**JOURNAL OF ALGEBRAIC STATISTICS**
Volume 13, No. 3, 2022, p. 4742 - 4746
https://publishoa.com
ISSN: 1309-3452

# A Review on Identifying Data Leakage Using Guilt Agent

**Dr. A.Vijay Kumar [1], Dr. Manjunath CR [2]**

*[1, 2] Associate Professor, CSE, FET, Jain Deemed to be University, Bangalore, Karnataka, India.*

**ABSTRACT**

In the present scenario individuals are utilizing the web all over and sharing information like records, pictures, video and so forth through the net and it has been heightening step by step. While sharing the information to each other there are a few greater issues like assaults, hacking and information spillages these are a few hardships which continues forever. To vanquish these conditions, we really want to keep the information spillages from the outsider particularly in the associations where the information is being shared. To keep the information from being spilled we apply the Culpability Specialist Technique that works on the likelihood of distinguishing information spillages and forestalls the information spillages.

**Keywords— Guilt Agent Method, Data leakages, Attacks and Hacking.**

## I. INTRODUCTION

Data leakage is an unauthorized broadcast of data within the organization to an external destination or recipient. This can be transferred electronically or physically.
Data Breach is caused when there is a weak password or the hacker hack the website and transmits the personal data to his account.
Most of the time data leakage happens through the social websites like Amazon, Flipkart, Snapdeal, Paytm, Phone pay etc when money is transmitted from one account to another. Then hacker hacks the websites and steals all the required information related to credit cards, account numbers, pan number, email ids, phone numbers etc. Through this information, hackers will try to transfer the money from the victim's account.
There are many types of data leakages like:

**Ill-intentional or malicious internal employees**: Data leakages are not only done by online mediums may also happen with company's employees, who sells, the company's internal information for their self-centered profits.

**Accidental data breaches any malicious**: Sometimes unauthorized data leakage may happen accidentally without intention or purposely by external agents.

**Malicious intent in Electronic communication**: There several firms that allows the organizations to share messages like chat rooms, cloud and other social networking sites, as a part of their daily roles. These data leakage programs work in the backend without letting users to know about their existence.

**Physical data leakage**: It may happen with the mistake of the employee of an organization, sometimes they will do intentionally also, in that case they may steal data on the drive. Similarly, they may upload some data into the cloud just by lending their system for a few mini tenses.
To overcome this data leakage issues there are several types of technology like content matching, image recognition, fingerprinting, and statistical analysis which monitor the data and prevent it from leakage.
There are some methods like watermarking entity sets, Data Allocation Strategies, Optimization methods, Fake objects, Data distributor Strategies which helps to Identify, detect and prevent the data and reduce the complexity.

## II. REVIEW OF LITERATURE

In the article titled “Implementation of Guilt Model and Allocation Strategy for Data Leakage Detection”, they have used Traditional methods like watermarking entity sets, fake object allocation for finding the data leakages. In this paper they have proposed a technique to detect any leakages of data by using machine learning algorithms and the Sales handy website.
In the article titled “A KNN-SVR Data Mending Method for Insufficient Data of Magnetic Flux Leakage Detection” they have used Magnetic Flux Leakage (MFL) for the detection of the data leakage. A specific data restoration method based on K Nearest Neighbor-Support Vector Regression (KNN-SVR) is introduced, which in fact reduce the training cost of SVR and really improve the correctness of the algorithm. This technique is experienced and the outcome confirmed that the anticipated technique can improve the accuracy rate of data by restoration deficient data with an acceptable time cost.

In the article titled “Detecting data semantic: A data leakage prevention approach” authors developed a data leakage detection system using a variety of share policy which evaluate the chance that the leaked data came from one or more agents.

In the case of secure transactions, only privileged users were able to access sensitive data using access control policies, in that way, data leakages were prevented. They also tried with inserting the fake records in the dataset; with this experiment the probability of identifying the data leakages was improved. The mechanism they have implemented on the cloud which comprises of several allocation strategies.

In the article titled “Sensitive data leakage detection in pre-installed applications of custom Android firmware” they have used UIT (unit Investment trust) ROM, to detect data leakage in customary Android firmware by analyzing dealings of pre-installed applications. The new results show that the system can become aware of all sensitive data leakage in our custom Android firmware. Secondly, it detects several pre-installed applications which leak perceptive data from 290 custom ROMs downloaded from the Internet.

In the article titled “Avoiding the data leakage and providing privacy to data in networking” they have used existing systems to solve the problem of the data leakage, but, this method does not provide security to the data or the system. Hence in the proposed method they have introduced the method which provides security to the data leak recognition system.

In the article titled “A Survey: Data Leakage Detection Techniques” they propose a model to draw and avert unauthorized users from accessing the cluster again. They use the LD (Leakage Detector) algorithm to spot the guilt agent responsible for data leakage. Researchers developed a Logger which helps to offer a unique id to each user and keep a track of the actions carried by them throughout the complete session. The present work may enhance the security in the Hadoop framework.

In the article titled “An enhanced model for preventing guilt agents and providing data security in a distributed environment” they implement a method, aimed at improving the odds of detecting such leakages when a distributor's sensitive data has been leaked by trustworthy agents and also to possibly identify the agent that leaked the data. By adding fake objects to the distributed set, the distributor can find the guilty party.

In the article titled “A probability-based model for data leakage detection using bigraph” In some Data Leakage Detection models used the ‘fake objects’ which are stored in the server database. The fake objects help to identify the user who has leaked the file. Every user has a probability to leak the file which probability is called the guilt probability. Those users who have the probability to leak the file are known as the guilt probability. Many models of detection of data leakage focus on the fake object which is included with the database to find out the leaker.

In the article titled “A Study of Data Allocation Problem for Guilt Model Assessment in Data Leakage Detection Using Cloud Computing” Contemporary business relies on sharing and transferring information among various stakeholders such as employees, owners (shareholders), creditors, and suppliers within or outside the organization. As the shared critical data can be leaked by some malicious entity, the persistence of preventing data misuse is expanded. Severe damage caused by sharing of sensitive information constitutes a threat to the organization's assets.

Protection of confidential data from unauthorized revelation is a matter of concern for any enterprise.

In the article titled “Detection and location for slow leakage of oil pipeline based on weighted logical inference and data fitting X Hu” A leakage location method for oil pipeline based on weighted logical inference and data fitting (WLIDF) are proposed, which can decrease the number of false alarms and improve the accuracy of slow leakage detection and location. When the leakage happens, the data fitting is used to decide the location and time of the leakage. According to the method of WLIDF, the slow leakage can be found on time and the position of the leakage point can be accurately located. The simulation and application show the effective and good performance of the proposed.

In this article “Fast Detection of Transformed Data Leaks” they have used the Utilize sequence alignment technique for detecting complex data leakage patterns. They have used Comparable sampling algorithm alignment and sampling-oblivious algorithm. These algorithms help to provide substantial speedup and high scalability of the design data-movement tracking approached which is not use in any techniques.

In this article “Privacy-Preserving Detection of Sensitive Data Exposure” this Provide privacy-preserving data-leak detection (DLD). They are using the MapReduce algorithm in which it detects a special set of important information digests and it is Capable to arbitrarily scale and use for public resources, which typically uses strong encryption.

In this article “DDSGA: A data-driven semi-global alignment approach for detecting masquerade attack” They are given the Data-Driven Semi-Global Alignment, DDSGA method to secure and upgrade the DDSGA scoring systems by adopting distinct alignment parameters for every user. DDSGA approach and results improving both the hit ratio as well as false-positive rates with an acceptable calculation overhead

Several Technical Challenges in Data Leak detection are as follows:

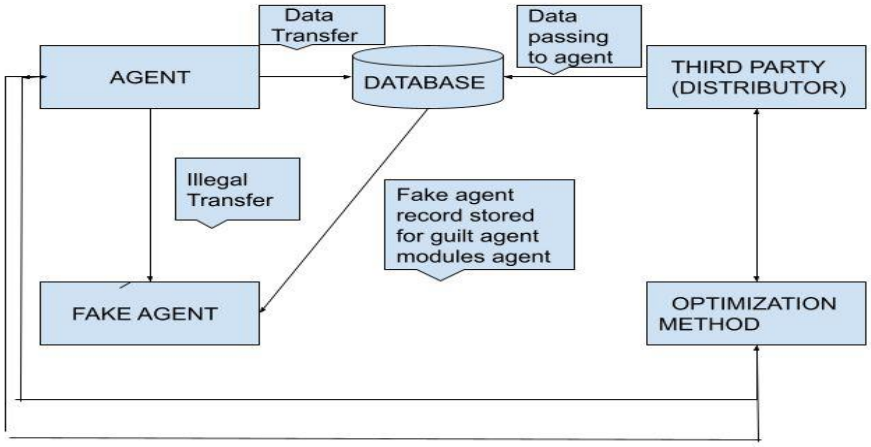
1. Scalability
2. Privacy Preservation
3. Accuracy
4. Timeliness

**JOURNAL OF ALGEBRAIC STATISTICS**
Volume 13, No. 3, 2022, p. 4742 - 4746
https://publishoa.com
ISSN: 1309-3452

## III. METHODOLOGY

The aim of this project is to find out the data leakage problem by the help of the Guilt agent method.
**Stage-1**: Studying and analyzing the different types of Guilt Agent Method.
S**tage-2**: Implementing the dataset.(mysql)
**Stage-3**: Implementing the Guilt Agent Method. (php language )
**Stage-4**: Compare and analyze the outcomes.

**System Design**




**Comparative Study**

| **Modules** | **Accuracy** | **Complexity** | **Capacity** | **Detection** | **Robust** | **Strength** |
|---|---|---|---|---|---|---|
| Data Allocation | high | low | low | Not easy | yes | Modification of data is not possible |
| Fake object | low | medium | Depend on hidden data | Not easy | no | Find the probability of data leakage by the agent |
| optimization | medium | low | Depend on data size | Depend on technology used | no | It reduces the information that used for the main user |
| Data distribution | high | medium | moderate | Depend on admin intelligences | yes | Permission is not grant to the unauthorized user |

## IV. CONCLUSION

In the real world, the organizations are facing problem with data leakage. The data may be seen in other laptops or websites. From this survey, we conclude that the data leakage detection has always been a threat to data security and integrity. To overcome this issue there are multiple methods such as data allocation with boolean filter, bigrap method etc.
Different types of tools like DLP network security and DLP data leakage which help the data secure from unauthorized users.

**JOURNAL OF ALGEBRAIC STATISTICS**
Volume 13, No. 3, 2022, p. 4742 - 4746
https://publishoa.com
ISSN: 1309-3452

**REFERENCES**

1. A KNN-SVR Data Mending Method for Insufficient Data of Magnetic Flux Leakage Detection X Zhang,- 2018 IEEE 7th Data 2018 - ieeexplore.ieee.org
2. Sensitive data leakage detection in preinstalled applications of custom Android firmware T Nguyen - 2017 18th IEEE International 2017 - ieeexplore.ieee.org.
3. Avoiding the data leakage and providing privacy to data in networking M Suryawanshi,2016 International Conference on 2016 - ieeexplore.ieee.org
4. A Conditional Variational Autoencoder Algorithm for Reconstructing Defect Data of Magnetic Flux Leakage J Zhang - 2020 Chinese Control And Decision ,2020 - ieeexplore.ieee.org
5. A Survey: Data Leakage Detection Techniques KS Wagh - International Journal of Electrical and Computer 2018 - academia.edu.
6. A Survey on Data Leakage Detection and Prevention, R Verma, Available at SSRN 2020 - papers.ssrn.com.
7. An enhanced model for preventing guilt agents and providing data security in distributed environment S Praveen kumar,.Conference on Signal 2016 - ieeexplore.ieee.org.
8. A probability based model for data leakage detection using bigraph, AK Singh - Proceedings of the 2017 the 7th International 2017 - dl.acm.org.
9. A Study of Data Allocation Problem for Guilt Model Assessment in Data Leakage Detection Using Clouding Computing AU Bhosale - International Journal of Scientific - 2017 the 7th International -Citeseer.
10. A probability based model for data leakage detection using bigraph, AK Singh - Proceedings of the 2017 the 7th International 2017 - dl.acm.org.
11. Detection and location for slow leakage of oil pipeline based on weighted logical inference and data fitting X Hu, - 2016 Chinese Control and Decision, 2016 - ieeexplore.ieee.org.
12. Implementation of Guilt Model and Allocation Strategy for Data Leakage Detection, P Gupta, - Citeseer, 2020 - ieeexplore.ieee.org.
13. Hamza Fawzi, Paulo Tabuada, and Suhas Diggavi. Secure estimation and control for cyber-physical systems under adversarial attacks. IEEE Transactions on Automatic Control, 59(6):1454–1467, 2014.
14. Elias Bou-Harb, Claude Fachkha, Makan Pourzandi, Mourad Debbabi, and Chadi Assi. Communication security for smart grid distribution networks. IEEE Communications Magazine, 51(1):42–49, 2013.
15. Christian Rossow. Amplification hell: Revisiting network protocols for ddos abuse. In NDSS, 2014.
16. Elias Bou-Harb, Mourad Debbabi, and Chadi Assi. A statistical approach for fingerprinting probing activities. In Availability, Reliability and Security (ARES), 2013 Eighth International Conference on, pages 21–30. IEEE, 2013.
17. Ashok Anand, Chitra Muthukrishnan, Aditya Akella, and Ramachandran Ramjee. Redundancy in network traffic: findings and implications. ACM SIGMETRICS Performance Evaluation Review, 37(1):37–48, 2009.
18. Yousra Chabchoub, Christine Fricker, and Philippe Robert. Improving the detection of On-line Vertical Port Scan in IP Traffic. IEEE 7th International Conference on Risks and Security of Internet and Systems (CRiSIS), 2012.
19. Ke Wang, Janak J Parekh, and Salvatore J Stolfo. Anagram: A content anomaly detector resistant to mimicry attack. In International Workshop on Recent Advances in Intrusion Detection, pages 226–248. Springer, 2006.
20. Zhaoyan Xu, Antonio Nappa, Robert Baykov, Guangliang Yang, Juan Caballero, and Guofei Gu. Autoprobe: Towards automatic active malicious server probing using dynamic binary analysis. In Proceedings of the 2014 ACM SIGSAC CCS, pages 179–190. ACM, 2014.
21. Antonio Nappa, Zhaoyan Xu, M Zubair Rafique, Juan Caballero, and Guofei Gu. Cyberprobe: Towards internet scale active detection of malicious servers. In In Proceedings of the 2014 NDSS, pages 1–15, 2014.

22. Andrei Broder and Michael Mitzenmacher. Network applications of bloom filters: A survey. Internet mathematics, 1(4):485–509, 2004.

23. B.H. Bloom. Space/time trade-offs in Hash Coding with Allowable Errors. Communications of the ACM, 13(7):422–426, 1970.

24. M. Naor and M. Yung. Universal One-Way Hash Functions and their Cryptographic Applications. ACM Symposium on Theory of Computing, 1989. Seattle, WA.

25. S. Joshua Swamidass and Pierre Baldi. Mathematical correction for fingerprint similarity measures to improve chemical retrieval. Journal of chemical information and modeling (ACS Publications), 47(3):952–964, 2007. doi:10.1021/ci600526a, PMID 17444629.

26. Shahabeddin Geravand and Mahmood Ahmadi. Bloom filter applications in network security: A state-of-the-art survey. Computer Networks, 57(18):4047–4064, 2013.

27. David Moore, Colleen Shannon, Geoffrey M Voelker, and Stefan Savage. Network telescopes: Technical report. Department of Computer Science and Engineering, University of California, San Diego, 2004.

28. Claude Fachkha and Mourad Debbabi. Darknet as a source of cyber intelligence: Survey, taxonomy, and characterization. IEEE Communications Surveys & Tutorials, 18(2):1197–1227, 2016.

29. Elias Bou-Harb, Mourad Debbabi, and Chadi Assi. Cyber scanning: a comprehensive survey. IEEE Communications Surveys & Tutorials, 16(3):1496–1519, 2014.

30. Malware Repositories. http://malshare. com/, http://www.freetrojanbotnet.com/, http://openmalware.org/. [22] Evan Cooke, Michael Bailey, Farnam Jahanian, and Richard Mortier. The dark oracle: Perspective-aware unused and unreachable address discovery. In NSDI, volume 6, 2006.

31. Jayanthkumar Kannan, Jaeyeon Jung, Vern Paxson, and Can Emre Koksal. Semi-automated discovery of application session structure. In Proceedings of the 6th ACM SIGCOMM conference on Internet measurement, pages 119–132. ACM, 2006.

32. Elias Bou-Harb. A probabilistic model to preprocess darknet data for cyber threat intelligence generation. In Communications (ICC), 2016 IEEE International Conference on, pages 1–6. IEEE, 2016.

33. Elias Bou-Harb, Mourad Debbabi, and Chadi Assi. On fingerprinting probing activities. Computers & Security, 43:35–48, 2014.

34. Elias Bou-Harb, Mourad Debbabi, and Chadi Assi. A systematic approach for detecting and clustering distributed cyber scanning. Computer Networks, 57(18):3826–3839, 2013.

35. Elias Bou-Harb, Mourad Debbabi, and Chadi Assi. Behavioral analytics for inferring large-scale orchestrated probing events. In Computer Communications Workshops (INFOCOM WKSHPS), 2014 IEEE Conference on, pages 506–511. IEEE, 2014.

36. Vern Paxson. Bro: a system for detecting network intruders in real-time. Computer networks, 31(23):2435–2463, 1999.