International Journal of Engineering Research and Development
e-ISSN: 2278-067X, p-ISSN: 2278-800X, www.ijerd.com
Volume 10, Issue 12 (December 2014), PP.14-22

Data Leak Protection Using Text Mining and Social Network Analysis

Ojoawo A. O., Fagbolu O.O., Olaniyan A.S., Sonubi T.A.
(Computer Science Department, the Polytechnic Ibadan, AOCE)

Abstract:- Data Leak prevention is a research field which deals with study of potential security threats to organizational data and strategies to prevent such threats. Data leaks involve the release of sensitive information to an untrusted third party, intentionally or otherwise while data loss on the other hand is disappearance or damage of data, inwhich a correct data copy isno longer available to the organization.Thesecorrespond toa compromise of data integrity oravailability. Data leak/loss has led to huge loss of revenue in the affected organisation and a threat to their continued existence. All organisations using electronic data storage are vulnerable to this attack. This research work is targeted at organisations with sensitive datasuch as Bank, Manufacturing industries, GSM operators, research centres, Military, Higher Educational Institutions and so on.The authorsanalyse the possible threats to organisational data and the parties that are involved in such threat, the impact of successful attack on an organisation,and current approaches to DLP.The authorsalso design a DLP modelusing "text mining" and "social network analysis", and suggested further research into "text mining" and "social network analysis"for effective future solution to DLP problems.In conclusion, implementation of this design with adherence to good data security practices and proactive strategies suggested in thispaper will significantly reduce the risk of such security threats.

Keywords:- Malicious hacking, cyber-attack, malware, thin-client, data repository, collaboration

I. INTRODUCTION

In the recent years there had been rivalrybetween the developed countries of the west and the newly emerging economies in Asian countries(such as China, India etc.).Consequently, there had been increase in threats to organizational data, ranging from cyber terrorism, malicious hacking, employee sabotage, fraud or theft, Denial of Service and the likes. The most severe of these threats is cyber-attack which is a new form of warfare employed by countries, organizations, companies and so on, battling for the control of markets, resources, and products. This method is employed to attack each other due to its cost effectiveness and anonymity of the attackers. For instance the economic powersof the world today frequently engage themselves in cyber-attacks, countries like the US, China, Russia, India and Iran, are very good examples. Computersand PDAs are now being used as a weapon in place of war equipment for attack.

Data Leak prevention is a research field which deals with study of potential security threats to organizational data and strategies to mitigate such attacks. Data leaks involve the release of sensitive information to an untrusted third partyintentionally or otherwise or an attacker (hacker) gaining unauthorized access to an organization's sensitive data, while data loss on the other hand is disappearance or damage of data, inwhich a correct data copy isno longer available to the organization, thesecorresponds toa compromise of data integrity oravailability. A number of Data leak prevention (DLP) products or techniques available attempt to mitigate some or all of these threats. Examples of the vendors of such products are Symantec, CA Technologies, Trend Micro,McAfee and so on. Data leak/loss prevention has received little attention in the academic research community. DLP is yet to be a solved problem because, current products are limited in what threats they address. Recent development on the increase of attacks on organizations had raised serious concern throughout the globe about the consequences of such attack. Today, reports of Cyber-attack often make News headline on international media frequently.

Data leak is a frequent activity that can goes on within an organization undetected until it becamepronounced. Data loss on the other hand is less frequent;however, it may be severe if no proper backup and data protection plan is in place. Data leak and data loss sometimes may not be malicious.Incidents such natural disaster destroying physical structures, careless data entry clerk entering erroneous inputs, careless placement of sensitive printed documents and the likes are not intentional.

Data Security Threats Relationship

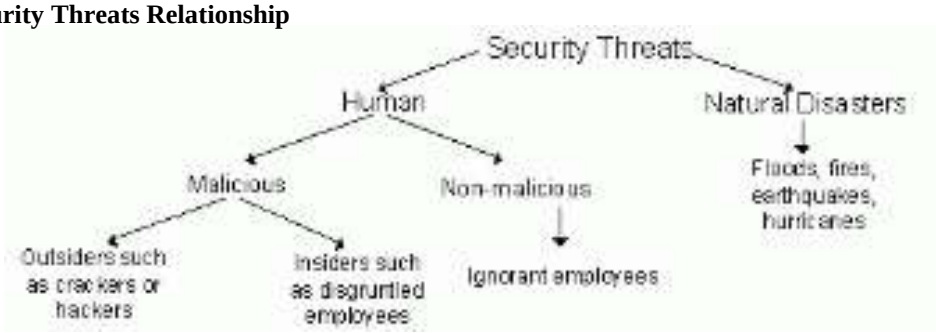


Fig.1 Source: Microsoft Corporation, Security Strategies2000

The hierarchical model in Fig.1 shows the relationship among the elements involved in Data threat.

Traditional data leak and data loss such as natural disasters, virus attacks, loss of data by careless employees etc. requires traditional DLP approach;these includes the use of fireproof cabinet, waterproof cabinets, to keep storage devices, and the use of password, access right control, watermarking, antivirus/anti-malware to protect program and data. However, sophisticated electronic data leak requiresa specialized approach.

II. ECONOMIC IMPLICATION OF DATA LEAK/LOSS

Companies can be held liable for the release of customer and employee information such as credit cards information,health records, social security numbers and so on and will be charged to pay huge compensation to the affected party. Furthermore, loss of proprietary information to competitors can result in loss of sales and may even threaten the existence of an organization. In addition to data leak, data loss can also inflict heavy losses on organisation. Loss of customer's information by a Bank or a cell phone operator for instance can lead to great financial loss. It can also lead to loss of trust by the customer, or damage the integrity of the affected organisation.

III. DATA PROTECTION LAW

Intellectual property Protection lawand some other some other relevant regulation tends to protect organisational data with tough penalties for the offenders. However some unscrupulous elements deliberately broke such because of the possibility that the authority in charge may not be able to apprehend them due to anonymity of some attack. In addition, such law is not effective in some countries, especially the Asian countries like China and India, because they do not really respect intellectual property law of the West.

IV. DATA LEAK CHANNELS

In DLP it is important to investigate data repositories and identify data leak channels. It is also very important to identify sensitive data repositories within an organization, since selecting suitable prevention techniques naturally depends on the repository in question. Employee's records, Customer records, proprietary source code and sensitive documents on network shares are a few examples of repositories. Different prevention techniques may be appropriate for different data states which are: 1. at rest (i.e. at the repository); 2. in motion (i.e. over the network), and 3. in use (i.e. at the endpoint).

Preeti Ramanet al argue that when the data is at rest, the repository can be protected with access control and audit, but when the data is in motion or in use, prevention using access control becomes increasingly difficult. For in motion and in use scenarios, the data leak prevention mechanism should be sufficiently context aware to infer the semantics of communication.

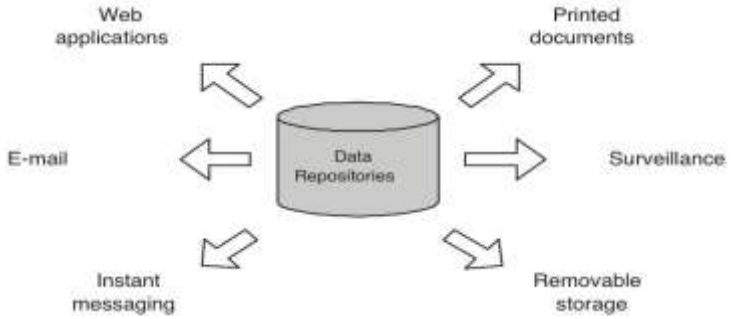


Fig. 2 Data leak channel as presented by Raman et al,

As shown in Figure 2, data leaks can occur in different ways such as Hardware theft, surveillance, and the mismanagement of printed documents. These are the traditional ways of data theft. Additionally, electronic communications such as instant messaging, web applications, social networking and email provides additional challenges. These electronic channels highly utilized in organizationsprovide means to quickly and easily send data to a third party. Traditional data leaks can be suitably prevented with traditional approaches, context aware techniques, which can infer who is communicating and what is being communicated and so on, are needed to prevent data leaks in electronic communications.

Data breaches in some organization were attributed to a number of factors as stated follows:

Code Injection: Poor programming of information systems and applications can leave organization vulnerable to various code injection attacks, or allow inappropriate information to beretrieved in legitimate database queries.Structured Query Language (SQL) injection is one of the most common attacktechniques for applications or websites that use SQL servers as back-end database.

Malware: Malware is designed to secretly access a computer system without owner's informedconsent. Sophisticated data-stealing malware may take various forms includingTrojan, spyware, screen scrappers, adware, etc. Users areusually infected during installation of other application software bundled withmalware or from malicious web sites. Download of freebies from the internet and installing it is the major source of such malware.

Phishing: Another data leakage channel is through the use of phishing sites as a lure to stealsensitive data from users. Phishing spam can be sent to staff's e-mailaddress. Once they are deceived to click the links in the malicious e-mails, theirbrowsers can be re-directed to fraudulent websites that mimic reputableorganisations, where users may unnoticeably leak their account name and passwordsto hackers. If the login credential to a organisation's web mail system is leaked, thehacker can authenticate himself or herself as the organisation's employee.

Malicious Insider: organisations sensitive data are also vulnerable to intentional data leakage performedby their internal users (e.g. employees, students). Motivations are varied, but usuallyfall into corporate espionage, financial interest, or a grievance with their employers.Sensitive data can be unauthorisedly transferred out through remote access, e-mail,instant messaging or FTP. Even if DLP solutions have been deployed within an organisation, these malicious insiders, especially IT personnel, can bypass therestrictions through sabotage of DLP systems. E.g. altering the DLP configuration tocreate backdoor, shutdown DLP services, physically cut off the power supply or declassifysensitive data.

Current approaches to Data Leak Prevention

Various companies have recently started providing data leak prevention solutions. While some solutions secure "data at rest" by restricting access to it and encrypting it, the best available solution relies on robust policies and pattern-matching algorithms for data leak detection. However, related academic work in data leak prevention focused on building policies, developing watermarking schemes, and identifying the forensic evidence for post-mortem analysis.

*Yasuhiro Kirihataet al*design a webcontent protection system to realize the protection of confidential webcontents. This system provides a special viewer application to view theencrypted content data and realize the prohibition of copying andtaking snapshots for the displayed confidential data. Ituses thedynamical encryption methodology by the intermediate encryptionproxy making it possible to protect the web contents generated dynamically.

Vachharajani et al provides a user-level policy language for hardware-enforced policies, which ensures that the sensitive data does not reach untrusted output channels through network communications, files, and shared memory. The proposed runtime information flow security system assigns predefined labels to the data and policies are enforced at the hardware level to ensure the data flow complies with the policies.

*Lars Bruckner et al***,** use data journals as a new kind of privacy enhancement technology to increase the user'sability to take advantage of his rights. Data journal is a tool that records the disclosureof personal data to services and collects related information about the serviceprovider's identity and its privacy policy. The authors describe how data journals work, howthe user can benefit from their usage, and their relation to other privacy enhancementtechnologies. They also describe two prototype implementations to show that data journals can be implementedon without changes to existing services or big changes of the user's browsing

*Lee et al.,*approaches data leak prevention from a forensics point of view and identifies the set of files needed to detect data leaks on a Windows operating system. The authors argue that delaying the collection of forensic data will have detrimental effects in the effectiveness of a data leak prevention system; hence, they

propose an efficient method to collect the basic information needed to detect data leaks by investigating five crucial system files: 1.the installation record file, 2. the system event log, 3. the windows registry, 4. the browser history, and 5. the core file in NTFS. Their approach is limited to file system-level data leaks on Windows platforms.

The current state-of-the-art in commercial data leak prevention focuses on pattern-matching, which suffers from the general shortcoming of misuse detection techniques; an expert needs to define the signatures. Given the elusive definition of data leaks, signatures should be defined per corporation basis, making the widespread deployment of current data leak prevention tools a challenge. On the other hand, the relevant academic work on data leak prevention and text mining takes a forensics approach and mainly focuses on post-mortem identification. Thus, there is a need to research further to detecting complex data leaks in real-time.

Historical records of Data leak/Loss

The table in Table 1 shows the reported cases of cyber-attack in Japan with date, target, and economic impact of the attack.

Database Data Loss Incidents

Loss Reported	Vertical	Impact
June 2011	Internet gaming	The company's Web site was hacked. No credit card information was exposed, but names, dates of birth, email addresses, and encrypted passwords were stolen. Almost 1.3 million customers were affected by this hack.
May 2011	Financial services	Employees of a bank resigned and took customer information to a competitor. They downloaded and printed confidential customer records from the bank's secure database prior to their leaving. The records included customer names, addresses, telephone numbers, Social Security numbers, dates of birth, bank account numbers, and additional personal information.
April 2011	Government	A former employee who worked for a school board was working with an identity theft ring. The person passed along information from a teacher certification database, which included names, Social Security numbers, and dates of birth. The information was used to fraudulently add people as authorized users to the victims' credit card and bank accounts.
January 2011	Retail	About 18,000 customers who used their credit cards on the retailer's Web site may have had their personal information exposed in a data security breach. A hacker was able to access the customer database and may have viewed billing information with names, addresses, telephone numbers, credit card numbers, expiration dates, security codes, and email addresses. Some people who were affected by the breach have reported fraudulent charges. People who made purchases offline are not at risk.
December 2010	Healthcare	A district attorney's office notified a hospital that its accounts payable system may have been breached. Vendors and employees who received checks between 1999 and 2011 may have had their names and Social Security numbers accessed by an unauthorized third party. The information seems to have been used to open electricity accounts.
December 2010	Retail	A hacker managed to obtain the retailer's email marketing list. People on the list were sent realistic-looking phishing emails that directed them to a Web page under hacker control. The only information that was stolen during the hack was the email list. People who fell victim to the phishing scam may have entered other personal information into the phony Web page.
September 2010	Education	A hacker from outside the United States may have accessed the information of over 100,000 applicants sometime in September. The information was mostly recruiting information and may have involved names, ACT and SAT scores, dates of birth, and Social Security numbers.
July 2010	Education	Someone gained unauthorized access to a database containing the names, Social Security numbers, and driver's license numbers of 93,000 school applicants, current and former students, parents, current and former faculty and staff, alumni, and donors. These records go back as far as 1987.
January 2010	Government	A state government commission says someone gained access to a computer server that holds more than 80,000 records containing employee information from a regulated industry. The person who hacked into the system was traced back to China and had used a computer with an external account. The server contains records including names, birth dates, and Social Security numbers.

Source: IDC, 2011

Table. 1

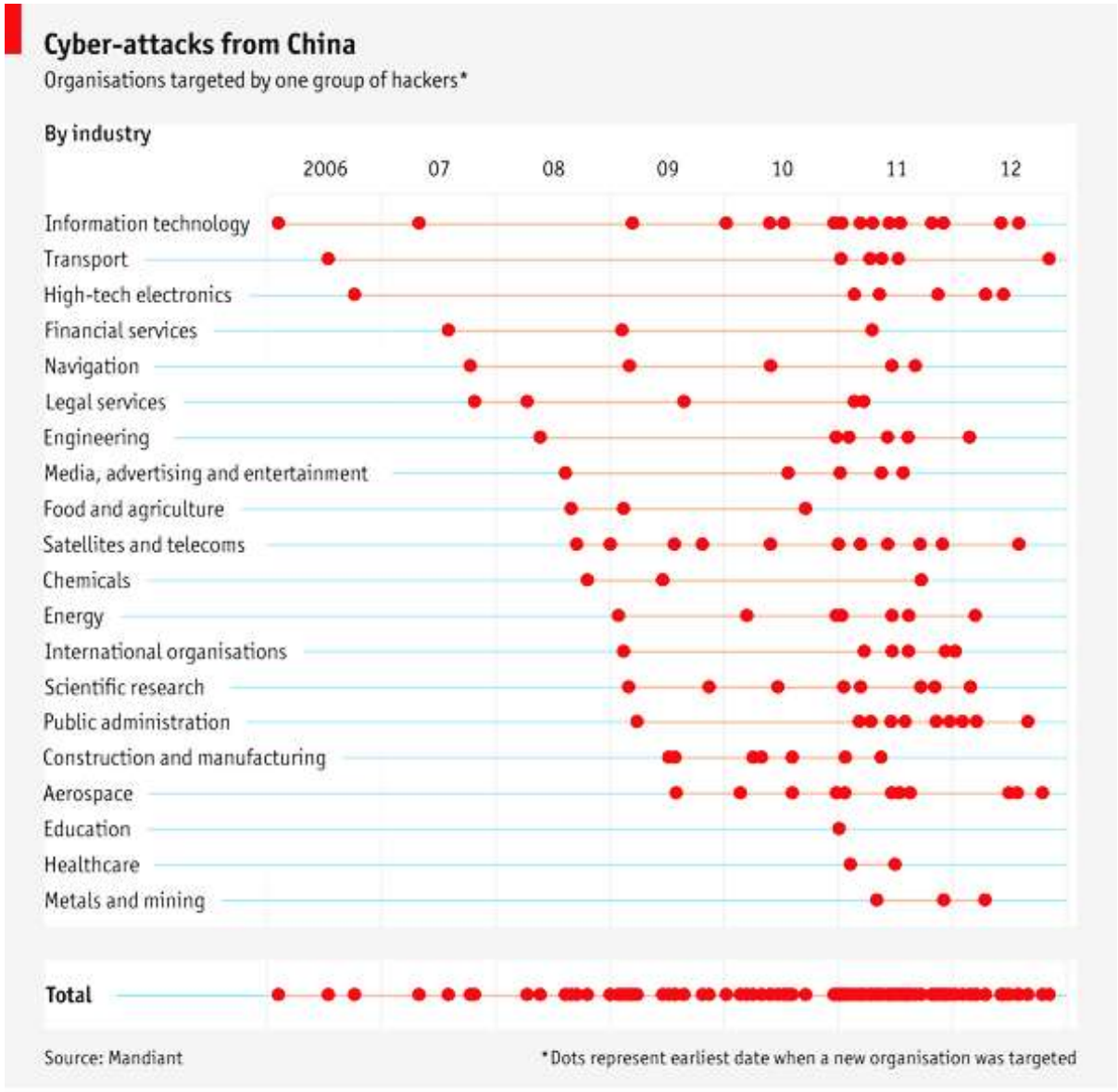
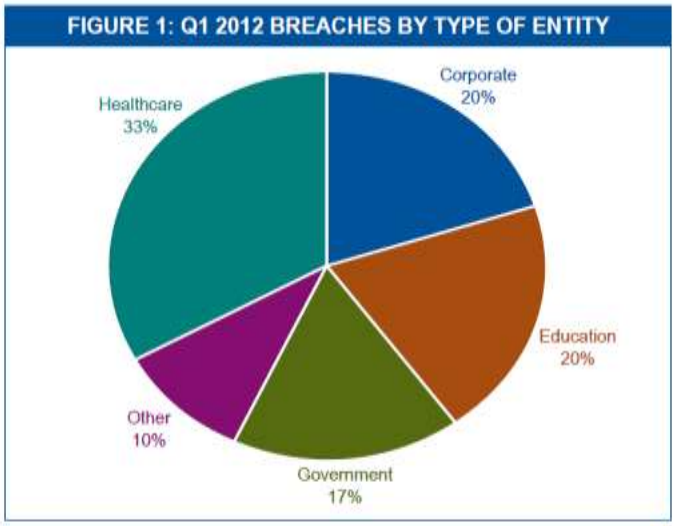


Table. 2

Table 2 above shows data on the threat from china and the targeted industry from 2006 to 2012, according to the table it is obvious data leak attack is on the increase at the same time from 2006 there is increase in the type of organization targeted. In addition, from table 2abovefor instance, there is increase in number of attack targeting educational institutions. However due to increase in awareness through staff training and implementation of security measures there had been a decrease on the impact of such attack. Also, according to the two figures in chart 1 below, the percentage of insider attack has also dropped compare to malicious hacking, this may be due to toughened penalties of security breach on the affected employees.



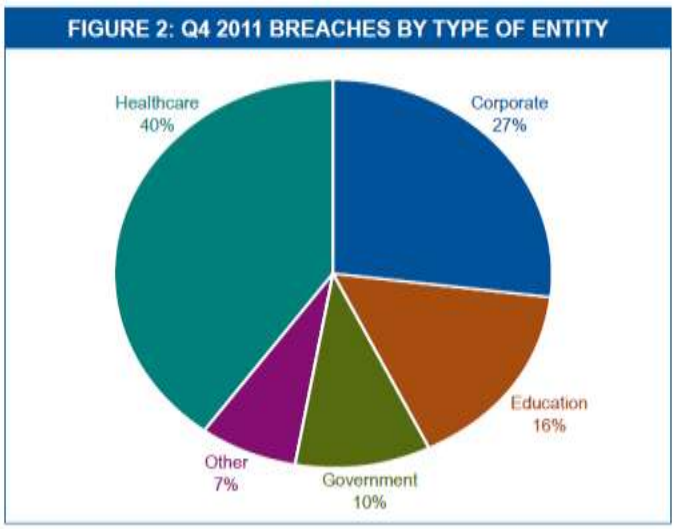


Chart 1

Challenges to Data Leak Prevention

Encryption

Different prevention mechanismsare needed to deal with different states of data. In particular,detecting and preventing data leaks in transit have majorchallenges due to encryption and the high volume of electroniccommunications. While encryption provides means to ensurethe confidentiality, authenticity and integrity of the data, it alsomakes it difficult to identify the data leaks occurring over encryptedchannels. Encrypted emails and file transfer protocolssuch as SFTP imply that complementing DLP mechanismsshould be employed for greater coverage of leak channels.Employing data leak prevention at the endpoint, outside theencrypted channel has the potential to detect the leaks beforethe communication is encrypted.

Collaboration

There is also a need to identify the collaboration parties.However, identifying the communities of collaboration is nota straightforward task. While a simple approach can considerusing the access control mechanisms e.g. to determine theprogrammers, managers, administrators etc. such approach isnot sufficient to capture heterogeneous groups where peoplecan belong to more than one group. Identifying a collaboration community should be a continuous task to care of changing and creation of new groups.

Access Control

Access control provides the first line of defence in DLP. Access control is only suitable for data atrest; it is difficult to implement it for data in transit and in use. This implies that the moment data is retrieved from the repository; it isdifficult to enforce access control. Furthermore, access controlsystems are not always configured with the least privilegeprinciple in mind.

Proposed solution to DLP problems

The biggest shortcoming of the state-of-art and the relevantprevious work is that they attempt to detect data leaks withoutan understanding of the communication context. However, thecomplex data leaks are in semantics (i.e. the content of theconversation) not in syntax. Thus, in order to address thesemantic gap problem in data leak prevention, new researchdirections should be explored to provide the semantic summarizationof communications. The main focus is identifyingin transit and in use data leaks, which are arguably morecomplex in nature. In this section, the authors review the text miningand social network analysis approaches, which will aid in building context aware DLP solutions to "in use" attacks.

Text Mining

Text mining is an exploratory data analysis techniquewhich aims to identify the natural groupings (i.e. "clusters")within a text body. Each cluster contains similar documents,according to a similarity metric.From a data leak prevention perspective, text can be collectedfrom numerous sources, such as email. Theclusters of text can serve as equivalence classes (contentsummaries), which can then be labelled to provide semanticmeaning. Thus, by applying clustering to email communications,it is possible to infer the subject of the communication ina privacy preserving manner. Based on the subjects that a usercommunicates about, a deviation from the 'usual' is flaggedand further analyzed for data leaks.Text mining, which places documents with similar propertieswithin the same group, have been utilized for summarizinglarge corpus of documents.

Chow etal. aimed to detect the inferences in sensitive documentsby applying various data mining algorithms to Enron emailcorpus. The inferences are determined based on co-occurrenceof terms in the text corpus. Similarly, Keila et. al. proposed a method for detecting deceptive emails, based onthe deception theory which suggests that people use fewerfirst person pronouns and more negative emotion and actionverbs. Singular value decomposition is utilized to visualizeemail messages and identify the outliers which correspond todeceptive emails. Applying text mining to data leak prevention involves monitoringcorporate email communications for a period of time toidentify the clusters of topics, in other words, communicationsubjects. The output of clustering may be difficult for a humanto comprehend without further processing such as in the caseof the commonly utilized k-means clustering. Thus the resulting visualization can be utilized toassign semantic meaning to the clusters manually or automatically.During deployment, when an email communications isprocessed, the most similar cluster is employed to assign thetopic of the email. If there exists a substantial deviation ofcommunication pattern (in terms of the context, frequency andthe involved parties), the resulting communication is flaggedfor further analysis.

Social Network Analysis

Social network analysis involves the mapping and measuringof relationships between people, groups and organizationsby representing the relationships in terms of nodes and connections.Social networks can be derived from communicationchannels such as email, forum discussions, and social networkingsites. Analysis of social networks can improve ourunderstanding of the relationships and groupings between theparties involved in electronic communications, email in particular.Thus the goal of social network analysis for data leakprevention is to identify the communication patterns withinthe organization and employ feedback from the administratorto identify unusual communications to uncover data leaks.*Diesner et al.* performed a social network analysis ofthe Enron email, which contains the email communicationsof top-level Enron employees before and during theEnron scandal. Applying social network analysis in data leak preventioninvolves monitoring the online collaborations (email, documentand code repositories) to discover the communities ofcollaboration. The discovered communities (i.e. social networks)are vital in identifying the collaborating parties suchas a team of developers working on the same code repositoryor a group of employees exchanging emails to perform atask (e.g. preparing for a meeting). Social network analysishas the potential to discover the collaborations which are notdocumented as a part of company policy or access control.Proper visualization of the communities can be presented tothe administrator for manual or automatic validation. Duringdeployment, if a substantial change in the social network isobserved, it is flagged for further analysis since it can reveal:(1) a dissolving social network (2) a merging social networkor (3) inclusion of an untrusted party, which is potentially adata leak.

Proposed solution to DLP problems

The biggest shortcoming of the state-of-art and the relevantprevious work is that they attempt to detect data leaks withoutan understanding of the communication context. However, thecomplex data leaks are in semantics (i.e. the content of theconversation) not syntax. Thus, in order to address thesemantic gap problem in data leak prevention, new researchdirections should be explored to provide the semantic summarizationof communications. The main focus is identifyingin transit and in use data leaks, which are arguably morecomplex in nature. In this section, the author review the text miningand social network analysis approaches, which willaid in building context aware DLP solutions to "in use" attacks.

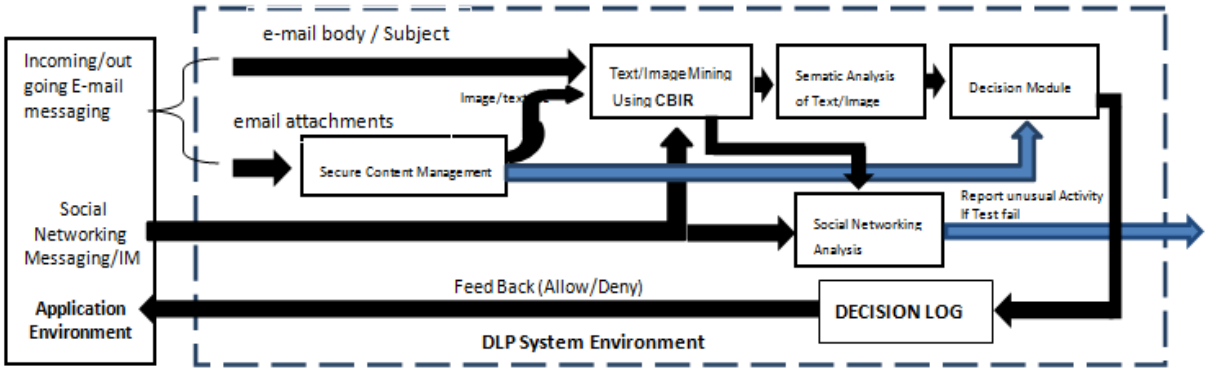


Figure. 4 Proposed DLP Model

The proposed DLP works by subjecting email message, social Networkingand Instant Messaging application to scrutiny before messages can be allowed to go out or come in. The email messages will be separated into two parts; the message body/title and the attachment.The attachment goes through secure content management module which checks for signature on the file to determine its classification, if it is classified as "restricted"then the message is denied access.On the other hand if it is image or document then it goes to text/image mining module, text or image or both will be extracted from the file using Content Based Image Retrieval (CBIR) and then proceed to text/image semantic analysis module. This is where the text and the image are processed semantically to know whether it's malicious or Non-malicious. The text/image analysis module is language sensitive and is capable of detecting the language of the text, thus the language used in writing the text is used in determining the meaning. The decision module decides whether to deny the message access or allow it. The email body is processed in the similar way except that it does not pass through secure content Management module.

The social networking analysis will be in form of investigative report for a particular period. The email messages sent and received and other messages exchanged within groups of collaborators are analysed to detect collaborations that are suspicious, and againstthe organisation's policies.

However, for data leaks the cannot be detected in real time, there must be a database in the text/image mining module which will record and store all incoming and outgoing e-mail messages, and sent/received text messages from the social networking web applications. These data will be used during periodic analysisof degree of collaborations among parties involved. After this analysis, any collaboration detected which are not documented as part of the organisation's policy is reported to the DLP administrator. This analysis can be done daily, weekly, monthly etc., depending on the requirement of the organisation.

LIMITATION

This design is targeted at the "in use" state of Data leak. If implemented, it is expected to detect some data leak in real time and other that cannot be detected immediately can be detected over a period of time. As it has been said earlier, there is no particular solution that can solve all DLP problems. In addition, detecting complex data leak in real-time still remain a challenge.

V. CONCLUSION

DLP is a multifaceted problem. Determining the sensitivedata to be protected, identifying the legitimate use of thedata and anticipating data leak channels require the internalbusiness logic of the corporation, thus, there is no particular solution that can solve all this problems. In addition to traditional data leak channelssuch as hardware theft, the widespread use of electroniccommunications such as email makes it easy to leak sensitivedata in a matter of seconds.Both data leak prevention and intrusion detection share thesame common goal, which is to detect potentially harmfulactivity. Thus, the commercial approach typically employssimilar techniques to solve data leak prevention. DLP is a substantially complex problem, when thethreat usually originates from the inside and to determine a data leak in real time is difficult. Sometimes data leaks can occur by accident betweenindividuals who are completely legitimate. The detection ofsuch data leak requires the understanding of semantics.Current state-of-art in data leak prevention mainly utilizesmisuse detection to detect data leaks, where a signature actsas a data leak description. However, misuse detection cannotscale well in data leak prevention since the data leak signatureshighly dependent on the internal business logic and should bedeveloped per organization to minimize false alerts andmaximize detection rate. Furthermore, misuse detection doesnot possess the sufficient context awareness to detect complexdata leak scenarios, where the data leak is in the semantics,not in syntax.In this paper, the author reviewed the current state-of-art,designa context awaredata leak prevention solutionusing text/image mining and social network

analysis.It is recommended thatprivacy of individuals is respected; only sematic meaning of the analysis result will be inferred. This allows data leak prevention to go beyondpattern matching and detect complex data leaks based who isinvolved in the communication and what information isbeing exchanged.

REFERENCES

[1]. **Information Technologies Promotion Agency (2011),**"10 Major Security Threats- Attacks are fast evolving...Is your security good enough?"Information-Technology Promotion Agency,Tokyo.

[2]. **J. White and D. Thompson,** 2006, "Using synthetic decoys to digitally watermarkpersonally-

[3]. identifying data and to promote data security," 2006International Conference on Security and Management, pp. 91–99.

[4]. **J. Diesner, T. L. Frantz, and K. M. Carley,** 2005 ,"Communicationnetworks from the enron email corpus "it's always about thepeople. enron is no different"," Comput. Math. Organ. Theory,vol. 11, pp. 201–228.

[5]. **Lars Bruckner, Jan Steffan, Wesley Terpstra, Uwe Wilhelm,** (2005) "Active Data Protection with Data Journals", GI-proceedings, Darmstadt pp. 269

[6]. **Microsoft corporation, (2000),**Security Strategies, Microsoft Corporation.

[7]. **National Institute of Standards and Technology**, (2011), Technology Administration U.S. Department of Commerce Special, Publication 800-12.

[8]. **N. Vachharajani, M. J. Bridges, J. Chang, R. Rangan, G. Ottoni, J. A.Blome, G. A. Reis, M. Vachharajani, and D. I.** 2004, "Rifle: Anarchitectural framework for user-centric information-flow security," Proceedings of the 37th annual IEEE/ACM InternationalSymposium on icroarchitecture. Washington, pp. 243-254.

[9]. **Preeti Raman, HilmiGüneşKayacık, and Anil Somayaji (2011),"**Understanding Data Leak Prevention",Annual Symposium On Information Assurance (Asia), New York,

[10]. **R. Chow, P. Golle, and J. Staddon,(**2008), "Detecting privacy leaks usingcorpus-based association rules," in KDD '08: Proceeding of the 14thACM SIGKDD international conference on Knowledge discovery anddata mining. New York, ACM, pp. 893–901

[11]. **Simon Liu, Rick Kuhn (2010), "**Data Loss Prevention", US National Institute of Standards and Technology, IEEE1520-9202/101520-9202/10.

[12]. **S. Lee, K. Lee, A. Savoldi, and S. Lee,**(2009), "Data leak analysis in acorporate environment," in ICICIC '09: Proceedings of the 2009 FourthInternational Conference on Innovative Computing, Information andControl. Washington,IEEE Computer Society, pp.38–43.

[13]. **Yasuhiro Kirihata, Yoshiki Sameshima, Takashi Onoyama, and Norihisa Komoda, (2011)**"Data Loss Prevention for ConfidentialWeb Contents and Security Evaluationwith BAN Logic",International Journal Of Computers, Tokyo, Issue 3, Volume 5, pp 414