A Novel Approach to Detect Data Leakage Detection and Image Deduplication Using DMA (Data Monitoring Algorithm)

Basheer. P
FDP Substitute lecturer, Department of Computer Science, KAHM Unity Women's College, Manjeri, Malappuram, Kerala, India

Abstract – In the recent years data leakage is detected as a main problem in distributed data mining. 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 security algorithms which helps to preserve user privacy in the distributed environment. There are several security techniques has been initiated for securing and protecting data's and image Deduplication. Recently proposed techniques are mostly focuses the issue or problem of privacy preserving. In this problem is occurred by data distributing and reconstruction of distributed data at the aggregate level. The aggregation level is order of mining. This research introduces and implements the invisible watermarking embedding algorithm for distribution reconstruction and rule verification for effective data leakage detection. This is more effective than the currently available method in terms of the level of information loss in data leakage detection. The work includes the investigation of agent guilt models that capture leakage scenarios that are not studied in the existing systems. The system developed with the main intension of achieves High Scalability through DMA. From the experimental results, detection accuracy of our proposed algorithm reaches 97.8% which is faster than the existing system.

Index Terms – Data Leakage, Deduplication, DMA.

1. INTRODUCTION

In the recent network Data Leakage is an important concern for the business organizations. Prevention of sensitive data from unauthorized entities and monitoring the data flow to avoid more security risks are the main goals of the security domain. The unauthorized or unpermitted disclosure may have serious consequences for an organization. In this consequence is match both long term and short term. To prevent from the unwanted access and transaction from happening, an organized effort is needed to control the information flow inside and outside the organization. Data leakage detection and prevention process are the important research issue, which is not always possible because several reasons [4]. Recent news and reports indicates 50 % of data's are leaked in the business sector either partially or fully. Protection of confidential data from being leaked to the public is a growing concern among organizations and individuals. Traditionally, confidentiality of data has been preserved using security procedures such as information security policies along with conventional security mechanisms such as firewalls, virtual private networks and intrusion detection systems. Unfortunately, these mechanisms lack pro-activeness and dedication towards protecting confidential data, and in most cases, they require predefined rules by which protection actions are taken. The result which is distributes the serious consequences that are confidential data. The confidential data is collects from different forms in different leaking channels. Recently, data leakage prevention systems (DLPSs) have been introduced as dedicated mechanisms to detect and prevent the leakage of confidential data in use, in transit and at rest. DLPSs use different techniques to analyze the content and the context of confidential data to detect or prevent the leakage [2]. Although DLPSs are increasingly being designed and developed as standalone products by IT security vendors and researchers, the term still ambiguous. In this study, a comprehensive analysis on the current DLPS mechanisms are discussed and finally proposed a data leakage prevention and detection scheme for email security. This chapter explicitly defines DLPS and categorizes active research directions in this field.

2. LITERATURE REVIEW

2.1 Related Work

In paper [8], authors proposed a method named as CoBAn a Context-based model for accidental and intentional data leakage prevention (DLP) and data leakage detection. The technique has the ability to detect small sections of confidential information with high success rate. This fully deal with the problem of rephrased texts, and it can be detected the modified contents. Authors contributed a novel approach for classification using the context based approach. The graph based schemes are used to match the key nodes. The CoBAn approach has several advantages like; this has the ability to detect the confidential information's in large scale documents. The visual graph idea provides easy understanding of the confidential terms and policies. The configuration process is simple and this can be customized according to the domain. The main demerits of the approach are the features used for the prevention. Additional and important features can improve the performance more. This is also suffers from Long running time. And it needs huge training samples to improve the accuracy.

In paper [9], authors concentrated on dynamic leakage detection scheme and implemented for video streaming application. There are several techniques were used to detect documental data's. This paper utilizes the video content leakage detection system using traffic pattern analysis. Using the video length and traffic patterns the authors detects the leakage. Then the proposed methodology is flexible for data leakage and accurate streaming content leakage detection. This improves the security and trust of the video content streaming networks. However, the technique is not adequate for the real time contents like documents, email etc.

In the paper [10], authors designed two new algorithms to detect the transformed data leaks. Transformations such as insertion, deletion processes result in the unpredictable leakage patterns. This may affect the sensitive information's. To solve this issue, efficient sequence comparison techniques are used. This consists of a special sampling algorithm and alignment algorithm. This calculates the similarity score between the sensitive or confidential data and the content. This improves the accuracy in data leak detection with low false rate. The main drawback of this paper is, it is not scalable and has several integrity issues. Based on the results of this paper, the sampling and alignment algorithms are enhanced in several researchers later.

In the paper [11], authors studied the effectiveness of statistical analysis techniques in confidential data semantics detection. Authors proposed a data leakage prevention classification technique, which is based on the term frequency (TFIDF). The TFIDF finds the terms and its frequency counts. The classification was based on computing the similarity between the documents and the category values. This model was tested against different scenarios in which the DLPS dealt with known, partially known and unknown data. The overall classification indicates encouraging outcomes across all scenarios. Further, a graphical representation of the classification results was applied using SVD abstraction. It is provides the analytical tool for semantic documents by visualization. The technique in this paper achieved a high score of 0.99 for both precision and recall. This technique is also suitable for the modified documents. This paper creates several issues and future directions for effective DLP. The techniques completely rely on the training samples. If the class distribution is not evenly distributed, then the result will be invalid. The algorithms creates class imbalance and classification speed related issues.

In the paper [12], authors developed Data leakage Prevention technique with time stamp approach. This is very important for giving permission to access a particular data, because in a particular period of time the data is confidential after the time stamp the same data could be non confidential, here authors developed an algorithm for data leakage prevention with time stamp. This technique collects the confidential and non-confidential documents and creates a cluster using k-means algorithm with cosine similarity function. For each cluster the key terms are identified using TFIDF. Then finally assigns the time stamp foe each document. This timestamp gives the deadlines of the document access. This approach can prevent data leakage within the time period, however the data leakage prevention for only fully leaked contents are developed. So partial data leakages are cannot be detected.

In the paper [13], sequence alignment techniques for detecting complex data-leak patterns are proposed. The algorithm is designed for detecting long and inexact sensitive data patterns. The technique proposed in the paper can only perform the data leakage detection in the network. It failed to detect the data leakage on a host. The sequence alignment techniques are based on aligning two sampled sequences for similarity comparison.

2.2 Data leakage prevention systems analysis techniques

Whether DLPSs are used for detection or prevention, usually there is an analysis phase involved in these tasks. Two main analysis techniques are used in DLPSs: context analysis and con-tent analysis. This section explains the differences between the two techniques and discusses some examples. The importance of content analysis compared to context analysis is exemplified. The significance of content statistical analysis as one of the state-of-the-art method in data leakage detection is also apparent.

A third technique called content tagging is used in some DLPSs. This technique is used to tag the file containing confidential data. Even with the most excessive alteration of content, such as changing format, compressing and encrypting, the same file tag can remain intact. Suen et al. (2013) introduced a technique called S2Loggeer to track files while travelling in the cloud. S2Loggeer is able to detect malicious actions, data leakages and data policy violation. Although this technique might seem robust, it can be bypassed if the same confidential data appears in a different unrecognized tag. Hence, content tagging can preserve the identity of the file but not the contained confidential data.

2.2.1 Context analysis

Context analysis uses metadata associated with actual confidential data. To keep track of confidential data, DLPSs perform contextual analysis of the transaction rather than of the actual data.

2.2.2 Content analysis

Since the main purpose of using DLPSs is the protection of confidential data, it is more important to focus on the content itself than on the context. This is what DLPSs with content analysis capabilities are trying to achieve. Content analysis in

DLPSs is done through three main techniques: data fingerprinting (including exact or partial match), regular expression (including dictionary-based match) and statistical analysis.

3. PROPOSED SYSTEM AND ITS CONTRIBUTIONS

3.1 Proposed System

The chapter entirely discusses the proposed research methodology and the absolute steps concerned in that proposed research work.

For effective data leakage prevention and data leakage detection 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 protect, detect 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

Here the data monitoring algorithm and probabilistic incremental program evaluation helps to monitor and track the data with rule matching process. Autoregressive is a stochastic process. In this process is used in statistical calculations in which future values are estimated based on a weighted sum of past values. The values of statistical calculation have an effect on current values. AHMM is helps to find the hidden aspects of the data transformations, which able to predict the exact data leaks.

Advantages of the proposed system:

- The proposed framework is very useful for detecting multiple common data leak scenarios.
- This helps to protect and detect the data leaks by customized rule.
- Have ability to Identify both partially leaks or fully leaked data and leakers with host monitoring and file movement tracking
- Applied in distributed network data security like mail server application with interactive and dynamic DLD and DLP.

The overall phases of the proposed system is divided into three segments, for each segment a new technique/algorithm is designed for the effective DLP and DLD process. The table 3.1 shows the process, algorithms for each process with its findings are described.

Table 3.1 proposed system outline

Process	Algorithms	Findings(outcome)	Modules
Host Data scrutinizer			
Pattern matching and predicting data leaks	AHMM (Auto Regressive HMM)-	Find the partial and fully leaked data's and leakers	1. Rule specification 2. Rule matching 3. Content/pattern matching. 4. Statistical analysis
For data monitoring and file watching process	Probabilistic incremental algorithm.	Detects the data movement in the host	1. File monitoring. 2. Probability calculation.
Rule updation process	Data-movement tracking and alerting algorithm	Prevents data leaks by embedding unknown objects	1. Fake object embedding 2. Rule updation

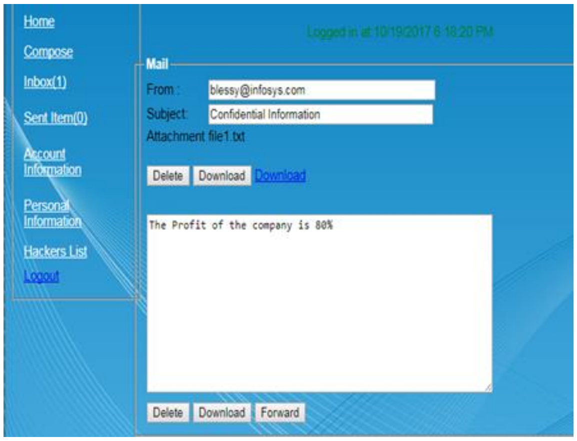
4. IMPLEMENTATION

4.1 Implementation Requirement Specification:

This section describes the implementation process. Implementation is the realization of an application, or execution of plan, idea, model, design of a research. This section explains the software, datasets and process which are used to develop the research. The proposed system has been successfully implemented in an Email application which has the following properties and functions.

4.1.1 Application:

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 are allocated. The server ID and the DNS (Domain Name Server) ids are collected and configured to create mail application. Ensuring email security is certainly not a small task. Countless mail servers and many millions of users are on the Internet today. Any system exposed to the Internet must be able to handle a large volume of traffic and also encounters a significant set of security risks. From several security risks, data leakage is more vulnerable and unsolved. How and where to address these risks plays a large part in determining the enterprise's overall security. All of these factors highlight the importance of taking a more holistic approach to email security to filter sensitive data's or otherwise inappropriate email content at multiple tiers of the network to ensure that all email is scanned and controlled.



From : blessy@infosys.com
Subject: Confidential Information
Attachment file1.txt

The Profit of the company is 80%

Fig 4.1 Email server process sample

Fig 5.1 shows the client side process, which contains inbox, sent items, compose process etc. The email process in the proposed system can be divided into two principal components i) Mail servers, which are hosts that deliver, forward, and store email. ii) Mail clients, which interface with users and allow users to read, compose, send, and store email. The Mail servers and user workstations running mail clients. The mail clients are sequentially targeted by attackers. Additionally, mail clients have been targeted as an effective means of forwarding sensitive data content into machines and propagating to other machines. The attackers are able to develop attack methods to exploit security weaknesses. Hence, it is important to consider the security issues of mail servers and mail clients including web based access to mail.

4.2 Software Specifications:

The system has used Visual Studio.Net framework. And C#.Net has been used for developing the front end and SQL Server for the back end. The reason for using C#.Net is its flexibility. For the experiment, An Intel I3 2.2 GHz processor with 2 Gb RAM was used to measure the execution time and detection speed.

Table 4.2 software specification

Operating System	Windows 10
Front End	**ASP.Net , C#**
Back End	**MS SQL Server**

The table 5.1 shows the proposed software specification of the proposed application development. The main reason of using .Net framework is, it's a complete GUI and have many unique features to deploy a high featured web application.

4.3 Process involved in the proposed system:

- **Mail server creation:** mail server application is created with n number of users, and with n number of rules. The mail server involved with the mailing process like composing,
- **Rule specification:** the rule specification is performed by the admin and the clients of the email server. This includes forward rights specification, sensitive data tracking, and other access rights.
- **Pattern matching**: the pattern matching process calculates the similarities between the source data and the transferred email data's based on the rule set. This returns the similarity score of the documents in terms of percentage. Based on the given threshold the data leak is described as partial leak or full leak.
- **Data monitoring and movement tracking:** whenever, the data set to be track or monitor, then the following two steps will be performed. One is document key or id is allocated and that is embedded to monitor the activity of the data.
    - Key embedding
    - Key based data movement tracking
- **Data leak protection and alerting:** the next process after monitoring is providing appropriate reports to the users and alerting them about the guilty users with their probability ratio. The data security related alerts are generated to the data owners.
- **Reports:** provides all research related reports in table format as well as graph format.

5. RESULTS AND ANALYSIS

5.1 PERFORMANCE ANALYSIS:

The experiments are designed so that the different parts of the work could be evaluated. 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 was compared with the existing algorithms based on the following parameters.

- **Specificity** –measures the proportion of negatives that are correctly identified.
- **Sensitivity**- measures the proportion of positives that are correctly identified
- **Accuracy** – Determines the correctness

5.1.1 Specificity:

The sensitivity parameter measures the negatives that are correctly identified. In the given dataset d, the **specificity** is the ability of the test to correctly identify those without the leakage (true negative rate) from the given mail data tracking reports.

Specificity = TN/(TN + FP) = (Number of true negative assessment)/(Number of all negative assessment)

Data count	**Specificity**
10	0.5
20	0.60
30	0.40

Table 5.1 Specificity of the proposed system

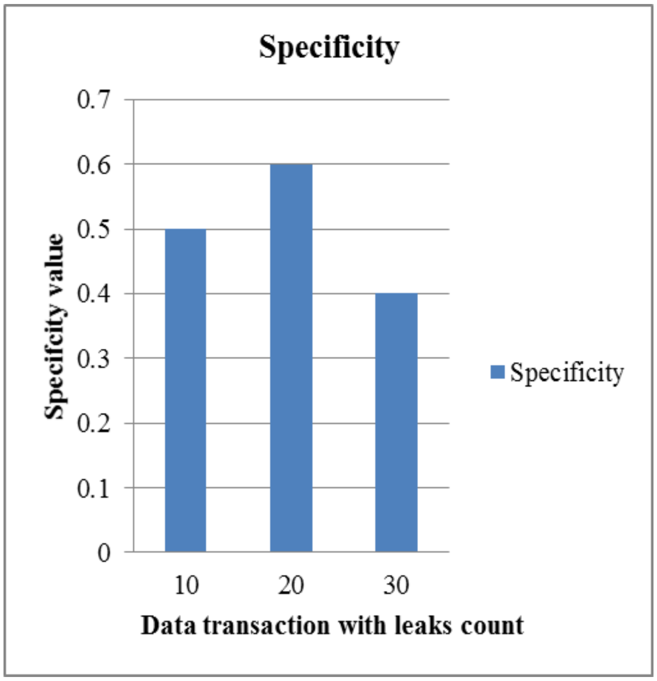


Fig 5.2 Specificity chart for the experiment

The fig 6.1 shows the specificity values for the proposed leak detection scheme for three different count experiments. This shows the specificity value is decreased when there is more number of leakers exists.

5.1.2 Sensitivity:

In health care analysis, sensitivity is the ability of a test to correctly identify those with the leakage (true positive rate),

Sensitivity = TP/(TP + FN) = (Number of true positive assessment)/(Number of all positive assessment)

Data count	**Sensitivity**
10	0.875
20	0.980
30	0.823

Table 5.3 Sensitivity of the proposed system

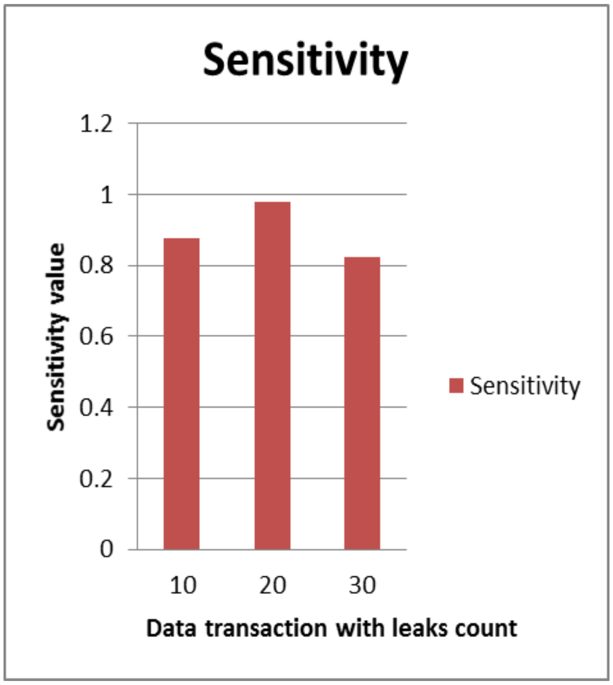


Fig 5.4 Sensitivity chart for the experiment

The fig 6.2 shows the Sensitivity values for the proposed leak detection scheme for three different count experiments. This shows the Sensitivity value is decreased when there is more number of leakers exists.

5.1.3 Accuracy:

The accuracy is the detection efficiency of data leakers. The accuracy is calculated using the following formula.

Accuracy = (TN + TP)/(TN+TP+FN+FP) = (Number of correct assessments)/Number of all assessments)

Data count	**accuracy**
10	0.80
20	0.92
30	0.83

Table 5.5 Accuracy of the proposed system

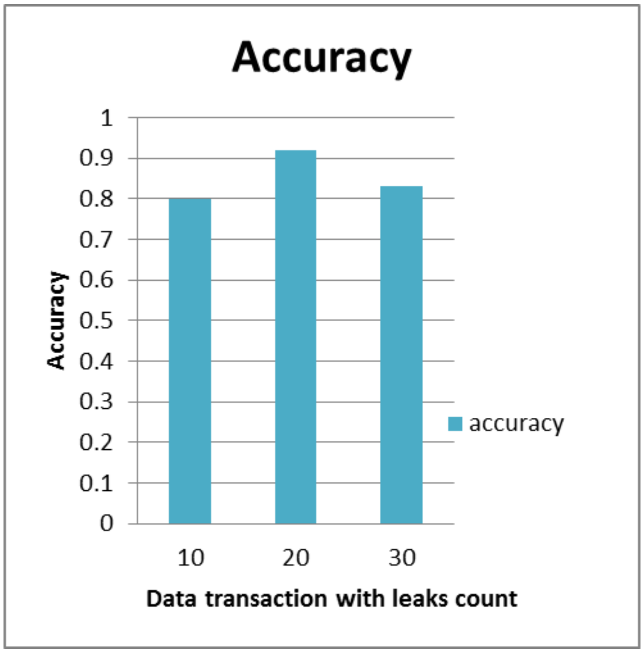


Fig 5.6 Accuracy chart for the experiment

The fig 6.3 shows the accuracy values for the proposed leak detection scheme for three different count experiments. This shows the accuracy value is different based on the number of transactions.

6. CONCLUSION

6.1 Conclusion

There is a huge need of protection against data leakage over sensitive data. Watermarking those sensitive data's for protection may create additional data stealing issues. And that is not a perfect way to prevent data form unknown access and transferred data leaks. The need of new prevention and detection technique may lead to the perfect data leaker detection.

In the research proposal the concept are added to detect data leakage and preventing sensitive data's among other users. This also includes an extended probability model named as DMA 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 been developed to prevent and identify the guilty users. To prevent this data leakage detection, the DLD and AHMM has been implemented in this concept.

REFERENCES

[1] Data loss db. Data loss statistics. Retrieved from ⟨http://datalossdb.org/⟩; 2015

[2] Mogull.Understandingandselectingadatalosspreventionsolution.Retrieved from ⟨https://securosis.com/assets/library/reports/ DLP - Whitepaper.pdf⟩; 2010.

[3] Boehmer, Wolfgang. "Analyzing Human Behavior Using Case-Based Reasoning with the Help of Forensic Questions." In Advanced Information Networking and Applications (AINA), 2010 24th IEEE International Conference on, pp. 1189-1194. IEEE, 2010.

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

[5] Yu, Fang, Zhifeng Chen, Yanlei Diao, T. V. Lakshman, and Randy H. Katz. "Fast and memory-efficient regular expression matching for deep packet inspection." In Architecture for Networking and Communications systems, 2006. ANCS 2006. ACM/IEEE Symposium on, pp. 93-102. IEEE, 2006.

[6] Hackl, Andreas, and Barbara Hauer. "State of the art in network-related extrusion prevention systems." Proceedings, 7th international symposuim on database engineering and applications 329-35, (2009).

[7] Alneyadi, Sultan, Elankayer Sithirasenan, and Vallipuram Muthukkumarasamy. "A survey on data leakage prevention systems." Journal of Network and Computer Applications 62, 137-152, (2016).

[8] Katz, Gilad, Yuval Elovici, and Bracha Shapira. "CoBAn: A context based model for data leakage prevention." Information sciences 262 137-158, (2014).

[9] Sudumbare, Pune. "Content Leakage Detection by Using Traffic Pattern for Trusted Content Delivery Networks." International Journal of Computer Science and Information Technologies, Vol. 5 (6), 7909-7913, 2014.

[10] Shu, Xiaokui, Jing Zhang, Danfeng Yao, and Wu-Chun Feng. "Rapid screening of transformed data leaks with efficient algorithms and parallel computing." In Proceedings of the 5th ACM Conference on Data and Application Security and Privacy, pp. 147-149. ACM, 2015.

[11] Alneyadi, Sultan, Elankayer Sithirasenan, and Vallipuram Muthukkumarasamy. "Detecting data semantic: a data leakage prevention approach." In Trustcom/BigDataSE/ISPA, 2015 IEEE, vol. 1, pp. 910-917. IEEE, 2015.

[12] Peneti, Subhashini, and B. Padmaja Rani. "Data leakage prevention system with time stamp." In Information Communication and Embedded Systems (ICICES), 2016 International Conference on, pp. 1-4. IEEE, 2016.

[13] 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. 3 528-542, (2016).