A HOLISTIC VIEW ON DATA PROTECTION FOR SHARING, COMMUNICATING, AND COMPUTING ENVIRONMENTS: TAXONOMY AND FUTURE DIRECTIONS

A PREPRINT

Ishu Gupta*
Cloud Computing Research Center
Department of Computer Science and Engineering
National Sun Yat-sen University
Kaohsiung, Taiwan
ishugupta23@gmail.com

Ashutosh Kumar Singh
Department of Computer Applications
National Institute of Technology
Kurukshetra, India
136119
ashutosh@nitkkr.ac.in

February 25, 2022

ABSTRACT

The data is an important asset of an organization and it is essential to keep this asset secure. It requires security in whatever state it is i.e. data at rest, data in use, and data in transit. There is a need to pay more attention to it when the third party is included i.e. when the data is stored in the cloud then it requires more security. Since confidential data can reside on a variety of computing devices (physical servers, virtual servers, databases, file servers, PCs, point-of-sale devices, flash drives, and mobile devices) and move through a variety of network access points (wireline, wireless, VPNs, etc.), there is a need of solutions or mechanism that can tackle the problem of data loss, data recovery and data leaks. In this context, the paper presents a holistic view of data protection for sharing and communicating environments for any type of organization. A taxonomy of data leakage protection systems and major challenges faced while protecting confidential data are discussed. Data protection solutions, Data Leakage Protection System’s analysis techniques, and, a thorough analysis of existing state-of-the-art contributions empowering machine learning-based approaches are entailed. Finally, the paper explores and concludes various critical emerging challenges and future research directions concerning data protection.

Keywords  Cloud Computing · Data Protection · Leakage Detection · Leakage Prevention · Information Security

1 Introduction

Data is recognized as the most vital asset of an enterprise because it defines each organization’s uniqueness [1]. In any organization, data is the main foundation of information, knowledge and ultimately the wisdom for correct decisions and actions [2,3]. It might be helping to cure a disease, boost a company’s revenue, make a building more efficient or be responsible for those targeted as we keep seeing [4]. If the data is relevant, complete, accurate, timely, consistent, meaningful, and usable, then it will surely help in the growth of the organization [5,6]. Companies that do not understand the importance of data management are less likely to survive in the modern economy. Therefore, it is essential to understand the importance of data management in companies [7,8]. Data should be carefully cultivated, managed, and refined into information that will allow an enterprise to better serve their community and ensure to remain viable in today’s competitive landscape [9,10].

In the real world scenario, a data distributor has to share the sensitive data of an entity among various stakeholders such as business partners, customers, and employees within or outside the organization’s premises for doing business [11] [12]. But the receiving entity may misuse this data and can leak it deliberately or inadvertently to some unauthorized third party [13,14]. Data leakage is the intentional or unintentional distribution of confidential or private
information or data to an unauthorized malicious entity [15, 16]. Critical data in various organizations include patient information, Intellectual Property (IP), personal information like credit card data, financial information and various other information depending upon the organization [17, 18].

Data leakage can cause a serious threat to the organization’s confidentiality as the number of incidents as well as the cost procured due to these leakages to continue to increase [19]. It poses a great challenge for any organization [20, 21]. It can destruct the enterprise’s reputation and goodwill, diminish shareholder’s value and decline the firm’s status and rank [22, 23]. Mostly precarious threats that any company faces, are because of insiders in the company, after all, as the insiders know the company’s internal system. To handle the attacks caused by insiders of the company is a most difficult task [24, 25].

It is essential to protect the confidential information as it increases the risk of falling the sensitive information in unauthorized hands and then it can be misused by unauthorized third party [26, 27]. Thus, it has become consequential for any organization to detect and prevent such leakage. Conversely, limiting the sharing of data in order to preserve the security and privacy of confidential information might reduce the organization’s performance. It influences the potential to perform the operation that can serve best to the organization and its customers [28]. Data leakage detection and prevention significantly play an inevitable role ineffective protection of the data. In this paper, concepts, challenges, existing solutions, research gaps, and future directions for sharing, communicating, and computing environments for any type of organization are discussed in detail.

Organization: Section 2 discusses the motivation for carrying out this study. Section 3 entails the challenges faced by the data protection system while protecting the data followed by the categorization of data protection solutions in section 4. Section 5 describes the concept of data leakage protection systems (DLPTSs) including data states, deployment schema, and DLPTSs analysis technique. An analysis and comparison of existing approaches and solutions along with their strengths and weaknesses are described comprehensively in section 6. Thereafter, the emerging challenges and future research direction in the field are entailed in section 7 followed by the conclusion and future scope of the work in section 8.

2 Motivation

In recent years, the internet and related technologies have grown rapidly [29, 30]. It offered the unrivaled capability to access and redistribute the digital data [31, 32]. It is easy to copy a huge amount of digital data at almost no cost that can be transferred in a very short time via the internet [33, 34]. The various organizations use this facility to enhance their capability by transferring data from one place to another but it involves a number of threats in transferring the confidential data as it can be made public by any malicious entity [35].

According to an exotic chronology of data breaches exposed by Privacy Rights Clearinghouse (PRC), 90, 61, 64, 759 records have been breached in the united state alone from 5, 278 data breaches made public since 2005 [36]. It is not hard to speculate that it is just a fingertip of an iceberg as most of the time, the data leakage cases are not recorded due to concern of restrictive penalties and loss of customer trust. In the year 2016, 4, 004, 004 records have been breached in the health care, medical providers, and medical insurance services organization alone and 300 breaches made public [37].

According to the annual cost of data breach study conducted by Ponemon Institute, the average consolidated cost of a data breach is $4 million. This study states that the cost procured for each stolen or lost record consisting of confidential and sensitive information increased from a consolidated average of $154 to $158 [38]. In highly regulated industries like healthcare, data breach cost is $355 per record which is $100 more than in 2013 [39]. Because of these reasons; the data leakage problem is increasing day by day.

A study reports that the number of leaked confidential data records has become 10 times within 4 years and it reached up to 1 : 1 billion in 2014 and it is kept on increasing as the number of users are increasing, also the malicious entity [40]. Fig. [41] presents the development of cyberattacks over time. It presents the recorded number of data breaches and records exposed in the United States between 2005 and 2020. In 2020, the number of data breaches in the United States amounted to 1001 with 155.8 million records exposed which is much higher compared to the past years.

The data leakage problem has reached a new dimension. There is a need for a general mechanism that can handle the data leakage problem by preventing the data leakage and by identifying the malicious entity responsible for data leakage. As the data is shared among multiple entities thus it becomes difficult to identify the malicious entity who has leaked the confidential information [42, 43]. In this paper, comprehensive solutions are explored that solve the data leakage problem by preventing the leakage and identifying the malicious entity who has leaked the organizational sensitive data.

2
3 Challenges in Data Protection

Like any other security mechanism, data protection systems also face many challenges when protecting confidential data. There are mainly seven challenges identified for data protection. To implement a successful data protection system, these challenges must be considered and addressed adequately [44, 45].

3.1 Leaking Channels

For Data-At-Rest and Data-In-Use, confidential data can be leaked through channels such as USB ports, CD drives, web services, and printed documents. Further, leaking channels associated with Data-In-Motion, such as web services and file sharing, might be extremely challenging, since these channels may be business prerequisites. To ensure maximum security for data passing through these channels, extensive traffic filtering must be carried out.

3.2 Human Factor

It is always difficult to predict human behavior because it is influenced by many psychological and social factors. Many human actions are affected by subjectivity in making decisions, such as defining the secrecy level of data, assigning access rights to specific users.

3.3 Access Right

It is important for DLPSs to be able to distinguish between different users based on their privileges and permissions. Without a proper definition of access rights, DLPSs cannot decide whether or not the data is being accessed by a legitimate user, and then it can be misused by a malicious entity.

3.4 Encryption and Steganography

Network-based DLPSs and DLDSs attempt to identify copies of confidential data using various analysis techniques, and then compare them to the original data. However, the use of strong encryption algorithms makes it very difficult or nearly impossible to analyze the data content. Furthermore, the steganography technique can be used to maliciously leak confidential data since it is highly likely to bypass traffic inspection mechanisms.
3.5 Data Modification

Some DLPSs and DLDSs use data hashing to check outgoing traffic and to monitor the outgoing data respectively. The problem with hashing is that any modification to the original document can lead to a totally different hash value, resulting in disclosure. Furthermore, some DLDSs use watermarking and steganography to hide the secret data/information that can be modified by some malicious entity.

3.6 Scalability and Integration

Like many other security mechanisms, data protection systems too can be affected by the amount of data being processed. DLPSs tend to have a poor integration/association with other security mechanisms in a network. This is because some of their mandatory features already exist in other solutions such as firewalls, IDSs, and proxy servers which may arise inconsistencies.

3.7 Data Classification

If the data is not classified into different levels, data protection systems will not be able to distinguish between confidential and normal data/traffic. Hence, without proper data classification, confidential data can easily be revealed even in the presence of data protection systems.

4 Data Protection Solutions

Data leakage prevention and data leakage detection solution are provided to protect the data as shown in Fig. 2.

Data protection (Leakage Handling Approach)
Data Leakage Prevention Systems (DLPS)
Data Leakage Detection System (DLDS)

Figure 2: Solution for data protection.

4.1 Data Leakage Prevention

Data leakage/loss prevention (DLP) is typically defined as any solution or process that identifies confidential data, tracks that data as it moves through and out of the enterprise, and prevents unauthorized disclosure of data by creating and enforcing disclosure policies [46, 47].

DLPSs are defined as designated analytical systems used to protect data from unauthorized disclosure at all states using remedial actions triggered by a set of rules. This definition contains three main attributes that distinguish DLPSs from conventional security measures. First, DLPSs have the ability to analyze the content of confidential data and the surrounding context. Second, DLPS scans be deployed to provide protection of confidential data in different states, that is, in transit, in use, and at rest. The third attribute is the ability to protect data through various remedial actions, such as notifying, auditing, blocking, encrypting, and quarantining. The protection normally starts with the ability to detect potential leaks through heuristics, rules, patterns, and fingerprints. The prevention then happens accordingly [48, 49].
4.2 Data Leakage Detection

Data leakage detection (DLD) is typically defined as any solution or process that identifies the unauthorized disclosure of confidential data. Data leakage detection systems (DLDSs) are specially designed systems that have the ability to monitor and protect confidential data, detect misuse of confidential data and identify the malicious entity responsible for data leakage [50, 51].

5 Data Leakage Protection System

A taxonomy of Data Leakage Protection System (DLPTS) is represented in Fig. 3 that represents, what data to be secured, deployment schemes, solution for data protection and the remedial action to be taken.

Figure 3: A taxonomy of Data leakage protection system (DLPTS).

5.1 Data State

Data states include “Data-At-Rest (DAR)”, “Data-In-Use (DIU)” and “Data-In-Motion (DIM)”.

Data-At-Rest is the type of data that is stored in repositories either locally or at remote locations. It consists of application databases, backup files, and file systems. It is normally protected by strong access controls, including physical and logical mechanisms.

Data-In-Use is the data that is accessible to the user in the forms of documents, emails, and applications. This type of data appears in plain text, so it can be easily interpreted and processed.

Data-In-Motion is the data being...
transmitted from one node to another. This type of data travels internally between nodes within the same network or externally between nodes that belong to different networks [52].

5.2 Deployment Schema

DLPS schemes are deployed at the endpoints to protect the data at rest and data in use and in the network to protect the data in motion/transit. Depending on the targeted data for protection, DLPS deployment can take many forms. For example, DLPSs that deal with Data-At-Rest are normally focused on protecting known data. The protection comes in the forms of preventing access to data based on predefined security policies. Also, this type of DLPS helps in protecting data at rest by encrypting entire file systems. Furthermore, it scans, identifies, and secures confidential data in data repositories while auditing and reporting security policy violations [53].

Protecting Data-In-Use requires built-in software that acts like a DLP agent on endpoints. This agent is responsible for disabling or enabling access to applications that deal with confidential data. It is also responsible for blocking confidential data transfer through portable media, that is, CDs, USB drives, and memory cards. Furthermore, it restricts copying, pasting, and editing of confidential data as well as restricting making hard copies through printers. It audits all activities related to confidential data access [54].

For Data-In-Motion, DLP appliances are normally used. They come with special processing capabilities to handle large amounts of data. This type of DLPS is responsible for inspecting all outbound traffic for confidential data. It also acts like a proxy server when accessing some applications with confidential data. Moreover, it proactively reports and alerts security administrators and users about potential data leaks. Finally, it coordinates with other security mechanisms such as Secure Sockets Layer (SSL) proxies and network firewalls [55].

5.3 DLPTSs Analysis Technique

Context-based DLPTSs focus on the metadata (context), such as size, timing, source, and destination surrounding confidential data, rather than the sensitivity of the content to detect any potential leaks. Content-based analysis DLPTSs are more common than and preferable to those that are context-based since it is more logical to focus on the protection of the data itself than on the surrounding context [55].

A typical content-based DLPTS works by monitoring sensitive data in its repository or on the go, mainly by using regular expressions, data fingerprinting, and statistical analysis. Regular expressions are normally used under a certain rule such as detecting social security numbers and credit card numbers. The problem with DLPTSs using regular expressions analysis is that they offer limited data protection and have high false-positive rates [55].

DLPTSs using data fingerprints have better coverage for sensitive data because they have the ability to detect and prevent the leakage of a whole document or parts of a document. However, traditional fingerprinting can lose track when the sensitive data is altered or modified [56, 57].

Although not widely used in DLPTSs, statistical analysis can work in a fuzzy environment where the sensitive data is not well structured. The main advantage of such a technique is the ability to identify sensitive documents even after extreme modification. In particular, DLPTSs with statistical analysis capabilities can use machine learning algorithms or Bayesian probability to identify altered documents. They can also use text-clustering techniques to construct scattered traces of sensitive data. N-gram analysis and term weighting analysis are the main statistical analysis techniques. N-gram statistical analysis is the way to analyze data based on the frequency of data of interest. It is a method based on machine learning that classifies enterprise documents as sensitive or not [54]. Term weighting is a statistical method that indicates the importance of a word within a document. It is normally used in text classification. A third technique called content tagging is used in some DLPTSs. This technique is used to tag the file containing confidential data. It can preserve the identity of the file but not the contained confidential data.

6 Current Approaches of DLPT

On the basis of the review, there are mainly seven approaches for data protection as represented in Fig. 4 that are summarized in Table 1 with their strengths and weaknesses.

The models that handle with data leakage problem by embedding the code in the document are presented in [58, 59, 60]. Backes et al. provide a generic data lineage framework called LIME for data flow across multiple entities having two principal roles i.e. owner and consumer. The guilty entity is identified by developing a data transfer protocol between two entities via considering an oblivious transfer, watermarking, and signature primitives. This method considers the possible data leakage and the corresponding constraints at the design stage. A model that identifies the
| Approach                      | Strengths                                                                                                         | Weaknesses                                                                                                                |
|-------------------------------|-------------------------------------------------------------------------------------------------------------------|---------------------------------------------------------------------------------------------------------------------------|
| Policy and access rights      | • suitable for any organization if access rights and data classification are properly established  
                              | • easy to manage  
                              | • suitable for Data-At-Rest and Data-In-Use  
                              | • strong prevention mechanism | • affected by improper data classification  
                              | | • affected by the access control policy in use  
                              | | • not a detective method, hence if a leak is happening the method is ineffective |
| Virtualization and isolation  | • requires small hardware implementation  
                              | • dynamic as it does not need regular administrative interference  
                              | • accessing sensitive data can use existing data classification | • not mature enough  
                              | | • produces considerable amount of overheads  
                              | | • not a detection method |
| Cryptographic approaches      | • strong cryptography can produce maximum security  
                              | • cryptographic methods are wieldy to use and have many options | • cryptography can secure sensitive data but may not deny its existence  
                              | | • does not detect data leak  
                              | | • confidential data can be accessed by weak credentials |
| Quantifying and limiting      | • goes beyond studying sensitive data, and focus on the leaking channels  
                              | • useful against specific types of attacks such as salami attacks  
                              | • effective for all data states | • does not ensure total blockage of the leaking channel  
                              | | • limited to specific situations or scenarios  
                              | | • can disrupt workflow |
| Social and behavior analysis  | • proactive data leakage prevention by detecting malicious relations  
                              | • suitable for all data states | • produces a high level of false-positive requires regular administrative interference  
                              | | • requires a huge amount of profiling and indexing |
| Data identification           | • very strong in detecting unmodified data  
                              | • very low false positive level for analysis using fingerprints  
                              | • some robust hashing can detect modified data | • extremely modified data cannot be detected  
                              | | • lacking semantic understanding |
| Data mining and text clustering | • can predict future data leaks powerful in detecting unstructured data  
                              | • less dependent on administrative help flexible and adaptable | • requires a great deal of processing requires learning phase, which means many false positives |
The guilty party who is responsible for leaking critical organizational data is given by Kumar et al. This method is able to find the exact client who leaked the data but it is unable to detect the guilty client if embedded code is modified or completely destroyed by the client.

The model introduced in [61, 62] helps to identify the guilty agent who leaks the data provided by the distributor. In [61] the model assesses the guilt of various agents. It consists of various algorithms for the distribution of data among multiple agents which improve the chances of identifying a leaker and then the fake objects are added to the distributed set. Data is allocated to various agents in such a manner that there will be minimum overlapping of data among various agents. Later on, if any agent leaked his confidential data at some unauthorized place then the guilty party can be found on the basis of the data allocated to the agents. Kumar et al. introduce the allocation strategies that work on the basis of No-wait model and increase the chances of identifying the guilty party. These methods have the benefit that the agent can not destroy the information after leaking the data as in the case of watermarking but it cannot identify the exact data leaker in the case when the same data is allocated to multiple agents or overlapping of data among agents increases.

The method introduced in [63] enables the data owner to depute the detection operation safely to a semi-host provider without exposing the critical data to the provider. This method is implemented using the fuzzy fingerprint technique to elevate data privacy at the time of data leak detection operations. The method provides the privacy-preserving Data Leakage Detection (DLD) solution to handle data leaks in situations where a set of sensitive data digests is utilized for detection.

To protect sensitive information from unauthorized parties, Bishop et al. present an information leakage detection (ILD) agent system in [64]. It is a mobile agent-based approach that brutalizes the process of discovering and coloring perceptive hosts file systems and observing the colored file system for detection of potential information leakage. A concept presented is based on assigning a sensitivity score to data sets for mitigating the data misuse and data leakage incidents in database system [65].

Other solution for data leaks includes sequence alignment techniques [66, 67, 68, 69], classification technique [70], network-based data leak detection solution [71, 72, 73]. In [67] solution for the detection of transformed data leaks is presented by detecting complex data leak patterns using sequence alignment techniques. A new class of health
information security intrusions that exploits physiological information leakage to compromise privacy is presented in [74]. A map-reduce framework is introduced by Liu et al. for the detection of exposed sensitive contents in [63]. A network-wide method based on black-box differencing for restraining and controlling the flow of confidential data within a network is presented in [75,76], which determines when the secret data is being leaked. [77,78,79] present the way for data leakage prevention. In [62], it is specified that solutions for data leakage preventions are mainly deployed to prevent unintentional leakage incidents. Table 2 summarizes the most relevant work of the field.

Table 2: Summary of most relevant work in the review.

| Paper and Category | Method | Analysis | Suitable for | Contribution | Limitation |
|--------------------|--------|----------|--------------|--------------|------------|
| Wuchner et al. [80], 2012 (Policy and access rights) | Detective/Preventive | Context | In use | UC4Win, a data loss prevention solution for Microsoft Windows operating systems. | Requires predefined policy. Cannot identify sensitive data. |
| Squicciarini et al. [81], 2010 (Policy and access rights) | Preventive | Context | In use | Introduces a three layers data protection framework. | Requires a pre-defined classification for data. Mis-classified sensitive data can be leaked. |
| Griffin et al. [82], 2005 (Virtualization and isolation) | Preventive | Context | In use | Proposes Virtual Trusted Domains (VTD) to offload processes to secure environments. | Imposes challenge to computational capabilities. |
| Wu et al. [53], 2011 (Virtualization and isolation/ Cryptographic approaches) | Preventive | Context | At rest/ In use | Introduces a combination of encrypted storage and virtual environment to prevent data leakage. | Suitable for data at rest only. Cannot prevent data leakage caused by privileged used. |
| Blanke [83], 2011 (Cryptography approaches) | Preventive | Context | At rest | Uses ephemeral encryption to protect data whenever accessed by a user. | Requires a pre-implementation of an encrypted file system. |
| Yoshihama et al. [84], 2010 (Quantifying and limiting) | Detective | Content/Context | In motion | Uses an application-level proxy to detect potential data leakage risks. | Cannot detect data leaked through covert channels. |
| Borders et al. [85], 2009 (Quantifying and limiting) | Detective | Content | In motion | Constrains the maximum volume of sensitive data in web traffic. | Unable to filter parts of URLs that contains random numbers to prevent cashing. |
| Suen et al. [86], 2013 (Quantifying and limiting/Social and behavior analysis) | Detective | Context | In motion | Uses S2 Logger to track files while traveling in the cloud. | Based on content tagging. Cannot track sensitive content. |
| Boehmer [87], 2010 (Social and behavior analysis) | Detective/Preventive | Context | In use | Uses case-based reasoning (CBR) in combination with directed acyclic graph (DAG) and Hamming similarity function. | Needs existing or synthetic compliance profiles for comparison process. |
| Authors | Type | Context | Phase | Content | Description |
|---------|------|---------|-------|---------|-------------|
| Shapira et al. [88], 2013 (Data identification) | Detective | Content | In use | Robust fingerprinting to overcome shortcoming in ordinary hashing. | Requires extensive data indexing for both sensitive and normal data. |
| Shu et al. [71], 2013 (Data identification) | Detective | Content | In motion | Uses message shingles/fuzzy fingerprints to detect in advertent data leak in network traffic. | Modified data can cause false negatives because the shingles fingerprints are different from the original ones. |
| Hart et al. [54], 2011 (Data mining and Text clustering) | Detective | Content | In use | Uses SVM Machine learning to classify documents to private and public. | Inflexible, Limited to two categories. |
| Lindell et al. [89], 2000 (Data mining and Text clustering) | Preventive | Content | At rest/In use | Sharing confidential data on the union of the entities databases, without releasing unnecessary information. | Theoretically proven but lacking practical experiments. |
| Hart et al. [54], 2011 (Data mining and Text clustering) | Detective | Context | In motion | Uses Partially Observed Markov Decision Processes (POMDPs) over decision epochs. | POMDP requires huge amount of calculations. |
| Gomez-Hidalgo et al. [91], 2010 (Data mining and Text clustering) | Detective/Preventive | Content | In motion | NER (named entity recognition) approach is used to identify and extract words from texts. | Named entity recognition could be affected by spelling mistakes and connected words. |
| Sokolova et al. [92], 2009 (Data mining and Text clustering) | Detective | Content | In motion | Uses support vector machine to classify enterprise documents as sensitive non-sensitive. | Not fixable because it classifies data to public or private only. |
| Parekh et al. [93], 2006 (Data mining and Text clustering) | Detective | Content | In motion | A new approach to enable the sharing information of suspicious payloads. | Polymorphic/obfuscated worms and mimicry attacks may create a big challenge. |
| Carvalho et al. [94], 2009 (Data mining and Text clustering) | Detective | Content/Context | In motion | Presents an extension –Cut Once- to "Mozilla Thunderbird". | Introduces high level of false positives, since it requires existing messages in the sent folder. |
| Zilberman et al. [95], 2010 (Data mining and Text clustering/Social and behavior analysis) | Detective/Preventive | Content/Context | In motion | Uses TF-IDF and cosine similarity to compute existing links between users. | High false positive rates because of short history between senders and recipients. |
| Carvalho and Cohen [96], 2007 (Data mining & Text clustering/Social & behavior analysis) | Detective | Content/Context | In motion | Predicts unintended message recipients. | Can create false outliers because of limited interaction history. |
7 Emerging Challenges and Future Research Directions

Data protection mechanism should make sure that the data is safe enough from all the internal as well as external threats so that there should not be any problem such as loss of data or data theft. Many methods and solutions have been proposed to protect the data from leakage/loss but still there exist some research gaps in the provided solutions. We need an ideal solution that can secure, manage confidential data and locate the malicious agent and activities. The following research gaps are identified based on the literature review:

1. **Increased Overhead:** When the data is stored at a third party or somewhere else in the database or repositories, the security mechanism such as encryption scheme, etc. are applied to the whole data which reduces the utility of data and increase the overhead of the service provider. Furthermore some of the approaches require a huge amount of calculations resulting in delay and increased overhead.

2. **Static Request handling:** Solution has been provided for the detection of a malicious agent by providing a guilty agent model.

3. **Single Objective Approaches:** The approaches provide the solution either for detection of guilty agent or prevention of loss of data. These approaches are suitable for data-at-rest or data-in-use or data-in-motion. Some approaches provide prevention solutions but cannot identify data as sensitive. For the better protection of data, there is a need for a hybrid approach that fulfills all the objectives.

4. **Involves Data modification:** The guilty agent is identified by hiding the information in the requested document which may involve modification of data. Furthermore, DLP uses techniques such as regular expression, fingerprinting, data mining/text clustering, etc. to identify the sensitive data, but if the information is too much modified then these techniques cannot extract the sensitive data.

5. **High false-positive rate:** The approaches used for DLP require a learning phase that may involve a high false-positive rate.

The following are the research directions in order to fill the identified research challenges:

1. To make a balance between data utility and security and reducing overheads by proper classification.
2. To formulate a Data leakage Detection System that handles the user request in a dynamic manner.
3. Efforts will be made in developing the system with high accuracy.
4. Development of a new multi-objective approach that provides both Data Leakage Detection and Prevention

8 Conclusions and Future Scope

Prevention of data disclosure to unauthorized entities is one of the main goals of information security. According to datalossdb (2015) report, in the year 2014, about 50% of recorded data leakage occurred in the business sector, about 20% occurred in the government sector and about 30% occurred in the health and education sectors. Although some reported leaks were not detrimental to organizations, others have caused several million dollars worth of damage. Business credibility is compromised when sensitive data is leaked to competitors. The developments to be carried out in research work will be helpful in business, governmental, private, etc. sectors with the following advantages: (i) Provide Data protection. (ii) Maintains Privacy and Security of data. (iii) Mitigates the risk of data leakage. (iv) Solves the controversy raised due to information leakage.

References

[1] I. Gupta and A. K. Singh, “An Integrated Approach for Data Leaker Detection in Cloud Environment,” *Journal of Information Science and Engineering*, vol. 36, pp. 993–1005, Sep. 2020.

[2] A. K. Singh and I. Gupta, “Online Information Leaker Identification Scheme for Secure Data Sharing,” *Multimedia Tools and Applications*, vol. 79, no. 41, pp. 31 165–31 182, November 2020.

[3] Khushbu, P. Nishad, V. Kashyap, and I. Gupta, “A Classification and Distribution Model for Data Leakage Prevention and Detection,” *International Research Journal of Modernization in Engineering Technology and Science*, vol. 3, no. 2, pp. 348–354, Feb. 2021.

[4] I. Gupta and A. K. Singh, “GUIM-SMD: Guilty User Identification Model using Summation Matrix-based Distribution,” *IET Information Security*, vol. 14, pp. 773–782, November 2020.
[5] G. Batra, H. Singh, I. Gupta, and A. K. Singh, “Best Fit Sharing and Power Aware (BFSPA) Algorithm for VM Placement in Cloud Environment,” in *2017 3rd International Conference on Advances in Computing, Communication & Automation (ICACCA) (Fall)*. IEEE, 2017, pp. 1–4.

[6] I. Gupta and A. K. Singh, “Dynamic Threshold based Information Leaker Identification Scheme,” *Information Processing Letters*, vol. 147, pp. 69 – 73, 2019.

[7] RingLead, “The Importance of Data Management In Companies,” 2021. [Online]. Available: https://www.ringlead.com/blog/the-importance-of-data-management-in-companies/#.WWLh09Tyu00

[8] I. Gupta, T. K. Madan, S. Singh, and A. K. Singh, “Stock Market Forecasting using Historical and Sentiment Analysis,” *CoRR*, 2022.

[9] Debbie King, “Association Analytics, Data is an Asset,” February 2013. [Online]. Available: https://associationanalytics.com/blog/data-is-an-asset/

[10] I. Gupta, V. Sharma, S. Kaur, and A. K. Singh, “An Efficient Parkinson’s Disease Prediction Model based on Random Forest Classification,” *CoRR*, 2022.

[11] I. Gupta, R. Gupta, A. K. Singh, and R. Buyya, “MLPAM: A Machine Learning and Probabilistic Analysis Based Model for Preserving Security and Privacy in Cloud Environment,” *IEEE Systems Journal*, vol. 15, no. 3, pp. 4248–4259, 2021.

[12] A. Acharya, H. Prasad, V. Kumar, I. Gupta, and A. K. Singh, “Host Platform Security and Mobile Agent Classification: A Systematic Study,” in *Computer Networks and Inventive Communication Technologies*, vol. 58. Singapore: Springer Singapore, 2021, pp. 1001–1010, data Engineering and Communications Technologies.

[13] I. Gupta and A. K. Singh, “A Framework for Malicious Agent Detection in Cloud Computing Environment,” *International Journal of Advanced Science and Technology (IJA*, vol. 135, pp. 49–62, Feb 2020.

[14] P. Godha, S. Jadon, A. Patle, I. Gupta, B. Sharma, and A. Kumar Singh, “Architecture, an Efficient Routing, Applications, and Challenges in Delay Tolerant Network,” in *2019 International Conference on Intelligent Computing and Control Systems (ICCS)*. IEEE, 2019, pp. 824–829.

[15] P. Godha, S. Jadon, A. Patle, I. Gupta, B. Sharma, and A. K. Singh, “Flooding and Forwarding Based on Efficient Routing Protocol,” in *International Conference on Innovative Computing and Communications*, vol. 1166. Singapore: Springer Singapore, 2021, pp. 215–223, advances in Intelligent Systems and Computing.

[16] I. Gupta, S. Mittal, A. Tiwari, P. Agarwal, and A. K. Singh, “TIDF-DLPM: Term and Inverse Document Frequency based Data Leakage Prevention Model,” *CoRR*, 2022.

[17] D. Saxena, I. Gupta, J. Kumar, A. K. Singh, and X. Wen, “A Secure and Multiobjective Virtual Machine Placement Framework for Cloud Data Center,” *IEEE Systems Journal*, pp. 1–12, 2021.

[18] I. Gupta and A. K. Singh, “SELI: Statistical Evaluation based Leaker Identification Stochastic Scheme for Secure Data Sharing,” *journal = IET Communications*, vol. 14, pp. 3607–3618, December 2020.

[19] P. Tiwari, S. Mehta, N. Sakhuja, I. Gupta, and A. K. Singh, “Hybrid Method in Identifying the Fraud Detection in the Credit Card,” in *Evolutionary Computing and Mobile Sustainable Networks*, vol. 53. Singapore: Springer Singapore, 2021, pp. 27–35, data Engineering and Communications Technologies.

[20] Khushbu, P. Nishad, V. Kashyap, I. Gupta, and A. K. Singh, “An Organized Study on Data Divulge Elimination and Discernment,” in *Computer Networks and Inventive Communication Technologies*, vol. 58. Singapore: Springer Singapore, 2021, pp. 569–578, data Engineering and Communications Technologies.

[21] U. Arora, S. Verma, I. Gupta, and A. K. Singh, “Implementing Privacy using Modified Tree and Map Technique,” in *2017 3rd International Conference on Advances in Computing, Communication & Automation (ICACCA) (Fall)*. IEEE, 2017, pp. 1–5.

[22] V. Sharma, S. Jalwa, A. R. Siddiqi, I. Gupta, and A. K. Singh, “A Lightweight Effective Randomized Caesar Cipher Algorithm for Security of Data,” in *Evolutionary Computing and Mobile Sustainable Networks*, vol. 53. Singapore: Springer Singapore, 2021, pp. 411–419, data Engineering and Communications Technologies.

[23] P. Agarwal, S. Mittal, A. Tiwari, I. Gupta, A. K. Singh, and B. Sharma, “Authenticating Cryptography over Network in Data,” in *2019 International Conference on Intelligent Computing and Control Systems (ICCS)*. IEEE, 2019, pp. 632–636.

[24] A. Kesharwani, A. Nag, A. Tiwari, I. Gupta, B. Sharma, and A. K. Singh, “Real-Time Human Locator and Advance Home Security Appliances,” in *Evolutionary Computing and Mobile Sustainable Networks*, vol. 53. Singapore: Springer Singapore, 2021, pp. 37–49, data Engineering and Communications Technologies.
[25] S. Jalwa, V. Sharma, A. R. Siddiqi, I. Gupta, and A. K. Singh, “Comprehensive and Comparative Analysis of Different Files Using CP-ABE,” in Advances in Communication and Computational Technology, vol. 668. Singapore: Springer Singapore, 2021, pp. 189–198, electrical Engineering.

[26] I. Gupta and A. K. Singh, “A Probabilistic Approach for Guilty Agent Detection using Bigraph after Distribution of Sample Data,” Procedia Computer Science, vol. 125, pp. 662 – 668, 2018.

[27] K. Kaur, I. Gupta, and A. K. Singh, “Data Leakage Prevention: E-Mail Protection via Gateway,” Journal of Physics: Conference Series, vol. 933, p. 012013, jan 2018.

[28] K. Gupta and I. Gupta, “A Comprehensive Study on Architecture, Security issues and Challenges in Cloud Computing,” International Journal of Scientific & Engineering Research, vol. 7, no. 12, pp. 118–120, Dec. 2016.

[29] A. Nag, A. Kesharwani, B. Sharma, I. Gupta, A. Tiwari, and A. K. Singh, “Potential and extension of internet of things,” in Second International Conference on Computer Networks and Communication Technologies (ICCNCT), vol. 44. Cham: Springer International Publishing, 2020, pp. 542–551, data Engineering and Communications Technologies.

[30] I. Gupta and A. K. Singh, “A Hybrid Technique for the Detection of Data Leakage in Cloud computing Environment,” in Ist International Conference on Science in Hindi, August 2017, vigyan Prakash.

[31] I. Gupta and K. Gupta, “Review on Intrusion Detection System Architectures in WSN,” International Journal of Scientific & Engineering Research, vol. 7, no. 12, pp. 111–115, Dec. 2016.

[32] I. Gupta, H. Mittal, D. Rikhari, and A. K. Singh, “A Multiple Linear Regression based Model for Average Temperature Prediction of A Day,” CoRR, 2022.

[33] I. Gupta, N. Singh, and A. Singh, “Layer-based Privacy and Security Architecture for Cloud Data Sharing,” Journal of Communications Software and Systems (JCOMSS), vol. 15, no. 2, 2019.

[34] K. Kaur, I. Gupta, and A. K. Singh, “A Comparative Study of the Approach Provided for Preventing the Data Leakage,” International Journal of Network Security & Its Applications, vol. 9, no. 5, pp. 21–33, 2017.

[35] I. Gupta, P. K. Yadav, S. Pareek, S. Shakeel, and A. K. Singh, “Auxiliary Informatics System: An Advancement towards a Smart Home Environment,” CoRR, 2022.

[36] “Privacy Rights Clearinghouse (PRC),” 2016. [Online]. Available: https://www.privacyrights.org/data-breaches

[37] “Chronology of Data Breaches,” 2016. [Online]. Available: http://educationnewyork.com/files/Chronology%20of%20Data%20Breaches%202005%20to%20Today.pdf

[38] “Security Intelligence: 2016 Ponemon Institute Cost of a Data Breach Study,” 2016. [Online]. Available: https://securityintelligence.com/media/2016-cost-data-breach-study/

[39] “CNBC: Data Breach Cost,” 2016. [Online]. Available: http://www.cnbcreports.com/2016/06/14/cost-of-data-breaches-hits-4-million-on-average-ibm.html

[40] R. B. S. (RBS), “2014 Data Breaches –A Billion Exposed Records –A New All Time High,” 2016. [Online]. Available: https://www.riskbasedsecurity.com/2015/02/2014-data-breaches-a-billion-exposed-records-a-new-all-time-high/

[41] Identity Theft Resource Center ©Statista 2021, “United States: 2005 to 2020 Sensitive Records Exposed,” 2021. [Online]. Available: https://www.statista.com/statistics/273550/data-breaches-records-in-the-united-states-by-number-of-breaches-and-records-exposed/

[42] I. Gupta and A. K. Singh, “Guilty Agent Detection through Probabilistic Analysis,” CoRR, 2022.

[43] V. Gautam, C. P. Yadav, R. Verma, I. Gupta, and A. K. Singh, “A Comparative Study on Data Leakage Detection and Prevention,” in IEEE International Conference on NanoElectronics, Nanophotonics, Nanomaterials, Nanobiotechnology & Nanotechnology (5NANO2021). IEEE, 2021.

[44] I. Gupta and A. K. Singh, “A Probability based Model for Data Leakage Detection using Bigraph,” in Proceedings of 7th International Conference on Communication and Network Security (ICCN), ser. ICCNS 2017. New York, NY, USA: Association for Computing Machinery (ACM), 2017, p. 1–5.

[45] I. Gupta, “Malicious Entity Detection for Protecting the Data using Watermarking and Probabilistic Approaches,” CoRR, 2022.

[46] I. Gupta and A. K. Singh, “A Confidentiality Preserving Data Leaker Detection Model for Secure Sharing of Cloud Data using Integrated Techniques,” in 2019 7th International Conference on Smart Computing Communications (ICSCC). Curtin University, Sarawak Malaysia: IEEE, 2019, pp. 1–5.
[47] R. Verma, V. Gautam, C. P. Yadav, I. Gupta, and A. K. Singh, “A Survey on Data Leakage Detection and Prevention,” in Proceedings of the International Conference on Innovative Computing & Communications (ICICC) 2020. SSRN, Elsevier, May 2020, pp. 1–7.

[48] A. Acharya, H. Prasad, V. Kumar, I. Gupta, and A. K. Singh, “MACI: Malicious API Call Identifier model to secure the Host Platform,” in Proceedings of the Seventh International Conference on Mathematics and Computing (ICMC-2021), ser. ICMC 2021, vol. 711. Singapore: Springer Singapore, 2021, advances in Intelligent Systems and Computing.

[49] I. Gupta, “A Comparative Study of the Approach Provided for Preventing the Data Leakage,” Other Topics Engineering Research eJournal, vol. 9, no. 5, September 2017. [Online]. Available: https://ssrn.com/abstract=3632686

[50] I. Gupta and A. K. Singh, “Data Leakage Detection in Cloud Computing Environment,” CoRR, 2022.

[51] A. K. Singh, I. Gupta, R. Verma, V. Gautam, and C. P. Yadav, “A Survey on Data Leakage Detection and Prevention,” in Proc. Int. Conf. Innov. Comput. Commun., 2020.

[52] I. Gupta and A. K. Singh, “Solutions for Organizational Data Security and Privacy in Real Environment,” CoRR, 2022.

[53] J. Wu, J. Zhou, J. Ma, S. Mei, and J. Ren, “An Active Data Leakage Prevention Model for Insider Threat,” in 2011 2nd International Symposium on Intelligence Information Processing and Trusted Computing, 2011, pp. 39–42.

[54] M. Hart, P. Manadhata, and R. Johnson, “Text Classification for Data Loss Prevention,” vol. 6794, 07 2011, pp. 18–37.

[55] R. Mogull, “Understanding and selecting a data loss prevention solution,” 2021. [Online]. Available: https://securosis.com/assets/library/publications/DLP-Whitepaper.pdf

[56] K. Kaur, I. Gupta, and A. K. Singh, “A Comparative Evaluation of Data Leakage/Loss Prevention Systems (DLPS),” in Proc. 4th International Conference Computer Science & Information Technology (CS & IT-CSCP), 2017, pp. 87–95.

[57] I. Gupta and K. Gupta, “Evaluation of Intrusion Detection Schemes in Wireless Sensor Network,” IOSR Journal of Computer Engineering, vol. 18, no. 2, pp. 60–63, Mar-Apr. 2016.

[58] N. Kumar, V. Katta, H. Mishra, and H. Garg, “Detection of Data Leakage in Cloud Computing Environment,” in 2014 International Conference on Computational Intelligence and Communication Networks, Nov 2014, pp. 803–807.

[59] M. Backes, N. Grimm, and A. Kate, “Data Lineage in Malicious Environments,” IEEE Transactions on Dependable and Secure Computing, vol. 13, no. 2, pp. 178–191, March 2016.

[60] K. N. Kaur, Diyaa, I. Gupta, and A. K. Singh, “Digital Image Watermarking Using (2, 2) Visual Cryptography with DWT-SVD Based Watermarking,” in Computational Intelligence in Data Mining, vol. 711. Singapore: Springer Singapore, 2019, pp. 77–86, advances in Intelligent Systems and Computing.

[61] P. Papadimitriou and H. Garcia-Molina, “Data Leakage Detection,” IEEE Transactions on Knowledge and Data Engineering, vol. 23, no. 1, pp. 51–63, Jan 2011.

[62] A. Shabtai, Y. Elovici, and L. Rokach, A Survey of Data Leakage Detection and Prevention Solutions. Springer Science & Business Media, 01 2012.

[63] F. Liu, X. Shu, D. Yao, and A. R. Butt, “Privacy-preserving scanning of big content for sensitive data exposure with mapreduce,” in Proceedings of the 5th ACM Conference on Data and Application Security and Privacy, ser. CODASPY ’15. New York, NY, USA: Association for Computing Machinery, 2015, p. 195–206.

[64] S. Bishop, H. Okhravi, S. Rahimi, and Y. Lee, “Covert Channel Resistant Information Leakage Protection Using A Multi-Agent Architecture,” IET Information Security, vol. 4, no. 4, pp. 233–247, December 2010.

[65] A. Harel, A. Shabtai, L. Rokach, and Y. Elovici, “M-Score: A Misuseability Weight Measure,” IEEE Transactions on Dependable and Secure Computing, vol. 9, no. 3, pp. 414–428, May 2012.

[66] X. Shu, D. Yao, and E. Bertino, “Privacy-Preserving Detection of Sensitive Data Exposure,” IEEE Transactions on Information Forensics and Security, vol. 10, no. 5, pp. 1092–1103, May 2015.

[67] X. Shu, J. Zhang, D. D. Yao, and W. Feng, “Fast Detection of Transformed Data Leaks,” IEEE Transactions on Information Forensics and Security, vol. 11, no. 3, pp. 528–542, March 2016.
[68] X. Shu, J. Zhang, D. Yao, and W.-c. 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*, ser. CODASPY ’15. New York, NY, USA: Association for Computing Machinery, 2015, p. 147–149.

[69] X. Shu, J. Zhang, D. Yao, and W.-C. Feng, “Rapid and parallel content screening for detecting transformed data exposure,” in *2015 IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS)*, 2015, pp. 191–196.

[70] C. Maldonado, “Data Leakage Detection Using Dynamic Data Structure and Classification Techniques,” *INGE CUC*, vol. 11, no. 1, pp. 79–84, 01 2015.

[71] X. Shu, J. Zhang, D. Yao, and W.-C. Feng, “Rapid and parallel content screening for detecting transformed data exposure,” in *2015 IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS)*, 2015, pp. 191–196.

[72] C. Maldonado, “Data Leakage Detection Using Dynamic Data Structure and Classification Techniques,” *INGE CUC*, vol. 11, no. 1, pp. 79–84, 01 2015.

[73] K. Kaur, I. Gupta, and A. K. Singh, “E-Mail Protection System to Prevent Data Leakage,” *Vigyan Prakash*, vol. 16, pp. 30–36, 2018.

[74] A. Mohsen Nia, S. Sur-Kolay, A. Raghunathan, and N. K. Jha, “Physiological Information Leakage: A New Frontier in Health Information Security,” *IEEE Transactions on Emerging Topics in Computing*, vol. 4, no. 3, pp. 321–334, July 2016.

[75] J. Croft and M. Caesar, “Towards Practical Avoidance of Information Leakage in Enterprise Networks,” in *Proceedings of the 6th USENIX Conference on Hot Topics in Security*, ser. HotSec’11. Berkeley, CA, USA: USENIX Association, 2011, pp. 7–7.

[76] J. Jung, A. Sheth, B. Greenstein, D. Wetherall, G. Maganis, and T. Kohno, “Privacy Oracle: A System for Finding Application Leaks with Black Box Differential Testing,” 01 2008, pp. 279–288.

[77] S. Alneyadi, E. Sithirasenan, and V. Muthukkumarasamy, “A Survey on Data Leakage Prevention Systems,” *Journal of Network and Computer Applications*, vol. 62, pp. 137 – 152, 2016.

[78] B. Hauer, “Data and Information Leakage Prevention Within the Scope of Information Security,” *IEEE Access*, vol. 3, pp. 2554–2565, 2015.

[79] S. Alneyadi, E. Sithirasenan, and V. Muthukkumarasamy, “Detecting Data Semantic: A Data Leakage Prevention Approach,” 08 2015, pp. 910–917.

[80] T. Wüchner and A. Pretschner, “Data Loss Prevention Based on Data-Driven Usage Control,” in *2012 IEEE 23rd International Symposium on Software Reliability Engineering*, 2012, pp. 151–160.

[81] A. Squicciarini, S. Sundareswaran, and D. Lin, “Preventing Information Leakage from Indexing in the Cloud,” in *2010 IEEE 3rd International Conference on Cloud Computing*, 2010, pp. 188–195.

[82] J. Linwood, T. Jaeger, R. Perez, R. Sailer, L. van Doorn, and R. Cáceres, “Trusted Virtual Domains: Toward Secure Distributed Services,” in *Proceedings of 1st IEEE Workshop on Hot Topics in System Dependability (HotDep’05)*, 06 2005.

[83] W. Blanke, “Data Loss Prevention Using An Ephemeral Key,” in *2011 International Conference on High Performance Computing Simulation*, 2011, pp. 412–418.

[84] T. Wüchner and A. Pretschner, “Data Loss Prevention Based on Data-Driven Usage Control,” in *2012 IEEE 23rd International Symposium on Software Reliability Engineering*, 2012, pp. 151–160.

[85] A. Squicciarini, S. Sundareswaran, and D. Lin, “Preventing Information Leakage from Indexing in the Cloud,” in *2010 IEEE 3rd International Conference on Cloud Computing*, 2010, pp. 188–195.

[86] J. Linwood, T. Jaeger, R. Perez, R. Sailer, L. van Doorn, and R. Cáceres, “Trusted Virtual Domains: Toward Secure Distributed Services,” in *Proceedings of 1st IEEE Workshop on Hot Topics in System Dependability (HotDep’05)*, 06 2005.

[87] W. Blanke, “Data Loss Prevention Using An Ephemeral Key,” in *2011 International Conference on High Performance Computing Simulation*, 2011, pp. 412–418.

[88] T. M. S. Yoshihama, T. Mishina, “Web-based Data Leakage Prevention,” in *Presented at IWSEC*, 2010.

[89] K. Borders and A. Prakash, “Quantifying Information Leaks in Outbound Web Traffic,” in *2009 30th IEEE Symposium on Security and Privacy*, 2009, pp. 129–140.

[90] C. H. Suen, R. K. Ko, Y. S. Tan, P. Jagadpramana, and B. S. Lee, “S2Logger: End-to-End Data Tracking Mechanism for Cloud Data Provenance,” in *2013 12th IEEE International Conference on Trust, Security and Privacy in Computing and Communications*, 2013, pp. 594–602.

[91] W. Boehmer, “Analyzing Human Behavior Using Case-Based Reasoning with the Help of Forensic Questions,” in *2010 24th IEEE International Conference on Advanced Information Networking and Applications*, 2010, pp. 1189–1194.

[92] Y. Shapira, B. Shapira, and A. Shabtai, “Content-based Data Leakage Detection Using Extended Fingerprinting,” *arXiv*, 2013.

[93] Y. Lindell and B. Pinkas, “Privacy Preserving Data Mining,” *Journal of Cryptology*, vol. 15, pp. 177–206, 2002.

[94] J. Marecki, M. Srivatsa, and P. Varakantham, “A Decision Theoretic Approach to Data Leakage Prevention,” in *2010 IEEE Second International Conference on Social Computing*, 2010, pp. 776–784.
[91] J. M. Gómez-Hidalgo, J. M. Martín-Abreu, J. Nieves, I. Santos, F. Brezo, and P. G. Bringas, “Data Leak Prevention through Named Entity Recognition,” in *2010 IEEE Second International Conference on Social Computing*, 2010, pp. 1129–1134.

[92] M. Sokolova, K. El Emam, S. Rose, S. Chowdhury, E. Neri, E. Jonker, and L. Peyton, “Personal Health Information Leak Prevention in Heterogeneous Texts,” in *Proceedings of the Workshop on Adaptation of Language Resources and Technology to New Domains*, ser. AdaptLRTtoND ’09. USA: Association for Computational Linguistics, 2009, p. 58–69.

[93] J. J. Parekh, K. Wang, and S. J. Stolfo, “Privacy-Preserving Payload-Based Correlation for Accurate Malicious Traffic Detection,” in *Proceedings of the 2006 SIGCOMM Workshop on Large-Scale Attack Defense*, ser. LSAD ’06. New York, NY, USA: Association for Computing Machinery, 2006, p. 99–106.

[94] V. R. Carvalho, R. Balasubramanyan, and W. W. Cohen, “Information Leaks and Suggestions: A Case Study using Mozilla Thunderbird,” in *Sixth Conference on Email and Anti-spam (CEAS)*, California USA, 2009.

[95] P. Zilberman, S. Dolev, G. Katz, Y. Elovici, and A. Shabtai, “Analyzing Group Communication for Preventing Data Leakage Via Email,” in *Proceedings of 2011 IEEE International Conference on Intelligence and Security Informatics*, 2011, pp. 37–41.

[96] V. R. Carvalho and W. W. Cohen, “Preventing Information Leaks in Email,” in *Presented at SDM*, 2007, pp. 68–77.