

Analysis and Prediction of Fluctuations for Sector Price Indices with Cross-Correlation and Association Mining Based Networks: Tehran Stock Exchange Case

Arash Negahdari Kia, Dr. SamanHaratizadeh and Dr. ZainaboldodaHeshmati

Abstract--- Network science has become an ever-increasing and interesting field of research in the recent decade by focusing on finding hidden knowledge in complex networks. This study of complex relationships in network structures has also gained a lot of interest in the world of finance and stock markets. The paper focuses on prediction in Tehran Stock Exchange (TSE) and analysis of the relationship between different sectors, looking into the market price indices data of different market sectors and their fluctuations over time. Four different network structures have been extracted from the TSE market data, two with association rules mining and two with Pearson cross-correlation. Using the correlation with different threshold cuts, different networks have been created and importance of market sectors have been analyzed using different network centrality measurements. Then, by using Apriori algorithms, association patterns in fluctuation of the price indices are extracted for building different directed networks. These directed networks are used in assessing current market dynamics as well as predicting future market price fluctuations that is tested through an evaluation method.

Keywords--- Association Rules Mining, Complex Networks, Fluctuation Prediction, Stock Market

I. INTRODUCTION

NETWORK science has shown enormous applications in many interdisciplinary fields in the recent decade. Any structure or natural phenomena that can be modeled as a network of nodes and edges can be studied by the means of graph theory and network science. This study is about using networks in the field of finance and economy, in particular, analysis of Tehran Stock Exchange (TSE) market and fluctuation prediction of different market sectors price indices. In this paper, association rules are used in an extra-ordinary way to study fluctuation patterns of the market.

There are many previous studies done mostly in recent decade that focus on analysis and prediction of financial markets with complex networks. The most cited research that was done in 2003, was about studying New York Stock

Exchange (NYSE) with the help of cross-correlation analysis to make networks of stock prices time series [1]. The researchers found minimal spanning trees of the networks helpful for a better understanding of the stock market topology. Complex networks analysis is a new tool for understanding many different aspects of financial markets that could not be fully understood before and makes an important role in new studies of financial markets. The importance of this new approach is discussed in a research by Gatti et al., [2].

Various applications of network science in finance were presented by Caldarelli et al., [3]. They showed some applications of graph theory methods that could be useful in finance and economy. With the spread of using complex networks in finance and economy, many studies tried to focus on different aspects of markets for building different kinds of networks. The researchers found out a better knowledge of the financial markets in many aspects by studying these networks. Some studies are presented in this section as examples of how the researches made networks out of financial data and what they found out with the help of networks science.

In many studies, network description of large markets were proposed. As an example, in a study by Garlaschelli et al., [4] network was made where stocks and shareholders were vertices and the edges of the network were weighted and corresponded to shareholdings.

In another study by Naylor et al., [5] two hierarchical methods were used, to develop a topological influence map for some currencies from a distance matrix. They used minimal spanning trees and ultra-metric hierarchical trees to understand the topology of complex networks for foreign exchange market and discussed the scale-free structures found out in the networks made. Correlation matrices of stock returns over time in NYSE were analyzed using spectral and network methods in another research by Heimo et al., [6]. In a study of Hank Seng stock market of Hong Kong, researchers extracted the hidden fluctuation patterns of the stock index from a directed network topology [7]. A review of the literature on small-world networks used in management and social science was done by Uzzi et al., [8] where they showed different interdisciplinary applications of small-world networks as previously discussed by Milgram in other fields of science [9]. In another research, researchers presented a reliable procedure to build networks from correlation matrix of different time series [10]. They used the correlation between time series to

Arash Negahdari Kia, Ph.D. Scholar, Faculty of New Sciences and Technologies, University of Tehran, Iran, E-mail:nkia.arash@ut.ac.ir

Dr. SamanHaratizadeh, Assistant Professor, Faculty of New Sciences and Technologies, University of Tehran, Iran, E-mail:haratizadeh@ut.ac.ir

Dr. ZainaboldodaHeshmati, Assistant Professor, Faculty of New Sciences and Technologies, University of Tehran, Iran, E-mail:zheshmati@ut.ac.ir

DOI: 10.9756/BIJIEMS.10437

build adjacency matrix based on different thresholds. Other researchers used transfer entropy to show direction and strength of information flow between stock indices time series [11]. It was a new way of building directed networks from time series.

Using statistical correlation between time series can only be used to make undirected networks. In this study, a new method of building directed networks has been used. Association rules mining in a dataset of different time series is used to find rules for the fluctuation patterns. These rules are then used for building a directed network for fluctuation analysis and prediction.

Yang et al., [12] did a research to investigate relations between six exchange rate time series by means of a visibility graph. A visibility graph is a network made by all of time series elements as nodes. The edges between any two nodes represents the visibility of two elements from each other in the time series plot. By this means, they constructed a network for every financial time series and analyzed the power law degree distribution and scale-free topology of exchange rate time series. Another study was done in Chinese stock market by Huang et al., [13], which analyzed the scale-freeness of the networks and centrality measurements of the nodes and cliques in the topology.

Another study which used complex networks to analyze correlations between prices of all US stock markets, used stocks as nodes [14]. Another research tried to solve the problem of reconstructing the tree-like structure of a network for the linear dynamic systems in financial markets [15]. The researchers used a distance function to calculate the closeness between the processes. Zhang et al., [16] did a study to analyze the time series of Shanghai stock index with the use of complex network theory. They showed that the network of the main series is fitted with a power law, and the network extracted from the return series is fitted by an exponential curve. The Brazilian stock market sectors were investigated by building the minimum spanning tree out of the correlation network in a study by Tabak et al., [17]. They showed that energy, finance and material sectors of the market were the most important sectors in the Brazilian market.

An important part of the study in this paper has used the sector indices time series of Tehran stock market for building networks which helps to understand the relationship between different sectors and helps to find the most important sectors of the market.

In another research, the complex network of stock trading data among investors of Shenzhen development bank was investigated [18]. The nodes were stock traders and the links presented the trade with the weight as the volume of it. The researchers made a network for each trading day and showed that networks present a power law degree distribution.

Random matrix theory has also been used in the study of financial markets [19]. The researchers have analyzed the clustering coefficients and component numbers of the network using random matrix theory. The data of Dow Jones Industrial Average (DJIA) and Tehran Stock Exchange (TSE) were used in their study. Another study established networks of cross-

shareholding for some companies in China in a period of time and analyzed the networks [20]. The researchers studied the topology of the cross-shareholding networks in an eight year period and discussed the differences of the networks before and after financial crisis in 2008. In another research a full transaction records of more than hundred stocks were used to build trading networks where nodes represented the investors and links connected sellers to buyers [21]. The research showed that degree distribution of these networks obeyed the power law and manipulated stocks could be distinguished from non-manipulated ones by a high lower band of the power law tail and high average degree. Directed network of ten important world's financial markets were made by use of partially directed coherence formula in a work of Allali et al., [22]. Partially directed coherence was used before in neuroscience studies to find out the causality between different processes.

Visibility graph of four macro-economic time series in China were studied in another study [23]. Similar to the previous study discussed about the visibility graphs, the researchers tried to capture new features from the time series and study the differences of the network structure before and after some economic policies of the China's government. They studied the small-world effect in the visibility graph of these time series. The relation between the variations of the structure and fluctuations of the Shanghai stock market was studied by Chunxia et al., [24]. The researchers used a moving window to scan through the stock prices time series for a period of time. The study of stock market returns in Europe was also done with the help of the complex networks [25]. The properties of the networks were extracted by the means of centrality measurement parameters. The correlation network between two important stock markets of Korea was studied and the two stock markets network topology were compared in a work by Ko et al., [26]. Importance of network science in the economy and financial market study was emphasized in the research of Farmer et al., [27]. The researchers discussed the importance of interdisciplinary studies in complex networks, economy, and finance and the effects of these studies on science, technology and society.

In another study, sudden changes in direction of stock market index was studied by making a new indicator using wavelet decomposition of the time series and using the correlation network with stocks as nodes and correlations as links [28]. The researchers showed that a combinational method of eigenvalues of adjacency matrix and their indicators could be used to find out the points where the stock index changes direction. Another research was done on 93 different stock markets by making the correlation network between them before and after the collapse of Lehman-Brothers [29]. The research studied the minimal spanning tree of the networks and analyzed them with an index called turbulence of the market that was calculated by eigenvector centrality measurement of the nodes in the networks. In another study, the correlation networks of different stock markets across the world were studied, before and after the financial crisis of 2008 [30]. The study was to find out which markets were more important in the networks by the help of centrality measurements. Association rules and K-means

clustering were used to make a good portfolio of stocks between China, Taiwan and Hong Kong stock markets [31]. The research showed the importance of using association rules in finding the relationship between different stocks. In another study, the researchers divided China into different regions with respect to GDP and constructed a correlation network [32]. They showed that the location and distance of the regions to each other plays an important role in connection between their nodes in the correlation network.

The prediction of fluctuations using networks was studied in a research, using a semi-supervised graph-based algorithm on a network of different stock markets, exchange rates, oil prices and some other financial time series [33]. The researchers claimed their method could see the interactions and cyclic effects of markets on each other. In our paper we try to build the network in another way and use our own algorithm based on the paths made between the markets in the network.

A research on relationship between credit market and stock market before and after subprime crisis was done by Lim et al., [34]. They used the network topology and random matrix theory and compared the eigenvalue of the network matrices. It was found out that the eigenvalue of credit market became bigger than the one for stock markets right before the crisis happened. After the crisis, the correlation between two markets became stronger than before. Another research tried to construct networks of important enterprises and firms in Euro region for both local regions and the whole Euro region [35]. The researches constructed directed networks and analyzed the changes of their topology in the crisis time period. Co-integration coefficient was used to make directed network of stock markets before and after financial crisis and collapse of Lehman-Brothers in a study by Yang et al., [36]. They found out that after the crisis, the impact of US stock market on other markets had reduced and China's markets impact had increased.

Constructing a network of market sectors by cross-correlation was done by Mai et al., [37] and it was shown that the degree distribution obeyed the power law with a little exponent. The researchers showed that the Industry sector had more impact on other sectors with the help of the built network. In our research we use the TSE data and correlation-based and association rules based networks to show important sectors in Tehran market.

In a study of fluctuations using network structure, researchers tried to investigate the characteristics of an exchange rate time series fluctuations with the methods of symbolization and coarse gaining [38]. They constructed the complex networks of relations between five foreign currencies with transforming methods. Prediction of the stock index movements by ten different data mining methods was studied in another work [39]. In a study by the authors of this research, the K-Nearest Neighbor method combined with Dynamic Time Warping distance functions was used for prediction of fluctuations in financial time series [40]. In this paper the authors try to solve this problem with a different approach and gain better results in some of the cases. The network structure used in this study can use the proper

external factors for prediction of the time series by finding relevant paths from external time series to the time series which is going to be predicted. In another research direct and recursive prediction models are discussed using mutual information for feature selection [41]. The study shows direct prediction superiority compared to recursive prediction. Our study also uses a direct approach for predicting fluctuations.

In the next section of the paper, we introduce our methodology of extracting directed and undirected networks out of stock exchange sector indices data. We analyze the networks by centrality measurement parameters and show the most important sectors that have more impact on others in TSE. Then we introduce a way of predicting fluctuations of the sectors stock indices by finding the paths in the directed network extracted from the association rules by Apriori algorithm. Section three is about the data gathering and preparation phase of this research. Section four talks about the evaluation methodology used to test our approach of predicting fluctuations of the indices time series. In section five the results are presented and discussed in detail and the last section concludes with some suggestions for further researches. The abbreviations of the sector names in TSE market that we used are presented in a table in appendix at the end of the paper.

II. METHODOLOGY

The methodology used in this paper consists of extracting networks of stock sector price indices out of TSE dataset, in order to find the important sectors from the networks in different aspects. It also provides a way to predict rising fluctuations of the price indices over time from the directed networks made by association rules and Apriori algorithm. At the end of this section a diagram of the whole research methodology is presented and explained.

A. Network Extraction Methods

In this study, first a correlation matrix between the sector indices time series is extracted as the adjacency matrix of the networks. By this method, and choosing a threshold it can be assumed that there is a link between sectors that have cross correlations bigger than the threshold. Two different undirected networks with two different thresholds for the correlation values between the sector indices data of TSE are extracted as an example to show how the proposed methods work in the TSE market. The correlation formula is shown in (1)[29].

$$r_{xy} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2 \sum_{i=1}^n (y_i - \bar{y})^2}} \quad (1)$$

In the correlation equation x and y are the two different sector indices time series and r is the correlation between them. The correlation, r , is a digit between -1 and +1. Two different thresholds of 0.7 and 0.9 have been used to check if r is greater than the thresholds or not. Any other thresholds could be used to make different networks. In this study using these thresholds caused more than half of the sectors appear in the networks (In the threshold of 0.7, 36 out of 38 sectors

appeared in the network). Due to one of the aims of this study to find out important sectors with more positive correlations with other sectors, the disappearance of some sectors in the network did not affect the results. It is obvious that any other threshold cuts could be used to make other networks. This study had to choose some threshold cuts to show numerical results of its methodology used in practice. It is clear that by using the threshold of -1 a complete graph is constructed and no knowledge of the market can be extracted out of it (The case of using a complete weighted correlation graph is different and is discussed in other researches explained in the introduction of this study). If the value of r is more than the threshold, then the corresponding element in the adjacency matrix of the network will be 1 otherwise 0. This means there will be a link between the nodes of two stock sector indices in the network. As mentioned, with lower thresholds the network will have more links and a link between two nodes shows positive correlation between the indices higher than the threshold. In other word, when a sector index rises, the other sector indices linked to it in the network will rise with more

probability. Once the two different correlation networks corresponding to two thresholds have been extracted, various centrality measurements are calculated on the networks in order to find the most important sectors in the TSE Market.

In the other phase of the research, a directed network is built out of the stock sector indices with the help of association rules and Apriori algorithm. First, each series are converted to binary series that only consists of zeros and ones. A rise in the value of the stock sector index from the previous value in the time series is represented by a 1, and a 0 indicates a fall in the value. After converting all the indices time series into these new series that show fluctuations, all these series are put as the columns of a new matrix. Each column of this matrix is a binary fluctuation time series of a different sector index and each row is a day in the stock market. Figure 1 presents a sample of converting the data for using in Apriori algorithm as described before for a 4 day period of time. It is obvious that after the conversion the new dataset will have one day less than the original dataset.

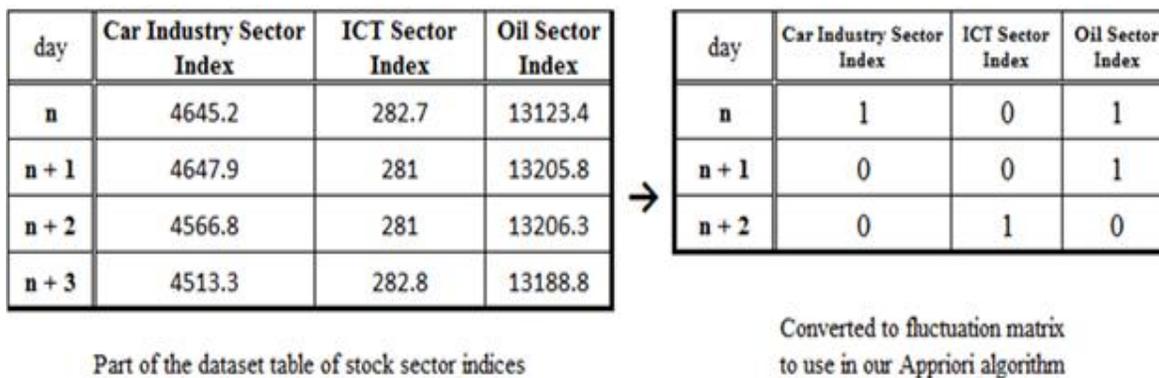


Figure 1: Converting the Dataset for Using in Apriori Algorithm

By using an algorithm of finding association rules in this matrix and looking at each row like a buying transaction in a store, all the couple indices that rise together are found. Rules like $A \rightarrow B$ that indicate if sector A , rises with a specific confidence and support value, the sector index B also rises are extracted. The Apriori algorithm is used to find out rules like this in the TSE market data. A brief description of the Apriori algorithm that is used in this research is presented in section 2.C.

Every association rule like $A \rightarrow B$ in the converted matrix means a directed link between the nodes A and B . Two different confidence values of 0.60 and 0.65 and a support of 0.1 are used to make two different directed networks for the TSE market. Again it should be explained that any other confidence and support can be used to make different directed networks. Using these thresholds for confidence and support made two networks with 36% (for 0.65 confidence level) and 84% (for 0.60 confidence level) of the sectors appear in the directed network. Using lesser confidence levels reveals lesser knowledge from the topology of the network and by using a confidence of zero a complete graph would have appeared. The centrality measures are calculated again for the two new networks and are also presented and discussed later.

B. Market Sector Analysis

Finding the most important sectors of the TSE market is done by centrality measure analysis of the nodes in the extracted networks. Centrality measurement is a good tool to find out which nodes are more important in the network [42]. In this work different centrality measures are used that are calculated for all the nodes of the networks that correspond to the market sectors.

Degree centrality calculates all the links that start or end from a node. If the links coming to a node are counted, it is called in-degree and if the links going out of the node are counted, it is called out-degree. The closeness centrality measures the mean distance from a node to other nodes. Another centrality measure used in this study is called betweenness. It measures the extent to which a node lies on paths between other couple of nodes. Eigenvector centrality and page rank are extensions to degree centrality. Not all the nodes have the same importance and having link to some nodes are more important than others. Eigenvector centrality and page rank are increased if having links to other important nodes [42], [43].

Network constraint is another measure that shows the extent to which a node links to other nodes that are already linked to each other. Betweenness and network constraint,

both try to find bridges in the network topology. Lower network constraint and higher betweenness indicate bridging[42], [43]. This means that in the results section, the nodes with the lowest network constraint or with the highest betweenness or degree are presented as important nodes (sectors in our case).

C. Association Rules and Apriori Algorithm

Association rules, present the relationship between different item sets in terms of occurrence. This means, if some items appear in a transaction (a record, or a row of matrix in our case), it can be assumed that some other items will also appear in the same transaction. Apriori is the name of a famous algorithm to find association rules in a dataset (The fluctuation matrix that was described in section 2.A in our case).

Apriori algorithm was presented by Agrawal and Srikant in 1994. They provided an algorithm for finding association rules in large database of sales transactions. The name of the algorithm comes from the fact that it uses prior knowledge to find frequent item sets in the database. Any given sequence of the items in the database is called an item set. The algorithm creates some candidate item sets with k items, which is shown as C_k . Those candidate itemsets that have repeated more than a proportion called support are frequent item sets. Frequent item sets with the length k are shown as L_k [44].

In the first step of Apriori algorithm, the C_1 item sets are gathered and L_1 is obtained from all C_1 item sets which are frequent (are repeated more than a predetermined number called support). Other steps of the algorithm are presented below:

- L_{k-1} is obtained.
- C_k is obtained from Cartesian product of $L_{k-1} \times L_{k-1}$.
- All C_k that have sub-itemsets which are not frequent can not be frequent themselves.
- L_k is obtained.

In this paper, after finding the L_2 itemsets, our work is finished with the Apriori algorithm. This is due to the nature of the rules that are going to be used in the proposed network extraction method (As mentioned before, rules like $A \rightarrow B$ are used that only consist of two items A and B). After finding L_2 item sets, all of $A \rightarrow B$ rules that have confidence level more

than a predetermined threshold are extracted to use in directed network building process. The confidence and support in the Apriori algorithm are defined in (2) and (3).

$$\text{Support}(A) = \text{proportion of itemset } A \text{ in dataset} \quad (2)$$

$$\text{Confidence}(A \rightarrow B) = \frac{\text{Support}(A \cup B)}{\text{Support}(A)} \quad (3)$$

D. Fluctuation Prediction with Directed Apriori Networks

After making the adjacency matrix of directed networks (having one in element (i, j) of the matrix if there is a rule $i \rightarrow j$, otherwise zero), the paths of length n can be found, by multiplying the matrix to itself, n times. If the adjacency matrix is called A , the element (i, j) of matrix A^n presents how many paths of length n are from i , to j .

As discussed before in section 2.A, a directed link from node i to node j shows that by a confidence threshold, if the market sector index i rises, the market sector index j rises in the same day (day or any other time unit that is used). Now if there is a directed path of length two, from i to j , and then from j to k , the fluctuation of i propagates to k through the node j . Suppose this hypothesis is going to be tested that the probability of market sector index k rising in the next day would be bigger than fifty percent, when there is a path of length two, from i to k . The reason of having this hypothesis is that, without any prior knowledge there is equal chance for rising or falling of the market sector index of k . But with the prior knowledge of having a path from i to j and j to k , it can be assumed that k rising in the next day may be more probable. The evaluation process in this study shows the truth of this hypothesis in case of directed sector index market of TSE. For paths of length more than two, the test methodology and evaluation process is extended and the results are also presented.

E. The General Schema of the Proposed Methodology

The overall research process diagram is presented in figure 2. All the boxes (sub-processes) have been described in detail in this paper. The boxes (sub-processes) in the diagram are labeled and also a brief description is given in this section to give a general idea of the whole process. These sub-processes are described in details in their corresponding section.

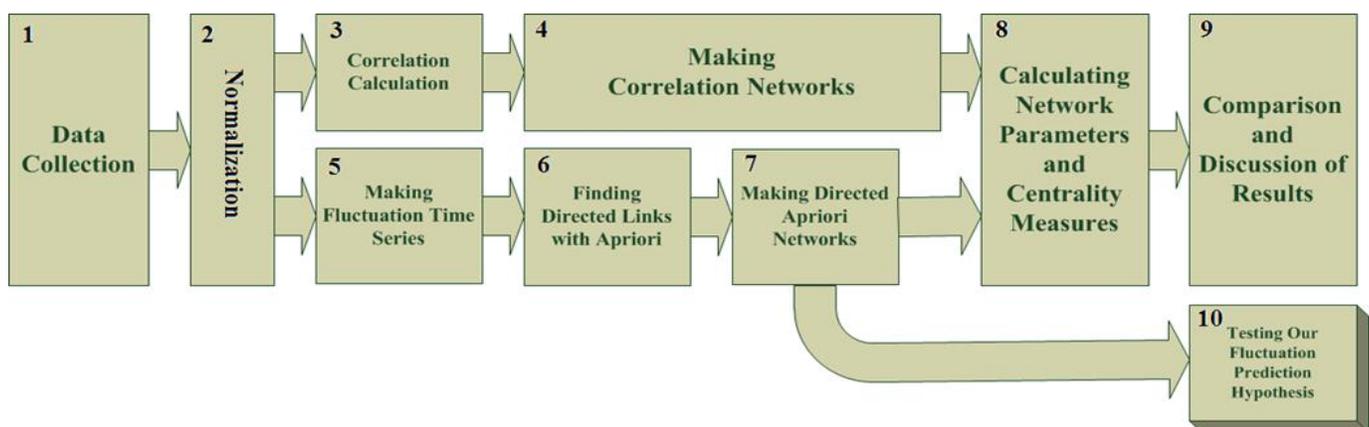


Figure 2: General Schema of the Proposed Methodology

Here is a brief description of the sub-processes in order of their box number:

1. Historical data collection for stock market sectors (described in section 3).
2. Normalizing the data (described in section 3).
3. Calculating the correlation between stock sector indices (described in section 2.A).
4. Extracting correlation networks from connecting the sectors with correlation higher than a threshold (described in section 2.A).
5. Converting the stock sector indices into fluctuation time series (described in section 2.A and in figure 1).
6. Applying Apriori algorithm on the rows of the fluctuation series matrix (having each day as a transaction like figure 1) for finding rules like $A \rightarrow B$ (described in section 2.C).
7. Extracting directed networks by connecting items (sectors) of left and right side of the rules (described in section 2.A).
8. Calculating network parameters and centrality measures for market sectors (Nodes of the Networks) for each network and sorting the results to find the most important nodes or sectors in different aspects in the market (described in section 2.B and results presented in section 5).
9. Comparing and discussing centrality measurement of market sectors for correlation and Apriori networks (presented in section 5).
10. Testing fluctuation prediction for n days later of market sector indices with finding paths of length $n+1$ in the directed Apriori network (described in section 4 and results presented in section 5). In this phase, the hypothesis of ability to predict fluctuations with directed network is tested in three different ways:
 - Using all the paths of length n between two nodes as the input of testing algorithm to evaluate our hypothesis.
 - Using all the paths of length n between two nodes apart from those that there is also a path of length one between the source and destination nodes.
 - Using 80 percent of the data to build an Apriori network, and using the other 20 percent for testing the hypothesis.

The evaluation process or box number 10 is explained in more details in section 4.

III. DATA PREPARATION

Data for market sector price indices are gathered from the official website of TSE¹. Price indices for every industrial, commercial, or service sector are calculated every day in TSE. Data gathered are for the period of 11th October 2009 to 11th May 2013. Considering the missing data (national holidays), 860 daily stock sector index data were gathered for 38 market sectors out of 43 market sectors. Five market sectors did not have enough data in the official site of TSE in the specified

time period.

One can consider the dataset as a matrix with 860 rows for each day that data was available and 38 columns for each market sector index. Because of the different value region of the indices, this dataset was normalized with (4).

$$Normalized = \frac{OriginalValue - Min(Index)}{Max(Index) - Min(Index)} \quad (4)$$

Due to the long names of market sectors in TSE, abbreviations were used for each of the official names of the sectors. These abbreviations are presented in appendix section.

IV. EVALUATION METHOD

A conclusion section is not required. Although a conclusion may review the main points of the paper, do not replicate the abstract as the conclusion. A conclusion might elaborate on the importance of the work or suggest applications and extensions. Include future enhancement.

Considering the directed networks made by the Apriori algorithm, a hypothesis is proposed along with an evaluation method to test it. The evaluation method is described as below:

- Hypothesis: If there is a path of length n between nodes A and B, in the directed networks, this means that if the value of index A rises, the value of index B rises $n-1$ days later. It is clear that $n > 1$.
- Evaluation Method Steps:
 - 1) Two variables of T (True) and F (False) are defined and set to zero. These variables count the number of true and false predictions of the fluctuation by our hypothesis in the directed network.
 - 2) All the directed paths of length n in the network are found. For this purpose, the adjacency matrix A will be multiplied to itself, n times. Any matrix element that is 1 or more than 1 shows there is 1 or more than 1 path of length n between the two nodes.
 - 3) Wherever there is a path of length n between A to B, the fluctuation dataset matrix (figure 1) receives our attention. Anytime corresponding column for A has the value 1 (meaning that A is rising), column B, $n-1$ rows ahead value is monitored. If this value is also 1, the variable T is increased one unit for a true prediction of fluctuation. If the value of the $n-1$ rows ahead of column B is 0 (meaning that despite the value of A that was rising $n-1$ days before, the value of B does not rise), the variable F is increased one unit as a false prediction.
 - 4) Finally, for each path of length n from any node A to the any node B, the hypothesis is tested, and the results of the tests for every path will be put in a test vector (the vector length will be the number of paths with length n). If more than 50 percent of the results in the test vector are 1, this means that directed network and our hypothesis

¹Tehran Stock Exchange Official Website, URL: <http://www.tse.ir/market/Shakhes.aspx>, [last access date: 2015-12]

for the fluctuation prediction can be used and is better than using a random classifier. From the test vector, the value of $T / (T + F)$ is calculated. This value presents the recall of the prediction model. Recall formula is presented in (5).

$$\text{Results from the test vector for each path length} = T / (T + F) = \text{Recall of the proposed prediction model} = \frac{\text{TruePositive}}{\text{TruePositive} + \text{FalseNegative}} \quad (5)$$

V. RESULTS AND DISCUSSION

In this study, two open-source applications are used. Graphviz (Graph Visualization Software) is used to depict the networks, and SNAP (Stanford Network Analysis Platform) is used to calculate various network parameters such as centrality and clustering. Correlation networks were extracted

with two different correlation thresholds of 0.7 and 0.9. The network parameters of correlation networks are presented in table 1.

Table 1: Network Parameters for Correlation Networks made by two Different Thresholds

Model	Number of Nodes	Number of Links
Threshold = 0.7	36	251
Threshold = 0.9	21	41

As it can be seen in table 1, with reducing the correlation threshold, more nodes (market sectors) will have the privilege of being in the network and the number of the links increases. It is obvious that with a threshold of -1, the network becomes a complete graph. Figure 3 and 4 present the visualization of the correlation networks with thresholds of 0.7 and 0.9.



Figure 3: Correlation Network of Tehran Stock Exchange Market Sectors with 70% Threshold

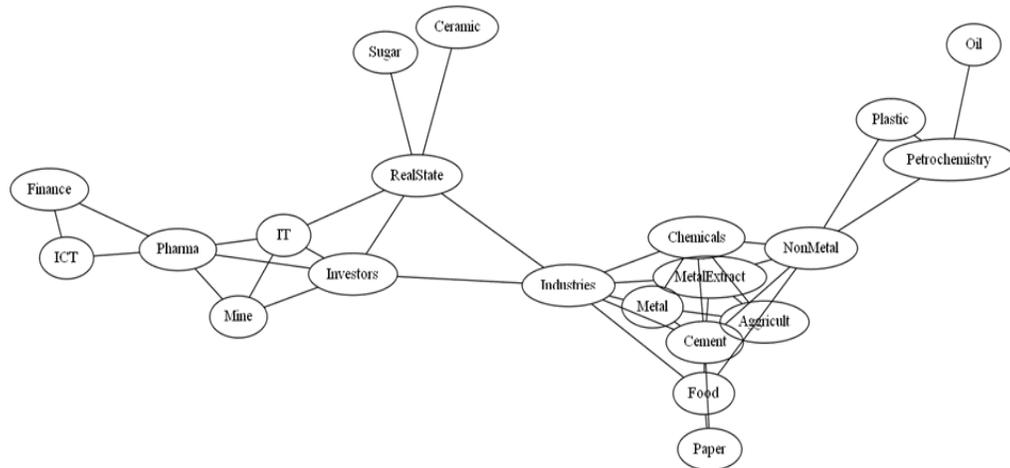


Figure 4: Correlation Network of Tehran Stock Exchange Market Sectors with 90% Threshold

As explained before, the Apriori networks are extracted with support level of 0.10 and confidence levels of 0.60 and 0.65. The Apriori network parameters are presented in table 2 and the visualization of the networks are shown in figures 5 and 6.

Table 2: Network Parameters for Apriori Networks made by Two Different Confidence Levels

Model	Number of Nodes	Number of Links	Nodes with Zero In-Degree	Nodes with Zero Out-Degree	Bi-Directional Links
Confidence ≥ 0.60	32	189	14	1	80
Confidence ≥ 0.65	14	46	3	1	18

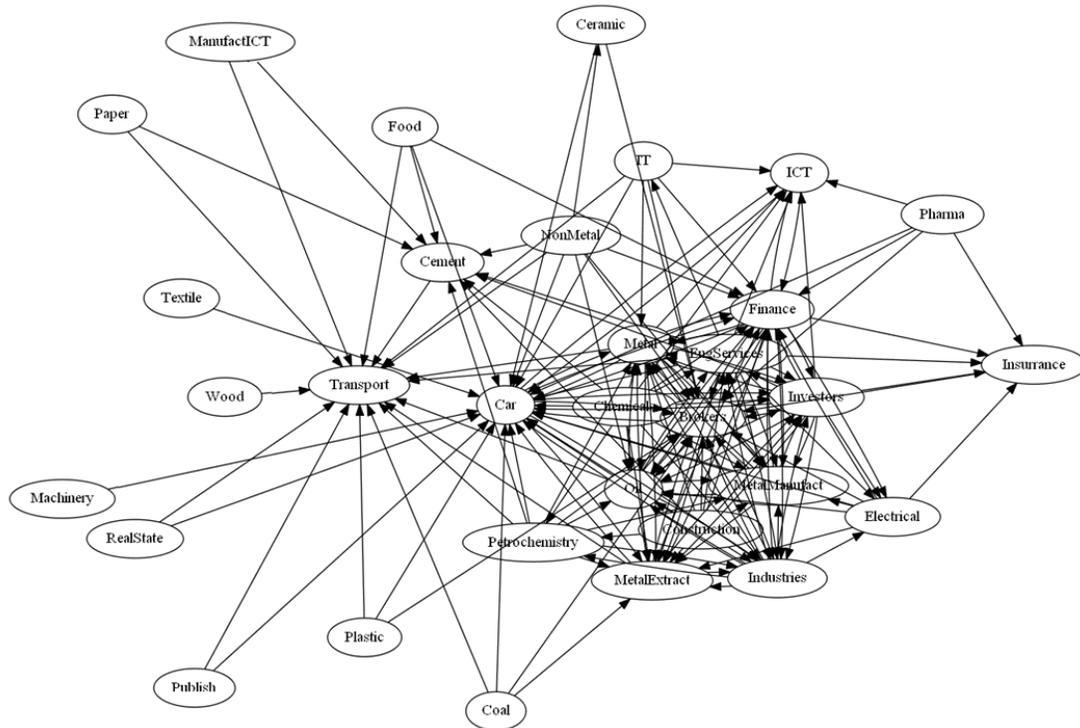


Figure 5: Apriori Network of Tehran Stock Exchange Market Sectors with Confidence Level of 60% or Higher

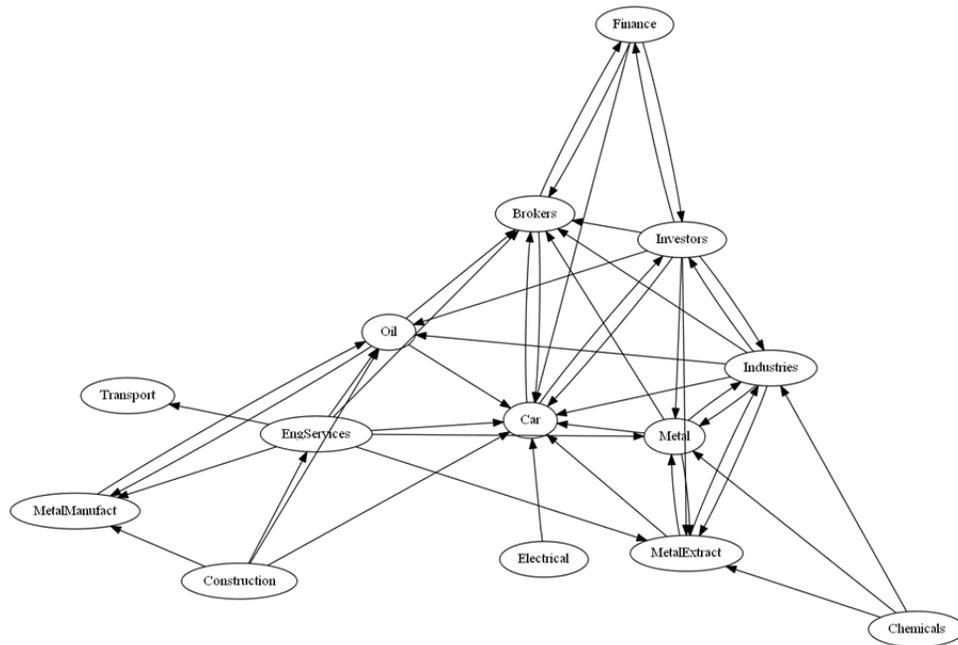


Figure 6: Apriori Network of Tehran Stock Exchange Market Sectors with Confidence level of 65% or Higher

In the networks extracted with Apriori algorithm, the more the confidence level is raised, the less are the links and nodes. This is due to extracting less association rules with higher confidences. The results of centrality measurement data sorted and calculated for two correlation networks and two Apriori networks are presented in table 3. These data are extracted after sorting the nodes according their importance in the

centrality and the most important node (market sector) is presented in table 3. The concept of importance in different centrality measurement methods has been discussed in section 2.B. The results are presented in abbreviation form and the full name of the market sectors in TSE are presented in the appendix section.

Table 3: Important Sectors Considering Network Centrality Measures

Model	Degree	Closeness	Betweenness	Eigen Vector	Network Constraint	Clustering Coefficient	Page Rank
Correlation network with threshold = 0.7	Industries	Industries	Finance	Industries	RealState	Car-Electrical Medical	Petrochemistry
Correlation network with threshold = 0.9	Industries	Industries	Industries	Cement	RealState	IT	Metal
Apriori network Confidence ≥ 0.60	Car	Car	Car	Car	Car	ManufactICT Paper-Ceramic	EngServices
Apriori network Confidence ≥ 0.65	Car	Car	Car	Car	Car	Industries	Investors

By looking at the table 3 it can be seen that in correlation networks the industrial companies (see appendix section) sector shows more correlation with other sector indices. This results that if some could only see one sector to analyze the fluctuations of the Tehran Stock Exchange Index (TEPIX), the industrial companies' sector index showed a better behavior of the TEPIX than other sector indices.

The results for the models made by Apriori algorithm on fluctuation time series presents a fact about the *vehicle and parts manufacturing* (abbreviated as car in our tables and figures) sector. This sector is affected by many of other sectors, meaning where they rise in value, the vehicle and

parts manufacturing rises also in the same day. From Apriori networks the paths can also be found out from a sector to another sector to understand the effects of a rise in a sector price index to other indices.

In table 4, the results of testing the fluctuation prediction hypothesis are presented. The values in the brackets show the second type of testing by eliminating all paths from A to B that also had a single link from A to B. In table 4 results show a high success except for paths of length three for networks of confidence level 60% (prediction of two days later fluctuation).

Table 4: Results of Testing the Prediction Hypothesis for all the Dataset With and Without Eliminating Paths that have a Single Link between Two Nodes. (The Results which come from Elimination are in Brackets)

Model	Number of Tests for Paths with length 2	Percentage of Successful Predictions	Number of Tests for Paths with length 3	Percentage of Successful Predictions	Number of Tests for Paths with length 4	Percentage of Successful Predictions	Number of Tests for Paths with length 5	Percentage of Successful Predictions
Apriori Network Confidence ≥ 0.65	85 (46)	98.8% (97.8%)	110 (66)	50% (48.5%)	117 (73)	60.4% (58.9%)	117 (73)	75.2% (71.2%)
Apriori Network Confidence ≥ 0.60	359 (188)	89.8% (87.7%)	448 (266)	45.8% (49.6%)	459 (277)	60.4% (51.3%)	459 (277)	61.2% (56.3%)

VI. CONCLUSION AND FURTHER RESEARCHES

In this study concepts of network science have been used to come up with a clearer understanding of the TSE market and relationship between its industrial and service sectors. The researchers also presented a network model using Apriori algorithm to predict the fluctuations of sector markets. It is found out that the industrial companies sector has the highest correlation with other sectors and the correlational structure of the TSE was extracted. By directed networks extracted from association rules, it became clear that the vehicle and parts manufacturing sector was the most influenced sector by other

sectors price fluctuation in Tehran Stock Market.

Finally, the fluctuation prediction model was tested and it was shown that this model can be used with directed Apriori networks built upon high confidences to predict the fluctuations and it works better on short-time predictions.

For fluctuation prediction out of the fluctuation dataset, other association rule mining algorithms and sequential pattern mining algorithms can be studied to see how they result. Also, other datasets of other markets can be the focus of future researches in predicting and analyzing the financial markets with our methodology of network construction.

APPENDIX: TABLE OF ABBREVIATIONS USED FOR THE MARKET SECTOR NAMES IN THE FIGURES AND TABLES

Full Name of the Sector	Insurances except Social Security	Oil and Gas Extraction	Tele-Communication	Tiles and Ceramics	Non-Metal Mining Products	Food Industry except Sugar	Financial Brokerages	Sugar	Pharmacologic Products	Leather and Shoes
Abbreviation	Insurance	Oil	ICT	Ceramic	Non Metal	Food	Brokers	Sugar	Pharma	Leather
Full Name of the Sector	Chemical Products	Real State	Multi Industries	Technological and Engineering Services	Coal Extraction	Transportation and Ware housing	Plastic	Information Technology	Medical Products	Agriculture
Abbreviation	Chemicals	Real State	Industries	EngServices	Coal	Transport	Plastic	IT	Medical	Agricult
Full Name of the Sector	Furniture	Construction	Metal Extraction	Financial Brokers	Textile	Tele-Communication Products	Mining	Paper Products	Car Industry	Machinery
Abbreviation	Furniture	Construction	MetalExtract	Finance	Textile	ManufactICT	Mine	Paper	Car	Machinery
Full Name of the Sector	Metal Products Manufacturing	Wood Products	Basic Metals	Publishing	Petro chemistry Products	Electrical Machines and Products				
Abbreviation	MetalManufact	Wood	Metal	Publish	Petrochemistry	Electrical				

REFERENCE

- [1] G. Bonanno, G. Caldarelli, F. Lillo, and R. N. Mantegna, "Topology of correlation-based minimal spanning trees in real and model markets," *Phys. Rev. E*, Vol. 68, no. 4, p. 46130, 2003.
- [2] D. D. Gatti, E. Gaffeo, and M. Gallegati, "Complex agent-based macroeconomics: a manifesto for a new paradigm," *J. Econ. Interact. Coord.*, Vol. 5, No. 2, Pp. 111–135, 2010.
- [3] G. Caldarelli, S. Battiston, D. Garlaschelli, and M. Catanzaro, "Emergence of Complexity in Financial Networks," *Lect. Notes Phys.*, Vol. 650, Pp. 399–423, 2004.
- [4] D. Garlaschelli, S. Battiston, M. Castri, V. D. P. Servidio, and G. Caldarelli, "The scale-free topology of market investments," *Phys. A Stat. Mech. its Appl.*, Vol. 350, No. 2–4, Pp. 491–499, 2005.
- [5] M. J. Naylor, L. C. Rose, and B. J. Moyle, "Topology of foreign exchange markets using hierarchical structure methods," *Phys. A Stat. Mech. its Appl.*, Vol. 382, No. 1, Pp. 199–208, 2007.
- [6] T. Heimo, J. Saramäki, J.-P. Onnela, and K. Kaski, "Spectral and network methods in the analysis of correlation matrices of stock returns," *Phys. A Stat. Mech. its Appl.*, Vol. 383, No. 1, Pp. 147–151, 2007.
- [7] P. Li and B.-H. Wang, "Extracting hidden fluctuation patterns of Hang Seng stock index from network topologies," *Phys. A Stat. Mech. its Appl.*, Vol. 378, No. 2, Pp. 519–526, 2007.
- [8] B. Uzzi, L. a. N. N. Amaral, and F. Reed-Tsochas, "Small-World Networks and Management Science Research: A Review," *Eur. Manag. Rev.*, Vol. 4, No. 2, Pp. 77–91, 2007.
- [9] S. Milgram, "The small world problem," *Psychol. Today*, Vol. 1, No. 1, Pp. 61–67, 1967.
- [10] Y. Yang and H. Yang, "Complex network-based time series analysis," *Phys. A Stat. Mech. its Appl.*, Vol. 387, No. 5–6, Pp. 1381–1386, 2008.
- [11] O. Kwon and J.-S. Yang, "Information flow between stock indices," vol. 1, no. 1, 2008.
- [12] Y. Yang, J. Wang, H. Yang, and J. Mang, "Visibility graph approach to exchange rate series," *Phys. A Stat. Mech. its Appl.*, Vol. 388, No. 20, Pp. 4431–4437, 2009.
- [13] W.-Q. Huang, X.-T. Zhuang, and S. Yao, "A network analysis of the Chinese stock market," *Phys. A Stat. Mech. its Appl.*, Vol. 388, No. 14, Pp. 2956–2964, 2009.
- [14] C. K. Tse, J. Liu, and F. C. M. Lau, "A network perspective of the stock market," *J. Empir. Financ.*, Vol. 17, No. 4, Pp. 659–667, 2010.
- [15] D. Materassi and G. Innocenti, "Unveiling the connectivity structure of financial networks via high-frequency analysis," *Phys. A Stat. Mech. its Appl.*, Vol. 388, No. 18, Pp. 3866–3878, 2009.
- [16] J. Zhang, H. Zhou, L. Jiang, and Y. Wang, "Network topologies of Shanghai stock index," *Phys. Procedia*, Vol. 3, No. 5, Pp. 1733–1740, 2010.
- [17] B. M. Tabak, T. R. Serra, and D. O. Cajueiro, "Topological properties of stock market networks: The case of Brazil," *Phys. A Stat. Mech. its Appl.*, Vol. 389, No. 16, Pp. 3240–3249, 2010.
- [18] Z.-Q. Jiang and W.-X. Zhou, "Complex stock trading network among investors," *Phys. A Stat. Mech. its Appl.*, Vol. 389, No. 21, Pp. 4929–4941, 2010.
- [19] a. Namaki, a. H. Shirazi, R. Raei, and G. R. Jafari, "Network analysis of a financial market based on genuine correlation and threshold method," *Phys. A Stat. Mech. its Appl.*, Vol. 390, No. 21–22, Pp. 3835–3841, 2011.
- [20] Y. Ma, X. Zhuang, and L. Li, "Research on the relationships of the domestic mutual investment of China based on the cross-shareholding networks of the listed companies," *Phys. A Stat. Mech. its Appl.*, Vol. 390, No. 4, Pp. 749–759, 2011.
- [21] X.-Q. Sun, X.-Q. Cheng, H.-W. Shen, and Z.-Y. Wang, "Distinguishing manipulated stocks via trading network analysis," *Phys. A Stat. Mech. its Appl.*, Vol. 390, No. 20, Pp. 3427–3434, 2011.
- [22] A. Allali, A. Oueslati, and A. Trabelsi, "Detection of Information Flow in Major International Financial Markets by Interactivity Network Analysis," *Asia-Pacific Financ. Mark.*, Vol. 18, No. 3, Pp. 319–344, 2011.
- [23] N. Wang, D. Li, and Q. Wang, "Visibility graph analysis on quarterly macroeconomic series of China based on complex network theory," *Phys. A Stat. Mech. its Appl.*, Vol. 391, No. 24, Pp. 6543–6555, 2012.
- [24] Y. Chunxia, X. Bingying, H. Sen, and W. Rui, "A study of the interplay between the structure variation and fluctuations of the Shanghai stock market," *Phys. A Stat. Mech. its Appl.*, Vol. 391, No. 11, Pp. 3198–3205, 2012.
- [25] P. Caraiiani, "Characterizing emerging European stock markets through complex networks: From local properties to self-similar characteristics," *Phys. A Stat. Mech. its Appl.*, Vol. 391, No. 13, Pp. 3629–3637, 2012.
- [26] J. S. Ko, G. Lim, and K. Kim, "Analyses of the structure of group correlations in Korean financial markets," *J. Korean Phys. Soc.*, Vol. 61, No. 11, Pp. 1751–1758, 2012.
- [27] J. D. Farmer, M. Gallegati, C. Hommes, a. Kirman, P. Ormerod, S. Cincotti, a. Sanchez, and D. Helbing, "A complex systems approach to constructing better models for managing financial markets and the economy," *Eur. Phys. J. Spec. Top.*, Vol. 214, No. 1, Pp. 295–324, 2012.
- [28] M. A. L. Caetano and T. Yoneyama, "A method for detection of abrupt changes in the financial market combining wavelet decomposition and correlation graphs," *Phys. A Stat. Mech. its Appl.*, Vol. 391, No. 20, Pp. 4877–4882, 2012.
- [29] R. B. Roy and U. K. Sarkar, "A social network approach to change detection in the interdependence structure of global stock markets," *Soc. Netw. Anal. Min.*, Vol. 3, No. 3, Pp. 269–283, 2013.
- [30] a. Sensoy, S. Yuksel, and M. Erturk, "Analysis of cross-correlations between financial markets after the 2008 crisis," *Phys. A Stat. Mech. its Appl.*, Vol. 392, No. 20, Pp. 5027–5045, 2013.
- [31] S.-H. Liao and S.-Y. Chou, "Data mining investigation of co-movements on the Taiwan and China stock markets for future investment portfolio," *Expert Syst. Appl.*, Vol. 40, No. 5, Pp. 1542–1554, 2013.
- [32] S. Hu, H. Yang, B. Cai, and C. Yang, "Research on spatial economic structure for different economic sectors from a perspective of a complex network," *Phys. A Stat. Mech. its Appl.*, Vol. 392, No. 17, Pp. 3682–3697, 2013.
- [33] K. Park and H. Shin, "Stock price prediction based on a complex interrelation network of economic factors," *Eng. Appl. Artif. Intell.*, Vol. 26, No. 5–6, Pp. 1550–1561, 2013.
- [34] K. Lim, M. J. Kim, S. Kim, and S. Y. Kim, "Statistical properties of the stock and credit market: RMT and network topology," *Phys. A Stat. Mech. its Appl.*, Vol. 407, Pp. 66–75, 2014.
- [35] O. Castrén and M. Rancan, "Macro-Networks: An application to euro area financial accounts," *J. Bank. Financ.*, Vol. 46, Pp. 43–58, 2014.
- [36] C. Yang, Y. Chen, L. Niu, and Q. Li, "Cointegration analysis and influence rank—A network approach to global stock markets," *Phys. A Stat. Mech. its Appl.*, Vol. 400, Pp. 168–185, 2014.
- [37] Y. Mai, H. Chen, and L. Meng, "An analysis of the sectorial influence of CSI300 stocks within the directed network," *Phys. A Stat. Mech. its Appl.*, Vol. 396, Pp. 235–241, 2014.
- [38] C.-Z. Yao, J.-N. Lin, X.-Z. Zheng, and X.-F. Liu, "The study of RMB exchange rate complex networks based on fluctuation mode," *Phys. A Stat. Mech. its Appl.*, Vol. 436, Pp. 359–376, 2015.
- [39] P. Ou and H. Wang, "Prediction of Stock Market Index Movement by Ten Data Mining Techniques," *Mod. Appl. Sci.*, Vol. 3, No. 12, Pp. 28, 2009.
- [40] A. Kia, SamanHaratizadeh, and HadiZare, "Prediction of USD / JPY Exchange Rate Time Series Directional Status by KNN with Dynamic Time," *Bonfring Int. J. Data Min.*, Vol. 3, No. 2, Pp. 12–16, 2013.
- [41] Y. Ji, J. Hao, N. Reyhani, and A. Lendasse, "Direct and recursive prediction of time series using Mutual information selection," *Comput. Intell. Bioinspired Syst.*, Vol. 3512, Pp. 1010–1017, 2005.
- [42] M. Newman, *Networks: an introduction*. Oxford University Press, 2010.
- [43] G. Creamer and S. Stolfo, "A link mining algorithm for earnings forecast and trading," *Data Min. Knowl. Discov.*, Vol. 18, No. 3, Pp. 419–445, 2009.
- [44] R. Agrawal and R. Srikant, "Fast algorithms for mining association rules," *Proceeding VLDB '94 Proc. 20th Int. Conf. Very Large Data Bases*, Vol. 1215, Pp. 487–499, 1994.



Arash Negahdari Kia did his B.S. in Computer Science at Department of Mathematics and Computer Science in University of Tehran during 2001-2005. He did his M.Sc. of IT engineering in e-commerce at School of Industrial Engineering, Iran University of Science and Technology during 2006-2009. He is presently pursuing his Ph.D. in IT engineering in the area of Data Mining and financial prediction under the guidance of Dr. SamanHaratizadeh, Assistant Professor in Department of Network Science and Technologies, University of Tehran. His Master Thesis was about prediction of foreign exchange rate time series with hybrid models of artificial neural networks and ARIMA. His Ph.D. Thesis is in the field financial prediction using a combination of network science and machine learning methods.



Dr. Saman Haratizadeh has received the doctorate degree in Artificial Intelligence, Software Engineering from Sharif University of Technology. He is currently working as Assistant Professor in Department of Network Science and Technologies, University of Tehran. His research interests, courses, and papers are in the field of Data Mining, Machine Learning, Network Science, Decision Support Systems, Bio-Informatics and development of intelligent tools for information processing, and automatic decision making.



Dr. Zainab Hoda Heshmati has received the doctorate degree in Telecommunications and Computer Engineering from University of Leeds, England. She is currently working as Assistant Professor in Department of Network Science and Technology, University of Tehran. Her research interests, courses, and papers are in the field of Health IT, Network Science, Data Mining, and e-Commerce.