

# An Efficient Data Enrichment Scheme for Fraud Detection Using Social Network Analysis

Soheil Jamshidi, Mahmoud Reza Hashemi

Internet Fraud Risk Assessment and Ubiquitous Detection Laboratory (iFRAUD)  
School of Electrical and Computer Engineering  
College of Engineering, University of Tehran, Tehran, Iran  
{s.jamshidi, rhashemi}@ut.ac.ir

**Abstract**—With the continuous fast paste of ecommerce growth and the ever increasing amount of electronic transactions, we witness a similar if not more significant increase in electronic fraud. Consequently, fraud detection systems need to update their methods and improve them by considering more sources of information to stay effective. In this paper, a data enrichment scheme is proposed which concentrates on using social network analysis to help the detection system by feeding information that is hidden in the relations among entities. Since one of the challenges of a real life electronic transaction system is the large amount of data and number of users, the proposed scheme is presenting an efficient method to update the social network, as well. Simulation results indicate that the proposed scheme is able to detect fraud scenarios that are not detected using typical anomaly detection methods based on the normal behavior of cardholders. Hence, providing a higher accuracy, while minimizing the updating procedure.

**Keywords**- Social Network analysis; Fraud detection; Data Enrichment; Data mining; update phase

## I. INTRODUCTION

Many industries including banking and financial sectors, insurance companies, government agencies, law enforcement and telecommunication companies are losing hundreds of millions of dollars to fraud every year. In 2011, just in North America, approximately \$3.4 billion was lost in ecommerce due to online fraud [1]. This loss has increased by about \$700 million since 2010 [1]. Having witnessed a severe increase in fraud attempts in recent years, fraud detection is more than ever an essential component to make these domains safe and save them from revenue losses. Despite significant efforts, since most existing techniques base their decision on just a small fraction of data to detect fraud, finding these misbehaviors can be very complicated. Moreover, as fraud detection methods improve, fraudsters become more sophisticated, as well. They commit fraud and reach their intentions behind a network of relations with trusted users and actors in order to deceive the security checks and fraud detection systems. For example, in the insurance domain, there are many instances where organized groups of fraudsters consisting of drivers, police officers, lawyers, garage mechanics and insurance workers form a fraud network. Unfortunately in many cases they are able to hide themselves from being detected. With all the different methods of committing fraud, fraud detection seems

not to be solved yet. Some of them need additional data sources to be able to detect these kinds of activities. One of the common features of committing fraud is the fact that fraudsters are humans, hence they should be considered as social entities with relations among them that may contain some additional insight about them. They can also be socially influenced [2]. This information can significantly help any fraud detection system, and ignoring them can strongly affect the time and accuracy of their decisions. This is where Social Network Analysis (SNA) can help. SNA has been used in detecting fraud in financial domains and its related industries such as insurance and online auctions to improve their trust and reputation systems [2][5][6][7][10][12].

Some researchers have used SNA in addition to conventional methods of data mining such as classification and clustering in order to improve their accuracy. Almeida [3] explains the use of social networks to help improve the fraud detection classifiers. He shows that in some cases, using extracted patterns from the social network of organizations and people can lead to a better classifier, and indeed to better results in fraud detection. Botelho and Antunes [2] tried to combine social network analysis with semi-supervised clustering methods. Their results indicate that semi-supervised clustering performs better when data is enriched with social network analysis.

In these models they follow the steps depicted in Fig. 1 to recognize the appropriate patterns for nodes with unknown labels and use them in data mining methods as new source of information.

First the social network is built from existing data and then this network is analyzed by using different methods, in order to find some knowledge patterns, and probability of being a fraudster; the fraud score is calculated based on these patterns and data is enriched with them. The result will be used as a new attribute in order to improve the accuracy of further investigations.

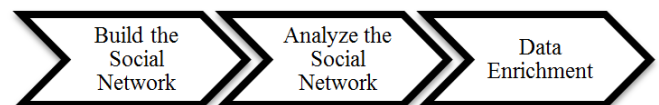


Figure 1: Steps in current models; start-up phase in our new scheme

This research was partially supported by a grant from the Research Institute for ICT (ITRC).

Although social networks are useful in many cases, but usually there is not enough information about relations that exist in the network of nodes. Often these relations are sparse, and there are many isolated nodes and just a small fraction of nodes are involved in this kind of relations. Therefore, fraud detection systems should not be dependent merely upon network analysis. Furthermore, as the network grows and new nodes and more transactions are added, the relations among them should be updated and form a new network. This can be a very time consuming and computationally intensive process especially for large networks. Most approaches that use SNA are not considering an update phase to keep their models up-to-date and are not able to cope with the huge amount of new data that needs to be processed continuously.

In this paper, a data enrichment scheme is proposed which concentrates on using social network analysis to help the detection system by feeding information that is hidden in the relations among entities. In the proposed method, the introduced steps of Fig. 1 are performed as the startup phase and a new phase for updating has been added. Since one of the challenges of a real life electronic transaction system is the large amount of data and the number of users, the proposed scheme is presenting an efficient method to update the social network, as well. Since the proposed updating scheme is not dependent on the enrichment method it can be used efficiently with any other SNA based scheme.

The rest of the paper is organized as follows. In section II, some of the applications of social network analysis for fraud detection in financial, insurance, online auction, reputation systems and related domains have been presented along with their weaknesses. Then in section III the new proposed model is introduced. Evaluation results are presented in Section IV, followed by the concluding remarks in Section V.

## II. RELATED WORKS

In the last decade, social networks were recognized to play a key role in researches on the behavior of humans and other social actors. Computer scientists have also taken advantage of social behavior of actors in their research area and developed fundamental researches on the area of social network analysis. A social network can be defined as a directed graph, where nodes are representing society agents and the edges correspond to the relationships among nodes [2].

On the other hand financial crimes are known to be one of the major problems of governments and are usually combined with network crimes. Money laundering and fraud are the most important kinds of financial crimes. The majority of fraud types are too hard to be executed individually, which forces fraudsters to create some relationships with fraudulent associations with the goal of committing fraud in a safer way. From the point of view of fraudsters, it will be safer to make connections to more trusted nodes and commit the fraud with the help of them, to deceive fraud detection systems [3]. Gomes [4] emphasizes that the social structure of those associations is nothing more than social networks. The huge amount of data and complexity of them is the main challenge in this domain. This challenge is confronted by integrating the visualization techniques and using social network analysis on financial activity networks.

In addition to approaches which use social network analysis metrics such as centrality, or using k-core to categorize the relations and network of activities, some researchers use SNA to enrich the data with additional sources and help to improve the accuracy of clustering and classification methods performed on data. Botelho and Antunes [2] tried to combine social network analysis with semi-supervised clustering methods. They use Badrank to distribute the information of known entities in a network and estimate the probability of being a fraudster for unknown entities. This probability will act as an additional attribute to be used in clustering. Almeida [3] used the DFS algorithm to determine the local social network of each node and then used the Gspan algorithm to find the common patterns. The involvement of nodes in these patterns will help estimating the actual node labels. For each new transaction or for each new node that enters the system, the above models need to be updated. But calculating the Badrank or performing DFS and Gspan on the entire network significantly increases the computational complexity and required time of the update phase. As a result these systems cannot meet the real time decision making requirements.

As fraud appears in many domains other than ecommerce such as online auctions, and insurance we will briefly review the methods and researches that leverage social networks to improve performance in such domains in the remaining of this section.

### A. Online Auctions

In Online Auctions, Trust and Reputation Systems (TRS) are in charge of keeping track of fraudulent activities and should make the reputation of fraudsters less than trusted users in order to provide a reliable environment for customers. Wang and Chiu [5] believe that the almost zero cost of creating multiple identities in online auction markets, can be the source of problematic transactions and online auction frauds. They conclude that SNA measurements such as K-core can be used as an effective indicator to differentiate the anomalies in the network structure caused by problematic transactions from normal structures. TrustDavis [6] is a reputation system that uses the social network of entities based on their references to each other. It is able to limit the damage caused by malicious collusions of dishonest participants for honest participants and also limits the changing or issuing of multiple identities for dishonest participants. Reference [7] focusses on the existence of credit speculation agents, who boost the score of fraudsters. They proposed an approach to analyze online crediting behaviors using SNA. Based on thresholds for In-degree, Out-degree, strength and interval of transactions in the system, they determine the probability that a node is engaged in trust fraud. Proposing novel measures of trustworthiness, such as credibility and density that are calculated by mining the topology of the network of seller-buyer relationships, with the goal to retrieve useful knowledge, is another approach of making systems more secure in this domain [10]. Although there are many researches that propose the use of social networks to handle misbehaviors in this domain, but in many online auctions, there is not enough (social) data to feed the social network based systems, and in many of them 89% of participants have just one transaction in common [11].

Therefore each new transaction can affect just a small fraction of nodes. As a result a procedure is needed to prevent the system to involve the entire network structure and nodes in the update phase.

### B. Insurance

In addition to frauds that are committed individually in this domain, there are many instances where organized groups of fraudsters consisting of drivers, police officers, lawyers, garage mechanics and insurance workers form a fraud network. In [12] an expert system is introduced that focuses on relations between participants of automobile accidents in order to fight against automobile insurance frauds. After constructing the network, it starts with detecting the suspicious components. It then determines the suspicious entities in each of those components by visualizing the social network of these entities. Clearly, using the hidden knowledge of relations as an additional source of data helps to make a more accurate model but there is no plan to update and rebuild the initial model in each round.

To summarize, although recent fraud detection methods are considering SNA as a tool to combat sophisticated fraudster relations, but none of them have addressed the necessity to update the system for new changes. While using SNA is useful for improving the accuracy, but in practice the huge amount of new incoming data and the necessity to update the created model based on them can prevent the system from being up-to-date and acting agile continuously. In this paper, we will use SNA to enrich the data with the hidden knowledge of relations and meanwhile make the update phase practical to keep the system up to date with less effort.

## III. PROPOSED APPROACH

As mentioned in the previous section, SNA is used in many researches in fraud detection domains, but most of them just use it in order to improve the accuracy of their models. As a result, not enough attention has been paid to the time cost and the need for real time decisions in some applications such as electronic payments or in reputation systems. Moreover, when a new node enters the network or a new transaction is made, an updating phase is needed in order to keep the system in a logically true state. In response to these needs, we introduce a new model in order to make updating more efficient.

The proposed model consists of two phases; startup and update. In the startup phase, the social network of involved elements in the organization is built. Then, the data model which is a graph in this case is created. Afterwards, this social network is analyzed to find useful patterns and relations. Similar to [2][3], the probability of being a fraudster, referred to as bad-score, is calculated based on the relations between a node and known fraudsters. Calculated probabilities will enrich the data in the form of new attributes and can improve detection rate. Fig. 1 illustrates these sub steps.

In the proposed approach and in order to make accessing, using and updating of these patterns more efficient, some extra information is stored in a repository referred to as Network Patterns Repository (NPR). This information consists of a score which is the probability of being a compromised node in the form of  $(A, n)$ , where A is the node and n is the bad-score.

The bad-score for normal nodes is set to zero. We will explain below how the bad-score is being calculated for a node that has been compromised with a fraudulent transaction. Also each relation is stored in the form of  $(A, B, level)$  where A and B are the source and target in this relation, and level is the number of hops (edges) between these two nodes. Finally, in the proposed scheme and in order to further reduce the complexity of the updating phase, by minimizing the search time, we store the set of all the nodes that are related to each node. After filling out the NPR, and before any update, the system has the current state of each node and its probability of being a fraudster or having fraudulent transactions. This information can be used as an extra attribute in order to improve the accuracy of data mining approaches which are adopted to detect fraud by determining the actual label of each node.

Afterwards, instead of repeating the first phase, it is more efficient to have a procedure to just apply the changes whenever needed, without rebuilding and reanalyzing the entire social network and data. This approach makes updating the model more efficient. Thus in this phase we just add the new relations to the NPR and update nodes whose suspicious degree is affected by this update. As illustrated in Fig. 2, this adds an update step to the conventional model. The above steps are explained in more details in the following sub-sections.

### A. Startup phase

Based on our model we first need to build the social network of nodes. Hereafter, we will explain the model based on the graph representation of a social network as in Fig. 3, where A is known to be a node with fraudulent transactions.

For instance the network of Fig. 3 can be represented as the set  $E = \{(A,B) (A,C) (B,D) (D,E) (C,B)\}$ . This set includes all the nodes that are directly related. But many nodes are related to A in the above example through one or more intermediate nodes. Considering all these indirect relations is time consuming and increases the SRP storage requirements. Hence, a threshold  $\theta$  is defined in order to control the number of levels that should be considered to be related to a node indirectly. The bigger the threshold, the more effort is needed to search and update NPR. For instance, for  $\theta = 2$ , A has no effect on the score of E in the above example. Hence in this case, NPR will consist of:  $\{(A,B,1), (A,C,1), (B,C,1), (B,D,1), (D,E,1), (A,D,2), (B,E,2), (C,D,2)\}$ , as mentioned before, the numbers represent the hops (edges) between nodes of each pair.



Figure 2: The proposed model; considering an update phase

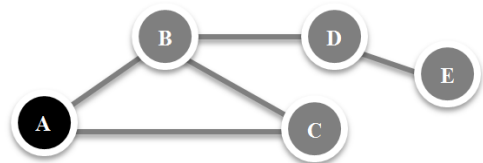


Figure 3: Relations of a subset of network in  $t_0$

The closer two nodes are, the more likely they have the same label. In other words, if two nodes are directly related a fraudulent activity in one of them should significantly influence the bad-score of the other. As a result, in this paper bad-score is calculated using (1):

$$bad - score(i) = \sum_{j=1}^M (\partial + 1) - h(i, j). \quad (1)$$

Where  $\partial$  is the threshold of being influenced and  $M$  is the set of nodes which are accessible from  $i$  in less than  $\partial$  hops and have had labeled as being compromised (having a fraudulent transaction or being a fraudster). The  $h$  function returns the number of hops between  $i$  and  $j$ . A higher  $h$  value decreases the bad-score value as expected.

With any changes in  $\partial$  or in the structure of the social network, for instance when a new node or a new edge is introduced, bad-score should be updated. The proposed efficient updating procedure will be explained in the next subsection.

The social network is represented as pairs, and more importantly, bad-score values are calculated for each node. Hence, the system can use the bad-score as an additional attribute to improve the accuracy of decision making and is ready for being used for fraud detection.

Bad-scores could be assigned to both nodes and transactions. If the mining method is used to detect fraudulent nodes, the calculated bad-score should be assigned to nodes. Otherwise the focus is on the transactions. In this case, for each node the bad-score is assigned to the transactions that have been committed at the same time as at least one fraudulent transaction of its neighboring nodes. This score is set to 0 for the remaining transactions of each node.

The enriched dataset can be mined using any clustering or classification method. We used the hybrid system proposed in [14] and added bad-score as a new attribute to the test data. Bad-score values are normalized with the range of [0-1]. This new source of information helps improving the accuracy of fraud detection by detecting some of the false negatives.

### B. Update Phase

After building up the system, filling out the NPR and calculating the node scores, new nodes are introduced to the system or new transactions are carried out with time. Hence, one needs to update the NPR. Most existing SNA methods either ignore the update phase or have to build the entire social network with any minor change, each time starting from scratch.

In our proposed model, each time a change occurs, only the new relations which affect bad-scores should be specified and stored. To determine the affected bad-scores, we only consider the directly or indirectly related nodes up to  $\partial$  hops. Once again, the bigger the threshold the more effort is needed for updating and consequently more accurate results will be achieved. But the improvement will have a limitation;  $\partial$  can be increased up to the maximum network diameter. For the network of Fig. 3, assuming that in  $t_1$  a transaction between A and D has been committed, a new relation between A and D is formed. Fig. 4 illustrates this condition.

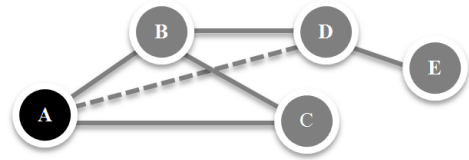


Figure 4: System in new state in  $t_1$

We should check the two ends of this new relation. If neither has been labeled to be compromised the new relation is added to NPR. In addition, the level values of all the nodes that are related to this new relation should be updated up to a distance of  $\partial$  hops. If either one of the ends of this relation has been detected to have fraudulent transaction (in this case node A), we store the new relation and levels as mentioned above, but we should also update the score values of all the nodes that are related to this new relation up to a distance of  $\partial$  hops. In this case the score of E is affected and changed to 1. Furthermore, a new relation is added to NPR as (A,E,2). TABLE I presents the NPR data records after the update for our example.

As described, when a new edge such as (A,D) is added nodes that are not affected by this change, such as B or C in the above example, do not get involved in the update procedure. This will save a lot of computation especially in real life networks that consists of millions of nodes and hundreds of millions of edges.

In this section we explained the proposed efficient data enrichment scheme. The proposed NPR enabled us not only to effectively use the network information for data enrichment and improve fraud detection rate, but also to provide an efficient framework to update only the necessary components. It should be noted that the proposed scheme can be used with any scoring algorithm or fraud detection technique. Although the NPR fields might need to be tuned to the used scoring or detection method.

TABLE I. Network Patterns Repository content after the update phase

Network Patterns Repository (NPR)													
Level info.													
A	B	1		D	E	1							
A	D	1		C	D	2							
A	C	1		B	E	2							
B	D	1		A	E	2							
B	C	1											
Related nodes													
A: {B,C,D,E} B: {A,C,D,E} C: {A,B,D} D: {A,B,C,E} E: {A,B,D}													
Scores info.													
A	0		B	2		C	2		D	2		E	1

#### IV. EXPERIMENTAL RESULTS

In the absence of public data sources in the financial domain, specially transactional datasets with information about social relations, we used the labeled financial data of PKDD'99 [13] in order to form the social relations of entities combined with synthetic transactions generated using the simulator in [14]. Although the PKDD'99 dataset labels are about loan payments and data has been gathered to find out which client deserves to be offered additional services to, but because of availability of financial transactions data, demographic information and validity of this dataset, it has been used here to form the social network of entities. Relations are based on demographic information, the branch where the customer account is held, and the date of each transaction. Typically, when a fraudulent transaction is detected in an account this increases the probability that other accounts in that branch have been compromised on the same date and time. Hence, account holders in a single branch that have transactions on the same date and time have been considered to be related together. This information will be used to improve the fraud detection rate.

In this dataset data has been gathered from 1993 to 1998, and relations are formed based on yearly aggregated data, so the yearly information can be used for the update phase.

Our dataset consists of about 150 selected nodes from 4500 nodes of the PKDD's dataset which have rich relations based on aggregated data of 1993. Using the information of next years, new nodes and relations are added to the dataset. Cardholders are categorized according to the amount of their aggregated transactions. For each node the transactions for one year are generated using the data generator of [14]. Then some random fraud cases are injected into the test data. Fraud cases have been selected such that they are very similar to normal transactions. This will increase the probability that they are not detected by conventional fraud detection methods such as [14].

Four different cases have been considered, in which the fraud cases are similar in terms of transaction amount (cases 1), transaction time (case 2), period among fraudulent transaction (case 3), and a combination of amount and period (case 4). The fraud cases were injected to only 10% of nodes. For evaluation purposes, we have assumed that 20% of these fraud cases have been detected and labeled. In TABLE II the characteristics of test data are demonstrated.

Labeled data are used to detect suspected transactions through social relations using proposed method. Based on relations, a score which represents the probability of being fraud activity is assigned to unknown transactions. Since all injected fraud cases are committed in a single month, the time slot is identical for all transactions. Hence each node's bad-score is assigned to all of its transactions committed in this month. Using this new attribute, the enriched data is used for fraud detection process. In section III this process is discussed in more detail.

TABLE II. Characteristics of test data

	Normal Transactions	Fraud Transactions	Socially Related Transactions
For 10% of cardholders	1000 Month 1 to 12	100 in average Months 10 to 11	50 in average Months 10 to 11

First the performance of the proposed model is compared with the hybrid system proposed in [14]. The results for all four cases are presented in Fig. 5. As mentioned above since the injected fraud cases were very close to normal cases they were not detected by typical anomaly detection methods such as [14]. Results indicate that the majority of these cases were detected using the relations among network entities.

In addition to accuracy, we should evaluate the proposed updating scheme in terms of its complexity. For this case it has been compared with the complexity of conventional SNA methods where the whole network is being updated for each new change. In our proposed scheme, we store all the required data to specify edges which can change a node's bad-score during the startup phase. As a results, the update phase for each node can be performed with a lower complexity of order  $O(E)$ , since there is no need to search for related nodes. In full update, and for  $\partial = 2$  each edge and also next level edges should be considered to update the scores. This results in a complexity of order  $O(E^2)$ . Similarly, the maximum complexity when the  $\partial$  is the network diameter has an order of  $O(E^\partial)$ . The results have been reported in Table III.

Although the results indicate a much lower complexity of the proposed scheme compared to full update, but it should be noted that it cannot handle nodes that have been removed or disconnected with time. This will be addressed in the next version of the proposed method in future works.

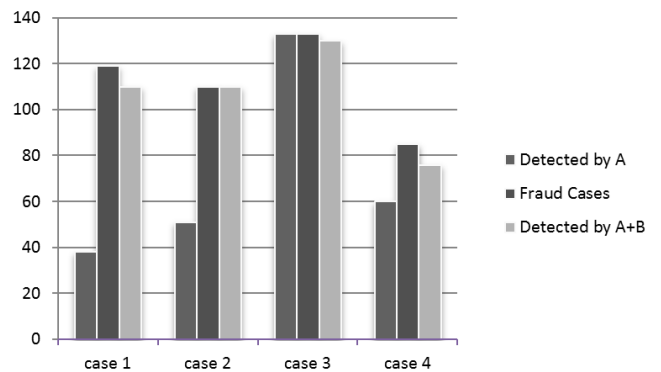


Figure 5: Accuracy comparison of Hybrid system of [14] and the proposed scheme. A represents the hybrid system and B represents the proposed scheme.

TABLE III. Complexity of algorithm for two type of update

	Startup	Update (for each node)
Full Update	$O(E^\delta)$	$O(E^\delta)$
Efficient Update	$O(E^\delta)$	$O(E)$

## V. CONCLUSION

The inherent dynamism of social networks, and any social network based system for that matter, emphasizes the need for a system that could be updated in an efficient way. In many existing models, the update phase is not considered at all. Hence, even though their initial model may be reasonably accurate, their lack of efficiently updating the model especially in real time systems makes them inapplicable.

In this paper, we introduced a new model for using the social network analysis to enrich data and to help make fraud detection more accurate. In this new model, some network information that could make the update phase more efficient has been stored in a repository, referred to in this paper as Network Pattern Repository (NPR). For each change, using this stored data only a portion of network information is being updated.

## REFERENCES

- [1] CyberSource's 2012 Online Fraud Report – 13th Annual Edition, available at: <http://www.cybersource.com/fraudreport2012>, Last access 27 May 2012.
- [2] J. Botelho and C. Antunes, "Combining social network analysis with semi-supervised clustering: a case study on fraud detection," *In proceeding of: Mining Data Semantics (MDS'2011) in conjunction with SIGKDD*, 2011.
- [3] M. P. San-Bento Almeida, "Classification for fraud detection with social network analysis," *MSc. Dissertation, Lisbon University of Technology*, 2009.
- [4] Gomes João Nascimento Social mining no combate à fraude [Journal]. - [s.l.] : TECH&BIZZ - Novabase, 2008.
- [5] J. C. Wang and C. Chiu, "Detecting online auction inflated-reputation behaviors using social network analysis," *Annual Conference of the North American Association for Computational Social and Organizational Science (NAACSOS 2005)*, 2005.
- [6] D. B. DeFigueiredo and E. T. Barr, "Trustdavis: A non-exploitable online reputation system," *Seventh IEEE International Conference on E-Commerce Technology*, pp. 274–283, 2005.
- [7] Z. Yanchun, Z. Wei, and Y. Changhai, "Detection of feedback reputation fraud in taobao using social network theory," *2011 International Joint Conference on Service Sciences*, pp. 188-192, 2011.
- [8] R. S. Burt, "Models of network structure", *Annual Review of Sociology*, Vol. 6, 1980, pp 79-141.
- [9] P. R. Monge, and E. M. Eisenberg, "Emergent communication networks", in *Handbook of Organizational Communication*, F. M. Jablin et al (eds.), Sage Publications, Newbury Park, 1987, pp. 305- 342
- [10] M. Morzy, "New algorithms for mining the reputation of participants of online auctions," *Internet and Network Economics*, pp. 112–121, 2005.
- [11] Y. Ku, Y. Chen, and C. Chiu, "A proposed data mining approach for Internet auction fraud detection," *Intelligence and Security Informatics*, pp. 238–243, 2007.
- [12] Šubelj L, Furlan Š, Bajec M. "An expert system for detecting automobile insurance fraud using social network analysis," *Expert Systems with Applications*, Vol. 38, No. 1, pp. 1039-1052, 2011.
- [13] "Discovery challenge guide to the financial data set", PKDD-99, 1999, <http://lisp.vse.cz/pkdd99>.
- [14] Leila Seyedhossein, Mahmoud Reza Hashemi, "A hybrid profiling method to detect heterogeneous credit card frauds", *7th International ISC Conference on Information Security and Cryptology (ISCISC'10)*, September 15-16 2010, Tehran, Iran.