Data Leakage Detection and Prevention: Review and Research Directions

Suvendu Kumar Nayak and Ananta Charan Ojha

Abstract Disclosure of confidential data to an unauthorized person, internal or external to the organization is termed as data leakage. It may happen inadvertently or deliberately by a person. Data leakage inflicts huge financial and nonfinancial losses to the organization. Whereas data is a critical asset for an organization, recurrent data leakage incidents create growing concern. This paper defines data leakage detection and prevention system and characterizes it based on different states of data, deployment points and leakage detection approaches. Further, this paper follows a systematic literature review considering a decade of the existing research efforts and makes a critical analysis thereof to highlight the issues and research gaps therein. The paper then proposes important research directions in the field of data leakage detection and prevention. This review helps fellow researchers and interested readers understand the research problem, appreciate the state-of-the-art techniques addressing the research problem, draw attention toward the research challenges and derive motivation for further research in this promising field.

Keywords Data leakage detection and prevention · Machine learning · Natural language processing · Information processing · Security and privacy

1 Introduction

In the present digital era, organizations are becoming more and more data-driven. Data and information are one of the most important assets for any data-driven enterprise. Data-driven decision making helps organizations manage the business operations well while keeping the customers and other stakeholders satisfied. Effective

S. K. Nayak (✉) · A. C. Ojha
Department of CSE, School of Engineering and Technology, Centurion University of Technology and Management, Bhubaneswar, Odisha, India
e-mail: suvendu.sonu@gmail.com

A. C. Ojha
e-mail: acojha2002@yahoo.co.in

D. Swain et al. (eds.), *Machine Learning and Information Processing*,
Advances in Intelligent Systems and Computing 1101,
https://doi.org/10.1007/978-981-15-1884-3_19

management practices in terms of data capture, processing, storage, usage and protection of data are indispensable for such organizations. In particular, the protection of data from loss and leakage is paramount to organizational policies and practices in order to stay ahead of business competition.

Data leakage is defined as the accidental or intentional distribution and disclosure of confidential or sensitive data to an unauthorized party. It may be caused by internal or external entities to the organization. Sensitive data of an organization may include financial data, corporate strategy information, trade secrets, intellectual property, data about future projects, personal data of customers and employees such as patient records, credit card data, biometric data and many such data depending upon the business and industry. Similarly, sensitive data for a government may involve data about internal security and law enforcement, military and defense secrets, relationships and transactions with political parties, confidential diplomatic engagements, etc.

While distribution and sharing of data are a necessary requirement for business operations, leakage of sensitive or confidential data results serious consequences such as heavy financial loss, damage of reputation and credibility, regulatory penalties, decrease of company share price and likes. According to Data Breach QuickView Report [1], the year 2015 reported all-time high 3930 incidents of data breach exposing 736 million records. The highest number of incidents accounted for the business sector was 47.2%, followed by education (13.9%), government (12.2%) and medical (6.8%). All other sectors combined were only 19.9% of the reported incidents. Several other reports and studies reveal that incidence of data leakage is frequent and significant in several organizations [2]. As per IBM Security's 2019 Data Breach Report [3], the global average cost of data breach is estimated to USD 3.92 million, and the most expansive sector health care accounts for USD 6.45 million. India accounts for USD 1.8 million as the total average cost of data breach with the highest cost in industrial sector estimated at USD 2.7 million.

Data leakage is a serious and growing security concern for every organization which creates a pressing need for leakage detection and prevention of sensitive data and information. Consequently, it drives ample research attentions toward development of effective solutions from both academia and industry. Although a plethora of research proposals and data leakage prevention systems available in the literature, there is a pressing need to find an effective approach to this problem of security and privacy of sensitive data and information [4, 5]. Thus, it remains an active field of research.

This survey paper provides a comprehensive understanding of the field of study and presents state-of-the-art approaches for data leakage detection and prevention problem. The paper is structured as follows. Section 2 discusses on phases of a data leakage detection and prevention system and presents a taxonomy to characterize data leakage detection and prevention systems. Section 3 provides a review of select research proposals in the field of study. Section 4 lists an array of research challenges and provides pointers for future research. Section 5 concludes the paper.

2 Data Leakage Detection and Prevention

Although a number of data leakage prevention solutions are available from software vendors, there is less clarity on the exact definition of a data leakage prevention system. While there is no commonly agreed definition of what exactly a data leakage detection and prevention (DLDP) system should be, an attempt is made here to avoid the ambiguity. It provides an understanding of the DLDP system and its characterization.

2.1 Phases of DLDP System

Data leakage detection and prevention systems aim at identifying sensitive data and information, monitoring its usages and movement inside and out of the organization and taking action to prevent unintentional or deliberate disclosure of it. As shown in Fig. 1, identifying the sensitive data is the first important step in the DLDP systems. What to be included and what not to be included within the scope of sensitive data are a very crucial decision to be taken by the organization. The DLDP system must be configured correctly; otherwise, the system may result in false negative, and the leakage of potentially sensitive data will go unnoticed. The sensitive data must be kept in a readable digital format so that the DLDP system can monitor its usage and movement through multiple channels such as email, instant messaging, file transfer, access using HTTP, blogs and copying to USB. While it is rather challenging to develop a system to monitor all possible channels of data leakage, most systems focus on critical channels such as email, instant messaging, file transfer, HTTP and USB copy. Several techniques such as content matching, image recognition, fingerprinting, and statistical analysis can be used by DLDP systems to detect sensitive data leakage during channel monitoring. Once the leakage is detected, the system may perform one or more of the following actions. It may simply log the incident for later analysis and investigation, notify the risky behavior of the individual involved in the leakage to designated authorities and block the data being transmitted.

2.2 Characterizing DLDP Systems

DLDP systems can be characterized mainly based on which data state they handle, where they are deployed and what approach they employ to detect leakage. The classification is shown in Fig. 2. Since techniques for monitoring data are different

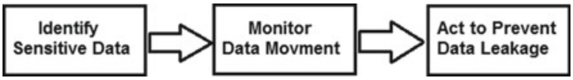

Fig. 1 Core phases of DLDP system (adopted from [21])

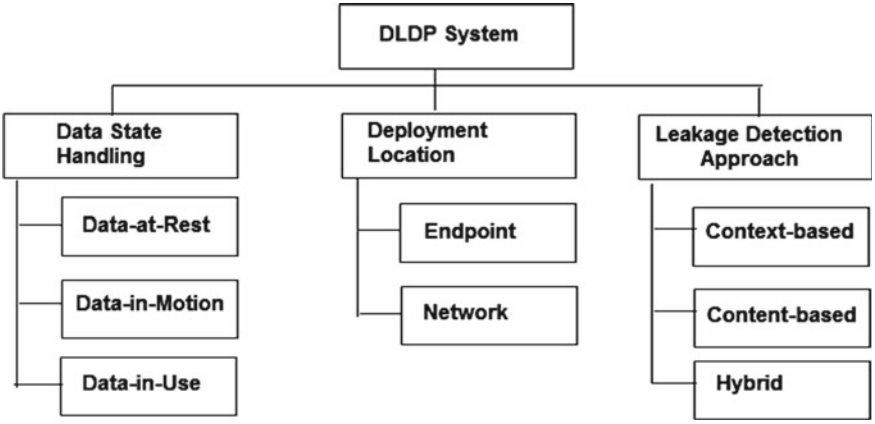

Fig. 2 DLDP system classification (adopted from [4])

for different data states, DLDP systems may be characterized based on three states of data such as Data-at-Rest, Data-in-Motion and Data-in-Use. Data-at-Rest can be defined as data in storage such as in file systems, databases, servers, desktop and laptops. Data-at-Rest can be local or remote to the DLDP system. A DLDP system may monitor Data-at-Rest at regular intervals or on demand. It checks if the sensitive data is encrypted or not. It checks if the access control policies are violated or not. Data-in-Motion can be defined as data that is passing through a network such as the Internet or local area network. A DLDP system usually monitors the network traffic by inspecting data packets to identify sensitive data traveling over the network using variety of protocols such as SMTP, FTP and HTTP. Data-in-Use can be defined as data that is being processed on endpoint devices. A DLDP system monitors the sensitive data and checks if it is being transmitted from one endpoint to another. Usually, the system monitors operations such as copy and paste between applications, screen capture, print, download to portable storage device such as USB drives and CD/DVD.

DLDP systems can be characterized based on their deployment location. A DLDP system can be deployed directly on endpoint devices or on network level. The DLDP system that is deployed on an endpoint device monitors and controls access to sensitive data, while a central server is responsible for policy creation and management, policy distribution and violation management, system administration and creation logs for incidents. On the other hand, a DLDP system deployed at network level inspects traffic based on leakage detection policy at multiple monitoring points in the network and blocks suspicious packets. DLDP systems can be further characterized based on the approaches employed to detect leakage of data. They can be context-based, content-based and hybrid. In a context-based approach, a DLDP system analyzes contextual metadata such as source, destination, sender, recipients, header, file type and format. A content-based system analyzes the data content using several techniques such as natural language processing, statistical analysis, fingerprint and keyword matching. A DLDP system using both content-based and context-based analysis is a hybrid one.

3 Literature Review

Several research efforts have been found in academic research literature to address the data leakage problem. Most of the works have been made to detect and prevent data leakage considering different states of data. These proposals mostly used either content analysis or context analysis for leakage detection. A few of them used the hybrid approach. The analysis techniques used in the works differ from one to another. This paper considers select research works carried out in the past decade based on their relevance. These are summarized below and their characterization is shown in Table 1.

Zilberman et al. [6] proposed an approach to detect email leakage based on topic identification and sender-recipient analysis in a group communication. A person in an organization belongs to several topic groups. The recipient of an email is classified as legal or illegal based on his/her association with the topic of the email being exchanged. The email content is analyzed to establish its topic using k-means clustering. The proposal was evaluated on a standard data set and provides 85% of

Table 1 Characterization of surveyed approaches

Proposal, Year	Leakage detection approach	Data state handling	Deployment location
Zilberman et al., 2011 [6]	Hybrid	Data-in-Motion	Endpoint
Shapira et al., 2013 [7]	Content-based	Data-in-Use	Endpoint
Shu et al., 2015 [8]	Content-based	Data-at-Rest, Data-in-Motion	Network
Costante et al., 2016 [9]	Hybrid	Data-at-Rest	Endpoint
Papadimitriou et al., 2011 [10]	Context-based	Data-in-Use, Data-in-Motion	Network
Alneyadi et al., 2016 [11]	Hybrid	Data-in-Motion	Endpoint
Shu et al., 2016 [12]	Content-based	Data-in-Use, Data-in-Motion	Network
Katz et al., 2014 [13]	Context-based	Data-in-Motion	Endpoint
Gomez-Hidalgo et al., 2010 [14]	Content-based	Data-in-Motion	Endpoint
Trieu et al., 2017 [15]	Content-based	Data-in-Motion, Data-in-Use	Endpoint
Lu et al., 2018 [16]	Hybrid	Data-in-Motion, Data-at-Rest	Network
Chhabra et al., 2016 [17]	Context-based	Data-in-Use	Network

accuracy. However, it suffers from a serious problem when the trained model does not have any information about a recipient.

Shapira et al. [7] proposed a content-based approach that uses well-known fingerprinting methods. The proposal uses *k*-skip *n*-gram technique and extracts fingerprints from confidential contents while skipping nonconfidential contents of a document to address the problem of traditional fingerprinting methods. It also takes care of rephrasing of contents and makes it possible to detect intentional leakage incidents. While it outperforms the traditional fingerprinting methods, it requires extensive data indexing when comes to implementation.

Shu et al. [8] proposed fuzzy fingerprint technique to enhance data privacy during leakage detection. The approach enables the data owner to safely delegate the content inspection task to a leak detection provider without disclosing the content. Although the network-based approach is efficient and provides satisfactory results under various data-leak scenarios, it suffers from computational complexity and realization difficulty.

Costante et al. [9] proposed a hybrid approach that combines signature-based and anomaly-based techniques to detect and prevent data leakage. It identifies insider threats by monitoring the activities of users and detecting anomalous behavior. Once the malicious behavior is detected, it is flagged up and the transaction is blocked. An attack signature is created and recorded to prevent such type of activities in the future. It uses a rule base which is updated automatically when a new anomaly is detected. While it attempts to combine both detection and prevention of leakage attacks using a rule-based technique, the system is flooded with false-positive alerts in the initial stage of system operation.

Papadimitriou et al. [10] studied guilty agent identification. The authors proposed data allocation strategies using which a distributer discloses sensitive data to a set of supposedly trusted agents. The distributer then assesses the likelihood that an agent is a guilty agent and responsible for the leakage when data is leaked. The approach is robust in case the released data is altered. However, the study does not capture adequate leakage scenarios.

Alneyadi et al. [11] proposed a hybrid method to detect potential data leakage in email communications of an organization. The method first performs context analysis using five contextual parameters in an email to measure RAI. The mail that scores high RAI is subjected to semantic analysis of its content in order to detect data leakage. The proposal showed encouraging results in detecting data leakage in emails. However, it suffers from implementation issues and consequently poor performance when it is not possible to capture all five contextual parameters in an organization.

Shu et al. [12] proposed a content-based approach that detects data leakage in transformed content. When the leaked data is modified, it becomes difficult to detect using usual n-gram technique. The problem is addressed using an alignment method in which a pair of algorithms, one sampling algorithm and another alignment algorithm, is used to compute a similarity score between the sensitive data sequence and the content sequence under inspection. The approach is efficient and results in high specificity; i.e., the percentage of true positive is very high than false-positive cases. However, it suffers from computational complexity.

Katz et al. [13] proposed a context-based model called CoBAn to detect data leakage in a controlled communication channel of an organization. During training phase, the model considers both types of documents, confidential and nonconfidential, and identifies various subjects dealt in these documents. Using k-mean clustering, it develops clusters of documents representing subject or context approximation. The model generates a context-based confidential term graph for each cluster. In detection phase, documents are analyzed and matched with one or more term graphs to estimate their confidentiality score. A document is detected confidential if its score crosses a predefined threshold. Although the model is able to find confidential data hidden in a document, the time complexity of the model is very high. Further, the model results in high rate of false positive which may not be acceptable in a real-world system.

Gomez-Hidalgo et al. [14] proposed a context-based approach that uses named entity recognition (NER) technique to detect data leakage of individuals and companies. Using Twitter data, experiments were conducted on a developed prototype which demonstrated encouraging accuracy level. However, the attributes considered in the NER technique are mostly homogenous.

Trieu et al. [15] proposed a method that used semantic and content analysis of documents to detect sensitive data leakage. The model uses document embedding to generate vector representation of a document or a text fragment. This vector representation is evaluated using a sensitivity corpus to find the sensitivity score of the document or text fragment. Experimental results show very high detection accuracy.

Lu et al. [16] proposed an approach for collaborative data leakage detection over distributed big data sets. The approach performs a privacy-preserving collaborative training on each owner's data which eventually trains and optimizes a global model. It uses a graph masking technique on the local weighted graphs representing local data and develops a trained template graph that represents global graph space. Then, the document to be tested is converted to graph and matched with the trained template graph to compute the sensitivity score of the test document. Although the approach can handle leakage detection with efficiency, the computational complexity remains high.

Chhabra et al. [17] studied data leakage detection in MapReduce computation in cloud computing environment. They used s-max algorithm to the reduced data in order to identify the guilty agent when any data leakage happens. They conducted a simulation for parallel processing of weather forecasting data using Hadoop framework and cloud analyst tool. However, the probability of indentifying the guilty agent is not very significant, and it reduces with increase in number of agents.

4 Challenges and Research Directions

DLDP systems face several challenges while preventing leakage of sensitive data [5]. One of them is abundant leaking channels such as email, social media, USB, printer, fax, smart phone and laptop. It is difficult to manage and secure all the channels. It is also very complex to model all possible activities of a person with sensitive data as well as leakage channels while developing a DLDP system. Another major challenge is the transformed data which is very hard to detect. When sensitive data is modified, its identity and patterns are changed making leakage detection very challenging.

The advent of big data and cloud computing has amplified the challenges of a DLDP system [18]. The system should be scalable to process massive data using parallel processing in a distributed environment. Real-time leakage detection is a requirement but a huge challenge for DLDP systems while dealing with big data. Anonymization and privacy preservation in big data are rather challenging to protect sensitive information [19]. Since data is kept in a remote location away from the data owner in a cloud environment, security and privacy remain a major concern which adds to the leakage detection problem. Multi-tenancy model in cloud computing offers threats of data leakage due to vulnerabilities in inter-tenant isolation [20]. Successful DLDP systems must be able to deal with the above said challenges in order to address growing concern of data leakage.

Additionally, there exist several areas with research opportunities which require further efforts from the researcher community. Deep learning has been successfully applied in various domains. The efficacy of deep learning can be exploited in both context and content analysis to detect data leakage and identify the insider threat with higher accuracy while achieving timely protection of sensitive data. In particular, deep learning-based leakage detection may be investigated in transformed data, wherein sensitive information is hidden in the exposed content. Cloud computing offers a new avenue for DLDP systems. Leakage detection and prevention can be offered as Software-as-a-Service (SaaS). Consequently, privacy preservation of the sensitive data becomes a major concern if DLDP system is offered as SaaS. Further, data leakage detection in MapReduce computation is a key research direction in cloud computing.

5 Conclusions

Data leakage is a persistent problem in organizations and inflicts grave consequences. It requires constant research efforts to mitigate the problem. The paper has reviewed several research contributions published in the recent past in order to portray the state-of-the-art techniques in the field of data leakage detection and prevention. The objective of the paper has been to provide a comprehensive reference to the field and attract research attention of fellow researchers toward it. The review reveals that the existing solutions are not satisfactory enough to tackle the perils of data leakage. In particular, the new challenges are thrown by big data and cloud computing invite further investigation. Nevertheless, organizational policies play a very effective role in curbing the menace of data leakage. Organizations should have concise policies to identify their critical data, and its handling since successful DLDP solutions heavily rely on classified information. Organizations should have policies to monitor access and activities on sensitive data at network level as well as on various endpoints. Additionally, a bare-minimum policy of encrypting sensitive data should be universally implemented across the organization.

References

1. Data breach quick view. 2015. Data breach trends. Available at https://www.riskbasedsecurity.com/2015-data-breach-quickview/. Accessed on 5 Sept 2019.
2. Data leakage news. Available at https://infowatch.com/analytics/leaks_monitoring. Accessed on 5 Sept 2019.
3. IBM security's cost of a data breach report 2019. Available at https://www.ibm.com/security/data-breach. Accessed on 5 Sept 2019.
4. Asaf, Shabtai, Yuval Elovici, and Lior Rokach. 2012. A survey of data leakage detection and prevention solutions, 1st ed. Springer: New York Heidelberg Dordrecht London. https://doi.org/10.1007/978-1-4614-2053-8.
5. Alneyadi, S., E. Sithirasenan, and V. Muthukkumarasamy. 2016. A survey on data leakage prevention systems. *Journal of Network and Computer Applications* 62: 137–152.
6. Zilberman, P., S. Dolev, G. Katz, Y. Elovici, and A. Shabtai. 2011. Analyzing group communication for preventing data leakage via email. In *Proceedings of 2011 IEEE international conference on intelligence and security informatics*, 37-41, 10–12 July 2011. Beijing, China: IEEE.
7. Shapira, Y., B. Shapira, and A. Shabtai. 2013. Content-based data leakage detection using extended fingerprinting. arXiv preprint arXiv:1302.2028.
8. Shu, Xiaokui, Danfeng Yao, and Elisa Bertino. 2015. Privacy-preserving detection of sensitive data exposure. *IEEE Transactions on Information Forensics and Security* 10 (5): 1092–1103.
9. Costante, E., D. Fauri, S. Etalle, J.D. Hartog, and N. Zannone. 2016. A hybrid framework for data loss prevention and detection. In *Proceedings of 2016 IEEE security and privacy workshops*, 324–333. IEEE Computer Society.
10. Papadimitriou, P., and H. Garcia-Molina. 2011. Data leakage detection. *IEEE Transactions on Knowledge and Data Engineering* 23 (1): 51–63.
11. Alneyadi, S., E. Sithirasenan, and V. Muthukkumarasamy. 2016. Discovery of potential data leaks in email communications. In *Proceedings of the 10th international conference on signal processing and communication systems (ICSPCS)*, 1–10. Gold Coast, Australia: IEEE.
12. Shu, Xiaokui, Jing Zhang, Danfeng Daphne Yao, and Wu-chun Feng. 2016. Fast detection of transformed data leaks. *IEEE Transactions on Information Forensics and Security* 11 (3): 1–16.
13. Katz, G., Y. Elovici, and B. Shapira. 2014. CoBAn: A context based model for data leakage prevention. *Information Sciences* 262: 137–158.
14. Gomez-Hidalgo, J.M., J.M. Martin-Abreu, J. Nieves, I. Santos, F. Brezo, and P.G. Bringas. 2010. Data leak prevention through named entity recognition. In *Proceedings of IEEE 2nd international conference on social computing*, 29–34, Minneapolis, USA.
15. Trieu, Lap Q., Trung-Nguyen Tran, Mai-Khiem Tran, and Minh-Triet Tran. 2010. Document sensitivity classification for data leakage prevention with twitter-based document embedding and query expansion. In *Proceedings of 13th International Conference on Computational Intelligence and Security*, 537–543, 15–18 Dec 2017. Hong Kong, China: IEEE.

16. Lu, Yunlong, Xiaohong Huang, Dandan Li, and Yan Zhang. 2018. Collaborative graph-based mechanism for distributed big data leakage prevention. In *2018 IEEE Global Communications Conference (GLOBECOM)*, 9–13, Abu Dhabi, UAE.
17. Chhabra, S., and A.K. Singh. 2016. Dynamic data leakage detection model based approach for MapReduce computational security in cloud. In *Proceedings of fifth international conference on eco-friendly computing and communication systems (ICECCS-2016)*, 13–19. IEEE.
18. Cheng, L., F. Liu, and D. Yao. 2017. Enterprise data breach: causes, challenges, prevention, and future directions. *WIREs Data Mining and Knowledge Discovery* 7: e1211. https://doi.org/10.1002/widm.1211. Wiley & Sons, pp. 1–14.
19. Basso, T., R. Matsunaga, R. Moraes, and N. Antunes. 2016. Challenges on anonymity, privacy and big data. In *Proceedings of seventh Latin-American symposium on dependable computing*, 164–171, Cali, Colombia, 19–21 October. IEEE Computer Society.
20. Priebe, C., D. Muthukumaran, D. O'Keeffe, D. Eyers, B. Shand, R. Kapitza, and P. Pietzuch. 2014. CloudSafetyNet: detecting data leakage between cloud tenants. In *Proceedings of the 6th edition of the ACM workshop on cloud computing security*, 117–128, Scottsdale, Arizona, USA, November 7–7, 2014. ACM.
21. Data Leakage Prevention (DLP)-ISF Briefing Paper. Available to https://www.securityforum.org/research/data-leakage-prevention-briefing-paper/. Accessed on 5 Sept 2019.