Social Network Data Analysis and Menaces Detection of Email users

Linta Jacob[1*],Mruthula. N. R [2]
[1*]PG Scholar, Dept. of computer science and engineering
[2] Asst Professor, Dept. of computer science and engineering
TKM Institute of Technology,Kollam,India

***Abstract* :- E-mail has emerged as the most important application on the Internet for communication of messages, the delivery of documents and carrying out of transactions. However email is an important communication mean in computer crime communication, So there is a need of email forensics. E-mail forensic analysis is used to study the source and content of e-mail message as evidence, identifying the actual sender, recipient and date and time it was sent, etc. to collect credible evidence to bring criminals to justice. This system proposed an email forensic method which is based on graph clustering method and social network analysis. The method which includes email decoding, attribute extracting, email communication graph generating ,data filtering and email clustering. Thereby identifying the suspect who have a criminal behaviour. The intent of the proposed system is to provide an assistance during forensic investigation.**

Keywords : Social Network Analysis (SNA), Text clustering, Email Forensic tool kit, Heirarchial Eucledian minimum spanning tree

I INTRODUCTION

Computer related crime activities are more now which is a great threat to network security. As an important method to avoid computer crime, digital forensics has to become an important component for the investigation of criminal cases. Email has become easy, efficient and economical means of communication. Large amount of email traffic is generated on daily basis however undesired increase in crimes are mediated via emails.E-mail has emerged as the most important application on Internet for exchange of messages,delivery of documents and carrying out of transactions. Email protocols have been secured through several security extensions such as PGP(Pretty Good Privacy),MIME and MAC. however, cyber criminals continue to misuse it for the illicit purposes by sending spam, phishing emails, spreading viruses, worms, hoaxes and Trojan horses etc . E-mail communication is often exposed to illicit uses due to mainly two limitations:There is rarely no encryption at the sender end and/or integrity checks at the recipient end.The widely used e-mail protocol Simple Mail Transfer Protocol [SMTP] lacks a source endorsement mechanism., the metadata in the header of an e-mail which contains information about the sender as well as the path which the message travelled can easily be synthesize.It is thus important to identify and avoid users and machines misusing e-mail service. E-mail forensic analysis is used to understand the source and content of e-mail message evidence for identifying the real sender, recipient and the time and date the email was sent, etc. to collect credible evidence to bring criminals to justice

Email data provides not only evidence of the flow of information through a network, but also an indication of actor relationships. This relational information may not only identify potential sources of evidence in cases involving many actors, but also when measured, it may provide a quantified assessment of a suspect's culpability in an event or set of events.There are many tools which may assist in the study of source and content of e-mail message so that an attack or the malicious intent of the intrusions may be investigated. These tools while providing easy to use browser format, automated reports, and other features, help to identify the origin and destination of the message, trace the path traversed by the message; identify spam and phishing networks, etc.
EmailTrackerPro analyses the headers of an e-mail to detect the IP address of the machine that sent the message so that the sender can be tracked down. It can trace multiple e-mails at the same time and easily keep track of them.

Adcomplain is a tool for reporting inappropriate commercial e-mail and usenet postings, as well as chain letters and "make money fast" postings. It automatically analyses the message, composes an abuse report, and mails the report to the offender's internet service provider by performing a valid header analysis.
This system proposed a method to achieve email forensic especially in email communication relationship.

II RELATED WORKS

To combat cyber crime,Email Forensics becomes to be an important research task.
D.V.Chandra Shekar[1] proposed that Naïve Bayes classification approach is useful for predicting user's behavior and to organize the emails according to users constraints
Farkhund Iqbal[2] formally define the problem of authorship attribution introduce a novel approach of authorship attribution and formulate a new notion of

write-print based on the concept of frequent patterns write-print is dynamically extracted from the data as combinations of features that occur frequently in a suspect's e-mails, but not frequently in other suspect's e-mails.

Hong Gou [3]analyzes the working principle of an email, discusses the construction mechanism of the keywords commonly used in the header field, and applies the analysis to forensic analysis.

John Haggerty[4] proposed a framework .It focuses on the triage and analysis of unstructured data to identify key actors and relationship with in an email network. It demonstrates the applicability of the approach by applying relevant stages to the enron email corpus

Liaquat .Khan[5] address the problem of authorship verification of textual documents and employ detectionmeasures that are more suited in the context of forensic investigation, we borrow the NIST's speaker recognition evaluation (SRE) framework The purpose of the SRE framework is not only to develop state-of-the-art frameworks for addressing the issues of speaker identification and verification but to standardize and specify a common evaluation platform for judging the performance of these systems as well.Next, the evaluation measures such as DCF, minDCF,and EER that are used in the SRE framework are more organization.the algorithm developed can analyse computer organization structure tailored to forensic analysis as compared to simple ROC andclassification accuracies, etc.

Minh Tuan Vu1[6] introduced a statistical rule-based method to create rules for Spam Assassin to detect spams. Spam Assassin is one of the most popular for deciding how likely an email message is spam. It filters spam based on content-matching rules. Each rule has its own score. If an email message gains enough scores (over the pre-defined threshold), it will be marked as spam.

O.de Vel [7] describe an investigation into email content mining for author identification for the purpose of forensic investigation.It focuses on the ability to discriminate between authors for the case of both aggregated email topics as well as across different email topics.An extended set of email document features including structural characteristics and linguistic patterns were derived and together with support vector machine learning algorithm

Paglievani[8] present a systematic process for email forensics which we integrate into the normal forensic analysis workflow, and which accommodates the distinct characteristics of email evidence. Our process focuses on detecting the presence of non-obvious artifacts related to email accounts, retrieving the data from the service provider, and representing email in a well-structured format based on existing standards. As a result, developers and organizations can collaboratively create and use analysis tools that can analyze email evidence from any source in the same fashion and the examiner can access additional data relevant to their forensic cases

Siti Rahayu[9]aims to produce the mapping process between the process and output for each phase in the Digital Forensic Investigation Framework.Existing digital framework will be reviewed and then the mapping is constructed.The result from the mapping process will prove a new framework to optimize the whole investigation process.

Sobiya Khan[10] proposed a framework to perform the email statistical analysis,email classification and the clustering,email author identification and email social network analysis

Tuan-Anh Nguyen[11] proposed a Sender Policy Framework (SPF) is an open standard specifying a technical method to prevent sender address forgery. This technique requires network administrators to create SPF records for their domain. Dynamic Sender Policy Framework (DSPF) approach, in which, the legal IP addresses of servers which send emails are collected and provided by a third-party. The database of SPF records can be updated automatically and can also be used among other email servers and email gateways. Using DSPF, clients may check the SPF records without any extra configuration of their DNS.

Ungsik Kim[12] proposed an algorithm based on the properties of social networks and spectral decomposition to distinguish spam and non-spam email.They also proposed a new edge partitioning method and a measure of centrality using the eigenvector of well-known Laplacian matrix.

Yanhua Liu[13]proposed an email forensics method based on graph clustering method and social network analysis.It analyze and mine email data of the suspicious users accounts using a method which can create email communication network graph for suspicious computer criminal

III SYSTEM ARCHITECTURE DESIGN

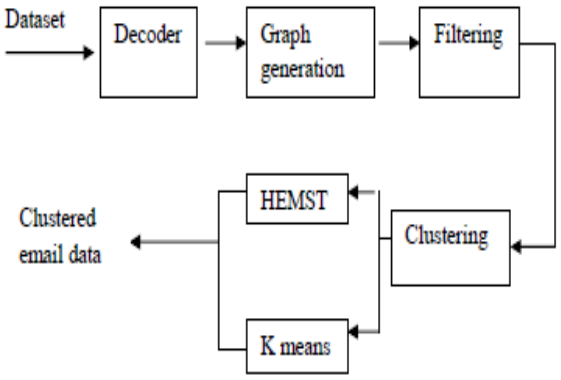

Fig 1 Email Forensics Architecture

IV IMPLEMENTATION

COLLECTING AND DECODING EMAIL DATA

Collect email data from different mail client.Different client may use different mail storage format and email account organization methods.From this it is able to find the relationship between email data files and email

accounts,For this purpose IAF decoder can be used for decoding purpose.Also there is a need to categorize every single email from one email account.Then get all user message data Most emails sent today are MIME (Multipupose Internet Mail Extensions) formatted. This allows emails to be sent with plain text and rich text/HTML versions, inline images, and attachments. MIME extensions can be added to a message in standard RFC/822 format so backward compatibility is achieved.

2 EXTRACTING EMAIL ATTRIBUTE FOR FORENSICS

Extracting the attributes required for the forensic from each email message.The attributes are 'from' address,'to'address,cc,subject,received date.The extracted attributes helps the investigator to obtain specified email relevant to the investigation

3 GENERATION OF COMMUNICATION NETWORK GRAPH

Using the email data,there is a need to construct an email communication graph using a convenient graph generating tool like graph#.Graph# is a graph layout framework. It contains some layout algorithms and a Graph Layout control for WPF applications.It helps to show the communication relationship more directly,so we realize the email communication network graph.In the network graph,the vertices represents the email accounts,The directed edge represents the communication between the users. The weight of the edge are the corresponding to number of email between the users

4 FILTERING OF EMAIL NETWORK GRAPH BASED ON MINIMUM TRAFFIC

For the filtering purpose,Set a threshold value,since the edge represents the email communication ,if the value of edge is less than the threshold value then delete that edge so it is possible to keep the email accounts more related to the group that have anomalous behaviour also can understand the email accounts of more traffic.

5 CLUSTERING ANALYSIS BASED ON HEMST

When the cluster has number of edges, the edges will influence our judgements.So there is a need to delete the redundant edges,making the cluster more closely.To determine which edge of the graph is most distant relative, HEMST algorithm can be used.HEMST produces a K partition of set of points for a given K.The algorithm constructs a minimum spanning tree of point set and remove the edges that satisfy the predefined criteria.The process is repeated until K clusters are produced.

Algorithm: HEMST(k)
Initialize $n_c \leftarrow 1$ //number of clusters
Let *S* be the point set
Let *e* be an edge in the EMST constructed from *S*
Let *we* be the weight of *e*
Let σ be the standard deviation of the edge weights
Let $ST = \varnothing$ be the set of disjoint subtrees of the EMST
Repeat
Construct an EMST from *S*
Compute the average weight w of all the edges
Compute the standard deviation σ of the edges
For each e $\in$ EMST
If we > w + σ
Remove *e* from EMST
nc ← nc + 1
ST = ST ∪ {T } //T is the new disjoint subtree
If the number of clusters nc is less than k,
remove nc − k longest edges so that nc = k
If nc < k
While nc ≠k
Remove the current longest edge
nc ← nc + 1
ST = ST ∪ {T ′} //T′ is the new disjoint subtree
Return k clusters
If the number of clusters nc is greater than k
If nc > k
Compute the centroid ci of each Ti ∈ ST
Find the representative ri ∈ Ti closest to ci
S = ∪Ti∈ST
{ri}
until nc = k
Return k clusters

6 CLUSTERING OF EMAIL DATA

For the clustering of email data ,there is a need to represent the documents as an array of numbers.The simplest way is to just represent as a vector of word counts.Each email document is represented as a vector using the vector space model.The vector space model is an algebraic model for representing text document as vectors of identifiers.

Eg:-TF-IDF weight ,Term Frequency for a word as the ratio of number of times the word occurs in the document as to the total no of words in the document.

IDF[Inverse document frequency]:-It is a measure of whether the term is common or rare across all documents.It is obtained by dividing the total number of documents containing by number of documents containing the term and then taking the log of quotient.

Finding similarity score:-Cosine similarity to identify the similarity score of the document.The method find cosine similarity takes 2 argument vector A and vector B as parameter which are the vector representation of document A and document B and returns a similarity score lies between 1 and 0 indicating that the document A and document B are completely similar.

V PERFORMANCE EVALUATION

On evaluating the proposed system it is found that cluster that exhibit the anomalous behavior can be identified more accurate than the previous method.HEMST algorithm used for the clustering purpose is much efficient as compared to other methods used in the previous work

V1 EXPERIMENT RESULT

Enron Email Dataset can be used as the dataset,IAF decoder is used for the decoding purpose.Using the extracted attributes ,Network graph is created using graph#.Set a threshold value in the case of filtering.Email data is clustered using HEMST and K means.Finally clustered the email data exhibit similar behavior.Cluster that exhibit the anomalous behavior that can be identified.

VII CONCLUSION

This system proposed a extensive forensic solution including email network creating based on communication ,weight calculation,filtering based -on minimum traffic ,clustering of email network using HEMST and the email data using K means algorithm which is to analyze the suspicious email data.These methods have good usability to email forensics

VI REFERENCES

[1] D.V.Chandra Shekar"Classifying and Identifying of Threats in E-mail Using Data Mining Techniques Proceedings of the International MultiConference of Engineers and Computer Scientists 2008 Vol I IMECS 2008
[2] Farkhund Iqbal "A Novel Approach of mining Write-prints for authorship attribution in email forensics.2008 Digital Forensics Research Workshop
[3] Hong Gou"Analysis of email header for forensic purpose" Published in Communication systems and Network Technologies(CSNT),2013 International Conference,Pages 340-344
[4] John Haggerty"A Framework for the forensic investigation of unstructured email relationship data"IJSCNS July 2011
[5] Liaquat A. Khan "Email authorship verification for forensic investigation" International Journel of Computer Science and Network Security,VOL.8,October 2010
[6] Minh Tuan Vu" Multilingual Rules for Spam Detection" Journal of Machine to Machine Communications, Vol. 1 , 107–122
[7] O.de Vel "Mining Email content for author identification forensic"IJCSNS international Journel of Computer Science and Network Security,April 2013,Vol 7
[8] Paglierani ,Mabey.M,"Towards comprehensive and collaborative forensics on email evidence"Networking Applications and Worksharing,2013 International conference Pages 11-20
[9] Siti Rahayu"Mapping process of Digital Forensics investigation framework.IJCSNS ,International Journel of Computer Science and Network Security,VOL.8,October 2008
[10]Sobiya khan"Email Data Analysis for theapplication to cyber forensic investigation using data mining IJSCNS 2012
[11] Tuan-Anh Nguyen" Spam Filter based on Dynamic Sender Policy Framework" Proceedings of the Asia- Pacific Advanced Network 2010 v-30 p 1-9
[12] Ungsik Kim"Analysis of personal email network using spectral decomposition" IJCSNS International Journal of Computer Science and Network Security, VOL.7 No.4, April 2010
[13]Yanhua Liu"An Email Forensics Analysis Method based on Social Network Analysis" International conference on cloud computing 2013