POLICY TO DETECT AND PRESERVE SENSITIVE DATA FROM DATA LEAK

Sushma Gaikwad[1], Asst.Prof.Shilpa Chougule[2]
[1,2]University of Mumbai
Email: Sushmavishwanath09@gmail.com[1], schougule@mes.ac.in[2]

Abstract

Nowadays data leakage is the big problem in front of the different kind of institution, organization, and security firms. Everyone knows what is important of sensitive data .The unveiling of sensitive data in storage and transmission creates a serious threat to organizational and personal security. The system propose a privacy-preserving data-leak detection model for preventing accidental data leak in network traffic as well as focus on designing a host-assisted mechanism for the complete data- revelation for organizations. Here Human misstep is the primary reason for data leak. The advantage of this system is without disclosing the sensitive data, information owner safely assign the detection operation to a DLD provider. In this paper the system used data leak detection (DLD) on the basis of summaries of information and depict how Internet service providers can offer their clients DLD as an extra service. It can be outsourced and be deployed in a semi-honest detection environment. The evaluation results show that this approach can support accurate detection with very less number of false alarms under various data-leak scenarios.

Index Terms— network security, Information security, Data leak, privacy.

I INTRODUCTION

Preserving confidential information is a prime concern for individuals and organizations alike, who stand to suffer very huge losses if sensitive data falls into the fallacious hands. On the basis of a report from Risk Based Security (RBS) the count of divulged sensitive Data records has increased drastically between the last Some years, i.e., from 412 million in 2012 to 822 million in 2013. Willfully arranged invasion, accidental leaks (e.g., forwarding secrete emails to unclassified email Accounts), and human Fault (e.g., allow the wrong Privilege) lead to disclosed large amount of information. Typical ways to preventing data loss are under two categories – host-based solutions and network-based solutions. Host-based approaches may include i) encode data when not used. ii) Detecting secret malware with antivirus scanning or monitoring the host and iii) Invoke behavior to restrict the transfer of sensitive data. But in this paper not using any of this method for host based approach as well as not using any use of virtualization or special hardware to ensure the system integrity of the detector. For network-based information exposure, researchers have developed data-loss prevention (DLP) systems. DLP systems work on outbound network traffic for known sensitive data, such a social security numbers, credit card number. For network data leak detection DPI is used. DPI is a process to scrutinize payloads of IP/TCP packets for scanning application layer data, e.g., HTTP header/content. When the quantity of sensitive information found in traffic passes a threshold, alerts are activated. For host based shortest path algorithm are used. DLD are honest-but-curious (aka semi-honest) therefore for data owner used summaries of information for detection method. These fingerprints have important properties, which intercept the provider from gaining knowledge of the sensitive data, and permit accurate comparison and detection. The DLD provider performs deep-packet inspection to identify whether these

summaries patterns exist in the outbound traffic of the organization or not, according to a quantitative metric. In this way, anyone needs new information spill discovery management that allows the suppliers to sweep content for holes without taking in the sensitive data.

II BACKGROUND OVERVIEW

In [2] this work, they present a new Map Reduce-based system to detect the occurrences of plaintext sensitive data in transmission and storage. The detection is distributed and parallel, capable of screening massive amount of content for exposed information. They address an important data privacy requirement. In this privacy-preserving data-leak detection, the content in data storage or network transmission scan by MapReduce nodes for leaks without learning what the sensitive data is. Specifically, the data privacy protection is realized with fast one-way transformation. Here this transformation needs the pre- and post-processing by the information owner for hiding and precisely identifying the matched items, respectively. The advantages of this paper are, this transformation supports the secure outsourcing of the information disclose detection to untrusted Map Reduce and cloud providers. But here a significant portion of the incidents are caused by unintentional mistakes of employees or data owners.

In [4] this paper Gyrus only analysis network traffic under protocols used by the secured applications, and it does not obstruct with traffic from other applications, such as background services, RSS feed readers, and BitTorrent consumer. Gyrus can support scheduled or asynchronous traffic like e-mail queued for sending in the future. From this evaluation, Gyrus exhibits grate performance and usability, while blocking all tested attacks. There are many potentially beneficial areas for future work. To simplify the process of supporting a new application by automating the analysis and generation of the UI and traffic signatures. The advantages of this paper Gyrus is very effective and proposes no observable delay to a users' communication with the preserved applications. But here Gyrus solves problem by relying on the semantics, but not the timing of user generated events.

In [5] this paper introduced and demonstrated Revolver, a novel approach and tool for detecting malicious JavaScript code similarities on a massive scale. Revolver's approach is based on identifying scripts that are similar and catching into account an Oracle's classification of each one script. By doing this, Revolver can pinpoint scripts those have high sameness but are classified differently (detecting likely evasion attempts) and improve the accuracy of the Oracle. This performed a massive scale evaluation of Revolver by running it in parallel with the popular Wepawet drive-by detection tool. This identified several cases of evasions that are used in the wild to evade this tool (and, likely, other tools based on similar techniques) and fixed them, improving this way the accuracy of the honey client. The advantage of this paper is this lack of application separation did not expose it as a concern. But here it may not be trusted with that data.

In [7] network-based data-leak detection (DLD) technique, the important feature of which is that the detection does not require the information proprietor to show the content of the sensitive data. Instead, only a small amount of specialize digests are needed. In comparison to host-based approaches, network-based data-leak detection focuses on examining the (unencrypted) content of outbound network packets information. For example, a naive solution requires to inspect every packet for the occurrence of any of the sensitive data defined in the database. Such solutions generate alerts if the sensitive data is found in the outgoing traffic However, this naïve solution requires to store sensitive data in plaintext at the network interface, which is highly undesirable. This is not efficient enough for practical data leak inspection in this setting.

In [10] this paper introduced a new approach for quantifying information leaks in web traffic. Instead of inspecting a message's data, the goal was to quantify its information content. The algorithms in this paper achieve precise results by discounting fields that are repeated or constrained by the protocol. This work focuses on web traffic, but similar principles can apply to other protocols. This analysis engine processes static fields in HTTP, HTML, and Java script to create a distribution of expected request content. It also executes dynamic scripts in an emulated browser environment to obtain complex request values.

The Hypertext Transfer Protocol (HTTP) is used for web browsing. The leak measurement method presented here focus on this protocol. Hear take advantage of HTTP and its communication with Hypertext Markup Language (HTML) documents and JavaScript code to quantify data leak capacity. The advantage of this paper is that it possible to identify smaller leaks. But Here Traffic measurement does not completely stop

This [9] paper introduces Storage Capsules, a new mechanism for protecting sensitive information on a local computer. The aim of Storage Capsules is to convey the same level of security as a mandatory access control system for standard applications running on a commodity operating system. The user's system can also downgrade from high-secrecy to low-secrecy by reverting to a prior state using virtual machine snapshots. Storage Capsules are similar to existing encrypted file containers, but protect sensitive data from malicious software during decryption and editing. The Capsule system provides this type of protection by isolating the user's primary operating system in a virtual machine. While it is accessing personal files, and retrogress its state to a snapshot taken prior to editing when it is finished, then Capsule system turns off the primary OS's device output. One main advantage of Storage Capsules is that they work with current applications running on commodity operating systems. The advantages of this paper is this system achieves this goal by taking a checkpoint of the current system state and before allowing access a Storage Capsule disabling device output. But here It does not rely on high integrity.

III RESEARCH PROBLEM

Information Security is the heart of organization, institution and security firms. Information security means protecting information and information systems from unauthorized handling, alteration, discovery, interruption, access, inspection, recording or destruction. Maintaining privacy in my personal communication is something everyone desires. Nowadays different kind of institution, organization, and security firms show that the number of data-leak instances has grown rapidly in recent years**.** Here reproduce four real-world scenarios where data leaks are caused by human users or software applications. i) Web leak - sensitive data post on wiki (Media Wiki) and blog (Word Press) pages by user. ii) Backdoor leak - A program (Glacier) on the user's machine (Windows 7) disclose sensitive data.iii) Browser leak- FFsniFF is a malicious Firefox extension which records the information in sensitive web forms, and emails the data to the attacker. iv) Key logging- leak a key logger EZRecKb exports intercepted keystroke values on a user's host. It connects to a SMTP server on the Internet side and sends its log of keystrokes periodically.

IV. PROPOSED SYSTEM METHODOLOGY

A. SYSTEM OVERVIEW.

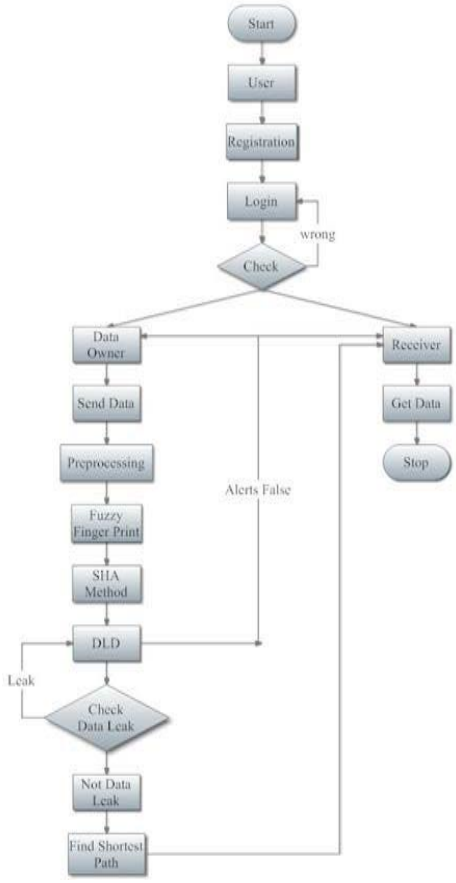

Fig.1. Data Flow Diagram

So, the above flow diagram can be explained as follows:

1. First the data owner Pre-process all data.
2. Create Fuzzy fingerprint of sensitive data using SHA-1 algorithm.
3. Send fingerprint to DLD.
4. On the basis of fingerprint DLD provider inspect network traffic.
5. Detect the Data leaks.
6. After that using shortest path algorithm for the host based approach.

7. Report all Data leak to the Data owner.
8. Data owner post-process all the data to find out it is a true leak or not.

B. ENHANCEMENT

The Existing process have only find out Data Leakage Process But here upgrade performance in paper so implement new concept Shortest path problem because it is help to host process and Deep packet inspection. So for shortest path Dijkstra's algorithms is used. SHA-1 is used for the creating summaries of information. For the security and encryption purpose AES algorithm are using.

Shortest Path Advantages**:** Hosting process time send some messages is traveled in multiple nodes that time take more time and some time attacking process chance available then cost is increase between the hosting approaches.

V. RESULTS

Figure 2 shows that ER stands for the Existing Result and PR Proposed Result. This Bar chart shows that the evaluation results show that this approach can support accurate detection with very small number of false alarms under various data-leak scenarios.

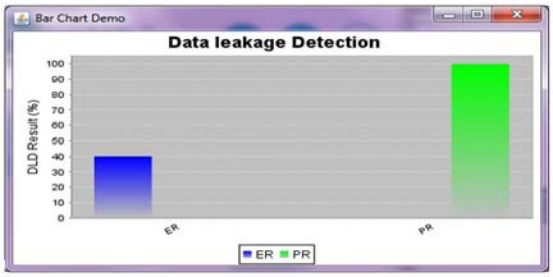

Fig.2 Data leakage Detection

Table1.1 Comparisons of traditional Approach and New Approach.

Analytics	ACCURACY	EFFICIENCY	FALSE ALARM	PRIVACY
TRADITIONAL APPROACH	LESS	LESS	MINIMUM NUMBER	LESS
NEW APPROACH	HIGH	HIGH	MINIMUM NUMBER	HIGH

Table1.1 shows that Current approach have better accuracy, efficiency and privacy than traditional approach as well as having small number of false alarm.

VI CONCLUSION

The system implemented, and evaluated a new privacy-preserving data-leak detection system that enables the information proprietor to safely deploy locally, or to delegate the traffic-inspection task to DLD providers without exposing the sensitive data. In this paper proposed privacy preserving data detection model using special digest and the exposure to sensitive data leakage is kept minimum. Here described its application in the cloud computing environments, where the cloud provider naturally serves as the DLD provider. It can support the accurate detection with low false positives.

VII ACKNOWLEDGMENTS

During the entire period of this survey, my survey paper would not have been materialized without the help of many people, who made my work so easier. It gives me proud privilege to complete this survey paper working under valuable guidance of Prof. Shilpa Chougule. She has been very supportive and patient throughout the process. She has been great help to me. I am also thankful to all staff members for providing all facilities and every help for smooth progress of paper work. I thank all others, and especially, friends and our family members who in one way or another helped me in the successful completion of this work.

REFERENCES

[1] Xiaokui Shu, Danfeng Yao, ―Privacy-Preserving Detection of Sensitive Data Exposure‖, IEEE Transcation on Information Forensic And Security, VOL. 10, NO. 5, MAY 2015.

[2] Fang Liu, Xiaokui Shu, Danfeng (Daphne) Yao and Ali R. Butt Privacy preserving scanning of big content for sensitive data exposure with MapReduce 2015.

[3] Ponemon Institute. (May 2013). 2013 Cost of Data Breach Study GlobalAnalysis.[Online].Available:https://www4.symantec.com/mktginfowhitepaper/053013_GL_NA_WP_Ponemon-2013-Cost-of-a-Data-Breach-

Report_daiNA_cta72382.pdf, accessed Oct. 2014.

[4] 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

[5] Kapravelos, Y. Shoshitaishvili, M. Cova, C. Kruegel, and G. Vigna, "Revolver: An automated approach to the detection of evasive web-based malware," in Proc. 22nd USENIX Secur.Symp., 2013, pp. 637–652.

[6] 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

[7] X. Shu and D. Yao, ―Data leak detection as a service,‖ in Proc. 8th Int.Conf. Secur. Privacy Common. Newt. 2012, pp. 222–240.

[8] JIANG, X., WANG, X., AND XU, D. Stealthy malware detection and monitoring through vmm-based "out-of-t box" semantic view reconstruction. ACM Trans. Inf. Syst. Secur. 13, 2 (2010)

[9]] K. Borders, E. V. Weele, B. Lau, and A. Prakash, "Protecting confidential data on personal computers with storage capsules," in Proc. 18th USENIX Secur. Symp., 2009, pp. 367–382

[10] K. Borders and A. Parkas, ―Quantifying information leaks in outbound web traffic,‖ in Proc. 30th IEEE Sump. Secure. Privacy, May 2009, pp. 129–140.

[11] H. Yin, D. Song, M. Agile, C. Kruegel, and E. Kirda, Panorama: Capturing system-wide information flow for malware detection and analysis,‖ in Proc. 14th ACM Conf. Compute. Commun. Secur., 2007,pp. 116–127.