A Novel Multiple Non Visible and Data Publishing with Provable Distance-Based Mining

K. Mowcika [#1], S.P.Santhoshkumar[#2], V.Anuradha[#3], G.Mariya Lavanya Sangeetha[#4]

#1UG Student, Department of CSE, Shree Sathyam College of Engg. and Tech, Sankari, India.
#2Assistant Professor, Department of CSE, Shree Sathyam College of Engg. and Tech, Sankari, India.
#3Assistant Professor, Department of CSE, Shree Sathyam College of Engg. and Tech., Sankari, India.
#4Associate Professor, Department of CSE, Shree Sathyam College of Engg. and Tech., Sankari, India.
[1]mowcika@gmail.com , [2]Spsanthoshkumar16@gmail.com,
[3]kavkamanu@gmail.com@gmail.com, [4]gsangit@gmail.com

***Abstract* -Data exchange and data publishing are becoming an inherent part of business and academic practices. Data owners, nonetheless, also need to maintain the principal rights over the datasets that they share, which in many cases have been obtained after expensive and laborious procedures. This project presents a right-protection mechanism that can provide detectable evidence for the legal ownership of a shared dataset, without compromising its usability under a wide range of machine learning, mining, and search operations. The algorithms also preserve important properties of the dataset, which are important for mining operations, and so guarantee both right protection and utility preservation. The project considers a right-protection scheme based on watermarking. Watermarking may distort the original distance graph. The proposed watermarking methodology preserves important distance relationships, such as: the Nearest Neighbors (NN) of each object and the Minimum Spanning Tree (MST) of the original dataset. It proves fundamental lower and upper bounds on the distance between objects post-watermarking. The application is designed using NET BEAN 6.8 as front end. The coding language used is Java 6.0. MS-SQL Server 2000 is used as back end database.**

***Keywords*—Watermarking, nearest neighbors (NN), minimum spanning tree (MST), restricted isometry property (RIP), Data Exchange.**

I. INTRODUCTION

DATA exchange and data publishing are becoming an inherent part of business and academic practices. Data owners, nonetheless, also need to maintain the principal rights over the datasets that they share, which in many cases have been obtained after expensive and laborious procedures. This work presents a right-protection mechanism that can provide detectable evidence for the legal owner- ship of a shared dataset, without compromising its usability under a wide range of machine learning, mining, and search operations. We accomplish this, by guaranteeing that order relations between object distances remain unaltered. To right protect we use watermarking. Watermarking allows the user to hide innocuous pieces of informa- tion inside the data. The value of watermarking becomes increasingly important because of the proliferation of digital content, and because of the ease of data sharing particularly through data clouds. However, traditional multimedia watermarking techniques considered only a single object and did not analyze distortions in the object relationships when deal- ing with watermarking multiple objects. Watermarking essentially adds noise to a given dataset, and so it may distort the original object distances. Therefore, our goal is two-fold: not only to provide right protection, but also to preserve important parts of the original object topology. We focus on the preservation of the following properties on the original distance graph: a) preservation of Nearest-Neighbor (NN) distances for every object, and b) preservation of the dataset's Minimum Spanning Tree (MST). By doing so, any mining or search task that depends on the previous properties will remain undistorted post-watermarking. There exist many mining and learning algorithms that depend on the objects' NN's or the MST of the dataset. Using the proposed right-protection methodology, the execution of these algorithms will remain the same before and after watermarking. We provide some concrete examples:

- Instance-based classifiers [2], [3] search for the NN's of a given query and assign the label based on a majority vote of the

neighborhood. Any classification scheme based on NN's will operate in the same way as on the original data.

- Many clustering algorithms utilize the MST, such as [4], [5]. Our technique also guarantees preservation of the MST post-watermarking.
- There exist many visualization methods and embedding techniques that use either the neighborhood or the MST. As one example, the popular ISOMAP dimensionality reduction technique [6] first creates the k-neighborhood of every object, and then projects the Minimum Spanning Tree distance relationships on a lower-dimensional space.

I. OVERVIEW OF OUR APPROACH

Our goal is to discover how to right-protect a dataset, but at the same time guarantee preservation of the outcome of important distance-based mining operations. We provide two variants: one that preserves Nearest-Neighbors (NN) and another that preserves the Minimum Spanning Tree (MST). Therefore, the output of any algorithm based on these two properties will be preserved after right protection. To guarantee this, we study the critical watermark intensity to both protect the dataset, as well as ensure that important parts of the object distance graph are not distorted. It is essential to discover the maximum watermark intensity for right protection. This provides assurances of better detestability and hence better security for the right protection scheme.

This gives us insight on how to design fast variants of our algorithms that still guarantee preservation of the NN and the MST, but operate significantly faster than the exhaustive algorithms. We demonstrate our findings primarily on image contour data from anthropology and the natural sciences. This is mainly for reasons of illustration, so that the effect of right protection can be more easily visualized. However, our approaches are applicable on any sequential numerical datasets (e.g., time series).

II. RIGHT PROTECTION VIA WATERMARKING

We describe first how watermarking mechanisms can embed a secret key (watermark) on a collection of objects. We demonstrate the techniques for 2D sequence data (image contours, trajectories, etc). We later demonstrate how to detect the watermark using a correlation filter. The embedded watermark should satisfy the following properties: 1) Detectable: the correlation distribution of the watermarked data with the correct key is sufficiently distinct from the distribution with a random key, thus allowing the conclusive determination of the watermark's presence; 2) Preservation of the NN and the MST: the power of the watermark is tuned in such a way so that the Nearest Neighbor of each object and the Minimum Spanning Tree of the distance graph of all objects does not change. 3) Robust to malicious attacks: the watermark is detectable even after data transformations (attacks). In the following sections we explain how the above requirements are satisfied by the proposed watermarking scheme.

a. Threat Model

We consider the following threat model: an attacker may modify the watermarked data so that the watermark cannot be detected. However the attacker may only modify the data to an extent such that the utility of the data is not sacrificed. We assume that an attacker: a) is knowledgeable of the algorithm but not of the secret key; b) may distort the data using geometric transformations, noise addition (in both time and frequency domains), data transcription (e.g., upsampling or downsampling); c) can also deploy other types of attacks, such as double-watermarking.

b. Watermark Embedding

A model can describe data trajectories or even image contour data which capture coordinates of a shape perimeter, as shown in Fig. 1. We use a spread-spectrum approach [8]. This embeds the watermark across multiple frequencies of each object and across multiple objects of the dataset. As such, it renders the removal of the watermark particularly difficult without substantially compromising the data utility.

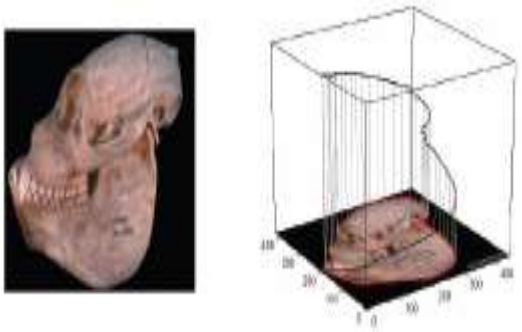

Fig1. Shape perimeter converted into 2D sequence

Definition 1:-(Multiplicative Watermark Embedding(W, p))

Assume a sequence x ∈ Cn with corresponding set of Fourier descriptors X, a watermark W ∈ Rn and power p ∈ [0, 1] which specifies the intensity of the watermark.

We can revert from the frequency domain back to the space domain and obtain the watermarked sequence using the inverse discrete Fourier transform. The robustness of the watermark embedding depends on the choice of coefficients. We embed the watermark in the coefficients that exhibit, on average over the dataset, the largest Fourier magnitudes. This makes the removal of the watermark difficult; in order for it to be masked out (e.g., by noise addition) it would mean that the dominant frequencies of the dataset have to be distorted. Thus, the dataset utility would have to be undermined. Fig. 2 shows the reconstruction of a shape from a dataset, when approximated using the highest energy coefficients. It is apparent that the high energy coefficients capture important characteristics of the dataset. When embedding the watermark, we exclude the first Fourier descriptor (the DC component) X1 from consideration, and leave it intact. The DC component captures the center of mass of object x and is therefore highly susceptible to translational attacks. For example, if a part of the watermark was embedded on the DC component of an object, then a simple translation would shift the center of mass of the object, thus erasing this part of the watermark without affecting the object's shape.

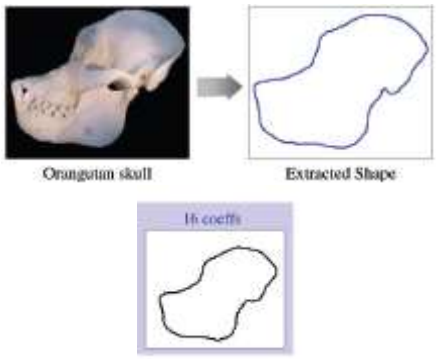

Fig 2. : Shape reconstruction for different number of Fourier coefficients that contain the highest energy.

Definition 2:- (Class of watermarks with l nonzero elements, compatible with dataset D (W_l(D))).

The class of watermarks with l non-zero elements, compatible Fig. 2. Shape reconstruction for different number of Fourier coefficients that contain the highest energy.
We can now write

$$\epsilon(x, \widehat{x}) = \frac{\|x - \widehat{x}\|}{\|x\|} = p\frac{\|\delta \cdot W\|}{\|x\|} \leq p. \quad \text{-- (1)}$$

This means that the watermark embedding introduces an error which is proportional to the embedding power and to the norm of those descriptors for which *Wj* 6= 0. Given this immediate relation between power and error, we will often refer to the relative error introduced by the watermark to quantify the amount of power used during the embedding process.

c. Watermark Detection

The detection process aims at discovering the presence of a particular watermark in a watermarked dataset. This involves measuring the correlation between a tested water mark and the watermarked dataset. The higher the correlation between the two, the higher the probability that the embedded watermark was the one tested. Because the watermark is embedded in all objects of a dataset, one option is to measure the correlation between the watermark and the average of the magnitudes of Fourier descriptors across all objects of the dataset. However, directly measuring the correlation may not be very effective under multiplicative embedding. The reason is that since we want to minimize distortion, a small embedding power is preferred, whence the magnitudes of the Fourier descriptors are dominated by the original level of the average.

Definition 3:-(Detection Correlation).

Let D be the original dataset, bD the watermarked dataset, and W a watermark (to be tested). The correlation between W and bD given the average magnitudes in the original dataset μ(D) is

$$\chi(W, D): =(\mu(D) - \mu(D)\mu(D))^{T} \cdot W, \text{----(2)}$$

where the division is element-wise, excluding elements for which μj(D) = 0, and where all vectors are taken to be column vectors.

In addition to the watermark W, vector μ(D) is also recorded, and they are used jointly as the key. The additional storage cost is minor in the

face of enhanced security. A malicious attacker with an incentive to remove the watermark may try to detect it by searching through different watermarks for the correct one.

d. Effectiveness Under Attacks

We consider here various attacks, and discuss the efficiency of our scheme:

e. Cryptographic attack:

This attack resorts to an exhaustive key search, which attempts to identify the key used for the embedding. This would mean searching over all possible W's; yet there are as many as 2l of them, when assuming knowledge of the l watermarked! Frequencies, and $(nl)2^{l}$ without such knowledge. Actually, this is the premise of all cryptographic systems; the encoding algorithm is known, the secret key unknown, but brute-force computation would simply take too long for anyone to break it within a realistic time frame.

f. Oracle attack:

Here, the attacker tries to reconstruct the non-watermarked data from the watermarked data when the watermarking detector device is also available (which is a big security breach on its own). See [9] for one example of this, which however does not apply to our case, because the detector is not publicly available.

III. PRESERVATION OF NN AND MST

Our right-protection scheme can guarantee preservation of the NN and the MST post watermarking. These are two important properties of the distance graph, because a number of mining, learning, and visualization algorithms are based on them. For example, preservation of the NN will result in preservation of search operations based on a query-by-example paradigm (e.g., multimedia search); instance-based classification tasks based on the Nearest Neighbors will also be retained. Computation of the MST is also a fundamental operation in many data analysis tasks; applications using the MST can be found in logistics [12],data clustering [13], visualization [14], and phylogeny construction in biological applications [15].

As an example, Fig. 3 shows the output of the visualization algorithm of [14], which is based on the Minimum Spanning Tree. The technique maps objects on the two dimensional plane while preserving exactly the MST distances. Using the 2D perimeter of a dataset containing skulls of primates, one can easily visualize the evolutionary path between the different species. Our technique will guarantee that this (and any other algorithm) based on the MST will produce same outputs, before and after watermarking.

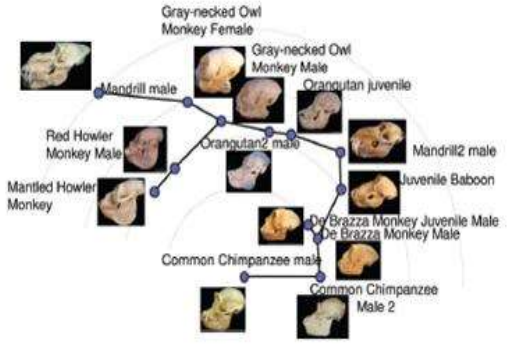

Fig. 3: Visualization algorithm based on the MST.

Definition 4:- (NN Preservation).

Given a dataset D and an object x ∈ D with Nearest Neighbor NN(x) 6= x in the distance graph of D, we say that object x preserves its Nearest Neighbor after the

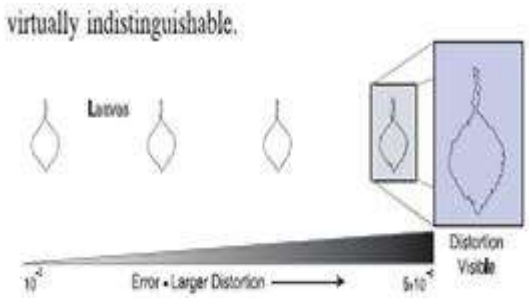

Fig. 4. Visual distortions for different watermark embedding powers, for the leaves dataset.

watermark embedding with watermark W and power p if

$$b\,Dp(x,NN(x)) \leq b\,Dp(x,y), \in y \in D, y \; 6= x. \quad \text{------ (3)}$$

We say that Nearest Neighbors are preserved if this holds for all x ∈ D.

Definition 5:-(MST Preservation).

Given a dataset D and two objects x, y v D such that the edge (x, y) is an edge of a Minimum Spanning Tree T of the distance graph of D, we say that the edge (x, y) is preserved in the MST after the watermark embedding with watermark W and power p if b

$$D\,p(x, y) \leq b Dp(u, v), \in u \in U(x,y), \in v \in V(x,y), \quad \text{----- (4)}$$

where U(x,y),V(x,y) are the two connected components T is split into after edge (x, y) has been removed. *We say that* the MST *T* is preserved *if all of its edges are preserved*. It is important to note that preservation of the MST after the watermark embedding does not necessarily imply preservation of Nearest Neighbors, and vice versa. This is because the MST edges do not necessarily involve the NN of each object; the MST edges include only those nearest neighbors that do not introduce a circle on the final MST graph.

IV. FAST NN-PRESERVATION

Like the existing system, proposed system also uses a spread-spectrum approach. In addition, different kinds of watermark embedding are applied to same data set so that the data set can be distributed to more types of users. Also, information receiving users are added so that only those authorized users can access their respective data.

V. MST-PRESERVATION ALGORITHM

A similar rationale applies for the MST preservation algorithm. The algorithm progressively removes infeasible powers, under which the MST properties are violated. Let T(D, E) be a Minimum Spanning Tree of the distance graph of dataset D, where E is the set of $|D| - 1$ edges composing the tree. If we remove an edge $e = (x, y) \in E$, we split the original tree into connected components Ue and Ve. Since T is a Minimum Spanning Tree, such edge e = (x, y) has the property of being a shortest edge that connects Ue with Ve.

Algorithm 2 MST-Preservation

1: INPUTS: D,W, pmin, pmax, τ
2: OUTPUT: p∗
3: T(D, E) = find MST of D (using Kruskal's algorithm)
4: for all e ∈ E do
5: feasible_powers(e) = [pmin, pmax]
6: for all u ∈ Ue do
7: for all v ∈ Ve do
8: feasible_powers(e) =
solve bD 2p (e) ≤bD 2p (u, v) | feasible_powers(e)
9: end for
10: end for
11: end for
12:p∗=max{p{e:p/∈feasible_powers(e)} ≤ τ · (|D| − 1)}

If for edge e = (x, y) we use D(e) to denote the Euclidean distance D(x, y), for every edge e ∈ E it holds that

$$D(e) \leq D(u, v)\ \forall u \in Ue, \forall v \in Ve. \quad \text{------ (5)}$$

This defining property of an edge e of the MST is preserved after the watermark embedding with watermark W and power p if and only if b

$$Dp(e) \leq bDp(u, v)\ \forall u \in Ue, \forall v \in Ve. \quad \text{------ (6)}$$

The MST-P Watermarking Problem can be solved again via a system of quadratic inequalities.

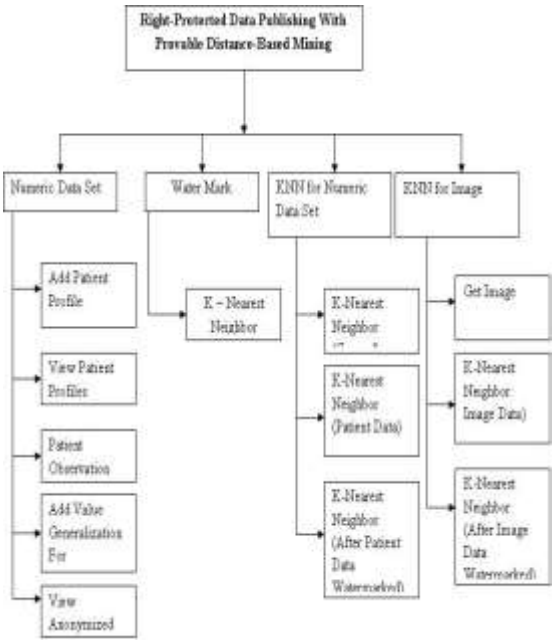

Fig. 5 :Right-Protected Data Publishing With Provable Distance-Based Mining

Advantages of MST-Preservation

a) Different users receive the different watermarked data.
b) Data about the receiving user is also embedded in watermarking information.
c) Simulation is applied on both image as well as numeric data set.

Complexity: The number of computations of quadratics, as well as the number of inequalities solved, is O(|D|3). The MST can be computed in O(n|D|2) time. If the pair wise distances in the

original dataset are stored, then computation of the coefficients of a quadratic takes O(l) time. Function solve can be computed in O(1) time. It follows that the time complexity of the algorithm is O(l|D|3 + n|D|2).

VI. EXPERIMENTAL EVALUATION

In all experiments we have used $p_{min} = 0$ and $p_{max} = 0.01$, so we allow up to 1% relative distortion due to watermarking. This corresponds to an SNR of 40db. We set $\tau = 0$, enforcing full maintenance of the NN and the MST.

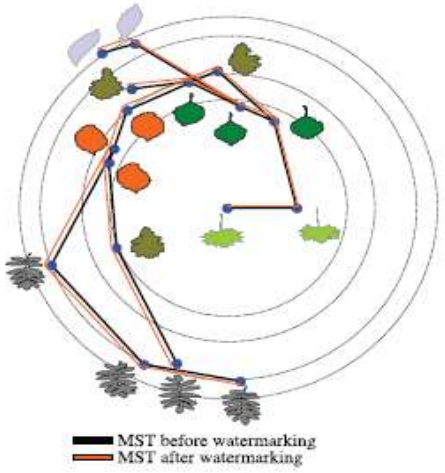

Fig.6:- MST preservation on the leaves dataset. Observe that the MST before (black lines) and after watermarking (orange lines) is not distorted.

ACKNOWLEDGMENT

Ms.K. Mowcika is currently pursuing Bachelor's Degree program in Computer Science and Engineering at Shree Sathyam college of Engineering and Technology, Anna University, India.

Mr. S.P.Santhoshkumar is currently working as an Assistant Professor in Department of Computer Science and Engineering at Shree Sathyam college of Engineering and Technology, Anna University, India.

Ms. V.Anuradha is currently working as an Assistant Professor and pursuing Doctoral Degree in Department of Computer Science and Engineering at Shree Sathyam college of Engineering and Technology, Anna University, India.

Ms. G.Mariya Lavanya Sangeetha is currently working as an Associate Professor and pursuing Doctoral Degree in Department of Computer Science and Engineering at Shree Sathyam college of Engineering and Technology, Anna University, India.

REFERENCES

[1] Spyros I. Zoumpoulis, Michail Vlachos, Nikolaos M. Freris, Member, IEEE and Claudio Lucchese, Right-Protected Data Publishing with Provable Distance-Based Mining" IEEE transactions on knowledge and data engineering, vol. 26, no. 8, august 2014.

[2] D. Aha, D. Kibler, and M. Albert, "Instance based learning algorithms," Mach. Learn., vol. 6, no. 1, pp. 37–66, 1991.

[3] C. G. Atkeson, A. W. Moore, and S. Schaal, "Locally weighted learning," Artif. Intell. Rev., vol. 11, pp. 11–73, Feb. 1997.

[4] N. Paivinen, "Clustering with a minimum spanning tree of scale-free-like structure," Pattern Recognit. Lett., vol. 26, no. 7, pp. 921–930, 2005.

[5] Y. Xu, V. Olman, and D. Xu, "Minimum spanning trees for gene expression data clustering," Genome Inform., vol. 12, pp. 24–33, 2001.

[6] J. B. Tenenbaum, V. de Silva, and J. C. Langford, "A global geometric framework for nonlinear dimensionality reduction," Sci., vol. 290, no. 5500, pp. 2319–2323, 2000.

[7] M. Vlachos, C. Lucchese, D. Rajan, and P. S. Yu, "Ownership protection of shape datasets with geodesic distance preservation," in Proc. 11th Int. Conf. EDBT, Nantes, France, 2008, pp. 276–286.

[8] I. J. Cox, J. Kilian, F. T. Leighton, and T. Shamoon, "Secure spread spectrum watermarking for multimedia," IEEE Trans.Image Process., vol. 6, no. 12, pp. 1673–1687, Dec. 1997.

[9] J.-P. M. G. Linnartz and M. van Dijk, "Analysis of the sensitivity attack against electronic watermarks in images," in Proc. 2nd Int. Workshop IH, Portland, OR, USA, 1998, pp. 258–272.

[10] F. Hartung, J. Su, and B. Girod., "Spread spectrum watermarking: Malicious attacks and counterattacks," in Proc. SPIE Security Watermarking Multimedia Contents, vol. 3657, San Jose, CA, USA, 1999.

[11] V. Solachidis and I. Pitas, "Watermarking polygonal lines using Fourier descriptors," IEEE Comput. Graph. Appl., vol. 24, no. 3, pp. 44–51, May/Jun. 2004.

[12] P. Das, N. R. Chakraborti, and P. K. Chaudhuri, "Spherical min imax location problem ," Comput. Optim. Appl., vol. 18, no. 3, pp. 311–326, 2001.
[13] G. Economou, V. Pothos, and A. Ifantis, "Geodesic distance and MST-based image segmentation," in Proc. 12th EUSIPCO, Vienna, Austria, 2004, pp. 941–944.
[14] M. Vlachos, B. Taneri, E. J. Keogh, and P. S. Yu, "Visual exploration of genomic data," in Proc. 11th Eur. Conf. PKDD, vol. 4702, Warsaw, Poland, 2007, pp. 613–620.
[15] S. J. Shyu, Y. T. Tsai, and R. C. T. Lee, "The minimal spanning tree preservation approaches for DNA multiple sequence alignment and evolutionary tree construction," J. Combinat. Optim., vol. 8, no. 4, pp. 453–468, 2004.

