

1 INTRODUCTION

Information, as an expression of knowledge, is probably the most valuable asset of humanity today. By enabling relatively cost-free, fast, and accurate access channels to information in digital form, computers have radically changed the way we think and express ideas. As increasing amounts of it are produced, packaged and delivered in digital form in a fast, networked environment, one of its main features threatens to become its worst enemy: zero-cost verbatim copies. The inherent ability to produce duplicates of digital Works at virtually no cost can now be misused e.g., for illicit profit (see Figure 1.2). This dramatically increases the requirement for effective rights assessment and protection mechanisms.

Different avenues are available, each with its advantages and drawbacks. Enforcement by legal means is usually ineffective, unless augmented by a digital counterpart such as Information Hiding. *Digital Watermarking* as a method of Rights Assessment deploys Information Hiding to conceal an indelible “rights witness” (“rights signature”, watermark) within the digital Work to be protected (see Figure 1.1). The soundness of such a method relies on the assumption that altering the Work in the process of hiding the mark does not destroy the value of the Work, and that it is difficult for a malicious adversary (“Mallory”) to remove or alter the mark beyond detection without destroying the value of the Work. The ability to resist attacks from such an adversary (mostly aiming at removing the embedded watermark) is one of the major concerns in the design of a sound watermarking solution.

There exists a multitude of semantic frameworks for discrete information processing and distribution. Each distinct data domain would benefit from the availability of a suitable watermarking solution. With the notable exception of software watermarking [1], the overwhelming majority of research efforts [2] [3] have been invested in the frameworks of signal processing and multimedia Works (e.g., images, video

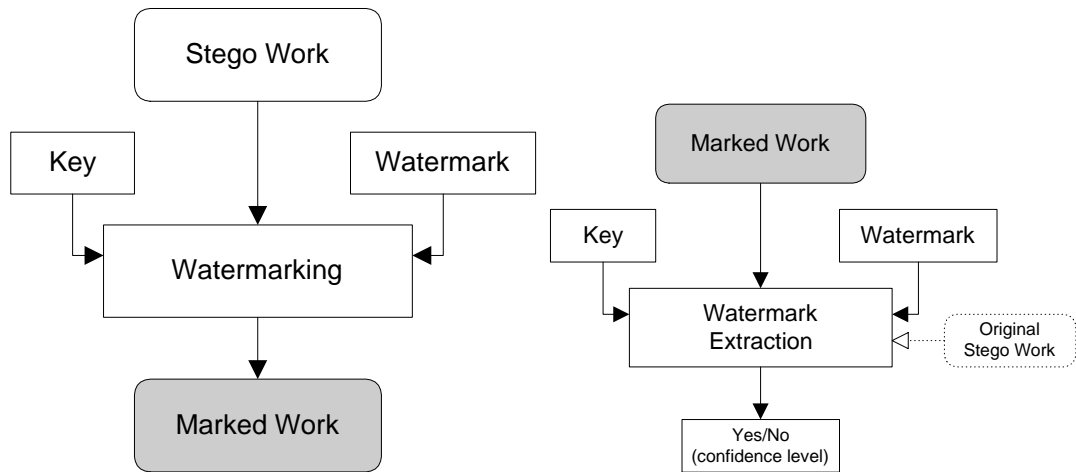


Figure 1.1. Introduction: (a) *Digital Watermarking* conceals an indelible “rights witness” (“rights signature”, watermark) within the digital Work to be protected. (b) In court, a detection process is deployed to prove the existence of this “witness” beyond reasonable doubt (confidence level) and thus assess ownership.

and audio). In this dissertation, we analyze information hiding as a rights assessment tool for discrete digital data types such as relational and time-series data. In this framework we propose a theoretical model and ask: are there any limitations to what watermarking can do? What are these and when can they be reached? (Chapters 2 and 7) We then design and analyze watermarking solutions for (i) numeric sets and relational data (Chapter 3), (ii) categorical data (Chapter 4), (iii) discrete sensor streams (Chapter 5) and (iv) semi-structures (Chapter 6).

1.1 Deployment Scenario

How does the ability to prove rights in court relate to our final desiderata, namely to *protect* those rights? Why not simply publish a digest of the Works to be protected in a newspaper, just before releasing them, enabling us to prove later on in court that at least they were in our possession at the time of publication. In the following we address these and other related issues.

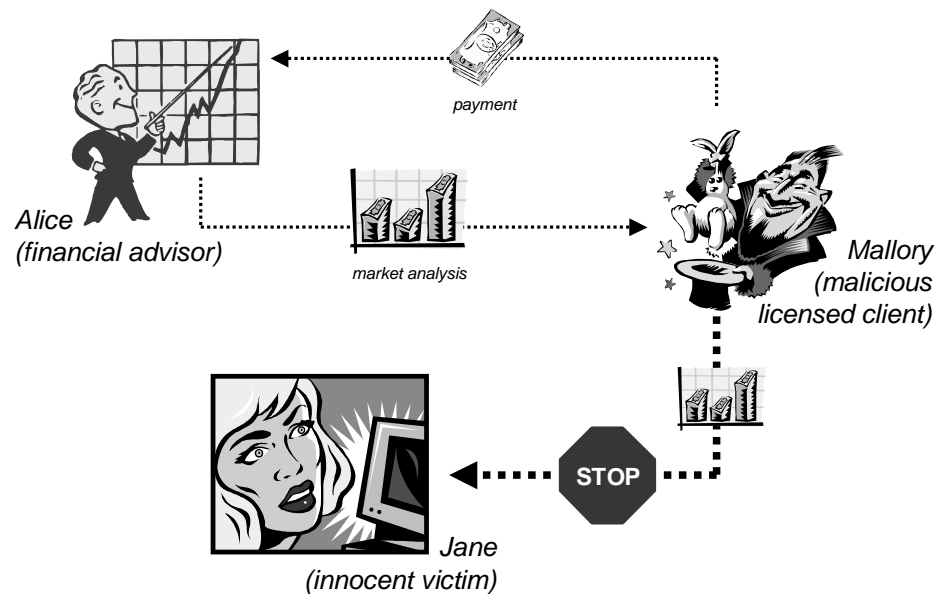


Figure 1.2. Introduction: Rights Assessment is useful when valuable content is to be sold/outsourced to potentially un-trusted parties, even if rightfully licensed.

1.1.1 Rights Protection through Assessment

The ability to prove/assess rights convincingly in court constitutes a deterrent to Mallory. It thus becomes a tool for rights protection if counter-incentives and legal consequences are set high enough. But because information hiding does not provide means of actual access control, the question of rights protection still remains. *How* are rights protected here?

It is intuitive that such a method would only work if the rightful rights-holder (Alice) actually knows about Mallory’s misbehavior **and** is able to prove to the court that: (i) Mallory possesses a certain Work X and (ii) X contains a “convincing” (e.g., very rare with respect to the space of all considered similar Works) and “relevant” (e.g., a string stating “(c) by Alice”) watermark.

What watermarking does not offer is a direct deterrent. If Alice does not have the knowledge of Mallory's illicit possession of the Work and/or if it is impossible to actually prove this possession in court beyond reasonable doubt, then watermarking cannot be deployed directly to prevent Mallory.

If, however, Information Hiding is aided by additional access control level levers, it can become very effective.

For example, if in order to derive value from the given Work (e.g., watch a video tape), Mallory has to deploy a known mechanism (e.g., use video player), information hiding could be deployed to enable such a proof of possession, as follows. One simple example would involve modifying the video player so as to detect the existence of a watermark and match it with a set of credentials and/or "viewing tickets" (that can be purchased) associated with the player's owner. If no match is found, the tape is simply not played back.

This is just one of many scenarios where watermarking can be deployed in conjunction with other technologies to aid in managing and protecting digital rights. Of course this scenario is simplistic and relies on the assumption that the cost of reverse engineering this process is far higher than the potential derived illicit gain. However this is essential in that it illustrates the game theoretic nature at the heart of the watermarking proposition and of information security in general.

Another example application of Resilient Information Hiding as a tool aiding rights management, would be its deployment to "track" license violators by hiding a specific mark inside the Work, a mark that uniquely identifies the party it was sold/outsourced to (fingerprinting). If the Work would then be found in the public domain, that mark could be used to assess the source of the leak (see Figure 1.3).

Watermarking is a game with two adversaries, Mallory and Alice. At stake lies the value inherent in a certain Work X , over which Alice owns certain rights. When Alice releases X she deploys watermarking for the purpose of ensuring that one of the following holds:

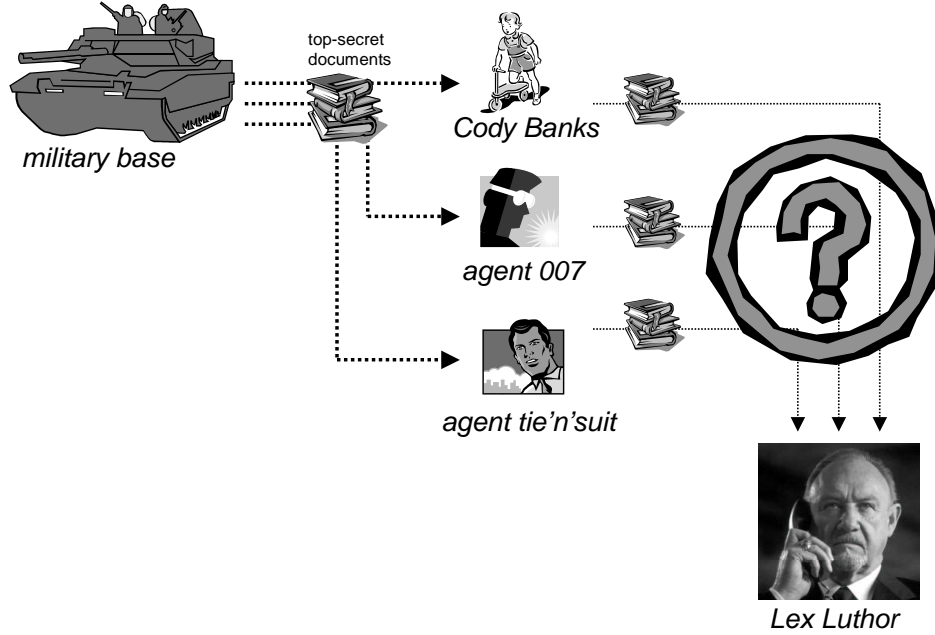


Figure 1.3. Introduction: A scenario where resilient information hiding for fingerprinting might reveal which secret agent leaked secret documents to Lex Luthor.

- she can always prove rights in court over any copy or valuable derivate of X (e.g., segment)
- any existing derivate Y of X , for which she cannot prove rights, does not preserve any significant value (derived from the value in X)
- the cost to produce such an un-watermarked (for which she cannot prove rights) derivate Y of X that is still valuable (with respect to X) is higher than its value

1.1.2 Information Hiding vs. Newspaper Digests

Apparently Alice could simply publish a (e.g., cryptographic) digest of X in a newspaper, thus being able to at least claim a time stamp of possession of X

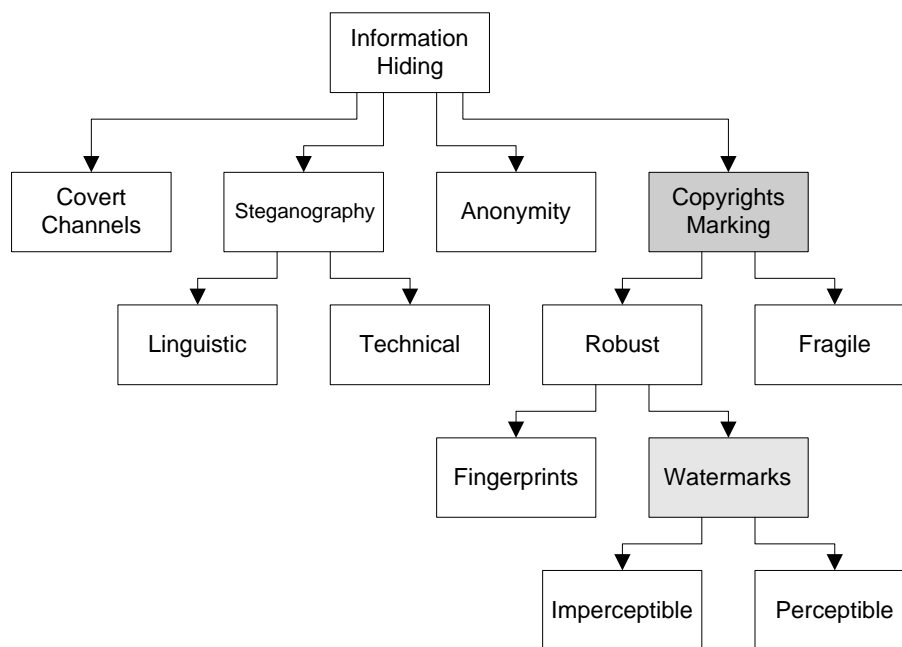


Figure 1.4. Introduction: Information Hiding classification according to Petitcolas et al [4]

later on. Why not deploy this as a rights assessment tool instead of information hiding? There are many reasons why it would not work, including (i) scalability issues associated with the need for a trusted third party (newspaper), (ii) the cost of publishing a digest for each released Work, (iii) scenarios when the fact that the Work is watermarked should be kept secret (stealthiness) etc.

Maybe the most important reason is that Mallory can now claim that his ownership of the Work precedes X 's publication date, and that Alice simply (modified it and) published a digest. It would then be up to the court to decide if Mallory is to be believed or not, hardly an encouraging scenario for Alice. This could work if there existed a mechanism for the mandatory publication of digests for each and every valuable Work, again probably impractical due to both costs and lack of scalability.

It becomes clear that deploying such aids as rights assessment tools makes sense only in the case of the Work being of value only un-modified. In other words if it does not tolerate any changes (without losing its value) and Mallory is caught

in possession of an identical copy, Alice can successfully prove in court that she possessed the original at the time of its publication, but she cannot prove more.

Now, considering that, in the case of watermarking, the assumption is that, no matter how small, there are modifications allowed to the Works to be protected, in some sense the two approaches complement each other. If no modifications are allowed, then a third-party “newspaper” service might work for providing a time-stamp type of ownership proof that can be used in court.

1.2 Watermarking vs. Watermarking

In existing research, the term “watermarking” denotes the use of information hiding techniques to (also) assess digital rights, overwhelmingly focused in the broader frameworks of signal processing and multimedia Works.

In this dissertation we (appropriately) re-use this term, to denote the same concept of using Information Hiding to provide proofs of rights. However this brings about the question of the specifics of the relationship between the actual research challenges and techniques deployed in both frameworks. Because, while the terms might be identical, the associated model, challenges and techniques are different, almost orthogonal: whereas in the signal processing case there usually exists a large noise bandwidth, due to the fact that the final data consumer is likely human (with associated limitations of the sensory system), in the case of discrete data types this cannot be assumed and data quality assessment needs to be closely tied with the actual watermarking process (see Section 2.2).

Another important differentiating focus in our research is the emphasis on the actual ability to convince in court as a success metric, unlike most approaches in the signal processing realm, that centered on bandwidth. We believe that, while bandwidth is a relevant related metric, it does not consider important additional issues such as malicious transforms and removal attacks. We are not as concerned with packing a lot of rights assessment information (i.e., watermark bits) in the

Works to be protected, as we are concerned with being able to both survive removal attacks and convince in court. We explore this more in Chapter 7 (and [5], [6]).

Maybe the most important difference between the two domains is that, while in a majority of watermarking solutions in the multimedia framework, the main domain transforms are signal processing primitives (e.g., Works are mainly considered as being compositions of signals rather than strings of bits), in our case data types are mostly discrete and are not naturally handled as continuous signals. Additionally, while (for example) discrete versions of frequency transforms can be deployed as primitives in information encoding for digital images [2], the basis for doing so is the fact that, although digitized (thus in discrete format), images are at the core defined by a composition of light reflection signals and are consumed as such (by the final human consumer). By contrast, arbitrary discrete data (e.g., categorical data) is naturally discrete ¹ and often to be ingested by a highly sensitive processing component (e.g., a computer rather than a perceptual system tolerant of distortions).

Thus, while the term “watermarking” will be used throughout this dissertation to denote the process of deploying information hiding for the purpose of rights assessment, in terms of actual models, challenges and techniques, it is to be distinguished from its use in the broader domain of signal processing and multimedia. And while similarities are always to be found (e.g., “Gaussian noise addition” in the multimedia case equates to a “random un-informed attack” in the discrete data case) we do not believe that “everything is a signal” for the purpose of rights assessment. Comparing efforts in the two domains can often result in comparing apples to oranges.

1.3 Summary of Contributions

The main contributions of this dissertation include: a theoretical model for rights assessment through information hiding for discrete digital data (Chapters 2 and 7), the design and analysis of watermarking solutions for numeric sets and relational

¹Unless we consider quantum states and uncertainty arising in the spin of the electrons flowing through the silicon.

data (Chapter 3), categorical data (Chapter 4), discrete sensor streams (Chapter 5) and semi-structures (Chapter 6).

1.3.1 Model

In Chapter 2 (and [5], [6]) we introduce a model for watermarking. We define fundamental concepts including: *usability domain* - a set of functionals quantifying a digital Work's value in terms of its specific use; *watermark* - an induced property of a watermarked Work O' , so rare, that if we consider any other Work O'' , "close-enough" to the original Work O , the probability that O'' exhibits the same property can be upper-bounded; watermark *vulnerability* - the ability of an attack to succeed against a watermarking scheme. One fundamental difference between watermarking and generic data hiding 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 in a manner as resilient and stealthy as possible, through a medium controlled by evil Mallory. Digital watermarking is deployed in court by Alice to prove rights over a given Work, usually in a scenario where Mallory benefits from using/selling that very same Work 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 be present (i.e., very unlikely to arise fortuitously). This defines a main difference from steganography: for its purpose, the 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. Thus, in watermarking the emphasis is on “detection” rather than “extraction”. Extraction of a watermark is usually a part of the detection but just complements the process up to the extent of increasing the ability to convince in court.

1.3.2 Numeric Relational Data

In Chapter 3 (and [7], [8], [9]) we introduce a solution for relational database content rights protection through watermarking. Rights protection for relational data is of ever increasing interest, especially considering areas where sensitive, valuable content is to be outsourced. A good example is a data mining application, where data is sold in pieces to parties specialized in mining it. Our solution addresses important attacks, such as subset selection, linear data changes, random alteration attacks, and data loss. We introduce `wmdb.*`, a proof-of-concept implementation and its application to real-life data, namely in watermarking the outsourced Wal-Mart sales data available at our institute.

The main challenges in this new domain derive from the fact that, since the associated data types do not have fixed, well defined semantics (as compared to multimedia) and may be designed for machine ingestion, identifying the available “bandwidth” for watermarking becomes as important as the actual encoding algorithms. Remember that one of the desiderata of watermarking is to insert an indelible mark in the object such that the insertion of the mark does not destroy the value of the object. Clearly, the notion of value or utility of the object is central to the watermarking process. This is closely related to the type of data and its intended use. For example, in the case of software the value may be in ensuring equivalent computation, and for text it may be in conveying the same meaning (i.e., synonym substitution is acceptable). Similarly, for a collection of numbers, the utility of the data may lie in the actual or the relative values of the numbers, or in their distribution (e.g., normal with a certain mean and variance). Because, one can always

identify some use of the data that would be affected by even a minor change to any portion of it, it becomes necessary that the intended purpose of the data to be preserved is identified and integrated in the watermarking process.

Our solution starts by receiving as user input a reference to the relational data to be rights-protected, a watermark to be embedded as a copyright proof, a secret key used to protect the embedding, and a set of data quality constraints to be preserved in the result. It then proceeds to watermark the data while continuously assessing data quality, potentially backtracking and undoing undesirable alterations that do not preserve data quality. Watermark embedding consists of two main parts: in the first stage, the input data set is securely partitioned into subsets of items; the second stage then encodes one bit of the watermark into each subset. If more subsets (than watermark bits) are available, error correction is deployed to result in an increasingly resilient embedding. The algorithms prove to be resilient to important classes of attacks, including subset selection, linear data changes, and random alterations.

The system design, (including the mechanisms evaluating data quality constraints through plugins), is outlined in Figure 1.5 (a). To exemplify the resilience of the method (e.g., to random alterations), in Figure 1.5 (b), a comparison is made between the case of uniformly distributed (i.e., values are altered randomly between 100% and 120% of their original value) and fixed alterations (i.e., values are increased by exactly 20%). In the case of fixed alterations the behavior demonstrates the self-healing ability of our method: as more and more of the tuples (past the 50% mark) are altered linearly, the watermark distortion decreases. When over 95% of the data is modified consistently and linearly, the watermark suffers only 7% alterations.

Another important experiment analyzes the ability to preserve classes in the resulting watermarked Work. Classification is extremely relevant in areas such as data mining, and we envision that many of the actual deployment scenarios for our relational watermarking application will require classification preservation. Classification preservation deals with the problem of propagation of the classes occurring in the original (input) data in the watermarked (output) version of the data. It

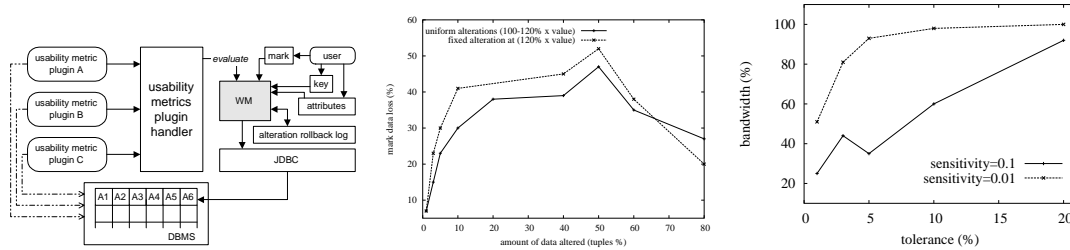


Figure 1.5. Introduction: Relational Data with Numeric Types – (a) The **wmdb.*** package. (b) Random attack (non-zero average) on a normally distributed data set. (c) Impact of classification preservation on the available watermarking bandwidth.

provides thus the assurance that the watermarked version still contains most (or within a certain allowed percentage) of the original classes. Figure 1.5 (c) depicts how classification can be preserved while making optimal use of the available bandwidth. For example, up to 90% of the underlying bandwidth can become available for watermark encoding with a restrictive 6% classification preservation goodness.

These results confirm the adaptability of our watermarking algorithm. As classification tolerance is increased, the application adapts and makes use of an increased available bandwidth for watermark encoding. This also shows that classification preservation is compatible with our distribution-based encoding method, an important point to be made, considering the wide range of data-mining applications that could naturally benefit from watermarking ability.

The main contributions in this chapter include: (i) a resilient watermarking method for numeric relational data, (ii) a technique for enabling user-level run-time control over properties that are to be preserved as well as the degree of change introduced, (iii) a complete, user-friendly implementation for numeric relational data, and (iv) the deployment of the implementation on real data, in watermarking the Wal-Mart Sales Database and the analysis thereof.

1.3.3 Categorical Data

In Chapter 4 (and [10], [11]) we introduce a novel method of watermarking relational data with categorical types. We discover new watermark embedding channels and design novel watermark encoding algorithms. We analyze important theoretical bounds including mark vulnerability. While fully preserving data quality requirements, our solution survives important attacks, such as subset selection and random alterations. Mark detection is fully “blind” in that it does not require the original data, an important characteristic especially in the case of massive data. We propose various improvements and alternative encoding methods. We perform validation experiments by watermarking the outsourced Wal-Mart sales data available at our institute. We prove (experimentally and by analysis) our solution to be extremely resilient to both alteration and data loss attacks, for example tolerating up to 80% data loss with a watermark alteration of only 25%.

Important new challenges are associated with this domain. One cannot rely on “small” alterations to the data in the embedding process. Any alteration may be significant. The discrete characteristics of the data require discovery of fundamentally new bandwidth channels and associated encoding algorithms. Our method proves to be resilient to important attacks, including subset selection and random alterations.

Our solution starts by discovering two domain-specific watermark embedding channels, namely (i) the *inter-attribute associations* and (ii) the *value occurrence frequency-transform* (attribute frequency histogram). Next, embedding methods to resiliently hide information in these channels are designed. The main method starts with an initial user-level assessment step in which a set of attributes to be watermarked are selected. Next, watermark encoding proceeds for each attribute pair (K, A) in the considered attribute set, by selecting a subset of “fit” tuples (determined directly by the association between A and K). These tuples are then considered for mark encoding. Mark encoding alters the tuple’s value according to secret criteria that induces a statistical bias in the distribution for that tuple’s altered

value. The mark decoding process relies on discovering this induced statistical bias. Yet another embedding method is available to counter extreme vertical partitioning attacks in which only a single attribute A is preserved in the result. If, intuitively, for massive data sets, the number of possible discrete values for A is much smaller than the data set size, then A contains many duplicate values. There is probably very little value associated with knowing the set of possible values of A . The main value in this scenario (in Mallory’s eyes) is (arguably) to be found in one of the only remaining characteristic properties, namely the value occurrence frequency distribution for each possible value of A . If we could devise an alternative watermark encoding method for this set then we would be able to associate rights also to this aspect of the data, thus surviving this extreme partitioning attack. In Chapter 3 we introduce a watermarking method for numeric sets that is able to minimize the absolute data alteration in terms of distance from the original data set. We propose to apply this method here to embed a mark in the occurrence frequency distribution domain. One concern we should consider is the fact that in the categorical domain we are usually interested in minimizing the *number* of data items altered whereas in the numeric domain we aim to minimize the absolute data change. It is fortunate that, because we now have numeric values modeling occurrence frequency, a solution minimizing absolute data change in this (frequency) domain naturally minimizes the *number* of items altered in the categorical value domain.

The experimental results include an analysis of the relationship between the amount of alterations required in the watermarking phase and a minimum guaranteed watermark resilience. It can be seen in Figure 1.6 (a) that with a decreasing number of encoding alterations (decreasing e) the vulnerability to random alteration attacks increases accordingly. This illustrates the trade-off between the requirement to be resilient and the preservation of data quality (e.g., fewer alterations). An experiment analyzing resilience to data loss is depicted in Figure 1.6 (b). We observe here the compensating effect of error correction. Compared to data alteration attacks, the watermark survives even better with respect to attack size (in this case data loss).

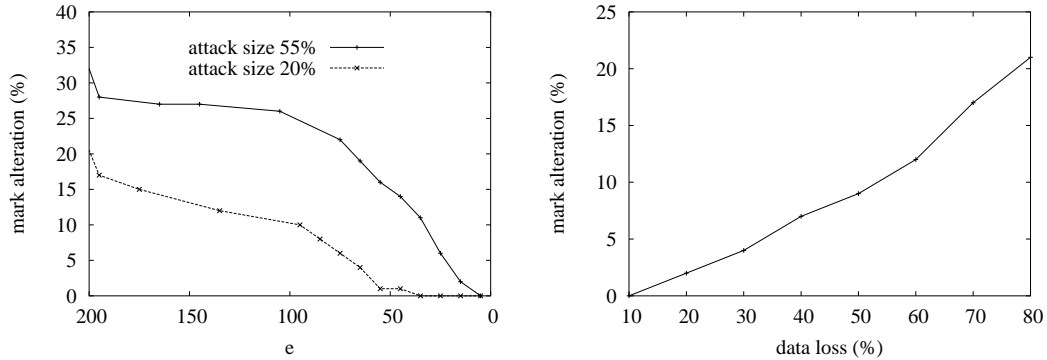


Figure 1.6. Introduction: Relational Data with Categorical Types
– (a) More available bandwidth (decreasing e) results in a higher attack resilience. (b) The watermark degrades almost linearly with increasing data loss.

The main contributions of this effort include: (i) the proposal and definition of the problem of watermarking categorical data, (ii) the discovery and analysis of new associated watermark embedding channels (iii) the design of novel encoding algorithms and (iv) their experimental analysis.

1.3.4 Sensor Streams

Today’s world of increasingly dynamic computing environments naturally results in more and more data being available as fast streams. Applications such as stock market analysis, environmental sensing, web clicks and intrusion detection are just a few of the examples where valuable data is streamed to its consumer. Often, streaming information is available on the basis of a non-exclusive, single-use customer license. One major concern, especially given the digital nature of the valuable stream, is the ability to easily record and potentially “re-play” parts of it in the future. If there is value associated with such future re-plays, it could constitute enough incentive for a malicious customer (Mallory) to record and duplicate segments of

data, subsequently re-selling them for profit. Being able to protect against such infringements becomes a necessity.

In Chapter 5 (and [12]) we introduce the issue of rights protection for streaming discrete (sensor) data through watermarking. This is a novel problem with many associated challenges including: the inability to perform multiple-pass random accesses to the entire data set, the requirement to be fast enough to keep up with the incoming stream rate, to survive instances of extreme sparse sampling and summarizations, while at the same time keeping data alterations within allowable bounds. We propose a solution and analyze its resilience to various types of attacks as well as expected domain-specific alterations, such as sampling and summarization. We implement a proof of concept software (wms.*) and perform experiments to assess these resilience levels in practice. Our method proves to be well suited for this new domain. For example, we can recover an over 97% confidence watermark from a sampled (e.g., less than 8%) stream. Similarly, our encoding ensures survival to stream summarization (e.g., 20%) and random alteration attacks with very high confidence levels, often above 99%.

A set of novel challenges present themselves in this domain. Any stream processing performed is necessarily both time and space bound. The time bounds derive from the fact that the processing has to keep up with incoming data. The space bounds are referring to the finiteness of any storage mechanism, when compared with the virtually infinite nature of streaming data. At the same time, any quality preservation constraints can be formulated only in terms of the current available data window; including any history information will come at the expense of being unable to store as much new incoming data. Moreover, the effectiveness of any rights protection method is directly related to its ability to survive legitimate domain specific transformations as well as malicious attacks. In this framework we deal with the following: (A1) summarization, (A2) sampling, (A3) segmentation (we would like to be able to recover a watermark from a finite segment of data drawn from the stream), (A4) scaling (there might be value in actual *data trends*, that Mallory could

still exploit, by scaling the initial values), (A5) addition of stream values and (A6) random alterations.

At an overview level, watermark embedding proceeds as follows: (a) first a set of “major” data extremes (actual stream max/min values) are identified in the data stream, extremes that feature the property that they (or a majority thereof) can be recovered after a suite of considered alterations (possibly attacks) such as (random) sampling and summarization. Next (b) a certain criteria is used to select some of these extremes as recipients for parts of the watermark. Finally (c), the selected ones are used to define subsets of items considered for 1-bit watermark embedding of bits of the global watermark. The fact that these extremes can be recovered ensures a consistent overlap (or even complete identity) between the recovered subsets and the original ones (in the un-altered data). In the watermark detection process (d) *all* the extremes in the stream are identified and the selection criteria in step (b) above is used once again to identify potential watermark recipients. For each selected extreme, (e) its corresponding 1-bit watermark is extracted and ultimately the global watermark is gradually re-constructed, by possibly also using an error correction mechanism. In summary, one of the main ideas behind our solution is the use of extreme values in the stream’s evolution as watermark bit-carriers. The intuition here lies in the fact that much of the stream value lies in its fluctuating behavior (and associated extremes), more likely to survive value-preserving, domain-specific transforms.

We performed experiments on watermark survival to a variety of transformations, including random alterations and combined sampling and summarization. In Figure 1.7 (a), random alterations are illustrated. Naturally, an increasing level of distortion results in decreasing detection. Nevertheless, for 50% of the data altered within 10% of the original value, we still detect a watermark bias of roughly 25 bits, yielding a very convincing false-positive rate of less than “one in thirty million”. In Figure 1.7 (b) we outline the impact of a *combined* transformation (sampling and summarization) on the watermark embedding. Because of the nature of both

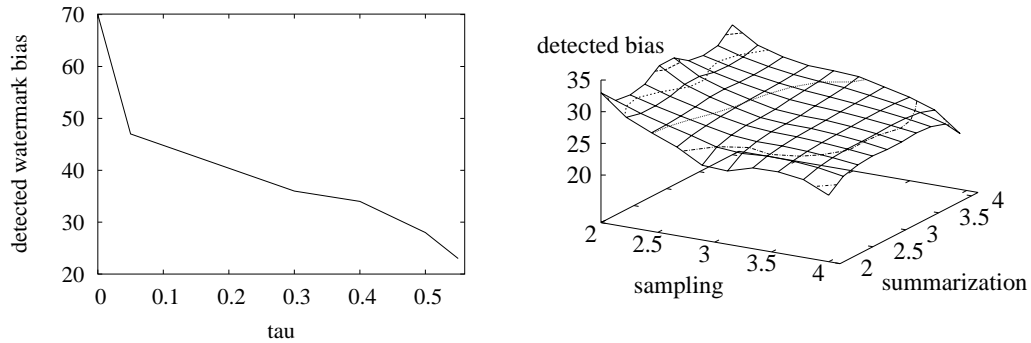


Figure 1.7. Introduction: Discrete Streaming Data – (a) Watermark survival to epsilon-attacks. (b) Watermark survival to combined sampling and summarization.

transformations and of the resilience featured in each case, the combination seems to be survived well. For example, 25% sampling, followed by 25% summarization still yields a watermark bias of up to 20, corresponding to a favorable, low false-positive rate of “one in a million”.

The main contributions in this chapter include: (i) the proposal and definition of the problem of watermarking discrete sensor streams, (ii) the discovery and analysis of new watermark embedding channels for such data, (iii) the design of novel associated encoding algorithms, (iv) a proof of concept implementation of the algorithms and (v) their experimental evaluation. The algorithms introduced here prove to be resilient to important domain-specific classes of attacks, including stream re-sampling, summarization (replacing a stream portion by its average value) and random changes.

1.3.5 Abstract Structures

While most existing work on watermarking is about specific kinds of media (in fact most papers were about image watermarking), in this part of our research, rather than dealing with single media at a time (image, audio, video), we explore a more

general notion of a modern document. Generalized documents are aggregates of multiple types of content – a document is a structured aggregation of many types of data, and the information content of the document is as much in the structure (the graph) as in the nodes (that contain particular data types). The fact that nodes in semi-structures are value-carrying, means that a watermarking algorithm can make use of their encoding capacity by using traditional watermarking, but the graph that “glues” these together is another central element of the watermarking process. We propose to combine and use these two facets (structural and node-content) to provide a solution for the watermarking of aggregates.

In Chapter 6 (and [13]) we discuss the watermarking of abstract structured aggregates of multiple types of content, such as multi-type/media documents. These *semi-structures* can usually be represented as graphs and are characterized by value lying both in the structure *and* in the individual nodes. Example instances include XML documents, complex web content, workflow and planning descriptions. We propose a scheme for watermarking abstract semi-structures and discuss its resilience with respect to attacks. While content-specific watermarking deals with the issue of protecting the value in the structure’s nodes, protecting the value pertaining to the structure itself is a new, distinct challenge. Nodes in semi-structures are value-carrying, thus a watermarking algorithm could make use of their encoding capacity by using traditional watermarking. For example if a node contains an image then image watermarking algorithms can be deployed for that node to encode parts of the global watermark. Also, given the intrinsic value attached to it, the graph that “glues” these nodes together becomes in itself a central element of the watermarking process that makes use of these two value facets, structural and node-content.

Multiple challenges are encountered in this framework, mostly derived from the requirement to survive domain-specific transformations and likely attacks by Mallory, including: elimination of value-“insignificant” nodes (A1), elimination of inter-node relations (A2), value preserving graph partitioning into independent usable partitions (A3), modification of node content, within usability vicinity (A4), addition of value

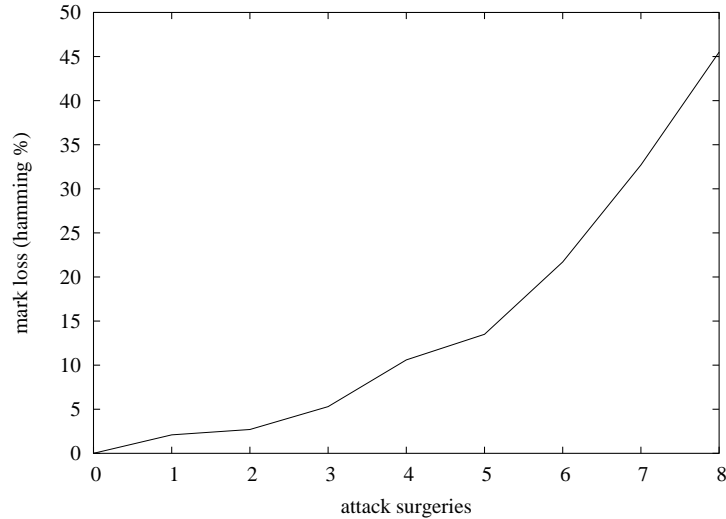


Figure 1.8. Introduction: Semi-structured Aggregates – Averaged watermark loss over 10 runs of an 8 bit watermark embedded into an arbitrary 32 node graph with 64 edges. Surgery attacks are applied randomly (node removals 60%, link addition 20%, link removal 20%). The labeling scheme was trained for 3 surgeries.

insignificant nodes (A5). Our solution is based on a canonical labeling algorithm that self-adjusts to the specifics of the content. Labeling is tolerant to a significant number of graph attacks (“surgeries”) and relies on a complex “training” phase at embedding time in which it reaches an optimal stability point with respect to these attacks. We perform attack experiments on the introduced algorithms under different conditions with very encouraging results. In Figure 1.8 we show the watermark behavior to data alteration in the case of a random artificially generated structure with 32 nodes and 64 edges. The embedded watermark is 8 bits long. The labeling scheme was trained for 3 surgeries. As the number of attack surgeries increases, the watermark degrades slightly. The results are averaged over 10 runs on the same graph with different random attacks. When 8 attack surgeries are applied to the graph we can still recover 60 – 65% of the watermark. One has to consider also the fact that an attacker is bound not to modify the structure beyond distortion limits.

1.3.6 Limits

The main desiderata and features of watermarking in a signal processing/multimedia framework, as outlined in [4] include: 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 Work should not interfere with each other; collusion attacks should not be possible; the watermark should survive any value-preserving transformation.

A common un-proved consensus has been implicitly assumed, namely that watermarking indeed lives up to its claimed features. [2, 3, 14] 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 [15]. 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” mainly for multimedia (i.e., images). Additional research [16–18] 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. 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 derive real-life recommendations for fine-tuning actual algorithms to increase their marking resilience.

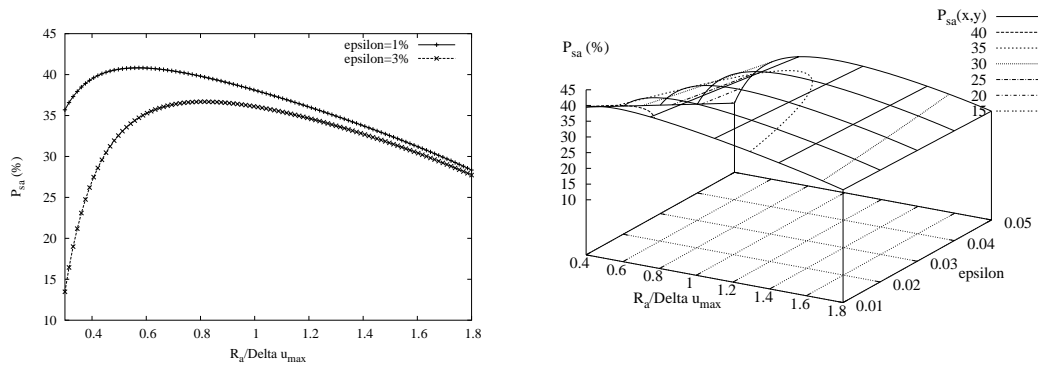


Figure 1.9. Introduction: Model of Watermarking – (a) No matter how sophisticated the watermarking method, there exists a random attack with a success probability of 33% and above (although we might not know what the attack is). It can be seen that a more court convincing ϵ_w value yields an even higher upper bound on attack success probability (2D cut through (b)). (b) The 3D evolution of the probability of a successful attack.

In Chapter 7 (and [6]) we explore these and other issues in the broader dissertation framework, for a broad class of watermarking algorithms. We use the previously introduced model to assess watermarking resilience and 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 algorithmic class framework, namely our research in watermarking numeric relational data (see Chapter 3). We discover that indeed there exist such limitations. More specifically, we identify an important *convince-ability trade-off*: the more “convincing” in court a watermarking method is, the higher the probability of success of a perfect attack. We further derive the *watermarking optimality principle* that states that the vulnerability of a watermarking scheme (in our considered class) is likely minimized when it yields watermarked results on the boundary of the maximum allowable usability vicinity of the original un-watermarked Works.

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. This is an inherent limitation of watermarking in general. Any additional knowledge can only improve on this probability. This is the case even if these algorithms conform to the optimality principle. Also, there seems to exist a “sweet spot” in which the probability of a successful attack is maximized. Mallory could make use of this by fine-tuning.

In summary, in this chapter we identify and analyze inherent limitations of watermarking, including the trade-off between two important watermarking properties: *being suitably “convincing” in court* while at the same time *surviving a set of attacks*. 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 thus becomes necessarily characterized by a significant non-zero probability of being successfully attacked. We discovered an optimality principle (quantified and proved for a broad class of algorithms) that postulates the minimization of vulnerability in specific data points.