IJCSE International Journal of Computer Sciences and Engineering Open Access
Review Paper Vol.-7, Special Issue, 4, Feb 2019 E-ISSN: 2347-2693

An Efficient Method to Discover Transformed Data Leak

S. Sindhuja[1*], M. Valliammai[2]

[1]M.Sc Computer Science, Idhaya College for Women, Kumbakonam, Tamilnadu, India
[2]M.Sc Computer Science, Idhaya College for Women, Kumbakonam, Tamilnadu, India

Corresponding Author: sindhu890@gmail.com

Available online at: www.ijcseonline.org

***Abstract*—**The computer system poses a serious threat to the organisational security due to the leak of sensitive data. According to the report of risk based security (RBS), the leaked sensitive data records has increased dramatically during last few years, (i.e.) from 412 million in 2012 to 822 million in 2013. These are caused only by the lack of proper encryption on files and documents and by human errors these causes data loss. Organisation has the responsibility of screening the content which is stored in the system as sensitive data. In this paper, we utilize two techniques, which are levenshtein-distance technique and luecene search framework. These two helps to detect the leakage of data and this technique is used for screening the data which are outsourced and it also keep an track of, who is transferring the data.

***Keywords*—**Sensitive data, Data leak, Data detection, levenshtein-distance, luecene search.

I. INTRODUCTION

In Networks data leak detection, content inspection, sampling, alignment, dynamic programming parallelism is performed. A report show that the number of leaked sensitive data records has grown 10 times in the last 4 years, and it reached a record high of 1.1 billion in 2014. A significant portion of the data leak incidents are due to human errors, for example, a lost or stolen laptop containing unencrypted sensitive files, or transmitting sensitive data without using end-to-end encryption. A recent Kaspersky Lab survey shows that accidental leak by staff is the leading cause for internal data leaks in corporates. The data-leak risks posed by accidents exceed the risks posed by vulnerable software. In order to minimize the exposure of sensitive data and documents, an organization needs to prevent clear text sensitive data from appearing in the storage or communication.

A screening tool can be deployed to scan computer file systems, server storage, and inspect outbound network traffic. The tool searches for the occurrences of plaintext sensitive data in the content of files or network traffic. It alerts users and administrators of the identified data exposure vulnerabilities. For example, an organization's mail server can inspect the content of outbound email messages searching for sensitive data appearing in unencrypted messages. Data leak detection differs from the anti-virus (AV) scanning (e.g., scanning file systems for malware signatures) or the network intrusion detection systems (NIDS) (e.g., scanning traffic payload for malicious patterns). AV and NIDS typically employ automated based string matching (e.g., Aho-Corasick, Boyer-Moore, which match static or regular patterns.

II. RELATED WORK

In [1], **X Shu, D Yao, E Bertin- "Privacy preserving detection of sensitive data exposure",** Statistics fromsecurity firms, research institutions andgovernment organizations show that thenumber of data-leak instances has grownrapidly in recent years. Among variousdata-leak cases, human mistakes are one ofthe main causes of data loss. There existsolutions detecting inadvertent sensitivedata leaks caused by human mistakes andto provide alerts for organizations. Acommon approach is to screen content instorage and transmission for exposedsensitive information. Such an approachusually requires the detection operation tobe conducted in secrecy. This secrecyrequirement is challenging to satisfy inpractice, as detection servers may becompromised or outsourced. In this paper,we present a privacy-preserving data-leakdetection (DLD) solution to solve the issuewhere a special set of sensitive data digestsis used in detection. The advantage of ourmethod is that it enables the data owner tosafely delegate the detection operation to asemi honest provider without revealing the ensitive data to the provider. We describehow Internet service providers can offertheir customers DLD as an add-on servicewith strong privacy guarantees.

In [2], **XiaokuiShu, Jing Zhang, Danfeng (Daphne) Yao, Wu-Chun Feng- "Rapid Screening of Kevin Border,**

Inc,AnnArbor-" Quantifying Information Leaks in Outbound Web Traffic", As the Internet grows andnetwork bandwidth continues to increase,administrators are faced with the task of keeping confidential information fromleaving their networks. Today's networktraffic is so voluminous that manualinspection would be unreasonablyexpensive. In response, researchers havecreated data loss prevention systems thatcheck outgoing traffic for knownconfidential information. These systemsstop naïve adversaries from leaking data,but are fundamentally unable to identifyencrypted or obfuscated information leaks.What remains is a high-capacity pipe fortunnelling data to the Internet. We takeadvantage of the insight that most networktraffic is repeated or determined byexternal information, such as protocolspecifications or messages sent by aserver. By filtering this data, we canisolate and quantify true information.flowing from a computer. In this paper, wepresent measurement algorithms for theHypertext Transfer Protocol (HTTP), themain protocol for web browsing. Whenapplied to real web browsing traffic, thealgorithms were able to discount 98.5% ofmeasured bytes and effectively isolateinformation leaks.

In [3], **SomeshJha- "Towards Practical Privacy for Genomic Computation",** Many basic tasks incomputational biology involve operationson individual DNA and protein sequences.
These sequences, even when anonymized, are vulnerable to re-identification attacks and may reveal highly sensitive information about individuals. We present a relatively efficient, privacy-preserving implementation of fundamental genomic computations such as calculating the edit distance and Smith- Waterman similarity scores between two sequences. Our techniques are cryptographically secure and significantly more practical than previous solutions. We evaluate our prototype implementation on sequences from the database of protein families, and demonstrate that it's performance is adequate for solving real world sequence-alignment and related problems in a privacy preserving manner. Furthermore, our techniques have applications beyond computational biology. They can be used to obtain efficient, privacy-preserving implementations for many dynamic programming algorithms over distributed datasets.

In [4], **Sailesh Kumar, BalakrishnanChandrasekaran, Jonathan Turner "Curing Regular Expressions Matching Algorithms from Insomnia, Amnesia, and Acalculia",** Theimportance of network security has growntremendously and a collection of deviceshave been introduced, which can improvethe security of a network. Networkintrusion detection systems (NIDS) areamong the most widely deployed suchsystem; popular NIDS use a collection ofsignatures of known security threats andviruses, which are used to scan eachpacket's payload. Today, signatures areoften specified as regular expressions; thusthe core of the NIDS comprises of aregular expressions parser; such parsersare traditionally implemented as finiteautomata. Deterministic Finite Automata(DFA) are fast, therefore they are oftendesirable at high network link rates. DFAfor the signatures, which are used in thecurrent security devices, however requireprohibitive amounts of memory, whichlimits their practical use. In this paper, weargue that the traditional DFA based NIDShas three main limitations: first they fail toexploit the fact that normal data streamsrarely match any virus signature; second,DFAs are extremely inefficient infollowing multiple partially matchingsignatures and explodes in size, and third,finite automaton are incapable ofefficiently keeping track of counts. Wepropose mechanisms to solve each of thesedrawbacks and demonstrate that oursolutions can implement a NIDS muchmore securely and economically, and atthe same time substantially improve thepacket throughput.

Disadvantages of existing system

- Inadvertent data leak.
- Malicious data leak.
- Data traffic and time consuming.
- Static filtering of authorized users.

III. METHODOLOGY

In our proposed system we propose a data-leak detection solution which can be outsourced from organization, we design and implement Lucerne search engine framework Levenshtein-distance technique to avoid data leak and also provide privacy preserving to Sensitive data. Two most important players in our proposed model is

- **Data Owner** owns the Sensitive data and authorizes the DLD provider to inspect the network traffic from the organizational networks for anomalies, namely inadvertent data leak.
- **Mail Server - DLD provider** inspects the network traffic forpotential data leaks.

We focus ondetecting inadvertent data leaks,and we assume the content in filesystem or network traffic isavailable to the inspection system.A supervised network channelcould be an unencrypted channel oran encrypted channel where thecontent in it can be extracted andchecked by an authority. Authority has the threshold for everycategorized position of users.

- In our security model, we assume that the analysis system is secure and trustworthy. Privacy-preserved data-leak detection can be achieved by leveraging special protocols and computation steps. It is another functionality of a detection system.
- We implement the web service to maintain the users and Sensitive content instead of data bases because of static implementation and rough data handling. Even the Sensitive data storage have to preserved from threatens in existing

system. For that purpose we used to maintain the Sensitive data in cloud.

Advantages of proposed system

• The implemented Levenshte in distance technique and Lucene search framework are used to avoid data leak.
• The DLD provider inspects the network for data leaks in the mail server.
• Threshold is given to all employees.
• Network traffic is reduced.

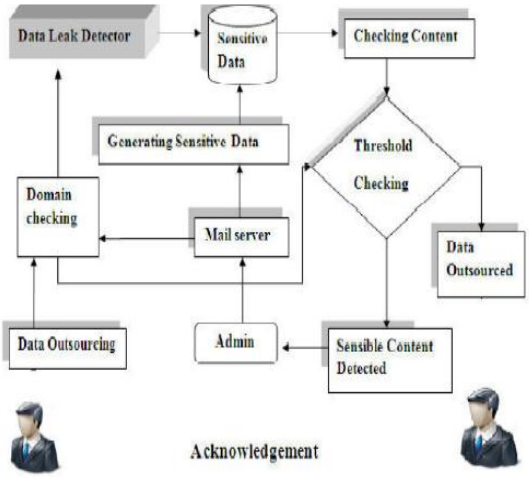

FIGURE 1: OVERALL SYSTEM ARCHITECTURE

IV. CONCLUSION

Fast detection of data-leakage framework to avoid sensitive data exposure and also provide privacy preserving to sensitive data. Lucerne search framework to detect the sensible data easily using indexing technique. Levenshte in distance algorithm to detect the shuffling of transferred mail content. To implement the own logics for detect sampling of transferred mail content appropriately. We implement threshold rate based on assigning and checking domains based user filtering technique. In future it will be created with the deleting of request list if it is checked or proceeded once.

REFERENCES

[1] X. Shu, D. Yao, and E. Bertino, "Privacy preserving detection of sensitive data exposure," *IEEE Trans. Inf. ForensicsSecurity*, vol. 10, no. 5, pp. 1092–1103, May 2015.
[2] K. Borders and A. Prakash, "Quantifying information leaks in outbound Web traffic," in *Proc. 30th IEEESymp. Secur. Privacy (SP)*, May 2009, pp. 129–140.
[3] S. Jha, L. Kruger, and V. Shmatikov, "Towards practical privacy for genomic computation," in *Proc. IEEE Symp.Secur.Privacy*, May 2008, pp. 216–230.
[4] S. Kumar, B. Chandrasekaran, J. Turner, and G. Varghese, "Curing regular expressions matching algorithms from insomnia, amnesia, and acalculia," in *Proc. 3rd ACM/IEEE Symp. Archit. Netw.Commun. Syst. (ANCS)*, 2007, pp. 155–164.

