CERIAS Tech Report 2003-07

ATTACKING DIGITAL WATERMARKS

by Radu Sion and Mikhail Atallah

Center for Education and Research in
Information Assurance and Security,
Purdue University, West Lafayette, IN 47907

Attacking Digital Watermarks *

Radu Sion, Mikhail Atallah
Center for Education and Research in Information Assurance,
Computer Sciences, Purdue University
West Lafayette, IN, 47907, USA
http://www.cs.purdue.edu/homes/sion
[sion, mja]@cs.purdue.edu

Abstract

This paper discusses inherent vulnerabilities of digital watermarking that affect its mainstream purpose of rights protection. We ask: how resistant is watermarking to un-informed attacks ?

There are a multitude of application scenarios for watermarking and, with the advent of modern content distribution networks and the associated rights assessment issues, it has recently become a topic of increasing interest. But how well is watermarking suited for this main purpose of rights protection ? Existing watermarking techniques are vulnerable to attacks threatening their overall viability. Most of these attacks have the final goal of removing the watermarking information while preserving the actual value of the watermarked Work.

In this paper we identify an inherent trade-off between two important properties of watermarking algorithms: *being "convincing enough" in court* while at the same time *surviving a set of attacks*, for a broad class of watermarking algorithms. We show that there exist inherent limitations in protecting rights over digital Works. In the attempt to become as convincing as possible (e.g. in a court of law, low rate of false positives), watermarking applications become more fragile to attacks aimed at removing the watermark while preserving the value of the Work. They are thus necessarily characterized by a significant (e.g. in some cases 35%+) non-zero probability of being successfully attacked without any knowledge about their algorithmic details. We quantify this vulnerability for a class of algorithms and show how a minimizing "sweet spot" can be found. We then derive a set of recommendations for watermarking algorithm design.

1 Introduction

Digital Watermarking [8] [9] [10] [12] [13] [14] [15] [19] [20] [29] [30] embeds un-detectable (un-perceivable) hidden information into digital Works mainly to protect the data from unauthorized duplication and distribution by enabling provable rights over the content. Efforts in the general area of watermarking for rights protection have recently been accelerated by the advent of on-line media distribution and data outsourcing. Mainstream research has focused on multi-media content watermarking (e.g. audio, video, images) but different other avenues have also been explored such as software [7] [18], natural language [2], relational databases [28], semi-structured data [25]. A common consensus has been implicitly assumed, namely that watermarking indeed lives up to its claimed features. In a seminal paper [20], the main desiderata and features of watermarking (for multimedia) are outlined generically: it should not degrade the perceived quality of the marked Work; the ability to detect the presence/content of a watermark should require the knowledge of a secret (key); different watermarks in the same

*Portions of this work were supported by Grants EIA-9903545, IIS-9985019 and IIS-9972883 from the National Science Foundation, Contract N00014-02-1-0364 from the Offi ce of Naval Research, and by sponsors of the Center for Education and Research in Information Assurance and Security.

Work should not interfere with each other; collusion attacks should not be possible; *the watermark in a certain Work should survive any value-preserving transformation on that particular Work.*

How well do algorithms and their associated implementations conform to these ideals ? [8, 13] present excellent area surveys as well as comprehensive examples of algorithms for watermarking (mainly) multi-media Works. We know now that arbitrary large collusion attacks cannot be defeated against [4]. Moreover, while most watermarking algorithms prove to be safe against a considered set of value-preserving transformations (e.g. JPEG compression) they certainly fail with respect to many others. This shortcoming can be directly traced back to the relativity of the "value" and "quality" concepts. Several (mostly experimental) efforts explored the ability to analyze and quantify the "goodness" of watermarking applications, resulting in various watermark benchmarking "suites" [1] mainly for multimedia (i.e. images). Additional research [5] [16] [17] aimed at analyzing concepts such as available bandwidth in the broader area of information hiding from a signal-processing, information-theoretic perspective, focusing mainly on various multimedia techniques. Recent research results [1] [24] [25] [28] in watermarking different types (i.e. including non-media) of digital Works gradually raised the need for a more complete understanding of generic watermarking algorithms (i.e. not necessarily in a specific data domain, e.g. images), their abilities and associated bounds. In watermarking generic data types one cannot count on a large noise-bandwidth, so common in the multimedia domain. The embedding abilities are severely constricted and must be utilized with a maximum of efficiency for the desired purpose. For example, in [28], watermarking relational data is often subject to a very restrictive set of data quality assessments making it often difficult to embed even a single bit watermark. This narrow embedding bandwidth raises major concerns with respect to theoretical watermarking vulnerability limits. After all, it seems that if the changes allowed to be performed are small and very application-specific, an attacker would be more likely to undo them or simply alter the Work enough to destroy the watermark while preserving sufficient value.

Thus one particular question becomes of interest, namely: *Are there theoretically assessable bounds on watermark vulnerability with respect to an arbitrary watermarking method ?* In other words, what is the inherent safety/vulnerability of a generic (i.e. with a minimum amount of assumptions, without considering implementation particularities) watermarking algorithm ? An answer to this question might afterward derive real-life recommendations for fine-tuning actual algorithms to increase their marking resilience. In this paper we present initial exploratory results for a broad class of watermarking algorithms. We introduce a model and associated principles of sound watermarking [2]. We the use this model to explore generic watermarking resilience and assess bounds.

While we believe it generalizes to a much larger class of algorithms, the quantitative part of our analysis is done within a well-defined algoritmic class framework, namely our research in watermarking relational data [28]. The contributions of this paper include: **(i)** the identification and analysis of inherent limitations of watermarking, including the trade-off between two important watermarking properties: *being enough 'convincing" in court* while at the same time *surviving a set of attacks*. This trade-off derives naturally from the inverse proportional nature of their relationship. **In the attempt to become as court convincing as possible, a watermarking application becomes more fragile to attacks aimed at removing the watermark, while preserving the value of the Work.** It becomes thus necessary characterized by a significant non-zero probability of being successfully attacked. **(ii)** the definition of watermark *vulnerability*, a quantifiable measure of generic watermarking safety. **(iii)** an optimality principle (quantified and proved for a broad class of algorithms) that postulates the minimization of watermark vulnerability in specific data points.

The paper is structured as follows. Section 2 defines the main model and foundations. Section 3 presents our

[1] StirMark: http://www.cl.cam.ac.uk/~fapp2/watermarking/stirmark
CheckMark: http://watermarking.unige.ch/Checkmark
OptiMark: http://poseidon.csd.auth.gr/optimark

[2] This model is an extension of our work in [26].

main analysis on watermark vulnerability for a certain class of algorithms. Section 4 discusses the broadening of the applicability of these results by relaxing class-defining assumptions. Section 5 concludes.

2 Foundations

One fundamental difference between watermarking [9] [10] [12] [14] [15] [19] [20] [23] [24] [25] [27] [29] [30] and generic data hiding [6] [16] [17], [22], resides in the main applicability and descriptions of the two domains. Data hiding in general and covert communication in particular, aims at enabling Alice and Bob to exchange messages [3] [11] [21] in a manner as resilient and stealthy as possible, through a medium controlled by evil Mallory. On the other hand, digital watermarking (especially for rights assessment) is deployed by Alice to prove rights over a piece of data, to Jared the Judge, usually in the case when Tim the Thief [3] benefits from using/selling that very same piece of data or maliciously modified versions of it.

In digital watermarking, the actual value to be protected lies in the Works themselves whereas information hiding usually makes use of them as simple value "transporters". Rights assessment can be achieved by demonstrating that a particular Work exhibits a rare property (read "hidden message" or "watermark"), usually known only to Alice (with the aid of a "secret" - read "watermarking key"). For court convince-ability purposes this property needs to be so rare that if one considers any other random Work "similar enough" to the one in question, this property is "very improbable" to apply (i.e. bound on false-positives). There is a threshold determining Jared's convince-ability related to the "very improbable" assessment. This defines a main difference from steganography: from Jared's perspective, specifics of the property (e.g. watermark message) are irrelevant as long as Alice can prove "convincingly" it is she who embedded/induced it to the original (non-watermarked) Work.

It is to be stressed here this particularity of watermarking for rights assessment. In watermarking the emphasis is on "detection" rather than "extraction". Extraction of a watermark (or bits of it) is usually a part of the detection process but just complements the process up to the extent of increasing the ability to convince in court. If recovering the watermark data in itself becomes more important than detecting the actual existence of it (i.e. 'yes/no answer') then this is a drift toward covert communication and pure information hiding (steganography).

2.1 Model

Let $\mathbb{D}$ be the domain of all possible Works to be considered for watermarking [4]. The value associated with such Works is owned by a given rights holder (possibly the Work creator). Watermarking (for rights protection) tries to protect the association between the value carrying Work and the identity of its rights holder.

More specific, watermarking deploys *information hiding* techniques in order to embed a rights "witness" *within* the Work to be considered. The embedding process usually distorts/changes the original Work. The level of distortion introduced is usually measured within a model of tolerable change from a data consumer perspective. For example if the final data consumer is to be a human, and the data domain is digital multimedia, the Human Sensory System's limitations are defining allowable distortion bounds in the watermarking framework. If the Work is a JPEG image, the allowable distortion could be defined by the human eye limitations. Works can exhibit different value levels when put to different uses. We need a way to express the different associated values of Works, in different *usability domains*.

Usability Domain: A *usability domain* is defined as a set of functionals, $e = \{f | f : \mathbb{D} \rightarrow [0, 1]\}$, quantifying intrinsic data value in terms of its specific use. In a real world algorithm the considered usability domain is

[3] Tim's middle name is *Mallory*.

[4] For example in case of digital media Works we can simply assume that $\mathbb{D}$ is the set of all variable sized strings over $\mathbb{B} = \{0, 1\}$.

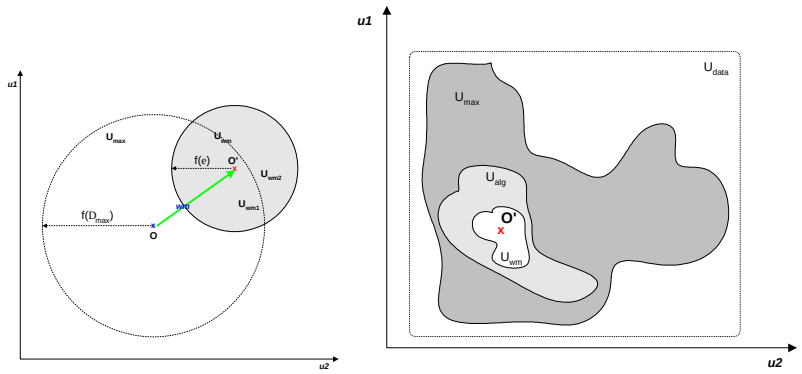

Figure 1: (a) A 2-dimensional view of an usability space. A point uniquely identifies a Work in $\mathbb{D}$ (e.g. coordinates in this space are DCT coefficients considered for watermark embedding). Watermarking results in a point O', a "watermarked" version of O and is naturally represented as a transform in the usability space. (b) Usability vicinities of a certain Work $O \in \mathbb{D}$ for a given marking algorithm. U_{data} is defined by the actual data type of the usability metrics. U_{max} is the maximal allowable usability vicinity with respect to the associated usability domain(s) (e.g. Human Visual System). The results (U_{alg}) of a valid marking algorithm with respect to a given Work and all other possible inputs should be contained within the maximal allowable usability vicinity (U_{max}) of the Work. U_{wm} is determined by ϵ_w.

constructed by mapping real world properties to actual parametrized functionals in e. Not.: Let the set of all usability domains be $\mathbb{U}$.

Usability: The *usability* metric of an Work $O \in \mathbb{D}$ corresponding to a domain $e \in \mathbb{U}$ ($e = \{f_1, f_2, ..., f_q\}$) is defined as $u(O, e)$ where $u : \mathbb{D} \times \mathbb{U} \rightarrow [0, 1]$. $u(d, e)$ is a combination of all the elements of e. For simplicity we will assume that $|e| = 1$. In this case we define $e = \{u\}$.

In real life settings each usability domain is characterized uniquely by the set of data items (or degrees of freedom) that are alterable for the purpose of watermarking. For example in the case of image frequency transform encodings such as the DCT transform, the main usability domain (associated with the human sensory system) is uniquely identified by the set of DCT coefficients considered in the watermarking process. This defines a multi-dimensional usability "space" in which each point uniquely identifies a Work in $\mathbb{D}$ and the process of watermarking becomes a simple translation from a point O to the watermarked version O' (see Figure 1 (a)).

The concept of usability enables the definition of a certain threshold below which the Work is not "usable" anymore in the given domain. In other words, it "lost its value" to an unacceptable degree. The notion of usability is related to *distortion*. A highly distorted Work (e.g. as result of watermark embedding or attacks) will likely suffer a drop in its distortion domain usability.

Usability Vicinity: Let $v \in \mathbb{U}$ be a usability domain and a maximum allowed change in usability Δu_{max}. We say that element $x \in \mathbb{D}$ is *in the radius Δu_{max} usability vicinity of $O \in \mathbb{D}$ with respect to v* [5] iff. $\Delta u(d, x, v) < \Delta u_{max}$.

Note that the usability vicinity of a certain Work $O \in \mathbb{D}$ with respect to a considered usability domain defines actually the set of possible watermarked versions of O with respect to it and Δu_{max}. In the usability space in Figure 1 (a) this vicinity translates into an shape "around" the original Work O, labeled U_{max}. Thus, the Work and any minor altered version of it (within usability vicinity bounds) can be uniquely identified by its locality within the considered usability domain (i.e. by it's usability "coordinates").

Watermark: Given an un-marked Work $O \in \mathbb{D}$ and the considered watermarked version of it, $O' \in \mathbb{D}$,

[5] And vice-versa. It is commutative.

where O' is within radius Δu_{max} usability vicinity of O, a key $k \in \mathbb{K}$ [6] , a usability domain $v \in \mathbb{U}$ and a maximum allowed change in usability Δu_{max}, a *watermark* can be asserted by a special property functional $w : \mathbb{D} \times \mathbb{K} \to \mathbb{B}$, defined by the following: $w(O, k) = 0$, $w(O', k) = 1$ and there exists $\epsilon_w \in (0, 1)$ such that we have $P(w(x, k) = 1, \forall x \in \mathbb{D}, x \neq O', x \in U_{max} | w(O', k) = 1) < \epsilon_w$. Not.: Let $\mathbb{W}_{\mathbb{D}}$ be the set of all w over a given $\mathbb{D}$.

In plain words, a watermark can be defined as a special induced (through watermarking) property (w) of a watermarked Work $O' \in \mathbb{D}$, so rare, that if we consider any other Work $x \in \mathbb{D}$, "close-enough" to the original Work O, the probability that x exhibits the same property can be upper-bounded. This is derived from the requirement to be enough "court-convincing" by bounding the rate of false-positives. In Figure 1 (a) this probability maps into an area "around" the watermarked Work O', labeled U_{wm}, zone in which the points directly correspond to Works that exhibit the watermark property. Of particular interest then becomes the intersection between U_{max} and U_{wm}, an area bounded by the watermarking construction (i.e. ϵ_w). This area defines the Works that are both usable (with respect to the original Work) and watermarked. Intuitively, one main challenge of watermarking becomes to find/derive O' such that, given only O' it will be very hard for an attacker (e.g. Mallory) to determine an $x \neq O'$ inside the usability vicinity of O, for which $w(x, k) = 0$.

Note: Throughout this paper we make certain simplifying assumptions. The usability spaces are discussed in a 2 dimensional context, the real world usability spaces are multi-dimensional (e.g. $O(n^2)$ in case of DCT encoding with a DCT matrix of size $n \times n$). We also assume a continuous nature of the usability vicinity shapes and propose extensions in Section 4.

Algorithm: A *watermarking algorithm* can be described as a functional $a : \mathbb{D} \times \mathbb{K} \to \mathbb{D} \times \mathbb{W}$, which, given as input an Work $O \in \mathbb{D}$ provides a watermarked version of the Work, O', and a property functional w that enables watermark detection. Not.: Let $\mathbb{A}_{\mathbb{D}}$ be the set of all a over a given $\mathbb{D}$.

Naturally, for a watermarking algorithm to be valid, its results, with respect to a given Work and all other possible inputs should be contained within the maximal allowable usability vicinity (U_{max}) of this Work, see Figure 1 (b).

Attack: Given a watermarking algorithm $a \in \mathbb{A}_{\mathbb{D}}$, an Work $O \in \mathbb{D}$ and its watermarked version $O' \in \mathbb{D}$, Δu_{max}, $\forall k \in \mathbb{K}$, a *successful watermarking attack* is defined by $z : \mathbb{D} \to \mathbb{D}$ such that $\Delta u(z(O'), O) < \Delta u_{max}$ and $w(z(O'), k) = 0$. In other words, an attack tries to maintain the attacked watermarked Work within the usability vicinity of the original non-watermarked version, while removing the watermark [7]. Not.: Let $\mathbb{Z}_w$ be the set of all attacks z for a given w.

Vulnerability: It is desirable to be able to quantify the ability of a certain attack to succeed against a watermarking scheme. This can be used as a metric of "goodness" of that specific scheme. The probability that a *particular* attack succeeds is not easy to compute as it depends on a variety of parameters, including the potentially infinite set of input watermarked Works. Instead of focusing on specifics, one solution would be to assess bounds for a generic class of algorithms and attacks.

We define the *single point vulnerability* of a certain watermarking method with respect to a given watermarked Work as the lower bound on the probability of success (e.g. watermark removal, resulting Work still "usable") of an attacker with no additional knowledge but the actual watermarked Work itself. That is, given a watermarked Work, what is the probability that an un-informed attack succeeds to remove the watermark and produce a result within the usability vicinity (U_{max}) of the original Work. It turns out that for a broad class of algorithms this can be quantified exactly. It naturally provides a measure of watermark fragility.

We define the *algorithm vulnerability* of a certain watermarking method as the average of the single point vulnerability over the entire input space. Although this value presents a great importance in assessing the quality of a watermarking method, it is not trivial to compute, especially in cases with infinite input spaces. We can

[6] Where $\mathbb{K}$ is the set of secrets (i.e. keys) involved in the watermarking process.

[7] We are focusing on the arguably most common class of attacks, having a final goal of removing the watermarking information while preserving the value.

define a *k-sampling approximation* of it that can be determined by taking k random samples (i.e. Works to be watermarked) from the input space and computing the algorithm vulnerability over this limited subset.

In the following we are analyzing single point vulnerability through P_{sa}, the probability of a successful attack, given any watermarked Work.

2.2 A Concise Overview on Watermarking Relational Data

In [28] we introduce a resilient watermarking method for numeric relational data. Here we present a very brief, distilled overview that should allow us to use this as an example without sacrificing any generality.

For simplicity, let us consider the input Work (i.e. relational data) being a set of n numbers $(s_i)_{i\in(1,n)}$. The output of the watermarking algorithm is another set of numbers $(v_i)_{i\in(1,n)}$, the watermarked version. The usability space is n-dimensional in this case. It suffices to know that watermark encoding occurs by slight modifications in the number set (altering some secret distribution characteristics) [8]. For more embedding algorithm details see [24, 28].

In this domain one common data quality metric is defined in terms of the mean squared error measured as the normalized sum of the square roots of all the performed (small) changes on the original Work, $MSE = \frac{\sum (s_i - v_i)^2}{n}$. The quality of the result is said to be satisfactory if the mean squared error is below a certain allowed upper bound Δu_{max}, in other words, if the resulting watermarked Work is not "too far away" from the original. In the associated n-dimensional usability space this quality assessment defines a vicinity U_{max} of the initial data set, in the shape of a sphere of radius Δu_{max}.

For simplicity, let's assume for now that U_{wm} (the zone in which we find Works that exhibit the watermark property) is continuous and also sphere-shaped (it is also a result of a MSE type of distance). We are relaxing this assumption later on. A 2-dimensional illustration of this is given in Figure 1 (a).

3 Analysis

Given a certain Work O (whose rights belong to Alice) and its watermarked version O', lets put ourselves in the position of Mallory attempting to attack O' and yield an un-watermarked Work O'', still usable with respect to O.

There exists a court-authority to which both Mallory and Alice would like to prove their rights ownership to. Lets assume this court mandates a certain maximal value on ϵ_w (see Section 2.1), for example 3% [9] , value which Mallory knows (usually public). There also exists an associated maximal allowable distortion bound, Δu_{max} (the radius of U_{max}), guaranteed on the resulting watermarked Work with respect to the original. This value is probably also public as it is included in the watermarked Work specification, delivered by Alice.

Mallory knows O' and it is only reasonable to assume that it is also aware of the quality metric (e.g. MSE, see Section 2.2) associated with the data domain of the Work in question. What Mallory doesn't know is the original Work O as well as its maximal usability vicinity (U_{max}).

What are our options in attacking ? Given the knowledge of O', it is the only natural starting point of our attack. It is also safe to assume that Alice encoded a watermark that does not exceed the ϵ_w false positive rate (modeled by $U_{max} \cap U_{wm}$), thus $||U_{max} \cap U_{wm}|| = \epsilon||U_{max}||$. Given this limited amount of available knowledge, the remaining thing to do is to randomly define a distance R_a from O' and "attack" by picking a point inside the resulting R_a radius vicinity of O', U_{attack}, see Figure 2.

Note: Because we know ϵ_w, we can estimate the minimal area of U_{wm} as being at least $\epsilon_w||U_{max}||$. This is why when choosing R_a we have to make sure that the resulting U_{attack} vicinity shape is larger than U_{wm}. Thus we

[8] The generality of this is warranted by the fact that many other existing algorithms [8] are proceeding similarly, by mapping the input Works (e.g. images) to a quantifiable domain (e.g. DCT coefficients, Least Significant Bit spaces) that is then to be "modulated" (i.e. altered) according to the watermark signal.

[9] The court is convinced only above a 97% true-pozitive rate.

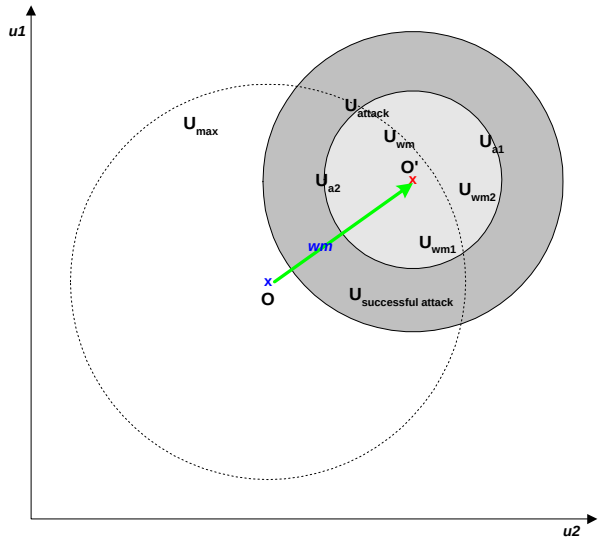

Figure 2: Mallory attacks.

have at least $R_a > \sqrt{\epsilon_w}\Delta u_{max}$. Otherwise, this, and the fact that U_{attack} is also centered in O' lead immediately to the conclusion that each point inside of U_{attack} corresponds to a watermarked Work, thus no attack success is achieved. Thus, we can improve our success probability even more by picking an "attack point" only within $U_{attack} \setminus U_{\epsilon_w}$, where U_{ϵ_w} is a vicinity of O' of radius ϵ_w. For space reasons this provision is not included in this analysis. An extension is discussed in Section 4.

Also because Δu_{max} is assumed to be public, Mallory can infer that there is no point in choosing an U_{attack} larger than required to just include U_{max} (all points outside of U_{max} are irrelevant anyway). Thus at the limit, $R_a < 2\Delta u_{max}$. In conclusion, we have

$$\frac{R_a}{\Delta u_{max}} \in (\sqrt{\epsilon}, 2) \tag{1}$$

Given the above, a successful attack is one which yields results in the area $U_{sa} = U_{max} \cap (U_a \setminus U_{wm})$. The probability of this becomes then

$$P_{sa} = \frac{||U_{sa}||}{||U_a||} = \frac{||U_{a2} \setminus U_{wm1}||}{||U_a||} = \frac{||U_{a2}|| - ||U_{wm1}||}{||U_a||} \tag{2}$$

We know from Section 2.1 that $\frac{||U_{wm} \cap U_{max}||}{||U_{max}||} = \frac{||U_{wm1}||}{||U_{max}||} = \epsilon_w$, $||U_{max}|| = \pi\Delta u_{max}^2$ and $||U_a|| = \pi R_a^2$, thus (2) becomes

$$P_{sa} = \frac{||U_{a2}|| - \epsilon_w||U_{max}||}{||U_a||} = \frac{||U_{a2}|| - \epsilon_w\pi\Delta u_{max}^2}{\pi R_a^2} \tag{3}$$

Convince-ability Trade-off: Equation (3) outlines the direct relationship between the probability of a successful attack and ϵ_w. *The smaller the ϵ_w value is (i.e. the more "convincing" in court, from a false-positive rate point of view), the higher the probability of success of an attack.* We call this the **watermark convince-ability trade-off**.

In the following, it helps knowing that the usability shapes are circular (see Section 2.2), the radius of U_{max} is Δu_{max} and the radius of U_a is R_a. $||U_{a2}||$ becomes

$$||U_{a2}|| = \Delta u_{max}^2 cos^{-1}(\frac{d^2 + \Delta u_{max}^2 - R_a^2}{2d\Delta u_{max}}) + R_a^2 cos^{-1}(\frac{d^2 - \Delta u_{max}^2 + R_a^2}{2dR_a}) -$$
$$\frac{1}{2}\sqrt{(-d + R_a + \Delta u_{max})(d + R_a - \Delta u_{max})(d - R_a + \Delta u_{max})(d + R_a + \Delta u_{max})} \quad (4)$$

where $d = d(O, O')$ is the distance between O and it's watermarked version (O') Because O' has to be within maximum allowable usability vicinity of the original, it is necessary that $d \leq \Delta u_{max}$. It can be seen that $||U_{a2}||$ decreases with the increase of d. It reaches minimum when d is maximal, namely when $d = \Delta u_{max}$, in other words, when the watermarking algorithm produces a resulting watermarked Work O' (asymptotically) on the boundary of U_{max}.

Optimality Principle: From (3) and (4) we derive the fact that *the vulnerability of a watermarking scheme (in our considered class) is minimized when it yields watermarked results on the boundary of the maximum allowable usability vicinity of the original un-watermarked Works.* We call this the **watermarking optimality principle**.

Recommendation: The optimality principle postulates (and we prove it for this example) the naturally occuring minimization of the successful attack probability on the vicinity boundary. This yields a recommendation to make the watermarking algorithms by design conform to this principle. Choosing this design in our relational database watermarking software increased its resilience to random ϵ-attacks (see [28]) by as much as 15%. From now on let us assume that we are going to conform to this principle, in other words $d = \Delta u_{max}$, in which case we have the revised version of (4)

$$||U_{a2}|| = d^2 cos^{-1}(1 - \frac{R_a^2}{2d^2}) + R_a^2 cos^{-1}(\frac{R_a}{2d}) - \frac{1}{2}R_a\sqrt{(2d - R_a)(2d + R_a)}$$

and (3) becomes

$$P_{sa} = \frac{d^2 cos^{-1}(1 - \frac{R_a^2}{2d^2}) + R_a^2 cos^{-1}(\frac{R_a}{2d}) - \frac{1}{2}R_a\sqrt{4d^2 - R_a^2} - \epsilon_w \pi d^2}{\pi R_a^2}$$

For a typical value of $\epsilon_w = 0.03$ (the court is convinced with a 3% false-pozitive rate) Figure 3 (a) depicts P_{sa} as a function of the $R_a/\Delta u_{max}$ ratio. It can be seen that there exists a clear maximum vulnerability spot (around $R_a/\Delta u_{max} = 0.4$) in which the probability of success of an attack exceeds 30-35% ! This result is not dependent on the technicalities of the watermarking method. It becomes so much more compelling as it occurs for *any* marking algorithm (using similar Works quality metrics, see Section 4) and any input. In this same figure we depicted the evolution of P_{sa} for $\epsilon_w = 0.01$. This more court-convincing setting (lower false-positive rate) results in even higher attack success rates (up to 40% and above).

From Mallory's perspective this is good news. It turns out that it *is* possible to defeat watermarking algorithms with a surprisingly high success rate, without any additional (insider's) knowledge [10]. This is the case even if these algorithms conform to the optimality principle outlined above. There seems to exist a "sweet spot" (characterized by $\frac{R_a}{\Delta u_{max}} \in (0.4, 0.6)$) in which P_{sa} is maximized. Mallory could make use of this by fine-tuning its attacks. This is confirmed in real data experiments [28] in which random attacks were more likely to succeed within this range.

4 Discussion

In the previous sections we analyzed a special class of watermarking algorithms considered to be simple and illustrative enough yet within the space and complexity constraints of the current scope. This class is defined by

[10] This is an inherent limitation of watermarking as a concept. From the attacker's perspective, any additional knowledge can only improve on this probability.

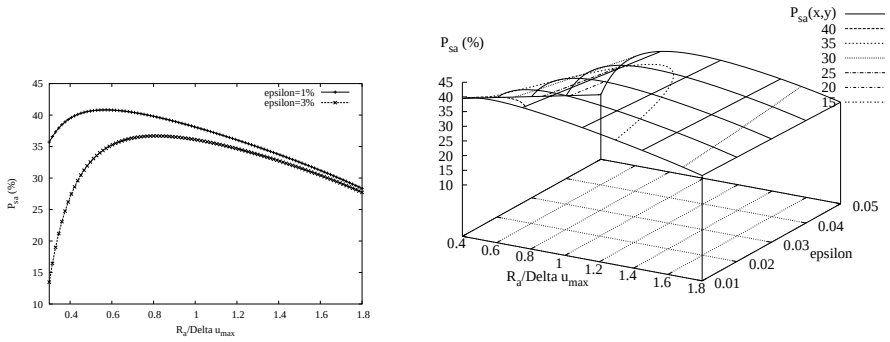

Figure 3: (a) No matter how sophisticated the watermarking method, there exists a random attack with a success probability of 35% and above (shown here for 2-dimensional usability spaces, see Section 4 for a discussion on high dimensional spaces). It can be seen that a lower ϵ value (more convincing in court) yields an even higher upper bound on attack success probability (2D cut through (b)). (b) The 3D evolution of P_{sa} with varying ϵ and $R_a/\Delta u_{max}$.

a set of assumptions. The quantitative details of our attack analysis are developed for a particular data quality metric, mean squared error. This resulted in the usability vicinity for the watermarking algorithm considered to be continuous and of a spheric shape. How restrictive are these assumptions ? Can the results be applied to broader classes of algorithms and how ? In the following we are discussing these and other issues and propose extensions.

High dimensional usability space.

In the above we considered a finite dimension of the usability space. In particular, the quantitative analysis was performed on 2D shapes. While many of real life applications feature only a finite dimensionality we can't help to wonder what happens (e.g. to U_{sa}) when the number of dimensions of the usability space grows. To the extreme, what happens when we go to infinity ? While a full fledged analysis is out of the current scope, the intuitive feel is (thanks to Ton Kalker, see acknowledgements) that in that case U_{sa} goes to zero. On the other hand, arguably, the U_{sa}/U_{attack} ratio has to be preserved. Thus it is unclear what the behavior would be in that case. Further research should explore this issue in more detail.

Shape. The assumption of a particular quality metric (mean squared error) determines directly the shapes of the analyzed Work's usability vicinities. How does this affect the generality of our analysis ? With respect to continuous shapes, while quantities may vary, the behavior of any watermarking algorithm is arguably ruled qualitatively by the *convince-ability trade-off*. This is an immediate result of the generality of (2) which does not depend on shape.

The validity of the *optimality principle* is intuitive for quality metrics resulting in convex usability vicinities. As the watermarked results are closer to the boundary of U_{max}, the probability of success of an uninformed attack decreases (see Figure 2). This is not obvious in the non-convex case. Figure 4 (b) illustrates a case where it does not work. Both O'_1 and O'_2 are (asymptotically) on the boundary of U_{max} but while O'_1 seems to minimize it's associated P_{sa}, O'_2 offers a highly attackable starting point for an uninformed Mallory[11]. A possible solution could be the construction of a "safe" subset of the shape's boundary. In this case, a watermarking algorithm should start by determining a subset of U_{max} in which watermarked Works result in a minimized P_{sa}. This is subject to further exploration.

[11]Because its attack vicinity ($U_{attack2}$) "intersects" comparably much of U_{max} as any interior point does. As discussed in the case of the optimality principle, interior points are by default non-optimal.

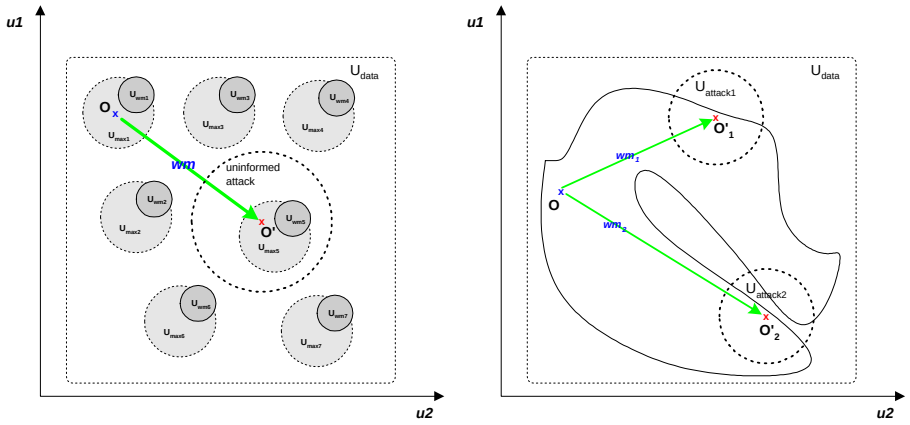

Figure 4: (a) A sparse maximum allowable usability vicinity, $U_{max} = \cup(U_{max_i})$, (b) A concave U_{max} does not respect the optimality principle.

Sparsity. How does our analysis suffer if the considered usability vicinities (e.g. U_{max}, U_{wm}) were to be sparse, "scattered" throughout U_{data} ? Sparsity in this domain is directly related to the function defi ning the maximal allowable change to the initial Work. While most known applications (e.g. wavelets, DCT for media) feature continuous usability vicinities [12], considering the case of sparsity becomes interesting as it might offer more generality. Sparsity is often associated with higher level semantic constraints. One scenario, image watermarking within a health care framework might present various legal and medical restrictions. For example, encoding a mark in a composite heart-disease image might be required not to result in alterations to the actual region containing the heart itself.

Given the sparse nature of the vicinity shapes (see Figure 4 (a)), it all boils down to the assumption made about the amount of knowledge available to Mallory. If Mallory is aware of the details of the actual cause of sparsity (i.e. function of allowable change), then it can deploy a virtually identical attack (with the one considered in our initial analysis) targeted at each sub-part composing the sparse distribution. In this case our analysis is identical.

The scenario becomes different if this knowledge is not available to Mallory. In this case the only viable option is to randomly attack within the known data domain. In this case the attack success probability has to suffer at least by a factor of U_{max}/U_{data} (assuming uniform distribution of the sparse fragments of U_{max} throughout U_{data}). Apparently this is good news for Alice. It seems that if the distortion constraints are of a higher semantic nature (i.e. resulting in sparse vicinities), successful watermark embedding [13] is less vulnerable (by a factor of U_{max}/U_{data}). This scenario requires more attention.

Partial Sparsity. A particular case occurs when only some of the vicinities are sparse, for example U_{wm}. This is equivalent to saying that the false positive upper bound, ϵ_w is "spread" throughout U_{max}. Then (see Figure 2) in (2) $||U_{wm1}||$ is to be replaced by $\epsilon U_{max}\frac{U_{a2}}{U_{max}} = \epsilon U_{a2}$ (assuming uniform distribution of U_{wm1} across U_{max}) and we yield $P_{sa} = (1-\epsilon)\frac{||U_{a2}||}{||U_a||}$, with a similar qualitative behavior to (2), thus our analysis holds.

Increasing P_{sa}. In the case of a continuous U_{wm}, Mallory can increase P_{sa} even more as follows. Knowing Δu_{max} and $||U_{wm1}|| = \epsilon_{wm}\pi\Delta u_{max}^2$ can immediately result in the computation of the radius R_{wm} of (the entire) U_{wm} with a formula similar to (4). Once R_{wm} is known, Mallory can increase P_{sa} by choosing a point

[12] Encoding ultimately occurs by altering a set of numbers within MSE types of constraints.

[13] Although probably less likely to succeed given these very constraints often hard to accommodate. For example most watermarking methods in the frequency domain [13] can not directly handle a constraint such as the health-care example above.

only within $U_{attack} \setminus U_{wm}$ which is now clearly specified. This increase to P_{sa} in (3) could be quantitatively significant, in our example, by a factor of $\frac{R_a^2}{R_a^2+R_{wm}^2}$.

5 Conclusions

We discovered and analyzed inherent limitations of watermarking. We introduced watermarking vulnerability, a quantifiable measure of safety with respect to un-informed attacks. We exemplified this theory with our work on watermarking relational databases. We proposed extensions to more generic frameworks. We recommended design choices meant to increase the safety of watermarking applications.

Further work is required in analyzing scenarios featuring concave and/or high dimensional usability domains. Various algorithms could be compared with respect to their vulnerability and improved according to the optimality principle.

Acknowledgements

We would like to thank Ton Kalker for an interesting insight into the high dimensionality issue.

References

[1] M.J. Atallah, V. Raskin, C. F. Hempelmann, M. Karahan, R. Sion, K. E. Triezenberg, and U. Topkara. Natural language watermarking and tamperproofing. In *Lecture Notes in Computer Science, Proc. 5th International Information Hiding Workshop 2002*. Springer Verlag, 2002.

[2] M.J. Atallah, V. Raskin (with M. Crogan, C. Hempelmann, F. Kerschbaum, D. Mohamed, and S. Naik). Natural language watermarking: Design, analysis, and a proof-of-concept implementation. In *Lecture Notes in Computer Science, Proc. 4th International Information Hiding Workshop, Pittsburgh, Pennsylvania*. Springer Verlag, 2001.

[3] W. Bender, D. Gruhl, N. Morimoto, and A. Lu. Techniques for data hiding. *IBM Systems Journal*, 35(3–4):313–336, 1996.

[4] Dan Boneh and James Shaw. Collusion-secure fingerprinting for digital data. *Lecture Notes in Computer Science*, 963:452–??, 1995.

[5] B. Chen and G. Wornell. An information–theoretic approach to the design of robust digital watermarking systems. In *Proc. Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP), Phoenix, AZ*, 1999.

[6] G. Cohen, S. Encheva, and G. Zemor. Copyright protection for digital data. In *Workshop on Watermarking and Copyright Enforcement*, Paris France, 1999.

[7] Christian Collberg and Clark Thomborson. Software watermarking: Models and dynamic embeddings. In *Principles of Programming Languages*, San Antonio, TX, January 1999.

[8] Ingemar Cox, Jeffrey Bloom, and Matthew Miller. Digital watermarking. In *Digital Watermarking*. Morgan Kaufmann, 2001.

[9] Ingemar J. Cox and Jean-Paul M. G. Linnartz. Public watermarks and resistance to tampering. In *International Conference on Image Processing (ICIP'97)*, Santa Barbara, California, U.S.A., 26–29 October 1997. IEEE.

[10] Scott Craver. On public-key steganography in the presence of an active warden. In *Information Hiding*, pages 355–368, 1998.

[11] Edward J. Delp. Watermarking: Who cares? does it work? In Jana Dittmann, Petra Wohlmacher, Patrick Horster, and Ralf Steinmetz, editors, *Multimedia and Security – Workshop at ACM Multimedia'98*, volume 41 of *GMD Report*, pages 123–137, Bristol, United Kingdom, September 1998. ACM, GMD – Forschungszentrum Informationstechnik GmbH, Darmstadt, Germany.

[12] Stefan Katzenbeisser (editor) and Fabien Petitcolas (editor). Information hiding techniques for steganography and digital watermarking. In *Information Hiding Techniques for Steganography and Digital Watermarking*. Artech House, 2001.

[13] Bob Ellis. Public policy: New on-line surveys: Digital watermarking. *Computer Graphics*, 33(1):39–39, February 1999.

[14] M. Kobayashi. Digital watermarking: Historical roots. IBM Research Report RT0199, IBM Japan, Tokyo, Japan, April 1997.

[15] P. Moulin and J. O'Sullivan. Information-theoretic analysis of information hiding. In *manuscript*, 1999.

[16] Pierre Moulin, M. K. Mihcak, and Gen-Iu (Alan) Lin. An information–theoretic model for image watermarking and data hiding. In *(manuscript)*, 2000.

[17] J. Palsberg, S. Krishnaswamy, M. Kwon, D. Ma, Q. Shao, and Y. Zhang. Experience with software watermarking. In *Proceedings of ACSAC, 16th Annual Computer Security Applications Conference*, pages 308–316, 2000.

[18] Fabien A. P. Petitcolas, Ross J. Anderson, and Markus G. Kuhn. Attacks on copyright marking systems. In David Aucsmith, editor, *Information Hiding: Second International Workshop*, volume 1525 of *Lecture Notes in Computer Science*, pages 218–238, Portland, Oregon, U.S.A., 1998. Springer-Verlag.

[19] Fabien A. P. Petitcolas, Ross J. Anderson, and Markus G. Kuhn. Information hiding - a survey. *Proceedings of the IEEE*, 87(7):1062–1078, July 1999. Special issue on protection of multimedia content.

[20] B. Pfitzmann. Information hiding terminology. In Ross Anderson, editor, *Information hiding: first international workshop, Cambridge, U.K., May 30–June 1, 1996: proceedings*, volume 1174 of *Lecture Notes in Computer Science*, pages 347–350. Springer-Verlag, 1996.

[21] Claude Shannon. Communication theory of secrecy systems. *Bell System Technical Journal*, 28:656–715, 1949.

[22] Radu Sion. Analog output-stage fingerprinting in DMAT/SDMI. In *Watermarking 2002, CERIAS TR 2001-53*, 2002.

[23] Radu Sion, Mikhail Atallah, and Sunil Prabhakar. On watermarking numeric sets. In *Proceedings of IWDW 2002, Lecture Notes in Computer Science, CERIAS TR 2001-60*. Springer-Verlag, 2002.

[24] Radu Sion, Mikhail Atallah, and Sunil Prabhakar. On watermarking semistructures. In *(submission for review), CERIAS TR 2001-54*, 2002.

[25] Radu Sion, Mikhail Atallah, and Sunil Prabhakar. Power: Metrics for evaluating watermarking algorithms. In *Proceedings of IEEE ITCC 2002, CERIAS TR 2001-55*. IEEE Computer Society Press, 2002.

[26] Radu Sion, Mikhail Attalah, and Sunil Prabhakar. Key pre-commitment in watermarking. In *(submission for review), CERIAS TR 2002-30*, 2002.

[27] Radu Sion, Mikhail Attalah, and Sunil Prabhakar. Rights protection for relational data. In *Proceedings of ACM SIGMOD*, 2003.

[28] G. Voyatzis, N. Nikolaidis, and I. Pitas. Digital watermarking: an overview. In S. Theodoridis et al., editors, *Signal processing IX, theories and applications: proceedings of Eusipco-98, Ninth European Signal Processing Conference*, pages 9–12, Patras, Greece, 1998. Typorama Editions.

[29] Jian Zhao and Eckhard Koch. A generic digital watermarking model. *Computers and Graphics*, 22(4):397–403, August 1998.