International Journal on Advanced Science Engineering Information Technology

Vol.6 (2016) No. 5
ISSN: 2088-5334

Detection and Prevention of Sensitive Data from Data Leak Using Shingling and Rabin Filter

Sushma Gaikwad[#], Shilpa Chougule[#], Shrikant Charhate[#]

[#] *Pillai HOC College of Engineering, and Technology ,Rasayani, Dist: Raigad, 410207 Maharashtra, India.*
E-mail: [1]sushmavishwanath09@gmail.com, [2]schougule@mes.ac.in, [3]scharhate@mes.ac.in

***Abstract*— Data leak is a major problem in all the organization of any land. A deliberate risk to institution and private security is the disclosure of secure data in transmission and storage. To check content for exposed sensitive data is the main aim for exposed sensitive data. There are large numbers of data-leak cases but human flaws are one of the main reasons of data leak. This paper proposed a data-leak detection model for preventing accidental and intentional data leak in network. If someone succeed to steal some kind of data and send that data to outsider then data owner has obtain to use two methods to find out guilty employee or leaker. This work suggests use of shingling and rabin filter system performs Data Leak Detection (DLD) and Prevention task. The results show that this approach can be effectively implemented in various organizations; however rigorous testing on various data division of such methods will be required to implement the same in sector of importance like defence and other even in large establishment.**

***Keywords*— Information security; Data leak; network security; privacy.**

I. INTRODUCTION

Now a days since all the organizations in various sectors are connected through system, information flows from various services to the end user hence, information security has become sensible to organizations and institutions. Information Security means protecting information and information systems from unauthorized handling, alteration, discovery, interruption, access, inspection, recording or destruction. Maintaining privacy in personal communication is something that everyone desires. The main aim of data leak detection is scanning content for exposed sensitive data. It can be outsourced. These deployed in a semi-honest detection environment.

Typical ways to preventing data loss are under two categories Network-based solutions [1] and host-based solutions [12]. Host-based approaches may include i) encode data [10] when not used. ii) Detecting secret malware [13] with antivirus scanning or monitoring the host and iii) Invoke behavior to restrict the transfer of sensitive data [8]. The disclosure of Network-based data is a serious threat to confidentiality.

Xiaokui Shu et.al. proposed network based data leak detection (DLD) technique[1]. The data leak detection provider who performs the process of data leak detection cannot be trusted; as he acts as third party in the system. Also he can send the data to an unknown user as because of the fact he might not be using any secure channel to send the data. Y. Jang et.al. (2014) proposed a traditional security system which is focused on attack detection But Gyrus framework is not applicable to all the application [4].

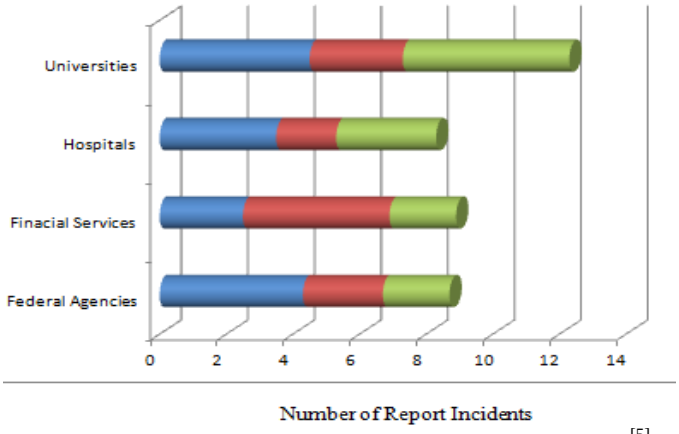

Fig.1Top Industry sectors experiencing multiple breaches [5]

Figure 1 shows a wider review of all industries reveals the above industry sectors as having organizations more likely to experience multiple incidents. These industries are hospitals, federal agencies, financial services and universities. The unintentional disclosure or mishandling of confidential data by internal employees is a significant factor. The proposed system detects files that contain confidential information and prevent them from leaving via the network.

It cannot pass sensitive data transfers to Universal Serial Bus (USB) drives and other removable media. This has application in cloud computing environment where cloud provider can offer their clients DLD as an extra service with minimal changes to the infrastructure and makes the cloud service more attractive.

There are three main types of sensitive information:

1) Personal information: Personal information is data that can be related to particular person and that, if disclosed, could result in harm to that person. Financial information, medicinal background, bio metric data such information includes personally identifiable, and unique identifiers such as passport or social security numbers.
2) Organizational information: Sensitive institutions information includes anything such as business operation relation that poses a risk to the company in question if discovered by a competitor or the other public. Such information includes trade secrets, acquisition plans, financial data and supplier and customer information. Unauthorized access are becoming integral to corporate security, as business contains increasing amount of data, different methods of protecting corporate information.
3) Classified information: Classified information pertains to a government body and is limited according to level of sensitivity (for example, restricted, confidential, secret and top secret).

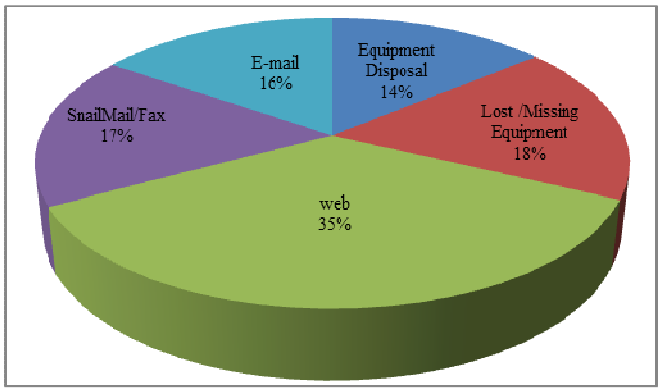

Fig. 2: Inside Accidental incidents [6]

A review of breach type in insider incidents also yields interesting results. In figure 2 shows accidental data loss due to activities such as errant website postings, careless equipment disposal or poor equipment management accounted for 63.6% of insider incidents.

Nadkarni et.al. presents Modern operating systems [3] have changed both the way users use software and the underlying security architecture. These two changes make accidental data disclosures easier. This approach was defeated by static detection of the malicious code using signatures. Kapraveloset.al. introduced and demonstrated Revolver [6] a novel approach and tool for detecting similarities between malicious JavaScript code on a large scale. It may not be trusted with that data

Stefan D et.al. proposed a technique secure multi-party computation (SMC)[8],[12] similar to string matching method. SMC is a cryptographic mechanism, which supports a wide range of fundamental arithmetic, set, and string operations as well as complex functions such as knapsack computation, automated troubleshooting, network event statistics, genomic computation and distributed data mining.K. Borders et.al. paper introduced a new approach for quantifying information leaks [9] in web traffic. The goal was to quantify its information content instead of inspecting a message's data. The advantage of this paper is that it possible to identify smaller leaks. But Here Traffic measurement does not completely stop information leaks from slipping by undetected. K .Borders et.al introduces Storage Capsules [13] a new mechanism for protecting sensitive information on a local computer but here it does not rely on high integrity. H. Yin et.al.. proposed a technique, Panorama [10] to detect and examine malware by capturing this fundamental trait. But here detecting malware and examining unknown code samples are insufficient and have significant shortcomings.

II. Material and Method

A. Problem Formulation

The Federal agencies, Universities, Financial services and hospitals contains large amount of various sensitive data. Data leak is a major problem in all the institute on of any land. Many modern networks contain highly sensitive data. e.g. Health care companies must store the all sensitive data or information of their client securely that should not be leaked outside their network. On the basis of a report from Risk Based Security (RBS) [7] between last some years the count of exposed sensitive data records has expanded tremendously, e.g. .from 412 million in 2012 to 822 million in 2013. There are various data-leak cases but accidental data leak are one of the main causes of data leak. This data leak may be occur because of human mistake as sometime forgetting to encrypt the data, sending an intentionally or unintentionally internal email and attachments to outsiders and sharing the data or application flaws.

B. System Design

The following figure 1 shows the architectural view of the proposed system. In this section, a model has been proposed which detect and prevent data leak. Here for detection purpose shingling and rabin filter method is used. For prevention purpose secure channel is used to send the data on network. If someone succeed to steal data and send that data to outsider then data owner used two methods fake object and probability to find out organizational leaker.□The proposed system is elaborated in Fig.1, which consists of data leak detection and prevention. Detection system is handled by data owner itself.□□Data owner distribute or upload the data to their organizational employee. □mployees must have their unique id and password to handle system.□To download the data which is sent by the data owner employee must enter unique secrete code for that file. In organization there are two types of employee one is admin and other is normal employee. Data owner share their files with admin and employees. Confidential files he can share with only admin and other local files share with other employees.□□□Admin can share only local files with

employees. He cannot share confidential files. Employee cannot share local files to other employee also. To detection and prevention purpose uses rabin and shingling method.

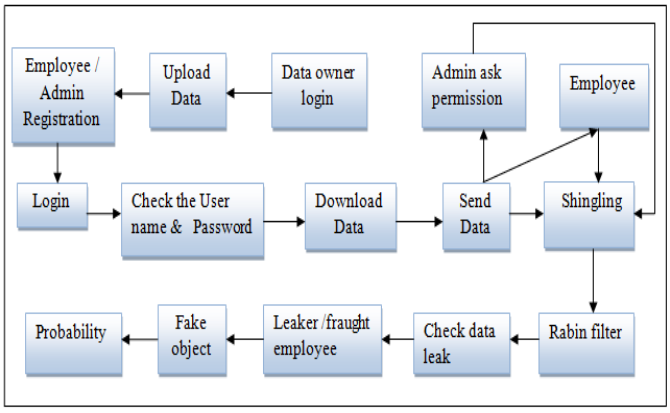

Fig. 3: System Architecture diagram

C. Enhancement

The Existing process has only found out Data Leakage Process. Here the proposed method implements prevention mechanism to upgrade the system performance. It also tries to find out data leaker using following methods.

1) Shingling method: w-shingling is a collection of unique "shingles" which contiguous sub sequences of tokens in a document. It is used to check the similarity of two documents. The w denotes the number of tokens in each shingle in the set. A shingle method [1] is also known as q-gram method. A shingle (q-gram) is a fixed-size sequence of contiguous bytes. For example, the 3-gram shingle set of string pqrstuvw consists of six elements {pqr, qrs, rst, stu, tuv,uvw}.Using shingling preserving the Local feature. Therefore, this approach can tolerate sensitive data modification to some extent.

2) Rabin Filter: Rabin filter [1] is utilized to satisfy requirement after shingling. In filtering, each shingle is treated as a polynomial q(x). Rabin algorithm works on fast polynomial modulus operation. Each coefficient of q(x), i.e., ci (0 < i < k), is one bit in the shingle. q(x) is mod by a selected irreducible polynomial p(x). The process shown in (1) maps a k-bit shingle into a p f -bit filter f where the degree of p(x) is p f + 1

$$f = c_1x^{k-1} + c_2x^{k-2} + \ldots + c_{k-1}x + c_k \bmod p(x) \ldots\ldots(1)$$

3) Dynamic mail box: If any employee wants to share their own file with anyone then he can use this dynamic mail box. This mail box have same structure like compose of Gmail, yahoo etc. The employee can send their personal file to anyone using the mail id given but they are not allowed to send the organizational file. This restriction on sending the file is implemented using two algorithms shingling and rabin filter which are placed on send button. These two methods prevent employee to share organizational file. If the employee try to send the file then he will get the error message that "You are not authorize to share this file", and notification of this employee will be send to the owner.

D. Testing Mechanism to identify data leaker

1) Fake Object: The data owner may be able to add fake objects [9] to the distributed data in order to improve his effectiveness in detecting guilty employee. Data owner can add any object which acts as real data. For example, objects may contain e mail addresses, and each fake e-mail address may require the creation of an actual inbox (otherwise, the employee may discover that the object is fake). The inboxes can actually be monitored by the data owner: if e-mail is received from someone other than the agent who was given the address, it is evident that the address was leaked.

2) Probability: If someone steals the data by using pen drive or stealing hard disk after that owner found same like that data on the web or somebody's laptop then owner uses probability method to find out leaker. In probability method [9] owner can match the content of own file and another file which he found in an unauthorized place (e.g., on the web or somebody's laptop). If content matches with some employees file based on the probability owner find out leaker.

III. Results and Discussion

The risk-based security [2] is one type of model which identifies the real or true risks to an organization's. It is providing not only security but the right security. On the basis of a report from Risk Based Security (RBS) [7] data leak incidence are increased day by day e.g. .from 266 million in 2012 to 1.00 billion in 2013.

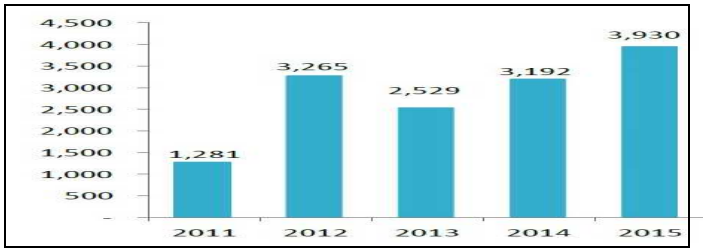

Fig. 4: Number of Incidents [2]

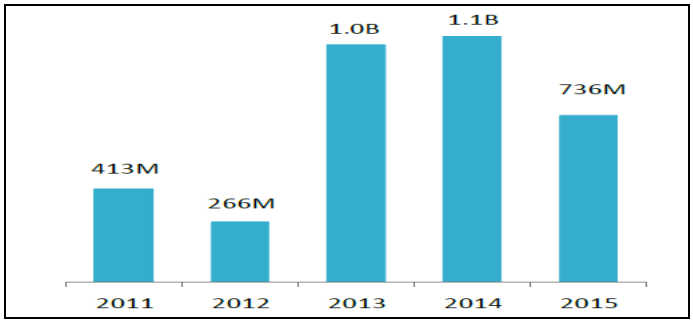

Fig. 5: Number of Records Exposed [2]

Here we studied Shingling and Rabin filter to detect and prevent data leak. The system has implemented, and evaluated a new data-leak detection system that enables the data owner to detect and prevent data leak as well as to find out guilty employee or leaker using fake object and probability method.

From the above study we obtained Process result. Figure 4 shows the process result based on file count of system. The proposed system shows positive output. There is no change for negative output. It shows that performance of proposed system is superior.

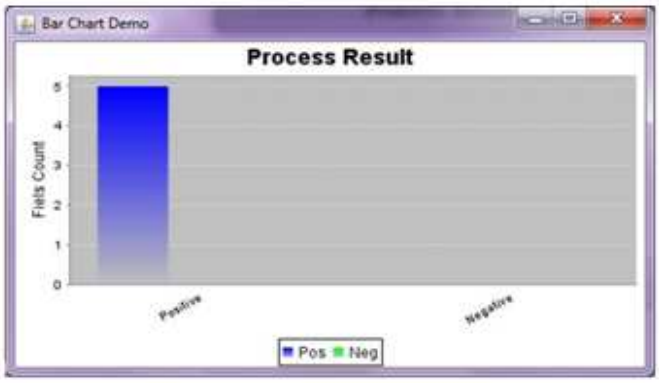

Fig. 4 Process Result based on file count

Here owner can share confidential file with only admin. But admin cannot share that file with any other employee. He can share only public file with other employee. The above result shows the superiority of the process in the proposed system over traditional approach.

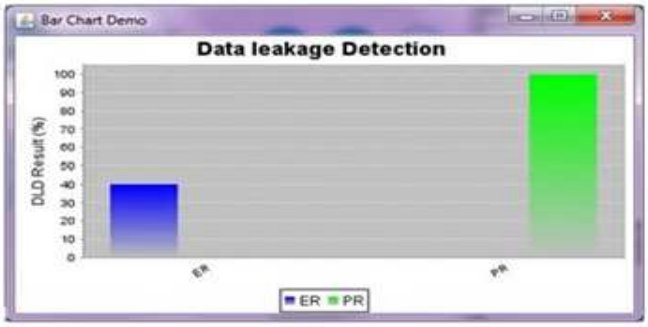

Fig. 5 Data Leakage Detection Rate

Figure 5 shows that evaluation result where ER stands for the Existing Result and PR Proposed Result. This evaluation results show that proposed approach can support accurate detection. The Y-axis shows detection rate of existing system as well as proposed system.

Figure 6 shows the line graph of existing data leak detection and current data leak detection. A line chart or line graph is a one type of chart. This displays information as a series of data points called 'markers' connected by straight line segments. From this graph we conclude that proposed system detect highly rate of data leak.

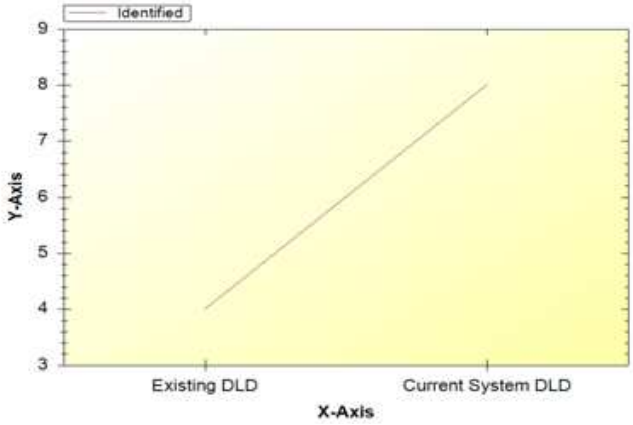

Fig. 6 Comparison between existing and current DLD system

TABLE I
COMPARISON OF TRADITIONAL APPROACH AND NEW APPOROACH

ANALYTICS	ACCURACY	EFFICIENCY	FALSE ALARM	PRIVACY
TRADITIONAL APPROACH	Less	Less	Maximum Number	Less
NEW APPROACH	High	High	Minimum Number	High

Table I shows that Current approach has better accuracy, efficiency and privacy than traditional approach as well as having small number of false alarm

IV. CONCLUSIONS

Preventing sensitive data from being compromised is important and practical research problem in all the fields. Maintaining privacy in personal or business communication is something everyone desires. The algorithms namely Shingling and Rabin filter discussed in this paper showed better result. The system implemented, and evaluated a new data-leak detection system that enables the data owner to detect and prevent data leak. The evaluation results show that this approach can support accurate detection with very small number of false alarms and can be effectively implemented in various organizations, however rigorous testing on various data division of such methods will be required to implement the same in sector of importance like defence and other even in large establishment.

REFERENCES

[1] X. Shu and D. Yao "Privacy-Preserving Detection of Sensitive Data Exposure" in IEEE Transaction on Information Forensics and security, Vol.10, No.5, May 2015.

[2] Risk Based Security."Data Breach QuickView: An Executive'sGuideto2013DataBreachTrends".[Online].Available:https://www.riskbasedsecurity.com/reports/2013DataBreachQuickView.pdf, accessed Oct. 2015.

[3] F. Liu, X. Shu, D. Yao, and A. R. Butt, "Privacy-preserving scanning of big content for sensitive data exposure with MapReduce," in Proc.ACM CODASPY, 2015.

[4] Y. Jang, S. P. Chung, B. D. Payne, and W. Lee, "Gyrus: A framework for user-intent monitoring of text-based networked applications". 23rd USENIX Secur. Symp., pp.79–93 Febuary 2014.

[5] Risk Based Security. "Data Breach QuickView: An Executive'sGuideto2013DataBreachTrends".[Online].Available:https://www.riskbasedsecurity.com/reports/2013DataBreachQuickView.pdf, accessed Oct. 2014.

[6] Kapravelos, Y. Shoshitaishvili, M. Cova, C. Kruegel, and G. Vigna Year: (2013) "Revolver: An automated approach to the detection of evasive web-based malware" in Proc. 22nd USENIX Secur. Symp. 2013, pp. 637–652. August 2014.

[7] Nadkarni and W. Enck, "Preventing accidental data disclosure in modern operating systems," in Proc. 20th ACM Conf. Comput. Commun. Secur., 2013, pp. 1029–1042.

[8] X. Shu and D. Yao, "Data leak detection as a service," in Proc. 8th Int. Conf. Secur. Privacy Commun. Netw, pp. 222–240. July 2012.

[9] Panagiotis Papadimitriou, Student Member, IEEE, and Hector Garcia-Molina, Member,(2011) Data Leakage Detection IEEE Transaction on Knowledge and data Engineering Vol. 23, NO. 1, January 2011.

[10] Stefan, D., Wu, C., YAO, D. And U, G."Cryptographic provenance verification for the integrity of keystrokes and outbound network traffic" In Proceedings of the 8th International Conference on Applied Cryptography and Network Security (ACNS) pp. 110–124 June 2010.

[11] J. Li, Q. Wang, C. Wang, N. Cao, K. Ren, and W. Lou, "Fuzzy keyword search over encrypted data in cloud computing," in Proc. 29th IEEE Conf. Comput. Commun., Mar. 2010, pp. 1–5.

[12] K. Borders and A. Prakash, "Quantifying information leaks in outbound web traffic," in Proc. 30th IEEE Symp. Secur. Privacy, May 2009,pp. 129–140.

[13] 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.

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