<table>
<thead>
<tr>
<th>Title</th>
<th>Toward a new cloud-based approach to preserve the privacy for detecting suspicious cases of money laundering in an investment bank</th>
</tr>
</thead>
<tbody>
<tr>
<td>Authors(s)</td>
<td>Le-Khac, Nhien-An; Kechadi, Tahar</td>
</tr>
<tr>
<td>Publication date</td>
<td>2014-06-12</td>
</tr>
<tr>
<td>Conference details</td>
<td>International Conference on Computational Science, (ICCS 2014), Cairns, Australia, 10-12 June 2014</td>
</tr>
<tr>
<td>Link to online version</td>
<td><a href="http://www.iccs-meeting.org/iccs2014/">http://www.iccs-meeting.org/iccs2014/</a></td>
</tr>
<tr>
<td>Item record/more information</td>
<td><a href="http://hdl.handle.net/10197/6553">http://hdl.handle.net/10197/6553</a></td>
</tr>
</tbody>
</table>
Toward a new cloud-based approach to preserve the privacy for detecting suspicious cases of money laundering in an investment bank

NhienAn LeKhac¹, and M-Tahar Kechadi¹

¹School of Computer Science & Informatics, University College Dublin
Belfield, Dublin 4, Ireland
{an.lekhac,tahar.kechadi}@ucd.ie

Abstract. Today, money laundering poses a serious threat not only to financial institutions but also to the nations. This criminal activity is becoming more and more sophisticated and seems to have moved from the cliché of drug trafficking to financing terrorism and surely not forgetting personal gain. Most international financial institutions have been implementing anti-money laundering solutions to fight investment fraud. On the other hand, cloud-based applications are merging daily and bringing to clients with lower cost of platforms and data storage, greater scalability and improved business continuity. Hence, more financial institutions aim to move their IT infrastructure to the cloud. However, accessing directly to the customer transaction datasets by a third party could be a confidential issue. This approach is more severe when these solutions are built by collaborating partners. Traditional methods are based on data access agreement but there is still a risk of infringing privacy. In order to preserve the privacy of datasets, different data disguising methods have been proposed. Nevertheless, analysing disguised datasets is a performance issue in the context of detecting suspicious money laundering cases where the real value of data has an important impact. Indeed, the results of analysis could also be a privacy issue. Within the scope of a collaboration project for developing a new cloud-based solution for the Anti-Money Laundering Units in an international investment bank, in this paper, we propose new cloud-based approach using data disguising methods applied in analysing transaction datasets. We also show that the creating relevant dimensions from the current ones is efficient for analysing transaction datasets in terms of both detecting suspicious case and privacy preserving.

Keywords: SaaS, cloud-based framework, privacy preserving; data mining; data disguising; anti-money laundering; clustering

1 Introduction

Money laundering (ML) is a process to make illegitimate income appear legitimate; this is also the process by which criminals attempt to conceal the true origin and
ownership of the proceeds of their criminal activity. Through ML, criminals try to convert monetary proceeds derived from illicit activities into “clean” funds using a legal medium such as large investment or pension funds hosted in retail or investment banks. This type of criminal activity is getting more and more sophisticated and seems to have moved from the cliché of drug trafficking to financing terrorism and surely not forgetting personal gain. Today, ML is the third largest “Business” in the world following the Currency Exchange and the Automobile Industry. Nowadays, it poses a serious threat not only to financial institutions but also to the nation. Some risks faced by financial institutions can be listed as reputation risk, operational risk, concentration risk and legal risk. At the society level, ML could provide the fuel for drug dealers, terrorists, arms dealers and other criminals to operate and expand their criminal enterprises. Briefly, nations care about ML because they care about their political and economic stability. Hence, the governments, financial regulators require financial institutions to implement processes and procedures to prevent/detect money laundering as well as the financing of terrorism and other illicit activities that money launderers are involved in. Therefore, anti-money laundering (AML) is of critical significance to national financial stability and international security. Traditional approaches to AML followed a labour-intensive manual approach because ML is a sophisticated activity with many way of laundering money. Furthermore, given that the volume of banking data and transactions have increased in number of ways, such approaches need to be supported by automated tools for detecting ML’s pattern.

Recently, there are AML approaches based on data mining techniques (DM) [4] that have been proposed and discussed in literature. Most of these approaches try to recognize ML patterns by different techniques such as support vector machine [5], correlation analysis [6], histogram analysis [6][9], clustering [8], etc. They aim to provide techniques for detecting a variety of ML by exploring a massive dimensionality of datasets including customers x accounts x products x geography x time. Despite these approaches are efficient, they just take into account testing datasets created artificially and typically. Besides, there has been growing concern that the use of DM technology may violate individual privacy when they access and analyse real datasets [16]. This problem becomes more and more severe as solutions provided by third-party companies, Even though these datasets are protected by data access agreement, there is a risk of infringing privacy, such as customer information [7], transactions.

Privacy preserving data mining (PPDM) is a new research domain, which has attracted the attention of many researchers in the last decade. This research field aims at developing models and techniques about aggregated data without direct access to all detailed information of individual transactions. One of the most useful methods applied in PPDM is the perturbation or data disguising that distorts datasets before they are mined with the intention of keeping the performance of data mining techniques. Some techniques can be listed as randomisation [17], cryptography-based [18], transformation [19], creating of new dimensions, etc. In our recent work [8], we show that clustering techniques can be applied to detect suspicious cases of ML. In order to apply efficiently clustering techniques in PPDM, data disguising methods should preserve the distance between data elements. Besides, there is still little
research of applying these methods on real datasets such as customer transaction from a bank.

Today, cloud-based applications and new capabilities are emerging daily and bringing them lower cost of entry, pay-for-use processor and data-storage models, greater scalability, improved performance, ease of redundancy and improved of business continuity. Global competitions, dynamic markets, and rapid development in the information and communication technologies are some of the major challenges in today’s financial industry. Hence, more and more financial institutes select cloud computing as a solution of their IT platforms and services. However, privacy preserving in accessing to customer data is a big issue with these institutes especially with the banks it would effect their reputation because accessing directly to the customer transaction datasets by a third party is a confidential issue. This approach is more severe when these solutions are built by collaborating partners. Traditional methods are based on data access agreement but there is still a risk of infringing privacy. In order to preserve the privacy of datasets, we should look at different data disguising methods. Besides, the results of data analysis, especially in the case of money laundering are also confidential data that are also subject to preserve the privacy.

Within the scope of a collaboration project for developing a new cloud-based solution for the Anti-Money Laundering Units in an international investment bank, we present a framework for cloud-based solution to detect the suspicious cases of ML, it can also preserve the privacy for confidential data. We propose moreover a new approach to disguise data: creating new dimensions of transaction datasets. This new approach to preserve the privacy for not only the transactional data but also the analysing results from using data mining techniques. We also compare our solution with another method of disguising data: transformation as this method allows us to distort data by preserving the distance between elements that is important in detecting suspicious cases of ML. We also show that in our approach where new dimensions created appropriately from current ones can be efficiently used to analyse transaction datasets.

The rest of this paper is organised as follows: Section 2 deals with background of our research. Section 3 presents our approach of a cloud-based framework for detecting ML. In Section 4, we look at methods for disguising data applied to detect suspicious cases of ML activities. We evaluate our methods with real customer transaction datasets in Section 5. We also analyse and discuss on results of our approach in this section. Finally, we conclude in Section 6.

2 Background

In this section, we firstly resume recent research on applying DM techniques for detecting suspicious patterns of ML. We then present briefly disguising techniques applied in PPDM. Next, we look at SaaS (Software as a Service) solutions for analysing transaction datasets.
2.1 Data mining techniques for AML

An approach for analysing data in AML is using support vector machine (SVM) [10]. In [11], authors proposed an extension of SVM to detect unusual customer behaviour. They present a combination of an improved RBF kernel [12] with the definition of distinct distant [13] and supervised/unsupervised SVM algorithms (C-SVM, one-class SVM). One-class SVM is an unsupervised learning approach used to detect outliers based on unlabelled training datasets, which is highly suitable for ML training sets. The advantage of this approach is that it can deal with heterogeneous datasets. Even though DM techniques show their efficiency in detecting suspicious cases of ML, they have been viewed as a threat to privacy when they need to analyse the real datasets, as the more complete and accurate the data the better the DM results. Consequently, it could lead to the potential misuse of data. In next section, we review some PPDM approaches in the literature.

2.2 Privacy Preserving Data mining

Basically, the PPDM models and techniques attempt to aggregate data without accessing to original information of individual data. The important key is to design methods that are effective without compromising security. Some of the most used techniques include: randomization, k-anonymity, cryptography, and transformation. The randomization method is a technique for privacy-preserving data mining in which noise is added to the data in order to mask the attribute values of records [17]. The noise added is sufficiently large so that individual record values cannot be recovered. Therefore, techniques are designed to derive aggregate distributions from the perturbed records. Subsequently, data mining techniques can be developed in order to work with these aggregate distributions. This method is traditionally used in surveys and is extended to PPDM. However, this method has a performance issue with outlier [17] as it treats all data points equally irrespective of their local density. Moreover, the distance between data points could not be preserved after the distorting process.

k-anonymity method [18] reduces the granularity of data representation with the use of techniques such as generalization and suppression. This granularity is reduced sufficiently that any given record maps onto at least $k$ other records in the data. In the method of generalization, the attribute values are generalized to a range in order to reduce the granularity of representation. In the method of suppression, the value of the attribute is removed completely. It is clear that such methods reduce the risk of identification with the use of public records, while reducing the accuracy of applications on the transformed data. However, detecting suspicious cases of ML with insufficient information on datasets could be a performance issue.

Oliveria et al [19] proposed an isometric transformation approach to disguise datasets in preserving the distance among them in order to receive the better accuracy of clustering results. Isometric transformation is a class of geometric transformation. An important feature of this transformation is that distances between objects are preserved in the moving progress in an $n$-dimensional Euclidean metric. Its formal definition can be found in [20]. Three popular isometric transformations are: translation,
rotation and reflection. In this approach, distances among data objects are preserved. However, it is difficult to analyse the behaviour of customers by using rotated values (distort values). We can, of course, apply clustering algorithms on these distort values but we need some steps of encrypt and decrypt on datasets, post-processing as well as interactive with AML experts to detect suspicious cases of ML. This approach has a performance issue when the analysing is carried out outside the financial institute.

2.3 SaaS for analysing transaction datasets

Recently, there are many SaaS solutions for analysing transaction datasets [21][22][23]. Most of them focus on analysing transaction datasets for application such as sale prediction, etc. However, to the best of our knowledge, there is not any SaaS developed for AML.

3 Cloud-based solution for detecting ML

In this section, we describe our cloud-based framework for detecting suspicious cases of ML. Our framework deals with two different scenarios of data storage: inside and outside of a financial institution. We discuss firstly challenges of a SaaS for analysing data in the context of privacy preserving.

3.1 Challenges of a SaaS for analysing confidential data

In the first scenario, confidential data is locally stored inside the financial institutions. So, they can use a SaaS AML solution to analyse their data instead of buying...
AML software. In fact, using SaaS in this case could lead to security issues as SaaS solutions communicate periodically with servers to exchange data/information. This communication can be performed over a security channel such as SSL/TLS to guarantee the privacy. However, the financial institutions have also to run their own data centre.

In the second scenario, confidential data can be stored in the cloud by using cloud-based data storage service. Although the cloud-based data storage has many advantages such as scalability, reliability, ease of access, it has also issues that some users will never feel comfortable with their data in the cloud such as its performance, security and data orphans. It is more severe in the case of confidential data of financial organisations. Users may abandon data in cloud storage facilities, leaving confidential data at risk. Besides, the sharing storage devices across multiple customers can lead to the data leaking. For example, when a file system deletes a file on disk, it simply marks the locations within which the file resides as available for use to store other files. If another customer comes along and allocates space on the disk for storage, s/he can examine the allocated space and may have access to previous deleted confidential data. Data wiping [24] can be used in this case to overwrite deleted files. We can have moreover different kinds of attacking such as: Distributed Denial-of-Service (DDoS), Packet Sniffing, Man-in-the-Middle, etc., [24]. A popular solution is to encrypt the data for both data storage in the cloud and data communication over the cloud. Data is only decrypted at end-point user side (Fig.1). In this case, if financial institutions use a SaaS AML to analyse their cloud-based storage data, this SaaS can only read encrypted data. However, in order to analyse data, SaaS solution has to decrypt it inside the cloud and this task also raises an issue of security. In this case, SaaS should have the capacity of analysing directly encrypted data and therefore techniques of privacy preserving data mining should be considered.

In fact, there is one more solution for SaaS AML approach where it can download data from the cloud to the end-user platform and perform required analysis. However, this solution does not scale well with large financial datasets and it has to deal with the same issues as the first scenario above.

3.2 Cloud-based framework for detecting Money Laundering

We present in this section our cloud-based framework for building AML solutions. As shown in Figure 2, in our framework, we assume that users store their data in cloud-based storage. However, this data is not encrypted by public-private keys paradigm but by privacy preserving techniques that will be described in Section 4. We also call these techniques as disguising data techniques. We have two scenarios here. For the first scenario, data is moved from the local data centre to cloud and the second scenario presume that data already exists in the cloud. In the first scenario, data is disguised before being sent to the cloud (Fig.2 (1), (8)). The updating of data consists of updating for both transactional data and disguised data. In the second one, a tool is developed to integrate in cloud-based storage application to disguise confidential data. Our framework supports to both two scenarios. We describe it with more details in following paragraphs.
There are two main components in our framework: data analysis (Fig.2 (II)) and data conversion (Fig.2 (I) & (III)). Our data conversion module can deal with two different scenarios as follows:

**Scenario 1: Moving data centre to cloud-based storage.** In this case, users want to move their data centre into the cloud. Normally, they use services from the cloud providers to perform this process. They can apply at the same time our disguised data application to distort their local data (Fig.2 (1)) and store it in the cloud (Fig.2 (8)). This disguised data application is developed as a plug-in (Fig.2 (III)) that can be integrated into not only our SaaS AML solution (SAML) but also into migrating solutions offered by cloud providers.

**Scenario 2: Disguising cloud-based data.** In this case, again, our disguised data application can be plugged-in to not only our SAML but also other data management tool of cloud providers to convert customers’ data to distort data (Fig.2 (1)).

In both scenarios, to distort data, our data conversion component uses methods discussed in the next section. The data analysis component (Fig.2 (II)) consists of statistical and data mining methods to analyse the disguised data to detect suspicious cases of ML.

### 4 Approaches for disguising data

In this section, we describe our approach for disguising data, which is implemented in our data-conversion component. This approach consists of creating new dimensions and the main purpose of this method is to disguise datasets in preserving the distance among them in order to receive the accuracy of both statistics and clustering results. We focus on statistics and clustering results because in the data analytic component of our framework, the AML method uses these two techniques.
Instead of analysing directly sensitive data, different cryptographic methods can be used to encrypt these datasets. However, preserving the distance among them is still a challenge. Creating new dimensions can be considered as a kind of cryptography where new dimensions are created by using existing dimensions such that their data created cannot disclose sensitive information. This approach can preserve the distance among datasets as analysing is carried out only on new created dimensions. The important challenge of this approach is that new dimensions should be meaningful in term of reflecting features needed for analysing. This challenge requires analysts not only to have a sufficient knowledge on the meaning of datasets via their dimensions but also to carry out further analysis depending on business requirements in order to create new efficient dimensions.

We present, in this section, the creation of new dimensions for the historical datasets. Generally, AML expert uses a decision tree using frequency and value of transactions as a marker; the thresholds for these markers are based on averages and the standard deviation. The following dimensions are normally considered: frequency of subscriptions, frequency of redemption, subscription value, redemption value, current balance. All these features are conditional on time: daily, weekly, monthly etc. However, the analysing directly of these dimensions would raise a concern of privacy. For instance, subscription/redemption value, balance, etc contain sensitive information. We can disguise these dimensions by either the rotation transformation as mentioned above or creating new dimensions. New dimensions in this case should reflect not only the relationship among the subscription/redemption/balance but also the behaviour of customers in the investment activities. Typically, the frequency of subscription and redemptions are less important than their value in term of the suspicious cases in the investment activities because there are many factors that influence the frequency of trades in investment activities such as political environment, market climate, fund prices, currency exchange rate, etc. In this case, values of subscription/redemption as well as the total value of the investors’ shares (balance) are significant dimensions in detecting suspicious cases. By analysing transaction datasets from different funds of the BEP bank, with the experience from its AML experts, we created new dimensions based on current significant ones above. Concretely, we defined six new dimensions: $\Delta_1, \Delta_2, \Delta_3, \Delta_4, \Delta_5$ and $\Delta_6$. $\Delta_1$ is the proportion between the redemption value and the subscription value conditional on time (daily, weekly, monthly, etc) and $\Delta_2$, the proportion between a specific redemption value and the total value of the investors’ shares conditional on time as below:

$$\Delta_1 = \frac{\alpha}{\beta} \bigg|_{\tau} \quad \text{Where} \, \alpha \leq \beta$$

$$\Delta_2 = \frac{\beta}{\theta} \bigg|_{\tau}$$

1. There are also other dimensions such as geography, nationality, type of transaction, etc. However, according to the experience of AML expert, they are secondary features comparing two the main ones listed above.

2. Real name of the bank and its fund names cannot be disclosed because of confidential agreement of the project.
where $\alpha_i$ is the subscription value and $\alpha_i \in [0, +\infty]$, $\beta_j$ is the redemption value and $\beta_j \in (0, +\infty)$, $\theta_h$ is the value of the investors shares and $\theta_h \in (0, +\infty)$, $\tau_k$ is time where $\tau_k \in \{\text{Day, Week, Month}\}$ depending on the investigating level (by day, by week or by month). Note that the value of the transactions (subscription or redemption) of each investor in an investment fund is aggregated by time (daily, weekly...). The definition of $\Delta_3$ is based on the proportion between the frequency of subscription and its average value. Concretely, if $\delta_{\tau j}$ the frequency of subscription conditional on time $\tau_j$ ($\tau \in \{\text{Day, Week, Month}\}$, $j$ is the time-step) and $\lambda$ is the average value of all $\delta$ from the first time-step until the time-step $j$. If the value of $\Delta_3$ is close to 1, then the frequency $\delta_{\tau j}$ is significantly high comparing to the average value $\lambda$. This is also a remarkable sign for a suspicious behaviour. The definition of $\Delta_4$, $\Delta_5$, $\Delta_6$ is the same as (3) and $\Delta_4$ is for the frequency of redemption; $\Delta_5$, $\Delta_6$ is for the amount of subscription and redemption respectively.

Briefly, the original datasets with subscription and redemption values with be disguised to six parameters. Table 1 gives an example of the original datasets and its relevant disguised data. By observing this table, we notice that it is very difficult to retrieve the original data from the disguised data.

Table 1. Original data (x100 USD) vs. disguised data.

<table>
<thead>
<tr>
<th>CID</th>
<th>Subscription</th>
<th>Redemption</th>
<th>$\Delta_1$</th>
<th>$\Delta_2$</th>
<th>$\Delta_3$</th>
<th>$\Delta_4$</th>
<th>$\Delta_5$</th>
<th>$\Delta_6$</th>
</tr>
</thead>
<tbody>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>10</td>
<td>12122.62</td>
<td>94594.82</td>
<td>0.0076</td>
<td>0.732</td>
<td>0.5</td>
<td>0</td>
<td>0</td>
<td>0.002</td>
</tr>
<tr>
<td>26</td>
<td>16000</td>
<td>15000</td>
<td>0.998</td>
<td>0.8</td>
<td>0.98</td>
<td>0.77</td>
<td>0.013</td>
<td>1</td>
</tr>
</tbody>
</table>

5 Experiments

Our cloud-based framework is being implemented. In this framework, the most important component is the data conversion, because it should guarantee both the accuracy of analysing data carried out by data analysis process and the privacy preserving of the data. Hence, in this paper, we evaluate first of all the performance the data conversion component. We use transactions from 2 of 15 funds administered by BEP bank with around one million transaction records of about 3 thousands customers in the last ten years. The original data is distorted in six new dimensions. Next, we perform analyse on the new distorted datasets with clustering technique and statistics technique. We compare also the results with using the transformation techniques.
We evaluate first of all the performance of our approach i.e. the capacity of analysing data based on new dimensions. As mentioned in the previous section, two most important parameters are $\Delta_1$ and $\Delta_2$. Hence, we start to evaluate these two parameters first.

By observing Figure 3, we can first notice that these new dimensions can hide sensitive information i.e. real value of investment (values of subscription/redemption/balance). Moreover, it is important to note that the $(\Delta_1, \Delta_2)$ reflects well on customer behaviours. For instance, the $\Delta_1$(week) of fund AB is in [0..0.23] and its $\Delta_2$(day) is in [0..1]. This can be explained as there are investment activities where customers redeem nearly all or all of their shares in a week ($\Delta_2$(week) $> 0.9$ and some $\Delta_2$(week) $= 1$). However, there is no subscription or subscription with a very small amount of value comparing to redemption of relevant weeks (their $\Delta_1$(week) ~

Fig. 3. Suspicious Factor ($\Delta_1$ and $\Delta_2$(week)) of different funds

Fig. 4. Suspicious Factor ($\Delta_1$ and $\Delta_2$(week)) of different funds
0). There is only one case where \( \Delta_1 \) is equal to 0.23 and its \( \Delta_2 \) is less than 0.05 i.e. there is subscription/redemption of a customer in that month (with their ratio = 0.23) but their value is small compared to the total value in shares (less than 5%). Basing on these analyses, we can conclude that there is no suspicious case in this fund on that period. A double check with AML Unit also confirms our conclusion. Meanwhile, there are cases with high value (\( \Delta_1, \Delta_2 \)) in the fund SK. Consequently, they are suspicious cases in this fund.

Next, we can apply K-means based algorithm (a modified version of K-means with fixed initial points [8]). Fig.4 shows show the clustering results of the fund SK based on \( \Delta_1 \) and \( \Delta_2 \) by week. By observing these results, we recognise that elements in Cluster2 (Fig. 4) are obtained high value in \( \Delta_1 \) and \( \Delta_2 \) and their values will be used as suspicious patterns. This analysis can be carried out outside the financial institute and clustering results are sent back where and further analyses are needed to find out the origin: they are real suspicious cases or this is only trends in investment activities such as exchange transactions [8].

![Fig. 5. Suspicious Factor (\( \Delta_1 \) and \( \Delta_2 \)(week)) of different funds](image)

Other parameters \( \Delta_3, \Delta_4, \Delta_5 \) and \( \Delta_6 \) can be used as additional parameters that can first of all support the AML expert in decision of suspicious cases of ML by combining them with the first two parameters. Besides, these parameters can also be used to hide the results of analysing. As mentioned in previous sections, financial institutes normally aim to distort not only their datasets but also hide the results of analysing as they do not want information such as there are some suspicious cases of money laundering detected by SaaS solution. Therefore, the meaning of each dimension from \( \Delta_1 \) to \( \Delta_6 \) is defined at user level, not at analysing level. So, at the analysing level, all dimensions are applied with the same techniques and send results back to users. Users can link results with the meaning of each dimension to interpreter results by using our Data interpreter component (Fig. 2 (IV)). For example, Figure 5 shows the clustering results of \( \Delta_2 \) and \( \Delta_3 \) and it is the same as \( \Delta_1 \) and \( \Delta_2 \) in terms of plotting but the their analysing are different.
We also compare with the transformation [19]. We apply the first disguising method, the rotation transformation, on our testing datasets. The number of dimensions is three: value of subscription, value of redemption and total value (balance). The rotation angle in this case is fixed at 122.5. Figure 6 shows the value of redemption/balance and value of subscription/redemption before (left) and after (right) the rotation from two funds $AB$ and $SK$ respectively. By observing these figures, we recognise that the rotation can hide the sensitive information: real value of subscription/redemption and balance. Distances among data objects are preserved. However, it is difficult to analyse the behaviour of customers by using rotated values (distort values), especially the negative values. We can, of course, apply clustering algorithms on these distort values but we need some steps of encrypt and decrypt on datasets, post-processing as well as interactively with AML experts to detect suspicious cases of ML. This approach has a performance issue when the analysing is carried out outside the financial institute.

As a conclusion, the creating of new dimensions is more efficient than the rotation transformation in the context of detecting suspicious cases of ML and it can be moreover performed outside the financial institute. This requires a strong knowledge on business requirements to decide which dimension will be created. However, it can be carried out by internal experts of financial institutes before sending disguised datasets to outside partners. Hence, our approach integrates knowledge from AML experts to create efficient and relevant dimensions.

6 Conclusion and Future Work

In the context of detecting suspicious case of ML in an investment bank, in this paper, we present a framework for cloud-based solution that can preserve the privacy for confidential data. We also propose a new approach to disguise data: creating new dimensions of transaction datasets. This method allows us to distort data by preserving the distance between elements that is the important impact to clustering analysis. We also showed that basing on knowledge of business requirements, new dimensions
created appropriately from current ones can be efficiently used to analyse transaction datasets. Moreover, this method can be applied in collaboration project where the creating of new dimensions (simple, business knowledge required) is carried out inside the financial institute. The experts from external institute can then analyse on these encrypted datasets and then send the feedback to the financial institute. Hence, cloud-based tool can use this approach for developing solutions for AML.

Our framework is developed as a SaaS. We are testing our SaaS with real world datasets. In the next step, more disguising techniques will be analysed in the same context. Experimental results for more datasets are also being produced and these will allow us to test and evaluate the robustness of our approach.

References