# Fraud Detection Using Transaction Data in ERP Systems

**Roheena Khan, Andrew Clark and George Mohay**

*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 transaction frequency 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. Primarily, we propose two automated techniques: (a) a univariate method, called Boxplot which is based on the sample's median and (b) a multivariate method which uses Euclidean distance, for detecting transaction frequency anomalies within each transaction profile. 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 obtain the optimal 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.**

***Index Terms*—Anomaly Detection, Enterprise Resource Planning, Fraud Detection, Security.**

## I. INTRODUCTION

Enterprise Resource Planning (ERP) systems are one of the most important IT developments to emerge in the 1990's. 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 nature 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 recent publications have discussed fraud prevention approaches such as role - based access control, segregation of duties, encryption, username and passwords, etc in different systems [3], [1] 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].

Auditors and fraud examiners generally review audit logs to detect frauds in ERP systems, which is a labour intensive task 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. ERP systems typically have role based access control over which permissions a user is allowed to perform. In addition to this, security policies assist in the segregation of related duties to reduce the opportunities to commit 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 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 anomalies in the frequency of: (a) each transaction type within a transaction profile. We identify such univariate outliers for each transaction type, using boxplots, a common graphical outlier detection technique; and (b) set of transaction types within a transaction profile - taking into account the entire set of transaction types. We identify such cases using Euclidean distance (ED), a prevalent distance measuring technique. The rationale here is to detect cases where a combination of individual transaction type frequencies in a transaction profile, may cause an outlier.

R. Khan is with the Science and Engineering Faculty, Queensland University of Technology, Brisbane 4001, Australia (telephone: 61-07-3138 4860, e-mail: rq.khan@student.qut.edu.au).

Dr A. Clark is with the Science and Engineering Faculty, Queensland University of Technology, Brisbane 4001, Australia (e-mail: a.clark@qut.edu.au).

Dr G. Mohay is with the Science and Engineering Faculty, Queensland University of Technology, Brisbane 4001, Australia (e-mail: mohaygm@gmail.com).

**ISBN: 978-0-9803267-6-5**

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 detection of univariate anomalies using Boxplots and multivariate anomalies using Euclidean distance are presented in Sections IIIA and IIIB respectively. The 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 problem of mining outliers being akin to finding a needle in a haystack). 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].

Outlier detection methods have been categorized into univariate and multivariate techniques. Most univariate outlier detection procedures are studied in the field of statistics. 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. It is used to rescale raw data into its equivalent standard score, that is in accordance with a measure of the overall data spread (known as the distributions standard deviation) [15]. For example: given $x_1, x_2, x_3 \dots, x_n$ as a sample from a dataset of size *n*, let $\bar{x}$ be the mean and *s* the standard deviation, an observation *x* is considered an outlier, if:

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

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, given that it will be far from the distribution's mean (where about 95% of the data lies), assuming 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. Although z-score mean and standard deviation estimates give a good idea of the data shape, the techniques reliance on the mean and the standard deviation makes it highly susceptible to outliers.

Another technique, that avoids the dependence on the distribution's mean and standard deviation, is called Dixon's test. The test calculates ratios of differences between an outlier (or suspect value) and it's nearest or next-nearest neighbour, compared to the range. For example: given $x_1, x_2, x_3 \dots, x_n$ as a sample from a normally distributed dataset of size *n*, where the data is arranged in ascending order, a single upper outlier is detected by:

$$T_{D(upper)} = \frac{x_n - x_{n-1}}{x_n - x_1} \qquad (2)$$

where $x_i$ is the lowest value in a dataset, $x_n$ is the highest value in the dataset and $x_{n-1}$ is the second highest value in the dataset. Once the formulae is applied, the resultant $T\_(D(upper))$ value is then compared with the pre-defined critical values for that particular sample size *n*. If the $T\_(D(upper))$ value is greater than its respective critical value, then the value in question (*xn*) is characterized as an outlier. Dixon's test can be used to detect a single lower outlier, where the data is arranged in descending order. Other versions of Dixon's test are proposed based on the different data distributions. Essentially, the test is designed to be effective in identifying a single outlier or an outlier pair (that is, two co-related outliers). Hence, the distribution and the number of expected outliers needs to be known.

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 (as required in (1)), the distribution parameters (as in (2)), the number of expected outliers (also needed in (2)) and the type (univariate or multivariate) of expected outliers [17].

In real world datasets, these factors or assumptions are often not known. Consequently, a more robust outlier detection technique is required. Thus, we adapt a well-known, graphical method for outlier detection, called boxplot. Among many other exploratory data analysis techniques, John Tukey [18] in his book, 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).

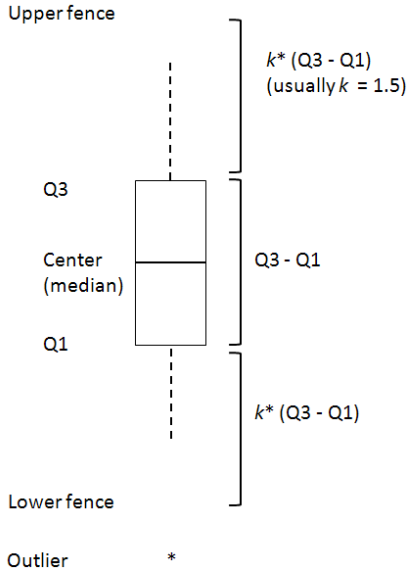


Fig. 1. Example of a boxplot (adapted from [18]).

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.

### *A. Multivariate Outlier Detection Techniques*

Similar to univariate outlier detection methods, 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 don't require the dataset to have a normal distribution. In this section, we provide 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 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:

$$ED(x_i, x_j) = \sqrt{(x_{i1} - x_{j1})^2 + (x_{i2} - x_{j2})^2 + \cdots + (x_{in} - x_{jn})^2} \quad (3)$$

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. The complexity of Euclidean distance runs in polynomial time $O(n)$. In the next section, we demonstrate the role of Euclidean distance in our proposed approach for identifying multivariate outliers.

## 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 [21] 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 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.

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). 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. 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).

## IV. EXPERIMENTAL RESULTS AND DISCUSSION

The approaches was 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 2008. 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.

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 | |

The dataset is extracted from an operational environment, and any alerts identified do not necessarily indicate fraud, but represent anomalous activity. The dataset consist of real users and activities and the detection of such activities demonstrates the effectiveness of the proposed approach. A thorough manual investigation of the anomalies is carried out for verification.

### 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. 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). Transaction profiles containing one transaction type are tp 1299, 1296, 1314, 1291 and 1316 (as mentioned in Table I). From the 9,300 $N_u$ or records in Table I, the boxplot has tagged 651 univariate outliers, for the six transaction types. 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 may just needs to review one anomaly to understand whether they indicate fraud or not. Since a transaction type performed twice by 86 users, doesn't 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 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 readily be identified at a quick glance of the boxplots. From the total number of records identified as outliers with 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 users job function or role in the organization. This can be verified by examining the SAP R/3 systems 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 the might be valid users who may have changed their job function or are promoted, meaning that they are assigned a new user-id for accessing the system and the outlying value 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 ED.

### C. *Discussion of Anomalies Detected With ED*

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 or zero-mean normalization method.

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 | |

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 atleast 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 greater than 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 number of users, transaction types and profiles, it may be useful for a fraud examiner or an auditor to deduce an 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.

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 |

For tp 1326, amongst the 66 user pairs that were above the $\Delta EDthreshold$, user 'agKRcVoNk' appeared 19 times, indicating that, this particular users activities are very different and potentially fraudulent as 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 only 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.

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 |

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 $\Delta EDthreshold$ greater than 5.3. On manual investigation of these two users we found that their 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 their extent of involvement in the total number of generated user pairs.

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.

## VI. 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].

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] Adam G. Little and Peter J. Best. A framework for separation of duties in an sap r/3 environment. Managerial Auditing Journal, 18(5):419–430, 2003.

[2] Prasad Bingi, Maneesh K. Sharma, and Jayanth K. Godla. Critical issues affecting an erp implementation. Information Systems Management, 16(3):7–14, 1999.

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

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

[5] R. Bolton and D. Hand. Statistical fraud detection: A review. Statistical Science, 17(3):235–249, 2002.

[6] W. Steve Albrecht, Chad Albrecht, and Conan C. Albrecht. Current trends in fraud and its detection. Information Security Journal: A Global Perspective, 17:2 – 12, 2008.

[7] Peter J. Best. Computer assisted auditing techniques. Technical report, Queensland University of Technology, 2007.

[8] Andrew Clark, George Mohay, and Peter Best. Integrated financial fraud detection in enterprise applications. Technical report, Information Security Institute, Queensland University of Technology, 2005.

[9] Joseph T. Wells. Fraud Casebook: Lessons from the Bad Side of Business. John Wiley & Sons, 2007.

[10] Roheena Khan, Malcolm Corney, Andrew Clark and George Mohay. Transaction mining for fraud detection in ERP Systems. Industrial Engineering and Management Systems, 9(2), pp. 141-156, 2010.

[11] Toby Lewis Vic Barnett. Outliers in statistical data. Wiley & Sons, 1994.

[12] E.W.T. Ngai, Yong Hu, Y.H. Wong, Yijun Chen and Xin Sun. The application of data mining techniques in financial fraud detection: A classification framework and an academic review of literature. Decision Support Systems, 2010.

[13] Richard J. Bolton and David J. Hand. Unsupervised profiling methods for fraud detection. Credit Scoring and Credit Control, VII, 2001.

[14] Tom Fawcett and Foster Provost. Adaptive fraud detection. In Data Mining and Knowledge Discovery, pages 291 – 316. Kluwer Academic Publishers, 1997.

[15] Toby Lewis Vic Barnett. Outliers in statistical data. Wiley & Sons, 1994.

[16] Ronald E. Shiffler. Maximum z scores and outliers. The American Statistician, 42(1):79–80, 1988.

[17] Edgar Acuna and Caroline Rodriguez. A meta analysis study of outlier detection methods in classification. Technical report, University of Puerto Rico at Mayaguez, 2004.

[18] John Wilder Tukey. Exploratory data analysis. Addison-Wesley Pub. Co, 1977.

[19] John Wilder Tukey, David Caster Hoaglin, Frederick Mosteller. Understanding robust and exploratory data analysis. John Wiley & Sons, 2000.

[20] Jiawei Han and Micheline Kamber. Data Mining Concepts and Techniques. Morgan Kaufmann, 2006.

[21] Ronald E. Shiffler. Maximum z scores and outliers. The American Statistician, 42(1):79–80, 1988.

[22] Martti Juhola Jorma Laurikkala and Erna Kentala. Informal identification of outliers in medical data. In 5th International Workshop on Intelligent Data Analysis in Medicine and Pharmacology, 2000.

[23] KPMG. Kpmg 2006 fraud survey. Technical Report 06.06.2009, 2006.