Authorization Data Leakage Detection Using Heuristic Guilt Based Analysis

N. Balasubramanian[1], S. Harishma[2]
[1]*Associate Professor MCA, Mohamed Sathak Engineering College, Kilakarai*
[2]*Final Year MCA, Mohamed Sathak Engineering College.*

***Abstract*—Data leakage poses a significant risk in environments where sensitive information is distributed among multiple recipients. Traditional mechanisms like watermarking can be circumvented or are infeasible due to data integrity constraints. This paper proposes a Data Leakage Detection System utilizing a Heuristic Guilt-Based Analysis approach. The system embeds subtle, individualized data variations to track unauthorized disclosures and assess the likelihood of a leak originating from specific agents. By analysing access patterns, behavioural deviations, and incorporating change point detection methods, the system effectively identifies guilty entities even in dynamic environments subject to concept drift. The model enhances traceability and accountability in data distribution, reducing the impact of insider threats and improving organizational data security practices.**

***Index Terms*—Data leakage detection, Heuristic guilt-based analysis, Insider threat, Concept drift, Process mining, Data security, Fingerprinting, Change detection.**

1.INTRODUCTION

In the modern era of digital information sharing, the secure handling of sensitive data is a paramount concern for organizations across various domains. With the rise of data-driven decision-making and the proliferation of data distribution among stakeholders—ranging from internal employees to third-party partners—the risk of unauthorized data disclosure has significantly increased. Traditional security mechanisms such as encryption and access control are effective in preventing external threats but often fall short when dealing with insider threats, where individuals with legitimate access may intentionally or inadvertently leak confidential information.

This paper presents a novel approach to addressing this challenge through the implementation of a Data Leakage Detection System that employs a Heuristic Guilt-Based Analysis methodology. Unlike watermarking techniques, which require modification of the original data, this system maintains data integrity while embedding imperceptible variations tailored for each data recipient. When a data breach occurs, the system analyses the leaked content to compute a probabilistic guilt score for each recipient, based on the similarity between their unique data version and the compromised file.

Additionally, the system integrates process mining principles to detect concept drift—situations where business processes evolve over time due to changes in operational dynamics, legislation, or external conditions. By monitoring behavioral patterns and adapting to such changes, the proposed system enhances the detection of anomalies and strengthens the identification of potential leakers.

Through heuristic analysis, pattern recognition, and probabilistic modelling, the proposed solution not only detects data breaches effectively but also helps organizations take timely corrective actions. This framework is particularly valuable in sectors like healthcare, finance, and government, where data sensitivity and compliance requirements are critical. Ultimately, the system fosters a secure environment for data sharing, reduces the risk of insider threats, and reinforces trust in digital information systems.

2.LITERATURE SURVEY

The increasing need for dynamic and secure data management systems in today's interconnected environments has motivated extensive research into data leakage detection and process mining. This section outlines key contributions in the field that inform the proposed approach.

Van Infrastructurean Milieu (2010) introduced the concept of a unified permit system in the Netherlands, streamlining administrative processes and illustrating the need for integrated systems to manage complex workflows securely and efficiently. While not directly related to data leakage, this work underscores the importance of centralized and coordinated systems, which are foundational to secure data governance.

Van der Aalst, Rosemann, and Dumas (2011) explored deadline-based escalation mechanisms in Process-Aware Information Systems (PAIS). Their framework provides insights into how time-sensitive processes can be monitored and modified dynamically to meet operational constraints. This is particularly relevant to data leakage detection, where timely identification of abnormal behaviour is critical.

Bose et al. (2011) addressed the issue of concept drift in process mining, highlighting how operational processes evolve over time due to external influences such as legislative changes and market dynamics. They proposed methodologies to detect and manage these shifts within recorded event logs, thus enhancing the reliability of anomaly detection in evolving environments.

Buijs, van Dongen, and van der Aalst (2012) advanced the concept of cross-organizational process mining. Their approach enables the comparison and alignment of process models and event logs across organizational boundaries, providing a framework for collective monitoring and anomaly identification in shared infrastructures.

Van der Aalst (2010) further discussed configurable services in the cloud, advocating for customizable yet standardized business processes across different entities. This flexibility is key to deploying scalable data leakage detection mechanisms that adapt to varying organizational needs while maintaining consistency in detection logic.

These foundational studies emphasize the criticality of dynamic, adaptive, and intelligent systems for effective process monitoring and data security. While much work has focused on process optimization and anomaly detection, fewer approaches tackle insider threats and data leakage using probabilistic reasoning. The proposed heuristic guilt-based analysis system addresses this gap by combining behavioral analytics, fingerprinting, and probabilistic modelling to identify potential data leakers within trusted networks.

2.1EXISTING SYSTEM

In the existing system, data leakage detection relies on the assumption that the process underobservation remains stable over time, and sufficient event logs can be used to derive a high-quality process model. This model is then employed for tasks such as performance analysis, compliance checking, and predictive analytics. Traditional methods like data watermarking are used, where unique codes are embedded into each copy of distributed data to trace leaks. However, these approaches have significant limitations. They often require modification of the original data, which may not be permissible in all scenarios due to the sensitive nature of the content. Moreover, malicious recipients can potentially remove or alter watermarks, rendering the detection ineffective. Consequently, the system falls short in environments where data must remain unaltered, or where insider threats exploit their legitimate access to disseminate confidential information without leaving detectable traces. These drawbacks necessitate a more advanced and resilient framework to effectively trace and manage data leaks, especially in dynamic and complex data-sharing ecosystems

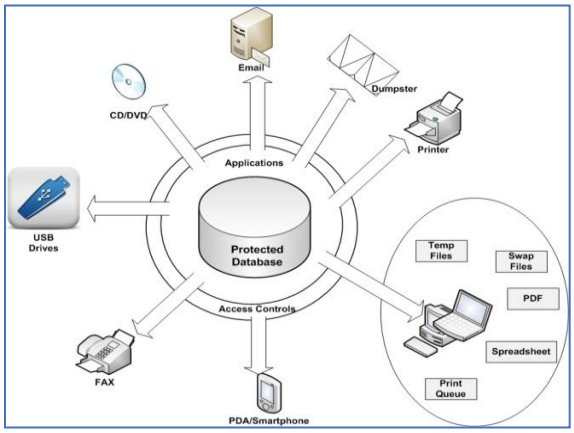
2.2PURPOSE OF THE WORK

The primary purpose of this work is to develop a robust system for detecting data leakage by leveraging a heuristic guilt-based analysis approach. In environments where sensitive data is distributed to multiple agents or users, the risk of unauthorized dissemination—whether deliberate or accidental—poses significant challenges to information security. Traditional methods, such as watermarking or data perturbation, often involve modifying the original data, which may not always be permissible or effective. The proposed system addresses this issue by introducing a novel data allocation and tracking mechanism that preserves the integrity of the original data while enhancing traceability. By distributing uniquely fingerprinted data to each agent and monitoring access patterns, the system can assign probabilistic guilt scores to users suspected of leaking data. Additionally, the integration of behavioral analysis and change-point detection allows the system to adapt to evolving threats and improve detection accuracy. This approach is particularly vital in sectors where data con-

fidentiality is paramount—such as finance, healthcare, and government—ensuring proactive protection against insider threats while maintaining data authenticity.

3.PROPOSED SYSTEM

The proposed data leakage detection system employs a heuristic guilt-based analysis framework to trace and mitigate unauthorized dissemination of sensitive data. Unlike traditional methods that rely heavily on data alteration techniques such as watermarking or perturbation, this system maintains data integrity by assigning uniquely modified yet imperceptible copies of data to different recipients. Upon detection of a leak, the system analysis the exposed content and correlates it with the distributed versions to identify likely sources. A key innovation lies in the integration of change point detection mechanisms—whereby time-series features derived from event logs are analysed to detect significant process changes or concept drifts. These features, representing control flow dependencies, enable the system to identify shifts in user behaviour or access patterns that may coincide with leakage events. The guilt assessment model further enhances detection accuracy by assigning probabilistic scores to agents based on the presence of fake objects—specially crafted data items embedded in distributed sets—which, if leaked, provide strong evidence of the responsible party. Collectively, this architecture not only bolsters insider threat detection but also supports dynamic organizational processes by offering adaptive and non-intrusive security mechanisms, making it suitable for high-stakes environments such as finance, healthcare, and government sectors.



4.MODULES

The Data Leakage Detection System employs a modular architecture to systematically identify unauthorized data disclosures. The first module is the Log-in/Registration system, which authenticates users and controls access to the application. This ensures that only authorized individuals can interact with the system, thereby establishing a secure perimeter from the outset. Following user authentication, the Data Transfer module handles the secure allocation and transmission of data from the distributor to agents. This module also captures instances of illegitimate transfers between agents, setting the groundwork for leak detection.

A core component of the system is the Guilt Model Analysis module. It implements a heuristic-based approach where each data item contains embedded identifiers (e.g., fake objects), allowing for the probabilistic tracing of data leaks. This module incrementally tracks data interactions and builds a profile of agent behaviours that are indicative of potential leakage.

The Change Point Detection module addresses the temporal dynamics of process behaviour. It identifies shifts or drifts in data access patterns by analysing event logs for significant variations. This is essential in environments where operational procedures evolve over time, potentially altering the risk profile of users.

Finally, the Agent-Guilt Model module synthesizes the data gathered from the guilt analysis and change detection stages to pinpoint suspicious agents. By correlating leaked data with known fingerprints and fake records, the system assigns a probabilistic guilt score to each agent. Visualization tools, including graphical representations, are used to enhance the interpretability of this analysis, supporting informed decision-making regarding culpability.

5.RESULT AND CONCLUSION

The implementation of the Data Leakage Detection System using Heuristic Guilt-Based Analysis demonstrated promising outcomes in accurately identifying sources of unauthorized data disclosures. By assigning uniquely traceable copies of data to different agents and integrating fake objects for verification, the system was able to calculate guilt scores and trace

data leakage with high reliability. Through the application of heuristic algorithms and behavioral analysis, instances of data misuse were successfully linked to responsible agents, even in the presence of modified or partially leaked data. The inclusion of a change detection module enabled dynamic monitoring of process behavior and the timely identification of concept drifts. Experimental analysis using real-world event logs from a Dutch municipality confirmed the system's capability to detect process deviations and adjust risk assessments accordingly. The approach not only enhanced traceability and accountability but also served as a deterrent against malicious insider activity. In conclusion, the proposed system effectively combines probabilistic modelling with behavioral analytics to strengthen data security, offering a scalable and adaptive solution to the growing challenge of data leakage in sensitive environments. Future enhancements are expected to further refine the detection of control-flow deviations by introducing additional analytical features and expanding the model's predictive capabilities.

6.FUTURE ENHANCEMENTS

1. Enhanced Feature Set for Concept Drift Detection
Currently, four features are used to detect control flow dependencies. Future work can expand this to include six or more advanced features, capturing more nuanced behavioral changes in data usage or access patterns. This would increase the granularity and accuracy of detecting concept drift in processes.
2.Integration of Machine Learning Models
Incorporate machine learning techniques like decision trees, random forests, or deep learning models to predict and classify potential leakers based on behavioral patterns, access frequency, and data modification traces.
3.Real-time Monitoring and Alerting System
Develop a real-time alert mechanism that uses the guilt score and user behavior to notify administrators immediately when abnormal activities are detected.
4.Advanced Visualization Dashboard
Create a visualization dashboard to represent data flow, agent activities, leak probabilities, and concept drift points dynamically. This would assist in rapid analysis and decision-making.
5.Support for Cross-Organizational Data Sharing
Extend the system to support and secure data shared across multiple organizations using configurable process models and cloud-based BPM tools, enabling safe collaboration without compromising data integrity.
6.Blockchain for Immutable Data Trails
Use blockchain technology to store data distribution records and access logs immutably, ensuring traceability and tamper-proof records for audits and legal processes.
7.Collusion Detection Algorithm
Enhance the guilt analysis by adding collusion detection capabilities that can identify coordinated leaks involving multiple agents.
8.User Behavior Profiling
Continuously update user profiles based on their access history and anomalies, using anomaly detection algorithms to improve the system's adaptability to insider threats.
9.Fake Data Object Optimization
Use optimization algorithms to determine the best number and type of fake data objects to insert for maximizing detection efficiency while minimizing system overhead.
10.Mobile and Distributed System Support
Extend the architecture to mobile devices and distributed systems to support broader data-sharing scenarios, including remote workforce environments.

REFERENCES

[1] van Infrastructure en Milieu, M.: All-in-one permit for physical aspects: (omgevingsvergunning) in a nutshell (2010)
[2] Sarbanes, P., G. Oxley et. al.: Sarbanes-Oxley Act of 2002 (2002)
[3] van der Aalst, W.M.P., Rosemann, M., Dumas, M.: Deadline-based Escalation in Process-Aware Information Systems. Decision Support Systems 43(2) (2011)492{511
[4] Bose, R.P.J.C., van der Aalst, W.M.P., Zliobait_e, I., Pechenizkiy, M.: HandlingConcept Drift in Process Mining. In Mouratidis, H., Rolland, C., eds.: InternationalConference on Advanced Information Systems Engineering (CAiSE 2011). Volume6741 of Lecture Notes in Computer Science., Springer-Verlag, Berlin (2011) 391{405

[5] Shokin, D.: Handbook of Parametric and Non-parametric Statistical Procedures.Chapman & Hall/CRC (2004)
[6] Bose, R.P.J.C.: Process Mining in the Large: Preprocessing, Discovery, and Diagnostics. PhD thesis, Eindhoven University of Technology (2012)
[7] CoSeLog: Congurable Services for Local Governments Project Home Page. www.win.tue.nl/cos log.
[8] Buijs, J.C.A.M., van Dongen, B.F., van der Aalst, W.M.P.: Towards Cross Organizational Process Mining in Collections of Process Models and their Executions. In Daniel, F., Baraki, K., Dust Dar, S., eds.: Business Process Management Workshops. Volume 100 of Lecture Notes in Business Information Processing., Springer-Verlag, Berlin (2012