IJREAT International Journal of Research in Engineering & Advanced Technology, Volume 2, Issue 1, Feb-Mar, 2014
ISSN: 2320 - 8791
www.ijreat.org

Fixed Grid DDM Based Watermarking Datasets for Process Mining

Tintu Mary Xavier[1], Mercy Paul selvan[2], Dr. A.Chandrasekhar[3]

[1] Department of CSE, Sathyabama University,Chennai
[2]Faculty of Computing, Sathyabama University, Chennai
[3]Department of CSE, St.Joseph's Engineering College,Chennai

Abstract-Watermark, a recognizable pattern applied on the existing datasets to identify authenticity. A watermark stored in a data file refers to a method for ensuring data integrity in tamper detection which has its own advantage. The focus is to define "usability constraints" for watermarking the data mining datasets in such a way that the watermark is not only robust and flexible but the knowledge contained in the dataset is also preserved based on user's preferences. Here, trying to utilize the option of matching data by using common characteristics found within the data set in the relational datasets. The proposed project is to create an innovative system to identify the fake tuples through which the data owner can easily identify the fake users re-using the distributed datasets. Uncontrolled data leakage put business in a vulnerable position. Data allocation strategies are proposed to improve the probability of identifying leakages. Storing data in a de-normalized way enhance security. Amending the fake tuples relevant to the data owner is one of the main ideas to safeguard the data. This can be automated by providing an innovative model of amending this kind of data in it. Modifying the existing data is another option by creating a clear watermark on the data and it should be easily recoverable one. Privacy on the existing data by cut shorting the columns from user view makes the column looks more secure one.

Keywords- Watermarking, data leakage, fake tuples.

I.INTRODUCTION

A data owner shares his datasets with mining agents to extract information from a datasets and transform it into an understandable structure for further use. During this sharing of datasets there may be a chance of dataset leakage and led the business into a vulnerable position. If cyber criminal cash out or sell this data for profit it leads to organization downfall. Therefore watermark has been applied for protecting ownership on the datasets.

The information to be embedded in a digital data in different formats such as audio, video, image, relational database, text and software is called a watermark [1]-[3].The signal in which the watermark is embedded is called host signal. A watermarking scheme consists of three steps, insertion, attack and detection. In insertion algorithm the host and the data to be embedded is selected and produces a watermark signal. Then the watermarked signal is transmitted to another person, and if he makes any modification then it is called an attack. Detection checks whether the watermark inserted was modified or not after the transmission. Watermarking is very useful in protecting ownership but it involves some modification of the original data and it can be destroyed if the data recipient is malicious.

The data can be secured more by the amendment of fake tuples. It improves the possibility to identify the agent that leaked the data. The fake tuples can be considered as a type of watermark for the entire set and it doesn't need modification of any individual values. If it proves an agent was given one or more fake objects that were leaked, then the distributor can be making sure that agent was guilty.

II.RELATED WORK

Mayuree *et al.* [4] reviewed four papers proposed by different authors on watermarking relational databases. Watermark can be applied to any database relation such that changes in a few of

the values in the attributes do not affect the applications. The watermark bit pattern which constitutes attributes within a tuple, bit positions in an attribute, and specific bit values are algorithmically determined under the control of a private key known only to the owner of the data. All the techniques proposed have the property of robustness and it is flexible and easy to find the leakage based upon the attributes. But they didn't explain how the knowledge about the schema and watermark will be given to the other user and not sure how the owner will identify the criticality of the data to be changed.

In the work of Kamran *et al.* [5], usability constraints were defined to preserve the knowledge contained in the datasets during watermarking a relation. Datasets have given as input to this model and features were extracted. Features were ranked using mutual information to understand the correlation of feature on predicting a class label. Then data have grouped using a parameter CPT (Classification potential threshold), so that features having same classification potential remains in the same group. Based on this local and global constraints were defined, where local constraints deal with features within a tuple and global constraints for the whole datasets. This watermarking scheme can work with both numeric and non-numeric datasets. But this work concerns only watermarking at datalevel, addition of fake tuples is meaningless. The data owner doesn't able to identify the fake users accurately.

The nature of data leakage and possibilities to avoid it has explained in [6]. It specified how the distributor can intelligently give data to agents in order to improve the chances of detecting a guilty agent. The addition of fake objects to the distributed set, leads the distributor to find the guilt agent easily. We can get the clear idea to know about the leakage and to detect the guilty agent easily. This paper didn't explain the kind of impact happening on the system while adding the fake objects. Touching or modifying the sensitive data is not an advisable move.

Viji *et al.* [7] proposed a novel data distortion method to increase the privacy level and correctness of the numerical data. The new introduced data perturbation techniques alter the confidential attribute of the data by the multiplication and addition of random noise. This method provides a high degree of privacy and clustering quality. But this technique deals only with the numerical data. It doesn't mention how a perturbation can be done on a non numerical data and higher dimensional data.

The paper [8] explains about the transformation or modification of data happening automatically due to mining of data or while storing the data in the warehouse. Tracing procedures takes advantage of known structure or properties of transformations when present, but also work in the absence of such information. By using Mining concepts the data modifications can be done while data storing and the structure also defined to trace the transformation. This paper didn't focus on the latest tools which will solve this kind of problem automatically and doesn't give clear explain on the security part in this tool.

Latanya *et al.* [9] put forward other techniques called generalization and suppression to safeguard the data. The data in the system is analyzed for generalization like replacing (or recoding) a value with a less specific but semantically consistent values. Suppression is not providing the data to the user. By using generalization and suppression techniques the data can be secured and semantically have consistent values. The major issue with this paper is that, there is no clear explanation on, how the data is going to be secured in suppression technique. If the data is not semantically linked, then this technique won't be effective.

In comparison, the focus of the current work is to create an innovative system that incorporates the concept of water marking and amendment of fake tuples in the data. This technique preserves datasets of the data owner and shares data based on the preference of the user request. A high level security can be achieved that safeguards data thereby avoiding cyber crime risk in datasets.

III.PROPOSED WORK

The core idea of proposed work is preserving the data from exposing to the end user. To achieve this an automated system is proposed which will take care of injecting the data

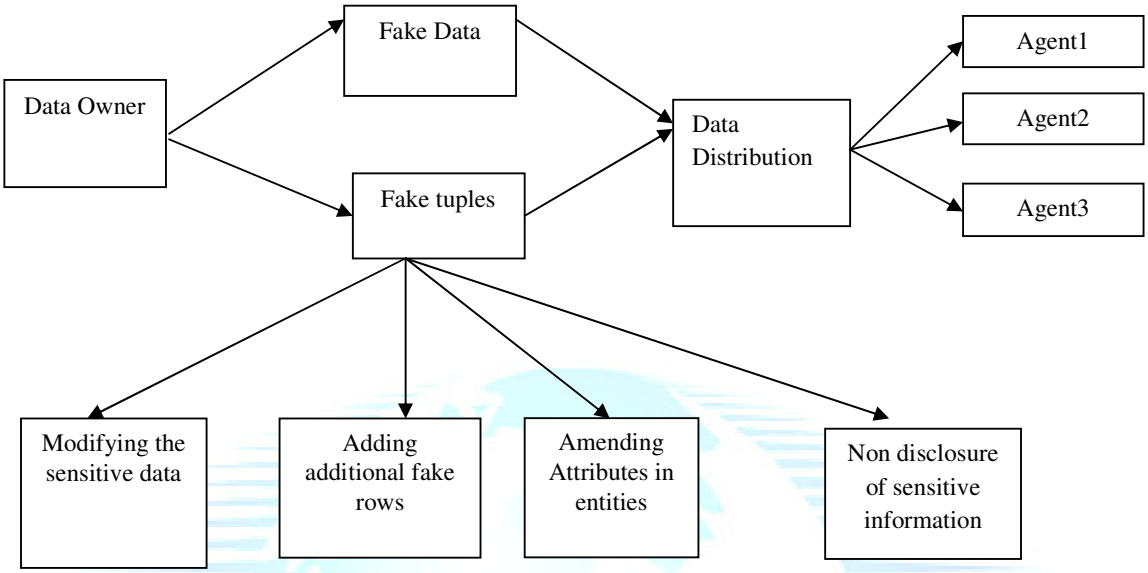

Fig 3.1 Architecture diagram of the proposed system

in a fashioned manner. A recognizable watermarking pattern based on the usability constraints has the property of robustness and flexibility as well as safeguard the data based on the user preference. Data owner act in the core part to identify the fake users reusing distributed datasets thereby using the automated system that identifies the fake tuples, this in turn prevents the leakage of data. A strategy named data allocation is proposed to improve the probability of identifying the leakage.

Preserving main columns from exposing for mining is the next level of security provided on the data. To enhance security, data will be stored in a denormalized way. An automated way of handling the data selection for distribution is proposed in this work. Morphing sensitive data and injecting row wise tuples using an automation process with a clear picture of data changes in the system is one of the newest and proposed features in this work.

The fig 3.1 shows the architecture diagram of the proposed system. A data owner enters into the system selects a distributor from the distributor list and the data to be transferred based on the user preference. Based on the selection automatic queries will be generated to retrieve the datasets. A unique watermarking will be applied on the selected datasets. Addition of fake tuple is one of

the watermarking strategies. The following algorithm explains the amendment of fake tuples to the distributed sets. This approach is based on the algorithm proposed by Vahab *et al.* [10]

Algorithm1:

Input: - D: dataset of the agent.
N: the number of fake tuples.
A: the set of attributes that are not Candidate keys.
Output:-D: dataset with fake tuples.
1. for each a_i in A
a) Read a fake data from the owner
b) Generate n fake values by adding different characters to the fake data. The generated value should be distinct.
2. Set of fake records, F with n tuples is generated.
3. While n>0 do
f= select a fake tuple at random from set F
D= D U {f}
F= F-{f}
n=n-1

The modification of sensitive data as mentioned in [5] is another option for watermarking. The data is modified in such a way

that it should be easily recoverable as well as feature should not be altered much. The identification of more sensitive information and not revealing it to the agent, makes the datasets more secured one.

After applying the watermark, the data is subjected for distribution. A mode of data processing whereby a terminal or group of terminals serves as a receptacle for data supplied by the CPU. The Data Distribution Service for Real-Time Systems (DDS) is an Object Management Group (OMG) machine to machine middleware standard that aims to enable scalable real-time dependable high performance and inter operable data exchanges between publishers and subscribers. The data are sent to the data distributors with unique watermarking. The distributors proceeds with the data on passing it to the different receiver ignorance of the watermarking. The distribution is based on the fixed grid DDM algorithm [11]. The routing space is partitioned into non-overlapping grid cells, and a multicast group is defined for each cell. A federate subscribes to the group associated with each cell that partially or fully overlaps with its subscriber regions. The result associates a region with several multicast groups in a fixed and pre-determined manner. A publish operation is realized by sending an update message to the multicast groups corresponding to the cells that partially or fully overlap with the associated publisher region. The fixed-grid approach eliminates the need to explicitly match publisher and subscriber regions.

The fixed-grid DDM algorithm implementation consists of three sub-procedures: Grid Initialization, Federate-to-Grid mapping, and Multicast Group creation·

Grid Initialization Sub-Procedure
BEGIN Procedure
Divide the routing space into grid cells G_i given dimension in routing space;
Each cell is uniquely identified by a cell ID;
Each grid cell maintains counters for number of publisher regions overlapping and their federate ID;
Each grid cell maintains counters for number of subscriber regions overlapping and their federate ID;
END Procedure

Federate-to-Grid Mapping Sub-Procedure
Initialization
For each grid cell G_i, the following variables are maintained.
Pub_Fed_ID: Array for storing the federate ID of each publisher region overlapping with G_i,
Pub_Region_Counter: Counter for number of publisher regions overlapping with G_i,
Sub_Fed_ID: Array for storing the federate ID of each subscriber region overlapping with G_i,
Sub_Region_Counter: Counter for number of subscriber regions overlapping with G_i.
BEGIN Procedure
For all federates F_i do
Begin
// For the publisher region P_i
For all grid cells G_i covered by publisher region P_i do
Begin
Add publisher region P_i information to grid cell G_i;
Increment the Pub_Region_Counter for grid cell G_i;
End
// For the subscriber region S_i
For all grid cells G_i covered by subscriber region S_i do
Begin
Add subscriber region S_i information to grid cell G_i;
Increment the Sub_Region_Counter for G_i ;
End
End
END Procedure

Create Multicast Group Sub-procedure
Initialization
MCG: Multicast group for DDM.
Each grid cell G_i is assigned a multicast group MCG_i.
BEGIN Procedure
For all grid cell G_i do
Begin
Add all publisher regions in grid cell G_i to MCG_i ;
Add all subscriber regions in grid cell G_i to MCG_i ;
End
END Procedure

The watermark detection algorithm checks whether the distributed datasets has been modified or not. The data is examined carefully for accuracy with the intention of verification. Robustness requirements of watermarks mandate that any attempt to remove or destroy the watermark should produce a remarkable degradation in data quality before the watermark is lost. A master copy of the data cum watermark data will be maintained by the Distributor and the watermarked original copy will be matched with the agent's copy to identify the threat in data leakage. The focus is to find the exact guilt.

The proposed model provides data security in high level and identification of leakage using a data allocation strategy is an effective way to safeguard the data. Unlike other models, this provides security at data level as well as column level. Usage of automated system drives the work more perfect than the manual process used in the existing.

IV.EXPERIMENT AND RESULT

We conducted an experiment with data leakage to find the correctness of probability of leakage and a guilty agent. A sample 15 datasets table the attribute title have morphed with a given value "ttt". This morphing have done to provide security for high secure column.

For each agent along with the actual datasets, a set of fake datasets also have added. Before distributing the data to the agents fake have taken, table 4.1 shows the datasets that have to be distributed to different agents . In the below

For example the values for the attributes Fname, Mname, Lname given by the user for agent1 were Linu ,Mary ,Joseph respectively. Using these details a set of fake tuples have generated for agent1 and it has shown in table 4.2. Similarly for the remaining two agents a set of fake tuples have created by the user given values. The fake tuples

Customer key	Tittle	Fname	Mname	Lname	Gender	Marital status	Email Address
11001	Ttt	Eugene	L	Hwang	M	S	eugenio@gmail.com
11002	Ttt	Ruben	Null	Torres	M	M	ruben35@gmail.com
11101	Ttt	Christy	Null	Zhu	F	S	christy12@gmail.com
11120	Ttt	Elizabeth	Null	Johnson	F	S	elizabeth5@gmail.com
11199	Ttt	Julio	Null	Ruiz	M	S	julio1@gmail.com
11289	Ttt	Janet	G	Alvareez	F	M	janet9@gmail.com
12100	Ttt	Macro	Null	Mehta	M	M	macro14@gmail.com
12250	Ttt	Rob	Null	Vehoff	F	M	rob4@gmail.com
12505	Ttt	Shannon	C	Carlson	M	M	shannon38@gmailcom
13401	Ttt	Jacquelyn	C	Suraez	F	S	jacquelyn20@gmail.com
13498	Ttt	Curtis	Null	Lu	F	S	curtis9@gmail.com
13565	Ttt	Lauren	M	Walker	M	M	lauren41@gmail.com
13589	Ttt	Ian	M	Jenkins	M	S	ian47@gmail.com
13602	Ttt	Sydney	Null	Bennett	M	S	sydney23@gmail.com
13612	Ttt	Chloe	Null	Young	F	M	chloe23@gmail.com

tuples have created for the three agents. The fake tuples are generated based on the values given by the user for the attributes specified in this model.

created for the agent2 and agent3 are shown in table 4.3 and table 4.4 respectively

TABLE 4.1 SAMPLE DATASETS TO BE DISTRIBUTED

TABLE 4.2 FAKE TUPLES FOR AGENT1

Customer key	Tittle	Fname	Mname	Lname	Gender	Marital status	Email Address
11100	Null	Linu	Mary	Joseph	M	M	linuent1@gmail.com
11200	Null	Linujaclin	Maryrb	Joseph	F	M	linuentjacqueline@gmail.com
11300	Null	Linu	Mary	Josephvazquez	M	M	ent1varquejoseph@gmail.com
11400	Null	Bernadlinu	Mary	Mehtajoseph	F	S	josephbrendalinu@gmail.com

TABLE 4.3 FAKE TUPLES FOR AGENT2

Customer key	Tittle	Fname	Mname	Lname	Gender	Marital status	Email Address
12200	Null	Prem	Kora	Dinakar	F	S	premkora2@gmail.com
12300	Null	Premwalter	Kora	Dinakar	M	M	premdinakar@gmail.com
12400	Null	Prem	Jkora	Dinakarchapman	M	S	korapremdin@gmail.com
12500	Null	Heatherprem	Kora	Wangedinakar	F	S	dinakarkoraprem@gmail.com

TABLE 4.4 FAKE DATA FOR AGENT3

Customer key	Tittle	Fname	Mname	Lname	Gender	Marital status	Email Address
13400	Null	Rajedward	kumar	Vijay	M	M	raj12entedward@gmail.com
13500	Null	Raj	Kumar	Vijaycox	F	S	Entcoxvijay12@gmail.com
13600	Null	Valerieraj	Kumar	Zhuvijay	F	M	Vijay12rientraj@gmail.com

TABLE 4.5 LEAKED DATA

Customer key	Tittle	Fname	Mname	Lname	Gender	Marital status	Email Address
11001	ttt	Eugene	L	Hwang	M	S	eugenio@gmail.com
11002	ttt	Ruben	Null	Torres	M	M	ruben35@gmail.com
11101	ttt	Christy	Null	Zhu	F	S	christy12@gmail.com
11120	ttt	Elizabeth	Null	Johnson	F	S	elizabeth5@gmail.com
11199	ttt	Julio	Null	Ruiz	M	S	julio1@gmail.com
11289	ttt	Janet	G	Alvareez	F	M	janet9@gmail.com
12100	ttt	Macro	Null	Mehta	M	M	macro14@gmail.com
12250	ttt	Rob	Null	Vehoff	F	M	rob4@gmail.com
12505	ttt	Shannon	C	Carlson	M	M	shannon38@gmailcom
13400	ttt	Rajedward	kumar	Vijay	M	M	raj12entedward@gmail.com
13401	ttt	Jacquelyn	C	Suraez	F	S	jacquelyn20@gmail.com
13498	ttt	Curtis	Null	Lu	F	S	curtis9@gmail.com
13500	ttt	Raj	Kumar	Vijaycox	F	S	Entcoxvijay12@gmail.com
13565	ttt	Lauren	M	Walker	M	M	lauren41@gmail.com
13589	ttt	Ian	M	Jenkins	M	S	ian47@gmail.com
13600	ttt	Valerieraj	Kumar	Zhuvijay	F	M	Vijay12rientraj@gmail.com
13602	ttt	Sydney	Null	Bennett	M	S	sydney23@gmail.com
13612	ttt	Chloe	Null	Young	F	M	chloe23@gmail.com

The generated fake tuple for an agent will be injected into the original data before the data is given to that agent. A master copy of the data cum watermark data will be maintained by the distributor. Later, if found a leaked data it will be compared with the watermarked original copy. Here the table 4.5 is an example of the leakage datasets. The model compares the data with the reference data and it founds a match and unmatched records. The graph in fig 4.1 shows the probability of leakage.

The model tries to find the exact culprit. The focus is to find among the three agents who will be having data similar to leaked datasets .The three agent's data will be compared with the leaked data. The fig 4.2 shows how much percentage each agents record match with the leaked one .The y-axis shows the number of records. From the graph it is clear that Agent3 is the guilty agent.

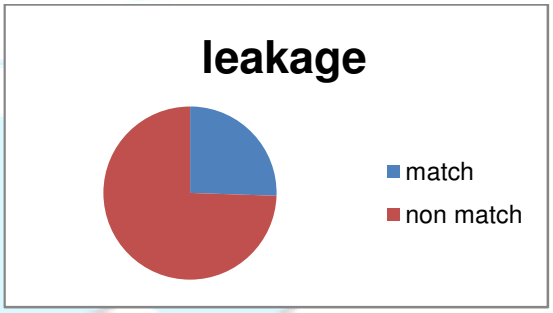

FIG 4.1 MATCH AND UNMATCH RECORDS

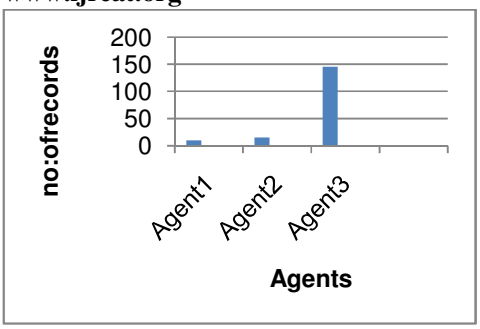

FIG 4.2 AGENT'S MATCH RECORDS

V. CONCLUSION

In this paper, we have developed a novel yet pragmatic framework for watering. It includes carefully designed, shifting and clustering in each of which handles a specific problem in watermarking. Fixed grid pre-processes tend to host images by adjusting the pixels into a reliable range for satisfactory reversibility. Datasets are included in shifting and clustering constructs of new watermark embedding and extraction processes for good robustness and low run-time complexity in the precise estimates the local sensitivity of grids and adaptively optimizes the watermark strength for a trade-off between robustness and invisibility. In contrast to representative methods, the proposed framework: 1) obtains comprehensive performance in terms of reversibility, robustness, invisibility, capacity and run-time complexity; 2) is widely applicable to different kinds of images; and 3) is readily applicable in practice.

In future, we will combine the proposed framework with the local feature water marked concepts for the efficiency to further improve robustness. In addition, it is valuable to integrate the merits of sparse representation and probabilistic graphical model into the designing of image watermarking.

REFERENCES

[1] R. Agrawal, P. Haas, and J. Kiernan, "Watermarking relational data:Framework, algorithms and analysis," *The VLDB Journal*, vol. 12, no.2, pp. 157–169, 2003.

[2] J. Palsberg, S. Krishnaswamy, M. Kwon, D. Ma, Q. Shao, and Y.Zhang, "Experience with software watermarking," in *Proc. 16thAnn.Computer Security Applications Conf.*, 2000, pp. 308–316.

[3] M. Atallah, V. Raskin, M. Crogan, C. Hempelmann, F. Kerschbaum, D. Mohamed, and S. Naik, "Natural language watermarking: design, analysis, and a proof-of-concept implementation," in *Information Hiding*. New York, NY, USA: Springer, 2001, pp. 185–200.

[4] Mayuree K.Rathva, Prof. G. J. Sahani, "Watermarking relational databases," *International Journal of Computer Science, Engineering and Applications (IJCSEA)* Vol.3, No.1, February 2013.

[5] M. Kamran, M. Farooq, "A formal usability constraints model for watermarking of outsourced datasets", *IEEE Transactions On Information Forensics And Security*, Vol. 8, No. 6, June 2013.

[6] Umamaheswari, S. Getha , H. A, "Detection of guilty agents", *Innovations in Emerging Technologies(*NCOIET*)*, 2011, pp. 23-26

[7] A.Viji Amutha Mary, Dr. T.Jebarajan, " A novel data perturbation technique with higher security ", *International Journal of Computer Engineering and Technology,* Vol 3, No. 2, 2012, pp. 126-132

[8] Cui, Yingwei, and Jennifer Widom. "Lineage tracing for general data warehouse transformations." *The VLDB Journal—The International Journal on Very Large Data Bases* 12.1 ,2003, pp. 41-58.

[9] L. Sweeney, " Achieving *k*-anonymity privacy protection using generalization and suppression," *International Journal on Uncertainty, Fuzziness and Knowledge-based Systems,* 10 (5), 2002, pp. 571-588.

[10] Vahab Pournaghshband, " A New Watermarking Approach for Relational Data ", *in proc. 46th Annual Southeast Regional conf. ,* 2008, pp. 127-131.

[11] Pankaj Gupta, Ratan K. Guha, " Design, Analysis, and Performance Evaluation of an Efficient Algorithm for Data Distribution Management in High Level Architecture," *CS-TR*-05-12, December 2005.