International Journal of Computer Information Systems and Industrial Management Applications.
ISSN 2150-7988 Volume 8 (2016) pp. 284-292
© MIR Labs, www.mirlabs.net/ijcisim/index.html

A Novel Method for Outlier Entity Detection in Email Communication Network (ECN)

Maqsood Mahmud
Department of Computer Sciences, ALMAAREFA Colleges of Science and Technology (MCST), Riyadh, Kingdom of Saudi Arabia
mmahmud@mcst.edu.sa

Abstract: **This paper is the extension of the paper stated in [1].In the extended paper, elaboration of outlier entities are discovered based on the behavioral dissimilarities. Fifteen different features were proposed for this process using a variant of Kth nearest neighborhood (KNN) algorithm. Outliers were the convicted email users in ENRON, US based company which were already been declared convicted. The results in the papers proved to be matched with the 80% of the convicted email users because few of the users were not detected due to the constraints of the algorithm. In top 19 outliers detection 3 convicted users were found out. It means 15% of the result was achieved in top 20 users. The bench marking is made with the existing convicted and declared email users. The proposed features proved to helpful in the detection of outliers in email communication network and can be further implemented for various other kind of communication networks.**

Keywords: Email Communication Networks (ECN), Early Warnings, Outliers Detection, Anomaly Prediction, Behavioral Dissimilarity, Kth Nearest Neighborhood algorithm KNN.

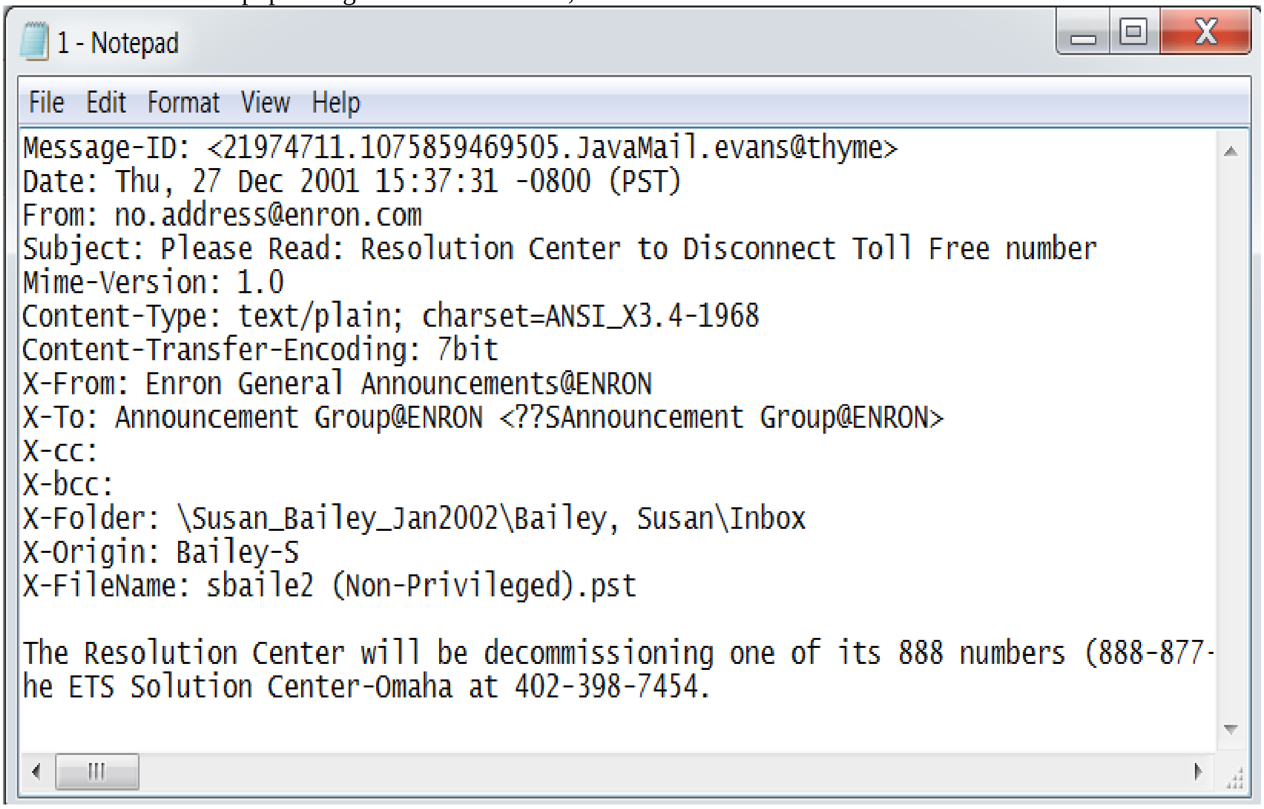
I. Introduction

Anomalies in online social networks can indicate asymmetrical and often illegitimate behavior. Detection of such anomalies has been used to identify malevolent individuals, including spammers, sexual predators, and online fraudsters [2]. Today the Internet is being used by e-criminals, fraudulent persons and scammers for criminal activity. Various researches into early warnings are now emerging. These early warning systems depend on the concept of recovering and sourcing data obtainable online or offline in the form of emails on an organization's server, news material or any data supplied by investigators. The concept behind of such warning devices is primarily rate of recurrence or frequency [3].

The increasing acceptance of online social media is leading to its extensive use among the online community for various usages. Some of the web forums are mainly being used for open dialogues on critical issues prejudiced by fundamental views. The radicalness of a user is captured by a measure based on the degree of match of the commented posts with a threat list. Eleven different collocation features are articulated to categorize the connotation among users, and they are lastly entrenched in a modified PageRank algorithm to produce a ranked list of radically significant users. The collocation theory is more effective to deal with such ranking problem than the textual and temporal similarity-based measures studied earlier [4].

The study of associations among entities like people, groups, or websites is systematically studied in social network analysis. Networks like telephone calls, e-mails or social networking sites like Facebook, Twitter etc., comprise social networks. A graph where nodes symbolize entities like people or groups, whilst the edges portray associations or data flow among the nodes forming a social network. Both visual and mathematical analysis of relationships is performed today in Social Network Analysis (SNA). Intelligence and investigative analysts noted recognition of significant patterns within various sectors of the network and featuring them as a significant work. The Social Communication Network (SCN) is now extensively used for business purposes. The abnormal behavior of entities can be easily identified through automated anomaly detection using outlier analysis. Abnormal entities are not always suspicious. The analyst can be helped through identification of anomalous behaviour and restricting the amount of data for visual inspection. Sometime visual inspection is onerous and also prone to errors. Sometimes social changes alter the user's actions and behaviour to escape detection. At a specific time, an irregular entity residue sufficiently dissimilar from the "expected" o "normal" behaviour can be seen in most of the entities visualized by the virtue of statistical analysis. Non-suspicious entities may be discarded by using the information in combination with other accessible meta-data. In an email network, it is usually seen that a person who directs or order a cluster or group of people is replicated in the majority of the emails streaming within the cluster. Consequently, this user may be noticed as an exceptional case due to significant interaction with their friends, but can be eradicated using additional information [21].

Now anomaly or irregularity detection is a highly developed field of research. It can be efficiently practiced in different areas like credit-card fraud; network intrusion detection, email depending network analysis etc. "Normal behavior is more pre-dominant than abnormal behavior" is the proposition on which the anomaly detection algorithms works. This phenomenon is shown through many network entities. For designing the proposed system, Enron dataset was chosen for testing purposes. A sample email is presented below in Figure 1. A number of features were selected out of the dataset to determine outliers in nodes. These nodes were deemed as malicious nodes as our features concentrated on deviated behaviours. These features are discussed in detail in later sections. Basically, three main files were created for extracting the information from dataset. The first file for (i) user information, the second for (ii) message information and third for (iii) interaction details information between users and messages. A runtime matrix was also generated to measure processed information from these three basic files based on the 15 selected features. After populating the Runtime Matrix, the square of Euclidean distance was calculated between a specific node/point with every node/point. The square of distance was taken for the sake of convenience. This process was performed with every node. Minimum distance is identified for every point, and then these minimum distances are listed in descending order. The top points are considered as outliers. These outliers give clues of malicious nodes/users.



```
Message-ID: <21974711.1075859469505.JavaMail.evans@thyme>
Date: Thu, 27 Dec 2001 15:37:31 -0800 (PST)
From: no.address@enron.com
Subject: Please Read: Resolution Center to Disconnect Toll Free number
Mime-Version: 1.0
Content-Type: text/plain; charset=ANSI_X3.4-1968
Content-Transfer-Encoding: 7bit
X-From: Enron General Announcements@ENRON
X-To: Announcement Group@ENRON <??SAnnouncement Group@ENRON>
X-cc:
X-bcc:
X-Folder: \Susan_Bailey_Jan2002\Bailey, Susan\Inbox
X-Origin: Bailey-S
X-FileName: sbaile2 (Non-Privileged).pst

The Resolution Center will be decommissioning one of its 888 numbers (888-877-
he ETS Solution Center-Omaha at 402-398-7454.
```

Figure 1. ENRON Dataset Sample

The objective of this paper is to find out a method for outliers detection in email communication networks. The outliers are anomalies or abnormalities in a system or a network. A set of thoughtful features needs to be selected to fit into the variant of Kth Nearest Neighbourhood (KNN) algorithm for outlier's detection in e-mail communication network designed of Enron dataset.

The rest of the paper is organized as follows. “Related Work” is in Section 2, “Proposed System” in Section 3, Section 4 describes “MUDS Architecture”, “Results & Discussions” in Section 5 and “Conclusion & Future Work” in Section 6. Acknowledgement and References follows after Section 6.

II. Related Work

Gupta and Dey, [21] proposed a proficient algorithm for irregularity determinations in social networks. Irregular individuals are discovered depending on their behavioural differences from other individuals. They chose 23 features for user’s characterization, divided in three categories, i.e., (i) outgoing features (ii) incoming features (iii) global features. Outgoing features and incoming features are based on the couplet of frequency of emails and length of messages/calls/emails of users, while global features characterize one-way interactions. These global features are crucial in determining anomalous user’s behaviour. We also used two of the sub-features in the global features. They applied kth to the nearest neighbour algorithm on their dataset with these rich features and found the top two irregular entities i.e. Jeffery Skilling and Kenneth Lay. They also compared their results with KOJAK on VAST 2008 dataset and proved better results. They also showed allocation of the irregular and non-irregular entities for the two datasets on the shifted feature space subsequent to implication of PCA. The top three principle components were chosen for their simulations for efficient visualization. For both situations, peak ten irregular entities were coloured red. The rest were coloured blue. Stacked Pie charts were also showed to provide explanations of anomalous behaviour. The differences in incoming and outgoing behaviour led the authors to conclusions of anomalous entities. The stack bar charts also showed that proportion of one way incoming and outgoing communications are greatly elevated for irregular entities then their neighbours. The weak point in their research is the non-specification of users with respect to their email IDs. If this point is addressed it can provide more fruitful results. In our research we addressed this point and specified the users with their identities, message IDs, subjects and contents as well.

Ramaswamy et al [18] have proposed a new technique for distance, depending outliers that rely on the distance of a point from a neighbour. Authors ranked every point based on its kth distance to its adjacent neighbour and declared the top n points in this position to be outliers. They used the concept of MBR (Minimum Bounding Rectangles) for distance calculation. The advantage of using MBR is to reduce computation for mining outliers in a large dataset. MAXDIST and MINDIST were calculated between various MBRs. Authors preferred Partition based algorithm over the “Blocked Nested-Loop Join” and “Indexed Based Join” algorithms because of its less computational and I/O cost. The partition based algorithm first generate partitions, calculates limits on Dk for points in each partition. Then it categorizes candidate partitions having outliers and calculates outliers from points in nominee partitions. Empirical experimentation on NBA (National Basketball) dataset has proved that partition based algorithm scale well with respect to dataset size and dataset dimensionality. Partition depending technique is 30 times quicker than “Indexed Based Join” technique and 180 times faster than “Blocked Nested-Loop Join” technique. Authors mathematically showed the solution of dimensionality problem but calculation of the distances between MBRs diagonals has prompted questions. MBRs needs to be more defined in the paper as it sometimes creates confusion in specifying the boundaries of MBRs and its diagonals. The purpose of this critique is that when we go for n dimensionality, the MBRs becomes MBP (Minimum Bounding Polygon) which needs to be addressed to give a solution to real life problems in n dimensions.

Nithi and Dey [20] proposed an efficient algorithm for anomaly detection from call data records. Anomalous users were detected based on dubious attribute values derived from their communication patterns. They proposed techniques for discovering irregular behaviour from big datasets, observing communication to any depths. The author's position is that usual behaviour is supplementary pre-dominant than irregular behaviour. The features that were selected by the author are: receiver ID, caller ID, call duration, time of call initialization, tower used, call type, SMS or voice and details of handset used. They only used incoming and outgoing features of the call. He did not used global features in this paper. Authors have focused on “calling patterns” and “interaction patterns” here and termed “call duration pattern” as the most important feature for anomaly detection. In our perspective, “call duration” cannot always be a determining feature for anomaly or outlier. It may be misleading in specific events where a normal user can be suspected as he prolonged a call depending on his mood or specific event. The author also applied Principle Component Generator to recognize the dimensions which the data exhibits upper limit variations. The kth closest neighbour algorithm [3] was implemented to find top n outliers. Authors also showed experimental results which were better than Kojak because proposed algorithm showed 80% results while Kojak showed 60% results. The interesting thing was that it dealt with real life dataset and found pleasing results.

In [5], authors have introduced the idea of a multi criteria weighted graph, similarity techniques and its application to observe characteristics of social networks which are brought under observation. The similarity method on the email network for which the likely outputs are shown. These output are based on terrorist networks that equipped and performed 9/11, 2001 intrusions. Weinstein et al [6] described initial results on modelling, detection, and tracking of terrorist groups and their intents based on multimedia data. The scenarios can be created by subject matter experts using a graphical editing tool. In [7], authors presented a new network irregularity detection method depending on wavelet analysis, estimated auto regressive and outlier discovery methods. To differentiate between network traffic behaviours, authors have presented 15 characteristics and implemented these features as the put-in signals in wavelet-based methods. Kurkovsky et al [8] have presented a multi-modal social networking architecture for the purpose of contributing geographic data amongst neighbouring individuals. The architecture gives individual with a conventional web-enabled with a voice-enabled ends or interfaces, which can be controlled over a cell phone.

Larsen and Vejin [9] used centrality measures. These methods were used to observe the destabilisation of network from multifaceted networks. They recently presented algorithms for building a chain of command of secret networks. It was because; investigators could inspect the construction of the makeshift networks in a way to weaken opposition and enemies. In [10] authors have provided a detailed study of the research relevant to the study of vibrant modelling and link forecasting of SCN. Irregularity discovery for discovering alteration in the performance of e-mail treatment were also presented. In [11], authors have presented a framework, which consists of components of information extraction, blog spider, visualization of network and its analysis. They implemented this structure to recognize and investigate a chosen set of 28 anti-Blacks hatred clusters on Xanga. This is one of the mainly trendy blog hosting sites. Lin and Chalupskyin [12] described a novel unsupervised framework to identify abnormal instances. In the second part of the paper, authors described an explanation mechanism to automatically generate human-understandable explanations for the discovered results.

In this work, Bhatia and Gaur, [13] proposed a novel algorithm breadth first clustering. This clustering used arithmetical method for population mining in community networks. This method continues in breadth first method and increasingly exposes social clusters from the networks. This method is straightforward, robust and can be configured without difficulty for major community networks. Hui-Yi and Hung-Yuan [14], highlighted the current, modern and advanced social networks and its prevailing impact on society i.e. Facebook. Authors have struggled to examine online friend matching and making models of Facebook individuals. This can be seen from comparing the relationship between the individual activities and website existing time of Facebook individuals. Further, it is on the basis of hypothetical foundation of “experiential marketing” and in perspective of the practice component. These components might be “sense experience”, “feeling experience”, “thinking experience”, “acting experience” and “related experience”. These kind of widely used social networks needs to be addressed because malicious entities can make their covert places in these social networks and can be easily used for criminal, movements provoking /rebellions or terrorist activities.

B. Carrier and E.H. Spafford [15] described the process of digital investigation. This paper highlights that digital inspection which has currently proven to be more widespread. Physical inspection existed for a long time and the physical practices are practiced on the digital investigation. The inspection of a computer device is analogous to a physical transgression sight. A bodily offence scene can be developed to recognize numerous parts of proofs. Blood on a place is one proof and can be examined to recognize the owner of the blood, what hit the subject, the site of the injured party, the place of the aggressor, and time of intrusion etc. Likewise, a fingerprint is one portion of proof that can be analyzed to exhibit self data and point of reference data about how an individual was confronted.

Numerous definitions exist for each vocabulary and these were selected since they most precisely adhere to the author's perspective on the issue. These are as follows. Physical Evidence: Physical substance that can determine an offence or a crime has been executed, can supply association between a crime and its victim. It can supply association between a crime and its perpetrator. The real, hard disk, computer, CD-ROM and PDA, are illustrations of physical evidence [16]. Digital Evidence: Digital data that can determine an offence or crime has been executed, can give a relation between a transgression and its victim. It can give a relation between a transgression and its performer. Data in memory, in computer USB, on the computer hard disk, or in a mobile phone are illustrations of digital evidence [16]. Physical Crime Scene: The physical circumstances where physical proof of a transgression or event is supposed to be found. The place where the first unlawful act took placed is the primary physical crime scene. Successive scenes are secondary physical crime scenes [17]. Digital Crime Scene: The imaginary circumstances developed by software and hardware where digital proof of a transgression or event is found. Circumstances where the first illegal action took placed are the primary digital crime scene and successive scenes are referred to as secondary digital crime scenes[17].

The literature reviews above lead us to the new method of outlier's detection in social networks and assessing the digital crime scene. In [21] authors give a direction of outlier's detection in phone call communication networks. Thus, email communication network outliers are a new scenario for the outlier detection with new feature

III. Proposed System

The system to be proposed will determine outliers and ultimately detect malicious nodes in the dataset. It was assumed earlier that users could be characterized by the number of emails sent or received, number of contacts, servers used for emails interactions etc. Statistical properties like the average number of emails sent per day, average number of contacts, etc. have also been used. However, a user's actions cannot be inspected by summation values or mean values. Individual interaction patterns can better pigeon-hole a user in the perspective of global behavioral features. For this purpose we considered incoming features, out-going features and global features in detail. The rule of outlier discovery describes the irregularity or anomaly discovery algorithm. An outlier is an observation that diverges to a great extent from other observations that a person feel that this observation stimulate doubts that it was produced by unlike methods. Unsupervised methods to recognize outliers are helpful in finding irregularities or uncharacteristic entities from huge datasets. An outlier is illustrated as follows [3]: Definition: Outliers of a set are the peak "n" data elements that are extreme from their kth adjacent neighbors. This distance measure is very proficient for examining spatial datasets. The following procedure or algorithm finds top n outliers or irregularities. The values “n” and “k” are given as inputs. This method finds all those entities, that are very dissimilar from their neighbors and that's why entitled to be outliers. The complexity of the below procedure or algorithm is O (N2). N is the total quantity of nodes. The procedure is stated below [21]:

Step 1: Select a value for k.

Step 2: For each node determine the distance from its kth closest neighbor using Euclidean distance.

Step 3: Organize the data points in descending order of the distance obtained in step 2.

Step 4: Select top n points as outliers.

The initial design of the system depends upon the three relational schemas (i) UserFile (ii) MessageFile (iii) UserInteractionFile. These are given below with the fields. UserFile(Uid, Name, EmailId); UserFile Schema is the User information extracted from dataset. It contains User ID (UId), Name of Nodes as (Names) and Email address of users as (EmailId) of all users in dataset. MessageFile(No, Mid, Subject, Size, Date, Time, Content); Message File Schema is about the message details. It contains Message ID as (Mid), subject of a message as (Subject), size of an email as (Size (KB)), Date, Time and Contents as well. Content Analysis will be discussed in future research work. ; UserInteractionFile; (SourceUser, DestUser, MsgId, CommType, Attachements); UserInteractionFile schema is about the interaction details of users with the corresponding emails sent and received. The first column is about the Source User who sends an email; the second column is about the Destination User who receives the message, augmented with message ID from Message detail file. Communication type is recorded to know the whether the communication is in the form of “TO” or “CC” or “BCC”. Attachment information is also noted to see the frequent or infrequent email attachments. It should be noted that we used Square of Euclidean distance for our convenience. In general, the distance between two points x and y in a Euclidean space$\Re$nis given by Equation 1.

$$d = \| x - y \| = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$$

The thoughtful features sets defined for our experimentation on the dataset are listed below in Table1. These features are about 15 in number. They are based on the behavioural aspect as well as self-identifiers. Table 1a and Table1b are about the listing of features used for the experimentation. Details of these features are discussed below.

Table 1a. Selected Set of Features for Behavioural Dissimilarities

NO.	**Feature's Name**	**Mathematical Symbol**	**Mathematical Representation of Features**
1	Out-degree	$(deg^{+}(u))$	$deg^{+}(u)=\lvert\{<u,v>\mid \forall v\in V\wedge <u,v>\in E\}\rvert$
2	In-degree	$(deg^{-}(u))$	$deg^{-}(u)=\lvert\{<v,u>\mid \forall v\in V\wedge <v,u>\in E\}\rvert$
3	In-Out Ratio	$R_{IO}\ (u)$	$R_{IO}(u)=\left\{\begin{matrix}\dfrac{deg^{-}(u)}{deg^{+}(u)},deg^{+}(u)\neq 0\\ \propto,otherwise\end{matrix}\right\}$
4	Out-In Ratio	$R_{OI}\ (u)$	$R_{OI}(u)=\left\{\begin{matrix}\dfrac{deg^{+}(u)}{deg^{-}(u)},deg^{-}(u)\neq 0\\ \propto,otherwise\end{matrix}\right\}$
5	Ratio Variance	$V_{R}\ (u)$	$V_{R}(u)=\lvert R_{IO}(u)-R_{OI}(u)\rvert$
6	Message Sent Time	$T_{S}\ (u)$	$T_{S}(u)=\dfrac{\sum_{i=1}^{deg^{+}(u)}t(u,v_{i})}{\sum_{i=1}^{deg^{+}(u)}f(u,v_{i})}$
7	Message Received Time	$T_{R}\ (u)$	$T_{R}(u)=\dfrac{\sum_{i=1}^{deg^{-}(u)}t(v_{i},u)}{\sum_{i=1}^{deg^{-}(u)}f(v_{i},u)}$
8	Message Sent Day	$D_{S}\ (u)$	$D_{S}(u)=\dfrac{\sum_{i=1}^{deg^{+}(u)}d(u,v_{i})}{\sum_{i=1}^{deg^{+}(u)}f(u,v_{i})}$
9	Message Received Day	$D_{R}\ (u)$	$D_{R}(u)=\dfrac{\sum_{i=1}^{deg^{-}(u)}d(v_{i},u)}{\sum_{i=1}^{deg^{-}(u)}f(v_{i},u)}$
10	Sent Messages with Attachments	$A_{s}\ (u)$	$A_{S}(u)=\dfrac{\sum_{i=1}^{deg^{+}(u)}a(u,v_{i})}{\sum_{i=1}^{deg^{+}(u)}f(u,v_{i})}$

Table 1b. Selected Set of Features for Behavioural Dissimilarities

1	Received Messages with Attachments	$\mathrm{A_R}\ (u)$	$A_R(u) = \frac{\sum_{i=1}^{deg^-(u)} a(v_i,u)}{\sum_{i=1}^{deg^-(u)} f(v_i,u)}$
2	Sent Message Size Mean	$\mathrm{M_S}\ (u)$	$M_s(u) = \frac{\sum_{i=1}^{deg^+(u)} s(u,v_i)}{\sum_{i=1}^{deg^+(u)} f(u,v_i)}$
3	Received Message Size Mean	$\mathrm{M_R}(u)$	$M_R(u) = \frac{\sum_{i=1}^{deg^-(u)} s(v_i,u)}{\sum_{i=1}^{deg^-(u)} f(v_i,u)}$
4	Least Contacted User	$\mathrm{LC}_d\mathrm{U}\ (u)$	$LC_dU(u) = \frac{\sum_{i=1}^{deg^-(u)} f(v_i,u)}{\sum_{i=1}^{deg^+(u)} f(u,v_i)}, wheref(v_i,u) = \begin{cases} 0, if, f(u,v_i) = 0 \\ f(v_i,u), otherwise \end{cases}$
5	Least Contacting User	$\mathrm{LC}_g\mathrm{U}(u)$	$LC_gU(u) = \frac{\sum_{i=1}^{deg^+(u)} f(u,v_i)}{\sum_{i=1}^{deg^-(u)} f(v_i,u)}, wheref(u,v_i) = \begin{cases} 0iff(v_i,u) = 0 \\ f(u,v_i)otherwise \end{cases}$

A. Feature's Description

In order to extract discriminative features for malicious entity identification, the e-mail communication network is modelled as a directed graph G = (V, E) where, V is the set of nodes representing users and E⊆V×V is the set of directed edges representing communications between them. A directed edge originating from a node u and terminating to a node v is represented as <u,v>. Once the communication graph is created, we have identified a set of 15 graph-based features explained briefly below. Out-degree: out-degree of a node is defined as the number of originating edges from it. For a node u, the out-degree is represented as deg+(u).In-degree: In-degree of a node is defined as the number of terminating edges to it. For a node u, the in-degree is represented as deg −(u) and it can be calculated using equation 3.In-Out Ratio(RIO):This feature shows the ratio of in-degree and out-degree of a user u. Out-In Ratio(ROI): This feature shows the ratio of out-degree and in-degree of a user u. Ratio Variance (VR):This feature is for knowing the difference between the two ratio's obtained from the above two features. Message Sent Time(TS): This feature is considered to model the message sent time pattern of a user. Message Received Time(TR): This feature is considered to model the message received time pattern of a user. Message Sent Day (DS):This feature is deemed to model the message sent day pattern of a user. Message Received Day (DR):This feature is perceived to model the message received time pattern of a user. Sent Messages with Attachments(AS):This feature is deemed to model the message sent with attachments of a user. Received Messages with Attachments (AR): This feature is perceived to model the message received with attachments of a user.

Sent Message Size Mean(MS):This feature models the mean of sent message sizes of a particular user. Received Message Size Mean(MR):This feature models the mean of received message's sizes of a particular user. Least Contacted User (LCdU):The feature shows directional behavior of communication for user u in an email network. Least Contacting User (LCgU): This feature shows directional behavior of communication for user u in an email network.

IV. MUDS System Architecture

Figure 2 shows our proposed conceptual prototype MUDS (Malicious Users Detection System) design. In this prototype design, "Mailboxes" are extracted from the Internet with the help of POP and IMAP. The "Mailbox Extraction" module is further connected with "Email Repository" in a unidirectional fashion while connected bidirectional with "User Interface". In the diagram "Transformation Sub System" module transform all emails in desired way to help in outlier detection. The "Intermediate Representation (IR)" module takes information from "Transformation Sub System" and stores in the next module called "Intermediate Representation" module. In IR, four relations are extracted (i) User Table (ii) Message Table (iii) Interaction Table (iv) Results Table. The IR module helps kth nearest neighborhood algorithm module to perform algorithm and extract results using Euclidean distances based on the above selected fifteen features. The "Computation" module has bidirectional relation with IR. "Outlier Ranking" module is attached with IR in a unidirectional way. It computes the malicious user and ranks them to reveal the accused malicious users. "User interface" is connected with "Outlier Ranking" module to interact and represent information to operator or intelligence agencies.

Figure 2 is a conceptual model for the MUDS. The system was developed using programming language C#. The MUDS Architecture was tested by Enron dataset. Enron dataset was a raw data and could not be used directly for MUDS. Cleaning of data required few modules to be developed in programming language. Extraction of required data from Enron dataset was performed by a function named "UserFile(Uid, Name, EmailId)". The contents of the emails were extracted using another module "MessageFile(No, Mid, Subject, Size, Date, Time, Content)". To calculate the distances between two points i.e. P1 and P2 a third module related to the email user interaction is needed. The name of the third module was given name as "UserInteractionFile(SourceUser, DestUser, MsgId, CommType, Attachements)".

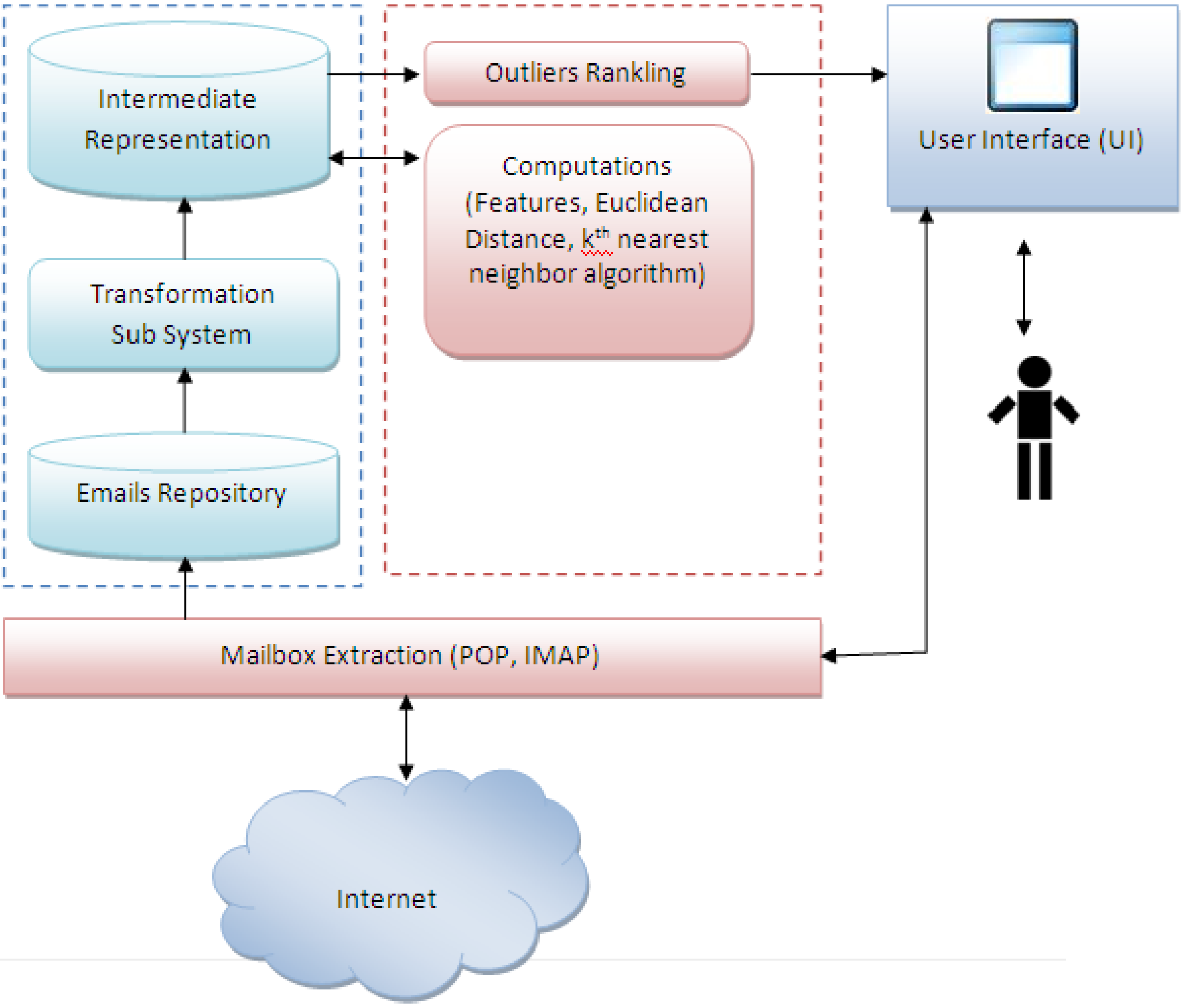

Intermediate Representation
Transformation Sub System
Emails Repository
Outliers Rankling
Computations (Features, Euclidean Distance, kth nearest neighbor algorithm)
User Interface (UI)
Mailbox Extraction (POP, IMAP)
Internet

Figure 2. Conceptual Malicious Users Detections System Architecture (MUDS)

A variant of KNN was used to achieve better results. The variant KNN contains robustness in the sense of calculating distances between two points. The highest distances are bubbled out on the top of the table as shown in Table 2. The demarcation of the suspects is made according to the existing benchmark or accused existing digital crime scene.

V. Results and Discussion

Results obtained from the MUDS were interesting to note as it gave three malicious users in the top 20 records. As can be seen in Table 2. Ms. Sally Beck and Mr. Richard Shapiro originally accused of Enron fraud are in top list of MUDS system. The third accused user Mr. John Lavorato is shown in 18th position. In above results 15% of the records show us our desired results. The three victims discussed above are in the list of the convicted email users who were sentenced in the court for their fraudulence in Enron Company. The results are crossed check with the published list of crime victims and were found true. Table 2 is a sample result from thousands of the records collected from Enron dataset. Result of 15% accuracy in top 20 records is meaningful and very important for the investigation agencies that are in search of crime scene. A single clue regarding primary or secondary crime scene may lead to open new gates for big crime scandals.

Table 2. MUDS Prototype Results

NO.	Point One (P1)	Point Two (P2)	Distances
1	jeff.dasovich@enron.com	veronica.espinoza@enron.com	9634.455078
2	sally.beck@enron.com	taffy.milligan@enron.com	4513.056641
3	richard.shapiro@enron.com	James.steffers@enron.com	2641.174561
4	janette.elbertson@enron.com	david.forster@enron.com	2272.520752
5	david.forster@enron.com	kay.chapman@enron.com	2241.006348
6	kay.mann@enron.com	tana.jones@enron.com	2103.902588
7	tana.jones@enron.com	kay.mann@enron.com	2103.902588
8	monika.causholli@enron.com	paul.kaufman@enron.com	1437.817627
9	kay.chapman@enron.com	mary.hain@enron.com	1248.970215
10	mary.hain@enron.com	kay.chapman@enron.com	1248.970215
11	mathew.lenhart@enron.com	mary.cook@enron.com	1075.147217
12	mark.guzman@enron.com	geir.scolber@enron.com	948.7890625
13	stephanie.panus@enron.com	christi.nicolay@enron.com	923.8248291
14	susan.scott@enron.com	liz.taylor@enron.com	896.3186035
15	steven.merris@enron.com	michael.mier@enron.com	865
16	craig.dean@enron.com	leaf.harasin@enron.com	828
17	rosalee.fleming@enron.com	taffy.milligan@enron.com	694.1584473
18	john.lavorato@enron.com	bill.williams@enron.com	689.1584473
19	liz.taylor@enron.com	drew.fossum@enron.com	673.3147583
20	drew.fossum@@enron.com	liz.taylor@enron.com	673.3147583

VI. Conclusions & Future Work

Outliers were successfully revealed by using variant of kth nearest neighbor algorithm. Features were thoughtfully chosen to meet the requirements of outlier detection and hence finding malicious users. These results can be used further for analysis of more complex social networks now e.g. Facebook, Twitter, LinkedIn or Orkut etc for detection of cybercrimes with help of anomalies in SCN. Partition based algorithm can also be used in case of very large social networks. Content analysis of email bodies may also addressed in the future to give stronger results of malicious users with proof from text rather than behavior analysis.

Acknowledgement

This work was supported by ALMAAREFA COLLEGES OF SCIENCE AND TECHNOLOGY (MCST), RIYADH, KINGDOM OF SAUDI ARABIA. The author is grateful for this support.

References

[1] Mahmud. M, Pathak.P, Pathak.V, Afridi.Z., "Detection of Criminally Convicted Email Users by Behavioral Dissimilarity" 6th International Conference on the Computational Aspects of Social Networks (CASoN 2015) Pietermaritzburg, South Afri ca. Vol. 409, pp. 429-439, 1-3 December 2015.

[2] Savagea.D,Zhanga.X, Yua.X, Choua.P, Wanga. Q., “Anomaly detection in online social networks”. Journal of Social Networks, Volume 39, Pages 62–70. October 2014.

[3] Henry L. Timothy Palmbach, and Marilyn Miller. Henry Lee's, *"Crime Scene Handbook*". Academic Press, 2001.

[4] Anwar.T, Abulaish. M. “Ranking Radically Influential Web Forum Users” IEEE Transactions on Information Forensics and Security Volume:10 , Issue: 6 1289 – 1298, Feb 2015.

[5] Tarapata. Z. and Kasprzyk. R., “An Application of Multicriteria Weighted Graph Similarity Method to Social Networks Analyzing,” 2009 International Conference on Advances in Social Network Analysis and Mining, vol. 1, Jul. 2009, pp. 366-368.

[6] Weinstein. C., Campbell. W, Delaney. B., Leary. G.O., and Street. W., “Modeling and Detection Techniques for Counter-Terror Social Network Analysis and Intent Recognition,” 2009.

[7] Lu. W., Tavallaee. M., and Ghorbani. A. a, “Detecting Network Anomalies Using Different Wavelet Basis Functions,” 6th Annual Communication Networks and Ser vices Research Conference (cnsr 2008), May. 2008, pp. 149-156.

[8] Kurkovsky. S., Strimple. D., Nuzzi. E., and Verdecchia. K., “Mobile Voice Access in Social Networking Systems,” Fifth International Conference on Information Tech nology: New Generations (itng 2008), Apr. 2008, pp. 982-987.

[9] Larsen. H.L. and Vej. N.B., “Practical Approaches for Analysis, Visualization and Destabilizing Terrorist Networks,” 2006.

[10] Negnevitsky. M., huey Lim. M.J., Hartnett. J., and Reznik. L., “Email Communica tions Analysis: How to Use Computational Intelligence Methods and Tools ?,” Sys tem, 2005, pp. 16-23.

[11] Mining. U.W.E.B., S. Network, A. To, S. The, E. Of, C. Communities, and I.N. Blogs, “Using web mining and social network analysis to study the emergence of cyber communities in blogs,” Computer.

[12] Lin. S.-de, and Chalupsky. H., “Discovering and Explaining Abnormal Nodes in Semantic Graphs,” IEEE Transactions on Knowledge and Data Engineering, vol. 20, 2008, pp. 1039-1052.

[13] Bhatia. M.P.S. and Gaur. P., “Statistical approach for community mining in social networks,” 2008 IEEE International Conference on Service Operations and Logitics, and Informatics, Oct. 2008, pp. 207-211.

[14] Hui-Yi. P., Ho and Hung-Yuan, “Use Behaviors and Website Experiences of Face book Community,” Statistics, vol. 1, 2010, pp. 379-383.

[15] Carrier. B. and Spafford. E.H., “Getting Physical with the Digital Investigation Pro cess,” International Journal, vol. 2, 2003, pp. 1-20

[16] Richard Saferstein. Criminalistics: An Introduction to Forensic Science. Pearson, 7 edition, 2000.

[17] Eoghan C. "Digital Evidence and Computer Crime". Academic Press, 2000.

[18] Ramaswamy.R Sridhar.R. “Efficient Algorithms for Mining Outliers from Large Data Sets,” In: Proceedings of the 2000 ACM SIGMOD international conference on Management of data, Texas, United States (2000).

[19] Qureshi. P.A.R., Memon. N., and Wiil. U.K., “EWaS: Novel Approach for Generat ing Early Warnings to Prevent Terrorist Attacks,” Second International Conference on Computer Engineering and Applications, 2010, pp. 410-414.

[20] Nithi. L. and Dey, “Anomaly Detection from Call Data Records,” Social Networks, 2009, pp. 237-242.

[21] Gupta. N. and Dey. L. , “Detection and Characterization of Anomalous Entities in Social Communication Networks,” 2010 20th International Conference on Pattern Recognition, Aug. 2010, pp. 738-741.

Author Biographies

The author is Assistant Professor at Department of Information Systems, ALMAAREFA Colleges Of Science and Technology (MCST), Riyadh, Kingdom of Saudi Arabia. He is member of IEEE . He did his PhD at University Technology Malaysia (UTM). He contributed in research in various International Conference and Journal mostly in IEEE, ACM and LNCS. He published two US Patents. He has served as reviewer for various international conferences and journals.