Data Leakage Detection Model for Finding Guilty Agents

J.P.Chavan[1], Ashwini Choutmol[2], Rahul Shelke[3], Satish Shelar[4]

[1,2,3,4] Department of Computer Engineering, Sinhgad Institute of Technology, Lonavala, India

Abstract: ***The following problem may occur in real world scenario: A data distributor has given sensitive data to a set of trusted agents (third parties or third persons).It may happen that some of the data is leaked and found in an unauthorized place (e.g., on the web or unauthorized person's laptop). The distributor must assess the probability of specified outcome that the leaked data came from one or more agents. In this paper, we implement methods aimed at improving the odds of detecting such leakages when a distributer's sensitive data has been leaked by trustworthy agents and also to possibly identify the agent that leaked the data. By adding fake objects to distributed set, the distributor can find the guilty party.***

Keywords: guilty agents, data distributor, data leakage, fake records, leakage detection.

1. Introduction

Data leakage is the unauthorized transmission of data or information from within an organization to unauthorized parties. Data leakage is defined as the distribution of private or sensitive data to an unauthorized person. Private data of companies and organization includes financial information, employee's personal information and other information depending upon the business. Sometimes sensitive data may hand over to supposedly trusted third parties. This increases the risk that confidential information will fall into unauthorized hands, whether caused by force or by mistake. The problem of data leakage is much more relevant and crucial nowadays as much of our information is available online through social networking sites and third party aggregators.[1]Recent years have seen an increasing number of agents being developed to extend the particular environment of human. Those agents, with their autonomous reasoning and decision-making capability, can participate in complex interactions on behalf of their owners. There is no single agent system. Instead, agents usually live in a society of agents, which is known as multi agent system. Usually, agents in MAS represent various stakeholders, each with distinct interests and goals. They try to pursue their own goals, even at the cost of others.

The 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.[3] An application where the original sensitive datacannot be perturbed is considered. Perturbation is a veryuseful technique where the data is changed in form of characters and made "lesssensitive" before being handed to agents. For example, onecan add random noise to certain attributes, or one canreplace exact values by ranges. However, in some cases it isimportant not to alter the original distributor's data. Forexample, if an outsourcer is doing our payroll, he must havethe exact salary and customer bank account numbers. Ifmedical researchers will be treating patients (as opposed tosimply computing statistics), they may need accurate datafor the patients [3]

2. Objective

A data failure is the unintentional release of secure information to an unauthorized environment. The goal is to estimate the likelihood that the leaked data came from the agents as opposed to other sources. Not only to we want to estimate the likelihood the agents leaked data, but we would also like to find out if one of them was more likely to be the leaker. The data allocation strategies help the distributor "intelligently" give data to agents. Fake objects are added to identify the guilty part, to address this problem four instances are specified. Depending on which the data request is provided. Depending upon the type of data request, the fake objects are allowed.

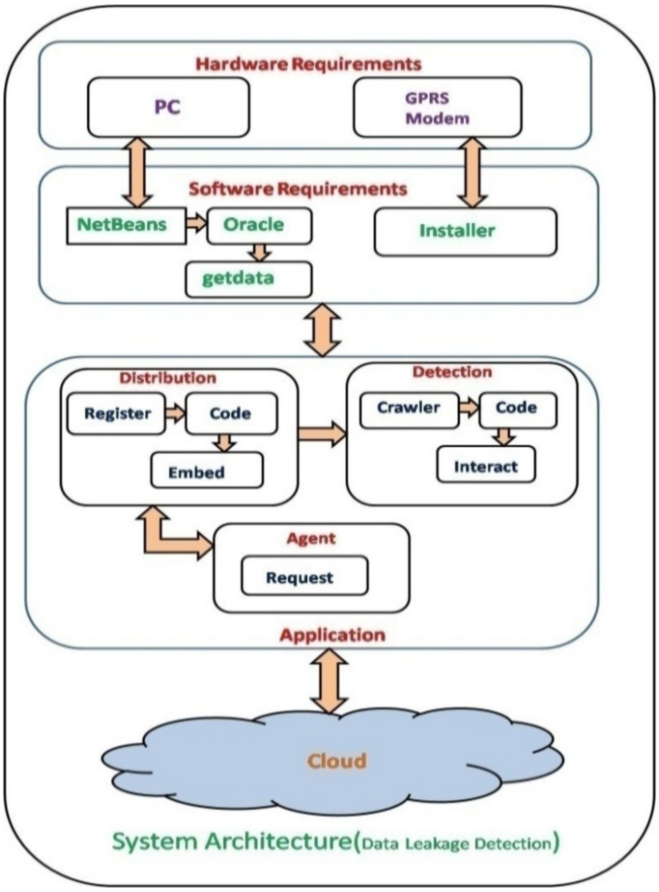

Figure 1: System Architecture

3. System Architecture

The system architecture of data leakage detection uses the cloud. The cloud is a large group of interconnected computers. These computers can be personal computers or network servers; they can be public or private. Here in the system distributor sends embedded document to the agent. And the system uses web crawlers to search the embedded document ifsome of the data is leaked and found in an unauthorized place. Web crawlers are programs or automated script that gather and locate information on the web in a methodical, automated manner.

4. Existing System

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.[4] Watermarks were initially used in images, video and audio data whose digital representation includes considerable repetition in messages. Watermarking aims to identify a data owner and, hence, is subject to attacks where a pirate claims ownership of the data or weakens a businessperson's claims.

5. Proposed System

It is possible to assess the likelihood that an agent is responsible for a leak, based on the overlap of his data with the leaked data and the data of other agents, and based on the probability that objects can be "guessed" by other means. This model is relatively simple, but it is considered that it captures the essential trade-offs. The algorithms which are presented implement a variety of data distribution strategies that can improve the distributor's chances of identifying a leaker. It is shown that distributing objects characterized by good judgment can make a significant difference in identifying guilty agents, especially in cases where there is large overlap in the data that agents must receive. In this project, the model for assessing the "guilt" of agents is developed. The option of adding "fake" objectsto the distributed set is considered. Such objects do not correspond to real entities but appear realistic to the agents.[6]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.

6. Problem Definition

The distributor's data allocation to agents has one and one objective. The distributor's constraint is to satisfy agents' 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.[6]The constraint is considered as strict. The distributor may not deny serving an agent request and may not provide agents with different perturbed versions of the same objects. For this fake object distribution is the only possible constraint relaxation. The detection objective is deal and tractable. The main objective to maximize the chances of detecting a guilty agent that leaks all his data objects.

7. Related Work

The guilt detection approach we present is related to the data provenance problem: tracing the lineage of an subject implies essentially the detection of the guilty agents.[5]It provides a good overview on the research conducted in this field. Suggested solutions are domain specific, such as lineage tracing for data Warehouses, and assume some prior knowledge on the way a data view is created out of data sources. Our problem formulation with objects and sets is more general and simplifies lineage tracing, since we do not consider any data transformation from Ri sets to S.As far as the data allocation strategies are concerned, our work is mostly relevant to watermarking that is used as a means of establishing original ownership of distributed objects. Watermarks were initially used in images, video and audio data whose digital representation includes considerable redundancy.

8. Modules of Data Leakage Detection System

A. 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.[5]

B. 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.[5] 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.

C. 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.

D. 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.

E. Agent Guilt Module

To compute this, 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 100individuals. 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 questionnaire 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 t can 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 pt'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.

9. Conclusion

In a perfect world there would be no need to hand over sensitive data to agents that may unknowingly or maliciously leak it. And even if we had to hand over sensitive data, in a perfect world we could watermark each object so that we could trace its origins with absolute certainty. Our model is relatively simple, but we believe it captures the essential trade-offs. Our future work includes the implementation of data allocation strategies for explicit data requests. We will also extend our work to handle agents' requests in an online fashion, i.e. when the number of agents and agent requests are not known in advance. The FCFS technique presented above can be extended to handle agents' requests in an online fashion as we know that the data allocation to an agent, in this case, does not depend on the agents that present their request after the first.

10. Acknowledgment

We express our thanks to all those who have provided us valuable guidance towards the completion of this project as part of the syllabus of bachelor's course. We express our sincere gratitude towards co-operative department who has provided us with valuable assistance and requirements for the system development.

We hereby take this opportunity to record our sincere thanks and heartily gratitude to our Project guide, Prof. J. P. Chavan for her useful guidance, making available to us her valuable knowledge and experience in making Automotive Telematics as project. We are also thankful to our project co-ordinator Prof. V. Dhavas for her constant enlightenment and motivation which has been highly instrumental in making our project. The acknowledgment will be incomplete if we do not record our sense of gratitude to our HOD of the Computer department, Prof. T. J. Parvat and Principal Dr. R. S. Gaikwad who gave us necessary guidance, encouragement by providing us all the facilities to work on this project.

References

[1] Aay Kumar, Ankit Goyal, Ashwani Kumar, Navneet Kumar Chaudhary, Sowmya Kamath S, "Comparative Evaluation of Algorithms for Effective Data Leakage Detection". Proceedings of 2013 IEEE Conference on Information and Communication Technologies (ICT 2013).

[2] Jiangjiang Wu, Jie Zhou, Jun Ma, Songzhu Mei, Jiangchun Ren, "An Active Data Leakage Prevention Model for Insider Threat", 2011 International Symposium on Intelligence Information Processing and Trusted Computing.

[3] Panagiotis Papadimitriou, Hector Garcia-Molina," Data Leakage Detection". IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 23, NO. 1, JANUARY 2011

[4] Panagiotis Papadimitriou, Hector Garcia-Molina, " A Model for Data Leakage Detection", IEEE International Conference on Data Engineering.

[5] Sandip A. Kale, Prof. S.V.Kulkarni," Data Leakage Detection". International Journal of Advanced Research in Computer and Communication EngineeringVol. 1, Issue 9, November 2012

[6] S. Umamaheswari, "Detection Of Guilty Agents", Proceedings of the National Conference on Innovations in Emerging Technology-2011Kongu Engineering College, Perundurai, Erode, Tamilnadu, India.17 & 18 February, 2011.pp.23-26.

Author Profile

Prof. J. P. Chavan received her B.E. and M.E. degrees in Computer Engineering from Shivaji University (2003-2006) and Pune University respectively.Currently she is Assistant Professor at Sinhgad Institute of Technology, Lonavala, India

Ashwini Choutmol pursuing her B.E. degree in Computer Engineering at Pune University in Sinhgad Institute of Technology, Lonavala 2014 Batch

Rahul Shelke pursuing his B.E. degree in Computer Engineering at Pune University in Sinhgad Institute of Technology, Lonavala 2014 Batch

Satish Shelar pursuing his B.E. degree in Computer Engineering at Pune University in Sinhgad Institute of Technology, Lonavala 2014 Batch