scientific reports

Check for updates

OPEN

Structure and dynamics of financial networks by feature ranking method

Mahmudul Islam Rakib[1], Ashadun Nobi[1] & Jae Woo Lee[2]✉

Much research has been done on time series of financial market in last two decades using linear and non-linear correlation of the returns of stocks. In this paper, we design a method of network reconstruction for the financial market by using the insights from machine learning tool. To do so, we analyze the time series of financial indices of S&P 500 around some financial crises from 1998 to 2012 by using feature ranking approach where we use the returns of stocks in a certain day to predict the feature ranks of the next day. We use two different feature ranking approaches—Random Forest and Gradient Boosting—to rank the importance of each node for predicting the returns of each other node, which produces the feature ranking matrix. To construct threshold network, we assign a threshold which is equal to mean of the feature ranking matrix. The dynamics of network topology in threshold networks constructed by new approach can identify the financial crises covered by the monitored time series. We observe that the most influential companies during global financial crisis were in the sector of energy and financial services while during European debt crisis, the companies are in the communication services. The Shannon entropy is calculated from the feature ranking which is seen to increase over time before market crash. The rise of entropy implies the influences of stocks to each other are becoming equal, can be used as a precursor of market crash. The technique of feature ranking can be an alternative way to infer more accurate network structure for financial market than existing methods, can be used for the development of the market.

The complex dynamic of financial market has been a place of interest for many researchers in last two decades[1–6]. Reconstructing or inferring an unknown network structure from the available monitored time series data has also been a foremost modern network science problem[7–9]. There are many approaches to analyze the time series of stocks for constructing the network. One of the approaches is the Pearson correlation of the returns of the stocks which has been used in last two decades[10]. In this approach, the threshold network is constructed assigning a threshold from the correlations of the stocks. The decision on the link presence or absence is based on pair-wise correlation between nodes. But the correlation between the given pair of nodes is linear. However, there can be also a nonlinear relationship between two stocks. To address this problem, the nonlinear correlation between stocks known as mutual information is also used to analyze the time series of financial indices in recent years[11–13]. Then, using mutual information, minimum spanning tree (MST), planar maximally filtered graph (PMFG) and also threshold networks are constructed and the network topologies are determined[14]. But mutual information takes univariate stance in correlation.

In this article, we use a new approach known as feature ranking of machine learning which can take a multivariate view on the correlation[15–17]. This approach was completely different than the existing correlation and mutual information technique. In recent past, this method has been used in logistic chaotic time series generated by using logistic map and to construct dynamic network[15]. Hence its real-world applicability was not checked. In this paper, we apply this approach in financial time series and construct financial network which has never seen before. We also did entropic analysis of the financial system using feature matrix which may be used as an early warning of the financial market crash. So, the feature ranking method is a novel approach in financial system.

One of the core techniques of machine learning is supervised learning[18]. In this method, the model tries to learn from a given dataset. In the dataset, there consists a dependent variable which is the target and a set of independent variables which are the features. Often, all the features don't influence the target with same degree. Generally, the target depends more on some features than some others. Taking this into account, we can rank the features according to their influence on the target. In machine learning, this phenomenon is known as feature

[1]Department of Computer Science and Telecommunication Engineering, Noakhali Science and Technology University, Sonapur, Noakhali 3814, Bangladesh. [2]Department of Physics, Inha University, Incheon, Republic of Korea. ✉email: jaewlee@inha.ac.kr

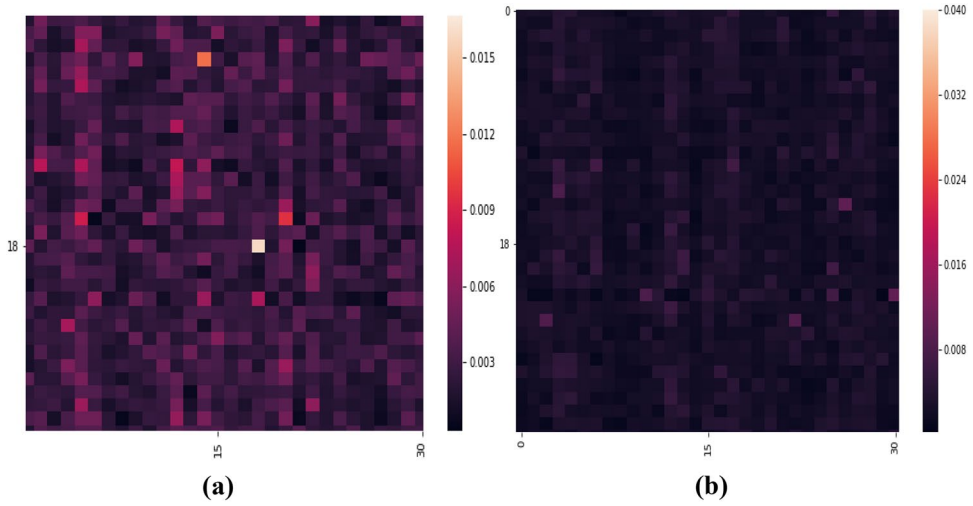

Figure 1. Feature ranking matrices of 30 companies out of 375 companies: (**a**) 2008 (**b**) 2011. Here, we chose 30 companies to show finer map. The light shaded color in heat map indicates the companies which influence more to other companies. The influences of companies during European sovereign debt crisis were higher than global financial crisis.

ranking[19]. There are different methods such as Random Forest[20], ReliefF[21], Decision Tree[22], Gradient Boosting[23], which calculate feature ranking implicitly. As features with low ranks usually have a very little or no influence on the target variable, it doesn't contribute to the accuracy of the machine learning model[15]. Rather they complicated the model yet causing a major degrading called over fitting. Hence, we can improve any machine learning model by simply discarding the insignificant features. We can use this whole process of modern complex network reconstruction method to identify the backend structure of stock markets.

Now, we will introduce how we apply the feature ranking approach on empirical time series. We consider every company as an individual dynamic system and monitor them, which in turn generate a time series. We assume the structure of the studied network to be hidden in a black box and try to reconstruct it using the time series provided by all the stocks individually. As we see, different analytic tools are available in order to reveal the black box. Among those tools, machine learning approaches seem more promising. In this paper, we propose our method of reconstructing networks for stock markets from discrete time series by applying the Feature Ranking[19], which has never been used with the financial time series data before. Here we use the monitored discrete time series data to measure how much the target is influenced by each feature and compute the feature ranking accordingly. Some features have such strong impact on the target that we can safely assume that those features have some connection edges with the target node[15]. We use logarithmic return r_t of the closing prices of stocks of the current day to predict the state or return r_{t+1} of the target stock in the next day. While predicting r_{t+1}, the importance of stocks in predicting the state of the target stock is calculated by the Random Forest model[20]. After applying the prediction model on the whole dataset D, we find the importance of all stocks in predicting the state of all other stocks. In this way, the feature importance matrix or feature ranking matrix F is calculated where stocks act as features. Here, the feature ranking matrix shows asymmetric properties which makes sense because a powerful stock company may have influence on a relatively less powerful company but vice-versa may not applicable. Hence, we represent a directed network in this paper. We calculate both static and dynamic threshold which gives us two different networks for each year. We then analyze the network using different topological properties. Some properties e.g. average clustering coefficient[24], Shannon entropy[25] shows good result for both static and dynamic network and capable of identifying crisis moments where other properties obtain good result on static network only. We also show the most influential companies during crisis moments. We show that, feature ranking approach infer more accurate network structure for financial market than existing methods.

Results

We calculate feature ranking matrix based on the machine learning models. The feature ranking matrix measures feature importance in terms of probability. We choose Random Forest which uses Gini Impurity or Information Gain approach which finds the best split. It calculates Gini feature importance implicitly which is nothing but the probability of contribution in prediction. Figure 1 shows the feature ranking matrices for two different crises 2008 and 2011 as shown in Figure 1a and b, respectively. The matrices are asymmetric which means unidirectional relationship between two stocks as giant stocks may influence most other stocks but may not be affected much by them all.

We calculate the dynamic threshold and static threshold from elements of the ranking matrix. Figure 2 shows how dynamic thresholds changes with one-year time window. During different crisis years, the dynamic threshold changes to higher values than any normal period. The trend shows that the threshold remains smaller in the beginning of the 'dot-com bubble' (2000) and gradually increased up to 2003. The dynamic threshold acts as a

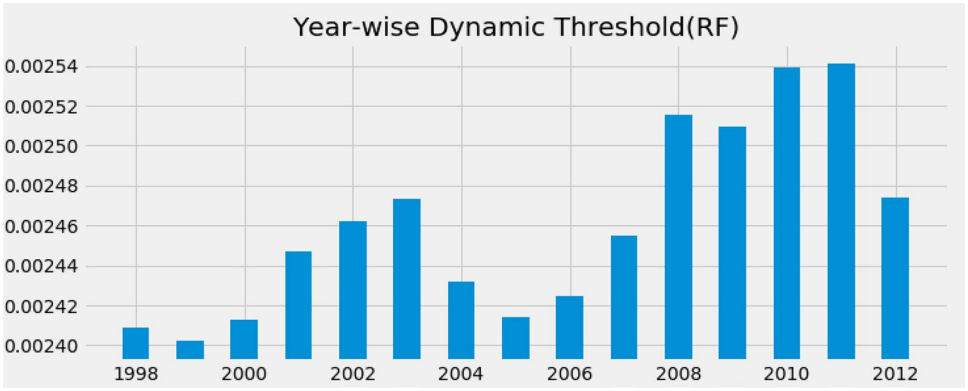

Figure 2. Fluctuation of dynamic threshold over different monitoring years.

good indicator for both the 2008 global finnacial crisis and 2011 European sovereign debt crisis as it holds peak values during these times. On the other hand, static threshold retains its value to 0.0026. As dynamic threshold assigned high values during crises, the finer network structure and topological properties are not seen. But as static threshold retains its value, it shows proper topological properties.

Topological properties

Let's consider the topological change of the financial network induced by the feature rank matrix**.** In a financial system of size N, the companies are known as the nodes of the network and set of links among the nodes depend on threshold. In this paper, the financial networks are constructed from feature ranking matrix F assigning a certain threshold θ. Here, we choose two static thresholds $\theta = 0.0026$ and 0.0027. One is mean value of F and another is near to mean value. Since, network is sensitive to the threshold value and for this reason, we took another threshold near mean of the feature ranking matrix. An edge will be added in the financial network if a feature importance is higher than the pre-determined threshold, that is, if $F_{ij} > \theta$, where $i,j = 1, 2\ldots,N$. We represent the network with directed graph in Fig. 3 **for 30 companies**. The network with dynamic threshold shown in Fig. 3b is densely connected than the network generated using static threshold in Fig. 3a. The threshold networks show heavy connections for static and dynamic threshold. The network does not show the scale-free behavior for the degree distribution.

Average clustering coefficient. The dynamics of average clustering coefficient of the threshold network will be measured over time. The average clustering coefficient is a measure of the compactness and robustness of a network. The clustering coefficient of a vertex i can be expressed as[26–28],

$$C_i = \frac{m_i}{n_i(n_i - 1)} \tag{1}$$

where n_i denotes the number of neighbors of vertex i, and m_i is the number of the edges existing between the neighbors of vertex i. C_i is equivalent to 0 if $n_i \leq 2$. The average clustering coefficient at a specific threshold for the entire network is defined as the average of C_i over all the nodes in the network, i.e.

$$\mathrm{C} = \frac{1}{\mathrm{N}}\sum_{\mathrm{i}=1}^{\mathrm{N}} \mathrm{C_i} \tag{2}$$

The average culstering coefficient of the financial threshold network of s&p500 is shown at mean threshold $\theta = 0.0026$ and 0.0027 and it remains fixed in all time windows. The samll change of threshold show the sharp change of the network structure and for this reason, we chose another threshold around mean. However, the trend is almost similar. The peak values of average clustering coefficent is found in three crises which are dot-com bubble in 2002, global finnacial crisis in 2008 and European sovereign debt crisis in 2011 shown in dotted line in Fig. 4a. The higher values of average clustering coeffiicent imply that the influence of one company to other is higher during crises. The influence of stocks during global and European crises is higher than dot-com bubble. Since, dot-com bubble hit to the technological companies, the lower average clustering coefficient in this period is appropriate.

Average shortest path. The average shortest path length of the threshold network is determined with the evolution of time. The characteristic path length or the average shortest pathlength in a cluster can be expressed as[26–28],

$$\bar{l} = \frac{1}{N(N-1)} \sum_{i,j} l_{ij} \tag{3}$$

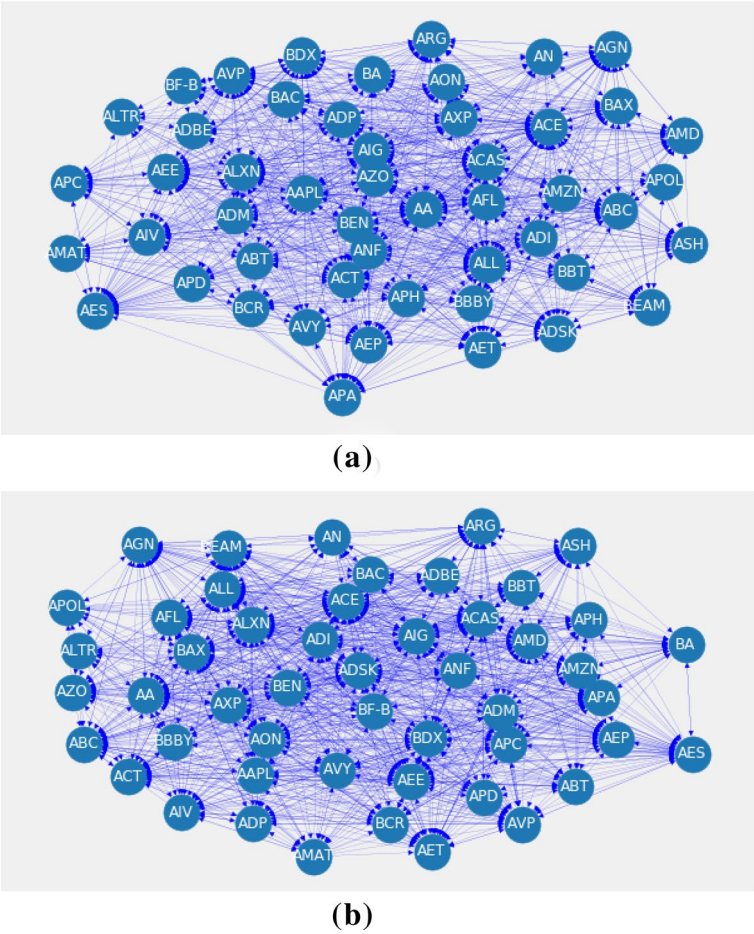

Figure 3. Visualization of directed network structure of 2008 for 50 stocks with: (**a**) Static threshold (**b**) Dynamic threshold. Here, we work on static threshold since it can identify financial crises properly. The higher indegree of a node imply the more influential company in the network. The names of the stock are given in appendix A.

where l_{ij} is the shortest path length between nodes i and j.The average shortest path length of the financial threshold network of s&p500 at threshold $\theta = 0.0026$ and $\theta = 0.0027$ is shown in Fig. 4b. The curve shows a comparatively higher path length in 1999 which indicate that nodes are loosely connected with each other in this year. Then it shows a decreasing trend up to 2003, which implies that nodes become closer during dot-com bubble. The lowest mean shortest path length was observed during ESD crisis (2010) which implies that the dependencies of companies with each other are higher in ESD crisis than global crisis in 2008. On other times, higher mean shortest path length is observed that implies, companies are less dependent to each other and indicates steady state of the market. We also observe that the higher the threshold, the bigger the shortest path is.

Network density. The network density is the ratio of the number of existing links to the maximum number of possible links, which can be determined as[26,29],

$$\rho = \frac{M}{[N(N-1)]} \tag{4}$$

where N is the total number of the nodes and M is the number of connecting links. The network density of the threshold network of s&p500 at threshold $\theta = 0.0026$ and $\theta = 0.0027$ is shown in Fig. 4c. At the beginning i.e. at 1998, the curve starts with a higher density indicating the effect of Asian and Russian financial crisis in 1997–1998. The local market also shows a higher trend in 2000–2003 due to the effects of dot-com bubble and September 11 attack, which then declined up to 2005. After that, the highest densities are observed during the mortgage crisis (2007), global crisis (2008) and ESD crisis (2010 and 2011), which indicates a tightly coupled and highly influential network during crisis. The higher value of network density is found in 2010 than global crisis as like average shortest path. During normal period, the curve shows lower density, indicating a loosely connected network while the market is in its calm state.

Figure 4. Network properties of the static threshold network for s&p500: (**a**) average clustering coefficient, (**b**) average shortest path length, (**c**) network density, (**d**) global reaching centrality, and (**e**) modularity.

Global reaching centrality. The global reaching centrality (GRC) is a global network quantity that calculates the flow of hierarchy of a complex network. It can be defined as[30],

$$GRC = \frac{\sum_{i \in V}[C^{max} - C(i)]}{N - 1} \tag{5}$$

where $C(i)$ is the local reaching centrality (LRC) of node i and C^{max} is the maximum value of LRC.The global reaching centrality of financial threshold network of s&p500 at threshold $\theta = 0.0026$ and $\theta = 0.0027$ is shown in Fig. 4d. The curve shows high value during all finnacial crisies from 1998 to 2012. The high values indicate

Top betweenness centrality			Year	Top influential nodes		
Company name	Company type	Percentage of links		Company name	Company type	Percentage of links
Diamond offshore drilling, Inc	Energy	94.4	2008	Diamond offshore drilling, Inc	Energy	94.4
ACE limited	Financial Services	91.7		ACE Limited	Financial Services	91.7
CenturyLink, Inc	Communication Services	95.7	2011	CenturyLink, Inc	Communication Services	95.7
Sprint Nextel Corp	Communication Services	94.1		Sprint Nextel Corp	Communication Services	94.1

Table 1. Top 2 central and most influential stocks of the year 2008 and 2011 of s&p500. The energy and financial services companies were the most leading in the market during global financial crisis while communication services companies were in ESD crisis.

maximal heterogeneous distribution of the *LRC* which implies maximal hierarchical state of the market during crisis. The sharp change of GRC is observed during subprime mortgage crisis in 2007 which indicate that the influential nodes are placed in the center of the network. This hierarchical state of the market sustains up to 2011 and the market comes back again in low hierarchical state during 2012.

Modularity. The true community structure in a network which can be quantified by using the phenomenon known as modularity Q_N can be expressed as[31,32],

$$Q_N = \frac{1}{2m}\sum_{i,j}\left(A_{ij} - \frac{k_i k_j}{2m}\right)S_{ij} \tag{6}$$

where k_i is the number of edges of node i, and $m = \frac{1}{2}\sum_i k_i$ is the total number of edges in the network, then the probability that two nodes i and j are connected by chance is $\frac{k_i k_j}{2m}$. A is the adjacency matrix, entries are in such a way that $A_{ij} = 1$ if node i connects to node j and $A_{ij} = 0$ otherwise. S is the modularity matrix; entries are in such a way that $S_{ij} = 1$ if nodes i and j belong to the same module, and zero otherwise. Here, S is calculated in such a way that i and j will belong to the same module if they are of the same type companies. The modularity of financial threshold network of s&p500 at threshold mentioned above is shown in Fig. 4e. As like other network parameter, the higher value of modularity is also found in all crises. In our analysis period, the higher modularity is found during dot-com bubble than other crises. Since this crisis was on the technological companies, the intra-community communication of other groups is incresed because they didn't want to communicate with the technogial companies. It indicates that intra-module communication increases and inter-module communication falls during crisis times. That is, same type companies depend more on each other during crisis.

Betweenness centrality. Betweenness centrality is the measurement that captures how much a given node is in-between others. The betweenness centrality can be defined as follows[33],

$$B(u) = \sum_{u\neq v\neq w}\frac{\sigma_{v,w}(u)}{\sigma_{v,w}} \tag{7}$$

where $\sigma_{v,w}(u)$ is the number of shortest paths (between any couple of nodes in the graphs, here node v and w) that passes through the target node u. $\sigma_{v,w}$ is the total number of shortest paths existing between any couple of nodes (here node v and w) of the graph.

Table 1 shows a comparison between top in-between central nodes of the largest subset network and top influential nodes i.e. nodes having most influential edges of two crisis years 2008 and 2011. It also shows the percentage of the number of influential links which is the ratio of number of influential edges of a node and maximal possible influential links. This comparison shows the companies that have most impact on the market during global market crisis in 2008 and ESD crisis in 2011. We found the same companies as central node of the network and the node that have the greatest number of influential edges during two severe crises given in Table 1. During global financial crisis, the Diamond Offshore Drilling and ACE limited were the most powerful companies and they are in the sector of energy and financial services respectively. However, the most leading companies during ESD crisis were Century Link and Sprint Nextel Corporation and they are in the sector of communication services. In this process, we can identify the most powerful companies of the market in any time which can be useful for risk management and portfolio investment.

Entropy. The feature ranking is nothing but the probability of contribution of a company in predicting the return of the target company. The probability of contribution is distributed among 375 companies and the companies which have higher probability imply the most influential in predicting the return of the target. Since total probability of all companies is one, we can calculate the entropy to understand the state of the market. The entropy is calculated as[25],

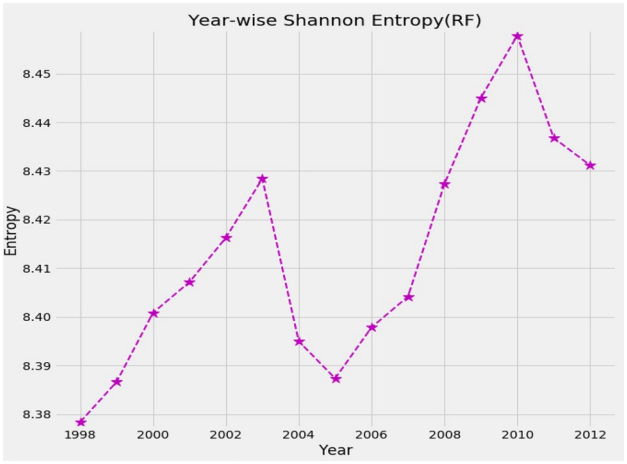

Figure 5. The change of entropy over time. The higher the entropy, the higher the risk of the market.

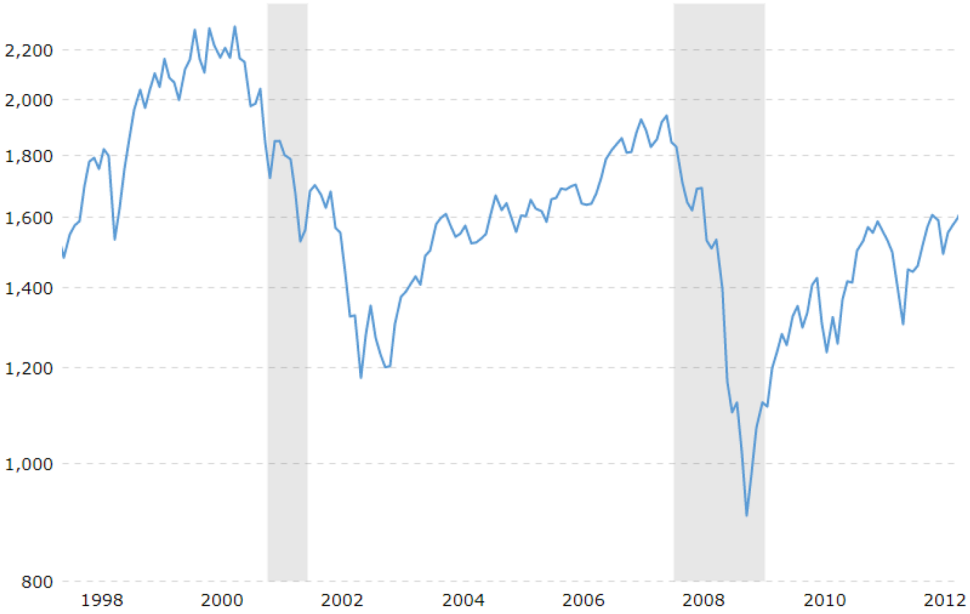

Figure 6. Historical records[35] of the S&P 500, where the dataset starts from 1998 to 2012. The shaded area indicates the financial crises.

$$S = -\frac{1}{N}\sum_{i,j=1}^{N} F_{ij} log_2\left(F_{ij}\right) \tag{8}$$

The higher entropy implies that the influences of the companies are becoming equal. We observe that the entropy is increasing from 1998 to 2003 as shown in Figure 5. It means that the market is going to unstable state. We show the index of S&P 500 in the analyzing period in Figure 6. We can compare the change of entropy and the evolution of the index in Figures 5 and 6. Similarly, before global and ESD crises, the entropy is rising and consequently, market falls in crisis. The rise of entropy over time can be used as an indicator of upcoming crisis.

Discussion

We analyze the daily time series of financial indices using a new approach known as feature ranking. This technique is different than the existing linear and non-linear correlation. The feature ranking is non-linear multivariate technique where multiple features are used to construct feature ranking matrix. We construct threshold network assigning a threshold from the mean of feature matrix and the network topologies are investigated over time. The dynamic change of network properties can identify the financial crises which reflect the crucial state of the market. We identify the companies which are more influential during severe state of the market. We

identify four powerful companies belonging to the sector of energy, financial and communication services in two big crises. These companies may shield to protect the crisis. Finally, the entropy estimates and it is seen to increase over time before market crash. The higher entropy can be an indicator for unrest state of the market. Our technique can be an efficient way to analyze the financial time series for development of the market. In future, we will apply this technique in other market using recent time series to monitor the state of the market.

Methods

Data analysis. We monitored the daily closing prices for 375 companies listed in the S&P 500 from 1998 to 2012. The data were collected from Yahoo Finance[34]. We then segmented the data using 1-year time window. During this period, the market has covered by different major crises such as 'dot-com' bubble in 2000, September11 attack in 2001, the crash of 2002, sub-prime mortgage crisis in 2007, Global financial crisis in 2008 and European sovereign debt (ESD) crisis in 2010 and 2011 as shown in Fig. 6.

Logarithmic return. The daily return of ith stock index on day t, $r_i(t)$, can be defined as,

$$r_i(t) = ln[I_i(t)] - ln[I_i(t-1)] \tag{9}$$

where $I_i(t)$ is the closing price of a stock index i on day t. Thus, we can observe and measure the time series of the dynamics of companies.

Feature ranking approach. Feature ranking is mostly a concept of Machine Learning. Machine learning studies algorithms which learn to make decisions or predictions from some given sample dataset, known as "training data". These algorithms improve their performance through "experience"[36]. For example, by observing many sample pictures of cat and dog, machine learning algorithm finds distinguishable patterns by which it learns to classify new dog or cat that was never seen. Machine learning is a data driven learning approach which can be used in the circumstances where human knowledge about the studied sector is limited. This is the core reason for which machine learning is widely used in various scientific disciplines, ranging from natural language processing[37], medicine and biology[38,39] to financial market analysis[40], image identification and classification[41]. Machine learning approaches are further divided into some categories on the basis of attributes and feedbacks available to the learning process. Supervised learning is one of such categories where training data consists of both input or independent variables and desired output label or dependent variable. A supervised learning algorithm analyzes the training input data and predicts corresponding output data which get compared with the labeled output data. The model penalized with some feedback value if any mismatch happens between predicted output and labeled output through which the model learns to predict more accurately. Here the input data is said to be the "feature" and the labeled output is called the "target". Hence, data in supervised learning is given in the feature-target representation. While predicting the target, all the features don't contribute with the same degree. Some features have more influence on target than some other while predicting the target. Hence, we can determine the importance of features or calculate feature ranks with the help of supervised learning. In this work we adopt feature ranking method introduced by Ref.[15]. We use regression model such as Random Forest[20] and XGBoost[23]. Both of them are Decision Tree based ensemble Machine Learning algorithm. This type of algorithms make prediction by recursively splitting or partitioning data on the basis of selecting attribute or feature. To find the best split of data, it uses method like Gini Impurity or Information Gain while selecting feature in each step. These methods also measure importance of features in the best split. For example, Gini Impurity calculates Gini feature importance implicitly which is nothing but the probability of contribution in prediction. Feature gets higher importance in prediction if selecting it led to a greater reduction in Gini Impurity. We use these techniques of calculating feature importance in our dataset. In every iteration, we select a node as target and calculate importance of all feature nodes F_i in predicting the target node, where i is the feature. After all iteration, we find the feature importance for all target nodes and this gives us the feature ranking matrix F. The matrix element F_{ij} indicates the importance of the feature j in predicting target i. The implementation detail of calculating feature ranking matrix is discussed in the following section.

We introduce the method to obtain the matrix element F_{ij} through Eqs. (10)–(12). We input the training data set D_i to the Machine Learning algorithm to obtain the feature element F_{ij}. Therefore the returns of the index j at the previous time t influence to the feature of the index i at time $t+1$, which represents the element of the feature matrix F_{ij}. From this feature matrix, we generate the directed networks showing the influence from the source node to target node.

Reconstruction method. We now briefly dive into our reconstruction method. In the above section, we explain how feature ranking method works. This method can effectively be applied to the problem of reconstructing a dynamical network of stock market from the monitored time series data of its node dynamics (generated using Eq. 9). We select a node (share company) of a dynamical network (stock market) and assume its state as the target, keeping all other node's dynamics as features to calculate their influence on the selected node. Now from here, we define a supervised learning model e.g. a regression model for predicting the state of the selected node (target company) from the dynamics of all the other nodes in the network. Important fact to note here that, our goal here isn't to build a predictive model, rather, we only try to construct the feature ranking matrix i.e. our main goal is to rank the importance of the other nodes to the selected one, because a highly ranked node is more likely to be connected to the selected node. We now only need to repeat this procedure for all the stock companies and we can reconstruct the entire network structure of the stock market i.e. which stock company is going to affect a particular stock company's state.

Let's consider $r_i(t+1)$ at time $t+1$. Its dynamics is influenced by the previous state represented as,

$$r_i(t+1) = f_i(r_1(t), r_2(t), \ldots, r_N(t)), i = 1, \ldots, N \tag{10}$$

where the interaction function f_i is unknown which can be modeled using the observed time series data and $N = 375$. We construct our training dataset D_i with $L-1$ instance from the monitored time series as,

$$D_i = \bigcup_{t=1}^{L-1} (r_1(t), r_2(t), \ldots, r_N(t); r_i(t+1)) \tag{11}$$

where $r_i(t+1)$ is our target index and the states of all the other indices at some prior time t is considered as feature index.

In this paper, we don't model the interaction function f, rather we will only focus on network structure of the stock market. To calculate the feature ranking of share companies, we consider the two mentioned algorithms, Random Forest[20] and XGBoost[23]. By applying any of the above-mentioned feature ranking algorithm R to the training data D_i,we get feature ranks for node i,

$$(F_{i1}, F_{i2}, \ldots, F_{iN}) = R(D_i) \tag{12}$$

where F_{ij} tells us the estimated impact of node j (share company j) on node i (share company i). By estimating this feature importance for all N share companies, we construct a feature ranking matrix F of dimension $N \times N$,

$$F = \begin{pmatrix} F_{11} & F_{12} & \cdots & F_{1N} \\ F_{21} & F_{22} & & \vdots \\ \vdots & & \ddots & \\ F_{N1} & \cdots & & F_{NN} \end{pmatrix} \tag{13}$$

We see that, the estimated feature ranking matrix shows asymmetric properties i.e. F_{ij} may not equal to F_{ji}. Hence, a company may influence some other companies but vice-versa may not applicable. This makes sense as powerful companies may have unidirectional influence. Hence, here we represent a directed network. We can simply assume that higher the value of F_{ij} is, more likely it is that the link $i \rightarrow j$ exists.

Finally, we come up with our desired reconstructed adjacency matrix $\widehat{A}$ simply by filtering out low ranked features by setting up a threshold value θ. We calculate both dynamic thresholds and static threshold. In dynamic thresholding, threshold value may change every year depending on the values of F, while static threshold retains its value each year. Dynamic threshold θ_D can be defined as,

$$\theta_D = \frac{1}{N} \sum_{i=1}^{N} \left\{ \frac{(N+1)}{2} \right\}_{F_{ij}} \tag{14}$$

where θ_D is a vector of per year threshold and N is the number of nodes. Here, $\left\{ \frac{(N+1)}{2} \right\}_{F_{ij}}$ is the median of row i in F where J indicates N columns and $\left\{ \frac{(N+1)}{2} \right\}$ is the index of the median in row i. On the other hand, static threshold can be defined as,

$$\theta_S = \frac{1}{N * N} \sum_{i,j=1}^{N} F_{ij} \tag{15}$$

where F_{ij} is an entry of row i and column j of the feature ranking matrix. We then calculate the reconstructed adjacency matrix $\widehat{A}$ from the feature ranking matrix F with the help of the estimated thresholds as follows,

$$\hat{A}_{ij} = \begin{cases} 0 \ if F_{ij} \leq \theta \\ 1 \ if F_{ij} > \theta \end{cases} \tag{16}$$

Received: 23 March 2021; Accepted: 10 August 2021
Published online: 02 September 2021

References

1. Mantegna, R. N. & Stanley, H. E. *Introduction to Econophysics*. (1999). https://doi.org/10.1017/cbo9780511755767.
2. Pelletier, D. Regime switching for dynamic correlations. *J. Econom.* **131**, 445–473 (2006).
3. Drozdz, S., Kwapień, J., Grümmer, F., Ruf, F. & Speth, J. Quantifying the dynamics of financial correlations. *Phys. A* **299**, 144–153 (2001).
4. Tastan, H. Estimating time-varying conditional correlations between stock and foreign exchange markets. *Phys. A Stat. Mech. Appl.* **360**, 445–458 (2006).
5. Rosenow, B., Gopikrishnan, P., Plerou, V. & Stanley, H. E. Dynamics of cross-correlations in the stock market. *Phys. A* **324**, 241–246 (2003).
6. Eleanor, X. X., Chen, P. & Wu, C. Time and dynamic volume-volatility relation. *J. Bank. Financ.* **30**, 1535–1558 (2006).
7. Yang, S. Networks: An introduction by M.E.J. Newman. *J. Math. Sociol.* **37**, 250–251 (2013).
8. Barabási, A.-L. Network science introduction. *Netw. Sci.* 1–27 (2016).
9. Porter, M. A. & Gleeson, J. P. *Dynamical systems on dynamical networks*, 49–51 (2016).https://doi.org/10.1007/978-3-319-26641-1_6.

10. Kenett, D. Y. *et al.* Dominating clasp of the financial sector revealed by partial correlation analysis of the stock market. *PLoS ONE* **5**, e15032 (2010).
11. Guo, X., Zhang, H. & Tian, T. Development of stock correlation networks using mutual information and financial big data. *PLoS ONE* **13**, e0195941 (2018).
12. Hartman, D. & Hlinka, J. Nonlinearity in stock networks. *Chaos* **28**, 083127 (2018).
13. Fiedor, P. Partial mutual information analysis of financial networks. *Acta Phys. Pol. A* **127**, 863–867 (2015).
14. Yan, Y., Wu, B., Tian, T. & Zhang, H. Development of stock networks using part mutual information and australian stock market data. *Entropy* **22**, 773 (2020).
15. Leguia, M. G., Levnajić, Z., Todorovski, L. & Zenko, B. Reconstructing dynamical networks via feature ranking. *Chaos* **29**, 093107 (2019).
16. Von Toussaint, U. Bayesian inference in physics. *Rev. Mod. Phys.* **83**, 943–999 (2011).
17. Zanin, M. *et al.* Combining complex networks and data mining: Why and how. *Phys. Rep.* **635**, 1–44 (2016).
18. Cunningham, P., Cord, M. & Delany, S. J. Supervised learning. In *Cognitive Technologies*, 21–49 (2008). https://doi.org/10.1007/978-3-540-75171-7_2.
19. Iguyon, I. & Elisseeff, A. An introduction to variable and feature selection. *J. Mach. Learn. Res.* **3**, 1157–1182 (2003).
20. Breiman, L. Random forests. *Mach. Learn.* **45**, 5–32 (2001).
21. Robnik-Šikonja, M. & Kononenko, I. An adaptation of {R}elief for attribute estimation in regression. In *Machine {L}earning: {P}roceedings of the {F}ourteenth International Conference (ICML'97)* 296–304 (1997).
22. Quinlan, J. R. Induction of decision trees. *Mach. Learn.* **1**, 81–106 (1986).
23. Chen, T. & Guestrin, C. XGBoost: A scalable tree boosting system. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* vols. 13–17-Augu 785–794 (2016).
24. Chalancon, G., Kruse, K. & Babu, M. M. Clustering coefficient. In *Encyclopedia of Systems Biology*, 422–424 (2013). https://doi.org/10.1007/978-1-4419-9863-7_1239.
25. Lesne, A. Shannon entropy: A rigorous notion at the crossroads between probability, information theory, dynamical systems and statistical physics. *Math. Struct. Comput. Sci.* **24**, e240311 (2014).
26. Nobi, A., Lee, S., Kim, D. H. & Leea, J. W. Correlation and network topologies in global and local stock indices. *Phys. Lett. Sect. Gen. Solid State Phys.* **378**, 2482–2489 (2014).
27. Solnik, B., Boucrelle, C. & Le Fur, Y. International market correlation and volatility. *Financ. Anal. J.* **52**, 17–34 (1996).
28. Pollet, J. M. & Wilson, M. Average correlation and stock market returns. *J. Financ. Econ.* **96**, 364–380 (2010).
29. Koldanov, A. P., Koldanov, P. A., Kalyagin, V. A. & Pardalos, P. M. Statistical procedures for the market graph construction. *Comput. Stat. Data Anal.* **68**, 17–29 (2013).
30. Mones, E., Vicsek, L. & Vicsek, T. Hierarchy measure for complex networks. *PLoS ONE* **7**, e33799 (2012).
31. Newman, M. E. J. Modularity and community structure in networks. *Proc. Natl. Acad. Sci. USA* **103**, 8577–8582 (2006).
32. Hintze, A. & Adami, C. Modularity and anti-modularity in networks with arbitrary degree distribution. *Biol. Direct* **5**, 32 (2010).
33. Perez, C. & Germon, R. Graph creation and analysis for linking actors: Application to social data. In *Automating Open Source Intelligence: Algorithms for OSINT*, 103–129 (2016). https://doi.org/10.1016/B978-0-12-802916-9.00007-5.
34. ^GSPC 3,876.50 -30.21 -0.77% : S&P 500 - Yahoo Finance. https://finance.yahoo.com/quote/%5EGSPC/chart?p=%5EGSPC.
35. S&P 500 Index - 90 Year Historical Chart | MacroTrends. https://www.macrotrends.net/2324/sp-500-historical-chart-data.
36. Learning, M., Mitchell, T. & Hill, M. Speeding! up! Decision! Tree! Learning. *Mach. Learn*, 639–644 (2012).
37. Processing, N. L. Light textual inference for semantic parsing. *Coling*, 1007–1018 (2012).
38. Furey, T. S. *et al.* Support vector machine classification and validation of cancer tissue samples using microarray expression data. *Bioinformatics* **16**, 906–914 (2000).
39. Guyon, I., Weston, J., Barnhill, S. & Vapnik, V. Gene selection for cancer classification using support vector machines. *Mach. Learn.* **46**, 389–422 (2002).
40. Huang, W., Nakamori, Y. & Wang, S. Y. Forecasting stock market movement direction with support vector machine. *Comput. Oper. Res.* **32**, 2513–2522 (2005).
41. Granitzer, M. *Hierarchical Text Classification using Methods from Machine Learning*. Master's Thesis, Graz University of Technology (2003).

Acknowledgements

This work was supported by the Ministry of Education of the Republic of Korea and the National Research Foundation of Korea (NRF-2020R1A2C1005334) and ICT ministry of Bangladesh (The tracking Number is 20FS34427).

Author contributions

M.I.R. designed and performed the numerical analyses, wrote the main text and supplementary information, and generated the figures. A.N. collected the data, reviewed the economic background and put the work into context. J.W.L. reviewed the graph-theoretic background. A.N. supervised the project. All authors reviewed the consistency of results and revised the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-021-97100-1.

Correspondence and requests for materials should be addressed to J.W.L.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.