DATA LEAKAGE DETECTION

SHIVA NANDHINI , P.NITHIN BHARADWAJ,
P.SAITEJAREDDY, P.A.SREE HARSHA
DEPARTMENT OF COMPUTER SCIENCE, SRM INSTITUE OF SCIENCE AND TECHNOLOGY, CHENNAI, INDIA.

ABSTRACT

Data or information is major thing and also valuable. Now-a-days this data leakage has becoming a big problem. To get rid of this first we have to differentiate the data into important and unimportant, identifying important data became a big challenge for the organizations. A system has been bloomed with help of data mining techniques, which identifies intimate data of an organization. The process of dealing out sensitive or confidential data from the dealer to the certitude third parties always happens regularly in this modern world. Generally, the confidential data which is leaked by the agents, and the specific agent who is in helm for the leaked data should always be detected at an premature stage. Thus, the detection of data from the dealer to agents is obligatory. The project accords a data leakage detection structure consisting of various allocation techniques and which guage the probability that the leaked data came from agents. For secure proceedings, allowing only authorized users to access intimate data through access control policies shall avoid data leakage by sharing data only with trusted parties. But, in some cases we have to share data with other partner companies, in such case we use unobtrusive techniques to find guilty agents.

Key words: Data leakage, Guilty agents, Distributor.

INTRODUCTION

Data leakage is the big challenge in front of the industries & different institutes.Eventhough we got many encryption techniques for the security of intimate data of organizations,there is a integrity drawback with the users of those problems.The tracing of the leaker is very tough for the adminstrator as he is one the emplyee with his system. It creates a lot many issues regarding the ethical values in the working environment of the office. The adminstrator or distributor can be able to analyse the situation of what data is being leaked and how many are leaking that, as opposed to having been independently gathered by other means. If we find anyone with data of ours they can easily escape by saying that it has been given by one of their friends.For example if you consider a cookie jar and a cookie is stolen by george it is easy to argue,but if we are able to catch him with more than one cookie in his hand then it is very difficult for him to say that he has not kept his hands in that cookie jar. If the the data distributor has the enough information about the agent that has been leaking the data then he can end his work with him or proceed in a legal way. In the project we are proposing a tehnique for knowing the 'guilt´ of agents. We are also going to give many distributing techniques for distributing the data so that they can help us in catching the guilty agent. Finally, we also consider the option of adding 'fake´ objects to the distributed set. These objects are not tied with any real entities but they look very natural to everyone. In a way,these objects which we added acts as watermark for the whole set,without modifying each of the members. If it shows an user was provided with more fake things that have been leaked, then the dealer can be more assured that culprit was guilty. Fake objects are improvised by the distributor so as to increase the option of finding agents that who are reason for leakage of data. Generally these fake objects may lay some affect on the correctness of the agents in what they did so, they are not that much capable [1]. This petrubing technique is not a present one and there are some drawbacks in that also but we have seen many using pertrubated objects, e.g.,adding watermark or some fake objects in some sensitive areas by pertrubing is done in the adminstrator data.The pertrubing of real objects becomes a problem because of the fake objects in some applications. A hospital may sent patient records to researchers who will give new treatment. Similarly, a company may had different partnerships withdifferent companys that requires sharing the customer data.different enterprises may give out source its data process to data must be sended to different company We call the owner of the data the distributor and the supposedly trusted third parties the agents. Our aim to detect the distributor's important data will be leaked by the agents, and also possible to observe the agent that particular leaked the data. We can consider application the sensitive data cannot able to perturbed. Perturbion is a very used technique where data can be modiy and also "less data" before being hand to agent.Take example , one can able to add random way of noise to particular sutiable in certain attribute, or else it can replace same value by ranges. However, in some cases it is important not to alter the original distributor's data.For an instance, for an outside person to do payment he should have the details of the account.There will be some medical researchers who will be doing tratments for them the data should be more accurate.coventionaly this watermark technique is used to detect the leakage e.g., a different code is implimented in each individual copy. Normally these watermarks are useful but, what happenes is we need to modify the original data which is a problem. This paper depicts about the unubtrusive techniques to detect the leakage with a set of objects.

EXIXTING SYSTEM

Generally, leakage detection problem is encountered by watermarking, e.g., a unique code is implanted in each distributed copy. To dispense copyright protection for vedio and digital audio data, two reciprocal techniques are being progressed: encryption and watermarking. We can retain our data and can say that is our datasets. So we can dispense watermark to each dataset. If that copy later comes across in the control of an unauthorized party, that agent can be identified. Watermarks can be capable in some instances, but we, needs some alteration of the intimate data. Moreover, in some instances the destroying of watermark is also possible if the agent is malicious. E.g. A hospital may provide patient records to researchers who will conceive new treatments. Similarly, a company may have partnerships of other companies that require sharing customer data. Another enterprise may utilize its data processing, so it must be handed over to many other companies. We call the possessor of the data dealer and the apparently certitude third parties the agents.

PROPOSED SYSTEM

The main focus point on which we are looking is to detect the intimate data leakage very fast, and if able to catch the culprit that leaked the data. We develop *unobtrusive* technique s for finding leakage of a bunch of objects. We progress a model for guaging the “culpability” of agents. We also provide algorithms for distributing instances to agents, to increases our chances of finding a culprit. Finally, we also providing the choice of implementing “fake” objects to the distributed thing. Such objects do not relate to today life examples but appear realistic to the agents. In a way,these objects which we added acts as watermark for the whole set, without modifying each of the members. If it shows an user was provided with more fake things that have been leaked, then the dealer can be more assured that culprit was guilty.

DISADVANTAGES

The above-mentioned techniques are very useful for detecting data leakage. But, hacking the data has been grown more. so, we need some new techniques to overcome this problem. There are some drawbacks in above techniques those are:

If we are using fake objects technique then, if the third party is malicious there is a chance of removing watermarks. There is a disadvantage like it is difficult to divide data to different distributors. For these fake objects creation, we need modify the data.suppose if we have Financial support like budget or employee's funds, such sensitive support cannot be able to altered at any alteration it can be lead to more financial critises of the company.

ADVANTAGES

in some areas it can also interacts and also to check if their any interactions it can match our intention, in this section we will study two simple scenarios as In this more Probability p and In this system both RI and S. In every particular situation we know the target that we need, i.e., T = S. By this two methods we can find the leaker and to whom the file has been sent.so, this makes more secure and also we can catch the agent with more evidence.

MODULES

1. Data Distributing Module

The main part of the project is the problem of providing intimate data so, how can he “wisely” provide data to agents so as to increase the chances of finding a guilty agent, Admin can provide the files to authorized user so, user can change it their account information etc. Agent can get the secret key information through mail. so as to improve the posibilities of finding agents that leak data.

2. Fake Object Module

The distributor fabricates and adds duplicate objects to the intimate data that he provides to agents. Fake objects are improvised by the distributor so as to increase the option of finding agents that who are reason for leakage of data. The distributor is skilled of adding fake objects to intimate data so as to revamp his effectiveness in finding guilty agents. The advantage of fake objects is stimulated by the use of “trace” records in mailing registry. In case of giving the fake secret key to get the file, the fake file is unbarred, and that duplicate details also posted to the mail. The fake object information will be shown.

C. Optimization Module

In the Optimization Module is distributor's data distribution to users has one curtailment and one intent. The agent's curtailment is to satisfy distributor's requests, by giving them with the amount of objects they have asked or with all posible objects that satisfy their needs. His idea of doing this process is to catch the culprit behind it. User can able to lock and unlock the files for secure.

D. Data Distributor Module

This module has given the intimate data to the unauthorised persons. Sometimes we find our data in unhabited area like web (e.g., on the web or somebody's laptop). The distributor should know the leakage of data whether it is done by more than one person so as to analyse the situation, as against to having been separatly gathered and see which file is leaked and fake user's details also.

E. Agent Guilt Module

To calculate this Guilt, we need an guessed for prospect that values in S can be “estimated” by the target. For specimen, say some of the objects in T are emails of every single person.an experiment can be conducted by a person having almost same experience and knowledge of an email in 100 emails of the people. For a case the person is capable of finding 90 email, from that we can estimate it took 0.9 seconds per email. Similarly, the person can only discover 20 that is around 2.0 if the object are the question of the bank count numbers. Naturally that this estimation pt can be said as the probability of finding the target. The formulas we used in the papaer are made easier to understand by assuming that all T objects has the same pt, which we call p. the equations are easily generalised by pt, althought it's difficult to display.After this we make two assumptions about the leakage events. First assumption says that the aim of leaking the data is not related other objects.

CONCLUSION

From the study we can finally say that the data leakage model is very useful to compare the current watermarking model. We can give more security to the 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 be provide more securityto use different algorithm to encrypit, whereas this model provides security plus detection technique. This model is very helpful in various industries, where data is distribute through any public or private channel and shred with third party. Now, industry & various offices can rely on this security & detection model.

ACKNOWLEDGEMENT

For all the efforts behind the paper work, I first and foremost would be happy to say my sincere appreciation to the staff of Dept. of Computer Sci.& Engg., for their extended help & suggestions at every stage of this paper. It is with a great sense of gratitude that I acknowledge the support, time to time suggestions and highly indebted to my guide Prof. S.Nandhini (my project guide), and Dr.jagadeshen (HOD). Finally, I pay sincere thanks to all those who indirectly and directly helped me towards the successful completion of the paper

REFERENCES

1) Chun-Shien Lu, Member, IEEE, and Hong-Yuan Mark Liao, Member, IEEE MultipurposeWatermarking for Image Authentication and Protection

2)Sandip A.Kale, Prof. Kulkarni S.V. (Department Of Computer Sci. &Engg,MIT College of Engg, Dr.B.A.M.University, Aurangabad(M.S), India, Data Leakage Detection: A Survey, (IOSR)

3) Panagiotis Papadimitriou 1, Hector Garcia-Molina 2 Stanford University 353 Serra Street, Stanford, CA 94305, USA P.P (1, 4-5) A Model for Data Leakage Detection

4)Web-based Data Leakage Prevention Sachiko Yoshihama1, Takuya Mishina1, and Tsutomu Matsumoto2 1 IBM Research - Tokyo, Yamato, Kanagawa, Japan fsachikoy

5) Joseph A. Rivela Senior Security Consultant P.P (4-6) Data Leakage: Affordable Data Leakage Risk Management

6) Data Leakage Prevention: A news letter for IT Professionals Issue 5 P.P (1-3)

7) Panagiotis Papadimitriou, Student Member, IEEE, and Hector Garcia-Molina, Member, Data Leakage Detection IEEE P.P (2-6)IEEE transactions on knowledge and data engineering, vol. 23, no. 1, JANUARY 2011

8)Archie Alimagno California Department of Insurance P.P (2- 7),The Who, What, When & Why of Data Leakage Prevention/Protection [13] An ISACA White Paper Data Leak Prevention P.P (3-7)

9) Mr.V.Malsoru, Naresh Bollam/ REVIEW ON DATA LEAKAGE DETECTION , International Journal of Engineering Research and Applications (IJERA) ISSN: 2248-9622 www.ijera.com Vol. 1, Issue 3,

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)RudragoudaG Patil, "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 Engineering201.