A Note on the Limits of Collusion-Resistant Watermarks

Funda Ergun*, Joe Kilian**, and Ravi Kumar***

Abstract. In one proposed use of digital watermarks, the owner of a document D sells slightly different documents, $D^1, D^2, \ldots$ to each buyer; if a buyer posts his/her document D^i to the web, the owner can identify the source of the leak. More general attacks are however possible in which k buyers create some composite document D^*; the goal of the owner is to identify at least one of the conspirators.
We show, for a reasonable model of digital watermarks, fundamental limits on their efficacy against collusive attacks. In particular, if the effective document length is n, then at most $O(\sqrt{n/\ln n})$ adversaries can defeat any watermarking scheme.
Our attack is, in the theoretical model, oblivious to the watermarking scheme being used; in practice, it uses very little information about the watermarking scheme. Thus, using a proprietary system seems to give only a very weak defense.

Keywords: watermarking, intellectual property protection, collusion resistance

1 Introduction

1.1 The General Problem

The very properties that have made digital media so attractive present difficult, not clearly surmountable, security problems. The ability to cheaply copy and transmit perfect copies of text, audio, and video opens up new avenues both for electronic commerce and for electronic piracy. The advent of ubiquitous high speed networks and network caching algorithms further amplifies this problem. Anyone will have the capability to cheaply distribute any movie, song, book, or picture (which we will generically call a *document*) in their possession to anyone else on the planet. The challenge is to maintain intellectual property in this environment.

* Bell Laboratories, 700 Mountain Avenue, Murray Hill, NJ 07974. email: `fergun@research.bell-labs.com`.
** NEC Research Institute, 4 Independence Way, Princeton, NJ. email: `joe@research.nj.nec.com`.
*** IBM Almaden Research Center/K53C San Jose , CA 95120-6099. email: `ravi@almaden.ibm.com`.

There are a number of approaches to this problem; we concentrate on methods related to digital watermarking, also known as digital fingerprinting. In one general approach, the media to be distributed is altered so that it contains a hidden "do not copy" signal (the "watermark"). Most or all of the hardware for viewing, copying, or transmitting the media look for this signal and prevent illicit use. Two major problems with this approach are preventing the construction of illicit hardware that ignores the safeguards and preventing the erasure of the hidden signal. The latter problem is aggravated by the fact that one has to effectively distribute oracles (e.g., copying machines) that give feedback as to whether the signal can still be detected.

We know of no such watermarking scheme that has survived a serious attack. Indeed, with one commercially distributed scheme for watermarking images, the mark was so delicate that owners would accidentally destroy it themselves (such as by resizing the image prior to selling it).

A less ambitious use of watermarking is to identify pirates after the fact. That is, nothing prevents a pirate from anonymously posting ones intellectual property to the web, but one should be able to identify who did so. The general approach is to, given a document D, perturb it in an unobtrusive manner to generate documents $D^1, D^2, \ldots$, giving each buyer a distinct copy. If the i-th buyer posts D^i to the web, the document owner can identify him/her as the pirate.

Innumerable schemes have been proposed for both uses; we refer to [3] for a discussion of many of these schemes.

1.2 Modeling Collusion Attacks

Of course, a pirate may not be so cooperative as to simply post its document unchanged. It may attempt to alter it or, perhaps in concert with others, combine several documents to produce a document that cannot be linked with any of the "original" marked documents.

The first theoretical modeling and treatment of collusion of attacks was given by Boneh and Shaw [1]. We instead use a model suggested by Cox *et. al.* [3]. We refer to [3, 5] for a more extensive introduction to this model, described briefly below.

First, we model a document D as a sequence of real numbers $\langle D_1, \ldots, D_n \rangle$. This should not be thought of as a literal and complete description of the document, but as an indication of the values of "critical values" that might be changed by the watermarking process. For example, they may be coefficients in a wavelet decomposition of an image or audio stream. In [3], it is posited that these should be orthogonal, independent attributes; [5] primarily analyzes the case where they are uniformly distributed. We do not make any such assumptions.

We model collusion attacks as follows. First, we model a watermarking scheme as a pair of functions Mark and Detect. $\mathsf{Mark}(D, m)$ defines a distribution on sequences $D^1, \ldots, D^m$, where m is the total number of documents produced; Mark may be viewed as a randomized procedure for producing $D^1, \ldots, D^m$.

A t-collusion attacker is modeled by a probabilistic polynomial time procedure Attack, and a distribution on distinct $i_1, \dots, i_t$. In all of our discussions we assume that $i_1, \dots, i_t$ is chosen uniformly from all t-element subsets. On input

$$(i_1, \dots, i_t, D^{i_1}, \dots, D^{i_t}),$$

Attack generates a distribution on its output, D^*.

On input $D, D^1, \dots, D^m, D^*$, Detect returns a distribution on its output $i \in [1, m] \cup \emptyset$. Returning an index indicates an accusation; returning $\emptyset$ indicates that no one has been caught. For notational simplicity, we omit the $D, D^1, \dots, D^m$ arguments when they are fixed or clear, writing simply Detect(D^*).

We now specify our requirements for Mark, Detect, and Attack. First, we consider the fidelity of the marked documents and the attacked documents. We require that $d(D^i, D) \leq \Delta/2$, where d denotes the Euclidean metric. We require a successful attack to achieve $d(D^*, D) \leq \Delta'/2$; the closer Δ' is to Δ, the better the attack. Intuitively, $\Delta/2$ indicates the degree to which the watermarking algorithm is willing to distort D, and $\Delta'/2$ indicates the amount of distortion past which the document is no longer worth stealing or protecting.

(We use $\Delta/2$ instead of Δ to simplify the analysis. By the triangle inequality, our condition enforces that $d(D^i, D^j) \leq \Delta$; this turns out to be the more natural condition to consider.)

Next, we consider the efficacy of the detection algorithm. Detect succeeds if it returns an $i \in \{i_1, \dots, i_t\}$. Detect can fail in two ways: (i) The owner can fail to identify any of the pirates by returning $\emptyset$ (a false negative), or (ii) the owner can falsely conclude that an innocent person is a pirate (a false positive). A false negative is unfortunate; a false positive is catastrophic. If one fails to catch a pirate 90% of the time, the 10% may deter some (but not all), but if one misidentifies an innocent person 1% of the time one may not be able to ever credibly accuse anyone of piracy.

1.3 Our Result

We show a generic attack procedure Attack that defeats all watermarking schemes for the above model. It is oblivious to the Mark and Detect schemes. It has the following properties:

1. The attack uses $t = \frac{\alpha}{\epsilon}\sqrt{n/\ln n}$ documents, where α is a parameter (the larger the parameter, the more effective the attack), and ϵ controls the fidelity of the attack (we ignore integer rounding issues).
2. With high probability, it produces an attack document D^* such that $d(D^*, D) \leq (\Delta/2)(1 + 2\epsilon^2 + o(1))$.
3. Suppose Detect succeeds with probability above, say $2/\sqrt{\ln n}$, then it must incur a false positive probability of $\Omega(n^{-c})$, for some c, where c depends on α. More general tradeoffs are implied by our analysis.

1.4 Related Work

Boneh and Shaw introduced the first formal model of collusion resistance. They consider a more abstract model in which one may insert a sequence of marks into a document; each mark has a value associated with it (most usually boolean) . They assume that if for all the documents available to the attacker, the i-th mark has the same value, then the attacker cannot remove this mark. If, however, two of the documents disagree on the value of the i-th mark, the attacker can change its value as it sees fit. In this model, they show upper and lower bounds for the collusion resistance as a function of the number of marks. Further improvements and additions to their basic scheme appear in [9, 10, 8].

It is impossible to directly compare this model and its models with that of Cox *et. al.* The model of [3] gives a more low-level model for watermarking. One pleasing aspect of our result is that it essentially matches to within a constant factor some lower bounds on collusion resistance proven by [5]. For the case where $m = n^{O(1)}$, they show that one can achieve collusion resistance of $\Omega(\sqrt{n/\ln n})$, given a very specialized assumption about the distribution of D (or given a very restricted class of attacks). Our bounds show that this is essentially the best one can hope for, regardless of the assumptions one makes about the distribution of the documents. In contrast, there is a substantial gap in the upper and lower bounds known for the Boneh-Shaw model.

Along a similar vein, Chor, Fiat, and Naor [2] introduce *traitor tracing* schemes. In their scenario, a large amount of data is broadcast, or made publicly available (say by DVD disks) in encrypted form; keys allowing the data to be decrypted are individually sold. Subsequent work in this area includes [6, 7]; a further twist on key protection is given in [4]. In one respect, these models have a similar flavor to the scenario we consider, in that one wishes to identify those who publish or resell their keys. This work, however, is intended for the regime where the plaintext is so large that it is hard to (re)broadcast it. Watermarking hopes to protect much smaller data (hundreds of kilobytes).

1.5 Road Map

In Section 2 we describe our attack. In Section 3 we analyze its efficacy. In Section 4 we present conclusions and open problems.

2 The Attack

Our attack is parameterized by a collusion parameter t and a noise parameter σ. We will analyze the case where $t = (\alpha/\epsilon)\sqrt{n/\ln n}$, α is some (typically constant) parameter, and $\sigma = \epsilon\Delta/(2\sqrt{n})$, where n is the length of the attacked document; i.e., the dimension of D, and ϵ is a (typically small constant) parameter. Let $N(\mu, \sigma^2)$ be the Gaussian (normal) distribution with mean μ and standard deviation σ.

Described in words, the colluding attack is to average the t vectors and perturb with a random Gaussian noise at each component. σ is to be determined later.

$\mathsf{Attack}_{t,\sigma}(i_1, \ldots, i_t, D^{i_1}, \ldots, D^{i_t})$

1. First, compute $\overline{D^*} = \frac{1}{t}\sum_{j=1}^{t} D^{i_j}$, where the sum is performed coordinate-wise. That is, each coordinate of $\overline{D^*}$ is set to be the average of the corresponding values of the sample documents.
2. Let n denote the length of $\overline{D^*}$. Choose $R = \langle r_1, \ldots, r_n\rangle$ by choosing r_j independently according to $N(0, \sigma^2)$, for $1 \leq j \leq n$. Compute $D^* = \overline{D^*} + R$.

Observe that in the abstract model, Attack uses no information about Mark, except for σ. We discuss more practical issues in Section 4.

There is a tension in our choice of t and σ. As we will see, the larger the values of t and σ, the more effective the attack. However, we would like to minimize the number t of adversaries (document copies) needed, and increasing σ weakens the fidelity of the attacked copy.

3 Analysis

We analyze the efficacy of Attack as a function of the parameters t and σ. First we analyze the fidelity of the attack, and then we show, for any choice of Detect, a tradeoff between the probability that it generates a false positive and the probability that it generates a false negative.

3.1 The Fidelity of the Attack

For the rest of our discussion, high probability mean with probability $1 - o(1)$ as n grows large.

Lemma 1. *Suppose that* $\sigma = \epsilon\Delta/\sqrt{n}$*. Then with high probability,* $d(D, D^*) \leq (\Delta/2)(1 + 2\epsilon^2 + o(1))$.

Proof. (Sketch) Consider the triangle formed by D, D^*, and $\overline{D^*}$. Let $a = d(D, D^*)$, $b = d(D^*, \overline{D^*})$, and $c = d(D, \overline{D^*})$. Let θ be $\angle DD^*\overline{D^*}$. Then $c^2 = a^2 + b^2 - 2ab\cos\theta$. First, by the convexity of the Euclidean norm, it follows that $a \leq \Delta/2$ (D^* is the centroid of points all within $\Delta/2$ of D). Now, $b^2 = \langle r_1^2, \ldots, r_n^2\rangle$ (where r_i is as in Attack); hence, b^2 is a χ^2 distribution with mean $\sigma^2 n = \epsilon^2\Delta^2$. Using simple bounds on the tail of χ^2 distributions, we have that with high probability, $b^2 \leq (1 + o(1))\epsilon^2\Delta^2$. It remains to bound the magnitude of $\cos\theta$. By the spherical symmetry of the distribution on R, θ has the same distribution as the angle between two random unit rays from the origin. For this case, it is well known that $|\cos\theta|$ is $O(\ln n/\sqrt{n})$ with high probability. Hence, with high probability,

$$c^2 \leq (\Delta/2)^2 + (1 + o(1))\epsilon^2\Delta^2 + O(\epsilon\Delta^2 \ln n/\sqrt{n}).$$

The lemma follows.

3.2 A Tradeoff between Errors

Attack ignores the values of $i_1, \ldots, i_t$. To simplify our notation,we assume without loss of generality that the attacking coalition is $1, \ldots, t$.

Suppose on D^* Detect outputs a valid value of $i \in \{1, \ldots, t\}$ with probability at least ρ, where the probability is taken over the randomness used by Attack and any randomness used by Detect. Assume without loss of generality that Player 1 is the player most often detected. Thus, $\mathsf{Detect}(D, D^1, \ldots, D^m, D^*) = 1$, with probability $\geq \rho/t$. The idea is to produce another document D' with a slightly different colluding set that does not include Player 1. When t is sufficiently large, D^* and D' cannot be reliably distinguished by Detect (or any other distinguisher). Hence Detect will output $i = 1$, yielding a false positive, with an unacceptably large probability.

Consider the output of Attack on $D^2, \ldots, D^{t+1}$. We define $\overline{D'}$ and D' by

$$\overline{D'} = \frac{1}{t} \sum_{i=2}^{t+1} D^i, \text{ and}$$
$$D' = \overline{D'} + N(0, \sigma^2)^n.$$

That is, D' is distributed according to the output of Attack. Note that D^1 is not part of the set that produces D'.

Fixing, $D, D^1, \ldots, D^m$, we consider D' and D^* as defining probability measures on the document space. We claim that $\mathsf{Detect}(D')$ still outputs 1 with unacceptably high probability if $\mathsf{Detect}(D^*)$ outputs 1 with a reasonably large probability.

We now proceed to show that there is a tradeoff between the false positive and false negative probabilities. First we define a parameterized set of problematic documents for which the false positive probability is low.

Definition 1. *Given probability measure D' and D^*, and a parameter γ, we define the* bad set B_γ *by*

$$B_\gamma = \{x \mid \Pr_{x \leftarrow D'}[x] \leq \gamma \Pr_{x \leftarrow D^*}[x]\}.$$

This set is bad for the attacker, because Detect can safely output 1 without incurring too large a probability of producing a false positive. Lemma 2 bounds the probability that Detect makes a false positive as a function of γ.

Lemma 2. $\Pr_{x \leftarrow D'}[\mathsf{Detect}(x) = 1] \geq \gamma \cdot \left(\frac{\rho}{t} - \Pr_{D^*}[B_\gamma]\right)$.

Proof. We have

$$\begin{aligned}\Pr_{x \leftarrow D^*}[\mathsf{Detect}(x) = 1 \wedge x \notin B_\gamma] &\geq \Pr_{x \leftarrow D^*}[\mathsf{Detect}(x) = 1] - \Pr_{x \leftarrow D^*}[x \in B_\gamma] \\ &\geq \frac{\rho}{t} - \Pr_{x \leftarrow D^*}[x \in B_\gamma].\end{aligned}$$

Thus,

$$\begin{aligned}
\Pr_{x \leftarrow D'}[\mathsf{Detect}(x) = 1] &\geq \Pr_{x \leftarrow D'}[\mathsf{Detect}(x) = 1 \wedge x \notin B_\gamma] \\
&\geq \gamma \cdot \Pr_{x \leftarrow D^*}[\mathsf{Detect}(x) = 1 \wedge x \notin B_\gamma] \\
&\geq \gamma \cdot \left(\frac{\rho}{t} - \Pr_{x \leftarrow D^*}[x \in B_\gamma]\right). \square
\end{aligned}$$

We now obtain, for some reasonable setting of parameters, a lower bound on the false positive probability.

Lemma 3. *Let* $t \geq \frac{\alpha}{\epsilon}\sqrt{n/\ln n}$ *and* $\sigma = \epsilon\Delta/(2\sqrt{n})$. *If* $\Pr_{x \leftarrow D^*}[\mathsf{Detect}(x) = 1] \geq \rho/t$, *for* $1/\rho = o(\ln n)$, *then*

$$\Pr_{x \leftarrow D'}[\mathsf{Detect}(x) = 1] \geq \frac{\epsilon}{\alpha}\rho n^{-\beta - 1/2}\sqrt{\ln n},$$

where $\beta = (2/\alpha)(1 + 1/\alpha)$ *and* n *is sufficiently large.*

Proof. For the proof, we set some of the parameters in the expression given in Lemma 2 and use the lemma to lower bound the probability of a false positive. The value of $\gamma > 0$ must be chosen to balance between two competing considerations imposed by the γ term and the $\rho/t - \Pr_{x \leftarrow D^*}[x \in B_\gamma]$ term. Intuitively, when γ is close to 1, then x is often in B_γ, but this is not so advantageous for the Detect; when γ is small, it is indeed good for Detect to have $x \in B_\gamma$, but this hardly ever happens.

We will choose γ such that $\Pr_{D^*}[B_\gamma] \leq \rho/(2t)$; γ will be n^β for some constant β. Since $\Pr_{x \leftarrow D^*}[\mathsf{Detect}(x) = 1] \geq \rho/t$, $\Pr_{x \leftarrow D^*}[\mathsf{Detect}(x) = 1 \wedge x \notin B_\gamma] \geq \rho/(2t)$. Then the probability of a false positive for document instances from D', will be at least $\gamma\rho/(2t)$.

Although each point x we consider is an n-dimensional quantity, we can exploit the spherical symmetry of n-dimensional Gaussian distributions as follows. Given a point x, let $x_\|$ denote the projection of x onto the line L connecting $\overline{D^*}$ and $\overline{D'}$. We define $d_\|(x)$ to be $d(\overline{D^*}, x_\|)$ if $\overline{D^*}$ is between $x_\|$ and $\overline{D'}$, and $-d(\overline{D^*}, x_\|)$ otherwise. We define $d_\perp(x)$ to be the distance from x to L.

Now, by the spherical symmetry of Gaussian distributions, we have

$$\Pr_{x \leftarrow D^*}[x] = c_n \exp\left(\frac{-d^2(x, \overline{D^*})}{2\sigma^2}\right) \text{ and}$$
$$\Pr_{x \leftarrow D'}[x] = c_n \exp\left(\frac{-d^2(x, \overline{D'})}{2\sigma^2}\right).$$

where c_n is some normalization constant, depending on n. Let $\delta = d(\overline{D^*}, \overline{D'})$; note that $\delta \leq \Delta/t$. By the Pythagorean theorem and elementary geometry, we have

$$d^2(x, \overline{D^*}) = d_\|^2(x) + d_\perp^2(x) \text{ and}$$
$$d^2(x, \overline{D'}) = (d_\|(x) + \delta)^2 + d_\perp^2(x).$$

Hence,

$$\Pr_{x\leftarrow D^*}[x] = c_n \exp - \left(\frac{d_{\|}^2(x) + d_{\perp}^2(x)}{2\sigma^2} \right) \text{ and}$$

$$\Pr_{x\leftarrow D'}[x] = c_n \exp - \left(\frac{(d_{\|}(x) + \delta)^2 + d_{\perp}^2(x)}{2\sigma^2} \right).$$

Let $b(x) \stackrel{\text{def}}{=} \dfrac{\Pr_{x\leftarrow D'}[x]}{\Pr_{x\leftarrow D^*}[x]} = \exp \dfrac{-(2d(x_{\|})\delta + \delta^2)}{2\sigma^2}$. By definition, $B_\gamma = \{x \mid b(x) \leq \gamma\}$. Let $\sigma = \epsilon\Delta/(2\sqrt{n})$, where ϵ is to be determined later. Thus, $\delta \leq 2\sigma\sqrt{n}/(\epsilon t)$, hence

$$b(x) \geq \exp - \left(\frac{2d(x_{\|})\sqrt{n}}{t\epsilon\sigma} + \frac{2n}{\epsilon^2 t^2} \right).$$

If $b(x) \leq \gamma$, we get

$$d(x_{\|}) \geq \sigma \left(\frac{\epsilon t}{2\sqrt{n}} \ln \frac{1}{\gamma} - \frac{\sqrt{n}}{\epsilon t} \right).$$

Thus, we can bound $\Pr_{x\leftarrow D^*}[B_\gamma]$ as

$$\Pr_{x\leftarrow D^*}[B_\gamma] \leq \int_{d(x_{\|}) \geq \sigma\left(\frac{\epsilon t}{2\sqrt{n}} \ln \frac{1}{\gamma} - \frac{\sqrt{n}}{\epsilon t}\right)} \frac{1}{\sqrt{2\pi}\sigma} \exp \frac{-x^2}{2\sigma^2},$$

which is upper bounded by

$$\frac{1}{\sqrt{2\pi}} \cdot \left(\frac{1}{\frac{\epsilon t}{2\sqrt{n}} \ln \frac{1}{\gamma} - \frac{\sqrt{n}}{\epsilon t}} \right) \exp -\frac{1}{2} \left(\frac{\epsilon t}{2\sqrt{n}} \ln \frac{1}{\gamma} - \frac{\sqrt{n}}{\epsilon t} \right)^2,$$

when $\dfrac{\epsilon t}{2\sqrt{n}} \ln \dfrac{1}{\gamma} - \dfrac{\sqrt{n}}{\epsilon t} > 0$. Here, we are exploiting the spherical symmetry of our n-dimensional Gaussian distribution: projecting onto a line gives a 1-dimensional Gaussian distribution.

We are interested in the case when this is at most $\rho/(2t)$. Now, $t = \frac{\alpha}{\epsilon}\sqrt{n/\ln n}$. Set $\gamma = n^{-\beta}$. Then, the above bound is at most

$$\frac{1}{\sqrt{2\pi \ln n}} \cdot \frac{1}{\left(\frac{\alpha\beta}{2} - \frac{1}{\alpha}\right)} \cdot n^{-\frac{1}{2}\left(\frac{\alpha\beta}{2} - \frac{1}{\alpha}\right)^2}.$$

If we set $\beta = (2/\alpha)(1 + 1/\alpha)$, then for large enough n this is less than

$$\frac{1}{\sqrt{2\pi n \ln n}} < \frac{\rho}{2t},$$

for $\rho = 1/o(\ln n)$.

We must also ensure that

$$\frac{\epsilon t}{2\sqrt{n}} \ln \frac{1}{\gamma} - \frac{\sqrt{n}}{\epsilon t} > 0.$$

When $t = \frac{\alpha}{\epsilon}\sqrt{n/\ln n}$, for a given α, our choice of $\beta = (2/\alpha)(1 + 1/\alpha) > 2/\alpha^2$ guarantees $\alpha\beta/2 - 1/\alpha > 0$.

3.3 The Final Calculation

Lemma 3 gives a criterion for when the output of

$$\mathsf{Attack}(i_1, \ldots, i_t, D^{i_1}, \ldots, D^{i_t})$$

will cause Detect to have a high false positive rate; however, these bad indices may be very uncommon, and almost never encountered by the Detect procedure, since we assume the adversary receives a uniformly chosen subset. It remains to bound how likely it is for such a bad $i_1, \ldots, i_t$ to be chosen. There are many ways of doing so; a very simple argument will make our point.

First, we show a high false positive rate under a different distribution of indices, defined by the following procedure.

1. Choose $I = i_1, \ldots, i_t$ uniformly,
2. Determine the j maximizing the probability that, after Attack produces D^*, Detect returns i_j.
3. Remove i_j from I and replace it with a new element, chosen at random (without replacement), giving I'.

The sets I, I' are completely analogous to $\{1, \ldots, t\}$ and $\{2, \ldots, t+1\}$ in the previous analysis. Lemma 2 implies that whenever Detect catches the colluders (correctly) on set I with probability ρ, for $\rho > 1/\sqrt{\ln n}$ (and n sufficiently large), it will falsely accuse someone with probability at least $\rho\phi$, where

$$\phi \stackrel{\text{def}}{=} \frac{\epsilon}{\alpha} n^{-\beta-1/2}\sqrt{\ln n}$$

when the attack is based on set I'. Note that the $1/\sqrt{\ln n}$ term can be replaced by any function $f(n)$ where $1/\ln n = o(f(n))$.

By a simple probability calculation, if Detect is successful with probability q (catches a correct colluder), when I is chosen uniformly (as in the procedure above), it will make a false accusation with probability $(q - 1/\sqrt{\ln n})\phi$ on sets chosen according to the distribution of I' in the procedure above.

We next observe that for any t-set I^*, $\Pr[I' = I^*] \leq t \Pr[I = I^*]$. That is, the distribution on I' assigns at most t times the weight to some subset than would the uniform distribution. To see this, note that for any I^* there are at most $t(m-t)$ values of I in the above procedure such that I' could possibly be equal to I^* (there are that many ways of swapping an index out). For each of these possibilities, the probability that I' is indeed equal to I^* is either 0 (when the index that needed to be swapped out was not the maximally accused index) or exactly $1/(m-t)$ (the probability of swapping the right index in).

By the "flatness" property we have shown for I', it then follows that if the false positive rate is $(q - 1/\sqrt{\ln n})\phi$ when the sets are chosen according to I', then the false positive rate is at least $(q - 1/\sqrt{\ln n})\phi/t$ when the sets are chosen uniformly.

4 Conclusion and Open Problems

We have shown that in the framework of [3], $O(\sqrt{n/\ln n})$ adversaries suffice to break the watermarking scheme. Within this framework, the attack is essentially oblivious to the actual watermarking method. In practice, a real document consists of much more than the n-vector assumed for the theoretical model; the relationship between a document and its corresponding n-vector may be more obscure. As soon as this correspondence (and a way of computing inverses) is figured out, our attack is applicable.

An interesting open question is to generalize our result for a general class of metrics. One criticism of the Euclidean distance is that it is not always a good measure of fidelity; one would like to choose ones notion of fidelity.

A more important open question is to properly model the "do not copy" problem for watermarking. Whereas for the problem we consider, the question is what the right bound for the adversaries should be, for the other problem it is unclear whether there is a theoretically defensible solution at all.

Acknowledgements

Uri Feige provided us with crucial assistance. Based on our preliminary notes, Steven Mitchell, Bob Tarjan, and Francis Zane have written and shared with us an alternate exposition of our methods.

References

1. D. Boneh and J. Shaw. Collusion-secure fingerprinting for digital data. *Proc. Advances in Cryptology — CRYPTO*, Springer LNCS 963:452–465, 1995.
2. B. Chor, A. Fiat, and M. Naor. Tracing traitors. *Proc. Advances in Cryptology — CRYPTO*, Springer LNCS 839:257–270, 1994.
3. I. Cox, J. Kilian, T. Leighton, and T. Shamoon. A secure, robust watermark for multimedia. *IEEE Transaction on Image Processing*, 6(12):1673–1687, 1997.
4. C. Dwork, J. Lotspiech, and M. Naor. Digital signets: Self-enforcing protection of digital information (preliminary version). *Proc. 28th ACM Symposium on Theory of Computing*, pp. 489–498, 1996.
5. J. Kilian, T. Leighton, L. R. Matheson, T. G. Shamoon, R. E. Tarjan, and F. Zane. Resistance of digital watermarks to collusive attacks. *Technical Report TR-585-98*, Department of Computer Science, Princeton University, 1998.
6. M. Naor and B. Pinkas. Threshold Traitor Tracing. *Proc. Advances in Cryptology — CRYPTO*, Springer LNCS 1462:502–517, 1998.
7. B. Pfitzmann. Trials of traced traitors. *Proc. 1st International Workshop on Information Hiding*, Springer LNCS 1174:49–64, 1996.
8. B. Pfitzmann and M. Schunter. Asymmetric fingerprinting (extended abstract). *Proc. Advances in Cryptology — EUROCRYPT*, Springer LNCS 1070:84–95, 1996.
9. B. Pfitzmann and M. Waidner. Anonymous fingerprinting. *Proc. Advances in Cryptology — EUROCRYPT*, Springer LNCS 1233:88–102, 1997.
10. B. Pfitzmann and M. Waidner. Asymmetric fingerprinting for larger collusions. *Proc. 4th ACM Conference on Computer and Communications Security*, pp. 151–160, 1997.