International Journal of Recent Trends in Engineering & Research
ISSN(ONLINE) : 2455-1457
IMPACT FACTOR : 3.344

Privacy Preserving Data Leakage Detection

Vidhya Priya.S.P[1], Aravinthan.A[2],Nancy Prathiba.P[3],Poornima.R[4]
[1]*Computer Science and Engineering,Kathir College Of Engineering*
[2,3,4] *Information Technology, Kathir College Of Engineering*

Abstract : The number of data-leakages has been grown up rapidly in recent years these are found by using the statistic firms, research institution and government organizations. The main reason for the data-leakage is human mistakes. The carelessactivities of human is one of the important cause of the data loss. The solutions for detecting sensitive data leaks caused by human mistakes. With the help of using rabin algorithm the sensitive data is protected. An approach is used to screen the content in storage and transmission for exposing sensitive data information. This approach needs the detection operation that has been conducted in secrecy. The detection servers may be satisfied in the secrecy requirements. By using rabin algorithm, we have provided the solution for privacy preserving data leakage detection. The main advantage in this method is that it enables the data owner to safely delegate the detection operation to a semihonest provider without revealing the sensitive data to the provider and it also reduces the false alarms.
Keywords:Data-leak,privacy,networksecurity,collectionintersection.

I. INTRODUCTION

According to a report from Risk Based Security (RBS) [2], the number of leaked sensitivedata records has increased dramatically during the last few years, i.e., from 412 million in 2012 to 822 million in 2013. Deliberately planned attacks, inadvertent leaks (e.g., forwarding confidential emails to unclassified email accounts), and human mistakes (e.g., assigning the wrong privilege) lead to most of the data-leak incidents [3]. Detecting and preventing data leaks requires a set of complementary solutions, which may include data-leak detection [4], [5], data confinement [6]–[8], stealthy malware detection [9], [10], and policy enforcement [11].Network data-leak detection (DLD) typically performs deep packet inspection (DPI) and searches for any occurrences. DPI is a technique to analyze payloads of IP/TCP packets for inspecting application layer data, e.g., HTTP header/content. Alerts are triggered when theamount of sensitive data found in traffic passes a threshold. The detection system can be deployed on a router or integrated into existing network intrusion detection systems (NIDS).

Straightforward realizations of data-leak detection require the plaintext sensitive data. However, this requirement is undesirable, as it may threaten the confidentiality of the sensitive information. If a detection system is compromised, then it may expose the plaintext sensitive data (in memory). In addition, the data owner may need to outsource the data-leak detection to providers, but may be unwilling to reveal the plaintext sensitive data to them. Therefore, one needs new data-leak detection solutions that allow the providers to scan content for leaks without learning the sensitive information.

Data mining is the process of discovering interesting patterns (or knowledge) from large amounts of data. The data sources can include databases, data warehouses, the Web, other information repositories, or data that are streamed into the system dynamically. . Data mining, the extraction of hidden predictive information from large databases, is a powerful new technology with great potential to

help companies focus on the most important information in their data warehouses. Data mining tools predict future trends and behaviors, allowing business to make proactive, knowledge-driven decisions.

It is also known as Knowledge Discovery in Databases (KDD). Credit ratings/targeted marketing given a database of 100,000 names, which persons are the least likely to default on their credit cards. Identify likely responders to sales promotions Fraud detection in which types of transactions are likely to be fraudulent, given the demographics and transactional history of a particular customer. Customer relationship management is in which of my customers are likely to be the most loyal, and which are most likely to leave for a competitor. The Rabin's algorithm present details of our solution and provide extensive experimental evidences and theoretical analyses to demonstrate the feasibility and effectiveness of our approach.

Given a set of data points, each having a set of attributes, and a similarity measure among them, find clusters such that Data points in one cluster are more similar to one another. Clustering can be used to generate class labels for a group of data which did not exist at the beginning. The objects are clustered or grouped based on the principle of maximizing the intra-class similarity and minimizing the inter-class similarity. Data points in separate clusters are less similar to one another. The various Similar Measures are Euclidean Distance if attributes are continuous. Other Problem-specific Measures. Unsupervised learning when old data with class labels not available e.g. when introducing a new product. Group/cluster existing customers based on time series of payment history such that similar customers in same cluster. The Key requirements that are used in clustering are need a good measure of similarity between instances and identify micro-markets and develop policies for each.

II RELATED WORK

We abstract the privacy-preserving data-leak detection problem with a threat model, a security goal and a privacy goal. First we describe the two most important players in our abstract model: the organization (i.e., data owner) and the data-leak detection (DLD) provider.

2.1 Modules:

- Data leak
- Distributor
- Agent's Sensitive Data
- Login using Rabin's Algorithm
- Get file information

2.1.1 Data leak

It mainly 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. In order to increase the chances of detecting agents that leak data.

2.1.2 Distributor

This module contains the distributor details such as distributor name and password. Distributor is an overall controller of this project. Distributor will sent the sensitive data information to the agent.

2.1.3 Agent's Sensitive data

Theagents are like as client or user. The agent will provide the personal information to the distributor. The sensitive data information contains the agent's name, address, phone number, mail id, account number, voter id number and bank name.

2.1.4 Login using Rabin finger print algorithm

This module contains login details such as user name, random number and image. When user will give these details, then it compared to the existing finger print image. These are same to the existing data after user accepts the next page information otherwise will not get the information.

2.1.5 Get file information

Thismodule contains the file information. This file contains the file id, name and text. Admin had passed the file to the agent. Agents will get that file details.

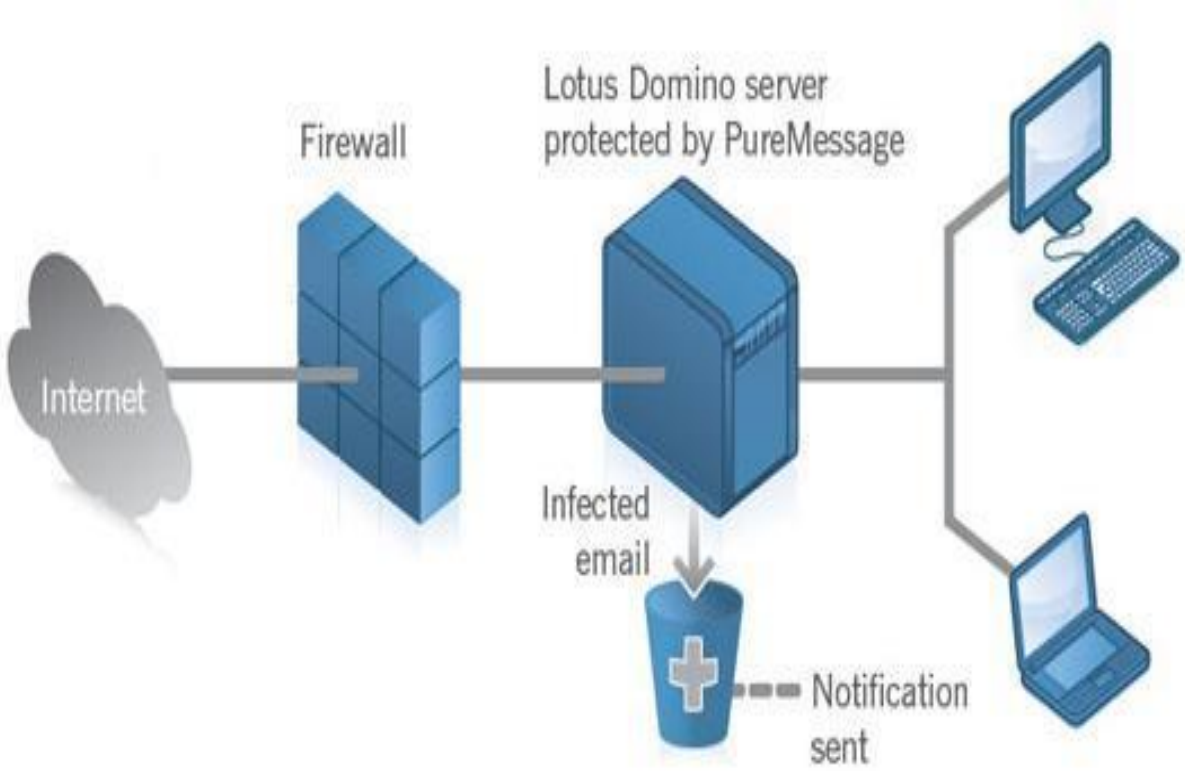

Fig 1: Architectural diagram to denote the Privacy preserving for DLD

III METHOD

*Rabin finger print algorithm:*A fingerprinting algorithm is a procedure that maps an arbitrarily large data item (such as a computer file) to a much shorter bit string, its fingerprint, that uniquely identifies the original data for all practical purposes just as human fingerprints uniquely identify people for practical purposes. This fingerprint may be used for data de-duplication purposes. Fingerprints are typically used to avoid the comparison and transmission of bulky data. For instance, a web browser or proxy server can efficiently check whether a remote file has been modified, by fetching only its fingerprint and comparing it with that of the previously fetched copy. This project used the Rabin finger print algorithm. This algorithm used for identify image are correct or else.

IV EXPERIMENTAL EVALUATION

Network data-leak detection (DLD) typically performs deep packet inspection (DPI) and searches for any occurrences of sensitive data patterns. DPI is a technique to analyze payloads of IP/TCP packets for inspecting application layer data, e.g., HTTP header/content. Alerts are triggered when the amount of sensitive data found in traffic passes a threshold. The detection system can be deployed on a router or integrated into existing network intrusion detection systems.

Straightforward realizations of data-leak detection require the plaintext sensitive data. However, this requirement is undesirable, as it may threaten the confidentiality of the sensitive information. If a detection system is compromised, then it may expose the plaintext sensitive data (in memory). In addition, the data owner may need to outsource the data-leak detection to providers, but may be unwilling to reveal the plaintext sensitive data to them. Therefore, one needs new data-leak detection solutions that allow the providers to scan content for leaks without learning the sensitive information. The disadvantages of fuzzy fingerprint algorithm are only less security is provided, these types of data can be easily hacked by using hacking tools and it provides more false alarms.

V PROPOSED SCHEME

By using Rabin Fingerprint Algorithm, we propose a data-leak detection solution which can be outsourced and be deployed in a semi honest detection environment. We design, implement, and evaluate our fuzzy fingerprinttechnique that enhances data privacy during data-leak detection operations. Our approach is based on a fast and practical one-way computation on the sensitive data. It enables the data owner to securely delegate the content-inspection task to DLD providers without exposing the sensitive data. Using our detection method, the DLD provider, who is modeled as an honest-but-curious (aka semi-honest) adversary, can only gain limited knowledge about the sensitive data from either the released digests, or the content being inspected.

Using our techniques, an Internet service provider can perform detection on its customers' traffic securely and provide data-leak detection as an add-on service for its customers. In another scenario, individuals can mark their own sensitive data and ask the administrator of their local network to detect data leaks for them. In our detection procedure, the data owner computes a special set of digests or fingerprints from the sensitive data and then discloses only a small amount of them to the DLD provider. The DLD provider computes fingerprints from network traffic and identifies potential leaks in them. To prevent the DLD provider from gathering exact knowledge about the sensitive data, the collection of potential leaks is composed of real leaks and noises. It is the data owner, who post-processes the potential leaks sent back by the DLD provider and determines whether there is any real data leak.

VI CONCLUSION

Rabin Algorithm is used to provide security for the sensitive data. This method has a special kind of digest which makes the sensitive data more secured. The disadvantages that are faced by the fuzzy fingerprint methods are overcome by the Rabin algorithm. From the privacy preserving data leakage detection we can reduce the false alarms also. The accuracy, efficiency, and privacy of our solution have been better than the fuzzy fingerprint methods. The login page has been provided with the image password by using Rabin's algorithm we can compare the images more effectively.

REFERENCES

1. Xiaokui Shu, Danfeng Yao, *Member, IEEE,* and Elisa Bertino, *Fellow, IEEE*"Privacy preserving data of sensitive exposure," in *Proc. 2015*, pp. 222–240.

2. Risk Based Security. (Feb. 2014). *Data Breach Quick-View: An Executive's Guide to 2013 Data Breach Trends*.[Online]. Available: https://www.riskbasedsecurity.com/reports/2013- DataBreachQuickView.pdf, accessed Oct. 2014.
3. Ponemon Institute. (May 2013). *2013 Cost of Data Breach Study: Global Analysis*. [Online]. Available: https://www4.symantec.com/mktginfo/whitepaper/053013_GL_NA_WP_Ponemon-2013-Cost-of-a-Data-Breach-Report_daiNA_cta72382.pdf, accessed Oct. 2014.
4. Identity Finder. *Discover Sensitive Data Prevent Breaches DLP Data Loss Prevention*. [Online]. Available: http://www.identityfinder.com/,accessed Oct. 2014.
5. K. Borders and A. Prakash, "Quantifying information leaks in outbound web traffic," in *Proc. 30th IEEE Symp. Secur. Privacy*, May 2009, pp. 129–140.
6. H. Yin, D. Song, M. Egele, C. Kruegel, and E. Kirda, "Panorama: Capturing system-wide information flow for malware detection and analysis," in *Proc. 14th ACM Conf. Comput. Commun. Secur.*, 2007, pp. 116–127.
7. K. Borders, E. V. Weele, B. Lau, and A. Prakash, "Protecting confidential data on personal computers with storage capsules," in *Proc. 18thUSENIX Secur. Symp.*, 2009, pp. 367–382.
8. A. Nadkarni and W. Enck, "Preventing accidental data disclosure in modern operating systems," in *Proc. 20th ACM Conf. Comput. Commun.Secur.*, 2013, pp. 1029–1042.
9. A. Kapravelos, Y. Shoshitaishvili, M. Cova, C. Kruegel, and G. Vigna, "Revolver: An automated approach to the detection of evasiveweb-based malware," in *Proc. 22nd USENIX Secur. Symp.*, 2013, pp. 637–652.
10. X. Jiang, X. Wang, and D. Xu, "Stealthy malware detection and monitoring through VMM-based 'out-of-the-box' semantic view reconstruction," *ACM Trans. Inf. Syst. Secur.*, vol. 13, no. 2, 2010, p. 12.
11. G. Karjoth and M. Schunter, "A privacy policy model for enterprises,"in *Proc. 15th IEEE Comput. Secur. Found. Workshop*, Jun. 2002, pp. 271–281.
12. J. Jung, A. Sheth, B. Greenstein, D. Wetherall, G. Maganis, and T. Kohno, "Privacy oracle: A system for finding application leaks with black box differential testing," in *Proc. 15th ACM Conf. Comput.Commun. Secur.*, 2008, pp. 279–288.
13. Y. Jang, S. P. Chung, B. D. Payne, and W. Lee, "Gyrus: A framework for user-intent monitoring of text-based networked applications," in *Proc.23rd USENIX Secur. Symp.*, 2014, pp. 79–93.
14. K. Xu, D. Yao, Q. Ma, and A. Crowell, "Detecting infection onset with behavior-based policies," in *Proc. 5th Int. Conf. Netw. Syst. Secur.*, Sep. 2011, pp. 57–64.
15. M. O. Rabin, "Fingerprinting by random polynomials," Dept. Math., Hebrew Univ. Jerusalem, Jerusalem, Israel, Tech. Rep. TR-15-81, 1981.
16. A. Z. Broder, M. Charikar, A. M. Frieze, and M. Mitzenmacher, "Min-wise independent permutations," *J. Comput. Syst. Sci.*, vol. 60, no. 3, pp. 630–659, 2000.
17. A. Z. Broder, "Some applications of Rabin's fingerprinting method," in *Sequences II*. New York, NY, USA: Springer-Verlag, 1993, pp. 143–152.
18. A. Z. Broder, "Identifying and filtering near-duplicate documents," in *Proc. 11th Annu. Symp. Combinat. Pattern Matching*, 2000, pp. 1–10.
19. A. Broder and M. Mitzenmacher, "Network applications of bloom filters: A survey," *Internet Math.*, vol. 1, no. 4, pp. 485–509, 2004.
20. G. Aggarwal *et al.*, "Anonymizing tables," in *Proc. 10th Int. Conf. Database Theory*, 2005, pp. 246–258.

