International Journal of Engineering Research and Technology. ISSN 0974-3154, Volume 13, Number 6 (2020), pp. 1173-1178
© International Research Publication House. https://dx.doi.org/10.37624/IJERT/13.6.2020.1173-1178

Database Intrusion Detection using Role and User Level Sequential Pattern Mining and Fuzzy Clustering

Indu Singh[1], Siddharth Singhal[2], and Vaibhav Kumar[3*]

[1,2,3]Department of Computer Science and Engineering, Delhi Technological University, Delhi, India.

Abstract

Majority of the present-day organizations store and use data on large scale for their functioning. Often this data is private and essential, and unauthorized access to the data can have serious financial and legal repercussions. While it is easier to capture malicious activities perpetrated by an external agent, it is relatively harder to detect transactions committed by company employee with some malicious intent. This is due to the awareness of the employee about the company database structure along with their authorized access privileges. We propose a novel intrusion detection system called user and role-level cluster-based intrusion detection system (URCIDS). We analyze the user behavior at role and user level, to check if the transaction under consideration is in accordance to the regular behavior of the user. We flag transactions that violate the general access pattern followed by the user. A detailed experimental analysis shows that we are able to gain a higher accuracy than most of the state-of-the-art methodologies proposed in this field of study.

Keywords: Database intrusion detection, user behavior, role behavior, data mining, role-based access control, risk assessment.

I. INTRODUCTION

Lately, large-scale extension in the use of network-based connectivity has presented the databases to countless dangers, and in this manner, it is of most extreme significance to shield the databases against attacks. Notwithstanding the risk of outer attacks, the important data stored in databases is at risk of inner attacks, resulting as a consequence of inside clients manhandling benefits of their access provided by the organisation. Intrusion Detection Systems (IDS)are utilized to recognize such security breaks in a database deciphering irregularities in client behavior to distinguish unapproved access attempts to get to the important data stored in databases made by the employees of the association.

IDS can be classified as non-signature-based intrusion detection and signature-based. The signature-based IDS distinguish attacks by utilizing contrasting the information and known examples of attacks [2], it in this manner neglects to identify previously unknown obscure dangers. Additionally, it becomes wasteful to record the data about marks of a wide range of attacks. Such frameworks require normal updates, and in this way are cannot capture the extraordinary attacks on different the system.

Then again, the non-signature-based IDS [3] contrast the information exchange and typical client-side interaction with databases, and assess the comparing deviation to characterize the contribution as an ordinary or pernicious movement. Such frameworks end up being powerful in handling with insider dangers. Through this paper, we expect to distinguish and forestall insider attacks in databases. The NIST rules [1] institutionalized RBAC for RDBMS by doling out every client a predetermined job, and in this way including a level of security for the database. RBAC restrains the client to a space of characteristics consequently limiting the client to get to the activities out of its degree. Countless IDSs are intended to handle outer attacks, however right now, we present a novel methodology user and role-level cluster-based intrusion detection system (URCIDS).

In our methodology, we process each incoming transaction on two levels, this has the added advantage of creating multiple levels of security. The two levels are used in parallel to obtain a combined score. In one of the levels, we analyze the incoming transaction after preprocessing it on a user level. The read and write sequences resulting from the incoming query are analyzed for the user’s history. The second level, analyses the fuzzy-cluster membership using role queries. The scores from these two levels are combines and a single score is created which embodies the risk level associated with this transaction.

The remainder of the research article is organized as follows. Section II depicts the related work. In Section III, the proposed approach and its novelties. Section IV presents the results. Section V represents the conclusion of the paper.

II. RELATED WORK

Various analysts have worked in the field of Network Intrusion Detection Systems, yet just the research in the areas of Database Intrusion detection system is less. A few frameworks for Intrusion Detection in working frameworks and systems have been developed, anyway they are not satisfactory in shielding the database from the insider attacks that pose a major threat. ID framework in databases work at transaction level, query level and client (job) level. Bertino et. al in [1] depicted the difficulties to guarantee information privacy, availability and integrity and the requirement of database security wherein the need of database IDS is to handle the threats that arise from inside dangers are available.

Database IDSs incorporate time-based (temporal) analysis of questions and Data conditions among characteristics, inquiries and exchange. Lee et al. [6] proposed a intrusion detection technique based upon temporal analysis, which joined time marks and recorded an update hole of fleeting characteristics. Any abnormality in update example of the quality was accounted for as an interruption in the method proposed. The seminal work done by Aggarwal et al. [2] in association rule mining helped in discovering information conditions among information traits, which was consolidated in the field of database security, particularly intrusion detection in databases.

Due to the underlying improvement of data dependency rule mining, DEMIDS [8], an abuse discovery framework for social database frameworks was put forward by Chung et al. Profiles created from the access patterns of users were synthesized from the review log and metrics for finding the distance were additionally utilized for identifying data items. These were utilized together so as to increase the spectrum of users. In any case, when the quantity of clients for a solitary framework gets significant, keeping up profiles turns into an excess method. Another problem with the system is that it assumes the framework for an given information schema.

Hu et al. [5] introduced a data mining based framework for identifying malicious queries, which utilized the static investigation of database review log to find conditions among qualities at exchange level and spoke to those conditions as sets of perusing and composing procedure on every datum thing. In another methodology put forward by Hu et al. [10], procedures of consecutive sample mining are known to be successfully applied on the transaction log, so as to recognize visit arrangements at the exchange level. This methodology helped in recognizing a spectrum of malignant transactions, which independently followed the client conduct.

The technique proposed by [13] broadens the methodology in [5] by doling out loads to all the procedure on information characteristics. The transactions act out of alignment with the manner with the information conditions were set apart as intrusive. The significant inconvenience of client allotted weights is that they are constant and irrelevant to other information characteristics. Kamra et. al [14] utilized a grouping strategy on a RBAC framework to frame profiles dependent on property get to which spoke to typical client conduct. An alert is raised when irregular conduct of that job profile is observed.

Y. Yu et. al.[20] presented a Intrusion Detection System based upon the application of fuzzy logic. A classifier is utilized an incoming event as anomalous or normal. The premise of classifier is shaped by the autonomous recurrence of every framework call from a procedure in ordinary conditions. The proportion of the likelihood of an arrangement from a procedure and the likelihood not from the procedure fills in as the contribution of a fluffy framework for the characterization.

A mixture method was portrayed by Doroudian et. al [16] to recognize intrusion at two levels: inter-transaction level and inter-transaction level. At the level of transaction, a lot of predefined expected transactions were indicated to the framework and a data mining calculation was put-forward at the entomb exchange level to discover conditions between the recognized transaction. The downside of such a framework is, that groupings which a lower frequency than the limit threshold are not considered. Along these lines, the rare sequences were totally disregarded by the framework, regardless of their significance. Therefore, the True Positive Rate tumbles down for the framework.

The above downside was augmented upon by Sohrabi et. al [13] who put-forward a new methodology named ODARDM, in which rules were detailed for lower recurrence thing sets, too. These principles were removed utilizing influence as the standard worth measure, which limited the fascinating information conditions. Therefore, True Positive Rate expanded and the False Positive Rate diminished. In late advancements, Ranao et. al[18] introduced a Query Access location approach utilizing Random Forest and PCA to lessen information dimensionality and create just important and unrelated information. As the dimensionality is decreased, both, the framework execution and True Positive rate increments

III. PROPOSED METHODOLOGY

New age associations manage gigantic amounts of information whose security is of prime importance. The information in databases includes characteristics depicting real life objects as entities in the database. The attributes have varying levels of sensitivity, for example not all traits are similarly critical with respect to the entities that are stored in a database. For instance, consider a database that stores information of individuals such as: biometrics, credit card and others. Therefore, in this manner, unapproved access to such critical information is of critical importance. This kind of information may only be accessed by certain employees and should be protected from others. Further, at times, a user may even have access to these attributes and decides not to access them in his day-to-day job, a sudden access of to these attributes could then be caused a malicious reason and therefore needs to be captured. In this paper we deal with such transactions and make sure that appropriate steps are taken by the system.

We propose an Intrusion Detection system, URCIDS that utilizes rules created in the learning stage and afterward distinguishes vindictive inquiries in the detection stage. Our methodology is novel in contrast with recently proposed frameworks, as the proposed calculation URCIDS is database adaptable. It progressively designates affectability to the tasks, in this manner rendering the learning stage confident and independent. Further, for each incoming transaction a score is calculated using role and user behavior-based attributes and a score based upon them is assigned to the transaction, leading to the calculation of risk associated with the incoming transaction.

A. *Terminology*

Transaction: A transaction maybe defined as a collection of queries issue by a user. We refer to each transaction by a unique ID. eg: A transaction consisting of n queries, TID = Q1, Q2, Q3,...Qn where Qk is the kth query of the transaction.

User Role: Each transaction is committed by an individual user. In our IDS system, each user is given permissions under a role, and this role is logged when the user commits a transaction. Therefore, at an individual level every user is assigned a role from a set of roles R = {r1, r2,...rm}.

Operation: We define an operation as an action which can lead to reading or writing an attribute from a relation. we define this symbolically by O(x) where, O(x) ∈ {w(write),r(read)}.

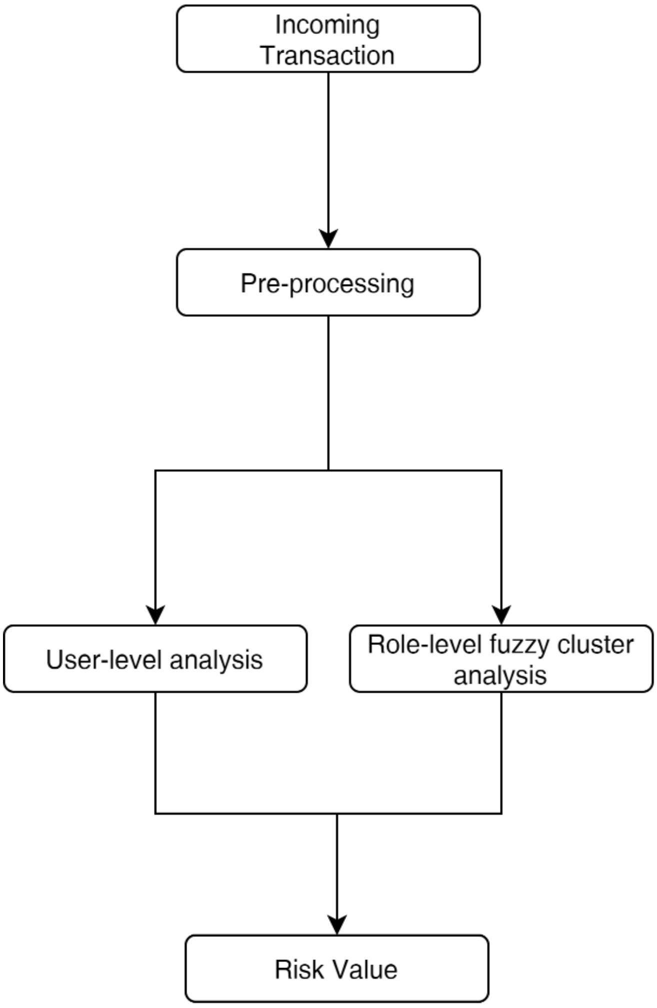

Fig.1: Architecture of detection phase.

Read Sequences: Read sequences are defined as {R(x1), R(x2), R(x3),...O(x)}, where, O(x) ∈ {w(write),r(read)}. It means that x1, x2, x3 etc. are read before O(x).

Write Sequences: Write sequences are defined as {O(x), W(x1), W(x2), R(x3),... } where, O(x) ∈ {w(write),r(read)}. It means that element x1, x2, x3 etc. are modified after the execution of the operation O(x).

Read Rules: Read rules are obtained from read sequences. They are defined as R(x1), R(x2), R(x3),... } → O(x), where x ∈ μ.

Write Rules: Write Rules are obtained from Write Sequences. They are defined as O(x) → {W(x1), W(x2), R(x3),... }, where x ∈ μ.

B. *Learning Phase*

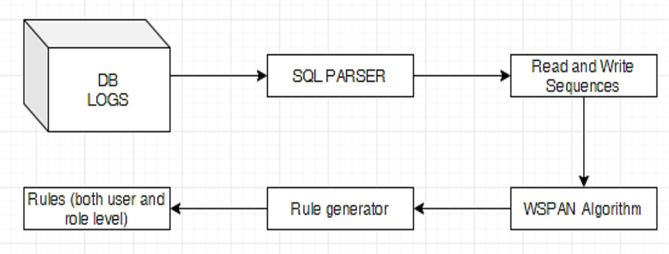

Fig.2 Learning phase

The steps consisting of learning phase are as shown in Figure 2. The SQL parser first parses the logs which are sourced from the DB history. The parser converts the queries from different roles and users into read and write sequences as defined earlier. These are then utilized for finding the most frequent patterns using the WSPAN algorithm, which creates frequent patterns based upon frequencies. Finally, these read and write sequences are then utilized for creating rules using the Rule generator. The rule generator then results in rules at both user and role level.

C. Detection Phase

Once the rules have been formulated in the previous step, they are now utilized in the detection phase. The incoming transactions are first pre-processed using the Parser which results in read and write sequences. The rules are then classified on two levels (as shown in figure 1.):

User level

Each incoming transaction is evaluated against the above generated rules for each user using WSPAN and ARM. For each attribute present in the transaction, we try to check if that transaction is in accordance with the Association rule for that attribute for the corresponding user. We use a modified version of Jaccard Index to calculate the degree of non- maliciousness of the user. We define it as the ratio of total rules followed by the transaction by the total rules applicable to the transaction. We denote this by USR(S_{it}) for an incoming transaction from user i with role t.

Role-level fuzzy cluster analysis.

Each of the read-write sequences from the pre-processing step are first classified on the basis of roles. The role set of read and write sequences is then defined as follows:

$$R_{si} = \{R(x) \text{ or } W(x) \text{ for element } x \text{ such that they belong to a user of role } i\}$$

Rule-based binary sequence (R_{bi}) for each role is defined as follows:

$$R_{bi} = \{x_i \text{ such that } x_i = 1 \text{ if } x_i \text{ belongs to } R_{si} \text{ else } x_i = 0\}$$

For instance, consider the following example.

$$R_{s1} = \{\{R(1),R(2),R(3)\},\{W(1),W(7),W(8)\}\},$$

Then,

$$R_{b1} = \{1,1,1,0,0,0,1,1\}$$

Further, a similar binary sequence is created for each incoming query. Let this sequence be defined as S_{it} . For incoming query from user i with role t.

The role-based score is then calculated as follows.

$$\mathbf{RB(S_{it})} = \frac{M_t(S_{it})}{\sum_{k=1}^{N} M_k(S_{it})}$$

$$\mathbf{M_t(S_{it})} = \frac{D(S_{it},R_{bt})}{\sum_{k=1}^{n} D(S_{it},R_{bk})}$$

Where, D is a suitable distance metric, in our study we have used the Manhattan distance. The scores from the aforementioned steps are the combined for quantifying risk, which is defined as follows:

Risk

We define the risk associated with each transaction as the reciprocal of the harmonic mean.

$$\text{Risk}(\mathbf{S_{it}}) = \frac{\text{USR}(S_{it}) + \mathbf{RB}(S_{it})}{2 * \text{USR}(S_{it}) * \mathbf{RB}(S_{it})}$$

IV. RESULT AND DISCUSSION

Dataset Description

We begin with a brief description of the dataset used for evaluating the performance of our proposed approach. One of the major roadblocks in this field of study is the unavailability of a standard dataset for training and testing of our proposed approach. The absence of a pubic dataset can be attributed to the confidentiality constraints imposed on various organizations. To overcome this problem, researchers generate their own synthetic data adhering to standard benchmark database schemas.

We conducted our experiments on banking database schema conforming to the TPC – C benchmark [12]. Our dataset consisted of two categories of transactions, i.e., malicious and non-malicious queries. In our RBAC paradigm, we assumed four different categories of roles and three different kinds of users in each role category. For each user we took two thousand malicious and five thousand non-malicious queries.

In order to generate the non-malicious queries, we gave each role access to a limited number of tables in the database schema. Thereafter, for each user in the role, we generated queries accessing the attributes in their corresponding table with a pre-specified seed value and order of access. For generating malicious queries we changed the frequency of access by modifying the seed value and also accessed attributes from tables outside their access privileges, thus deviating from the genuine pattern of access of a user.

Performance Discussion

We created a sixty forty split of our generated dataset for training and testing purposes. We evaluated the True Positive, False Positive, True Negative, and False Negative values for our testing dataset. Subsequently we calculated the precision, recall, and F1-score of our proposed approach.

We tested the role level model followed by a combination of the role level and user level model. For evaluating the performance of both the approaches we varied the total number of transaction under consideration. We considered seventy, eighty, ninety, and hundred percent of the total transactions for our purpose. The general trend observed was increase in the F1 score values for both the approaches with increase in the number of transactions. This can be attributed to the fact that with increase in available data, the model s able to derive more useful information in the form of association rules which can improve the accuracy of classification of the malicious and non-malicious queries. As depicted in figure 3, the F1 score for role based model increased from 0.64 to 0.73. Figure 4 provides an overview of the F1 score values for a combination of role and user level based model, which increased from 0.88 to 0.94 with increase in the number of transactions. The higher F1 score for our proposed model clearly depicts the superiority of our approach over the conventional role-based approach.

We also evaluated the performance of our proposed approach for different values of support threshold for our pattern mining algorithm. In figure 5, we can observe that the accuracy of our approach increased monotonically till 0.7 threshold, thereafter, it began to decrease. We can attribute this trend to the fact that for a very small threshold value the algorithm identifies unimportant patterns that do not occur frequently and judges the incoming query according to these patterns increasing the amount of False Positives, while at very high threshold values important patterns get ignored leading to steep increase in the False Negative values.

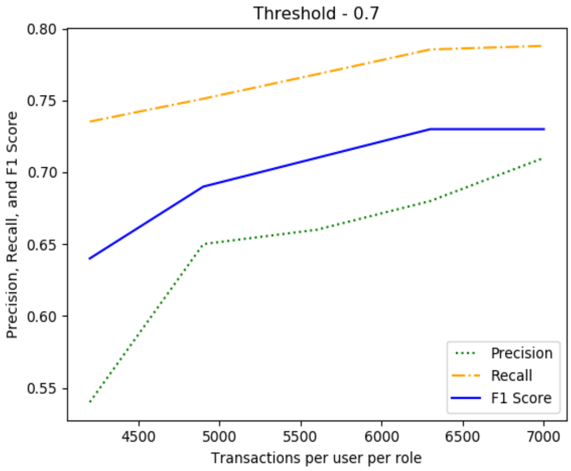

Figure 3: Precision, Recall and F1 Score for Role Based IDS

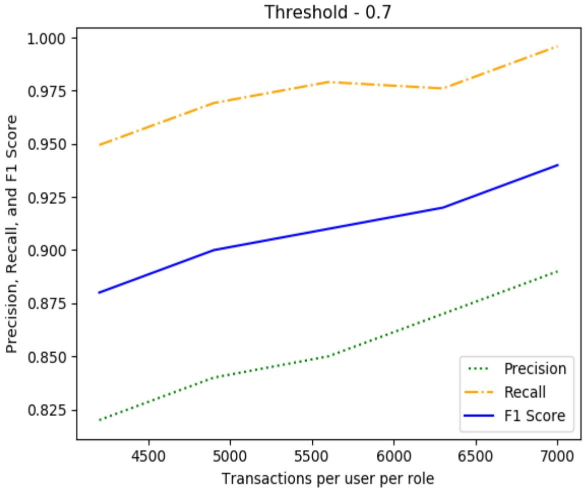

Figure 4: Precision, Recall and F1 Score for URCIDS

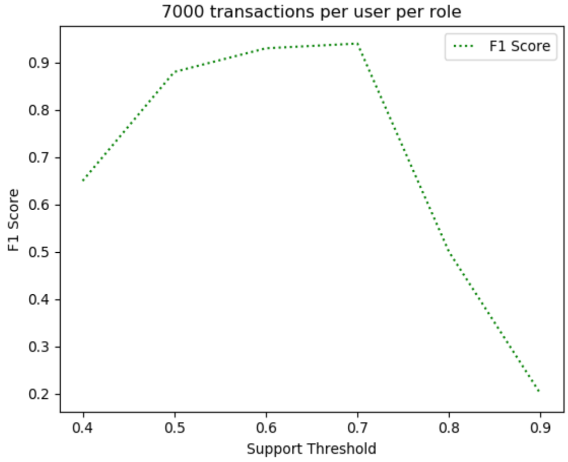

Figure 5: F1 Score for different support threshold values

Table 2 provides a comparative study between our proposed approach and different detection methodologies proposed in the field of study. As depicted in the table, we observe that our intrusion detection system outperforms the existing approaches. This can be accounted to the fact that we take into consideration the differences and irregularities in access patterns of the different users of the same role.

Table 2: Comparison of our approach with related works

Approach	**F1 Score**
Hashemi et al. [19]	0.82
Majumdar et al. [15]	0.78
Elisa Bertino et al. [13]	0.92
URCIDS	0.94

V. CONCLUSION AND FUTURE WORK

In this paper, we propose a novel approach URCIDS for detecting malicious transactions performed by an insider. In our approach we account for the irregularities and differences in access pattern of different users belonging to the same role, thus obtaining superior accuracy when compared with existing Intrusion Detection Systems performing analysis at only Role level.

In future we aim to improve the granularity of existing user attributes and add new attributes for performing analysis at user level. We also aim to integrate methods for mitigating the damage caused by malicious transactions that were not detected by the system.

REFERENCES

[1] D. Ferraiolo, R. Sandhu, S. Gavrilla, D. Kuhn, R. Chandramouli, ”NIST standard for role-based access control”, ACM Transactions on Information and System Security (TISSEC), Vol 4, Issue 3, 2001.

[2] R. Agrawal, T. lmieliiski, and A. Swami, ”Mining Association Rules between Sets of Items in Large Databases”, in Proceedings of the 1993 ACM SIGMOD International Conference on Management of data, 1993.

[3] H S Vaccaro and G. E. Liepins, ”Detection of Anomalous Computer Session Activity”, In Proceedings of the 1989 IEEE Symposium on Security and Privacy, pages 280-289, Oakland, California, 1-3 May 1989.

[4] Bertino and R. Sandhu, ”Database Security Concepts, Approaches, and Challenges”, IEEE Transactions on Dependable and Secure Computing, Vol. 2, No. 1, pp. 2-19, 2005.

[5] Y. Hu and B. Panda, ”Identification of Malicious

Transactions in Database Systems", in Proceedings of the International Database Engineering and Applications Symposium (IDEAS '03), 2003.

[6] V. C. S. Lee, J. A. Stankovic, and S.H. Son, "Intrusion Detection in Real-time Database Systems Via Time Signatures", in Proceedings of the 6th IEEE Real Time Technology and Application Symposium (RTAS),pp. 124-133, 2000.

[7] D.E. Denning, "An Intrusion-Detection model", IEEE Transactions on Software Engineering, Vol. SE-13, pp. 222-232, Feb. 1987.

[8] C. Y. Chung, M. Gertz, and K. Levi , "DEMIDS: A Misuse Detection System for Database Systems", in Third Annual Working Conference on Integrity and Internal Control in Information Systems, pp.159-178, 1999.

[9] G. Lan, T. Hong, H. Lee, "An efficient approach for finding weighted sequential patterns from sequence databases", Applied Intelligence, vol. 41, no. 2,pp. 439-452, 2014.

[10] Y. Hu and B. Panda, "A Data Mining Approach for Database Intrusion Detection", in Proceedings of the ACM Symposium on Applied Computing, pp. 711-716, 2004.

[11] Y. Hu and B. Panda, "Mining inter-transaction data dependencies for database intrusion detection", Innovations and Advances in Computer Science and Engineering, Springer, 2010.

[12] TPC-C benchmark: http://www.tpc.org/tpcc/default.asp

[13] Sun, Y., Xu, H., Bertino, E., Sun, C.: 'A data-driven evaluation for insider threats', Data Science and Engineering, 2016, 1, (2), pp. 73–85.

[14] A. Kamra, E. Bertino, and E. Terzi, "Detecting anomalous access patterns in relational databases", The International Journal on Very Large Data Bases, Springer, 2008, ISSN:1066-8888.

[15] A. Srivastava, S. Sural, and A. K. Majumdar, "Weighted intra transactional rule mining for database intrusion detection", in Proceedings of the Pacific-Asia Knowledge Discovery and Data (PAKDD), 2006.

[16] M. Doroudian and H.R. Shahriari, "A Hybrid Approach for Database Intrusion Detection at Transaction and Inter-Transaction Levels", 6th Conference on Information and Knowledge Technology (IKT), pp. 1-6, 2014.

[17] Y. Yu and H. Wu, "Anomaly Intrusion Detection Based upon Data Mining Techniques and Fuzzy Logic." IEEE Conference on Systems, Man and Cybernetics, 2012.

[18] C. A. Ranao and S. Chao, "Anomalous query access detection in RBAC-administered databases with random forest and PCA", Journal Information Sciences, Volume 369, Issue C, Pages 238-250, 2016.

[19] Hashemi, S., Yang, Y., Zabihzadeh, D., Kangavari, M.: 'Detecting intrusion transactions in databases using data item dependencies and anomalyanalysis', Expert Systems, 2008, 25, (5), pp. 460–473.

[20] Mina Sohrabi, M. M. Javidi, S. Hashemi, "Detecting intrusion transactions in database systems: a novel approach", Journal of Intelligent Info Systems 42:619-644 DOI 10.1007 Springer 2014.