ISSN No: 2348-4845
International Journal & Magazine of Engineering, Technology, Management and Research
A Peer Reviewed Open Access International Journal

Data Leakage Detection

Saima Faiyaz
B.Tech,
Department of CSE,
Lords Institute of Engineering and Technology.

M. Jyothi
B.Tech,
Department of CSE,
Lords Institute of Engineering and Technology.

G.Vaishnavi
B.Tech,
Department of CSE,
Lords Institute of Engineering and Technology.

Saba Sultana
Assistant Professor,
Department of CSE,
Lords Institute of Engineering and Technology.

ABSTRACT

We study the following problem: A data distributor has given sensitive data to a set of supposedly trusted agents (third parties). Some of the data is leaked and found in an unauthorized place (e.g., on the web or somebody's laptop). The distributor must assess the likelihood that the leaked data came from one or more agents, as opposed to having been independently gathered by other means. These methods do not rely on alterations of the released data (e.g., watermarks). In some cases, we can also inject realistic but fake‖ data records to further improve our chances of detecting leakage and identifying the guilty party.

1.INTRODUCTION:

Data leakage is the unauthorized transmission of data or information from within an organization to an external destination or recipient. . Data leakage is defined as the accidental or intentional distribution of private or sensitive data to an unauthorized entity. Sensitive data of companies and organization includes intellectual property, financial information, patient information, personal credit card data and other information depending upon the business and the industry. We propose data allocation strategies hat improve the probability of identifying leakages. A data distributor has given this sensitive data to a set of supposedly trusted agents (third parties). Some of the data are leaked and found in an unauthorized place. The distributor must assess the likelihood that the leaked data came from one or more agents, as opposed to having been independently gathered by other means. We propose data allocation strategies (across the agents) that improve the probability of identifying leakages. These methods do not rely on alterations of the released data. In some cases we can also inject "realistic but fake" data records to further improve our chances of detecting leakage and identifying the guilty party.

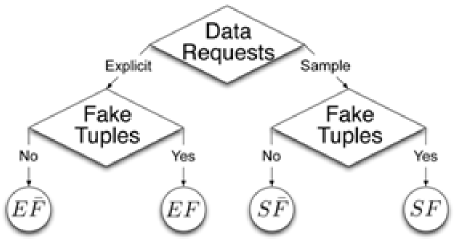
Our goal is to detect when the distributor's sensitive data have been leaked by agents, and if possible to identify the agent that leaked the data. We call the owner of the data the distributor and the supposedly trusted third parties the agents. Our goal is to detect when the distributor's sensitive data have been leaked by agents, and if possible to identify the agent that leaked the data.

2. PROBLEM SETUP AND NOTATION

2.1 Entities and Agents

Let the distributor database owns a set S= {t1, t2,…., tm} which consists of data objects. Let the no of agents be A1, A2,..., An. The distributor distributes a set of records S to any agents based on their request such as Sample or explicit request.

- Sample request Ri= SAMPLE (T, mi): Any subset of mi records from T can be given to Ui [1].
- Explicit request Ri= EXPLICIT (T;condi): Agent Ui receives all T objects that satisfy condition.



Dig: Leakage Problem Instances

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. After giving objects to agents, the distributor discovers that a set S of T has leaked. This means that some third party called the target has been caught in possession of S. For example, this target may be displaying S on its web site, or perhaps as part of a legal discovery process, the target turned over S to the distributor.

Since the agents (A1, A2, ..., An) have some of the data, it is reasonable to suspect them leaking the data. However, the agents can argue that they are innocent, and that the S data was obtained by the target through other means.

2.2 Guilty Agents

Guilty agents are the agents who had leaked the data. Suppose the agent say Ai had leaked the data knowingly or unknowingly. Then automatically notification will be the send to the distributor defining that Agent Ai had leaked the particular set of records which also specifies sensitive or non-sensitive records. Our goal is to estimate the likelihood that the leaked data came from the agents as opposed to other sources.

3. RELATED WORK

The guilt detection approach we present is related to the data provenance problem

[1] Tracing the lineage of an S object implies essentially the detection of the guilty agents.

[2] It provides a good overview on the research conducted in this field.

Suggested solutions are domain specific, such as lineage tracing for data Warehouses

[3] And assume some prior knowledge on the way a data view is created out of data sources.

Watermarks were initially used in images

[4] Video

[5] Audio data

[6] Whose digital representation includes considerable redundancy?

Our approach and watermarking are similar in the sense of providing agents with some kind of receiver identifying information. Finally, there are also lots of other works on mechanisms that allow only authorized users to access sensitive data through access control Policies. Such approaches prevent in some sense data leakage by sharing information only with trusted Parties.

4. RESULTS OF DATA LEAKAGE DETECTION MODEL

4.1 Agent Guilt Model

To compute this PrfGijSg, we need an estimate for the probability that values in S can be “guessed” by the target. For instance, say some of the objects in T are emails of individuals. We can conduct an experiment and ask a person with approximately the expertise and resources of the target to find the email of say 100 individuals. To simplify the formulas that we present in the rest of the paper, we assume that all T objects have the same pt, which we call p. Our equations can be easily generalized to diverse pt’s though they become cumbersome to display.

4.2 Guilt Model Analysis

In order to see how our model parameters interact and to check if the interactions match our intuition, in this section, we study two simple scenarios. In each scenario, we have a target that has obtained all the distributors
Objects, i.e., T ¼ S.

B.1 Impact of Overlap between Ri and S

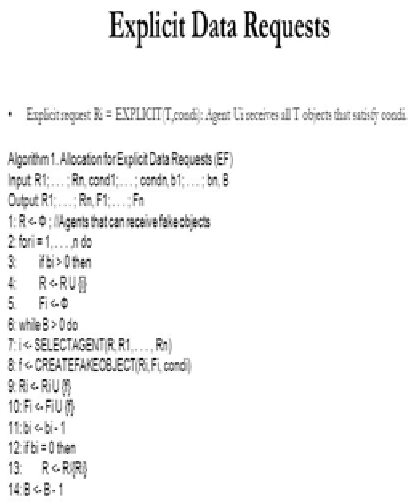
In this section, we again study two agents, one receiving all the T ¼ S data and the second one receiving a varying fraction of the data. The probability of guilt for both agents, as a function of the fraction of the objects owned by U2, i.e., as a function of j R2 \ Sj = j Sj. In this case, p has a low value of 0.2, and U1 continues to have all 16S objects. Note that in our previous scenario, U2 has 50 percent of the S objects. We see that when objects are rare (p ¼ 0:2), it does not take many leaked objects before we can say that U2 is guilty with high confidence. This result matches our intuition: an agent that owns even a small number of incriminating objects is clearly suspicious. The same scenario, except for values of p equal to 0.5 and 0.9. We see clearly that the rate of increase of the guilt probability decreases as p increases. This observation again matches our intuition: As the objects become easier to guess, it takes more and more evidence of leakage (more leaked objects owned by U2) before we can have high confidence that U2 is guilty. In, we study an additional scenario that shows how the sharing of S objects by agents affects the probabilities that they are guilty. The scenario conclusion matches our Intuition: with more agents holding the replicated leaked data, it is harder to lay the blame on any one agent.

5. DATA ALLOCATION STRATEGIES

The data allocation strategies used to solve the problem of data distribution as discussed in previous sections exactly or approximately are provided in the form of various algorithms. The algorithms are provided here.

ISSN No: 2348-4845
International Journal & Magazine of Engineering, Technology, Management and Research
A Peer Reviewed Open Access International Journal

5.1Explicit Data Request



Explicit Data Requests

- Explicit request Ri = EXPLICIT(T,condi): Agent Ui receives all T objects that satisfy condi.

```
Algorithm 1. Allocation for Explicit Data Requests (EF)
Input R1; . . . ; Rn, cond1; . . . ; condn, b1; . . . ; bn, B
Output R1; . . . ; Rn, F1; . . . ; Fn
R <- Φ ; //Agents that can receive fake objects
for i = 1, . . . ,n do
    if bi > 0 then
    R <- R U {i}
    Fi <- Φ
while B > 0 do
i <- SELECTAGENT(R, R1, . . . , Rn)
f <- CREATEFAKEOBJECT(Ri, Fi, condi)
Ri <- Ri U {f}
Fi <- Fi U {f}
bi <- bi - 1
if bi = 0 then
    R <- R/{Ri}
B <- B - 1
```

It is a general algorithm that is used by other algorithms. Our goal is to detect when the distributor's sensitive data has been leaked by agents, and if possible to identify the agent that leaked the data. Furthermore, watermarks can sometimes be destroyed if the data recipient is malicious.

6. EXISTING SYSTEM

We consider applications where the original sensitive data cannot be perturbed. Perturbation is a very useful technique where the data is modified and made "less sensitive" before being handed to agents.However, in some cases it is important not to alter the original distributor's data. Traditionally, leakage detection is handled by watermarking, e.g., a unique code is embedded in each distributed copy. If that copy is later discovered in the hands of an unauthorized party, the leaker can be identified. Watermarks can be very useful in some cases, but again, involve some modification of the original data. Furthermore, watermarks can sometimes be destroyed if the data recipient is malicious.

7. PROPOSED SYSTEM

After giving a set of objects to agents, the distributor discovers some of those same objects in an unauthorized place. At this point the distributor can assess the likelihood that the leaked data came from one or more agents, as opposed to having been independently gathered by other means.If the distributor sees "enough evidence" that an agent leaked data, he may stop doing business with him, or may initiate legal proceedings. In this project we develop a model for assessing the "guilt" of agents. We also present algorithms for distributing objects to agents, in a way that improves our chances of identifying a leaker. Finally, we also consider the option of adding "fake" objects to the distributed set. Such objects do not correspond to real entities but appear.If it turns out an agent was given one or more fake objects that were leaked, then the distributor can be more confident that agent was guilty.

7.1 Watermarking:

We describe a digital watermarking method for use in audio, image, video and multimedia data. The watermark is difficult for an attacker to remove, even when several individuals conspire together with independently watermarked copies of the data. It is distortions such as digital-to-analog and analog-to-digital conversion, resampling, quantization, dithering, compression, rotation, translation, cropping and scaling. The same digital watermarking algorithm can be applied to all three media under consideration with only minor modifications, making it especially appropriate for multimedia products.

7.2 Steganography

Steganography is a technique for hiding a secret message within a larger one in such a way that others can't discern the presence or contents of the hidden message. A plain text message may be hidden in one of two ways, the method of steganography conceal the existence of the message, whereas the outsiders of cryptography render the message unintelligible to transformation of the text. Steganography serves as a means for private, secure and sometimes malicious communication.

8. MODULES:

In this project we have the following modules:
Module 1: Administrator Module
Module 2: Employees Module
Module 3: Data leakage detection module.
Module 1: Administrator Module
All the privileges of the website are only available with the administrator. Admin has privileges to accomplish the following responsibilities/tasks: Manage Users (Add, Edit, Delete,, Assign Permissions). Lock / UnLock.
Manage Articles (Add, Edit, Update, Delete).
Manage Categories (Add, Edit, Update, Delete).
Send Messages.
Upload Documents.

Module 2: Employees Module

Some of the privilages are restricted to the employee by the administrator.

Only little permission is available with the employees. Employees have the following task/ responsibilities. Employees have Read-only access to the content.
. Employees can Read the articles
. Downloads the documents
. Read the messages which are send by the Admin
. Discuss the content in the Discussion Board.

Module 3: Data leakage detection module.

The main scope of this module is provide complete information about the data/content that is accessed by the users within the website. Continuously observation is made automatically and the information is send to the administrator so that he can identify whenever the data is leaked. Forms authentication techniques are used to provide security to the website in order to prevent the leakage of the data. Above all the important aspect providing proof against the guilty objects. The following techniques are used.

– Fake Object Generation.
– Water Marking.

9. OBJECTIVE

The distributor must assess the likelihood that the leaked data came from one or more agents, as opposed to having been independently gathered by other means. These methods do not rely on alterations of the released data (e.g., watermarks). In some cases we can also inject “realistic but fake” data records to further improve our chances of detecting leakage and identifying the guilty party.

10. CONCLUSION

From this study we conclude that the data leakage detection system model is very useful as compare to the existing watermarking model. We can provide security to our data during its distribution or transmission and even we can detect if that gets leaked. Thus, using this model security as well as tracking system is developed. Watermarking can just provide security using various algorithms through encryption, whereas this model provides security plus detection technique.Our model is relatively simple, but we believe that it captures the essential trade-offs. The algorithms we have presented implement a variety of data distribution strategies that can improve the distributor’s chances of identifying a leaker. We have shown that distributing objects judiciously can make a significant difference in identifying guilty agents, especially in cases where there is large overlap in the data that agents must receive. Our future work includes the investigation of agent guilt models that capture leakage scenarios.

REFERENCES

[1]. P.Buneman, S.Khanna, and W.C.Tan, ”Why and Where: A Charaterization of Data provenance,” Proc. Eighth Int’l Conf. Database Theory(ICDT ‘01’),J.V. den Bussche and V.Vianu,eds.,pp.316-330,Jan.2001.

[2].P.Buneman and W.C.Tan,”Provenence in Databases”, Proc ACM SIGMOD, pp.1171-1173,2007.

[3].Y.Cui and J.Widom, ”Lineage Tracing For General Data Warehouse Transformations,” The VLDB J.vol.12, pp.41-58,2003.

[4].J.J.K.O.Ruanaidh, W.J.Dowling, and F.M.Boland,” Watermarking Digital Images For Copyright Protection”, IEE Proc.Vision,Signal and Image Processing, vol.143, no.4, pp.250-256,1996.

[5].F.Hartung and B.Girod,”Watermarking of Uncompressed and Compressed Video,” Signal Processing, vol.66, no.3, pp.283-301,1998.

[6].S.Czerwinski, R.Fromm,and T.Hodes,”Digital Music Distribution and Audio watermarking,”
http://www.Scientificcommons.org/43025658,2007.

[7].S.Jajodia, P.Samarati, M.L.Sapino,and V.S. Subrahmanian,” Flexible Support For Multiple Access .