Detecting Fraud Using Transaction Frequency Data

Roheena Khan, Andrew Clark, George Mohay, Suriadi Suriadi
Science and Engineering Faculty
Queensland University of Technology
Brisbane 4001, Australia

***Abstract*—Despite all attempts to prevent fraud, it continues to be a major threat to industry and government. In this paper, we present a fraud detection method which detects irregular frequency of transaction usage in an Enterprise Resource Planning (ERP) system. We discuss the design, development and empirical evaluation of outlier detection and distance measuring techniques to detect frequency-based anomalies within an individual user's profile, relative to other similar users. Primarily, we propose three automated techniques: a univariate method, called Boxplot which is based on the sample's median; and two multivariate methods which use Euclidean distance, for detecting transaction frequency anomalies within each transaction profile. The two multivariate approaches detect potentially fraudulent activities by identifying: (1) users where the Euclidean distance between their transaction-type set is above a certain threshold and (2) users/data points that lie far apart from other users/clusters or represent a small cluster size, using k-means clustering. The proposed methodology allows an auditor to investigate the transaction frequency anomalies and adjust the different parameters, such as the outlier threshold and the Euclidean distance threshold values to tune the number of alerts. The novelty of the proposed technique lies in its ability to automatically trigger alerts from transaction profiles, based on transaction usage performed over a period of time. Experiments were conducted using a real dataset obtained from the production client of a large organization using SAP R/3 (presently the most predominant ERP system), to run its business. The results of this empirical research demonstrate the effectiveness of the proposed approach.**

Keywords—anomaly detection; enterprise resource planning systems; fraud detection

I. INTRODUCTION

Enterprise Resource Planning (ERP) systems are one of the most important IT developments to emerge in the 1990s. More and more organizations are now adopting ERP systems, with most of the Fortune 1000 firms having installed ERP systems to run their businesses [2]. An ERP system is a packaged software solution that aims to automate and integrate the core business processes of an organization. Whilst ERP systems provide numerous benefits to organizations, due to their complexity they are vulnerable to many internal and external threats [1].

Since the advent of ERP systems, researchers have typically focused on fraud prevention rather than detection. Many publications have discussed fraud prevention approaches such as role based access control, segregation of duties, username and passwords, etc in different systems [1], [3] and [4]. Although many organizations employ fraud prevention techniques, they only prevent simple kinds of fraud from occurring and are not enough on their own [5]. Complex fraud schemes built over time, involving various applications, are difficult to prevent. Nevertheless only a few publications deal with fraud detection approaches in ERP systems [6], [7]. Another driver for better fraud detection particularly in ERP systems, is the shift towards service oriented architectures. These architectures allow a higher degree of automation of business processes, which may lead to more cases of fraud as the number of human checks are reduced and the number of entry points into the system are increased [8].

The audits conducted by auditors and fraud examiners to detect fraud in ERP systems are generally very labour intensive requiring time, effort and resources [9]. They need to have a good understanding of the business, ERP software and its features to conduct effective audits. As audits are conducted periodically, generally once every financial year, fraud is only detected towards the end of the year. According to the KPMG fraud survey [23], the average time to detect fraud is 18 months. Automated fraud detection approaches provide a possibility of real time detection which can be conducted continuously therefore identifying frauds as soon as they are perpetrated and reducing the overall financial losses and time to detect fraud.

An important analysis carried out by auditors is the investigation of outliers or anomalies in the types of transaction performed by users, their frequency and the transaction amounts. In this paper, we propose the use of continuous and automated outlier detection and distance measuring techniques to detect frequency-based anomalous behavior. The intention is to identify activity which may be indicative of financial fraud. We use the term, *transaction type* to represent a single activity in the system, and a *transaction profile* (tp) to denote a set of distinct transaction types that one or more users have performed [10]. A transaction profile may be associated with one or many users and each user is associated with exactly one transaction profile (as discussed in [10]). In particular, we detect per-user anomalies in the frequency of each transaction type within a transaction profile. We identify such univariate outliers for each transaction type, using boxplots, a common graphical outlier detection technique. We also detect per-user anomalies using a set of transaction types within a transaction profile - taking into account the entire set of transaction types. We identify such cases using the Euclidean distance (ED), a prevalent distance measuring technique and a clustering algorithm; k-means. The rationale here is to detect cases where

ISSN (Print): 2204-0595
ISSN (Online): 2203-1731

a combination of individual transaction type frequencies in a transaction profile, may cause an outlier.

The next section describes the related work in the field. The paper follows with a discussion of the proposed approach, using transaction profiles, in Section III. The methodology of detection of univariate anomalies using Boxplots and multivariate anomalies using Euclidean distance is presented in Section III. The dataset, experiments and a discussion of the results are presented in Section IV. The paper concludes with a brief discussion on the current work and future directions presented in Section V.

II. Related Work

Typically outliers are considered as noise or errors in data, that may need to be discarded; usually in a preliminary step before carrying out further data analysis. In our case, outliers may signify users that behave in a suspicious or irregular manner, and these rare and suspicious events are more interesting than the frequently occurring ones. Barnett et al.'s [11] classical definition of an outlier is, "an observation that appears to deviate markedly from other members of the sample in which it occurs". Ngai et al. [12] in their work argue that there is a lack of research on the application of outlier detection techniques to fraud detection, perhaps due to the complexity of detecting outliers. The authors suggest that in the field of fraud detection, outlier detection is highly suitable for distinguishing fraudulent data from authentic data, and thus deserves more investigation [12].

Outlier detection methods are also referred to as anomaly or novelty detection methods, and have been employed to identify credit card [13] and telecommunications fraud [14]. For example, Fawcett et al. [14] propose a rule-based method for detecting fraudulent usage of cellular phones based on profiling customer behaviour. Their system, called DC-1, constructs a fraud detection tool in three stages:

- rules are generated to distinguish fraudulent calls from legitimate ones;
- these rules along with a set of templates are used to build a collection of profiling monitors. Each monitor examines behaviour based on one learned rule. In other words, these monitors, profile the typical behaviour of each account in accordance with a rule, and describe any deviations from the typical behaviour;
- the system finally weighs the monitor outputs and generates an alarm if there is sufficient evidence of a fraudulent activity [14].

Clustering algorithms too have been applied to detect anomalous behavior. Oh and Lee [25] detect anomalies in audit trail data by profiling the transactions executed by the users. They propose a method of clustering the activities of transactions generated by a user and detect anomalies based on each user's profiles.

Outlier detection methods have been categorized into univariate and multivariate techniques. In the next section, we investigate related univariate statistical and graph-based methods.

A. Univariate Outlier Detection Techniques

Perhaps one of the most accepted statistical outlier detection techniques is the use of standard scores, also known as Z-scores. These are used to rescale raw data into its equivalent standard score, that is in accordance with a measure of the overall data spread (the distribution's standard deviation) [11]. For example: given $x_1, x_2, x_3 \ldots, x_n$, let $\bar{x}$ be the mean and s the standard deviation, an observation x is considered an outlier, if:

$$z = (|x - \bar{x}|)/s \; > k \tag{1}$$

where k is the outlier threshold, generally a value of 3 or even 4 standard deviations above the mean. The justification is that an outlier will have a relatively large standard score (z), given that it will be far from the distribution's mean (where about 95% of the data lies), assuming a normal distribution. However, Shiffler [16] shows that a k value of 3 or 4 precludes the existence of outliers in samples of size $n \leq 10$, or $n \leq 17$, respectively. The Z-score mean and standard deviation estimates also give a good idea of the data shape.

Barnett and Lewis [11], present a comprehensive review of statistical outlier detection methods with mathematical proofs and discordancy tests for detecting outliers, where the underlying distribution of the data is known (called parametric techniques). In other words, outliers are observations that deviate from the model assumptions. Unfortunately, these methods depend on many assumptions, such as the knowledge of the distribution, the distribution parameters, the number of expected outliers and the type (univariate or multivariate) of expected outliers [17].

In real world datasets, these factors or assumptions are often not realistic. Consequently, a more robust outlier detection technique is required. Thus, we adapt a well-known, graphical method for outlier detection, called the boxplot. Among many other exploratory data analysis techniques, John Tukey [18], proposed the concept of a boxplot. Boxplot is a non-parametric (or distribution-free) technique, which is based on the five-number summary: lower extreme, lower quartile, median, upper quartile and upper extreme. Figure 1 illustrates an example of a boxplot. The middle line across the box, divides the data into two halves, indicating the median (see Figure 1). The rectangular box around the median, depicts the inter-quartile range (IQR), that is the distance between the 25% percentile (or lower quartile: Q_1 and the 75% percentile (or upper quartile: Q_3), that is $Q_3 - Q_1$. From the upper and lower quartile, dashed lines extend in either directions, called whiskers, representing k times the interquartile range and stop at the data point closest to this limit. Points beyond this limit are tagged as outliers. k corresponds to values of 1.5 and 3, for mild and extreme outliers, respectively. The boundaries of k are portrayed by the lower and upper fences, computed as $Q_1 - k(Q_3 - Q_1)$ and $Q_3 + k(Q_3 - Q_1)$ respectively. The values of 1.5 and 3 for k are also known as inner and outer fences and are selected based on a normal distribution. However, the authors [19] argue that these k values have shown to successfully tag outliers in several datasets (and therefore defined as a non-parametric technique).

ISSN (Print): 2204-0595
ISSN (Online): 2203-1731

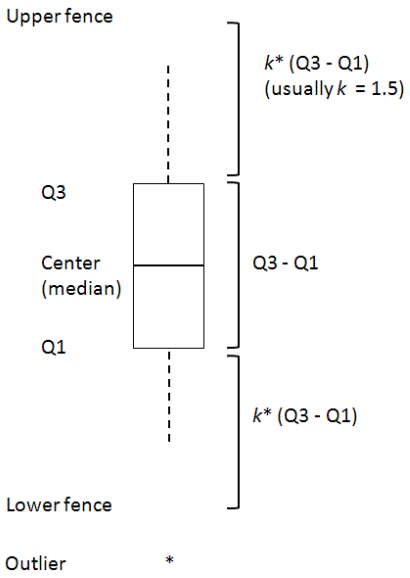

Fig. 1. Example of a boxplot (adapted from [18]).

On visual inspection, the boxplot shows several data features such as the: location - shown by the median, spread - shown by the length of the box or the quartiles, skewness - for instance: if the median is much closer to the lower quartile than to the upper quartile, indicating that the data is positively skewed, tail length - shown by the points at which the whiskers stop: determined primarily by the most extreme data values that are within the outlier cutoffs or fences; and outliers of the data - depicted by a plus (+) sign outside *k* [19].

In the next section, we present a detailed review of several related multivariate statistical and distance-measuring techniques, along with the motivation for using Euclidean distance.

B. Multivariate Outlier Detection Techniques

Multivariate techniques are categorized into (a) statistical methods that are typically parametric (depending on the data distribution, for example: Mahalanobis distance) and (b) data-mining based methods, which are often non-parametric (that is, they do not rely on the data distribution parameters, for example: Euclidean distance and k-means clustering techniques) [17]. We choose non-parametric measures for detecting multivariate outliers because they do not require the dataset to have a normal distribution. This section provides a brief overview of related statistical and data-mining multivariate outlier detection methods, and present the motivation for choosing the selected method.

In order to determine whether an observation is an outlier or not, we incorporate in our proposed anomaly detection approach, a prevalent distance measure, which does not rely on the distribution mean and the variance-covariance, called the Euclidean distance. The Euclidean distance, $ED(x_i, x_j)$, between two p-dimensional instances: $x_i = (x_{i1}, x_{i2}, \dots, x_{ip})$ and $x_j = (x_{j1}, x_{j2}, \dots, x_{jp})$ can be calculated as: $\sqrt{(x_{i1} - x_{j1})^2 + (x_{i2} - x_{j2})^2 + \cdots + (x_{in} - x_{jn})^2}$.

The Euclidean distance satisfies the following mathematical distance conditions [20]:

- $d(i,j) \geq 0$: distance is a non-negative number;
- $d(i,i) = 0$: distance of an object to itself is zero;
- $d(i,j) = d(j,i)$: distance is a symmetric function; and
- $d(i,j) \leq d(i,h) + d(h,j)$: going directly from object i to object j in space is no more than making a detour over any other object h (called the triangular inequality).

The above conditions also hold true for detecting anomalies within the transaction frequencies of each transaction type with a transaction profile.

Though the main objective of clustering techniques is to segment data into related clusters, many clustering algorithms such as k-means and k-mediods are also used for detecting multivariate outliers [21]. Multivariate outliers are denoted as data points that lie far apart from any other clusters and/or represent a small cluster size. Clustering algorithms are generally non-parametric and therefore do not assume an underlying data distribution model. They determine clusters based on a distance measuring function such as the Euclidean distance. We have selected the k-means method for our experimental analysis, as it is the most well-known and commonly-used clustering algorithm. The clustering algorithm takes two parameters as input - *k*, the number of clusters or partitions to be made (needs to be pre-set) and *X*, the dataset or the matrix. The algorithm aims at minimizing the sum of the object-to-centroid distances, thereby making the resulting k clusters as compact and as separate as possible [20]. In the next section, we demonstrate the validity and effectiveness of Boxplot, Euclidean distance and the k-means clustering technique for detecting univariate and multivariate outliers, within transaction profiles.

III. Anomaly Detection Approach

In order to detect univariate anomalies, we flag users whose frequency of a particular transaction type is much higher compared to other users who have performed the same transaction type within that particular transaction profile. We identify anomalous transaction frequencies by constructing a boxplot for each transaction type in a profile, where the threshold for the anomalous user transaction frequencies *tf* , are set with *k*.

The objective is to flag the most suspicious or highly unusual usage of transaction types, hence we set the value of *k* to 3 (which is recommended for the detection of extreme outliers that lie outside the outer fence). This implies that no outliers are detected in transaction profiles with a small number of users. Shiffler [16] suggests that a boxplot may wrongly identify some observations as outliers in datasets which have a small sample size ($n > 10$). Thus, we decided to set the minimum number of users in a transaction profile to be greater than ten. It may be noted, that the anomaly type focuses on the upper quartile and upper fence only and not the lower quartile.

ISSN (Print): 2204-0595
ISSN (Online): 2203-1731

We detect multivariate anomalies in frequency usage of the set of transaction type(s) present within a transaction profile. The Euclidean distance between the frequency of each transaction type, between each pair of users within a transaction profile is calculated (where the multiple variables are the different transaction types). Euclidean distance above a certain threshold value, $\Delta EDthreshold$, is used as a criterion to flag users based on all the transaction types performed and their frequencies. These users may or may not have univariate outliers in each feature (that is each transaction type) within a transaction profile, but the whole observation (set of transaction type frequencies), may result in a multivariate outlier. We automatically set a threshold value based on the mean of the highest distances between users within all transaction profiles in the dataset. The technique flags pairs of users within transaction profiles. To find out if one or both users within a pair of users are anomalous within a transaction profile, we flag for further investigation user(s) that occur the most number of times amongst the user pairs which are above the $\Delta EDthreshold$ (implying that their transaction usage is different from all others within that profile).

For the multivariate k-means analysis, we detect data points that lie far apart from any other clusters and/or represent a small cluster size. Our algorithm groups similar objects together, based on the principle that objects within a cluster have higher similarity (based on the Euclidean distance) amongst themselves in comparison to objects in another cluster. We visually represent the clusters using Matlab's silhouette plots (detailed in Section IV D) to identify the anomalous users.

For detecting multivariate outliers, we consider transaction profiles which are associated with at least two transaction types. These users typically belong to one role as they have performed the same transaction type set over a period of time.

IV. Experimental Results and Discussion

The approaches were implemented in Matlab, using the boxplot and Euclidean distance functions and parameters. All results were generated and analyzed using Matlab.

A. Dataset

To assess the effectiveness of the proposed method, we performed experiments on a real dataset collected for a period of about eight months between the 17th of June 2008 and 16th of February 2009. The dataset contains 81,047 records and 9,383 users, who have performed 17 different transaction types. All usernames in the dataset were anonymized. To improve the overall detection mechanism, transaction profiles with a user group of more than ten users were selected. Amongst the 68 transaction profiles, 10 profiles were identified with a user group of ten or more users. These transaction profiles represent one to three distinct transaction types. The dataset is extracted from an operational environment, and consists of real users and activities. We identify anomalous activities alerted by use of the techniques described above in order to demonstrate the effectiveness of our approach. A careful examination of the anomalies is then carried out for verification.

TABLE I. Univariate Outliers

tp_i id	Users in tp (N_u)	No. of t (N_t)	Transaction (t) names	Total Outliers (N_o)	Visual Outliers (N_{me})	N_{me} t names
1302	11	2	XK01, XK02	None	None	None
1299	14	1	XK02	None	None	None
1297	25	2	invoice_approved, requisition	1	1	invoice_ approved
1296	282	1	requisition	15	1	requisition
1314	568	1	conf_approval_1	29	1	conf_ approval_1
1326	583	2	goods_received, requisition	27, 135	1	requisition
1327	663	3	goods_received, invoice_approved, requisition	7, 24, 48	1	goods_ received
1291	716	1	invoice approved	57	2	invoice_ approved
1320	1892	2	goods_received, invoice_approved	35, 148	None	-
1316	4546	1	goods_received	125	None	-
	9,300			651	7	

B. Discussion of Anomalies Detected With Boxplots

Univariate outliers were identified with boxplots for each transaction type in a profile. Table I summarizes the results of the experiments – showing the transaction profile id, the number of users in each transaction profile (N_u), the total number of transaction types (N_t), the transaction names of the transactions in the profile, the total number of records identified as outliers with boxplot (N_o), the most extreme outliers identified from the visual impression of the boxplots (N_{me}) and the transaction names of the most extreme outliers.

Boxplots were constructed for all transaction profiles with more than ten users (listed in Table I). The median value (as portrayed on the boxplots) for all transaction types is around one, in almost all cases, as the majority of users have performed the transaction type only once, during the period for which the data has been extracted (in other words, the box plots show no lower quartile because the median and the lowest frequencies are the same or very close to equal). Table I consists of six unique transaction types: XK01, XK02, invoice_approved, requisition, goods_received and conf_approval_1, out of the 17 transaction types in the dataset. Transaction profiles containing one transaction type are tp 1299, 1296, 1314, 1291 and 1316 (see Table I). From the 9,300 N_u or records in the dataset, the boxplot has tagged 651 univariate outliers, for the six transaction types (as shown in Table I). These 651 outliers also include a count of all the users for a particular frequency value. For example in a tp: an outlier value of frequency 2 may be denoted by a single plus (+) sign on the boxplot for a particular transaction type, but it may consist of say, 86 users. An investigator will need to review each anomaly to understand whether they indicate fraud or not. Since a transaction type performed twice by 86 users, does not seem like an anomaly, the investigator can either adjust the threshold parameters or exclude these from further investigation. In general, it may be observed that the transaction profiles with the largest user groups had the highest

ISSN (Print): 2204-0595
ISSN (Online): 2203-1731

number of outlying values. The threshold value could be further adjusted for these transaction profiles to obtain the optimal number of outliers. The overall percentage of $N_o$is reasonably small, contributing to about 7% of the dataset. However, N_{me} equates to a much smaller number of outliers (7), constituting about 0.07% (7/9300) of the dataset. These most extreme outliers can be readily identified at a quick glance of the boxplots. From the total number of records identified as outliers with the boxplot, the most extreme outliers identified from the visual impression of the boxplots, were selected based on the following criteria:

- the distance (as observed from the visual impression) or frequency displayed on the y-axis, from the most extreme outlier value to its second or next data point in the boxplot; and
- if the transaction frequency value of the outlier is particularly high.

For an auditor, it is interesting to investigate the most extreme outlying values from the visual impression of the boxplots. The most prominent outlying value is observed in transaction profile 1296. The outlier consists of a user who has performed only one transaction type (that is, requisition), 56 times, which is significantly higher compared to the remaining 281 users in the profile. Similarly the most extreme outlier in transaction profile 1314, represents a user who has just performed configuration approvals, 453 times. The profile consists of 568 users, and no user except for this particular user has performed configuration approvals more than 259 times. Transaction profile 1291, has two outlying values, where both users have performed invoice approvals, 115 and 95 times, respectively. Although these frequency values may not necessarily be regarded as high transaction usage, they appear distant – and hence anomalous, compared to the other group of users depicted in the boxplot.

On the contrary, in transaction profiles 1316 and 1320, the outlying values are close to each other, and thus may not represent potentially fraudulent activities or perhaps, they may both be fraudulent. In transaction profile 1297, the most extreme outlier represents a user who has executed 26 invoice approvals. Though the boxplot has marked the transaction as an outlier, the overall low transaction usage of the transaction types in this profile may suggest that it may not be outlier.

The high transaction usage of a particular transaction type may imply that the transaction is one of the main responsibilities defined by the user's job function or role in the organization. This can be verified by examining the SAP R/3 system's user-role and role-transaction type tables. Such users who have performed only 1 transaction type for the entire period are interesting to investigate, as they might be valid users who may have changed their job function or are promoted, meaning that they are assigned a new role for accessing the system and the outlying values are transactions performed with their previous role. Or perhaps they might be synthetic user ids created by valid users to perform fraudulent activities - anyway they are anomalous. Evidently, these anomalous values require an in-depth analysis.

In the next section, we discuss the multivariate anomalies detected using Euclidean distance.

C. *Discussion of Anomalies Detected With Euclidean Distance*

For the multivariate approach, the dataset is stored in MySQL for analysis. For manual analysis and investigation of the flagged users, multiple SQL queries and reports are generated. Prior to running the distance measuring techniques we normalize the dataset using the z-score method (discussed in Section II A).

To improve the overall detection mechanism, we select transaction profiles that have at least two transaction types and ten users for our analysis. It may be observed from Table I that among the ten transaction profiles, only five fulfill the formulated criteria. Transaction profiles containing one transaction type - that is tp ids 1299, 1296, 1314, 1291 and 1316 are excluded for the current multivariate analysis (these profiles were included in the univariate analysis). The remaining five transaction profiles represent at least two distinct transaction types (see Table II). Table II presents a summary of the experimental results – showing the transaction profile id, the number of users in each transaction profile (N_u), the total number of transaction types (N_t), the transaction names of the transactions in the profile, the maximum Euclidean distance between any two users (or a pair of users) within the profile, the mean of all Euclidean distances for each pair of users, the total number of records identified as multivariate outliers (N_{mo}), based on the Euclidean distance threshold, and for comparison, we have also included the most extreme outliers identified from the visual impression of the boxplots (N_{me}) from Table I.

The Euclidean distance was calculated for pairs of users within the five transaction profiles, to detect multivariate outliers. Based on the mean of the highest Euclidean distances in each of the profiles (shown in Column 5 of Table II), we set the Euclidean distance threshold value to 5.3. Consequently, as the threshold value is increased, the total number of records identified as multivariate outliers decreased – representing only the most anomalous users. However, with different datasets, different numbers of users, transaction types and profiles, it may be useful for a fraud examiner or an auditor to deduce an

TABLE II. MULTIVARIATE OUTLIERS

tp_i id	Users in tp (N_u)	No. of t (N_t)	Transaction (t) names	Highest ED	Mean ED	Multi-variate outliers (N_{mo})	Uni-variate outliers (N_{me})
1302	11	2	XK01, XK02	4.24	1.40	None	None
1297	25	2	invoice_approved, requisition	5.6	1.97	2	1
1326	583	2	goods_received, requisition	5.6	1.20	1	1
1327	663	3	goods_received, invoice_approved, requisition	5.7	2.14	1	1
1320	1892	2	goods_received, invoice_approved	5.5	1.70	1	None
						5	

ISSN (Print): 2204-0595
ISSN (Online): 2203-1731

appropriate threshold value from the highest Euclidean distances in each transaction profile. In our dataset, the minimum Euclidean distance between a pair of users, in all five transaction profiles is zero. This may occur, for example when users in a transaction profile perform the same two or three transaction types with the same frequency (in our case a frequency value of one or two). From the user pairs that have a $\Delta EDthreshold$ greater than 5.3, we identify and flag users that occur the highest number of times within each transaction profile. A total of five multivariate outliers are detected using the Euclidean distance measuring technique. No outliers are flagged in tp ids 1302 as the highest ED value is below 5.3. In tp 1297, two users are flagged as they appear the highest number of times amongst all user pairs which have a Euclidean distance of 5.3 in this profile. Interestingly, all the other 23 users in the profile have done both the transaction types less than or equal to 13 times, however the two flagged users have performed one of the transactions types only once and the other, 30 and 26 times. These users are suspicious as one of the transaction types has the highest frequency values in the profile, whilst the other has only been executed once.

For tp 1326, amongst the 66 user pairs that were above the $\Delta EDthreshold$, user 'agKRcVoNk' appeared 19 times, indicating that, this particular user's activities are very different and potentially fraudulent when compared to all other users in the dataset. On manual analysis of the transaction profile, compared with all other 579 users in the transaction profile, the flagged user (anonymized user name: 'agKRcVoNk'), appears most anomalous (as shown in Table III). Table III, shows the anonymized username and the transaction frequencies of the goods received and requisition transaction types. We pick a sample of three users for demonstration purposes, other users amongst the 579 in the transaction profile exhibit a similar behaviour. It may be observed that while three of the users ('cpRfDYZ0X', 'U13GQSjxJ' and 'arGVUHzWg') have performed the goods received transaction most during the period for which the dataset has been extracted, user: 'agKRcVoNk' is the user who has executed the requisition transaction the most. One assumption may be that this user's main responsibility is to perform the requisition transaction as part of their job function, however, the frequency usage appears anomalous and merits further investigation.

The technique flags one user in tp 1327 as anomalous, where the ED is greater than 5.3. This particular user appears around 200 times in the user pairs which have a ΔEDthreshold greater than 5.3. On manual investigation of User 'w1ElHuBUAp' we found that the transaction usage pattern differed considerably compared to other users within the transaction profile. One particular transaction type has been performed a lot more times than the other two transaction types (as depicted in Table IV). Table IV presents the anonymized username and the transaction frequencies of the goods received, requisition and invoice approved transactions, performed by the user. For an auditor or fraud examiner, this user is perhaps the most interesting or potentially suspicious due to the extent of their involvement in the total number of generated user pairs.

TABLE III. MULTIVARIATE OUTLIERS IN TP ID 1326

Anonymized user name	t_1(goods_received)	t_2(requisition)
cpRfDYZ0X	78	3
agKRcVoNk	**3**	**75**
U13GQSjxJ	25	1
arGVUHzWg	51	1

TABLE IV. MULTIVARIATE OUTLIER IN TP ID 1327

Anonymized user name	t_1(goods_ received)	t_2(requisition)	t_3(invoice_ approved)
w1ElHuBUAp	3	97	3

TABLE V. MULTIVARIATE OUTLIERS IN TP ID 1320

Anonymized user name	t_1(goods_ received)	t_2(invoice_ approved)
SBjThyxGU	6	724

In tp 1320, user 'SBjThyxGU' is flagged. On manual analysis of the transaction frequency values of this user, we found some very unusual behaviour - where the goods_received transaction is only performed six times, while the invoice_approved transaction has been performed 724 times - being the highest frequency in this transaction profile (see Table V). Table V presents the anonymized username and the transaction frequencies of the goods received and invoice approved transaction types performed by the user. This user needs to be investigated by auditors to confirm if the behaviour is fraudulent or not.

In the next section, we evaluate and discuss the effectiveness of employing the k-means clustering algorithm for the detection of multivariate outliers within transaction profiles.

D. Discussion of Anomalies Detected With K-means

As mentioned earlier, the experiments are conducted using Matlab's built-in k-means function, which takes two parameters as input – k, the number of clusters and *X*, the dataset. We select five transaction profiles from the dataset for our multivariate outlier analysis. For each of these five transaction profiles: tp's 1302, 1297, 1326, 1327 1320, which have at least two transaction types and ten users. We perform an array of experiments with a different number of pre-selected clusters (k values). We use Matlab's silhouette plots to measure the quality and strength of each particular grouping of records or cluster, for the k parameter values of 2-5. The silhouette value for each point in the dataset, is a measure of how similar that point is to points in its own cluster, compared to points in other clusters [24]. The silhouette values range from:

- +1, representing points that are very distant from neighboring clusters, through
- 0, indicating points that are not distinctly in one cluster or another, to
- -1, signifying points that are probably assigned to the incorrect cluster.

In other words, +1 indicates that the point is well classified within a cluster - implying that the point is much closer, on

ISSN (Print): 2204-0595
ISSN (Online): 2203-1731

average, to the other members in its cluster than to the members of the neighbouring clusters, 0 indicates that the object lies between clusters and -1 indicates that the point is poorly classified within a cluster - that is, on average, the members of the neighbouring cluster are closer to the point than the members of its own cluster [15].

In order to create clusters that are as compact as possible, we repeat the clustering algorithm 500 times using the 'replicates' parameter in Matlab. This ensures that the algorithm returns the best results with the optimal cluster, for each transaction profile.

1) Detecting Anomalies in Transaction Profile 1302

Tp 1302 consists of 11 users who have all performed two transaction types: XK01 and XK02. Figures 2 (a), (b), (c) and (d) show the silhouette plots depicting different grouping structures for k partitions of 2 - 5. The y-axis shows the cluster number and the x-axis displays the silhouette values. Each grouping or cluster indicates the number of users in that cluster (represented by the vertical bars) and their silhouette value (suggesting how similar a point is to points in its own cluster compared to points in other clusters). The users that belong to each cluster in the silhouette plots (cluster structures shown in Figure 2) are summarized in Table VI (a), (b), (c) and (d), respectively. At a quick glance of the silhouette plots in Figure 2, it can be seen that a cluster of size 3 suits well transaction profile 1302 as: (i) there are no negative silhouette values – which signify that points are wrongly assigned to a cluster, (ii) most points indicate a silhouette value of around 1, signifying that the clusters are well separated from each other, and (iii) the grouping structure efficiently detects a cluster with only one observation - defining an outlier.

2) Detecting Anomalies in Transaction Profile 1297

Next, we examine transaction profile 1297, which comprises 25 users who have all performed two transaction types: requisition and invoice approvals. The silhouette plots constructed for *k* values of 2 to 5 are shown in Figure 3 and the number of users within each cluster are summarized in Table VII. Looking at this set of plots, it may be observed that most users (in Figures 3 (a), (b), (c) and (d)) exhibit similar behaviour and belong to a single cluster. Perhaps a high number of users in an individual cluster in all plots, may suggest normal or legitimate behaviour. Plots (b) and (c), that is, where the number of clusters is 3 and 4 respectively, show a few members that are allocated to unrelated clusters. On closer investigation of the results, we found that the outlying observations in C1(in plot (a)), C1 (in plot (b)), C3 (in plot (c)), and C2, C3, (in plot (d)) are exactly the same 3 users. It may be observed that a cluster of size 5 is well suited for transaction profile 1297, where the outlying observations are in C1and C3. On further investigation of these two clusters, we verified that they indeed indicate abnormal or irregular behaviour, since the transaction types have been performed with the highest frequency values within the transaction profile.

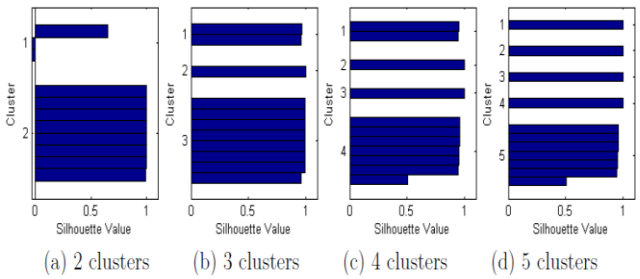

Fig. 2. Silhouette plots depicting *k* partitions of 2 - 5 in tp 1302.

TABLE VI. TABLES SHOW USERS IN EACH CLUSTER FOR 4 DIFFERENT GROUPINGS IN TP 1302

C_i	u_i
C_1	u_1, u_2, u_4
C_2	$u_3, u_5 - u_{11}$

(a)

C_i	u_i
C_1	u_1, u_4
C_2	u_2
C_3	$u_3, u_5 - u_{11}$

(b)

C_i	u_i
C_1	u_1, u_4
C_2	u_2
C_3	u_3
C_4	$u_5 - u_{11}$

(c)

C_i	u_i
C_1	u_1
C_2	u_2
C_3	u_3
C_4	u_4
C_5	$u_5 - u_{11}$

(d)

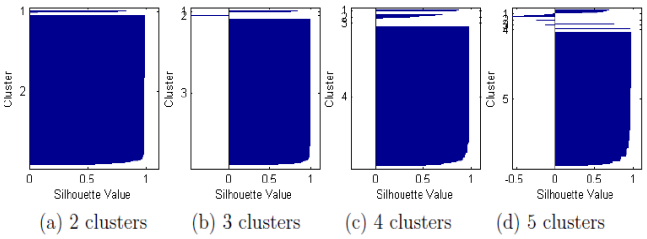

Fig. 3. Silhouette plots depicting *k* partitions of 2 - 5 in tp 1297.

TABLE VII. TABLES SHOW USERS IN EACH CLUSTER FOR 4 DIFFERENT GROUPINGS IN TP 1297

Number of clusters (k)	Number of users in each cluster				
	C_1	C_2	C_3	C_4	C_5
2	10	573	-	-	-
3	10	4	569	-	-
4	7	21	3	552	-
5	30	3	9	4	537

3) Detecting Anomalies in Transaction Profile 1326

Transaction profile 1326, consists of 583 users, who have performed two transaction types: goods received and requisition. Figure 4 shows the discovered clusters for 2-5 partitions and Table VIII, presents the number of users in each of the four different groupings. Since the number of users within this transaction type is higher, we created silhouette plots for 2 to 10 clusters. However, when the cluster size was set to ≥5, more and more observations had a negative silhouette value (therefore, Figure 4 depicts clusters 2 to 5). From the plots, it can be observed that for this particular transaction profile, a cluster size of 4 is optimal. From the visual patterns revealed in the silhouette plots, a similar trend to transaction profile 1297 may be observed - where most users belong to a single cluster. Looking at the large number of users in these single clusters (more than 500 users in each as shown in Table VIII), it may be assumed that these user activities represent normal behaviour. As users in the transaction profile have performed the same transaction types, they only differ in terms of the usage or frequency of those transaction types.

ISSN (Print): 2204-0595
ISSN (Online): 2203-1731

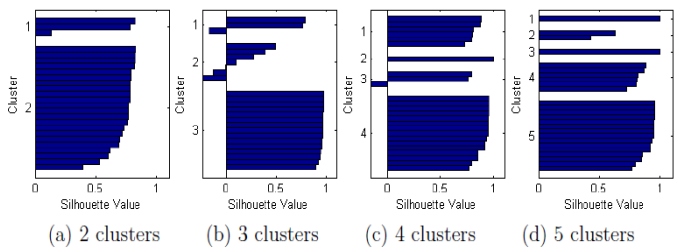

Fig. 4. Silhouette plots depicting *k* partitions of 2 - 5 in tp 1326.

TABLE VIII. TABLES SHOW USERS IN EACH CLUSTER FOR 4 DIFFERENT GROUPINGS IN TP 1326

Number of clusters (k)	Number of users in each cluster				
	C_1	C_2	C_3	C_4	C_5
2	3	22	-	-	-
3	3	7	15	-	-
4	6	1	3	15	-
5	1	2	1	6	15

From an auditor's viewpoint, these small clusters encompassing about 10 overlapping users across all plots (in Figure 4), portray potentially suspicious behaviour and merit closer investigation. On a manual investigation of these observations in the dataset, we found that these users have performed both transaction types with exceptionally high frequencies.

4) Detecting Anomalies in Transaction Profile 1327

This transaction profile is associated with 663 users and three distinct transaction types, namely goods_received, invoice_approved and requisition. Similarly, in transaction profile 1327, most users belong to an individual, large cluster (C2 as shown in Figure 5 (a)). However, it is interesting to note that when the value of k is set to 3 another group of around 130 users is fragmented from the larger cluster, implying that there

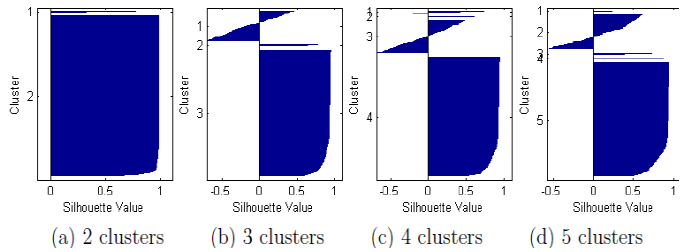

Fig. 5. Silhouette plots depicting *k* partitions of 2 - 5 in tp 1327.

TABLE IX. TABLES SHOW USERS IN EACH CLUSTER FOR 4 DIFFERENT GROUPINGS IN TP 1327

Number of clusters (k)	Number of users in each cluster				
	C_1	C_2	C_3	C_4	C_5
2	8	655	-	-	-
3	129	8	526	-	-
4	12	5	138	508	-
5	5	150	12	4	492

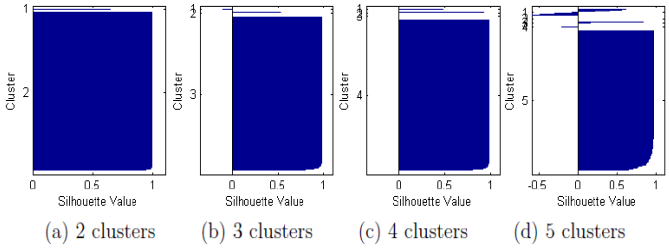

Fig. 6. Silhouette plots depicting *k* partitions of 2 - 5 in tp 1320.

TABLE X. TABLES SHOW USERS IN EACH CLUSTER FOR 4 DIFFERENT GROUPINGS IN TP 1320

Number of clusters (k)	Number of users in each cluster				
	C_1	C_2	C_3	C_4	C_5
2	9	1883	-	-	-
3	9	18	1865	-	-
4	18	5	4	1865	-
5	93	7	16	8	1768

may be at least two main categories of users in this profile that differ in their frequency of transaction types. When the number of partitions is set to 4 (see plot (c)), another small group of users emerge into a separate cluster.

The results are most interesting when the number of clusters is set to 5 (in plot (d)), whereby C1, C3 and C4 represent three small groups of users - representing the characteristics of an outlier. Based on the empirical analysis of the silhouette plots, the results were identical if k equals 6 - 10. On a manual verification of the results, we found that amongst the 3 transaction types, the 4 users in the smallest cluster, C4 (where k = 5), have performed only one transaction type (invoice approved) with unusually high frequency values (of around 100), while the other two transactions have been performed about 5 times. Likewise, the 5 users in C1 and the 12 users in C3, have relatively large frequency values for only the requisition transaction. Also of note are the 150 users in C2 who have on an average performed all transaction types 30 - 60 times, whilst the 492 users in C5 have predominantly executed all transaction types with a frequency of 20 or less.

5) Detecting Anomalies in Transaction Profile 1320

Transaction profile 1320, contains the maximum number of users (1,892) in the dataset, who have performed the goods received and invoice approval transactions. Parallel to the previous four cases (that is, tp 1302, 1297, 1326 and 1327), most members in this profile belong to a single cluster (see Figure 6). As mentioned earlier, we believe that a single large cluster may indicate the normal usage or frequency of transaction types for the particular organizational role, represented by this transaction profile. The set of generated clusters (see Table X) are well-defined and the groupings remain unaffected as the data is further divided (where $k = 2 - 4$). Two outlying clusters with 9 and 18 users are detected in the four different grouping structures, with strikingly high frequency values. Nevertheless, an additional cluster of 93 users occurs when *k* is set to 5, these users are grouped into a separate cluster as they have performed transactions with a

ISSN (Print): 2204-0595
ISSN (Online): 2203-1731

frequency of 25- 30. The users in the largest cluster, C5, have executed all transactions with a frequency of about 10.

In summary, we observed that the multivariate analysis using the Euclidean distance and the k-means clustering algorithm yielded some similar results. For example: lets consider transaction profile 1297, where user 'SoaJjPi1l' has performed two transaction types: requisition and invoice_approved. This particular user has been flagged for its frequency usage (26) of the requisition transaction in both methods: the multivariate Euclidean distance (detected in tp 1297 – listed in Table II, Row 2) and the k-means analysis (depicted by C1 in Figure 3 (c) and Table VII, Row 1). Also user '862BhD247' who has performed the invoice_approved transaction type 30 times within the same transaction profile 1297 has been detected in both the multivariate techniques: the Euclidean distance (detected in tp 1297 – listed in Table II, Row 2) and the k-means analysis (depicted by C1 in Figure 3 (c) and Table VII, Row 1).

Both multivariate approaches have their pros and cons, however, we believe that for our dataset, the ED multivariate method is more suited, primarily for two reasons. (1) With the ED multivariate analysis, the threshold is automatically set based on the mean of the highest distances between users within all transaction profiles in the dataset. While for the clustering approach, the number of clusters needs to be pre-determined and manually set. (2) With the k-means analysis, the number of pre-determined clusters effects the number and sensitivity of alerts generated. On the other hand, the ED approach detected a small number of more sensitive alerts.

V. Conclusion and Future Work

In this paper, we have presented two main contributions: (a) the detection of univariate anomalies within transaction profiles using boxplots and (b) the detection of multivariate anomalies using Euclidean distance. The experimental results suggested that the techniques have successfully tagged anomalous, potentially fraudulent behaviour in the dataset. The flagged outliers were manually investigated and verified from the dataset.

An inherent limitation of the approach is that boxplots are an informal method, that is, they are not statistically verified for the detection of outliers. Nevertheless, the constraints and assumptions set by statistical approaches (as mentioned previously), has made the use of boxplot suitable for many large practical applications [22].

The multivariate analysis using the ED and the k-means algorithm also yielded many significant results. An auditor or fraud examiner would use all three techniques together to detect different types of anomalies (as each technique has its own benefits). While each alert detected is significant in its own right, more interesting users may be flagged in two or more techniques.

In addition, for the k-means clustering technique we observed a smooth trade-off, where if the number of clusters is reduced, users quickly collapse into unrelated groups, while if the number of preselected clusters is increased, users tend to fall into their own single clusters. In addition, it can be noted that most users in the dataset collapsed into a single large cluster - suggesting that users in the dataset have performed related activities (a similar phenomenon occurs with all our clustering experiments). The smaller clusters, with a few observations represent data points that are far apart from the vast majority of observations. For an auditor, the users in these outlying clusters may be interesting to investigate, as they have performed transaction types with markedly high frequencies, thus portraying unusual, and potentially suspicious behaviour.

Our future work will focus on incorporating time analysis into the anomaly detection. At the moment, our transaction profiles are based on transaction types and frequencies only without regard for the period during which the transaction types are performed. This will naturally affect both the nature of transaction profiles and also the processing involved. It will provide the benefit of being able to detect much more subtle differences - possibly anomalies - amongst users.

References

[1] A. G. Little and P. J. Best, "A framework for separation of duties in an SAP R/3 environment," *Managerial Auditing Journal*, vol. 18, no. 5, pp. 419–430, 2003.

[2] P. Bingi, M. K. Sharma, and J. K. Godla, "Critical issues affecting an ERP implementation," *Information Systems Management*, vol. 16, no. 3, pp. 7–14, 1999.

[3] Y. F. Musaji, *Integrated Auditing of ERP Systems*. New York: John Wiley and Sons, 2002.

[4] J. D. O'Gara, *Corporate Fraud: Case Studies in Detection and Prevention*. New Jersy: John Wiley and Sons, 2004.

[5] R. Bolton and D. Hand, "Statistical fraud detection: A review," *Statistical Science*, vol. 17, no. 3, pp. 235–249, 2002.

[6] W. S. Albrecht, C. Albrecht, and C. C. Albrecht, "Current trends in fraud and its detection," *Information Security Journal: A Global Perspective*, vol. 17, no. 2, pp. 2–12, 2008.

[7] P. J. Best, "Computer assisted auditing techniques," Queensland University of Technology, Brisbane, QLD, 2007.

[8] A. Clark, G. Mohay, and P. Best, "Integrated financial fraud detection in enterprise applications," Information Security Institute, Queensland University of Technology, Brisbane, QLD, 2005.

[9] J. T. Wells, *Fraud Casebook: Lessons from the Bad Side of Business*. New Jersy: John Wiley and Sons, 2007.

[10] R. Khan, M. Corney, A. Clark, and G. Mohay, "Transaction mining for fraud detection in ERP Systems," *Industrial Engineering and Management Systems*, vol. 9, no. 2, pp.141-156, 2010.

[11] V. Barnett and T. Lewis, *Outliers in Statistical Data*, 3rd ed. England: Wiley and Sons, 1994.

[12] E.W.T. Ngai, Y. Hu, Y.H. Wong, Y. Chen, and X. Sun, "The application of data mining techniques in financial fraud detection: A classification framework and an academic review of literature," *Decision Support Systems*, vol. 50, no. 3, pp.559-569, 2011.

[13] R. J. Bolton and D. J. Hand, "Unsupervised profiling methods for fraud detection" in *Proc. Credit Scoring and Credit Control VII*, London, 2001, pp. 5-7.

[14] T. Fawcett and F. Provost, "Adaptive fraud detection," *Data Mining and Knowledge Discovery*, vol. 1, no. 3, pp.291–316,1997.

[15] P. Flach, *Machine Learning: The Art and Science of Algorithms that Make Sense of Data*. Cambridge: University Press, 2012.

[16] R. E. Shiffler, "Maximum z scores and outliers," *The American Statistician*, vol. 42, no. 1, pp.79–80, 1988.

[17] E. Acuna and C. Rodriguez, "A meta analysis study of outlier detection methods in classification," University of Puerto Rico, Mayaguez, 2004.

ISSN (Print): 2204-0595
ISSN (Online): 2203-1731

[18] J. W. Tukey, *Exploratory Data Analysis*. Addison-Wesley, 1977.

[19] D. C. Hoaglin, F. Mosteller, and J. W. Tukey, *Understanding Robust and Exploratory Data Analysis*. New York: John Wiley and Sons, 2000.

[20] J. Han and M. Kamber, *Data Mining Concepts and Techniques.Morgan Kaufmann*, 3rd ed. MA: Elsevier, 2006.

[21] O. Maimon and L. Rokach, *Data Mining and Knowledge Discovery Handbook*, 2nd ed. New York: Springer, 2005.

[22] M. Juhola, J. Laurikkala, and E. Kentala, "Informal identification of outliers in medical data," in *5th Int. Workshop on Intelligent Data Analysis Medicine and Pharmacology*, 2000.

[23] KPMG, "KPMG 2006 fraud survey," Australia, 2006.

[24] MATLAB, "K-means clustering," 2010.

[25] S. H. Oh and W. S. Lee, "An anomaly intrusion detection method by clustering normal user behavior," *Computers and Security*, vol. 22, no. 7, pp.596–612, 2003.

ISSN (Print): 2204-0595
ISSN (Online): 2203-1731