PRESERVING SENSITIVE DATA BY DATA LEAKAGE PREVENTION USING ATTRIBUTE BASED ENCRYPTION ALGORITHM

P.Margathavalli\textsuperscript{1}, R.Manjula\textsuperscript{2}, V.Pramila\textsuperscript{3}, R.Priya\textsuperscript{4} & P.Abirami\textsuperscript{5}

\textsuperscript{1}Assistant Professor Department of IT, PEC
\textsuperscript{2}\textsuperscript{3}\textsuperscript{4}\textsuperscript{5} Student Member/IT, Pondicherry Engineering College

\texttt{marapriya@pec.edu}
\texttt{priyal2th41@pec.edu}
\texttt{manjula12th29@pec.edu}

Abstract--- To a large extent we are not quite aware of the real risks and how to avoid them. In most of the organizations and companies, the sensitive data include intellectual property (IP), financial information, patient information, bank accounts, credit-card passwords and other information depending on the particular organization and the industry. The serious obstacles for various companies occur due to data leakage as the number of incidents and the cost to those experiencing increases continuously. The common occurrence of data leakage are connected to our everyday lives -- the way we save, store and handle the data. The risk behind data leakage is that the sensitive information will be stolen by the unauthorized person. Hence it is very important to manage the risk of sensitive data exposure by detection and preventing it. As a result, organizations are forced to produce beyond the traditional models of securing the data and locking down specific parts of the IT infrastructure in order to formulate their data protection goals. Data leakage detection in the organization premises helps to detect the unaware transfer of intellectual property from owner to unauthorized person. The Detection process is carried using Fuzzy fingerprint method. The drawback of the detection process is that it only depicts whether the leakage happens or not. Overcoming the drawback, prevention method was proposed where we can keep the sensitive data locked. In this paper, we provide the data leakage prevention to prevent the sensitive data using Attribute based encryption algorithm (ABE). The advantage of this method is to prevent the organizational properties that are too sensitive from the unauthorized hands.

Keywords: Data security, leakage prevention, privacy, intellectual property, ABE, attributes.

I. INTRODUCTION

In 2014, over 90 percent of data leakage had been reported which would be prevented if the organization maintains a better strategies of data security. Data Leakage is defined as the transfer of data within an organization to the unauthorized destination. There are various causes of data leakage. They are:

\begin{itemize}
  \item Intentional action
  \item Unintentional action
  \item Failure
  \item Disaster
  \item Crime
\end{itemize}

Therefore, data leakage causes several problem to the organization as well as the way in which it occur have become more complex and numerous.

Data security is necessary to stop the issues of the data leakage. It keeps the protected access of data from the unauthorized users. A serious attacks to the Organizations and personal security is caused due to the sensitive data exposure [23]. Data leakage detection is the mechanism of screening the data in which undergoes the purpose of storage and transmission. There are various scanning algorithms for data leakage detection. The network data leak detection is the technique that performs inspection of deep packets and finds for the patterns of sensitive data and the efficiency is higher. It undergoes the analyses of each and every packets in the network for the sensitive data. When the sensitive data is found, alerts will be occurred. The undesirable requirement of data leakage detection is sensitive data in the form of plaintext. The exposure of sensitive data happen is the compromised detection system. The purpose of the data owner is to obtain data leakage detection without revealing the plaintext sensitive data to the providers. Hence, new methods of data leakage detection helps to screen the content without the reading the sensitive information [1] [2].

The data leakage detection method can be obtained and be moved in the semi-honest environment. We implemented the data leakage detection using fuzzy fingerprint method. It improves the security of the sensitive data. One way computation is followed on the sensitive data in the fast and practical manner. It enhances the data owner the event of content checking to the providers without showing the sensitive data. In this detection method, the DLD
(Data leakage detection) provider is formed as a sincere and unusual person. During inspection, he can only gather little information about the sensitive data. Internet Service Provider (ISP) plays a major role in the detection technique by making data leakage detection as an add-on service to all customers. In various other setting, each and every one can select their sensitive data and request for data leakage detection of their local network to the administrator. The detection method, the data owner obtains sensitive data in the form of fingerprints or set of digests. The data owner make known only limited data to the DLD provider. In the network traffic, the purpose of the DLD provider is to calculate the fingerprint and to indicate the data leak in them. The gathering of real knowledge by the DLD provider should be stopped. This can only be done by the data owner who scan the content send to the DLD provider and finds whether data leakage occur. This data leakage detection method should maintain continuous scanning of the content without allowing the sensitive data viewed by the DLD provider [1]. This data leakage detection method should maintain continuous scanning of the content without allowing the sensitive data viewed by the DLD provider. It is accomplished only when the data owner is alert all time.

Though the sensitive data is kept secure by data leakage detection but sometimes the data may leaked to the unauthorized users due to carelessness of the data owner. Hence it is very important to carry out the prevention technique. This allows to prevent the data leakage from the unauthorized users. In this paper, we design, implement and propose the prevention technique using Attribute based Encryption (ABE) Algorithm. The encryption of the sensitive data is more important as it was used in various websites on the Internet [3]. Attribute-based encryption is the public-key encryption where the secret key and the cipher-text are based on the attributes (e.g. the place he lives, or the kind of occupation he does). In such a situation, the cipher-text decryption is possible only if the attributes set of the user key matches the attributes of the cipher-text. The Attribute based encryption is based on collision-resistance. It holds multiple keys should be able to access data if at least one individual key has the grants access. Here the cipher-text is developed based on the attributes and the private key is associated with it. The private key is used to download the data if it matches with the attributes. This helps to prevent the sensitive data from the unauthorized users securely. No other unauthorized users cannot use the sensitive data unnecessarily. Therefore, this method enhances the data security to the greater rate.

II. RELATED WORK

We have made detailed analysis about the techniques that have been implemented so far in order to improve the security of the sensitive data in the organizations. The advancement in the privacy preserving have been developing day by day. The identification of similar spam messages in the collaborative setting was based on shingles in Rabin fingerprint. The unique technique of data leak detection is considered where the DLD provider is not fully trusted. The requirement of the privacy preserving does not appear in the virus scan. The additional privacy requirement makes the data leakage detection more challenging in the data security world. This requirement makes the amount of sensitive information gathered by the DLD provider limited. The partial digests is based on the sensitive data. The fuzzy fingerprint is new to the detection of data leakage and the Attribute based encryption is new to the data leakage prevention. Both this work prevent the sensitive data from the unauthorized users and enhances the security to the higher rate. The proposed system is the efficient solution that prevents the data leakage in the organization that makes the company to reach height in the economy.

There have been many advances in understanding the privacy needs [25] or the privacy requirement of security [26]. In this paper, we find out the privacy requirements in an outsourced data-leak detection service and provide a systematic solution to enable privacy-preserving DLD services. Shingle with Rabin fingerprint [15] was used previously for identifying similar spam messages in a collaborative setting [27], as well as collaborative worm containment [28], virus scan [29], and fragment detection [30]. In comparison, we tackle with the unique data-leak detection problem in an outsourced setting where the DLD provider is not fully trusted. Such privacy requirement does not exist in above models, e.g., the virus signatures are non-sensitive in the virus-scan paradigm [29]. We propose the fuzzy fingerprint approach to meet the special privacy requirement and present the first systematic solution to privacy-preserving data-leak detection with convincing results. Our fuzzy fingerprint method differs from these solutions and enables its adopter to provide data leak detection as a service. The customer or data owner does not need to fully trust the DLD provider using our approach. Bloom filter [19] is a space-saving data structure for set membership test, and it is used in network security from network layer [34] to application layer [35].

The fuzzy Bloom filter invented in [36] constructs a special Bloom filter that probabilistically sets the corresponding filter bits to 1’s. Although it is designed to support a resource-sufficient routing scheme, it is a potential privacy-preserving technique. We do not invent a variant of Bloom filter for our fuzzy fingerprint, and our fuzzification process is separate from membership test. The advantage of separating fingerprint fuzzification from membership test is that it is flexible to test whether the fingerprint is sensitive with or without fuzzification.

Besides fingerprint-based detection, other approaches can be applied to data-leak detection. If the sensitive data size is small and the patterns of all sensitive data are enumerable, string matching [37], [38] in network intrusion detection system (such as SNORT [39] or Bro [40]) can be used to detect data leaks. Privacy-preserving keyword search [41] or fuzzy keyword search [42] provide string matching approaches in semi-honest environments, but keywords usually do not cover enough sensitive data segments for data-leak detection. Anomaly detection in network traffic can be used to detect data leaks. [5] detect any substantial increase in
the amount of new information in the traffic, and entropy analysis is used in [43]. We present a signature-based model to detect data leaks and focus on the design that can be outsourced, thus the two approaches are different.

Another category of approaches for data-leak detection is tracing and enforcing the sensitive data flows. The approaches include data flow and taint analysis [6], legal flow marking [44], and file-descriptor sharing enforcement [8]. These approaches are different from ours because they do not aim to provide a remote service. However, pure network-based solution cannot handle maliciously encrypted traffic [45], and these methods are complementary to our approach in detecting different forms (e.g., encrypted) of data leaks.

Besides our fuzzy fingerprint solution for data-leak detection, there are other privacy-preserving techniques invented for specific processes, e.g., DNA matching [46], or for general purpose use, e.g., secure multi-party computation (SMC). Similar to string matching methods discussed above, [46] uses anonymous automata to perform comparison. SMC [47] is a cryptographic mechanism, which supports a wide range of fundamental arithmetic, set, and string operations as well as complex functions such as knapsack computation [48], automated trouble-shooting [49], network event statistics [50], private information retrieval [52] genomic computation [53], private database query [54], private join operations [55], and distributed data mining [56]. The provable privacy guarantees offered by SMC comes at a cost in terms of computational complexity and realization difficulty. The advantage of our approach is its concision and efficiency. The comparison of the data leakage detection algorithms are shown below:

<table>
<thead>
<tr>
<th>List of Papers</th>
<th>Techniques and algorithms used in</th>
<th>Merits</th>
<th>Demerits</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data leak detection as a service</td>
<td>Fuzzy fingerprint</td>
<td>the sensitive data from the DLP provider, yet they do not cause additional false alarms for the data owner, as it can quickly distinguish true and false leak instances.</td>
<td>the exposure of the sensitive data is kept minimum during the detection</td>
</tr>
<tr>
<td>risk-based security</td>
<td>ABC algorithm</td>
<td>This approach makes easier to understand and observe the level of risk. Methods of calculation are simple to understand and implement.</td>
<td>Without an automatic tool the process can be really difficult to implement. There are no standards and universally accepted information for</td>
</tr>
<tr>
<td>Cost of Data Breach Study: Global Analysis</td>
<td>genetic algorithm</td>
<td>Security level is better determined based on the three elements: availability, integrity and confidentiality</td>
<td>Estimated cost of the measure that should be implemented are not calculated. The most important areas of risk are evaluated.</td>
</tr>
<tr>
<td>Discover Sensitive Data Prevent Breaches DLP Data Loss Prevention</td>
<td>String matching algorithms</td>
<td>Management performance can be closely watched. Data accuracy improves as the organization gains Experience.</td>
<td>The performance of risk management are hard to follow because of their subjectivity</td>
</tr>
<tr>
<td>Quantifying information leaks in outbound web traffic</td>
<td>precise algorithm</td>
<td>The values of risk impacts are based on subjective opinions of people involved. The process handles a long time</td>
<td>The evaluation of risk and its result are subjective. It is possible that the reality is not defined correctly because of the subjective perspective of the author</td>
</tr>
<tr>
<td>Panorama: Capturing system-wide information flow for malware detection and analysis</td>
<td>Floyd's cycle-finding algorithm</td>
<td>By storing your data online you are reducing the burden of your hard disk, which means you are eventually saving disk space.</td>
<td>Improper handing can cause trouble: You must need your user-id and password safe to protect your data as if someone knows or even guess your credentials.</td>
</tr>
<tr>
<td>Protecting Confidential Data on Personal Computers with Storage Capsules</td>
<td>VM (virtual Machine) methods</td>
<td>Easy Communicatio n and Speed. Ability to Share Files, Data and Information</td>
<td>Breakdowns and Possible Loss of Resources.</td>
</tr>
<tr>
<td>Preventing accidental data disclosure in modern operating systems</td>
<td>Karn's algorithm</td>
<td>Hacking attempts or virus / spyware attacks from the internet will not be able to harm physical computers. External possible attacks are prevented.</td>
<td>networks causing losses of thousands of dollars each year</td>
</tr>
</tbody>
</table>
Revolver: An automated approach to the detection of evasive web-based malware,

Backpressur e routing

Files can be stored on a central computer (the file server) allowing data to be shared throughout an organisation.

There might be illegal activities that will occur that you need to be aware and be careful all the time.

"Gyrus: A framework for user-intent monitoring of text-based networked applications

UI Automation methods

Files can be backed up more easily when they are all on a central fileserver rather than when they are scattered across a number of independent workstations.

Lack of Independence

Fingerprinting by random polynomials

Rabin fingerprint and string matching technique

Networks also allow security to be established, ensuring that the network users may only have access to certain files and applications.

Risk of Theft or Vandalism. Corruptable Storage Device or Media

Min-wise independent permutations

AltaVista web index software

This approach is definitely very convenient. The user gets the opportunity to reliably and quickly (using standard means) limit access to confidential information for other household members or colleagues who also use the computer.

, if the password is lost or reset, or if the operating system fails or is reinstalled, it becomes impossible to gain access to the EFS-encrypted files on the drive

Table 2.1 Comparison of data leakage detection techniques

III. PROPOSED WORK

In the existing work, the sensitive data leakage in the organization is detected using the Fuzzy fingerprint method which has more obstacles in the data security. The detection process does only the work of whether the leakage occur or not. This is very simple process and sensitive data can be leaked easily to the unauthorized users due to the carelessness of the data owner which reduces the economy of the organization. Hence, in order to overcome this disorder it is very essential to prevent the data from the leakage. This helps to prevent the data leakage from the unauthorized users. There are various algorithm for preventing the data leakage in the organization. These algorithms are unique in its features and are used according to their requirements of the organization to prevent the sensitive data. In this paper, we design, implement and propose the prevention technique using Attribute based Encryption (ABE) Algorithm. The encryption of the sensitive data is more important as it was used in various websites on the Internet. Attribute-based encryption is a type of public-key encryption in which the secret key of a user and the cipher-text are dependent upon attributes (e.g. the country he lives, or the kind of subscription he has). In such a system, the decryption of a cipher-text is possible only if the set of attributes of the user key matches the attributes of the cipher-text. A crucial security aspect of Attribute-Based Encryption is collusion-resistance. An adversary that holds multiple keys should only be able to access data if at least one individual key grants access. Here the cipher-text is developed based on the attributes and the private key is associated with it. The private key is used to download the data if it matches with the attributes. This helps to prevent the sensitive data from the unauthorized users securely. No other unauthorized users cannot use the sensitive data unnecessarily. Therefore, this method enhances the data security to the greater rate.

Recently, more attention has been attracted by a new public key primitive called Attribute-based encryption (ABE). ABE has essential advantage over the traditional PKC primitives as it achieves flexible one-to-many encryption instead of one-to-one. ABE is envisioned as an important tool for addressing the problem of secure and sensitive data sharing and access control. In an ABE system, a user is identified by a set of attributes. A secret key based on a set of attributes ω, can decrypt a cipher-text encrypted with a public key based on a set of attributes ω′ only if the sets ω and ω′ overlap sufficiently as determined by a threshold value t. A party could encrypt a document to all users who have certain set of attributes drawn from a pre-defined attribute universe. For example, one can encrypt a blood group wanted document to all donors of that specific blood group from a particular locality of specific age group. In this case the document would be encrypted to the attribute subset {“B+”, “Karaikudi”, “Age 20-25"}, and only users with all of these three attributes in the blood bank can hold
the corresponding private keys and thus decrypt the document, while others cannot. There are two variants of ABE: Key-Policy based ABE (KP-ABE) and cipher-text Policy based ABE (CP-ABE) [1,2,3,4].

In KP-ABE, the cipher-text is associated with a set of attributes and the secret key is associated with the access policy. The encryptor defines the set of descriptive attributes necessary to decrypt the cipher-text. The trusted authority who generates user’s secret key defines the combination of attributes for which the secret key can be used. In CP-ABE, the idea is reversed: now the cipher-text is associated with the access policy and the encrypting party determines the policy under which the data can be decrypted, while the secret key is associated with a set of attributes. Besides, there is an increasing need to protect user privacy in today’s access control systems. In some critical circumstances, the access policy itself could be sensitive information. Therefore, we propose an attribute – based encryption scheme where encryptor specified access policies are hidden. Even the legitimate decryptor cannot obtain the information about the access policy associated with the encrypted data more than the fact that she can decrypt the data.

In its key-policy, the primitive enables senders to encrypt messages under a set of attributes and private keys are associated with access structures that specify which cipher-texts the key holder will be allowed to decrypt. In most ABE systems, the cipher-text size grows linearly with the number of cipher-text attributes and the only known exceptions only support restricted forms of threshold access policies. This paper proposes the key-policy attribute-based encryption (KP-ABE) schemes allowing for non-monotonic access structures (i.e., that may contain negated attributes) and with constant cipher-text size. Towards achieving this goal, we show that a certain class of identity-based broadcast encryption schemes generically yields monotonic KP-ABE systems in the selective set model. We then describe a new identity-based revocation mechanism that, when combined with a particular instantiation of our general monotonic construction, gives rise to the truly expressive KP-ABE realization with constant-size cipher-texts. The downside of these new constructions is that private keys have quadratic size in the number of attributes. On the other hand, they reduce the number of pairing evaluations to a constant, which appears to be a unique feature among expressive KP-ABE schemes.

There is a trend for sensitive user data to be stored by third parties on the Internet. For example, personal email, data, and personal preferences are stored on web portal sites such as Google and Yahoo. The attack correlation center, dshield.org, presents aggregated views of attacks on the Internet, but stores intrusion reports individually submitted by users. Given the variety, amount, and importance of information stored at these sites, there is cause for concern that personal data will be compromised. This worry is escalated by the surge in recent attacks and legal pressure faced by such services. One method for alleviating some of these problems is to store data in encrypted form. Thus, if the storage is compromised the amount of information loss will be limited. One disadvantage of encrypting data is that it severely limits the ability of users to selectively share their encrypted data at a fine-grained level. Suppose a particular user wants to grant decryption access to a party to all of its Internet traffic logs for all entries on a particular range of dates that had a source IP address from a particular subnet. The user either needs to act as an intermediary and decrypt all relevant entries for the party or must give the party its private decryption key, and thus let it have access to all entries. Neither one of these options is particularly appealing. In an ABE system, a user’s keys and cipher-texts are labeled with sets of descriptive attributes and a particular key can decrypt a particular cipher-text only if there is a match between the attributes of the cipher-text and the user’s key. While this primitive was shown to be useful for error-tolerant encryption with biometrics, the lack of expression seems to limit its applicability to larger systems. Therefore, the attribute based encryption makes the data security well versed and increases the economy by keeping the sensitive data away from the unauthorized person. The Attribute based encryption makes the prevention process simpler and enhances the security of the sensitive digests of the organization. The prevention process using ABE makes the security level of the sensitive data rises to the higher rate. We develop a much richer type of attribute-based encryption cryptosystem and demonstrate its applications.

In our system each cipher-text is labeled by the encryptor with a set of descriptive attributes. Each private key is associated with an access structure that specifies which type of cipher-texts the key can decrypt. We call such a scheme a Key-Policy Attribute-Based Encryption (KPABE), since the access structure is specified in the private key, while the cipher-texts are simply labeled with a set of descriptive attributes. We note that this setting is reminiscent of secret sharing schemes (see, e.g., [3]). Using known techniques one can build a secret-sharing scheme that specifies that a set of parties must cooperate in order to reconstruct a secret. For example, one can specify a tree access structure where the interior nodes consist of AND & OR gates and the leaves consist of different parties.
Any set of parties that satisfy the tree can reconstruct the secret. In our construction each user’s key is associated with a tree-access structure where the leaves are associated with attributes. A user is able to decrypt a cipher-text if the attributes associated with a cipher-text satisfy the key’s access structure. Therefore, the disadvantages of the data leakage detection is overcome by the prevention technique Attribute based encryption.

IV. IMPLEMENTATIONS AND RESULTS

We have implemented the data leakage prevention technique using Attribute based encryption algorithm in which the technique is based on the attributes of the sensitive data. When the data owner send the data to the receiver, a secret key is generated based on the attributes of the sensitive data. The secret key helps to download the data sent by the data owner. This prevents the data leakage from the unauthorized users. This improves the data security of the organization. If any unauthorized users, tries to access the sensitive data an alert message in the form of cipher-text is sent to the data owner. Proposed system consists of 4 phases, Setup Phase, Key Generation Phase, Encryption Phase and Decryption Phase. This phases does the data leakage prevention in step by step process.

V. CONCLUSION

We proposed Attribute based encryption method, the data leakage prevention method. Attribute-based encryption is a type of public-key encryption in which the secret key of a user and the cipher-text are dependent upon attributes (e.g. the country he lives, or the kind of subscription he has). In such a system, the decryption of a cipher-text is possible only if the set of attributes of the user key matches the attributes of the cipher-text. After matching the attributes of the cipher-text, the secret key is generated which helps to download the file which was sent secretly by the data owner. Hence, it maintains the privacy between the data owner and the data receiver. Therefore, the data leakage prevention is finally done using the ABE method which facilitates the data security in the organization. Finally, the obstacles in the data security is overcome in this paper by implementing the data leakage prevention. Hence, this changes the entire data security world to reach a greater extent and used for data transmission in various fields. This also improves the economy of the organization by preventing the sensitive data of one organization from the other. Therefore, the data leakage prevention method should be followed and maintained in each and every organization to stay unique in their features from the other and also to improve the economy.

REFERENCES


[15] Benoit Libert, Ellie de Pinafelt, Nattapon Gattrapadung researching center japan 2009 Expressing key-policy attribute based encryption with constant-size cipher texts vol94, no.5 pp 678-987


