A Survey Paper on Data Leak Detection using Semi Honest Provider Framework

Chinar Bhandari[1], Dr. Srinivas Narasim Kini[2]

[1]M.E (Computer) Department of Computer Engineering, Jayawantrao Sawant College of Engineering, Pune, India. Savitribai Phule Pune University, Pune, Maharashtra, India -411007

[2]Assistant Professor (Computer) Department of Computer Engineering, Jayawantrao Sawant College of Engineering, Pune, India. Savitribai Phule Pune University, Pune, Maharashtra, India -411007

Abstract: ***Current Statistics from various security organization research firms and government institutes suggest that there has been a rapid growth of data leak in past 8 years. Thera are various reasons for data leaks amongst which human errors is most endorsing of all. There are many ways in which we can have an audit trail verifying for data leak in networks, but still not all consider human errors as an important check factor, which makes them more prone to fail. Some of the common solutions include keeping a copy of sensitive data at the providers end and maintaining an audit trial for checking whether there is any data leakage in networks, which in turn notifies the organization about data leakage. All these methods in turn are prone to data attacks, as the keep copy of data while auditing.So keeping these errors in mind, we present a novel data leak detection (DLD) framework where we maintain a set of important data or sensitive data digests which is used for the purpose of detection of any data leakage in the network. So this method proves to be non-erroneous due to the fact that the owner does not completely have to expose the sensitive data to the semi honest provider for data leak detection method. The owner has data control rights i.e. he can decide what part of data to be revealed to the semi honest provider and which not to reveal.***

Keywords: Data leak Detection Method (DLD), semihonest provider (SHP), network security, framework

1. Introduction

According to current Statistics from various security organization research firms and government institutes suggest that there has been a rapid growth of data leak in past 8 years. There are various reasons for data leaks amongst which human errors is most endorsing of all. There are many ways in which we can have an audit trail verifying for data leak in networks, but still not all consider human errors as an important check factor, which makes them more prone to fail. Some of the common solutions include keeping a copy of sensitive data at the providers end and maintaining an audit trial for checking whether there is any data leakage in networks, which in turn notifies the organization about data leakage. All these methods in turn are prone to data attacks, as the keep copy of data while auditing. Also methods like deep packet analysis which searches for any relating data patterns. In this method, payloads of TCP/IP packets is analyzed for any alter in data or any data pattern matching which may form sensitive data collection in network. This data is then compared with the threshold value, and if it exceeds the threshold value, the detection system alerts or notifies the organization.

The data leak detection system can be used as a framework for internet service providers or it can also be integrated into the network. All these methods require a plain text or log form of data which is used to check for any data leakage by the 3rd party provider. But exposing the data to these provider may in turn result to data leakage during the audit trial for data leak detection method. So to avoid this, we propose a novel data leak detection (DLD) framework where we maintain a set of important data or sensitive data digests which is used for the purpose of detection of any data leakage in the network. So this method proves to be non-erroneous due to the fact that the owner does not completely have to expose the sensitive data to the semi honest provider for data leak detection method. The owner has data control rights i.e. he can decide what part of data to be revealed to the semi honest provider and which not to reveal. In this survey paper, we propose a data leak detection framework which can be used as a semi honest provider in the network itself or can also be outsourced. In this system we are implementing a fuzzy finger print technique that is an additional security check parameter for data leakage method.

This method proves to be faster than any another method and is based on one way computation of exposure of sensitive data. It gives the data owner rights of integrating data specific content securely to the DLD without actually exposing the sensitive data. So, this ensures that the semi honest provider has a very less amount of knowledge of the actual sensitive data. We have also given provisions wherein individual can themselves mark their sensitive data and ask the admin of their local repository to check for any data leak. In the solution procedure, we compute a method where the owner of the data contains a set of fingerprints or information digests of his own from the marked data, and can expose a small amount of part of the sensitive digest to the semi honest provider. The provider will then check for any data leak detection in that part of digest, where the digest is composed of real leaks and noise which is added by the data owner in order to assure that the provider do not acquire exact knowledge of the data. The provider sends back the potential leaks if found any by him.It is the data owner who post-checks for any data leaks in real time, as he has the knowledge of the noise added by him.

2. Literature Survey and Related Work

There has been many work going on recently in solutions for the data leak detection method. There are many ways in which we can have an audit trail verifying for data leak in networks, but still not all consider human errors as an important check factor, which makes them more prone to fail. Some of the common solutions include keeping a copy of sensitive data at the providers end and maintaining an audit trial for checking whether there is any data leakage in networks, which in turn notifies the organization about data leakage. In [1] the author Xiaokui Shu, Danfeng Yao proposes a method where the data owner has data control rights i.e. he can decide what part of data to be revealed to the semi honest provider and which not to reveal. In this survey paper, we propose a data leak detection framework which can be used as a semi honest provider in the network itself or can also be outsourced. In this system we are implementing a fuzzy finger print technique that is an additional security check parameter for data leakage method.

This method proves to be faster than any another method and is based on one way computation of exposure of sensitive data. It gives the data owner rights of integrating data specific content securely to the DLD without actually exposing the sensitive data. So, this ensures that the semi honest provider has a very less amount of knowledge of the actual sensitive data. We have also given provisions wherein individual can themselves mark their sensitive data and ask the admin of their local repository to check for any data leak. In the solution procedure, we compute a method where the owner of the data contains a set of fingerprints or information digests of his own from the marked data, and can expose a small amount of part of the sensitive digest to the semi honest provider. The provider will then check for any data leak detection in that part of digest, where the digest is composed of real leaks and noise

But the problem with this method is that the noise factor is set static in this case and can be improved using dynamic addition of noise at each audit trail the provider is performing at a given time interval.

In [2], authors Rabin and Shingle proposed a method where fingerprint was used as main parameter for checking data leak in network where each audit trail were checked for spam message.

Also authors Kleinberg and Papadimitriou [3], suggested that the data privacy should be categorized into sensitive and non-sensitive data so that it will help the third party provider to find data leak in the network.

In [4] author gave a solution where virus scan module was used to find the data leak in the network. The virus scan module was non-sensitive information and was just a trail module for data leak detection.

There were many solutions offered by the institutes like in [5] the Symantec DLP was proposed by the institute which was a good option for the organization to integrate the module, but lacked the capacity to outsourced. Also in [6] the institute proposed Global Velocity, a novel data leak detection solution, which incorporated many features like exposing a limited amount of data to the provider. But again this solution failed because it cannot be outsourced.

Also there are methods [7] where the solution consists of tracing and enabling the sensitive or important data work flows. This method aims toward integrating the solution in the router itself, which is different from the current versions which aims towards remote service.

3. Proposed Work

In this section, we will propose the algorithmic steps with the help of below data flow diagram: -

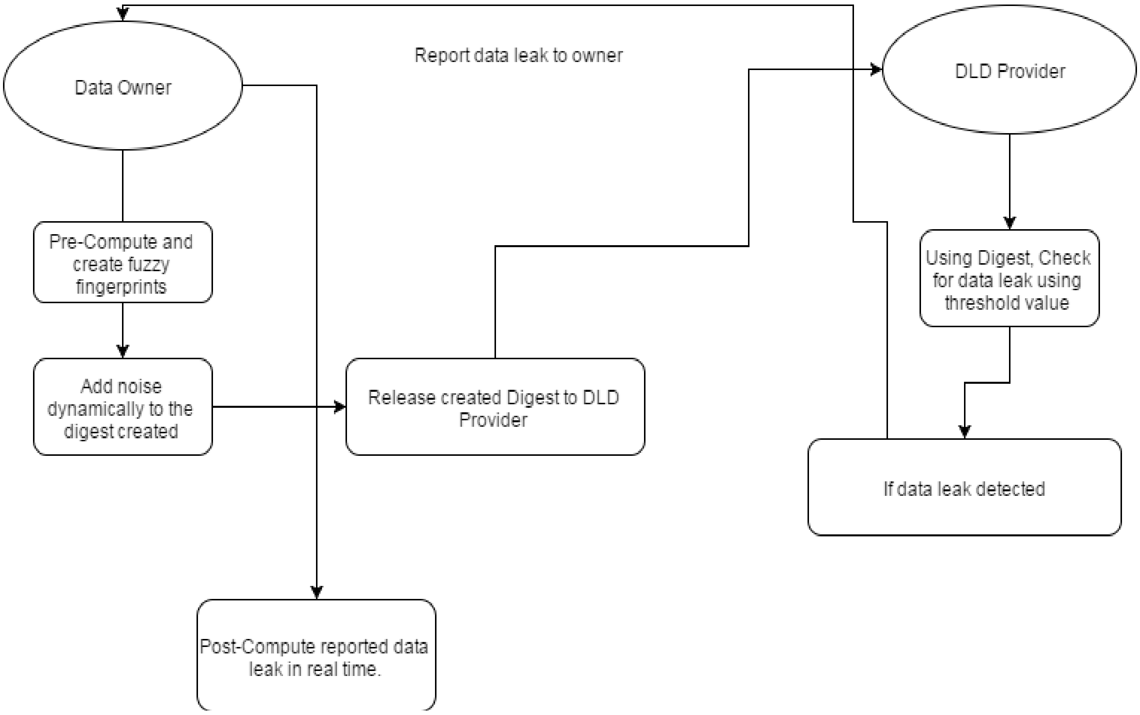
Report data leak to owner
Data Owner
DLD Provider
Pre-Compute and create fuzzy fingerprints
Using Digest, Check for data leak using threshold value
Add noise dynamically to the digest created
Release created Digest to DLD Provider
If data leak detected
Post-Compute reported data leak in real time.

Figure 1: Data flow diagram for the System

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

1) First the data owner will mark his set of sensitive data.
2) Secondly he will pre-compute all data and create a set of fuzzy fingerprint along with set of data digest.
3) He will add noise to the exposed digest, in order to assure that the semi honest provider does not gain complete knowledge about the sensitive data.
4) Then he will release the digest to the semi honest provider, to keep a track of any data leak detection in the network.
5) The DLD provider on receiving the digest, will start to check for any data leak using the digest.
6) If the provider finds any data leak in the current traffic network, he will notify it to the data owner.
7) The data owner on receiving the notification, will post-compute the data digest neglecting the noise he added, to check whether there was any data leakage in real time.

So, we are trying to give the data owner control rights i.e. he can decide what part of data to be revealed to the semi honest provider and which not to reveal. In this survey paper, we propose a data leak detection framework which can be used as a semi honest provider in the network itself or can also be outsourced. In this system we are implementing a fuzzy finger print technique that is an additional security check parameter for data leakage method.

This method proves to be faster than any another method and is based on one way computation of exposure of sensitive data. It gives the data owner rights of integrating data specific content securely to the DLD without actually exposing the sensitive data. So, this ensures that the semi honest provider has a very less amount of knowledge of the actual sensitive data. We have also given provisions wherein individual can themselves mark their sensitive data and ask the admin of their local repository to check for any data leak. In the solution procedure, we compute a method where the owner of the data contains a set of fingerprints or information digests of his own from the marked data, and can expose a small amount of part of the sensitive digest to the semi honest provider. The provider will then check for any data leak detection in that part of digest, where the digest is composed of real leaks and noise

4. Conclusion

In this survey paper we have seen that despite of enormous solutions for the data leak detection method, each method failed due to some or the other reasons to satisfy a full-fledged solution which can be integrated in the current network or can also be outsourced. Thus we try to suggest data leak detection using semi honest provider framework which can be both integrated within the network and also has provisions to be outsourced. Our framework proves to be better suggestion than any other solution.

References

[1] Xiaokui Shu, Danfeng Yao, “Privacy-Preserving Detection of Sensitive Data Exposure”, pages 1092-1103, 2015.

[2] M. O. Rabin, “Fingerprinting by random polynomials,” Dept. Math., Hebrew Univ. Jerusalem, Jerusalem, Israel, Tech. Rep. TR-15-81, 1981.

[3] J. Kleinberg, C. H. Papadimitriou, and P. Raghavan, “On the value of private information,” in *Proc. 8th Conf. Theoretical Aspects RationalityKnowl.*, 2001, pp. 249–257.

[4] F. Hao, M. Kodialam, T. V. Lakshman, and H. Zhang, “Fast payloadbased flow estimation for traffic monitoring and network security,” in *Proc. ACM Symp. Archit. Netw. Commun. Syst.*, Oct. 2005, pp. 211–220.

[5] Symantec. *Data Loss Prevention (DLP) Software*. [Online]. Available: http://www.symantec.com/data-loss-prevention/, accessed Oct. 2014.

[6] Global Velocity Inc. *Cloud Data Security From the Inside Out*. [Online]. Available: http://www.globalvelocity.com/, accessed Oct. 2014.

[7] 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. Secure* 2007, pp. 116–127.

Author Profile

Mr. Chinar C. Bhandari, is currently pursuingM.E (Computer) from Department of Computer Engineering, Jayawantrao Sawant College of Engineering, Pune, India. Savitribai Phule Pune University, Pune, Maharashtra, India -411007. He received his B.E (Computer) Degree from AISSMS IOIT, Pune, India. Savitribai Phule Pune University, Pune, Maharashtra, India -411007. His area of interest is mobile computing, web mining and network security.

Asst Prof. Dr. Srinivas Narasim Kini, received his PhD Degree from Cochin University of Science and Technology, Thrikkakara, South Kalamasserry, Cochin. He received his M.E (Computer) Degree from B.M.S. College of Engineering, Basavanagudi, Bangalore, India. He received his B.E (Computer) Degree from K L E Society's College of Engineering Udyambaug Belgaum, India. He is currently working as Asst Prof (Computer) at Department of Computer Engineering, Jayawantrao Sawant College of Engineering, Pune, India. Savitribai Phule Pune University, Pune, Maharashtra, India -411007. His area of interest is network security, data mining etc.