Stock Correlation Analysis Based on Complex Network

Shutian Li, Yi Yang, Caihong Li, Lian Li
School of Information Science & Engineering
Lanzhou University
Lanzhou, China
Shutian.lisht13@lzu.edu.cn

Xiangquan Gui
College of Computer and Communication
Lanzhou University of Technology
Lanzhou, China
xqgui@lut.cn

***Abstract*—Aim at the study of the relationship of correlation between the stocks of SSE(Shanghai Stock Exchange) companies and the movement of SSE complex index , by using the mathematical statistics method and correlation coefficient analysis, select each stock's daily closing price as object, establish the complex network of stocks. Through preprocessing the data sets, we used R Studio get the mean correlation coefficient matrix, which is the foundation of the financial network. The results confirm the utility of the correlation coefficient analysis as a stability indicator for SSE stock market, since it proves that the correlation of stocks is different according to the crisis and prosperous of stock market in times. Its development is opposite to the market index, the interaction between the stocks attenuates when the complex index rises and enhances when the index falls. We also used K-means clustering algorithm to classify these nodes into different communities and found that the structures of networks vary widely according to their correlations. The method of this paper and the model it proposed is not only for the selection of our data sets, but also can be generalized to other fields of research.**

Keywords—complex network; stock data; stock networks; stock correlations

I. INTRODUCTION

The network is ubiquitous in people's daily life, and it can be the specific entity, but also the abstract and untouchable relationship. The research of complex network comes from a branch of graph theory of discrete mathematics, focused on planar network[1] in early stage. However, with the rapid development of computer processing speed and computing ability, it was found that plenty of networks in the real world[2] has a small world characteristics, in which the dynamic behavior is complex. The so-called complex network[3,4], is a large, irregular, dynamic, and complicated network. It grows rapidly in the past decade and provides a new perspective for people to know the world well.

From the perspective of the development of the stock market, the fluctuation trend and magnitude of the stock market index is closely related to macro economy. Complex networks highly summarized the characteristics of complex systems[5], is currently recognized as one of the more suitable modeling of complex systems. As a complex economic system, stock markets[6] can be abstracted into a complex network and it can be analyzed by using complex network. Each stock can be used as a node in the network, the price linkage between stocks as links in the network. Since there is a certain correlation[7] between the stock and the correlation, we studied the correlation coefficient between stocks. Through the comparison of the average correlation coefficient and the stock market index, we found that the relationship between these two is anti correlated, that is, the stock market index lower, the average correlation coefficient is bigger, that is to say, the n coefficient is bigger, that is to say, the greater the impact of the stocks of mutual influence, vice versa.

II. DATA ACQUISITION AND PROCESSING

All data used in this paper was retrieved from the Yahoo[8]. In order to facilitate the study, the data were processed as follows:

- For missing days of data, such as when listing corporation has important information to be published, when the securities regulatory authorities believe that the listing corporation shall clarify and notice the issues related to the company's significant impact, when listing corporation is suspected of illegal and needs to be investigated, we assume that it is not much different of previous day's data, in the range of allowable error.
- Filling the missing data with previous data.
- We eliminated those suspended for more than 21 days of stock data, thus avoiding the calculation of correlation coefficient when the denominator is zero.

III. RESEARCH STRATEGY

Correlation coefficient is used to reflect the correlation between the variables and the degree of statistical indicators. The correlation coefficient is calculated by the method of product difference, and is based on the two variables and their respective mean values, and the correlation between the two

978-1-5090-1997-7/16/$31.00 ©2016 IEEE

variables is reflected by the product of two dispersion. By building a network model[9] of the closing price of the stock, we can calculate the correlation coefficient between different nodes, so that the correlation coefficient between any two stocks can be known. Through calculating the average correlation coefficient, the relationship between the market index and the nodes in the network is observed.

IV. COMPLEX NETWORK MODEL OF BUILDING STOCKS

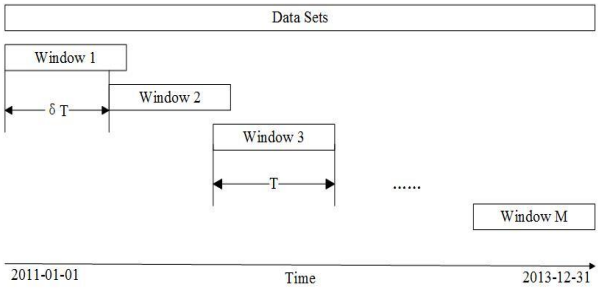
A. Selecting the Stock Data Sets

At present, the Shanghai Composite Index put the market average price earnings ratio standard to use, which contains all the members of the Shanghai Stock A shares, will be a lot of poor performance, small size, high price stocks included, it lead to higher market profitability. We choose the stocks of Shanghai 180 Index samples, which are selected according to the scientific and objective sampling method, are more representative.

The data set used here consists of daily closure prices for 130 stocks that have been continuously members of Shanghai 180 index since January 2011. Our time frame spans over 714 days, ending on December 2013. It contains energy, materials, industry, consumer discretionary, consumer staples, health care, financial, information technology, telecommunication services, utilities amount to 10 major industries.

B. Defining the Parameter δT

In order to investigate the dynamics of the Shanghai stock market, we divided these data into *M* windows[10], denominated *t=1,2,...M* of width T, that is, the number of daily returns in *M*. The experimental results show that the data is not



continuous, so we have a certain overlap[11], shifting further at length *δT*.

Figure 1. The time window with width T and overlap δT

C. The Definition of the Correlation Coefficient

The basic information in our data consists of *N* assets with closing price $P_i(\tau)$ for asset *i* on day τ. Then , the logarithmic return is given as

$$r_i(\tau) = lnp_i(\tau) - lnp_i(\tau - 1) \tag{1}$$

The τ is from second day of the time window *T* to the first day of trading time window *t+1*, resulting in the return vector r_i^t for the time window *t, t=1,2,3，.....M*. In order to quantify the degree of similarity between assets *i* and *j* at time *t*, we define the correlation coefficient of *i* and *j* in time window *T* as logarithmic return is given as

$$\rho_{ij}^{\ t} = \frac{E(r_i^t r_j^t) - E(r_i^t)E(r_j^t)}{\sqrt{\left(E(r_i^{t^2}) - E(r_i^t)^2\right)\left(E(r_j^{t^2}) - E(r_j^t)^2\right)}} \tag{2}$$

The correlation coefficient[11] $-1 \le \rho_{ij}^{\ t} \le 1$, *i* and *j* are perfect correlation when $\rho_{ij}^{\ t} = 1$ and completely counter correlation when $\rho_{ij}^{\ t} = -1$.

D. Selection of parameter T and delta δT

We introduce two parameters *T* and *δT* for constructing financial network, the choice of these two parameters also affects the final experimental results. Onnela et al[12,13,14], have used this correlation coefficient and time window method to construct the asset map. At the same time, the selection of the window width *T* is a trade-off, the data exist a lot of noise when *T* is too small while the data is too smooth when *T* is much more large. They found that it is the optimal choice when window step size *T=21* days and *T=1000*. But this is clearly not suitable for our data assets.

We choose the *T=200 , 100,* the window step size *T=21* days (fixed in a month's time), analysis the probability density of their corresponding mean correlation coefficient, which is defined as

$$\tilde{\rho} = \frac{1}{N(N-1)/2} \sum \rho_{ij}^{\ t} \tag{3}$$

E. Determine the Relationship Matrix

Firstly, we use the Eq. (1) established symmetric matrix $\rho_{ij}^{\ t}$. Then we obtain the $N \times N$ matrix C^t according to Eq. (2), which is completely characterized by *N(N-1)/2* correlation coefficients, is a non-diagonal matrix. It can be seen from the value of $\rho_{ij}^{\ t}$ and C^t that almost all of the correlation coefficient between stocks is nonzero, it indicates that there are connections between each stock. The correlation coefficient of some stocks are larger, it shows that the fluctuations of these stocks are more likely to spread in the market, and the influence on the rest of the stocks are also bigger. It shows little relevance to other stocks when $\rho_{ij}^{\ t}$ is small.

F. Analysis of Experimental Results

In Fig.2, the SSE(Shanghai Stock Exchange) complex index rises at about April, 2011, and obviously drops at February, 2012, and sharply falls at January, 2013 and the middle of 2013. Through carefully observation and

comparison, we find that the rise and fall of curve, and the peak and valley in Fig.2 is exactly opposite to Fig.3 and Fig.4.

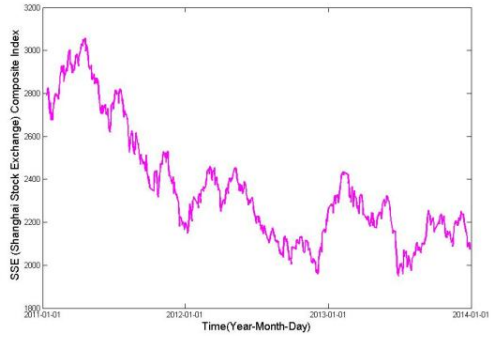


Figure 2. SSE (Shanghai Stock Exchange) Complex Index between 2011 and 2013

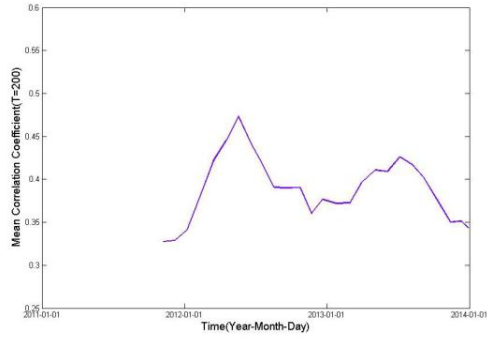


Figure 3. The mean correlation coefficient when T=200

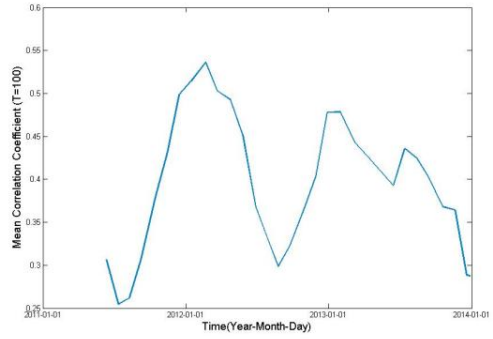


Figure 4. The mean correlation coefficient when T=100

Three distinct phases are clearly visible. First, the correlation in Fig.3 and Fig.4 falls in succession of the economy rise in 2011 and then rises to the peak in subsequent recession on the financial markets in 2012. Afterwards, markets started to grow again, which is indicated by a lower and less volatile mean, and then continued to fall until the end of year 2012. The final phase in our data is the big increase in short-term and a slump followed by, with $\tilde{\rho}$ dropping extraordinarily after a brief rise.

As can be seen from the Fig.3, when *T=200*, it misses a part of data when SSE Complex Index continuously fell down, and that the peak is too sharp to show that it reflects too sensitive to our data. In Fig.4, curve is relatively smooth, so *T=100* is a relatively good choice for our data sets.

G. The Analysis of Network Community Structure Based on K-means Algorithm

Normally, there will be some small groups of close interrelated in huge networks, just like the human society in different groups. In complex networks, called the "community". Excavating the community of the network is very important to understand the structures and properties of the network[15].

There are two basic types of algorithms in searching for community structure in complex networks: hierarchical clustering algorithm and Image segmentation algorithm. The K-means algorithm is a kind of clustering algorithm based on division proposed by Mac Queen, and is currently one of the most widely used clustering algorithm.

If the original data set is extracted as $\{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n\}$, and each $\boldsymbol{x}_i$ is a d-dimensional vector, K-means clustering aims to classify the original data into k classes under the conditions of a given k $(k \leqslant n)$.

$\boldsymbol{S} = \{\boldsymbol{s}_1, \boldsymbol{s}_2, \ldots, \boldsymbol{s}_k\}$, in the numerical model, calculate the minimum value of following expression:

$$\arg\min_{s} \sum_{i=1}^{k} \sum_{x_j \in s_i} \left\| \mathbf{x}_j - \boldsymbol{\mu}_i \right\|^2 \tag{4}$$

Here, $\boldsymbol{\mu}_i$ is the average value of $\boldsymbol{s}_i$. The general steps of the algorithm are as follows:

- Randomly select k elements as k cluster centers..
- Respectively calculate the dissimilarity of the rest of the elements to k centers , and then classified these elements to the lowest cluster dissimilarity.
- According to the clustering results, re-computing center of k clusters respectively. The method is to take arithmetic mean value of each element's dimension in one cluster.
- Clustering all the elements with the new centers.
- Repeat step fourth, until the clustering result does not change.
- Output the results.

Select the network with T=100, set *k=8* to cluster with 130 stock nodes in time window t=10 (market downturn) and t=19 (market boom).

Through the analysis of the clustering results, we find that , as we expected, the nodes in the same community are basically in the same or similar industry. In time window t=10, the correlation between nodes are larger, we can see from Fig.5 that communities are more concentrated while the largest community with 37 nodes and the smallest one with 5 nodes. This shows that the collapse of some hub stock nodes collapse

would affect a lot of stock nodes in the network seriously, or even make the network structure changes. In Fig.6, the distribution of community is relatively average and there are 7 communities contain more than 10 members, this is because in time window t=19 the correlation between nodes are smaller at market boom times.

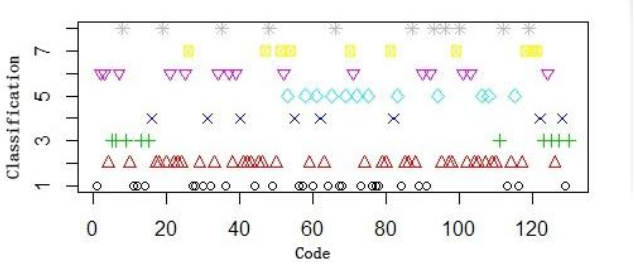


Figure 5. The result of clustering in time window t=10

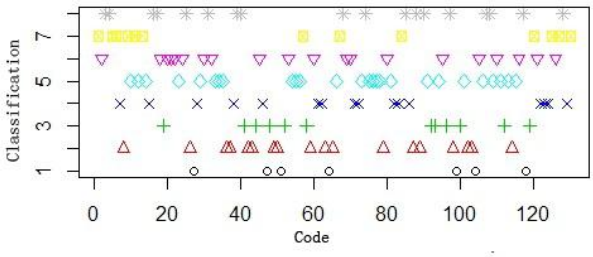


Figure 6. The result of clustering in time window t=19

V. CONCLUSIONS

Above all, we may safely draw the conclusion that the correlation between stocks are connected with the market index, it presents a inverse relationship. In other words, the interaction between the stocks attenuates when the complex index rises and enhances when the index falls. At the same time, the distribution of nodes is loose in times of boom and centralized in times of downturn.

This paper established mathematical model and solved specific problems by using the method of correlation coefficient analysis, so the reliability of the conclusion is higher. In this paper, the undirected network[15] is constructed to study the correlation of stocks, the method and the model it proposed can also be generalized to other fields of research.

In the future research, we can introduce other parameters such as the modularity[16] of networks and the weighted clustering algorithm[17,18,19,20] to investigate stock network stability in times of crisis and boom.

ACKNOLEDGEMENT

The authors would like to thank the Natural Science Foundation of P. R. of China (61300230), open fund of Guangxi Key laboratory of hybrid computation and IC design analysis and the Fundamental Research Funds for the Central Universities for supporting this research.

REFERENCES

[1] Tumminello M, Aste T, Di Matteo T, et al. A tool for filtering information in complex systems[J]. Proceedings of the National Academy of Sciences of the United States of America, 2005, 102(30): 10421-10426.

[2] Watts D J, Strogatz S H. Collective dynamics of 'small-world'networks[J]. nature, 1998, 393(6684): 440-442.

[3] Wang X F, Chen G. Complex networks: small-world, scale-free and beyond[J]. Circuits and Systems Magazine, IEEE, 2003, 3(1): 6-20.

[4] Boccaletti S, Latora V, Moreno Y, et al. Complex networks: Structure and dynamics[J]. Physics reports, 2006, 424(4): 175-308.

[5] Newman M E J. Detecting community structure in networks[J]. The European Physical Journal B-Condensed Matter and Complex Systems, 2004, 38(2): 321-330.

[6] Nobi A, Maeng S E, Ha G G, et al. Effects of global financial crisis on network structure in a local stock market[J]. Physica A: Statistical Mechanics and its Applications, 2014, 407: 135-143.

[7] Sinha S. Are large complex economic systems unstable?[J]. arXiv preprint arXiv:1009.0972, 2010

[8] http://finance.yahoo.com/.

[9] Heiberger R H. Stock network stability in times of crisis[J]. Onnela J P, Chakraborti A, Kaski K, et al. Asset trees and asset graphs in financial markets[J]. Physica Scripta, 2003, 2003(T106): 48.

[10] Costa L F, Rodrigues F A, Travieso G, et al. Characterization of complex networks: A survey of measurements[J]. Advances in Physics, 2007, 56(1): 167-242.

[11] Blondel V D, Guillaume J L, Lambiotte R, et al. Fast unfolding of communities in large networks[J]. Journal of Statistical Mechanics: Theory and Experiment, 2008, 2008(10): P10008.

[12] Tumminello M, Lillo F, Mantegna R N. Correlation, hierarchies, and networks in financial markets[J]. Journal of Economic Behavior & Organization, 2010, 75(1): 40-58.

[13] Keskin M, Deviren B, Kocakaplan Y. Topology of the correlation networks among major currencies using hierarchical structure methods[J]. Physica A: Statistical Mechanics and its Applications, 2011, 390(4): 719-730.

[14] Newman M E J. Mixing patterns in networks[J]. Physical Review E, 2003, 67(2): 026126.

[15] Darong Lai, Hongtao Lu, Christine Nardini. Enhanced modularity-based community detection by random walk network preprocessing[J]. Physical Review E, 2010, 23(2): 066118.

[16] Mantegna R N. Hierarchical structure in financial markets[J]. The European Physical Journal B-Condensed Matter and Complex Systems, 1999, 11(1): 193-197.

[17] Bonanno G, Caldarelli G, Lillo F, et al. Networks of equities in financial markets[J]. The European Physical Journal B-Condensed Matter and Complex Systems, 2004, 38(2): 363-371.

[18] Chi K T, Liu J, Lau F C M. A network perspective of the stock market[J]. Journal of Empirical Finance, 2010, 17(4): 659-667.

[19] Onnela J P, Chakraborti A, Kaski K, et al. Dynamic asset trees and Black Monday[J]. Physica A: Statistical Mechanics and its Applications, 2003, 324(1): 247-252.

[20] Onnela J P, Chakraborti A, Kaski K, et al. Dynamics of market correlations: Taxonomy and portfolio analysis[J]. Physical Review E, 2003, 68(5): 056110.