

ONLINE ANAMOLY FILE DETECTION

Saikumar S¹, N.Rajeshwari²

¹Dept. of MCA, Bangalore Institute of Technology, Bengaluru, India.

²Dept. of MCA, Assistant Professor, Bangalore Institute of Technology, Bengaluru, India.

Abstract: Security of data is important. Most desktop application users place a greater emphasis on the value of data and how it connects to their products. Security services for system must typically be offline. There is a chance that it could be hacked and altered if it is online. But everything is done online these days. As a result, we must strengthen data and file security. Many businesses develop internal applications to guard against data leaks. In order to stop data leaking, this document suggests a secure file sharing mechanism. One of the most crucial and sensitive pieces of information for IT organizations is source code. Source code breaches can result in tampering with the goods and services they provide, causing considerable financial losses for the company. In order to prevent the unauthorized sharing of sensitive data like source code, this system intends to create a secure in-file sharing method. For this, the system makes use of Principal Component Analysis.

Keywords: File Sharing, Prevention, Principal Component Analysis, Pattern matching, Anomalies

I. INTRODUCTION

Security is the first priority for any organization. The use of information technology is a major concern for businesses. Since most systems are online, there is a good chance that a data-stealing attack will take place. When sensitive information is made public, it may be misused and interfered with, which could lead to a decline in client satisfaction and a loss of revenue for a company. It is usually necessary to disclose sensitive information to a third party. Therefore, developing a secure sharing system is essential. Data leakers might also be inside staff members. It's crucial to monitor the information that employees are sharing as a result. One type of sensitive data for an IT company is source code. Therefore, only those with permission should have access to it. The intent of this study is to generate a system that tracks the data that employees share internally. The limitation of the existing method is that it is impossible to know what data or files are being transmitted. Principal Component Analysis is used in our system to determine the file's content. If the detection system discovers the sharing of a code file, the user will be prohibited. Even if the user copies the code to a text file in an attempt to share it, the system examines the text file's contents and halts the process if it detects any code. PCA, is a technique that enables you to more easily analyze and examine a smaller set of "summary indices" to summarize the information contained in massive data tables. The information can be statistical descriptions of the properties of production samples, chemical substances or reactions, process time points in a continuous process, batches from a batch process, biological people, or DOE-protocol trials. PCA is applicable in the case that follows.

PCA can be used when the input characteristics have extensive dimensions (for example, several variables).

- The PCA technique is useful for data analysis when features or variables are multi-collinear.
- **Denosing** and data reduction

In our system, the data set is handled as a file with a large number of lines. It is possible to check for code on each line separately when there are fewer lines. However, as the number of lines grows, it becomes difficult and time-consuming to examine each line for code. Therefore, we are using PCA to reduce a large number of lines into a small sample that contains all of the pertinent data. When a match is discovered, it then compares the sampling results with a set of specified keywords that are present in another data collection. It is seen as unauthorized file sharing that is occurring. A list of specified keywords located in another data set is then compared to the PCA result. When a match is found, it is considered that illegal file sharing is taking place. The proposed system is more accurate and effective than the current system.

II. RELATED WORK

Anomaly detection examines deviations from this "typical" behavior as evidence of fraudulent attacks by using models of how people and apps are expected to behave. This method is complementary to abuse detection, which compares a stream of audited events with a number of attack descriptions to look for signs that one of the modeled attacks is taking place.

Attack patterns diverge from typical activity, which is a fundamental premise behind anomaly identification. Anomaly detection also presupposes that this "difference" can be quantified. Numerous techniques, such as data mining for network traffic, statistical analysis for audit records, and sequence analysis for operating system calls, have been proposed to evaluate diverse data streams under these assumptions.

That determine the detection of parameters in the studied data are especially pertinent to the work presented here. The framework created by Lee et al., for instance, offers instructions on how to retrieve features that are helpful for creating intrusion classification models. The method makes use of labeled data to determine the ideal collection of features for malware detection.

The approach described in this study is comparable to Lee's in that it does link analysis and classification on the data using a group of selected features. The approach is distinct, since neither the features nor the detection thresholds are obtained from the labeling of assaults in the training data.

III. METHODOLOGY

A. Principal Component Analysis

A machine learning method called as PCA is used to minimize dimensionality. It uses an orthogonal transformation to change observations of correlated features into a set of linearly uncorrelated data. The newly changed features are the Principal Components. Typically, PCA looks for the surface with the lowest dimensionality onto which to project the high-dimensional data.

Significant features can be created by linearly reducing a large number of correlated variables into a smaller number of uncorrelated ones. The eigenvectors of the covariance/correlation matrix, also known as the principle components, are used to project (dot product) the original data into the smaller PCA space (PCs).

B. Implementation of PCA

Using PCA, the method separates the enormous dataset into more manageable chunks. There may be several lines in a file. It takes time to check each line, which lowers system performance. Using PCA, it is condensed to a smaller set while preserving the pertinent data in the dataset. The output of the PCA is contrasted with the predefined keywords. These keywords in programming languages are reserved. Without the reserved terms, a language cannot be written. It is assumed that a code file was attempted to be shared with an unauthorized party when these keywords turn up in PCA's reduced data. With the use of this technology, data leaks may be located and stopped while also ensuring data security.

IV. MODELLING AND ANALYSIS

SYSTEM ARCHITECTURE

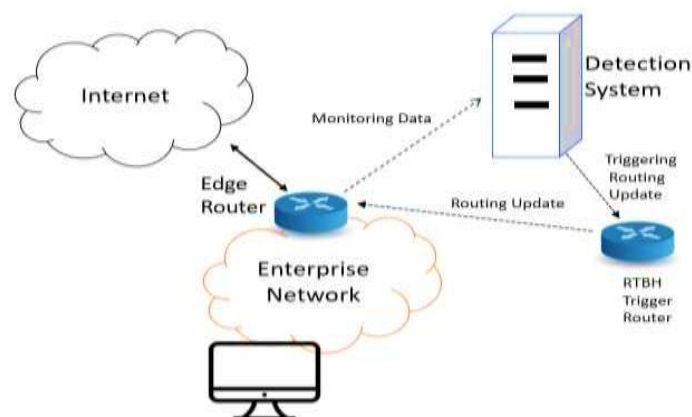


Fig 1: System Architecture

The system that monitors and detects any illegal file sharing is the detection and identification component of the above architectural design. When it does, it notifies the system administrator, who may then restrict the user's activity by blocking his IP, thereby effectively outlawing any form of sharing on the system.



DATA FLOW

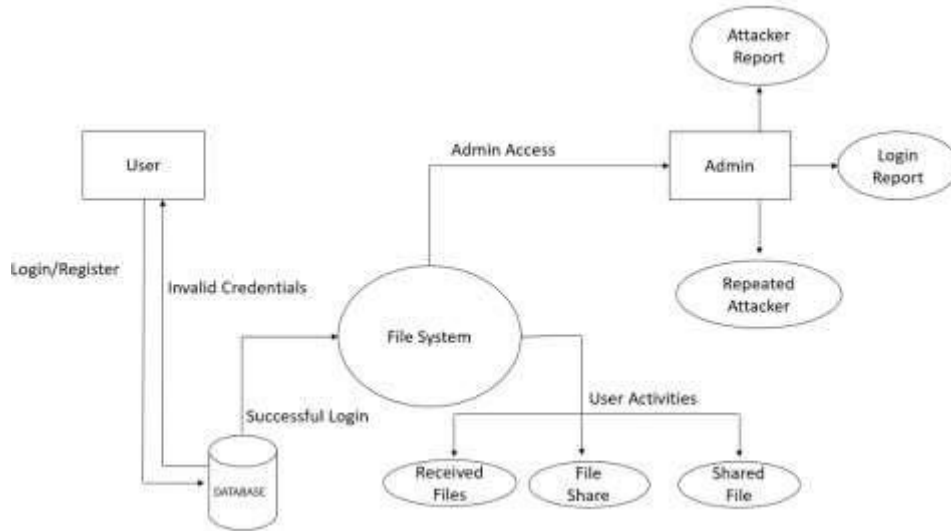


Fig 2. Data Flow Diagram

The proposed system has the following functionalities

The user must log in using the appropriate credentials in order to access the system. The user can share files with another user over the network after properly logging in. When a file is shared, it goes through a detection process that looks within to see if it contains any sensitive information. If any delicate information is discovered, the system alerts the administrator.

The administrator has access to information on the users who attempted to disclose sensitive information. The system administrator can then block his IP address to stop any such sharing in the future.

V. RESULT

A. User Registration Page



Fig 1: User Registration Page

This page displays the user registration page where the user can get registered themselves

B. User Dashboard



Fig 2. User Dashboard

The user's dashboard, where they can exchange files with other users and examine received files, will be displayed after a successful login..

C. File Share Page



Fig 3. File Share

On this page, the user can choose the file to share and the recipient with whom the file will be shared.

D. Shared Files



Share Member Name	File Name	File	Download File
anilkumar	Address	Chemical Industry Systems Pvt(Ltd) No.1, 75th Street, Ye-Flats, Avaloknagar, Chennai-63.	Download
anilkumar	addt	Chemical Industry Systems Pvt(Ltd) No.1, 75th Street, Ye-Flats, Avaloknagar, Chennai-63.	Download
anilkumar	addt		Download

Fig 4. List of shared files

The list of all shared files that the user has shared is shown on this page. This page contains details about the shared file, including the shared file's name, description, and a re-download option.

E. Admin Login

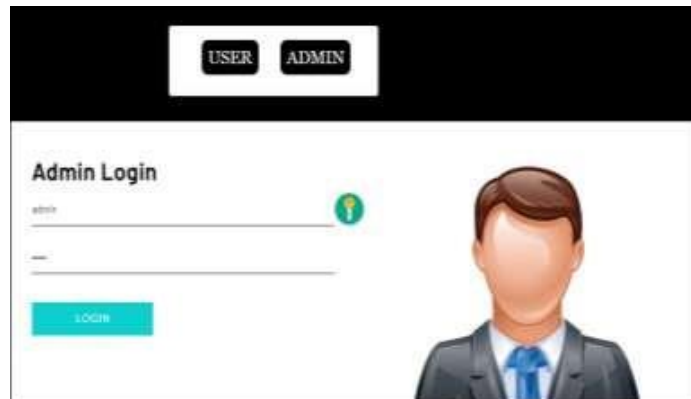


Fig 5. Admin Login Page

This page contains the login page for admin where the admin can login into system using proper credentials

F. Admin Dashboard

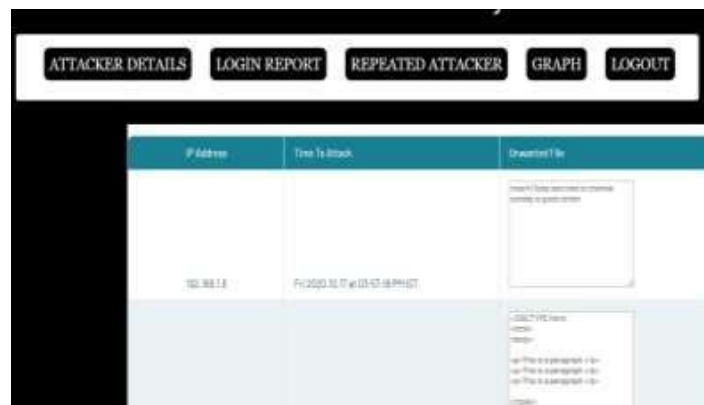


Fig 6. Admin Dashboard

On this page, the admin dashboard is visible. This dashboard offers a few choices. In the attacker details section, information is shown about the IP address from which a fraudulent file-sharing attempt was made.

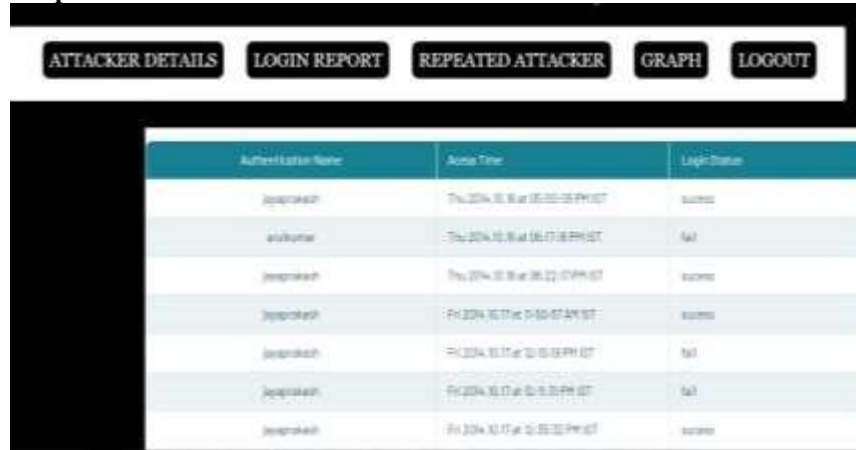
G. Block User



Fig 7. Option to block user

This page displays the details of the user who tried to share sensitive data multiple times and the admin has the option to block the user by blocking his/her IP address

H. Login Reports



The screenshot shows a web interface with a navigation bar containing buttons for 'ATTACKER DETAILS', 'LOGIN REPORT', 'REPEATED ATTACKER', 'GRAPH', and 'LOGOUT'. Below the navigation bar is a table with the following data:

Authenticator Name	Action Time	Login Status
jeppiaiah	Thu, 20th Jun 2022 09:59:07	Success
ankuram	Thu, 20th Jun 2022 09:59:07	Fail
jeppiaiah	Thu, 20th Jun 2022 09:59:07	Success
jeppiaiah	Fri, 22nd Jun 2022 09:59:07	Success
jeppiaiah	Fri, 22nd Jun 2022 09:59:07	Fail
jeppiaiah	Fri, 22nd Jun 2022 09:59:07	Fail
jeppiaiah	Fri, 22nd Jun 2022 09:59:07	Success

Fig 7. Displays list of users logged in This page displays the list of users who logged into the system

CONCLUSION

The mechanism for preventing the sharing of files containing sensitive information was suggested in this study. This technology aids in locating the source of data leakage and eliminating the source from the system, thereby guarding against the release or alteration of any sensitive information. The organization may benefit from the large-scale implementation of this project because it offers an easy solution to prevent data leaks. Principal Component Analysis implementation made it simpler to reduce the size of huge data sets, enabled quicker and more precise file detection, and decreased the overall processing time.

REFERENCES

1. G. Xie et al., "Fast low-rank matrix approximation with locality sensitive hashing for quick anomaly detection," in Proc. IEEE INFOCOM, May 2017, pp. 1–9.
2. K. Xie et al., "Fast tensor factorization for accurate Internet anomaly detection," IEEE/ACM Trans. Netw., vol. 25, no. 6, pp. 3794–3807, Dec. 2017,
3. H. Huang, H. Al-Azzawi, and H. Brani. (Feb. 2014). "Network traffic anomaly detection." [Online]. Available: <https://arxiv.org/abs/1402.0856>
4. D. Jiang, Z. Xu, P. Zhang, and T. Zhu, "A transform domain-based anomaly detection approach to network-wide traffic," J. Netw. Comput. Appl., vol. 40, no. 2, pp. 292–306, Apr. 2014.