Fraud Detection in Information Leakage

Poonam Sahoo
University of Pune, India
Computer Engg.
e-mail: poonam.sahoo00@gmail.com

Kanchan Garud
University of Pune, India
Computer Engg.
e-mail: garudkanchan11@gmail.com

Sarika Kate
University of Pune, India
Computer Engg.
e-mail: katesarika2@gmail.com

Sharayu Padmane
University of Pune, India
Computer Engg.
e-mail: sharayupadmane@gmail.com

Guided by: Prof. Manisha Marathe
Dept. of IT and Computer Engg.
PVG COET, Pune, India.
Email: manisha_marathe@yahoo.com

Abstract— In the field of business, the owner having some crucial data may need to share it with third-parties. These trusted third-parties may use this data for their own benefit causing reputational and monetary damage to the owner’s company. If some of the shared data is discovered at some illegal place, it is quite possible that one or more third party agent is responsible for such information leakage.

The owner must identify the leakage at the earliest and possibly the source of leakage. This paper presents two methods to handle such information leakage namely watermarking and Identifying Guilty agent using probability.

Keywords- Watermarking, Fake records, Guilty agent, Leaked dataset, sensitive data, Probability of guilty agent , Allocation strategies, data privacy, leakage model.

I. INTRODUCTION

While doing business, crucial and private data is sometimes shared with the supposedly trusted third parties. For example, an Insurance Company may give its customer data for further processing to other agents like what is the average salary of people who are getting themselves insured for amount greater than Rs.100000.

We call the owner of the data distributor and the supposedly trusted third parties agents. These agents may behave in such a way so as to achieve their own objectives even at the cost of others.

If any agent gives this crucial information to other unauthorized vendors without informing the distributor then the data is said to be leaked by the agent. The distributor’s goal is to detect such data leakage and to identify the agent responsible for it.

II. WATERMARKING METHODOLOGY

Nowadays, the digital assets such as software, images, video, audio and text are pirated which is a strong concern for owners of these assets. The protection schemes for such assets are based upon the insertion of digital watermarks into them. In this process, a particular object or record from the data is selected for the purpose of watermarking satisfying criteria. The criterion is these marks should have insignificant impact on the usefulness of data. The procedure of watermarking introduces small errors into the object being watermarked. These intentional errors are called marks and all the marks together constitute the watermark. These marks are chosen in such a way that it has least impact on the data and placed such that a malicious user cannot destroy them.

In traditional technique, leakage detection is handled by watermarking which is method of implanting a unique code on each of the distributed copy. When this copy is later discovered in the hands of an unauthorized party, the leaker can be identified.

i. System

The present system watermarks a database relation $R$ whose scheme is $R(P,A_0,\ldots,A_{v-1})$ where $P$ is the primary key attribute. It is assumed that all $v$ attributes $A_0,\ldots,A_{v-1}$ are candidate key attributes for marking. Each attribute is numeric and a small number of changes in the $E$ least significant bits of an attribute are unnoticeable.

The watermark insertion algorithm employs a cryptographically secure pseudo-random sequence generator $G$ together with a secret key. For each tuple, $G$
is initiated with the concatenation of the secret key and the value of the tuple’s primary key attribute. Then \( G \) is used together with a user-specified “gap parameter” to determine whether or not to mark the tuple; the gap parameter controls the number of tuples that are marked. If a tuple is selected for marking, then \( G \) is further used to determine which of the \( v \) attributes will contain the mark, which of the \( E \) least significant bits of the selected attribute will be marked, and the actual value of the marked bit.

ii. Drawbacks of watermarking

However, there are two major disadvantages of the above algorithm:
1. It involves some modification of data i.e. making the data less sensitive by altering attributes of the data. This alteration of data is called perturbation. However in some cases, it is important not to alter the original distributed data. For example, if an agent is doing the payroll, he must have the exact salary. We cannot modify the salary in this case.
2. The second problem is that these watermarks can be sometimes destroyed if the recipient is malicious.

III. PROBABILISTIC APPROACH

The main objective of this approach is to analyze the guilt of every agent which can be responsible the leakage.

The approach mainly deals with customer data which is important for companies like Life insurance companies. If data leakage happens, they can’t afford to lose their valuable customers through which they earn their income as well as reputation.

i. System

The whole system is divided into three parts:
1. Data allocation strategies
   In our system, there are such data allocation strategies [3] that involve addition of some realistic but fake objects while distributing data to the agents in order to improve the chances of finding the guilty agent.
2. Mail Detection system
   The system basically deals with customer information like email-IDs. When customer data is leaked, the third party tries to create contact with these customers through mail. Our mail detection system keeps a track of the fake email-ids added at the time of data allocation and distribution to agent and when an unidentified mail having advertisement content arrives in the mailbox, the system informs the administrator about it.
3. Probability calculation
   In the final stage, when the threshold value of unidentified mails is crossed, the system calculates the probability of each agent being guilty with the help of these fake records [2] and the one having the highest probability is the Guilty agent. The threshold value is the minimum value required to trigger the system for the calculation.

ii. Problem definition

The distributor has to distribute data in an efficient manner such that leakage can be detected.

There are many cases where alteration to the original data can’t be done. In such cases, we can add some realistic records similar to the dataset which don’t exist in reality. For example, In a Life Insurance Agency the distributor can’t alter the personal and contact details of the customer. In order to watermark such data, some fake customer records are acceptable since no real customer matches with this record.

There are two major constraints in this system:
1. The distributor has to satisfy agent’s request by providing them with the number of objects they request. The distributor may not deny serving an agent request and may not provide different perturbed versions of the same objects.
2. The second constraint is to allocate data in such a way that the guilty agent is tractable.

iii. Problem Setup and Notation

Entities and Agents: A distributor owns a set \( T = \{t_1; \ldots; t_n\} \) of valuable data objects. The distributor wants to share some of the objects with a set of agents \( U_1; U_2; \ldots; U_n \), but does not wish the objects be leaked to other third parties. The objects in \( T \) could be of any type and size, e.g., they could be tuples in a relation, or relations in a database. An agent \( U_i \) receives a subset of objects \( R_i \subseteq T \), determined either by a sample request or an explicit request:

- Sample request \( R_i = \text{SAMPLE} (T; m_i) \): Any subset of records from \( T \) can be given to \( U_i \).
- Explicit request \( R_i = \text{EXPLICIT} (T; \text{condi}) \): Agent \( U_i \) receives all the \( T \) objects that satisfy condition.

IV. MODULE DESCRIPTION

- Database Maintenance
  The sensitive data which is to be distributed to the agents is maintained in this database.
Agent Maintenance
The registration details provided by the agent as well as the data which is given to it by the distributor is maintained.

Addition of fake Objects
The distributed is able to add fake objects in order to improve the effectiveness in detecting the guilty agent.

Mail Detection system
The monitoring of mails on the fake records distributed is done by this system.

Data Allocation
In this module, the original records fetched according to the agent’s request are combined with the fake records generated by the administrator.

Calculation Of Probability
In this module, the each agent’s request is evaluated and probability of each agent being guilty is calculated.

V. SYSTEM FLOW

In the beginning, agent registers into the system by entering its personal details and the request of data. The system then extracts the requested data from the main database and performs the addition of fake records according to the request. It then provides the data to the agent.

If any of the agent leaks the data to some unauthorized vendor, then the vendor will try to establish a contact with the customers by sending them advertising mails. The job of the mail detection system is to monitor these incoming mails on the email addresses of the fake records continuously.

If the system detects unauthorized mail crossing the threshold value, then it starts its process of probability calculation. The threshold value is the minimum number of mails which have to be detected to trigger the calculation. It will check the presence of these fake records in each of the agent and accordingly will evaluate the probability of each agent being guilty. The agent having the maximum probability will be the guilty agent.

VI. CONCLUSIONS AND FUTURE WORK

Scope of the above system can be extended by handling the generation of fake records dynamically according to the agent’s request. Accuracy of the calculation of the probability becomes more, if we have a method to get the complete dataset.

In real world, there is always a need to handover crucial data to third parties and we can’t distribute such data by watermarking each and every object which also involves modification. Hence, we have to implement a system where data is not modified yet watermarked which will signify the ownership of the distributor.

REFERENCES


[5] Detection of Guilty Agents, S. Umamaheswari #1, H. Arthi Geetha #2 #1.2M.E II Year, Department Of Computer Science, Coimbatore Institute of Engineering and Technology; Coimbatore, Tamilnadu, India