Secret Detection of Sensitive Data Leakage

Miss. Pooja Kolte
Post Graduate Student,
Department of Computer Engineering,
RMD Sinhgad School of Engineering,
Savitribai Phule Pune University, Pune, India.
itzpoojakolte@gmail.com

Abstract—Surveys from many years have shown that many data leakages has been found due different problems like malicious attacks, hacking, different attacks. Approximately 28% of all data leakages are due to human mistakes which are increasing now a day. There exists lot of techniques to find out the leakages. Some cloud also provides this data leak detection (DLD) as an add-on service but they are semi-honest means they provide information about leakages but also attempt to get sensitive data of our organization. So proposed work will securely detect leakages and also provide privacy for our sensitive data. This proposed model uses fuzzy fingerprint algorithm to find out the human data leakages securely.

Keywords—Data leak detection; Privacy Preservation; Fuzzy Fingerprint; Shingles; Network Security.

I. INTRODUCTION

According to survey, there are many attacks occur for data leakages but we will now just focus on human mistakes which increases dramatically from 412 million in 2012 to 822 million in 2013[1][7]. This network data leak detection searches for data leakages in network with deep packet inspection by comparing data provided by organization and data traffic flow in network. It generates trigger to organization or data owner when data leak detection (DLD) will found same data as organization provided in network then data owner check for accuracy of comparison and finds that it is exact leakage or not. So for this earlier bloom filter like techniques are used to find out data leakages.

Previously data owner provides plaintext as a sample to find out leakages. Data leak detection (DLD) providers detects data leak but also attempt to get that sensitive data which is unauthorized for that DLD providers. Cloud provides this service as add-on service but these are sometimes semi-honest. Therefore we want a secure and privacy preserving technique for detection. The proposed fuzzy fingerprints allow us to provide private and secure DLD. It permits data owner to delegate sensitive data in secure manner. Using our method data owner can trust on DLD providers to find leakages out. This problem of the lack of support for privacy-enhancing data-leak detection has not been systematically addressed in the security literature. Here they design, implement, and experimentally evaluate an efficient technique that enhances the data privacy during the data-leak detection operations. This method is based on a fast and practical one-way computation and does not require any expensive cryptographic operations.

In this model, the data owner computes a special set of digests or fingerprints from the sensitive data, and then discloses only a small amount of digest information to the DLD provider. These fingerprints have important properties, which prevent the provider from gaining knowledge of the sensitive data, while they enable accurate comparison and detection. The DLD provider performs deep packet inspection (DPI) to identify whether these fingerprint patterns exist in the outbound traffic of data owners organization or not. The DLD provider is trusted to perform inspection on network traffic, but may attempt to learn the information about the sensitive database provided by the data owner, or to discover the leaked data easily from the network traffic. Existing work on cryptography-based multi-party computation is not efficient enough for practical data leak inspection in this setting.

In this paper, proposed model presented the detailed work to demonstrate the feasibility and effectiveness of approach. Paper contributions are summarized as follows.

1) Privacy preserving detection of sensitive data leaks from DLD providers.
2) Secure detection of data leak without knowing to unauthorized system.

This technique provides design, implementation and evaluation of fuzzy fingerprint for privacy preserving data-leak detection. This paper is organized as follows. Data-Leak Detection service has its some privacy problems for sensitive data of data owner or organization and motivation and overview is illustrated in section 1 introduction. Section 2 includes literature review of existing methodologies based on survey papers. Section 3 explains the proposed model to detect data-leak privately. This paper also includes demonstrated evaluation of proposed work in section 4. The final section provides a conclusion for our proposed approach and future work. At last referred papers are listed out which are explained in related work.

II. RELATED WORK

Many techniques related to data leak problem have been found as a solution. In this system [2] Larry proposed an algorithm for anonymous sharing of private data among N parties is developed. This technique is used iteratively to assign these nodes ID numbers ranging from 1 to N. This assignment is anonymous in that the identities received are unknown to the other members of the group. Resistance to collusion among other members is verified in an information theoretic sense when private communication channels are used. This assignment of serial numbers allows more complex data to be shared and has applications to other problems in privacy preserving data mining, collision avoidance in communications and distributed database access. The required computations are distributed without using a trusted central authority. The new algorithms are built on top of a secure sum data mining operation using Newtons identities and Sturms theorem. An algorithm for distributed solution of certain polynomials over finite fields enhances the scalability of the algorithms. Markov chain representations are used to find statistics on the number of iterations required, and computer algebra gives closed form results for the completion rates.

The model proposed in [3] use data processing functions are expected as a key-issue of knowledge-intensive service functions in the cloud computing environment. Cloud computing is a technology that evolved from technologies of the field of virtual machine and distributed computing. However, these unique technologies bring unique privacy and security problems concerns for customers and service providers due to involvement of expertise in data to be processed. So they have proposed the cryptographic protocols preserving the privacy of users and confidentiality of the problem solving servers.

In system [4] they have presented an approach for quantifying information leak capacity in network traffic [3]. Instead of trying to detect the presence of sensitive data an impossible task in the general case their goal is to measure and constrain its maximum volume. Advantage of the insight that most network traffic is repeated or determined by external information, such as protocol specifications or messages sent by a server. By filtering this data, can isolate and quantify true information owing from a computer. Here this model, also present measurement algorithms for the Hypertext Transfer Protocol (HTTP), the main protocol for web browsing. When applied to real web browsing traffic, the algorithms were able to discount 98.5% of measured bytes and effectively isolate information leaks. In the future, They plan to implement similar leak measurement techniques for other protocols. E-mail (SMTP) will probably be the most challenging because a majority of its data is free-form information from the user.

The new approach [5] is based on black-box differencing run two logical copies of the network, one with private data scrubbed, and compare outputs of the two to determine if and when private data is being leaked. To ensure outputs of the two copies match, build upon recent advances that enable computing systems to execute deterministically at scale and with low overheads. They believe our approach could be a useful building block towards building general-purpose schemes that leverage black-box differencing to mitigate leakage of private data. This modified Linux kernel would not be an intrusive solution for network with high demands for security, a more system transparent design would be ideal necessary.

One of them is bloom filter [6]. Bloom filter is space efficient data structure which support membership queries. It allows false positives but sometimes its overweight neglects false triggers. Bloom is verified and used in different ways as per requirement. Yoshida introduced spectral bloom filter [8] which is extension of bloom filter. It uses second filter to handle elements of minimum counter in the filter to improve accuracy of the resulting estimates. It helps to keep explicit lists only drawback with this is false positives.

To overcome this drawback our solution helps it out without providing false positives. It only generates trigger in positive leakages found. DLD providers uses fuzzy fingerprint algorithm for this private and secure direction.

III. PROPOSED WORK

A. DLD Model

This privacy-preserving data-leak detection method supports practical data-leak detection as a service and minimizes the knowledge that a DLD provider may gain during the process. Fig. 1 illustrates the six operations between the data owner and the DLD provider in our protocol, which include PREPROCESS run by the data owner to prepare the digests of sensitive data, data owner RELEASE digests to the DLD provider, MONITOR and DETECT for the DLD provider to collect outgoing traffic of the organization, compute digests of traffic content, and identify potential leaks, REPORT for the DLD provider to return data leak alerts to the data owner where there may be false positives (i.e. false alarms), and POSTPROCESS for the data owner to pinpoint true data leak instances.

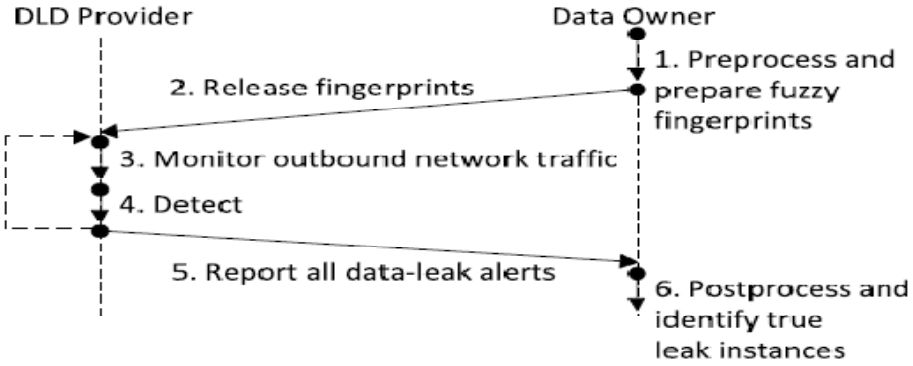

Figure 1. Data-Leak Detection Model

B. Algorithm

Step 1: Start

Step 2: Data owner does preprocessing of data on each part of sensitive data. The fuzzy fingerprint set ***S**** is obtained by PREPROCESS. The data owner keeps ***S*** for use in the subsequent POSTPROCESS operation.

Step 3: This operation is run by data owner. The fuzzy fingerprint set ***S**** is released to the DLD provider for detection procedure.

Step 4: The DLD provider monitors the network traffic ***T*** from the data owner's organization.

Step 5: This operation is run by the DLD provider on each ***T**** as follows. If DETECTION on ***T**** yields an alert, the DLD provider reports the set os detected candidate leak instances ***T**** to the data owner.

Step 6: After receiving ***T****, the data owner check every ***f' ∈ T**** to see whether it is in ***S.*** A precise likelihood of data leaking is computed at the data owner's side.

Step 7: End

C. Shingles and Fingerprints

The proposed technique of DLD provider uses fuzzy fingerprint and shingles to provide encryption to the sensitive data of data owner. Now we will see Rabin fuzzy fingerprint in details. The Rabin fingerprint is a method of implementing fingerprints using polynomials on a finite field. It was proposed by Michael O. Rabin.

Rabin fingerprint generates short and hard-to-reverse digests on plaintext. Sliding window [9] is used to generate fragments of processed data. It preserves local features of data and also provided noise tolerance property. Rabin includes shift, XOR and table lookup like mathematical calculations. Also supports fast random fingerprint selection for partial disclosure. Fuzzy will not be completed without shingles. Shingles use fixed size q-grams and divides digests into fixed size fragments. For example {pqrstuvw} is divided into 4-grams so it gives five shingles as {pqrs}, {qrst}, {rstu}, {stuv}, and {tuvw}. The use of shingle only will not satisfy our requirement so we need to use Rabin fingerprint and shingles together for best effect.

IV. EVALUATION

To prove our solution is the better one we have made some experimental evaluation. Following graph shows implementation of fingerprint algorithm. This model is implemented fuzzy fingerprint framework in Python, including packet collection, shingling, Rabin fingerprinting. Implementation of Rabin fingerprint is based on cyclic redundancy code (CRC). In experiments, the shingles are in 8-byte, and the fingerprints are in 32-bit (33-bit irreducible polynomials in Rabin fingerprint). Set up a networking environment in VirtualBox, and make a scenario where the sensitive data is leaked from a local network to the Internet. Multiple users' hosts (Windows 7) are put in the local network, which connect to the Internet via a gateway (Fedora). Multiple servers (HTTP, FTP, etc.) and an attacker-controlled host are put on the Internet side. The gateway dumps the network traffic and sends it to a DLD server/provider (Linux). Using the sensitive-data fingerprints defined by the users in the local network, the DLD server performs off-line data-leak detection.

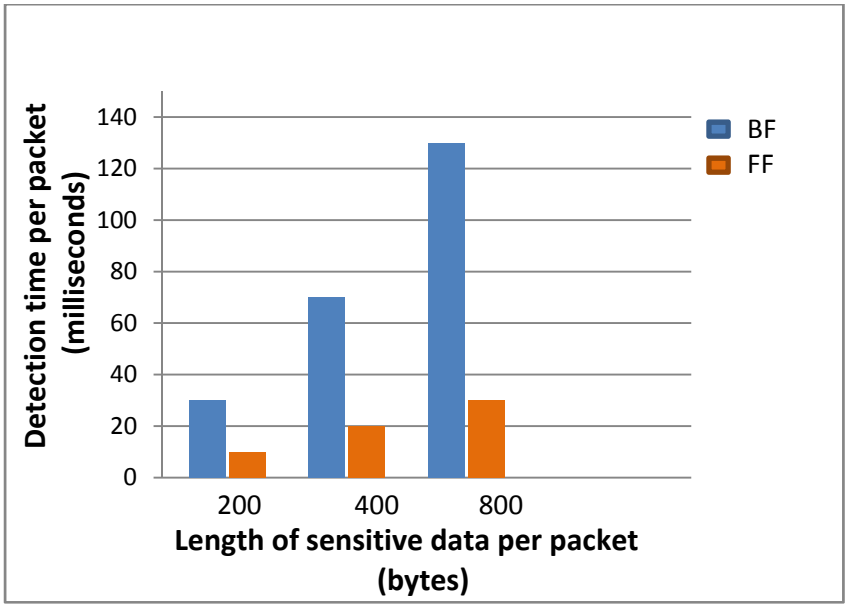

Figure 2. Detection time between Bloom filter and Fuzzy fingerprint

We can see in this Fig. 2 that as length of sensitive data increases detection time of bloom filter [6] is increases dramatically whereas detection time of our proposed fuzzy fingerprint increases very negligibly with sensitive data packet increment. It shows proposed fuzzy fingerprint data leak detection runs faster than bloom filter.

V. CONLUSION

Rapidly increasing data leak cases in now a day taken into account and so to detect the leakages here proposed a novel fuzzy fingerprint system for data-leakage detection in organization's network traffic. Using special digests, the exposure of the sensitive data is kept to a minimum during the detection. Fuzzy Fingerprints are generated to detect data leakages in network with privacy preserving strategy from semi-honest data-leak detection (DLD) Provider. Also this system validated the accuracy, privacy, and efficiency of this model.

VI. FUTURE SCOPE

As this model provides security and privacy for network-based leakages only so for future work, need to focus on designing a host-assisted mechanism for the complete data-leak detection for large-scale organizations.

ACKNOWLEDGMENTS

My special thanks to the experts my guide Miss. Kanchan Varpe who have contributed towards my survey and helped me out to make this survey paper.

REFERENCES

[1] Zixue Cheng, Peng Li, Junbo Wang, and Song Guo, Senior Members, IEEE, “Privacy preserving detection of sensitive data exposure", IEEE Transactions on Emerging Topics in Computing 2168-6750 (c) 2015 IEEE.

[2] G. Chun, S. Ihm, P. Maniatis, M. Naik, and A. Patti, “Privacy preserving data sharing with anonymous id assignment," in Proc. ACM 6th Int. Conf.on Computer syst., 2013, pp. 301314.

[3] Hiroshi Yamaguchi, Masahito Gotaishi, Phillip C.Y. and Shigeo Tsujii, “Privacy preserving data processing" (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 5 (4) , 2014, 5700-5704.

[4] Kevin Borders, Atul Prakash,\Quantifying information leaks in outbound web traffic", School of Computer Science Carnegie Mellon University Pittsburgh, PA 15213 May 2009 CMU-CS-12-148.

[5] Jason Croft, Matthew Caesar, Xin Liu, and Wenjuan Gong,\Towards practical avoidance of information leakage in enterprise networks",International Journal of Distributed Networks,10 November 2011.

[6] Karthik Kumar , Jibang Liu , Yung-Hsiang Lu, Bharat Bhargava, “Bloom filter applications in network security: A state-of-the-art survey", Springer Science Business Media, LLC 2013.

[7] Vandana singla, “Data leak detection as a service: challenges and solutions", Proceedings of 07th IRF International Conference, 22nd June-2012.

[8] Yoshida R., Cui Y., Sekino T., Shigetomi R., Otsuka A., Imai H., “Practical searching over encrypted data by private information retrieval", Proceedings of the Global Communications Conference GLOBECOM, 2010.

[9] Jian Liu, Kun Huang, Hong Rong, Huimei Wang and Ming Xian, “Privacy-preserving public auditing for regenerating-code-based cloud storage ", IEEE Transactions on Information Forensics and Security, 2015.