RESEARCH ARTICLE OPEN ACCESS

An Improved Method for Preventing Data Leakage in an Organization

Neetu Bose[1], Dr. N. Vishwanath[2]
[1]*M.Tech Scholar, CSE, Toc H Institute Of Science & Technology Ernakulam,India.Pin-682313*
[2]*Professor, CSE, Toc H Institute Of Science & Technology Ernakulam,India.Pin-682313*

ABSTRACT
Data is one of the most important assets an organisation has since it denes each organisations unique- ness.It includes data on members and prospects, their inter- ests and purchases, your events, speakers, your content, social media, press, your staff, budget, strategic plan, and much more. As organizations open their doors to employees, part- ners, customers and suppliers to provide deeper access to sensitive information, the risk sassociated with business increase. Now, more than ever, within creasing threats of cyber terrorism, cor- porate governance issues, fraud, and identity theft, the need for securing corporate information has become paramount. Informa- tion theft is not just about external hackers and unauthorized external users stealing your data, it is also about managing internal employees and even contractors who may be working within your organization for short periods of time. Adding to the challenge of securing information is the increasing push for corporate governance and adherence to legislative or regulatory requirements. Failure to comply and provide privacy, audit and internal controls could result in penalties ranging from large nes to jail terms. Non-compliance can result in not only potential implications for executives, but also possible threats to the viability of a corporation. Insiders too represent a sign cant risk to data security. The task of detecting malicious insiders is very challenging as the methods of deception become more and more sophisticated. There are various solutions present to avoid data leakage. Data leakage detection, prevention (DLPM) and monitoring solutions became an inherent component of the organizations security suite.DLP solutions monitors sensitive data when at rest, in motion, or in use and enforce the organizational data protection policy.These solutions focus mainly on the data and its sensitivity level, and on preventing it from reaching an unauthorized person. They ignore the fact that an insider is gradually exposed to more and more sensitive data,to which she is authorized to access. Such data may cause great damage to the organization when leaked or misused. Data can be leaked via emails, instant messaging, le transfer etc. This research is focusing on email data leakage monitoring, detection and prevention. It is proposed to be carried out in two phases: leakage detection through mining and prevention through encryption of email content.

***Keywords*-**Data misuse, Insider threat, Text analysis, Security measures, Misuseability

I. INTRODUCTION

Data leakage is dened as the accidental or unintentional distribution of private or sensitive data to an unauthorized entity. Sensitive data in companies and organizations include intellectual property (IP), nancial information, patient infor- mation, personal credit-card data, and other information depending on the business and the industry. Data leakage poses a serious issue for companies as the number of incidents and the cost to those experiencing them continue to increase. Data leakage is enhanced by the fact that transmitted data (both inbound and out- bound), including emails, instant messaging, website forms, and le transfers among others, are largely unregulated and unmonitored on their way to their destinations. The potential damage and adverse consequences of a data leakage incident can be classied into two categories: Direct and Indirect Losses. Direct losses refer to tangible damage that

Is easy to measure or to estimate quantitatively. Indirect losses, on the other hand, are much harder to quantify and have a much broader impact in terms of cost, place, and time. Direct losses include violations of regulations such as those protecting customer privacy resulting in nes, settlements or customer compensation fees litigation involving lawsuits loss of future sales costs of investigation and remedial or restoration fees. Indirect losses include reduced share price as a result of negative publicity damage to a companys goodwill and reputation customer abandonment and exposure of intellectual property such as business plans, code, nancial reports, and meet- ing agendas to competitors.

Data leakage is an error condition in information systems in which information is destroyed by failures or neglect in storage, transmission, or processing. Information systems im- plement backup and disaster recovery

equipment and processes to prevent data loss or restore lost data. Data leakage is distinguished from data unavailability, such as that arise from a network outage.

While attacks on computers by outside intruders are more publicized, attacks perpetrated by insiders are very common and often more damaging. Insiders represent the greatest threat to computer security because they understand their organizations business and how their computer systems work. They have both the condentiality and access to perform these attacks. An inside attacker will have a higher probability of successfully breaking into the system and extracting critical information. The insiders also represent the greatest challenge to securing the company network because they are authorized a level of access to the le system and granted a degree of trust. The task of detecting malicious insiders is very chal- lenging as the methods of deception become more and more sophisticated. According to the 2010 Cyber Security Watch Survey 26 percent of the cyber-security events, recorded in a 12-month period, were caused by insiders. These insiders were
the most damaging with 43 percent of the respondentsreporting that their organization suffered data loss. Of the attacks, 16 percent were caused by theft of sensitive data and 15 percent by exposure of condential data. Email is one of the most popular forms of communication nowadays, mainly due to its efciency, low cost, and compatibility of diversied types of information. An insider can leak the condential data through email. There is a need of mining the email and nding a way to detect what potential damage it can cause to an organization if leaked.

II. RELATED WORKS EARLIER DONE WORKS INCLUDES

A. Monitoring System Call Activity

One approach of Detecting insider misbehavior is to mon- itor system call activity and watch for danger signs or unusual behavior. Nam Nguyen and Peter Reiher, Geoffrey H.Kuenning [1] describes an experimental system designed to test this approach. They tested the systems ability to detect common insider misbehavior by examining le system and process- related system calls. Their results shows that this approach can detect many such activities.In this paper, they want to look at these raw data in a different manner: the relationships between users and les, users and processes, and processes and les. The work described here indicates some promising directions for detecting misbehavior by insiders,as well as intruders whose initial penetration has gone unnoticed. The patterns of le accesses by many programs are sufciently regular that attackers trying to misuse them will quickly be noticed. Similarly, process-calling behavior is sufciently regular for large classes of programs to serve as a good indicator of misbehavior. Other patterns of le systems access and process behavior do not appear to be good candidates for detecting attacks. Individual users have too much variability in their access patterns to allow simple statistical methods to detect suspicious changes in behavior, without also triggering alerts in many innocuous situations. Similarly, some processes, by their nature, tend to fork a wide variety of other processes. Thus, a system that merely implemented the effective tools discussed in this paper would not catch all attacks. However, these techniques do appear to be useful candidates to include in a system that monitors computers for possible misbehavior. A system that merely implemented the effective tools discussed in this paper would not catch all attacks. However, these techniques do appear to be useful candidates to include in a system that monitors computers for possible misbehavior. The experience with the data gathering and monitoring steps shows that the necessary data can be gathered and threats checked without causing noticeable delays to the user. There are challenges to collecting and managing the large quantities of data necessary to build good models, but these challenges can be overcome.

B. Document Control System

Another approach for mitigating the risk of an unauthorized person accessing sensitive document is the implementation of mandatory access control (MAC) models and frameworks. The MAC mechanism regulates user access to data accord- ing to predened classications of both the user (the subject) and the documents (the object). The classication is based on partially ordered access classes (e.g., top secret, secret, condential,unclassied).Both the objects and the subjects are labeled with one of these classes, and the permission is granted by comparing the subject access class to that of the ob- ject in question. JunghoEom ,NamukKim,Sunghwan Kim and TaiMyoungChung [2] proposed Document Control system.In this paper,they architected a document control system for monitoring leakage of important documents related to military information. The proposed system inspects all documents when they are downloaded and sent. It consists of 3 modules; au- thentication module, access control module and watermarking module. The authentication module checks insider information for allow to log on system. The access control module control access authorization to do operations by insiders according to their role and security level. The

watermarking module is used to track transmission path of documents. The document control system controls illegal information ow by insiders and does not allow access to documents which are not related to the insiders duties.

C. Inserting Fake Objects(Perturbation

S.W. Ahmad Dr G.R.Bamnote [3] developed a model for assessing the guilt of agents. They also present algorithms for distributing objects to agents, in a way that improves the chances of identifying a leaker. Finally,they considered the option of adding fake objects to the distributed set. Such objects do not correspond to real entities but appear realistic to the agents. In a sense, the fake objects acts as a type of watermark for the entire set, without modifying any individual members. Perturbation is a very useful technique where the data is modied and made less sensitive before being handed to agents.

D. Misuse Detection In Databases

In recent years, several methods have been proposed for mitigating data leakage and data misuse in database systems. These methods can generally be classied as syntax-centric or data-centric. The syntax-centric approach relies on the SQLex- pression representation of queries to construct user proles. For example, a query can be represented by a vector of features extracted from the SQL statement, such as the query type (e.g., SELECT or INSERT), and the tables or attributes requested by the query. Celikelet [4] presented a model for risk management in distributed database systems. The model is used to measure the risk poses by a user in order to prevent her from misusing or abusing her role privileges. In the model, a Risk Priority Number (RPN) is calculated for each user, which is the product of the Occurrence Rating (OR) that reects the number of times the same query was issued with respect to the other users in the same role; the Severity Rating (SR) that measures the risk by referring to the quality of the data the user might get from the queries she issued; and the Detection Rating (DR) indicates how close the behavior of the user is to the behavior of users in other roles. Another syntax-centric method is the framework to enforce access control over data streams that dene a set of secure actions (e.g., secure join) that replaces any unsecure action (e.g., join) the user makes. When a user issues an unsecure action, the appropriate secure action is used instead, and by addressing the user permissions, retrieves only data that this user is eligible to see. The data-centric approach focuses on what the user is trying to accessinstead of how she expresses it. With this approach, an action is modeled by extracting features from the obtained result-set. Since they are dealing with data leakage, they assume that analyzing what a user sees (i.e., the resultsets) can provide a more direct indication of a possible data misuse. An interesting work presents a data-centric approach and considers a querys expression syntax as irrelevant for discerning user intent; only the resulting data matters. For every access to a database, a statistical vector (S-Vector) is created, holding various statisti- cal details on the result-set data, such as minimum, maximum, and average for numeric attributes, or counts of the different values for text attributes. Evaluation results showed that the S-Vector signicantly outperforms the syntax centric approach presented in Yaseen and Panda also proposed a data-centric method that uses dependency graphs based on domain expert knowledge. These graphs are used inorder to predict the ability of a user to infer sensitive information that might harm the organization using information she already obtained. Then, utilizing dependency graphs, the system prevents unauthorized users from gaining information that enables them to infer or calculate restricted data they are not eligible to have. Closely related to this line of works are the preventive approaches. The insider prediction tool uses taxonomy of insider threats to calculate the Evaluated Potential Threat (EPT) measure. This measure tries to estimate whether a users action is correlated with a part of the taxonomy that is labeled as malicious. The EPT is calculated by considering features describing the user, the context of the action and the action itself. In addition, the tool uses a set of malicious actions that were previously discovered. To prevent insiders from misusing their privileges, Bishop and Gates suggested the Group-Based Access Control (GBAC) mechanism, which is a generalization of RBAC mechanism. This mechanism uses, in addition to the users basic job description(role),the user characteristics and behavioral attributes such as the times he normally comes to work or the customers with whom she usually interacts. As already mentioned, none of the proposed methods consider the sensitivity level of the data to which the user may be exposed. This factor can greatly impact the outcome when trying to estimate the potential damage to the organization if the data is leaked or misused. Consequently, they adopted the data-centric approachthe data retrieved by a user action is examined and its sensitivity level computed.

III. PROPOSED SYSTEM

Collection of large amount of data and managing it is a challenge in method of monitoring system calls, also embed- ding a unique code in each

distributed copy can be sometimes destroyed if the recipient is malicious. Watermarking system a simple technique of data authorization, there are various software which can remove the watermark from the data and makes the data as original. There is a need of applying a better concept of misuseability weight measure which can serve as complementary to present data leakage detection systems. The existing solutions for detecting data leakage focuses mostly on restricting outsiders and to the databases. Measuring potential damage caused in case of a document is leaked in an organisation is a complementary method which can be embedded with existing solutions present for data

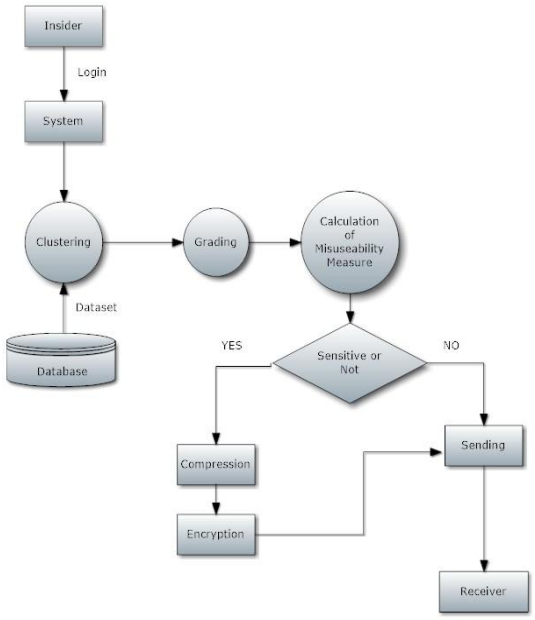

Fig. 1. Data leakage prevention Architecture

Leakage preventions. The proposed system focus on email data leakage. Constitutes email mining and encrypting the sensitive mails. It consists of two phases, first phase involves clustering and classifying the documents whereas second phase involves maintaining data integrity and encrypting and compressing of the emails.

Advantage

- Applying mining to email content will give better results.
- Checking data integrity.
- Encrypting and compressing the sensitive documents will restrict the leakage.

System Overview

The first phase consist of mining the emails. The dataset is used to create the training set. Dataset consists of emails from various employees in an organisation. Clustering of the documents, grading and classifying it to find a score i.e. misuseability weight measure. The second phase consist of checking the data integrity and encrypting and compressing the sensitive documents.

IV. HIGH LEVEL DESIGN

Data leakage prevention Architecture constitute the data leakage prevention system. User login into the system and gets the access to use it. Already existing email dataset is used to train the system. Mining is used here for classifying the email documents. Focus here is to classify the documents into sensi- tive and non-sensitive data. The emails are clustered first using self-organising mapping which is neural network based cluster- ing. Once the documents are clustered it can be graded now. Grading is done using TFIDF (Term Frequency and Inverse Document Frequency).Where using a trained dataset TFIDF is used to grade the documents into sensitive andnon-sensitive documents by finding the similarity between the documents. After grading the documents, classification procedure is carried out using TM score. If document is found to be sensitive it goes through compression and encryption. Compression of the email is carried out using G zip compression. Encryption is done using RSA. The procedure at receiver side is decryption of RSA.

Also before sending the mail, hash the document content and will store it in database. At the receiving end, again the document content is hashed, and the new hashed value and the hashed value in the database is compared. If both are same, concludes that the data is untouched. Or if the hashed values are different, the data integrity is lost.

V. LOW LEVEL DESIGN THE MODULES INCLUDED IN THE SYSTEM ARE:

A. Clustering

1) *K-means:* K-means algorithm features quick clustering and easy operation, and is applied to the cluster analysis of several data such as Texts, images and others; but this algorithm tends to terminate iterative process quickly to only obtain a partial optimal results, and fluctuate the clustering result because of random selection of the initial iterative center point. Due to the fact that the clustering is often applied to the data of the cluster quality the end-users cant judge and this fluctuation is difficult to be accepted in the application, it is of great significance to improve the quality and stability of clustering results in the analysis of the Text cluster. For K-means problems

concerning the selection of initial point and the sensitivity of isolated point in the Text clustering, combining the advantages of SOM and K-means algorithms will enhance the stability and quality of the Text clustering of the algorithm.

Steps for K-means clustering algorithm are:

(a) Select n objects as the initial cluster seeds on principle;
(b) Repeat (c) and (d) until no change in each cluster;
(c) Reassign each object to the most similar cluster in terms of the value of the cluster seeds;
(d) Update the cluster seeds, i.e., recompute the mean value of the object in each cluster, and take the mean value points of the objects as new cluster seeds.

2) *Self organising map:* Clustering the documents will provide a better picture of the data present.Aim is to classify the documents into sensitive and non sensitive data. The emails are clustered using self- organizing mapping which is neural network. A self-organizing map (SOM) is a type of artificial neural network (ANN) that is trained using unsupervised learn- ing to produce a low- dimensional (typically two-dimensional), discretized represen- tation of the input space of the training samples, and called a map. A self-organizing map consists of components called nodes or neurons. Associated with each node is a weight vector of the same dimension as the input data vectors, and a position in the map space.It describes a mapping from a higher-dimensional input space to a lower- dimensional map space. The procedure for placing a vector from data space onto the map is to find the node with the closest (smallest distance metric) weight vector to the data space vector.

Algorithm 1: Self Organizing MAP

1: Randomize the map's nodes' weight vectors;
2: Traverse each input vector in the input data set;
3: **while** Traverse each node in the map **do**
4: Use the Euclidean distance formula to find the similarity between the input vector and the map's node's weight vector;
5: Track the node that produces the smallest distance (this node is the best matching unit, BMU);
6: **end while**
7: Update the nodes in the neighborhood of the BMU (including the BMU itself) by pulling them closer to the input vector;

$$Wv(s + 1) = Wv(s) + (u, v, s)(s)(D(t) - Wv(s))$$

8: **while** $s < \lambda$ **do**
9: Increase s;
10: repeat from step 2
11: **end while**

Where,

- *s* is the current iteration
- λ is the iteration limit
- *t* is the index of the target input data vector in the input data set
- D(*t*) is a target input data vector
- *v* is the index of the node in the map
- W_V is the current weight vector of node v
- is the index of the best matching unit (BMU) in the map is a restraint due to distance from BMU, usually called the neighborhood function
- A is a learning restraint due to iteration progress.

3) Combination of SOM and K-means*:* The main process to cluster the texts by using the clustering combination algo- rithm of SOM and K-means is below:

Firstly apply the commonly used vector spatial model to represent the text information, delete the unusable words with the conventional method, and simplify the characteristic items according to TF-DF rules to obtain the texts characteristic set, secondly calculate the weights of various characteristic items and express the text in the form of vectors, thirdly input the vectors of the text set for SOM algorithm and cluster the texts through SOM (the number of SOM networks input nodes equals to the dimension of the text vectors, while the number of SOM networks output nodes equals to the number of the texts categories) to obtain a group of output weights, and finally initialize K-means algorithms cluster centers with this group of weights and implement K-means algorithm to cluster the text sets.

B. Grading

Grading of the documents is done using TFIDF (Term frequency inverse document frequency). It is a statistical measure used to evaluate how important a word is to a document in a collection or corpus. The importance increases proportionally to the number of times a word appears in the document but is offset by the frequency of the word in the corpus.

The tf-idf weight is composed by two terms: the first com- putes the normalized Term Frequency (TF), aka. The number of times a word appears in a document, divided by the total number of words in that document; the second term is the Inverse Document Frequency (IDF), computed as the logarithm of the number of the documents in the corpus divided by the number of documents where the specific term appears.

- TF: Term Frequency, which measures how frequently a term occurs in a documen t. Since every document is different in length, it is possible that a term would appear much more times in long documents than shorter ones. Thus, the term frequency is often divided by the document length (aka. the total number of terms in the document) as a way of normalization:
  TF(t) = (Number of times term t appears in a docu- ment) / (Total number of terms in the document).
- IDF: Inverse Document Frequency, which measures how important a term is. While computing TF, all terms are considered equally important. However it is known that certain terms, such as ”is”, ”of”, and ”that”, may appear a lot of times but have little importance. Thus we need to weigh down the frequent terms while scale up the rare ones, by computing the following:
  IDF(t) = log_e(Total number of documents / Number of documents with term t in it).

Assigning the misuseability score requires the assistance of a domain expert. Since the number of documents can be very high, it would be infeasible to have the domain expert to assign the score to each and every document. Therefore, a procedure is presented in which intelligently select a reasonable amount of documents that are presented and manually graded by the domain expert. Then, by modelling the experts rankings apply the model on the remaining documents and estimate their individual misuseability score.

Domain expert ranking:

Expert need to answer this in one of the option given. If leaked, how much damage would the document can cause the company?

[1 - None/ 2 - Not Signicant/ 3 - Signicant/ 4 - Very Signicant/ 5 - Critical]

Here now few documents are graded, rest of the documents in each cluster can be graded using cosine similarity.

Misuseability Measure

Misuseability measure is calculated using TM score (tex- tual misuseability score). TM score calculation is divided into two phase: In the first phase each document is assigned with an individual score which is independent of the rest of the documents. This individual TM score embeds the amount of the data in the document and the sensitivity of the data. In second phase accumulated TM score is calculated which considers the number of times the document is exposed to the user.Accumulated tm score is calculated as the addition of the number of times the document has been accessed and the individual tm score which is obtained in grading.

C. Compression And Encryption

If document is found to be sensitive it goes through compression and encryption. Compression of the email is carried out using G zip compression. GZIP performs best on text-based content, often achieving compression rates of as high as 50-70% for larger files.Encryption is done using RSA. RSA involves a public key and private key. The public key can be known to everyone, it is used to encrypt messages. Messages encrypted using the public key can only be decrypted with the private key.

Also before sending the mail, hash the document content and will store it in database. At the receiving end, again the document content is hashed, and the new hashed value and the hashed value in the database is compared. If both are same, concludes that the data is untouched. Or if the hashed values are different, the data integrity is lost.

VI. CONCLUSION

Data leakage poses a serious issue for companies as the number of incidents and the cost to those experiencing them continue to increase. Data leakage is enhanced by the fact that transmitted data (both inbound and out- bound), including emails, instant messaging, website forms, transfers among others, are largely unregulated and unmonitored on their way to their destinations. The insiders were the most damaging to their organisation and responsible for exposure of confidential data.

Mining the documents and assuring the data integrity will make the data leakage system more efficient. Proving clus- tering and classification through supervised learning will give better results for finding the documents which can cause potential damage to an organisation. Once the classification has been done for the documents, encrypting and compression of the document content will assure data privacy.

REFERENCES

[1]. N. T. Nguyen, P. L. Reiher, and G. H. Kuenning,Detecting insider threats by monitoring system call activity Proc. Inf. Assurance Workshop, Jun. 2003, pp. 45-52.

[2]. J. H. Eom, N. U. Kim, S. H. Kim, and T. M. Chung,An architecture of document control system for blocking information leakage in military information system Int. J. Secur. Appl., vol. 6, no. 2, pp. 109114, 2012.

[3]. Miss S.W. Ahmad, Dr G.R.Bamnote Data

leakage detection and data prevention using algorithm International Journal Of Computer Science And Applications Vol. 6, No.2, Apr 2013.

[1]. S. Mathew, M. Petropoulos, H.Q. Ngo, and S. Upadhyaya,Data-Centric Approach to Insider Attack Detection in Database Systems International Proc. 13th Conf. Recent Advances in Intrusion Detection, 2010.

[2]. C.M. Fung, K. Wang, R. Chen, and P.S. Yu, Privacy-Preserving Data Publishing: A Survey on Recent Developments ACM Computing Surveys, vol. 42, no. 4, pp. 1-53, 2010.

[3]. S. J. Stolfo, S. Hershkop, L. H. Bui, R. Ferster, and K. Wang,Anomaly detection in computer security and an application to file system accesses Foundations of Intelligent Systems. Berlin, Germany: Springer-Verlag, 2005, pp. 1428.

[4]. Y. Shapira, B. Shapira, and A. Shabtai. Content-based data leakage detection using extended fingerprinting,2013.

[5]. T. F. Gharib, M. M. Fouad, A. Mashat, and I. Bidawi,Self organizing map-based document clustering using WordNet ontologies Int. J. Com- put. Sci. Issues, vol. 9, no. 1, pp. 8895, 2012.

[6]. M. Gafny, A. Shabtai, L. Rokach, and Y. Elovici, OCCT: A one-class clustering tree for implementing one-to-many data linkage IEEE Trans. Knowl. Data Eng., vol. 26, no. 3, pp. 682697, Mar. 2014. Conference on.

[7]. A. Harel, A. Shabtai, L. Rokach, and Y. Elovici,M-score: A misuse- ability weight measure IEEE Trans. Dependable Secure Comput., vol. 9, no. 3, pp. 414428, May/Jun. 20 12.