Toward a New Cloud-Based Approach to preserve the Privacy for Detecting Suspicious Cases of Money Laundering

N-A Le-Khac and M-Tahar Kechadi

School of Computer Science & Informatics, University College Dublin
Dublin, Ireland

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, analyzing 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 are efficient for analysing transaction datasets in terms of both detecting suspicious case and privacy preserving.

I. 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 [1]. This type of criminal activity is getting more and more sophisticated. Nations care about ML because they care about their political and economic stability. 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. Recently, there are AML approaches based on data mining techniques (DM) [2] 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 [6], correlation analysis [3], histogram analysis [3][5], clustering [8], etc. However, there has been growing concern that the use of DM technology may violate individual privacy when they access and analyse real datasets [12]. 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 [4], transactions. Besides, privacy preserving data mining (PPDM) aims at developing models and techniques about aggregated data without direct access to all detailed information of individual transactions. However, there is still little research of applying these methods on real datasets such as customer transaction from a bank.

 ISBN : 978-81-925233-5-4

Today, cloud-based applications and new capabilities are emerging daily and briging 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. 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 affect 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.

In this paper, 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 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 II deals with background of our research. Section III presents our approach of a cloud-based framework for detecting ML and methods for disguising data applied to detect suspicious cases of ML activities. We evaluate our methods with real customer transaction datasets in Section IV. We also analyse and discuss on results of our approach in this section. Finally, we conclude in Section V.

II. Background

A. Data Mining Techniques for AML and Privacy Preserving Data mining

An approach for analysing data in AML is using support vector machine (SVM) [6]. In [7], authors proposed an extension of SVM to detect unusual customer behaviour. They present a combination of an improved RBF kernel [8] with the definition of distinct distant [9] and supervised/unsupervised SVM algorithms (C-SVM, one-class SVM). Even though DM techniques show their efficiency in detecting suspicious cases of ML, they could lead to the potential *misuse* of data.

On the other hand, the PPDM models and techniques attempt to aggregate data without accessing to original information of individual data. Some of the most used techniques include: randomization [13], kanonymity [14], cryptography, and transformation [15][16]. However, these methods have a performance issue with outlier [13], distance preservation or it is difficult to analyse the behaviour of customers by using transformed values. These approaches also has a performance issue when the analysing is carried out outside the financial institute.

B. SaaS for Analysing Transaction Datasets

Recently, there are many SaaS solutions for analysing transaction datasets [17][18][19]. Most of them focus on analysis transaction datasets for application such as sale prediction, etc. However, to the best of our knowledge, there is not any SaaS developed for AML.

III. Cloud-Based Solution for Detecting ML

A. 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 center.

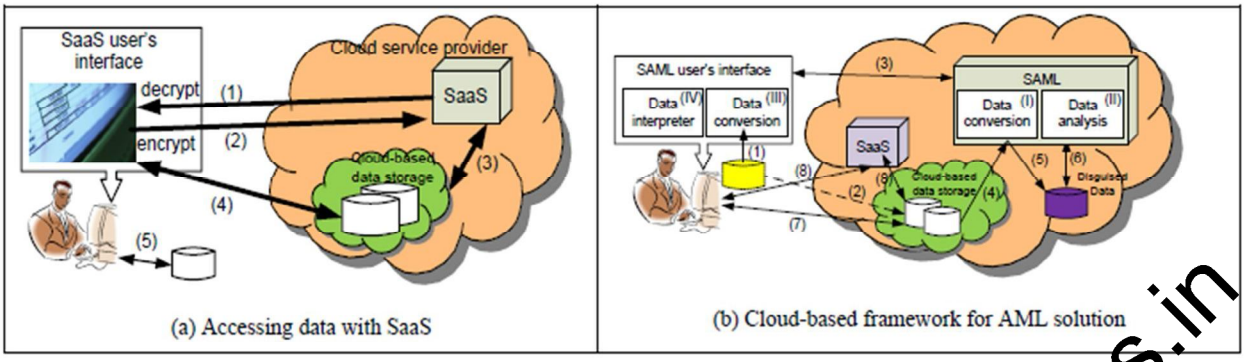

Figure 1. SaaS for analysing confidential data

In the second scenario, confidential data can be stored in the cloud by using cloud base 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 other customers come along and allocate space on the disk for storage, they can examine the allocated space and may have access to previous deleted confidential data. We can have moreover different kinds of attacking such as: Distributed Denial-of-Service (DDoS), Packet Sniffing, Man-in-the-Middle, etc., [20]. 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.1a). 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.

B. 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.1b (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.1b (II)) and data conversion (Fig.1b (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.1b (1)) and store it

in the cloud (Fig.1b (8)). This disguised data application is developed as a plug-in (Fig.1b (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.1b (I)).

In both scenarios, to distort data, our data conversion component uses methods discussed in the next section. The data analysis component (Fig.1b (II)) consists of statistical and data mining methods to analyse the disguised data to detect suspicious cases of ML.

C. Approaches for Disguising Data

In this sub-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. Generally, AML expert normally considers the following dimensions: frequency of subscriptions, frequency of redemption, subscription value, redemption value, current balance. All these features are conditional on time: daily, weekly, monthly etc. However, analysing directly of these dimensions would raise a concern of privacy. Based on the experience from AML experts, we created new dimensions based on current significant ones above. In fact, 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. 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. If the value of $\Delta 3$ is close to 1, then the frequency of subscription is significantly high comparing to its average value. 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.

V. Experiments

Our cloud-based framework has been 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. 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 2, 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. 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.

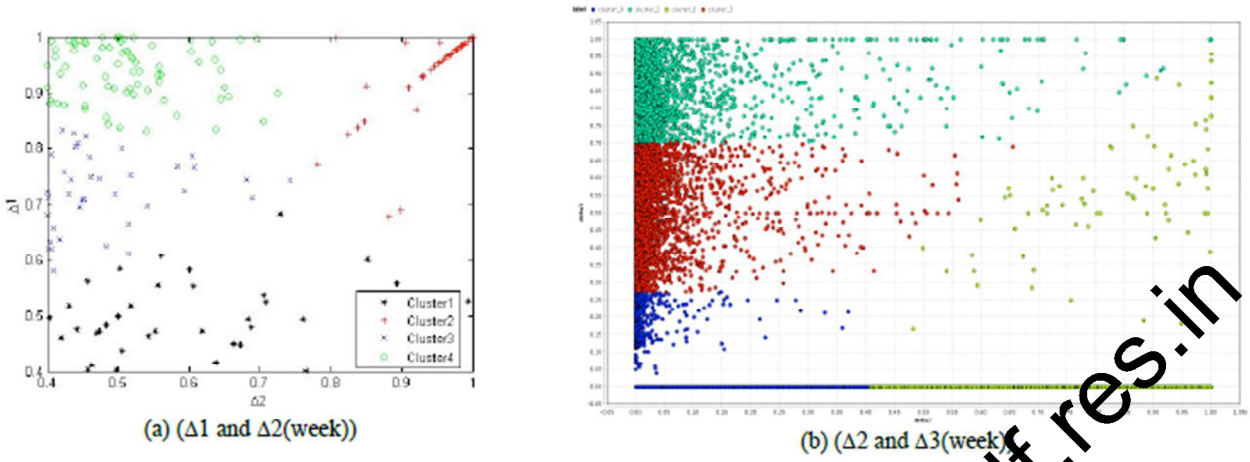

Figure 2. Suspicious Factor of SK fund

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]. Other parameters Δ3, Δ4, Δ5 and Δ6 can be used as additional parameters that can first of all support the AML expert in detection 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 Δ1, to Δ6 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.1b (IV)). As a conclusion, the creating of new dimensions is efficient 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.

VI. Conclusion

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. This framework allows us to distort data by preserving the distance between elements that is the important impact to clustering analysis. 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 was developed as a SaaS. More experiments are being carried out 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

1. Genzman L, Responding to organized crime: Laws and law enforcement.In H.Abadinsky (Ed.), CA: Wadsworth, 1997; 342.
2. Han J., Kamber M., *Data Mining: Concept and Techniques*. Morgan Kaufmann publishers, 2nd Eds., Nov. 2005.

3. Zang Z., Salermo J.J., Yu P. S., *Applying Data mining in Investigating Money Laundering Crimes*, SIGKDD'03, Washington DC, USA, August 2003: 747-752.
4. Le-Khac N-A., Markos S., O'Neill M., Brabazon A., Kechadi M-T., *An Efficient Search Tool for an Anti-Money Laundering Application of an Multi-National Bank's Dataset*, IKE '2009, LA, USA, July 13-16, 2009.
5. Jain R., Kasturi R., Schunck B.G., *Machine Vision*, Prentice Hall, 1995.
6. Scholkopf B., *A short tutorial on kernels*. Microsoft Research, Tech Rep: MSR-TR-200-6t, 2000
7. Kingdon J., *AI Fights Money Laundering*, IEEE Transactions on Intelligent Systems, 2004, : 87-89.
8. Scholkopf B. Plattz J., *Estimating the support of a high dimensional distribution*, Neural Computing, Vol. 13, No. 7, 2001.
9. Wilson D.R, Martinez T.R., *Improved Heterogeneous distance functions*. Journal of Artificial Intelligence Research
10. Tang J., *A Framework on Developing an Intelligent Discriminating System of Anti Money Laundering*, International Conference on Financial and Banking, Czech Rep., 2005
11. Vidyashankar G.S et al., *Mining your way to combat money laundering*. DM Review Special Report, Oct 2007
12. Vaidya J. ,Clifton C., Zhu M., *Privacy and Data Mining*. Privacy preserving data ming, Springer Verlag, 2006:. 1-6
13. Agrawal D., Aggarwal C.C., *On the Design and Quantification of Privacy Preserving Data Mining Algorithms*, ACM PODS Conference, Wisconsin, USA, 2002.
14. Pinkas B., *Cryptographic Techniques for Privacy-Preserving Data Mining*, ACM SIGKDD Explorations, 4(2), 2002
15. Oliveira S. R. M., Zaïane O. R., *Achieving Privacy Preservation when Sharing Data for Clustering*. Secure Data Management, VLDB 2004 Workshop, SDM 2004, Toronto, Canada, August 2004.
16. Croft H.T. , Falconer K. J., Guy R. K., *Unsolved problems in Geometry*. Vol.2, Springer Verlag, 1991
17. BigML, https://bigml.com
18. Alteryx, http://www.alteryx.com
19. Asterdata, http://www.asterdata.com
20. Jamsa K., Cloud Computing, John & Barlett Learning, 2013, pp.131-134