A NOVEL METHOD FOR TRANSFORMED DATA LEAK DETECTION AND PREVENTION IN DISTRIBUTED HETEROGENEOUS DATA SHARING NETWORKS

[1]C. Mercy Praba, [2]Dr.G. Satyavathy
[1]M.Phil Scholar, [2]Assistant Professor
[1]Department of Computer Science
[1]Sri Ramakrishna College of Arts and Science for Women, Coimbatore, TamilNadu, India

Abstract : In the recent years Data Leakage Detection as a main problem in distributed network, which has heterogeneous users. The increasing ability to trail and collect large amounts of data with the use of technology has lead to an interest in the development of data protection algorithms, which helps to preserve user sensitive data's in the distributed environment. There are several Data Leakage Detection and Data Leakage Prevention techniques have initiated for securing and protecting data's. A recently proposed technique addresses the issue of data leakage in transferred data's by sampling and matching techniques. However, the method can only perform detection part within the application level. There is several other ways to leak the transferred data within the host rather than the application. The proposed method designed and developed a **Host Data Scrutinizer** framework for distribution data security, which have developed in an email server for reconstruction secure email transfers. This includes different types of algorithms to detect and prevent data leaks either partially or fully. This is more effective than the currently available method in terms of the level of information loss. The proposed approach performs prediction and probability finding in order to identify and protect data from leakers. The experimental result shows the effective results on the distributed data in email security.

***IndexTerms* - Information Security, Data Leakage Detection, Data Leakage Prevention, Data Scrutinizer.**

I. INTRODUCTION

Protection of confidential data [1] from being leaking to the public is a growing concern among organizations and individuals. Managing Security and reduce the risk is very vital in todays fast running world [10]. Traditionally, confidentiality is preserved using security procedures such as information security policies along with conventional security mechanisms such as Firewalls, Virtual Private Networks and Intrusion Detection Systems [2][3]. Unfortunately, these mechanisms lack pro-activeness and dedication towards protecting confidential data and they required predefined rules. This can result in serious consequences, as confidential data can appear in different forms in different leaking channels [4]. Therefore, there has been an urge to mitigate these drawbacks using mechanisms that are more efficient. Data Leakage Prevention Systems (DLPSs) is dedicated mechanisms to detect and prevent the leakage of confidential data in use, in transit and at rest [5]. DLPSs use different techniques to analyze the content and the context of confidential data to detect and prevent the leakage. Although DLPSs are increasingly being designed and developed as standalone products by IT security vendors and researchers, the term still ambiguous. In this paper, a Data Leakage Detection and Data Leakage Prevention scheme for email security is proposed. This paper explicitly defines proposed DLP in the distributed heterogeneous data sharing networks. Nowadays data publishing is a common process. However, the issue of data sharing is difficult when the data is more sensitive and secure. Sometimes sensitive data may leaked and found in unauthorized places, so every data needs more concern when the data is distributed among the multi-party. The data should be safe and secure from unauthorized users. Data distributor affected a lot because of the data leakage problem. Some simple and recent research and news show the authorized users have shared the secret data with some other unauthorized users. Therefore, improving the privacy and identifying the data leaker with proof is the main aim of the proposed system. For example, a company may have partnerships with other companies that require sharing customer data. In addition, an algorithm has presented for distributing objects to agents, in a way that improves the chances of identifying a leaker. The Objective of the current proposal is to achieve a complete data security and applying for rule-based data protection from data leakage in the distributed environment. The proposed system aims at protecting the data against an adversary who has the knowledge and partial rights of at most m items in a specific transaction and also provide a cost-effective non-cryptographic data protection against private data- sharing and rule-based data leakage protection and leaker detection.

Protecting owner information is also an important goal, which performed by designing an effective algorithms to ensure better protection against data disclosure. The proposed system also aims at developing a new method, which has the ability to detect and prevent partial or full data leakage in the email server. Another goal is to detect when the distributor transmits a sensitive data, a user may leak the data using other applications. In this case, the detection of partial or full leaked data will be more helpful. The proposed system aims to detect the leaker when they try to leak the whole or a part of file of the owners. Maximize the chances of detecting a guilty agent that leaks all his data objects by applying the prediction strategies.

II. PROBLEM DEFINITION

In the literature, a set of DLD and DLP techniques was proposed [6]. Recently a Content Inspection Technique has proposed to detect the data leaks of sensitive information in the content of files or network traffic. The Data Leak Detection approach is based on aligning two sampled sequences for similarity comparison. The Sequence Alignment Technique [7] is used for Data Leakage Detection and Complex Data-leak Patterns. The algorithm is designed for detecting long and inexact sensitive data patterns. This detection is performed with a comparable sampling algorithm, which allows one to compare the similarity of two separately sampled sequences. This system achieves good detection accuracy in recognizing transformed leaks. The Sequence Alignment algorithm can only track the application level data leakage rather than another type of Data-Movement based tracking. The existing system failed to perform both detection and prevention of data leaks. In the existing sequence alignment and sampling techniques were used to only detect the transformed data leaks. However, the application was not performed a complete data leak detection and protection. The data leakage within the application can be easily detected, where data movement based detection has not yet performed well. The existing DLD and DLP need more computations to detect and prevent data leaks.

There are several challenges associated with Data Leakage Prevention (DLP), data authorization with Partial and Full Data Leakage Detection (DLD) [8]. First, the Data Leakage Detection provider gains knowledge about the sensitive data when the data contains a leak.

The challenge is how to authorize the data distributor and restrict the degree of information that can be learned by the owner in case of data leaks. The provider has access to the plaintext packet payload. The second challenge is how to make the detection accuracy high. In literature, it is evident that Data Leakage Detection and Prevention methods are inadequately studied. Most of the current methods suffer from serious limitations, especially when the confidential data is evolving. This is because they mainly depend on inflexible techniques. Even with some robust fuzzy fingerprinting [9] and statistical analysis, the confidential data semantics can be leaked using various obfuscations. Therefore, a potential research question in this area is “how to detect semantically the content of confidential data in order to prevent data leakage”. An effective future DLPS should have the ability to classify confidential data semantically even if it is evolving. Although some researchers insist on relying it is hard to protect the semantics without knowing the content. Current DLPSs maintain copies or references of confidential data. This allows successful identification of leaks when they are occurring. Unfortunately, this is not sufficient since confidential data can be created without going through classification procedures. DLPSs should have the ability to heuristically detect such data without the need for managing exact copies of existing and new data. In addition, detection techniques require extensive analysis that includes deep content inspection and comprehensive indexing.

III. HOST DATA SCRUTINIZER FRAMEWORK

For effective Data Leakage Detection and Data Leakage Prevention in the distributed systems, a new protection framework is designed and developed. The proposed framework is named as HDS (Host-based Data Scrutinizer), which helps to detect, protect and tracks the data leaks partially and fully using a set of algorithms. This includes the following algorithms to achieve the same.

- Data Monitoring Algorithm and Probabilistic Incremental Program Evolution.
- Data Movement Tracking and Alerting Technique to prevent data leaks
- AHMM (Auto-Regressive HMM(Hidden Markov Model))- for pattern matching and prediction of partial and full data leaks in the email application

Data Monitoring Algorithm and Probabilistic Incremental Program Evolution help to monitor and track the data with rule matching process. Auto-regressive method is used to predict the future based on the past behavior. An autoregressive process operates under the premise that past values have an effect on current values. AHMM is helping to find the hidden aspects of the data transformations, which able to predict the exact data leaks. The proposed systems overcome the above drawback by adapting new data leak detection and prevention method. The proposed system effectively deployed a new technique to find the data leak within the application transforms and from data movements. It developed to secure a complete mail server without data leak and ability to detect transferred partial or full leaks. The proposed system is more effective and very useful for detecting multiple data leak scenarios. This helps to detect and protect the data leaks by customized rule. The framework has the ability to identify both partial leaks or fully leaked data’s and leakers with Host Monitoring and File Movement Tracking. This can be applied in distributed network security like Mail Server Application with interactive and dynamic DLD and DLP. The overall phases of the proposed system are divided into three segments, for each segment a new technique/algorithm is designed for the effective DLD and DLP process.

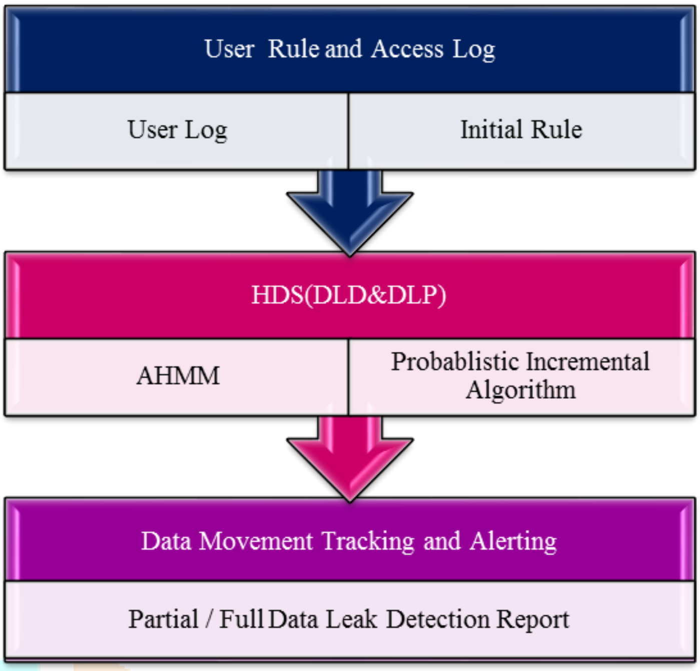
User Rule and Access Log
User Log
Initial Rule
HDS(DLD&DLP)
AHMM
Probablistic Incremental Algorithm
Data Movement Tracking and Alerting
Partial / Full Data Leak Detection Report

Figure 1 Proposed HDS Architecture

Figure 1 shows the overall architecture of the proposed HDS system, which takes user initial rule specification and the transfer details for Data Leak Detection and Prevention. The proposed system collects the user's activity log in online and offline and gets the rules for data protection from the users. That is described on the first phase. These details are passed to the HDS process, which performs the detection of hidden patterns and calculates the probability of data being leaked by a user by using their log. Using the hidden factors and data movement tracking information's, the system find the data leak either partially or fully. It returns the percentage for the leaked data.

Algorithm: AHMM
Input: Rules List (R), transformed data D, t-threshold
Output: Matching Report M, matching score S, SL-Suspected List.
Steps:

```
1: Get the rule r1,r2…rn from R
2: Read the data D
3: for Ri = 1, . . . , n (total rules) do
   Get pattern P from (Di).
   Match the rules with D and do
   If(Ri==”sensitive”)
       Generate_Pattern P(Di)
        Match_pattern S=sim(Pi, Pj)
    return S.
    if (S>t)→SL(Di)
4: if SL(Di)>t declare as Data leak
5: else repeat step 3
```

The probability can be calculated by, Q = q1q2 . . .qN a set of states

A = [aij]NxN a transition probability matrix A, each aij representing the probability of moving from state i to state j. q0,qend a special start state and end state which are not associated with observations. Here the states represented as partially leak, fully leak, and normal.

The proposed system performs the AHMM process to find the hidden states from the data movement and user log reports. So it's a collaborative process, which incrementally performs the data leakage from the other set of algorithms. This initially gets the rule list, transaction log and the threshold value from the user. The proposed AHMM gives the matching score with the suspected list as output. For each rule, and data D, it matches the sensitive rules. If the data is specified as sensitive, then the message id is used to track the content. This content will be matched with the other set of suspected contents and provides the matching score. The step shows the matching process. And it checks the threshold in the next step and finds the data leak. The next iteration of AHMM is based on the rules it will find the probability score. For every pattern and its frequency of pattern Wi, the frequent tag in a document is calculated and divided by its frequency score.

From the data movement, the Probabilistic Incremental Algorithm generates used with the probability of each transaction sequence. This begins with the original dataset and user log. For every user u, and their previous mail, the sequence number will be matched. For example if user A transmits their sensitive data to user B, the sequence is A-B, and the sequence of the content is "account details". If the same data with the same label or modified content has transmitted to User C, then the sequence will be A-B-C, if the data is not directly transmitted to the C from user B. then the suspected list from the AHMM is used to find the sequence of the transaction. From the probability calculation, it gets the sum value from the user log. For example, the deleted count, copied count of a particular file is used and the count of redundancy describes the total transaction log count. From the two data it finds the probability of being guilty user among others.

IV. RESULTS AND DISCUSSION

A. Implementation

The implementation performed by developing an email server, which poses different features and functions. To design an email server, the high-configured hardware's and storage systems have allocated. The experiments have designed with various features. These include the evaluation of the performance of the proposed system with the email server data's. The performance of the proposed work HDS Scheme has compared with the existing algorithms based on the following parameters.

B. Results

The comparison made with the recent approach such as Sequence Alignment Solution (Subsequence-Preserving Sampling Algorithm) and Alignment Algorithm (Recurrence Relation in Dynamic Programming). The figure below shows the results and comparison of the proposed system.

TABLE 1.0 Comparison Table

No.of Samples (Transaction count in numbers)	**Detection Process**			
	Alignment and Sampling		**HDS (AHMM+PIA)**	
	True Positive (%)	**False Positive (out of 1)**	**True Positive (%)**	**False Positive (out of 1)**
10	89.4	0.4	97.2	0.098
20	88.8	0.3	98.2	0.105
30	88	0.45	98.3	0.13
40	80.3	0.48	96.1	0.12
50	82.8	0.5	91.8	0.3

From the results shown in Table 1.0, it is found that proposed technique increased the True Positive Rate significantly and reduces the False Positive Rate to an acceptable extent.

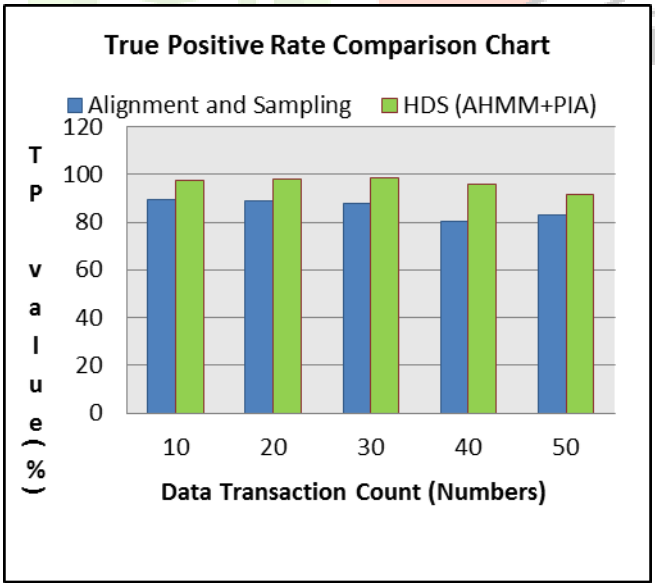

Figure 2 Performance Measure in Terms of True Positive Rate

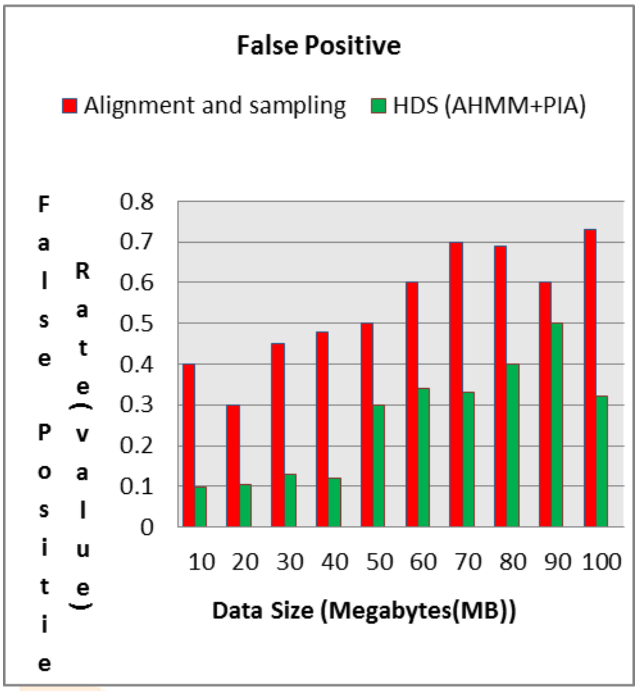

Figure 3 Performance Measure in Terms of False Positive Rate

The False positive rate of the proposed system is quite high because some normal data with some extra sensitive data merged could be considered as a leakage, but only the similarity score determines the percentage of data leaked. The reduction in the false positive rate of the proposed system is mainly due to the HDS (AHMM&PIA) process.

TABLE 2.0 Time Comparison Table

Type	**Alignment and Sampling** (Milliseconds)	**HDS(AHMM &PIA)** (Milliseconds)
Data Movement Tracking Time	Null	1.5
Rule Matching Time	8	3.4
Similarity Calculation Time	5	2.6

There are 1000 transfer records data available with normal and leakage data's from the total data,10 percent of the data has been chosen as a leak for analysis and experimentation.

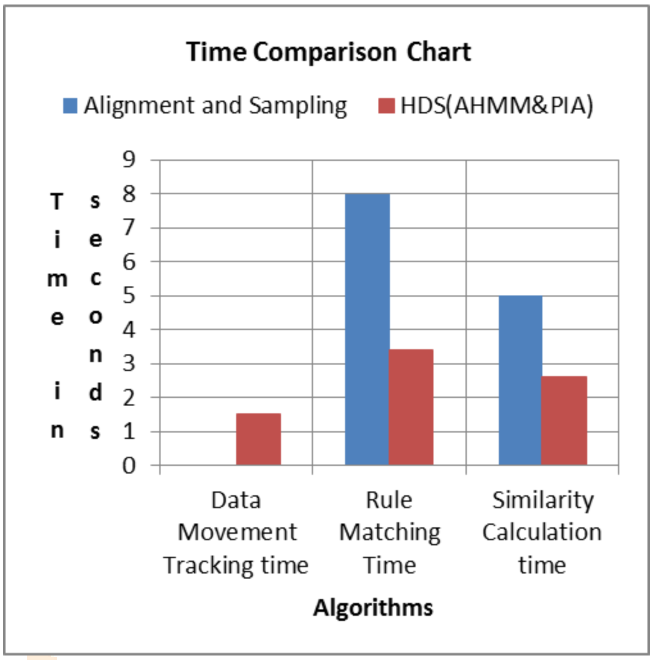

Figure 4 Time Comparison between Existing Alignment and Sampling and Proposed HDS(AHMM&PIA)

The HDS(AHMM&PIA) is designed to keep track of the hidden leakage patterns. It is observed that the performance is very hopefully compared to the existing methods that have been explored in the above figure 3 and 4.

V. CONCLUSION

There is a huge need for protection against data leakage over sensitive data. Watermarking those sensitive data's for protection may create additional data-stealing issues. This is not a perfect way to prevent data from unknown access and transferred data leaks. The need for new detection and prevention technique may lead to the perfect Data Leakage/Leaker Detection. In the research proposal, the concept is added to detect data leakage and preventing sensitive data's among other users. This also includes an extended probability model named as Probabilistic Incremental Algorithm (PIA) with Auto Regressive Hidden Markov Model (AHMM) for effective data protection. This algorithm predicts the future data leakers by analyzing the user data log and behavior. The guilty user-finding phase helps to track all the data leakers. The Data Movement Tracking has developed to identify and prevent the guilty users. The proposed research analyzed the data leaker prediction based on the activity log. There is a chance for low-level accuracy. The problem is the extension of the allocation strategies so that they can handle user requests in an online social network fashion. The presented strategies assume that there is a fixed set of the user with requests known in advance.

REFERENCES

[1] Garfinkel, Robert, Ram Gopal, and Paulo Goes. "Privacy protection of binary confidential data against deterministic, stochastic, and insider threat." *Management Science* 48, no. 6 (2002): 749-764.

[2] Kuwatly, Iyad, Malek Sraj, Zaid Al Masri, and Hassan Artail. "A dynamic honeypot design for intrusion detection." In *Pervasive Services, 2004. ICPS 2004. IEEE/ACS International Conference on*, pp. 95-104. IEEE, 2004.

[3] Resmi, A. M., and R. Manicka Chezian. "An extension of intrusion prevention, detection and response system for secure content delivery networks." In *Advances in Computer Applications (ICACA), IEEE International Conference on*, pp. 144-149. IEEE, 2016.

[4] Lin, Lang, Wayne Burleson, and Christof Paar. "MOLES: malicious off-chip leakage enabled by side-channels." In *Proceedings of the 2009 International Conference on Computer-Aided Design*, pp. 117-122. ACM, 2009.

[5] Shabtai, Asaf, Yuval Elovici, and Lior Rokach. *"A survey of data leakage detection and prevention solutions".* Springer Science & Business Media, 2012.

[6] C. Mercy Praba, and Dr.G. Satyavathy, "A Technical Review on Data Leakage Detection and Prevention Approaches." *Journal of Network Communications and Emerging Technologies (JNCET),www.jncet. org* ,Volume 7, Issue 9, September 2017.

[7] Shu, Xiaokui, Jing Zhang, Danfeng Daphne Yao, and Wu-Chun Feng. "Fast detection of transformed data leaks." *IEEE Transactions on Information Forensics and Security* 11, no. 3 (2016): 528-542.

[8] Guevara, César, Matilde Santos, and Victoria López. "Data leakage detection algorithm based on task sequences and probabilities." *Knowledge-Based Systems* 120 (2017): 236-246.

[9] Shu, Xiaokui, Danfeng Yao, and Elisa Bertino. "Privacy-preserving detection of sensitive data exposure." *IEEE transactions on information forensics and security* 10, no. 5 (2015): 1092-1103.

[10] Dr. G. Satayavathy and C. Mercy Praba, "A Study on Cyber Physical System and Network Security." CiiT International Journal of Networking and Communication Engineering, Vol 9, No 3, March 2017, 0974-9713.