Applied Artificial Intelligence
An International Journal

ISSN: 0883-9514 (Print) 1087-6545 (Online) Journal homepage: https://www.tandfonline.com/loi/uaai20

Feature Extraction Method Based on Social Network Analysis

Zahra Karimi Zandian & Mohammad Reza Keyvanpour

To cite this article: Zahra Karimi Zandian & Mohammad Reza Keyvanpour (2019) Feature Extraction Method Based on Social Network Analysis, Applied Artificial Intelligence, 33:8, 669-688, DOI: 10.1080/08839514.2019.1592347

To link to this article: https://doi.org/10.1080/08839514.2019.1592347

APPLIED ARTIFICIAL INTELLIGENCE
2019, VOL. 33, NO. 8, 669–688
https://doi.org/10.1080/08839514.2019.1592347

Taylor & Francis
Taylor & Francis Group

Feature Extraction Method Based on Social Network Analysis

Zahra Karimi Zandian[a] and Mohammad Reza Keyvanpour[b]

[a]Data Mining Lab, Department of Computer Engineering, Alzahra University, Tehran, Iran; [b]Department of Computer Engineering, Alzahra University, Tehran, Iran

ABSTRACT

Due to rapid development of Internet technology and electronic business, fraudulent activities have increased. One of the ways to cope with damages of them is fraud detection. In this field, there is a need for methods accurate and fast. Therefore, a novel and efficient feature extraction method based on social network analysis called FEMBSNA is proposed for fraud detection in banking accounts. In this method, in order to increase accuracy and control runtime in the first step, features based on network level are considered using social network analysis and extracted feature is combined with other features based on user level in the next phase. To evaluate our feature extraction method, we use PCK-means method as a basic method to learn. The results show using the proposed feature extraction as a pre-processing step in fraud detection improves the accuracy remarkably while it controls runtime in comparison with other methods.

Introduction

Fraud is a deliberately deceptive and misleading activity that is different from definitions of normal behavior. Fraud detection describes the methodologies deployed to investigate allegations of fraud. It is more reactive than proactive (Petrucelli 2013). In other words, when fraud occurs, it can be detected by different methods, as in the event of unauthorized use of another person's personal information. Therefore, fraud detection involves a review of historical transactions to identify indicators of a nonconforming transaction (Cendrowski et al. 2007).

On the one hand, in fraud detection area, studies show one of challenges of many existing methods is not to consider features based on user level and network level simultaneously to learn and investigating these two kinds of features can help to increase the accuracy of fraud detection methods.

On the other hand, the Computer Industry has seen a large growth in technology particularly in access, storage, and processing. This, combined with the fact that there are huge amounts of data to be processed has paved the way for data analysis and mining to derive potentially useful information. Various

CONTACT Zahra Karimi Zandian z.karimizandian@yahoo.com Data Mining Lab, Department of Computer Engineering, Alzahra University, Tehran, Iran.

demands ranging from commercial to military need to analyze data in an efficient and fast manner (Foschi et al. 2002). One of the issues related to data is to convert raw data into a set of useful features and another one is to identify the best and most useful features to analyze and extract (Guyon and Elisseeff 2006). Feature extraction can be the pre-processing step of Data Mining. Feature extraction is to extract patterns and derive knowledge from large collections of data with identification and extraction of unique features for a particular domain. Though there are various features available, the aim is to identify the best features and thereby extract relevant information from the data (Foschi et al. 2002). Today feature extraction is used in many fields such as image processing, text mining, signal processing, and pattern recognition.

Therefore, in this paper to cope with this challenge, we propose a novel and efficient feature extraction method based on social network analysis for fraud detection in banking accounts. In this method, in order to increase accuracy and control runtime in the first step, features based on network level are considered using social network analysis and extracted feature is combined with other features based on user level in next phase. To evaluate our feature extraction method and learn, PCK-means method will be used.

The rest of the paper is organized as follows: In Section 2, related works are discussed. In Section 3 the proposed feature extraction method is introduced. Evaluation results are presented in Section 4, followed by the concluding remarks in Section 5.

Related Works

Jamshidi et al. (Jamshidi and Hashemi 2012) propose a new feature extraction method based on social network analysis called bad-score to improve fraud detection. The proposed method is created from four phases: building social network, analyzing patterns, storing patterns, and updating. In this paper, various features of transactions are used to detect fraud. Carneiro et al. (Carneiro, Figueira, and Costa 2017) developed a method to detect fraud in credit cards that combine manual and automated classification. In this paper, features and properties of credit cards are used. Save et al. (Save et al. 2017) devised a novel system for credit card fraud detection based on decision tree with a combination of Luhn's algorithm and Hunt's algorithm using features of credit cards. Behera et al. (Behera and Panigrahi 2017) propose a fraud detection method based on fuzzy clustering and neural network using features of credit cards. Botelho et al. (Botelho and Antunes 2011) developed a feature that is obtained from social network called badRank to help improve the fraud detection using semi-supervised learning. Chiu et al. (Chiu et al. 2011) propose features extracted from social network as the input of fraud detection classifiers. In (Almeida, 2009), by analyzing social network, the patterns that are common to fraudulent entities are identified

and each entity is described by its original features plus another one for each pattern. Finally, these features are used in classification methods. Subelj et al. (Šubelj, Furlan, and Bajec 2011) use some features extracted from social network to detect fraud. In (Panigrahi et al. 2009) the use of features obtained from transaction history databases and the current and past behavior of credit cards to detect fraud is proposed. Sadaoui et al. (Sadaoui and Wang, 2015) propose a real-time framework that observes the progressing auctions to be able to take actions on time and set a fraud score for each user. This fraud score represents the user's behavior in past auctions. In (Zaslavsky and Strizhak 2006) a fraud detection method based on neural network is proposed. Self-organizing map algorithm is used to extract cardholders' behavior and to learn and classify this behavior. Krivko (Krivko 2010) uses features based on the user to propose a model to detect fraud. The proposed data-customized approach combines elements of supervised and unsupervised methodologies aiming to compensate for the individual deficiencies of the methods. Chang et al. (Chang and Chang 2010) use changes of behavior in each user to detect fraud. In this paper, clustering techniques are used to distinguish changes in behavior.

Reviewing the proposed methods in fraud detection area and classification, we proposed in (Zandian and Keyvanpour 2016) fraud detection methods which, based on the features of entities, can be divided into two categories: fraud detection methods based on user-level features and fraud detection methods based on network level features (Figure 1). In fraud detection methods based on user-level features, it is sufficient to investigate inherent and exclusive features derived from a specific component (Zandian and Keyvanpour 2016). According to the classification presented in (Zandian and Keyvanpour 2016), in methods based on network level features, features of each component are obtained considering a component position along with other components. The features then participate in fraud detection. These methods usually use the connections between components to obtain

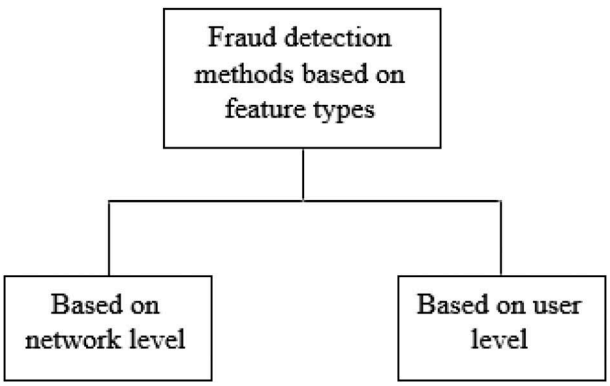

Figure 1. Classification of fraud detection methods based on feature types (Zandian and Keyvanpour 2016).

new features. To achieve this goal, social networks comprising these components are paid attention to, and useful information are obtained from them (Zandian and Keyvanpour 2016).

According to (Zandian and Keyvanpour 2016), the speed of detection in methods based on user level is higher than that in methods based on network level. In contrast, the complexity of methods based on network level is more than that of the methods based on user level (Gaol et al. 2013). Also, the accuracy of these two kinds of methods is not high (Jamshidi and Hashemi 2012).

As a result, feature extraction method proposed in this paper uses a combination of these two kinds of features to increase the speed and accuracy of fraud detection.

FEMBSNA: An Efficient Feature Extraction Method Based on Social Network Analysis

As it has been mentioned in (Zandian and Keyvanpour 2016), features of the components under consideration can be divided into two categories: network-based and user-based. Features based on the network level are features which consider components obtained in the presence of other components and include a set of components that are related to each other according to their relationships with others (Lin et al. 2012; Lin and Khomnotai 2014; Šubelj, Furlan, and Bajec 2011). While features based on the user level are features that belong to a certain component without regard to the relationship between that component and others (Chang and Chang 2010; Panigrahi et al. 2009; Phua et al. 2010; Yu and Lin 2013). Combining algorithms each of which has focused on various aspects of information hidden in the data can help detect fraudulent accounts more accurately (Zandian and Keyvanpour 2016), because in the first type of algorithms, existence or non-existence of relationships between the data is ignored and in the second type of algorithms, individual frauds are not considered to be important. Therefore, in the proposed feature extraction method, we have used both feature types to detect fraudulent accounts. The challenge of many existing methods in this area is not to consider these two feature types simultaneously.

The block diagram for the proposed system of feature extraction (FEMBSNA) has been shown in Figure 2. Accordingly, as shown in Figure 2, FEMBSNA involves two steps:

- In step 1, in order to provide and obtain new features, a social network of financial interactions is created and analyzed. The proposed method in this paper uses transactions record as input for the CAA phase.
- In step 2, features obtained from CAA phase and saved in NDR and accounts record stored in SDR are used as inputs for the CF phase. In

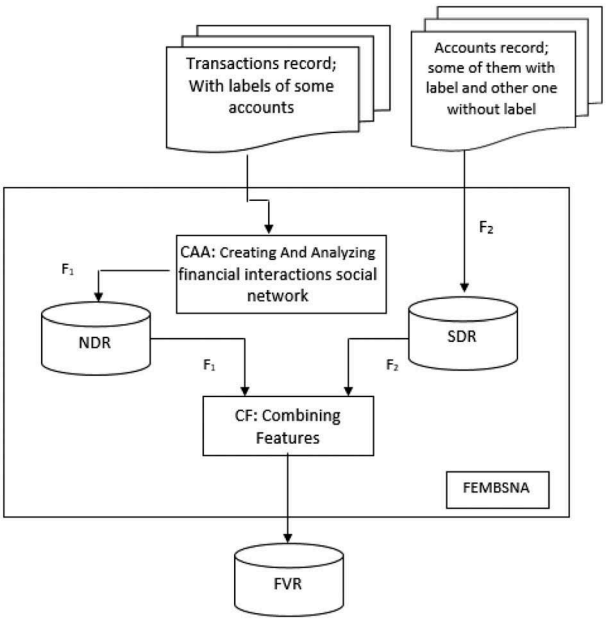

Figure 2. Block diagram of FEMBSNA.

the CF phase, features and network data belonging to accounts obtained from Phase 1 are combined with features based on the user level of the existing accounts record called simple. These features are then shown as accounts features vectors and saved in FVR.

Using criteria that demonstrate possible scenarios of fraudulence and factor increasing the risk of fraud accounts can have an essential role in increasing fraud detection accuracy (Jamshidi 2014). In order to achieve this goal, CAA is proposed and used in this paper. More specifically, the feature that is extracted from this network shows the score of fraudulence of each account because of the relationships with other fraudulent accounts. As illustrated in Figure 3, this phase is based on creating and analyzing financial interactions in social network. In the first step financial interactions social network is created by receiving transactions record and network data is achieved and saved in network data repository (NDR). This repository is used to save network data in the next step.

Description of observations used in Figures 2 and 3 are presented in Table 1.

Creating Financial Interactions Social Network

In this paper, an implicit social network called financial interactions social network is used. What is important in detecting fraud in financial interactions is financial transactions between accounts (Jamshidi and Hashemi

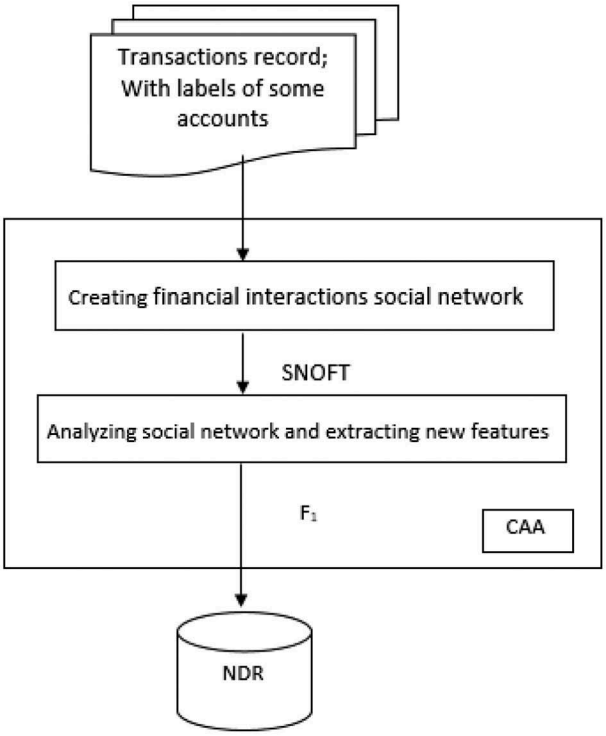

Figure 3. Bock diagram of proposed method of CAA.

Table 1. Description of observations used in Figures 2 and 3.

Observation	Description
FEMBSNA	Feature Extraction Method Based on Social Network Analysis
FVR	Feature Vectors Repository
NDR	Network Data Repository
F_1	Feature 1
F_2	Feature 2
SDR	Simple Data Repository
CAA	Creating And Analyzing financial interactions social network
CF	Combining Features
SNOFT	Social Network Of Financial Transitions

2012). Thus, by considering accounts as the nodes (Kosorukoff and Passmore 2011) of financial interactions social network and financial transactions as edges (Kosorukoff and Passmore 2011), hidden features in this network can be extracted. According to the proposed method in this paper, accounts for which at least a single transaction exists have a relationship (Aggarwal 2011). Another remarkable property in the proposed network is that if the account of the receiver of the transaction and its sender account are different, the relationship between accounts is directed (Wasserman and Faust 1994). In Figure 4, an example of this kind of network is showed. Another important

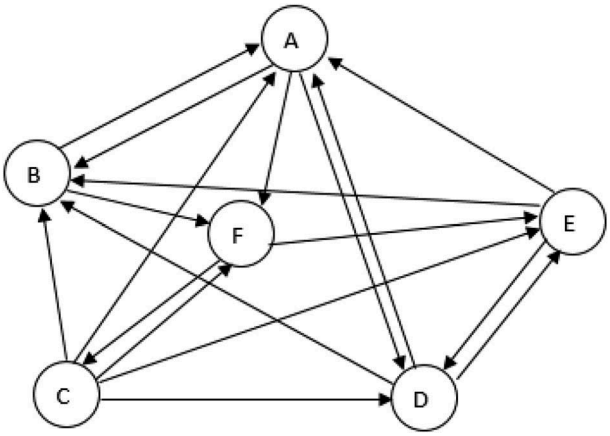

Figure 4. An example of directed social network.

feature of the proposed network, in addition to being directed, is having weight. The relationship between any two nodes with each other is not equal and as a result to obtain the nodes' fake scores they will not have the same influences. These weights (Wasserman and Faust 1994) can depend on various factors but what is considered in this paper is the number and total value of transactions between any two accounts affecting the weight of the edge between them. It seems likely that an account that is related to with a fraudulent account for many times with a high total value has a higher possibility of being fraudulent than an account that is related to a fraudulent account less frequently with lower total value. As a result, edges between nodes are also weighted.

As mentioned before, the proposed network is weighted and the weights of the edges affect the number and total value of transactions from account i to account j. The number and amount of transactions to total number and total amounts of transactions show their impacts on the social network and as it is clear in Equations (1)-(3) factors affecting the weights of edges are the relative number (RNT) and relative sum of the amounts of transactions (RSTA) from account i to account j.

$$RNT_{ij} = NT_{ij}/TNT \tag{1}$$

$$RSTA_{ij} = STA_{ij}/TSTA \tag{2}$$

$$weight_{ij} = \alpha.RNT_{ij} + (1-\alpha).RSTA_{ij}\ 0 \leq \alpha \leq 1 \tag{3}$$

Applying these two suggestions to financial interactions social network affects the accuracy of fraud detection. The pseudo code for the financial interactions social network phase in the proposed CAA method is presented in Figure 5.

	Algorithm: creating financial interactions social network **Input:** transactionData, accountData **Output:** *relationship matrix*
1	Assign accounts as nodes and transactions as directed edges between nodes in social
2	network
3	For each two nodes
4	Unite all edges and save the number and amount of transactions between them
5	Calculate the weight of edge using Equations (1), (2) and (3)
6	End
7	Create matrix including edges (two nodes) and their weights

Figure 5. Pseudo code for creating financial interactions social network phase in the proposed CAA method.

Analyzing Social Network and Extracting New Features

In analyzing social network phase, using special criteria social network created is analyzed and new features for each account is extracted. Generally, algorithms used to analyze social network focus on entities and nodes of the social network (Jamshidi 2014) while what seems to be more important is the information hidden in relationships between accounts in fraud detection in bank accounts. For example, Hits (Kleinberg 1999) and PageRank (Haveliwala 1999) which are basic algorithms for social network analysis focus on determining the centrality of web pages. Other examples from this kind of algorithms are BadRank (Botelho and Antunes 2011) and gspan (Almeida, 2009) that pay attention to the nodes of the potential network and do not consider relationships and their complexity and conditions involved in financial interactions. In front of these algorithms, the algorithms have proposed that have paid attention to relationships in the simple social network (Jamshidi 2014). Hence, in this paper, the directed and weighted network (Wasserman and Faust 1994) is analyzed to propose new features.

A new feature called Fake_score proposed in this paper shows the fraudulence score. Thus, a higher score of an account means that the account has a stronger relationship with fraudulence accounts. According to the proposed criterion in (Jamshidi 2014; Jamshidi and Hashemi 2012), this criterion, in general, depends on three factors:

- Distance from fraudulence nodes;
- Sum of the degrees of the nodes existing in the paths;
- Number paths ended to fraudulence nodes;

Equation (4) shows how to calculate the Fake_score.

$$\begin{aligned} Fake\_score = {} & A * PathElement(i) + B * DegreeElement(i) \\ & + (1 - (A + B)) * \ EndPointElement(i) \\ & 0 \leq A \leq 1, \ 0 \leq B \leq 1 - A \end{aligned} \tag{4}$$

As shown in Equation (4), to calculate the Fake_score for each account, we use the weighted mean of three factors PathElement, DegreeElement and EndPointElement that define Distance from fraudulence nodes, Sum of the degrees of the nodes in the paths and Number paths ending in fraudulence nodes. The degree of importance of each factor can change in various situations to calculate the Fake_score.

Since the main purpose of this study is to propose a new feature extraction method to increase the accuracy and speed of fraud detection, it must be noted that each relationship is not necessarily important. Considering a node with a number of weighted indegrees or outdegrees (Wasserman and Faust 1994) that are more than a special threshold is not useful and investigating the paths including this node is not necessary. In this paper, the average number of indegrees and outdegrees of all nodes in the network are special thresholds for a number of weighted indegrees and outdegrees of each node, respectively. Another point investigated here is that if the distance between the examined node and the fraudulence nodes is higher than a threshold, that node does not seem to be dangerous. This is because fraudsters always try to show normal behavior not to be detected quickly to achieve their goal. The maximum of the path (Wasserman and Faust 1994) can also be changed from 2 up to the network diameter (Wasserman and Faust 1994) and as mentioned in (Jamshidi and Hashemi 2012), the suitable value to achieve the highest accuracy is the average length of all paths in the network because whenever the search space depth in the network becomes more than this value, the possibility of fraud through relationships with fraudulent accounts becomes lower and this search uses more time. In contrast, whenever the search space depth becomes less than the usual value, accuracy becomes lower; hence, we have to instate a trade-off between accuracy and speed. Thus, the length of paths (Kosorukoff and Passmore 2011) investigated here has been set to 4 for the approved data set. In equations, the length of this path has been shown by Ψ.

Distance from Fraudulence Nodes Factor

In the proposed criteria, in order to apply the effect of distance from fraudulent nodes, PathElement component is used. Equations (5)-(7) show how to calculate this factor.

$$PathElement(i,F) = \sum\nolimits_{AllPath\_With\_Lenght \le \psi\_Between(i,F)} y \tag{5}$$

$$y = \begin{cases} 1 & SOWOP > (AOWOP) * \psi \\ SOWOP/((AOWOP) * \psi) & else \end{cases} \tag{6}$$

$$PathElement(i) = \frac{\sum_{AllFraudAccounts} PathElement(i, FraudAccount)}{NOP\_With\_Lenght \le \psi\_Between(i, AllFraudAccounts)} \tag{7}$$

This distance is influenced by the number of aligned edges between investigated node (i) and fraudulent node (F) and the weight of the mentioned edges. These aligned edges together are the paths in the directed graphs. An example for the existence of paths between an investigated node (i) and a fraudulent node (F) has been shown in Figure 6. For each path between i and F, Equation (6) is calculated and after that according to Equation (5) these amounts are calculated together. As mentioned in Equation (7) according to the proposed method PathElement of node i will be obtained by mean of the calculated PathElement between i and each fraudulent node F. The more this value is, the more its effect is and the less this value is, the less its effect is on Fake_score of the investigated account.

According to Equation (6), it is clear that this factor has the highest effect on the Fake_score if the sum of the weights of existing edges in the path is more than the sum of the weights of edges in a path with length ψ with the weight that is the average of all of the weights in network.

Sum of the Degrees of Nodes Existing in the Paths Factor

Degrees of existing nodes influence the Fake_score by DegreeElement component. Sum of the indegrees and outdegrees (Trudea, 2013) in the path is examined separately to calculate the value of this component and based on

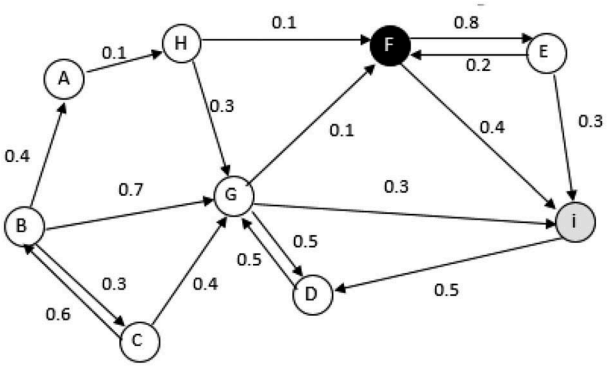

Figure 6. An example for the existence of paths between an investigated node (i) and a fraudulent node (F).

their weights determine the sum of indegrees and outdegrees separately in a normal path. In this paper, a normal path is considered but the weighted indegree and outdegree of each node in that path (Wasserman and Faust 1994) is the mean of the weighted indegrees and the mean of weighted outdegrees in all paths of that network with length ψ. According to Equations (9)-(11), if this relation in the path between investigated node and a fraudulent node for all indegrees and outdegrees are at least equal to 1, the DegreeElement value for that node will equal 0. It means that this component will have the least effect on the Fake_score of that node. Otherwise, if both of these ratios are less than 1, the value of this component is the maximum of these two ratios for the path; Finally, if only one of these ratios is at least equal to 1, the DegreeElement value for that node will equal the minimum of the ratio of indegrees in the path to indegrees in a normal path and the ratio of outdegrees in the investigated path to outdegrees in a normal path. After calculating the DegreeElement for all paths between the investigated node and a fraudulent node, as mentioned in Equation (8) with regard to a fraudulent node, the DegreeElement for that node will be calculated by sum of the obtained DegreeElement from all paths between the investigated node and that fraudulent node. Consequently, in order to calculate the total DegreeElement for each node, the obtained DegreeElement values from all paths between the investigated node and all fraudulent nodes are averaged (Equation (12)).

$$DegreeElement(i,F) = \sum\nolimits_{AllPath\_With\_Lenght \le \psi\_Between(i,F)} x \tag{8}$$

$$x = \begin{cases} 0 & if\ z \ge 1\ and\ s \ge 1 \\ max\left(\frac{1}{z}, \frac{1}{s}\right) & if\ z < 1\ and\ s < 1 \\ min\ \left(\frac{1}{z}, \frac{1}{s}\right) & else \end{cases} \tag{9}$$

$$z = \frac{\sum_{AllNodesInPath} WOIE}{(AOWOE\ ) * (NID) * (PathLenght - 1)} \tag{10}$$

$$s = \frac{\sum_{AllNodesInPath} WOOE}{(AOWOE) * (NOD) * (PathLenght - 1)} \tag{11}$$

$$DegreeElement(i) = \frac{\sum_{1}^{AllFraudAccounts} DegreeElement(i, FraudAccount)}{NOP\_With\_Lenght \le \psi\_Between(i, AllFraudAccounts)} \tag{12}$$

Table 2. Description of observations used in Equations (4)-(13).

Observation	Description
SOWOP	Sum_Of_Weights_Of_Path
AOWOE	Average_of_Weights_of_Edges
NOP	Number_of_Paths
NID	NormalInDegree
WOIE	Weight_Of_InputEdge
NOD	NormalOutDegree
WOOE	Weight_Of_OutputEdge

Number Paths Ending in Fraudulence Nodes Factor

As mentioned before, the number of fraudulent nodes that have relationships (Aggarwal 2011) with other nodes is important to calculate the Fake_score. According to the proposed method in this paper, when two nodes have a relationship with each other, the distance (Wasserman and Faust 1994) between those two nodes in the network is utmost ψ. So, the third component affecting the Fake_score called the EndPointElement is defined. According to Equation (13), if all nodes related to the investigated node are frauds, the maximum value for this factor that equals 1 will be obtained. In contrast, if none of the related nodes are frauds, this component has the least effect on the Fake_score feature.

$$EndPoint\ Element(i) = \frac{|\{node\ w | w \in RelatedNodes\_with\_i \&\&\ Label(w) =' fraud'\}|}{|RelatedNodes\_with\_i|} \tag{13}$$

Description of observations used in Equations (4)-(13) is expressed in Table 2.

Pseudo code of analyzing the social network and extracting new feature phase in the proposed CAA method is presented in Figure 7.

Experiments

Dataset

In the absence of public data sources in the financial domain, especially transactional datasets with information about social relations, we used the financial data of PKDD'99 (Berka and Sochorova 1999). This dataset is used to evaluate many methods in different fields (Buda et al. 2017; Frank, Moser, and Ester 2007; Zall 2015; Zhang and Tay 2016). Due to the availability of financial transactions data, demographic information and validity of this dataset, it has been used here to test our proposed method. We have used transactions table to form social network and accounts table to extract simple data. We have also applied some changes on transactions table like

```
Algorithm: Analyzing social network and extracting new feature
Input: relationship matrix
Output: Fake_score
For each node- fraudulent node
 For each fraudulent node
  Calculate PathElement(node- fraudulent node, fraudulent node) using Equation (5)
  Calculate DegreeElement(node- fraudulent node, fraudulent node) using Equation (8)
  End
  Calculate PathElement (node- fraudulent node) using Equation (7)
  Calculate DegreeElement (node- fraudulent node) using Equation (10)
  Calculate EndPointElement (node- fraudulent node) using Equation (11)
  Fake_score (node-fraudulent node) using Equation (4)
End
For each fraudulent node
    Fake_score (fraudulent node)=1;
end
```

Figure 7. Pseudo code of analyzing social network and extracting new feature phase in the proposed CAA method.

Table 3. Characteristics of used dataset.

Characteristic	Quantity
Number of accounts	387
Number of transactions	2070
Number of features of transactions	5
Number of features of accounts	3

eliminating transactions that are not transactions for transferring money or those accounts whose information. As shown in Table 3, our dataset consists of about 387 accounts selected from the accounts table and 2070 transactions from the transactions table. Each transaction has five features: trans-id, source_account-id, destination_account-id, amount, and date. Each account also has three features: account-id, district-id, and date.

Transactions data are used to calculate the Fake_score for accounts through social relations using the proposed method. Based on relations, a score which represents the probability of a fraud activity as a new feature is assigned to unknown accounts.

Evaluation Criteria

For performance evaluation of the proposed feature extraction methods for fraud detection, popular criteria are used: True Negative (TN) rate, False Positive (FP) rate, False Negative (FN) rate, precision, recall (also called True Positive (TP) rate), F_1score, accuracy, and runtime.

- TNrate: as Equation (14) shows, it is the proportion of negatives that are correctly identified as such.

$$TNrate = \frac{TN}{TN + FP} \tag{14}$$

- FPrate: as stated in Equation (15), it is the proportion of negatives that are wrongly identified as positives.

$$FPrate = \frac{FP}{FP + TN} \tag{15}$$

- FNrate: the proportion of positives that are wrongly identified as negatives (Equation (16)).

$$FNrate = \frac{FN}{FN + TP} \tag{16}$$

- Precision: as shown in Equation (17), it is the number of items correctly labeled as belonging to the positive class (TP) divided by the total number of elements labeled as belonging to the positive class (i.e., the sum of true positives and false positives, which are items incorrectly labeled as belonging to the class)

$$precision = \frac{TP}{TP + FP} \tag{17}$$

- Recall: the number of true positives divided by the total number of elements that actually belong to the positive class (i.e., the sum of true positives and false negatives, which are items which were not labeled as belonging to the positive class but should have been) (Equation (18)).

$$recall = TPrate = \frac{TP}{TP + FN} \tag{18}$$

- F_1score: as stated in Equation (19), it is the harmonic mean of precision and recall.

$$F_1score = \frac{2 * precision * recall}{precision + recall} = \frac{2TP}{2TP + FP + FN} \tag{19}$$

- Accuracy: the proportion of positives and negatives that are correctly identified as such (Equation 20).

$$accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{20}$$

- Runtime: the time used to perform the method completely, obtain the results, and label the data.

Experimental Results

For evaluating FEMBSNA, we have compared PCK-means with Fake_score feature obtained from FEMBSNA with other four methods: bad-score feature proposed in (Jamshidi 2014), Fake_score obtained from undirected social network, Fake_score obtained from unweighted social network and without feature extraction phase based on eight criteria expressed in the section on evaluation criteria.

Our method is compared with what is proposed in (Jamshidi 2014) because the feature extracted in that paper has also been obtained from social network using relationships among the nodes (i.e., network level feature) but from a simple, undirected, and unweighted network.

As shown in Figure 8(a) and (b), TNrate and FPrate corresponding to PCK-means with the Fake_score feature obtained from FEMBSNA is higher than other methods. This means the proposed FEMBSNA has had a major effect on increasing TNrate, decreasing FPrate, detecting non-fraud accounts correctly, and reducing wrong alarms. This is because of correct and appropriate restrictions applied on Fake_score calculation such as considering weighted and directed network and ignoring non-effective factors in detecting frauds correctly. The length of the paths (ψ) has been set to 4 in order to control runtime and as some useful information hidden in longer paths may not be considered and relationships among some nodes may be ignored. Therefore, as shown in Figures 8(c) and 9(a), the rate of detection of fraud accounts (recall) is less and thus the amount of FNrate is more in the proposed method than other methods. Lack of feature extraction phase in learning has led to labeling as positive (fraud) the data that are not fraud and so its TNrate is lower and its recall is more than other. But in PCK-means with bad-score using the method proposed in (Kosorukoff and Passmore 2011), the recall of this method and TNrate amounts are modest and lie between our method and PCK-means method without feature extraction phase because it has used simple network and it is possible to calculate higher scores for a node because of finding relationships between that node and a fraud node while this is not distinguished in our methods. As shown in Figure 9(b) and with regard to Equation (17), because the difference between TP and FP amounts is more in PCK-means with bad-score than others,

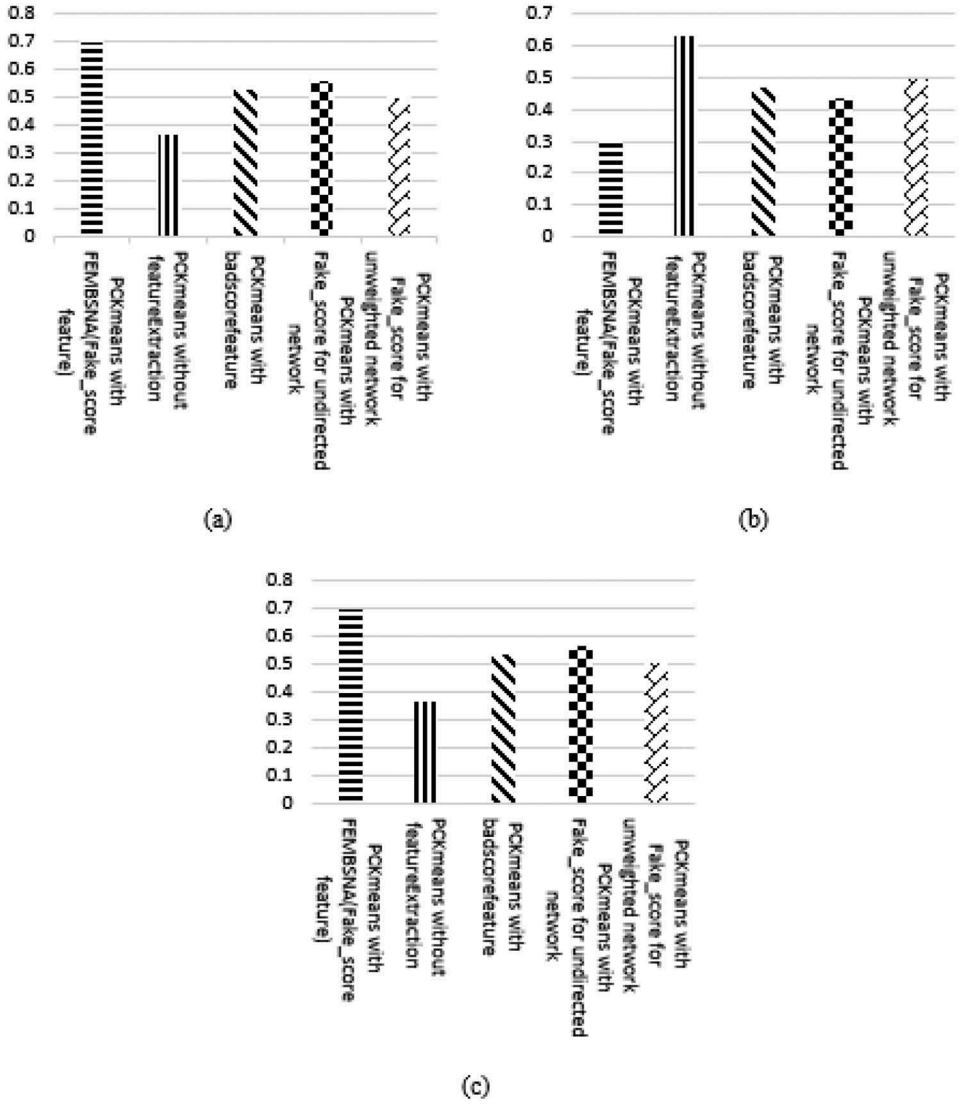

Figure 8. Comparison between the proposed method and other methods based on (a) TNrate, (b) FPrate, (c) FNrate.

the precision criterion for this method is better. All studied feature extraction methods in the paper have tried to reduce wrong detections in addition to increasing correct detections and also adding a feature showing probability of being fraud causes fraud detection rate (whether TP or FP) is lower than PCK-means without feature extraction phase. While TP is the only influencing parameter on F_1score, so as shown in Figure 9(c) and by paying attention to Equation (19), PCK-means without feature extraction phase is better based on F_1 score criterion. As shown in Figure 9(d), PCK-means with FEMBSNA is better than others based on accuracy. Because in this paper, the proposed method aimed to detect fraud and non-fraud correctly and simultaneously. Therefore,

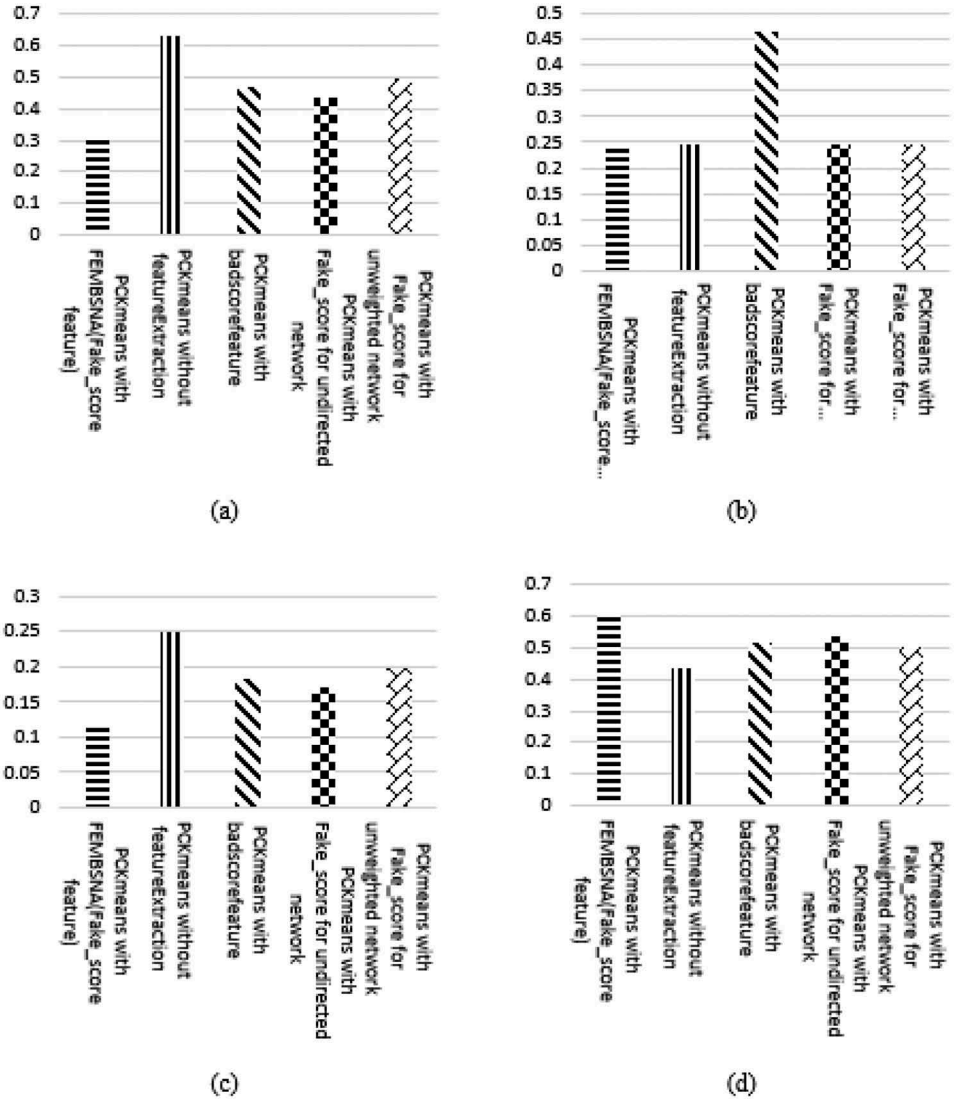

Figure 9. Comparison between the proposed method and other methods based on (a) recall, (b) precision, (c) F_1score, (d) accuracy.

we used the feature based on network level inside features based on user level. We also used weighted and directed network; using weighted network improved TNrate (Figure 8(a)) and using directed network improved recall (Figure 9(a)).

As shown in Fig. 15, it is clear that the runtime of PCK-means without feature extraction phase is very low. But among other methods, the study in (Jamshidi 2014) to obtain bad-score feature has used a simple network with less processing but the runtime of PCK-means with FEMBSNA method is similar. Runtime of PCK-means with Fake_score using unweighted network and undirected network is much higher because of higher complexity of the network and the large volume of the calculations.

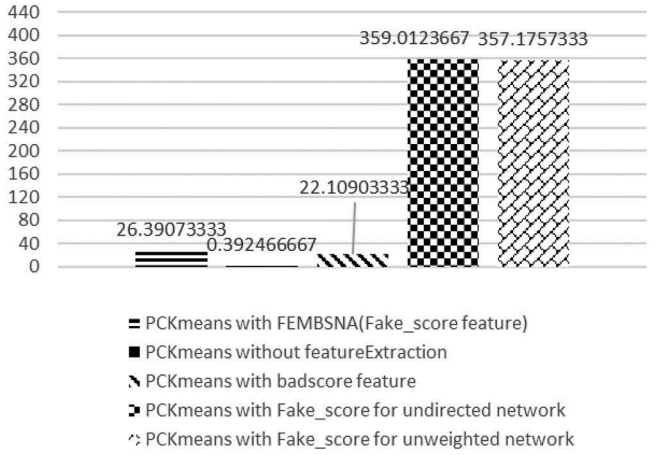

Figure 10. Comparison between the proposed method and other methods based on runtime.

Table 4. experimental results.

	TNrate	FPrate	FNrate	Precision	Recall	F_1score	Accuracy	Runtime(sec)
PCKmeans with FEMBSNA (Fake_score feature)	0.7	0.3	0.7	0.24	0.3	0.12	0.60	26.39
PCKmeans without featureExtraction	0.37	0.63	0.37	0.24	0.63	0.25	0.43	0.39
PCKmeans with badscorefeature	0.54	0.47	0.53	0.47	0.47	0.18	0.52	22.11
PCKmeans with Fake_score for undirected network	0.57	0.43	0.57	0.24	0.43	0.17	0.53	359.01
PCKmeans with Fake_score for unweighted network	0.50	0.49	0.50	0.25	0.50	0.20	0.50	357.17

Results obtained from experiments and shown in Figures 8–10 are come in Table 4 based on considered methods and evaluation criteria that are expressed in Section 4.2 by details. Numbers are rounded up to two decimal digits in this table.

Conclusion

In fraud detection area, there are two important challenges: accuracy and speed of detection (Raj and Portia 2011; Seeja and Zareapoor 2014). In this paper, a novel feature extraction method called FEMBSNA as a pre-processing step was proposed which uses user-level features and network level features (Zandian and Keyvanpour 2016). In this method, financial interactions social network is first created and analyzed and a new feature is extracted and then this feature combines with user-level features. This network is weighted and directed. It was shown in the experimental results that the use of this method as the pre-processing step for fraud detection improves the accuracy of detection remarkably while the runtime of fraud detection method is controlled and kept within an acceptable level compared to other methods.

References

Aggarwal, C. C. 2011. An introduction to social network data analytics. In *Social network data analytics*, 1–15. Charu C. Aggarwal (Ed.). Springer, Boston, MA.

Almeida, M. P. 2009. Classification for fraud detection with social network analysis. Masters Degree Dissertation, Engenharia Informática e de Computadores.

Behera, T. K., and S. Panigrahi. 2017. Credit card fraud detection using a neuro-fuzzy expert system. In *Computational intelligence in data mining*, 835–43. Behera, H. S., and D. P. Mohapatra (Eds.). Springer, Singapore.

Berka, P., and M. Sochorova. 1999. Discovery challenge guide to the financial data set. *PKDD-99*. https://sorry.vse.cz/~berka/challenge/pkdd1999/berka.htm9.

Botelho, J., and C. Antunes. 2011. Combining social network analysis with semi-supervised clustering: A case study on fraud detection. *in the Workshop of Mining Data Semantics (MDS'2011) in Conjunction with SIGKDD* 1–7. San Diego, CA, USA.

Buda, T. S., T. Cerqueus, C. Grava, and J. Murphy. 2017. ReX: Representative extrapolating relational databases. *Information Systems* 67:83–99. doi:10.1016/j.is.2017.03.001.

Carneiro, N., G. Figueira, and M. Costa. 2017. A data mining based system for credit-card fraud detection in e-tail. *Decision Support Systems* 95:91–101. doi:10.1016/j.dss.2017.01.002.

Cendrowski, H., L. Petro, J. Martin, and A. Wadecki. 2007. *The handbook of fraud deterrence*. John Wiley & Sons. Hoboken, New Jersey.

Chang, W. H., and J. S. Chang. 2010. Using clustering techniques to analyze fraudulent behavior changes in online auctions. in 2010 International Conference on Networking and Information Technology. 34–38. Manila, Philippines.

Chiu, C., Y. Ku, T. Lie, and Y. Chen. 2011. Internet auction fraud detection using social network analysis and classification tree approaches. *International Journal of Electronic Commerce* 15:123–47. doi:10.2753/JEC1086-4415150306.

Foschi, P. G., D. Kolippakkam, H. Liu, and A. Mandvikar. 2002. Feature extraction for image mining. International Workshop on *Multimedia Information Systems (MIS 2002)*. Tempe, Arizona, USA.

Frank, R., F. Moser, and M. Ester. 2007. A method for multi-relational classification using single and multi-feature aggregation functions. in European Conference on Principles of Data Mining and Knowledge Discovery. 430–37. Warsaw, Poland.

Gaol, F. L., S. Kadry, M. Taylor, and P. S. Li. 2013. Recent trends in social and behaviour sciences. in Proceding of the 2nd International Congress on Interdisciplinary Behaviour and Social Sciences, (ICIBSoS 2013). JAKARTA. INDONESIA.4–5.

Guyon, I., and A. Elisseeff. 2006. An introduction to feature extraction. In *Feature extraction*, 1–25. Guyon I., Nikravesh M., Gunn S., and L.A. Zadeh (Eds.). Springer, Berlin, Heidelberg.

Haveliwala, T. 1999. Efficient computation of PageRank. Technical report, Stanford University. Stanford, California.

Jamshidi, S. 2014. Developing a dynamic multi-level model for creatinging a behavioral profile to detect fraud in electronic payments. M.Sc. School of Electerical and Computer Engineering, Tehran Univesity, Tehran.

Jamshidi, S., and M. R. Hashemi. 2012. An efficient data enrichment scheme for fraud detection using social network analysis. in Sixth International Symposium on Telecommunications (IST). 1082–87. Tehran, Iran.

Kleinberg, J. M. 1999. Authoritative sources in a hyperlinked environment. *Journal of the ACM (JACM)* 46:604–32. doi:10.1145/324133.324140.

Kosorukoff, A., and D. L. Passmore. 2011. *Social network analysis: Theory and applications*. Passmore, D. L. Pennsylvania State University, USA.

Krivko, M. 2010. A hybrid model for plastic card fraud detection systems. *Expert Systems with Applications* 37:6070–76. doi:10.1016/j.eswa.2010.02.119.

Lin, J. L., and L. Khomnotai. 2014. Using neighbor diversity to detect fraudsters in online auctions. *Entropy* 16:2629–41. doi:10.3390/e16052629.

Lin, S. J., Y. Y. Jheng, and C. H. Yu. 2012. Combining ranking concept and social network analysis to detect collusive groups in online auctions. *Expert Systems with Applications* 39:9079–86. doi:10.1016/j.eswa.2012.02.039.

Panigrahi, S., A. Kundu, S. Sural, and A. K. Majumdar. 2009. Credit card fraud detection: A fusion approach using Dempster–Shafer theory and Bayesian learning. *Information Fusion* 10:354–63. doi:10.1016/j.inffus.2008.04.001.

Petrucelli, J. 2013. *Detecting fraud in organizations: Techniques, tools, and resources*. John Wiley & Sons. Hoboken, New Jersey.

Phua, C., V. Lee, K. Smith, and R. Gayler. 2010. A comprehensive survey of data mining-based fraud detection research. *arXiv preprint arXiv:1009.6119*.

Raj, S. B. E., and A. A. Portia. 2011. Analysis on credit card fraud detection methods. in 2011 International Conference on Computer. Communication and Electrical Technology (ICCCET). 152–56. doi:10.1016/j.colsurfb.2010.08.036

Sadaoui, S., X. Wang, and D. Qi. 2015. A real-time monitoring framework for online auctions frauds. in International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems. 97–108. Seoul, Korea (Republic of).

Save, P., P. Tiwarekar, K. N. Jain, and N. Mahyavanshi. 2017. A novel idea for credit card fraud detection using decision tree. *International Journal of Computer Applications* 161:6–9. doi:10.5120/ijca2017913413.

Seeja, K., and M. Zareapoor. 2014. FraudMiner: A novel credit card fraud detection model based on frequent itemset mining. *The Scientific World Journal* 2014:1–10. doi:10.1155/2014/252797.

Šubelj, L., S. Furlan, and M. Bajec. 2011. An expert system for detecting automobile insurance fraud using social network analysis. *Expert Systems with Applications* 38:1039–52. doi:10.1016/j.eswa.2010.07.143.

Trudeau, R. J. 2013. *Introduction to graph theory*. Courier Corporation. United States.

Wasserman, S., and K. Faust. 1994. *Social network analysis: Methods and applications*, vol. 8. Cambridge university press. Cambridge, United Kingdom.

Yu, C. H., and S. J. Lin. 2013. Fuzzy rule optimization for online auction frauds detection based on genetic algorithm. *Electronic Commerce Research* 13:169–82. doi:10.1007/s10660-013-9113-4.

Zall, R. 2015. A semi-supervised learning based method for classification of multi-relational data. M.Sc. Faculty of computer Engineering, Alzahra University, Tehran.

Zandian, Z. K., and M. Keyvanpour. 2016. Helpful and efficient framework for classification and analysis of various fraud detection approaches from the perspective of time and features. in 4th International Conference on Applied Research in Computer Engineering and Signal Processing. https://www.civilica.com/Paper-CEPS04-CEPS04_005.html.

Zaslavsky, V., and A. Strizhak. 2006. Credit card fraud detection using self-organizing maps. *Information and Security* 18:48.

Zhang, J., and Y. Tay. 2016. Dscaler: Synthetically scaling a given relational database. Proceedings of the VLDB Endowment 9:1671–82.