Guilt Model Process for Identifying Data Leakage and Guilty Agent in Data transmission

[1]S.Jenila, [2]K.Sivasankari, [3]R.Arudselvi, [4]J.Maria Monica, [5]B.Saranya
[1,2]M.E – Software Engineering, [3,4,5] M.E – Computer Science and Engineering
Periyar Maniammai University, Thanjavur, Tamil Nadu, India

ABSTRACT

For specific data processing activities a data distributor has to give the sensitive data to a set of third party agents. The distributor’s copy of data that had been transmitted to agents may be leaked by any of them. It is necessary to identify the agents that have leaked data. To improve the chances of detecting leakage data and the guilty agents, private object is created for each record that is sent to various agents. The agent guilt model is used to find the probability of identifying agents that have leaked information. Thus it helps the distributor to turn away from that agent and also to protect the responsive data. Further process can clear the data if the agent had sent data to unauthorized person.

Keywords

Data leakage, guilt model, private object

1. INTRODUCTION

The sensitive data which are significant in many fields needs to be processed in order to obtain profound information from it. Some companies have to give data to other enterprises for outsource its data. The outsourcer may be an expert to accomplish payroll of a company with details such as customer salary, account number etc. or else a researcher who can get patient details from hospital to formulate new medicines. Information such as customer details, patient data, source code, trade secrets, design specifications forecasts and budgets in spreadsheets can be leaked out. The data are possessed by the distributor and are given to trusted third parties called agents. An enterprise data leak is a scary proposition. Security practitioners have always had to deal with data leakage issues that arise from email, and other Internet channels. But now with the proliferation of mobile technology, it's easier than ever for data loss to occur, whether accidentally or maliciously. there are plenty of tools on the market for keeping mobile and stationary data from leaving the company surreptitiously, the best ones use a combination of prevention methods, such as a detection engine and a data blocker. The distributor’s sensitive data can be leaked by the agents to some unauthorized persons. A data leakage is the unintentional release of secure information to an untrusted environment. The guilty agents should be detected and the data must be protected from them. he option of adding “private” objects to the distributed set.

The technique of embedding some image or code is used in the copy of the data to be distributed to agents. Watermarks can be very useful in some cases, but involve some modification of the original data. Furthermore, watermarks can be sometimes destroyed by some of the malicious data recipients. By this the leaked data is identified in any irrelevant place. If some of those same objects are found in an unauthorized place, the distributor can determine that the leaked data came from any of his agents. If the distributor gets evidence that an agent leaked data, he may stop doing business with that agent. A model of agent guilt model is assessed to find the guilty agent. Different stakeholders that have their own distinct and sometimes conflicting objectives are given data by the agents within multi-agent systems. Nowadays, the complex interactions and important decision making activities are done by the agents without the involvement of the distributor. They would behave in such a way in order to achieve their own objectives at the low cost or with the cost of others. Mobile devices present yet another challenge for data leakage. USB keys, Bluetooth devices or removable CD drives, for example, can all circumvent network controls without a system administrator's knowledge.

In addition to traditional malware defenses, encryption and access controls play a huge role in protecting sensitive data from insiders no matter where the data rests or how it being acted upon. Equally important is the ability to filter, log, and take action on outbound traffic and downloads, which is commonly referred to as Data Leakage Protection. For example, other authorized users, such as business partners who hold or handle sensitive information, will need to prove they’re also using the encryption and security standards you’ve laid out for protecting sensitive data. With the increase in technological processes for many complex applications, there is no single agent system used in any of the enterprises or organizations. Instead, agents live in an enlarged system known as multiagent system. Usually, agents in multiagent system represent various stakeholders and agents, each with discrete interests and objectives. Some traditional techniques are also used for the detection of data leakage. But those techniques modify the sensitive data and the modified copy only received by the agents for further processing.

2. RELATED WORK

The data leakage depends on the source from which the data is taken and the process of extracting data from it, which are given as the provenance of the data [1]. It determines the quality and amount of trust one places on the results [2]. We consider applications where the original sensitive data cannot be perturbed. The idea of perturbing data to detect leakage is not new. In most cases, individual objects are perturbed, i.e., by adding random noise to sensitive salaries, or adding a watermark to an image. In this case, perturbing the set of distributor objects by adding fake elements is done. In some applications, fake objects may cause fewer problems that perturbing real objects. For example, say the distributed data objects are medical records and the agents are hospitals. In this case, even small modifications to the records of actual patients may be undesirable. Perturbation is a very useful technique where the data are modified. The data can be made “less sensitive” before being handed to agents [8].

One can add random noise to certain attributes, or one can replace exact values. Watermarks were initially used in images and the alterations performed by the watermark embedding should be unidentifiable by the human visual system in images [6], and audio data [3], which is either a possible approach to balancing these concerns or a failure waiting to happen depending on artists or consumers. Traditionally, leakage detection is handled by watermarking; a watermark applied to an object modifies the item being watermarked. It aims to identify a data owner and is subject to attacks where a pirate claims possession of the data which weakens the merchant's asserts. If the object to be watermarked is not modified, then a watermark cannot be inserted. With respect to [7], particular data uses and metrics of quality, it is of utmost importance that the watermarking process not interferes with the final data consumer requirements. An attacker with the intent of removing/altering the watermark is now faced with the fact that any further alterations performed have an increased likelihood of making the data invalid.

In such cases, methods that attach watermarks to the distributed data are not relevant. Other works on mechanisms that allow only authorized users to access sensitive data through access control policies [4]. If we use probability that can be guessed by the third party that are higher than the true values, we will know that the agents will be guilty [5].

3. DATA LEAKAGE PROBLEM

The data leakage can occur due to improper access control in systems containing sensitive data, inappropriate transfer of information via e-mail, file transfer etc., The lack of access control process may lead to weak safety of data records. Some of the data may be leaked from the stolen computers, mobile devices, portable storage devices and laptops. A desktop stolen from an administrative system can contain personal details may have included name, date of birth, mailing address, and Social Security number, depending on the service being provided to that agent.

The process by which the data are leaked is illustrated in fig. 1. The data distributor can have a data set D={d_1,d_2,...d_n} which are to be distributed to trusted agents A_1,A_2,...A_n. The agents responsibility is to process the data based on the distributor's requirement. It is possible that some of the agents can send the data to unauthorized agent A_u.

The data distributor has given authorization to agents A_1,A_2,...A_n for accessing data and also can give permission to share data within that group of agents. But sometimes these agents may distribute data to unauthorized agent A_u. Some of the agents A_2 and A_n are forwarding the sensitive data to the agent A_u. If the data is shared between distributor and agents then there will be an authorized flow of data otherwise it will be considered as leakage of data. For example, say that one of the objects in D represents a customer X. Perhaps X is also a customer of some other company, and that company provided the data to the target.

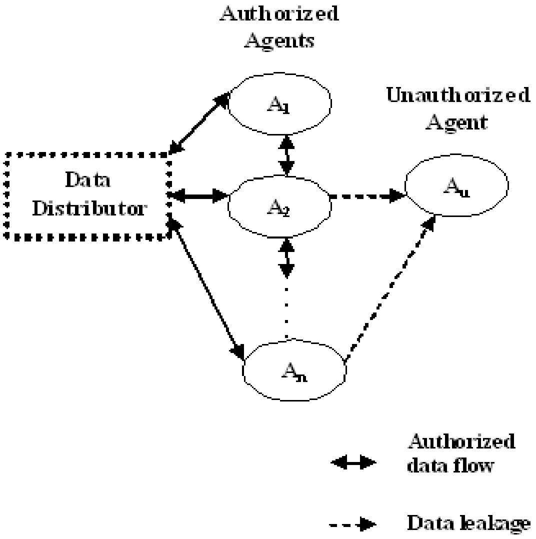

Figure 1. Data leakage

For instance, if one of the D objects was only given to agent A_1, while the other objects were given to all agents, we may suspect A_1 more. We say an agent A_1 is guilty and if it contributes one or more objects to the target. There may be any number of A_u those are not granted permission by the distributor for accessing the data. To simplify the formulas that we present in the rest of the paper, we assume that all D objects have the same probability, which we call p. Our equations can be easily generalized to diverse p though they become clumsy to display. This process should be detected by the distributor through some data leakage identification techniques. This includes the technique of embedding any code with the copy of the data to be distributed to agents. However, these are easily cracked by the receiver. So this should be monitored at the distributor side itself. The data transaction at the agents is observed and private objects are created for each distribution.

4. IDENTIFICATION METHOD

The data distributor having a data set D={d_1,d_2,...d_n} which are to be distributed to any of the agents A_1,A_2,...A_n. The agents are given data to process based on the distributor's requirement. The identification method includes creation of private objects for the data that are sent from each agent. Detection would be assured only if the distributor gave no data object to any agent to attain perfect privacy and security. We use instead the following objective as to maximize the chances of detecting a guilty agent that leaks all his data objects. The particular data which are can be detected and the guilty agents who are involved in this activity can also be determined. The model for assessing the guilt of agents is developed. The process of adding private objects for the distributed set is considered. Such objects do not correspond to real entities but generated when data are sent to the agents. In a sense, the fake objects send data without modifying any individual data. 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. The probability of occurrence of the repeated private object for the particular agent depicts the agent who involved in leakage of data.

4.1 Private objects

The private objects are represented as $f_1,f_2,...f_n$ for each data. The distributor creates and adds private objects for the data that he distributes to agents. If data d_1 transmitted from distributor or from any of the agents the private object f_1 is created for it. The irrelevant transfer of such data will create replication of the private object. Similarly, the distributor may want to limit the number of private objects received by each agent so as to not arouse suspicions and to not adversely impact the agents' activities. Thus, we say that the distributor can send up to bi fake objects to agent Ai. The data which occurred most number of times is identified and the agents who sent it also be distinguished. That agent might have shared the data within the authorized group. Thus the leaked data can be identified.

The process of transmitting same data within the agents also creates replication of private object. The distributor may be able to create private objects for the distributed data in order to improve the effectiveness in detecting guilty agents. However, private objects may impact the correctness of what agents do, so they may not always be allowable. The idea of perturbing data to detect leakage is not new. However, in most cases, individual objects are disturbed by adding random noise to sensitive data or adding a watermark to an image. In this case, we concerned the set of distributor objects by adding private elements.

4.2 Guilty agent

The agent who involved in irresponsible activity can be detected by the probability value. The probability of agents who are not authorized to receive data to the total number of agents for whom the particular data had been sent. The probability value depicts the agents who are guilty. Before we present the general formula for computing the probability $Pr\{G_i|S\}$ that an agent U_i is guilty, we provide a simple example. Assume that the distributor set D, the agent sets R_s, and the target set S are:

$D = \{d_1, d_2, d_3\}$; $R_1 = \{d_1, d_2\}$; $R_2 = \{d_1, d_3\}$; $S = \{d_1, d_2, d_3\}$: In this case, all three of the distributor's objects have been leaked and appear in S. Let us first consider how the target may have obtained object d_1, which was given to both agents. The target either guessed d_1 or one of A_1 or A_2 leaked it. We know that the probability of the former event is p, so assuming that probability that each of the two agents leaked d1 is the same, we have the following cases:

- The target guessed d_1 is leaked with probability p,
- Agent A_1 leaked d_1 to S with probability (1- p)/2, and
- Agent A_2 leaked d_1 to S with probability (1- p)/2.

Similarly, we find that agent A_1 leaked d_2 to S with probability 1- p since he is the only agent that has d_2. The agent who has the highest probability value had leaked data to many unauthorized persons. Such a guilty agent can be avoided by the distributor. This will help protecting the sensitive data being leaked further. The transaction with the particular guilty agent should be terminated. D contains 16 objects: all of them are given to agent A_1 and only eight are given to a second agent A_2. We calculate the probabilities $Pr\{G_1|S\}$ and $Pr\{G_2|s\}$ for p in the range [0, 1].

The data is protected by blocking it when there occurs a transmitting process with refused persons. Hence the declined receiver could not get the data sent from the guilty agent. Each agent has enough of the leaked data that its individual guilt approaches 1. However, as p increases in value, the probability that A_2 is guilty decreases significantly: all of A_2's eight objects were also given to A_1, so it gets harder to blame A_2 for the leaks. On the other hand, A_2's probability of guilt remains close to 1 as p increases, since A_1 has eight objects not seen by the other agent. At the extreme, as p approaches 1, it is very possible that the target guessed all 16 values, so the agent's probability of guilt goes to 0.

As the objects become easier to guess, it takes more and more evidence of leakage i.e., more leaked objects owned by A_2, before we can have high confidence that A_2 is guilty. We study an additional scenario that shows how the sharing of S objects by agents affects the probabilities that they are guilty.

5. CONCLUSION

The data needs to be processed and can be used in some significant applications. Hence the distributor has to give the sensitive data to trusted third party agents. In many of the cases there are chances for data to be leaked by the guilty agents. The various traditional techniques which are developed to detect leaked data could not be beneficial to all processes. This method of creating private objects provide improved results in finding the data which are leaked as well as the agent who leaked the data.

The agent guilt model is used to find the probability of guilt of the agent. The probability of data which are most time sent by the agent can also be calculated. Thus the distributor can efficiently identify the guilty actions using this method. This can be further improved by blocking the data when the guilty agent tries to send it to unauthorized agents.

6. REFERENCES

[1] P. Buneman, S. Khanna, and W.C.Tan, "Why and Where: A Characterization 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, "Provenance in Databases," Proc. ACM SIGMOD, pp. 1171-1173, 2007.

[3] S. Czerwinski, R. Fromm, and T. Hodes, "Digital Music Distribution and Audio Watermarking,", 2007.

[4] S. Jajodia, P. Samarati, M.L. Sapino, and V.S. Subrahmanian, "Flexible Support for Multiple Access Control Policies," ACM Trans. Database Systems, vol. 26, no. 2, pp. 214-260, 2001.

[5] P. Papadimitriou and H. Garcia-Molina, "Data Leakage Detection," IEEE Trans. on Knowledge And Data Engineering, Vol. 23, No. 1, Jan. 2011.

[6] 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.

[7] R. Sion, M. Atallah, and S. Prabhakar, "Rights Protection for Relational Data," Proc. ACM SIGMOD, pp. 98-109, 2003.

[8] L. Sweeney, "Achieving K-Anonymity Privacy Protection Using Generalization and Suppression," http://en.scientificcommons.org/43196131, 2002