentially small in $n^{1/3}$. Finally, Lemma 3 implies that the absolute value of the third term is at most $\frac{\alpha^2}{k}\sqrt{n} + O\big(n^{1/6}/k\big)$. Combining our estimates, we see that

$$\overline{s}\ \geq\ \frac{\alpha^2}{k}\ \sqrt{n}\ (\frac{1}{\delta} - 1 - \frac{\alpha^2}{\delta^2 k})\ -\ O\,(\frac{n^{1/6}}{\sqrt{k}})$$

except for a probability exponentially small in $n^{1/3}$.

Suppose the owner chooses the similarity threshold t as suggested in Section 3.3, namely $t = 2\alpha\sqrt{\ln(m/\sqrt{p})}$, where p is the desired upper bound on the probability of a false positive. Then $\overline{s} > t$ with high probability if

$$k \leq \frac{\alpha}{2}\left(1/\delta - 1 - O(n^{-1/12})\right)\sqrt{n/\ln(m/\sqrt{p})}\ -\ O(1)$$

If the average similiarity is greater than the threshold, then the attack document will test positive against at least one of the watermarks in the copies accessible to the attacker, and the attack will fail. This proves

Theorem 1. As noted in Section 3.3, the owner can set $t = 2\alpha\sqrt{\ln\, m}$ and have the probability of a false positive go to zero as $m \to \infty$. This means that for $k \le c\sqrt{n/\ln m}$, where c is an appropriate constant, a k-way collusive attack will fail with high probability. Equivalently, if $\ln m \le c^2\, n/k^2$, a k-way collusive attack will fail with high probability. This gives Corollary 1.

4.4 Comments About the Analysis

There are a few details missing in the analysis just completed. To make the argument entirely rigorous, we should have computed separate bounds with associated out-of-bound probabilities for each term in the sum for $\overline{s}$, and then added the bounds and added the probabilities. These details are straightforward, and we leave them as an exercise.

Let us make a few observations about our upper bound on k. The chance of a false positive affects the leading-order term in our estimate (but only logarithmically). The chance of a false negative, on the other hand, affects only lower-order terms. Thus we can make the chance of a false negative extremely small without affecting the maximum allowable k, at least asymptotically.

One significant restriction in our model is the assumption that V is normally distributed. We can completely eliminate this assumption and, indeed, *any* assumption on the distribution of V, if we limit the class of allowed attacks. Specifically, let us suppose that the attack is *shift-invariant*. By this we mean that V^* is computed in such a way that $X^* = V^* - V$ does not change if all the $X^{(j)}$ remain fixed but V changes. The averaging attack $V^* = \frac{1}{k}\sum_{j=1}^{k} W^{(j)}$ is shift-invariant, as is any attack that selects each component v_i^* of V^* from among $\{w_i^{(j)} | 1 \le j \le k\}$. Such an attack uses no information about the prior distribution of V. This means that for purposes of the analysis we can replace the actual distribution of V with *any* distribution, and such a replacement will not affect the attacker's chance of success. With this in mind, if we redo the analysis assuming an $N(0, \gamma^2)$ distribution on the components of V (instead of $N(0,1)$), we obtain the following bound on k:

$$k \le \frac{\alpha}{2}\left(\frac{1}{\delta} - \frac{1}{\gamma} - O(n^{-1/12})\right)\sqrt{n/\ln(m/\sqrt{p})} \;-\; O(1)$$

Letting $\gamma \to \infty$, which we can do because the bound on k is valid for *any* $\gamma \to 0$, the bound on k for shift-invariant attacks becomes

$$k \le \frac{\alpha}{2\delta}\left(1 - O(n^{-1/12})\right)\sqrt{n/\ln(m/\sqrt{p})} \;-\; O(1)$$

and this bound requires no assumption about the distribution of V.

We conclude this section with a table showing conservative bounds on the allowed value of k for selected values of n, m, δ, and overall failure probability p. We have divided p equally among the three

ways of failing, namely, probability $p/3$ of a bad distributed copy, $p/3$ of a false positive, and $p/3$ of a false negative. We computed the table values using standard tail estimates for $N(0,1)$, χ_n, and χ_n^2, applied to the estimates in Sections 3.2 and 3.3 for choosing α and t. For general collusion attacks the table values are based on the following inequality, which governs false negatives:

$$\frac{t}{\alpha} \;\leq\; \frac{\alpha}{\delta(k+\alpha^2)\sqrt{n}}\;\chi_n^2 - \left(\frac{1}{\sqrt{k+\alpha^2}} + \frac{\alpha}{\delta(k+\alpha^2)\sqrt{n}}\;\chi_n\right)\; N(0,1) - \frac{\alpha}{k}\;\chi_n$$

For shift-invariant attacks, if we redo the analysis as suggested above, using an $N(0,\gamma^2)$ distribution on the components of V and letting $\gamma \to \infty$, we obtain the following inequality governing false negatives, which was used to compute the k_s column in Table 3.

$$\frac{t}{\alpha} \leq \frac{\alpha}{\delta k\sqrt{n}}\;\chi_n^2 - \left(\frac{1}{\sqrt{k}} + \frac{\alpha}{\delta k\sqrt{n}}\;\chi_n\right) N(0,1)$$

n	m	δ	p	k_g	k_s
1000	10	0.1	10^{-1}	4	5
1000	10^4	0.1	10^{-2}	2	2
3000	10^3	0.07	10^{-2}	5	6
3000	10^5	0.07	10^{-3}	4	4
10000	10^2	0.03	10^{-2}	15	15
10000	10^5	0.03	10^{-3}	10	10
10000	10^6	0.03	10^{-4}	8	8

Table 3: Number of misappropriated copies protected against collusive attacks. For general attacks, the number is k_g; for shift-invariant attacks, the number is k_s.

The values in Table 3 show that, even under our conservative assumptions, watermarks based on Gaussian noise offer non-negligible protection against collusive attacks, for realistic parameter values. Much experimentation is needed to determine accurate parameters and collusion resistance in practice.

5 Summary, Remarks, and Extensions

In this concluding section, we first summarize our results. We then make some remarks about the assumptions of our model and about possible extensions of our work.

We analyzed the collusion resistance of a watermarking method based on adding Gaussian random noise. There were three main parts of our analysis. First, we showed that if α, the watermark strength, is chosen to be at most $\delta/\left(1+\sqrt{2\,\ln(m/p)/n}\right)$, then all distributed copies will be good with probability

at least p. Second, we showed that if the similarity threshold t is set equal to $2\alpha\sqrt{\ln(m/\sqrt{p})}$, then the probability of any false positive between the owner's watermarks or the attacker's implied watermark and any independent owner watermark will be at most p. Observe that this false positive result does not depend on n, and that neither of these results depends in any way on collusion or the lack of it.

Third, we showed that an attacker in possession of k different marked copies has a very small chance of producing an unmarked good copy if $k \leq \frac{\alpha}{2}\left(\frac{1}{\delta} - 1 - O(n^{-1/12})\right)\sqrt{n/\ln(m/\sqrt{p})} - O(1)$, if the owner has chosen t as specified above and p is the probability bound for false positives. The bound on k relies on the assumption that the original document is normally distributed. We sketched a proof, however, that for the restricted class of shift-invariant attacks, an attacker needs at least $\frac{\alpha}{2\delta}(1 - O(n^{-1/12}))\sqrt{n/\ln(m/\sqrt{p})} - O(1)$ different marked copies to have a non-negligible chance of producing a good unmarked copy. This last result makes *no* assumption on the distribution of V, the original document.

It is natural to ask whether other watermark distributions than Gaussian would work as well. In particular, what about a uniform distribution over a finite interval? The answer is that uniformly distributed watermarks are subject to successful collusive attacks with as few as six copies, independent of n. One attack on uniformly distributed watermarks is as follows. For each component i, compute $\max_i = \max\{w_i^{(j)}|j\}$, $\min_i = \min\{w_i^{(j)}|j\}$, $\text{midpt}_i = (\max_i + \min_i)/2$, and $\text{ave}_i = \text{ave}\{w_i^{(j)}|j\}$. If $\text{midpt}_i < \text{ave}_i$, set $v_i^* = \min_i$. Otherwise, set $v_i^* = \max_i$. The intuition behind this attack is to set each component v_i^* of V^* to correlate negatively with as many of the $w_i^{(j)}$ as possible. The weakness of a uniform distribution is that it has sharp boundaries in each component; with multiple copies, an adversary can get a very good estimate of the original document. This particular attack produces a negative similarity measure against all watermarks for as few as five copies, but risks goodness. Scaling back the power of the attack by moving v_i^* toward midpt_i preserves goodness while still erasing the watermarks with as few as six copies. Thus uniformly distributed watermarks are much less secure against collusive attacks than normally distributed ones.

A more general question is weather *any* watermarking scheme can offer better collusion resistance than the one we have analyzed. The answer is no. Specifically, F. Ergun, J. Kilian, and R. Kumar (unpublished notes, 1997) have studied the power against any watermarking method of an attack that averages k different copies and adds an appropriate amount of Gaussian noise. (For $k = 2$, the Karakos-Papamarcou attack [9] is a generalization of this one.) Their analysis, as refined and tightened by S. Mitchell, R. Tarjan, and F. Zane (unpublished notes, 1997), shows that this attack will produce a unwatermarked but good copy with high probability, if $k = \Theta(\sqrt{n/\ln m})$.

Let us review some of the assumptions of our model for analyzing Gaussian watermarks. A major assumption we have made is that the components of the original document V are independently distributed according to $N(0, 1)$. Although we have shown that for a restricted class of attacks (shift-

invariant attacks) no such distributional assumption is necessary, for general attacks we do need such an assumption.

Concerning the assumption that the document components are independent, Steve Mitchell (unpublished notes, 1997) claims that our analysis can be extended to the case of non-independent components if the covariance of their distributions is known or can be estimated.

Concerning the assumption of an $N(0,1)$ distribution for each component, we observe that the definition of "components" is up to the owner, and in any given setting it may be possible to transform the document representation so that it has components that are distributed approximately according to $N(0,1)$, even if this is not true of the original representation. For example, if one can sample the document distribution, one can estimate the mean and variance, and use the estimates to transform the representation so that each component has an approximately zero mean and approximately unit variance. We also conjecture that our analysis can be extended to other distributions besides the normal one.

A second key assumption in our model is that the perceptual threshold can be measured, at least approximately, by root-mean-square error, $\|V^* - V\|$. The truth of this assumption, for various representations of various kinds of media, needs to be tested by experiment. We have observed that, for still images, small amounts of cropping and rescaling change the DCT components enough to erase watermarks, without affecting the appearance of the image significantly. This issue needs much more experimental study, if our theoretical results are to be translated into a robust, practical watermarking scheme.

We have also assumed that the watermark detector is under the control of the owner, so that the attacker has only one chance at an attack. If, on the other hand, the watermark detector is public to the extent that an attacker can query it for the presence or absence of a mark, then an attack by repeated trials is possible. To guard against such an attack, the owner must raise the similarity threshold to reduce the chance of an attacker successfully creating a false positive by multiple trials. This by itself reduces the protection against collusion. In addition, the attacker has an increased chance of creating a false negative by multiple trials, and the false negative analysis must be modified to take this into account. The needed modifications to the analysis are straightforward, but they are beyond the scope of this paper, and we leave them for future work. The bottom line is that, by setting the threshold appropriately, the owner guards against a polynomial number of attacks while yielding a constant factor in the amount of collusion protection (the number of different marked copies the attacker needs to have a non-negligible chance of success).

A more powerful and sophisticated type of repetition attack is proposed by Cox and Linnartz [6]. In this attack a marked document is degraded until the mark is barely detectable, and individual components are modified to determine the direction of change that eliminates watermark detectability. This

component-by-component information is then used to erase the fingerprint.

A final issue concerns numerical errors. We have assumed continuous distributions for the original document and the watermarks. In practice, of course, all this information must be represented in finite precision. To determine the required amount of precision requires a calculation that is beyond the scope of this paper, and which we leave to future work. A rough estimate suggests, however, that $O(\log n)$ bits suffice, as long as m, the number of copies distributed, is polynomial in n.

Acknowledgement

We thank Steve Mitchell for many insightful observations and fruitful discussions.

References

[1] M. Abramowitz and I. Stegan, eds., *Handbook of Mathematical Functions*, Dover Publications, Inc., New York, 1965, p. 11.

[2] G. R. Blakely, C. Meadows, and G. B. Purdy, "Fingerprinting long forgiving messages," *Crypto '85*, Springer-Verlag, Berlin, 1985, p. 180.

[3] D. Boneh and J. Shaw, "Collusion-secure fingerprinting for digital data," *Crypto '95, Lecture Notes in Computer Science* **963**, Springer-Verlag, Berlin, 1995, pp. 452–465.

[4] B. Chor, A. Fiat, and M. Naor, "Tracing traitors," *Crypto '94, Lecture Notes in Computer Science* **839**, Springer-Verlag, Berlin, 1994, pp. 257–270.

[5] I. Cox, J. Kilian, T. Leighton, and T. Shamoon, "Secure Spread Spectrum Watermarking for Multimedia", *IEEE Trans. on Image Processing* **6** (1997), to appear.

[6] I. Cox and J.-P. Linnartz, Public watermarks and resistance to tampering, *Proc. IEEE Conf. on Image Processing*, 1997, CD-ROM.

[7] W. Feller, *An Introduction to Probability and Statistics*, Vol. 1, John Wiley & Sons, Inc., New York, 1994, p. 175.

[8] N. L. Johnson, S. Kolta, and N. Balakrishnan, *Continous Univariate Distributions*, Vol. 1, 2nd ed., John Wiley & Sons, Inc., New York, 1994.

[9] D. Karakos and A. Papamarcou, "Some results on the information capacity of authenication channels," Dept. of Electrical Engineering, University of Maryland, College Park, MD, manuscript, 1997.

[10] B. Pfitzmann and M. Waidner, "Asymmetric fingerprinting for larger collusions," 4^{th} *ACM Conf. on Computer and Communications Security*, 1997, to appear.

[11] B. Pfitzmann and M. Waidner, "Anonymous fingerprinting," Research Report RZ 2881, IBM Research, 1996.

[12] B. Pfitzmann and M. Schunter, "Asymmetric fingerprinting (extended abstract)," *EUROCRYPT '96, Lecture Notes in Computer Science* 1070, Springer-Verlag, Berlin, 1996, pp. 84–95.

[13] D. L. Wallace, "Bounds on normal approximations to Student's and the chi-square distribution," *Annals of Math. Stat.* **30** (1959), pp. 1121-1130.

Appendix: Bounds on the Tail Probabilities of χ_n

Lemma 2

$$P(\chi_n < \sqrt{n} - y) \leq \frac{1}{2y\sqrt{\pi}}\ e^{-y^2} \text{ for } 2 \leq y < \sqrt{n},\ n \geq 12 \text{ even}$$

Proof: The probability density function of χ_n for $x \geq 0$ is

$$P_{\chi_n}(x) \ = \ \frac{e^{-x^2/2}x^{n-1}}{2^{n/2-1}\Gamma(n/2)}$$

where Γ is the gamma function ([8], p. 417). Since n is even by assumption,

$$\begin{aligned}\Gamma(n/2) = (n/2-1)! \quad &= (2/n)(n/2)! \\ &\geq (2/n)\sqrt{2\pi(n/2)}(n/(2e))^{n/2} \\ &\qquad\qquad \text{by Stirling's approximation} \\ &= (2\sqrt{\pi}/\sqrt{n})\,(n/(2e))^{n/2}\end{aligned}$$

Thus $P_{\chi_n}(x) \leq \frac{1}{\sqrt{\pi}}\ e^{(-x^2+n)/2}(x/\sqrt{n})^{n-1}$.

We claim that for $x \leq \sqrt{n} - 2$ and $n \geq 12$, this is no more than the probability density function of $N(\sqrt{n},\ 1/2)$. The Lemma follows from this claim and from Lemma 1 (Section 3.1), using the substitution $y = \sqrt{n} - x$ and the equality $N(\sqrt{n},\ 1/2) = \sqrt{n} + (1/\sqrt{2})N(0,1)$.

The probability density function of $N(\sqrt{n},\ 1/2)$ is

$$P_{N(\sqrt{n},\ 1/2)}(x) = \frac{1}{\sqrt{\pi}}\ e^{-(x-\sqrt{n})^2}$$

Thus to show that $P_{\chi_n}(x) \leq P_{N(\sqrt{n},1/2)}(x)$ it suffices to prove

$$x^2/2 - n/2 - (n-1)\ln\,(x/\sqrt{n}) \geq (x - \sqrt{n})^2$$

This is equivalent to

$$\begin{aligned}-(n-1)\ln\,(x/\sqrt{n}) \quad &\geq \ x^2/2 - 2x\sqrt{n} + 3n/2 \\ &= \ (x-\sqrt{n})^2/2 - x\sqrt{n} + n\end{aligned}$$

Making the substitution $y = \sqrt{n} - x$ for $0 \leq x < \sqrt{n}$, the desired inequality becomes

$$-(n-1)\ln(1 - y/\sqrt{n}) \geq y^2/2 + y\sqrt{n}$$

Since $\ln\,(1 - y/\sqrt{n}) > -y/\sqrt{n} - y^2/(2n) - y^3/(3n^{3/2})$, the desired inequality will follow from

$$(n-1)(y\sqrt{n} + y^2/(2n) + y^3/(3n^{3/2})) \geq y^2/2 + y\sqrt{n}$$

This is equivalent to

$$y^3/(3\sqrt{n}) - y/\sqrt{n} - y^2/(2n) - y^3/(3n^{3/2}) \geq 0$$

which is in turn equivalent to

$$y^2(1-1/n) - 3y/(2\sqrt{n}) - 3 \geq 0$$

Simple algebra verifies this inequality for $y \geq 2, n \geq 12$. □

QED

Lemma 3

$$P(\chi_n > \sqrt{n} + y) \leq \frac{1}{2y\sqrt{\pi(1-y/(3\sqrt{n}))}}\ e^{-y^2(1-y/(3\sqrt{n}))}$$
$$\textit{for } 0 \leq y < 3\sqrt{n},\ n \geq 1$$

Proof: Wallace[13] showed that for $x \geq 0, n \geq 1$,

$$P(\chi_n^2 > x) \leq P(N(0,1) > w(x))$$

where $w(x) = (x - n - n\ln(x/n))^{1/2}$. Letting $y = \sqrt{x} - \sqrt{n}$, i.e. $x = (\sqrt{n} + y)^2$, the lemma will follow from Lemma 1 and the inequality $w(x) \geq \sqrt{2}y(1-y/(3\sqrt{n}))^{1/2}$, i.e. $w(x)^2 \geq 2y^2(1-y/(3\sqrt{n}))$. But

$$\begin{aligned} w(x)^2 &= 2\sqrt{n}\ y + y^2 - 2n\ln(1+y/\sqrt{n}) \\ &\geq 2\sqrt{n}\ y + y^2 - 2n(y\sqrt{n} - y^2/(2n) + y^3/(3n^{3/2})) \\ &\qquad \text{since } \ln(1+y/\sqrt{n}) \leq y/\sqrt{n} - y^2/(2n) + y^3/(3n^{3/2}). \end{aligned}$$

Thus $w(x)^2 \geq 2y^2 - 2y^3/(3\sqrt{n}) \ = \ 2y^2(1-y/(3\sqrt{n}))$. □

QED