ISSN:0975-9646

Chandni Bhatt et al, / (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 5 (2) , 2014, 2556-2558

Data Leakage Detection

Chandni Bhatt

PG Student (CSE)
GHRAET India

Prof.Richa Sharma

Asst.Proffessor(CSE)
GHRAET India

***Abstract-* A data distributor has given sensitive data to a set of supposedly trusted agents. Sometimes data is leaked and found in unauthorized place e.g., on the web or on somebody's laptop. For example, a hospital may give patient records to researchers who will devise new treatments. Similarly, a company may have partnerships with other companies that require sharing customer data. Another enterprise may outsource its data processing, so data might be given to various other companies. The owner of the data are called as *distributors* and the trusted third parties are called as *agents.* Data leakage happens every day when confidential business information such as customer or patient data, company secrets, budget information etc..are leaked out. when these information are leaked out, then the companies are at serious risk. Most probably data are being leaked from agent's side. So , company have to very careful while distributing such a data to an agents. The Goal of Our project is to analyze "how the distributer can allocate the confidential data to the Agents so that the leakage of data would be minimized to a Greater Extent by finding an guilty agent".**

Keywords-distributor , fake object , data leakage , watermarking ,guilt agent.

INTRODUCTION

In the course of doing business, sometimes sensitive data must be handed over to supposedly trusted third parties. For example, a hospital may give patient records to researchers who will devise new treatments. Similarly, a company may have partnerships with other companies that require sharing customer data. Another enterprise may outsource its data processing, so data must be givento various other companies. 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 has been *leaked* by agents, and if possible to identify the agent that leaked the data. 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. For example, one can add random noise to certain attributes, or one can replace exact values by ranges . However, in some cases it is important not to alter the original distributor's data. For example, if an outsourcer is doing our payroll, he must have the exact salary and customer bank account numbers. If medical researchers will be treating patients (as opposed to simply computing statistics), they may need accurate data for the patients.

MODULES OF DATA LEAKAGE SYSTEM

I- *Data Allocation Module*

The main focus of our project is the data allocation problem as how can the distributor "intelligently" give data to agents in order to improve the chances of detecting a guilty agent, Admin can send the files to the authenticated user, users can edit their account details etc. Agent views the secret key details through mail. In order to increase the chances of detecting agents that leak data.

II- Fake Object Module

The distributor creates and adds fake objects to the data that he distributes to agents. Fake objects are objects generated by the distributor in order to increase the chances of detecting agents that leak data. The distributor may be able to add fake objects to the distributed data in order to improve his effectiveness in detecting guilty agents. Our use of fake objects is inspired by the use of "trace" records in mailing lists. In case we give the wrong secret key to download the file, the duplicate file is opened, and that fake details also send the mail. Ex: The fake object details will display.

III-Optimization Module

The Optimization Module is the distributor's data allocation to agents has one constraint and one objective. The agent's constraint is to satisfy distributor's requests, by providing them with the number of objects they request or with all available objects that satisfy their conditions. His objective is to be able to detect an agent who leaks any portion of his data. User can able to lock and unlock the files for secure.

IV-*Data Distributor Module*

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 Admin can able to view the which file is leaking and fake user's details also.

V- Agent Guilt Module

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. If this person can find say 90 emails, then we can reasonably guess that the probability of finding one email is 0.9. On the other hand, if the objects in question are bank account numbers, the person may only discover say 20, leading to an estimate of 0.2. We call this estimate pt, the probability that object tcan be guessed by the target. 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 p t's though they become cumbersome to display. Next, we make two assumptions regarding the relationship among the various leakage events. The first assumption simply states that an agent's decision to leak an object is not related to other objects.

LITERATURE REVIEW

I-Watermarking Technique

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. E.g. A hospital may give patient records to research who will devise new treatments. Similarly, a company may have partnerships with other companies that require sharing customer data. Another enterprise may outsource its data processing, so data must be given to various other companies. We call the owner of the data the distributor and the supposedly trusted third parties the agents.

A.Disadvantages:

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.

II-Data Allocation Strategy

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 (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. We also present algorithm for distributing object to agent. 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. Perturbation is a very useful technique where the data is modified and made 'less sensitive´ before being handed to agents. We develop unobtrusive techniques for detecting leakage of a set of objects or records. In this section 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 realistic to the agents. In a sense, the fake objects acts as a type of watermark for the entire set, without modifying any individual members. 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. Today the advancement in technology made the watermarking system a simple technique of data authorization. There are various software which can remove the watermark from the data and makes the data as original.

III- DATA LEAKAGE DETECTION USING CLOUD COMPUTING

In the virtual and widely distributed network, the process of handover sensitive data from the distributor to the trusted third parties always occurs regularly in this modern world. It needs to safeguard the security and durability of service based on the demand of users. The idea of modifying the data itself to detect the leakage is not a new approach. Generally, the sensitive data are leaked by the agents, and the specific agent is responsible for the leaked data should always be detected at an early stage. Thus, the detection of data from the distributor to agents is mandatory. This project presents a data leakage detection system using various allocation strategies and which assess the likelihood that the leaked data came from one or more agents. For secure transactions, allowing only authorized users to access sensitive data through access control policies shall prevent data leakage by sharing information only with trusted parties and also the data should be detected from leaking by means of adding fake record`s in the data set and which improves probability of identifying leakages in the system. Then, finally it is decided to implement this mechanism on a cloud serverKey to the definition of cloud computing is the ―cloud‖ itself. For our purposes, The cloud is a large group of interconnected computers. These computers can be personal computers or network servers; they can be public or private. For example, Google hosts a cloud thatconsists of both smallish PCs and larger servers. Google's cloud is a private on(that is, Google owns it) that is publicly accessible (by Google's users).This cloud of computers extends beyond a single company or enterprise. The applications and data served by the cloud are available to broad group of users, cross-enterprise and cross-platform. Access is via the Internet. Any authorized user can access these docs and apps from any computer over any Internet connection. And, to the user, the technology and infrastructure behind the cloud is invisible. It isn't apparent (and, in most cases doesn't matter)whether cloud services are based on HTTP, HTML, XML, Java script, or other specific technologies. From Google's perspective, there are six key properties of cloud computing.

CONCLUSION

In a perfect world there would be no need to handover sensitive data to agents that may unknowingly ormaliciously leak it. And even if we had to hand over sensitivedata, in a perfect world we could watermark eachobject so that we could trace its origins with absolutecertainty. However, in many cases we must indeed workwith agents that may not be 100% trusted, and we maynot be certain if a leaked object came from an agent orfrom some other source, since certain data cannot admitwatermarks. In spite of these difficulties, we have shown it is possibleto assess the likelihood that an agent is responsible for a leak, based on the overlap of his data with theleaked data and the data of other agents, and based onthe probability that objects can be "guessed" by othermeans. Our model is relatively simple, but we believeit captures the essential trade-offs. The algorithms wehave presented implement a variety of data distributionstrategies that can improve the distributor's chances ofidentifying a leaker. We have shown that distributingobjects judiciously can make a significant difference inidentifying guilty agents, especially in cases where thereis large overlap in the data that agents must receive. Our future work includes the investigation of agentguilt models that capture leakage scenarios that are notstudied in this paper. For example, what is the appropriatemodel for cases where agents can collude andidentify fake tuples? A preliminary discussion of sucha model is available in . Another open problem isthe extension of our allocation strategies so that they canhandle agent requests in an online fashion (the presentedstrategies assume that there is a fixed set of agents withrequests known in advance).Another and most problem to be solved is protecting data before getting leaked.

REFERENCES

[1*]* *P. Papadimitriou and H. Garcia-Molina*, "Data leakage detection," IEEE Transactions on Knowledge and Data Engineering, pages 51-63, volume 23, 2011.

[2*]* *R. Agrawal and J. Kiernan*, "Watermarking Relational Databases," Proc. 28th Int'l Conf. Very Large Data Bases (VLDB '02), VLDB Endowment, pp. 155-166, 2002

[3] *Hartung and kutter*,‖Watermarking technique for multimedia data‖2003.

[4] *Chun-Shien Lu*, Member, IEEE, and *Hong-Yuan Mark Liao*, Member, IEEEMultipurpose Watermarking for Image Authentication and Protection

[5] *Mr.V.Malsoru, NareshBollam*/ REVIEW ON DATA LEAKAGE DETECTION, International Journal of Engineering Research and Applications (IJERA) ISSN: 2248-9622 www.ijera.com Vol. 1, Issue 3, pp.1088-1091 1088 | P a g e.

[6] *YIN Fan*, *WANG Yu*, *WANG Lina*, *Yu Rongwei*. A Trustworthiness-Based Distribution Model for Data Leakage Detection: Wuhan University Journal Of Natural Sciences.

[7] *L. Sweeney*. Achieving k- anonymity privacy protection using generalization and suppression. International Journal on Uncertainty, Fuzzyness and Knowledge-based Systems-2002

[8] *S. Czerwinski, R. Fromm*, and *T. Hodes*.Digital music distribution and audio watermarking.

[9] *P. Buneman*, *S. Khann*a and *W.C. Tan*. Why and where: A characterization of data provenance. ICDT 2001, 8th International Conference, London, UK, January4-6, 2001, Proceedings, volume 1973 of Lecture Notes in Computer Science, Springer, 2001

[10] *L. Sweeney*. Achieving k- anonymity privacy protection using generalization and suppression. International Journal on Uncertainty, Fuzzyness and Knowledge-based Systems-2002

[11] *Edward P. Holden, Jai W. Kang, Geoffrey R. Anderson, Dianne P. Bills*,Databases in the Cloud: A Work in Progress,2012

[12*]* *Michael Miller*, ‖Cloud Computing‖ Web-Based Applications that change the way you work and Collaborate Online,Pearson Education,2012

[13] *Rudragouda G Pati*l, "Development of Data leakage Detection Using Data Allocation Strategies International Journal of Computer Applications in E ngineering Sciences[VOL I, ISSUE II, JUNE 2011

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

[15] *N.Sandhya , K. Bhima , G. Haricharan Sharma*," Exerting Modern Techniques for Data Leakage Problems Detect ".International Journal of Electronics Communication and Computer Engineering-2012, Volume 3, Issue 1.

[16] *ArchanaVaidya, PrakashLahange, Kiran* More, *ShefaliKachroo&NiveditaPandey*. "DATA LEAKAGE DETECTION ". IJAET, 2012