2008 International Symposium on Nonlinear Theory and its Applications
NOLTA'08, Budapest, Hungary, September 7-10, 2008

Winner-Take-All Correlation-Based Complex Networks for Modeling Stock Market and Degree-Based Indexes

Chi K. Tse*, Jing Liu† and Francis C. M. Lau

Department of Electronic and Information Engineering, The Hong Kong Polytechnic University, Hong Kong
†Also with State Key Laboratory for Software Engineering, Wuhan University, Hebei, China
*Correspondence: cktse@eie.polyu.edu.hk

Abstract— **Complex networks are constructed to study correlations between the closing prices for all US stocks that were traded from July 1, 2005 to August 30, 2007. The nodes are the stocks, and the connections are determined by cross correlations of the variations of the stock prices and price returns within a chosen period of time. Specifically, a winner-take-all approach is used to determine if two nodes are connected by an edge. The network thus formed is a full network of stock prices giving full information about their interdependence. We find that the distribution of the number of connections follows a power law. Such power-law distribution is also found in several variations of complex networks formed by considering price returns and trading volumes. The results from this work clearly suggest that the variation of stock prices are strongly influenced by a relatively small number of stocks. We propose a new approach for selecting stocks for inclusion in stock indices and compare it with existing approaches.**

I. Introduction

Fluctuations of stock prices are not independent, but are highly inter-coupled with strong correlations with the business sectors and industries to which the stocks belong. Recently, analyses based on network models have been proposed for studying the correlations of stock prices [1]–[6]. The usual approach involves a procedure of finding correlation between each pair of time series of stock prices, and a subsequent procedure of constructing a network that connects the individual stocks based on the levels of correlation. The resulting networks are usually very large and their analysis is rather complex. In much of the previous work, networks of relatively small size were constructed [6]–[8], and specific filtering processes were applied to further reduce the complexity. In particular, the method of Minimal Spanning Tree (MST) has been used for filtering networks, resulting in simpler forms of graphs that can facilitate analysis. The MST reduction is a topology based approach, which removes edges drastically by retaining only those that fit the MST criterion. With reduced complexity, Vandewalle *et al*. [2] observed a scalefree degree distribution in MST filtered networks of US stock prices. The topological change in the MST structure of networks of US stock prices has also been studied by Onnela *et al*. [6] who found variation in the value of the power-law exponent of the scalefree degree distribution of the MST filtered networks for "business as usual" and "crash" periods. Notwithstanding the reduction of complexity by introducing MST filtering to correlation-based networks, essential information about the internal structure is inevitably lost. In order to retain more information about the networks, less drastic filtering may be applied, e.g., using Planar Maximally Filtered Graph (PMFG), as proposed by Tumminello *et al*. [7]. However, both MST and PMFG suffer substantial loss of information as edges of high correlations are often removed while edges of low correlations are retained just because of their topological conditions fitting the topological reduction criteria. The usefulness of such MST or PMFG filtered networks is thus greatly reduced, especially in respect of their ability to identify the levels of correlation among stock prices.

In this paper, we consider a full network of correlation-based connections which retains all information of the internal structure that reflects the interdependence of the stock prices. The calculation of cross correlation values is similar to that adopted in Onnela *et al*. [8], but here we use a winner-take-all approach in establishing edges of the network, which makes binary decision on connecting two stock prices according to the truth value of their cross correlation being larger than a threshold value. Specifically, we examine the closing prices of 19807 stocks (out of 51835) which were traded each trading day from July 1, 2005, to August 30, 2007. Our aim is to construct networks that connect stock prices having similar variation profiles over a given period of time. Basically we examine the time series of the daily stock prices and establish connections between any pair of stocks. If the *cross correlation* of the time series of the daily stock prices of two stocks is greater than a threshold (e.g., 0.9), we consider that the two stocks are "connected". This simple winner-take-all criterion for establishing connections can also be applied to daily price returns, daily trading volumes, etc. in addition to daily closing prices. For instance, when cross correlation is taken between two time series of price returns, a different network can be formed. We will show in this paper that the full networks of stock prices, price returns and volumes are scalefree, and will report the power-law exponents along with a number of network parameters found from the US stocks that were traded in the period stated above.

Traditionally, stock market indexes are used to reflect about the market variations and the levels of market capitalization [9]–[12]. Commonly used indexes are the Standard & Poor 500 Index, Dow Jones Indexes, and Nasdaq Indexes. Because power-law distributions have been found in the stock prices, we know that a small number of stocks are having strong influence over the entire market, and we therefore propose that

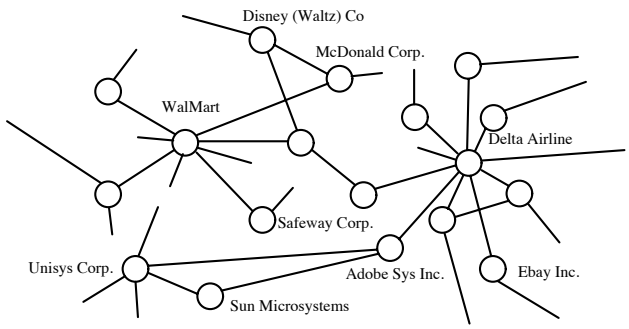

Fig. 1. Illustration of network of stock prices

stocks corresponding to nodes of high degrees can be used to compose a new index that can naturally and adequately reflect the market variation. We will evaluate the correlation of this new index with other existing indexes.

We will begin with a quick review of scalefree networks in Section II. We will then introduce the connection criterion and the network construction procedure in Section III. Results will be presented in Section IV, and some conclusions will be drawn in Section VII.

II. Review of Complex Networks

The study of complex networks in physics has aroused a lot of interest across a multitude of application areas. A key finding is that most networks involving man-made couplings and connection of people are naturally connected in a scalefree manner, which means that the number of connections follows a power-law distribution [13]. Scalefree power-law distribution is a remarkable property that has been found across of a variety of connected communities [14]–[17] and is a key to optimal performance of networked systems [18].

A network is usually defined as a collection of "nodes" connected by "links" or "edges" [14]. If we consider a network of stock prices, then the nodes will be the individual stocks and a link between two nodes denotes that the two stocks being connected display some "similarity". The number of links emerging from and converging at a node is called the "degree" of that node, usually denoted by k. So, we have an average degree for the whole network. The key concept here is the distribution of k. This concept can be mathematically presented in terms of probability density function. Basically, the probability of a node having a degree k is $p(k)$, and if we plot $p(k)$ against k, we get a distribution function. This distribution tells us about how this network of stock prices are connected. Recent research has provided concrete evidence that networks with man-made couplings and/or human connections follow power-law distributions, i.e., $p(k)$ vs k being a straight line whose gradient is the characteristic exponent [16]–[17]. Such networks are termed *scalefree networks*.

III. Network Construction and Winner-Take-All Connection Criterion

We consider a network of US stock prices of 19807 nodes. Each node corresponds to one of the stocks traded between July 1, 2005 to August 30, 2007. An illustration is shown in Fig. 1. For each pair of stocks (nodes), we will evaluate the cross correlation of the time series of their daily stock prices, daily price returns and daily trading volumes.

Let $p_i(t)$ be the closing price of stock i on day t and $v_i(t)$ be the trading volume of stock i on day t. Then, the price return of stock i on day t, denoted by $r_i(t)$, is defined as

$$r_i(t) = \ln\left[\frac{p_i(t)}{p_i(t-1)}\right] \tag{1}$$

Suppose $x_i(t)$ and $x_j(t)$ are the daily prices or price returns or trading volumes of stock i and stock j, respectively, over the period $t = 0$ to $N-1$. We now compare the two time series with no relative delay. In other words, x_i and x_j are compared from $i = 0$ to $N-1$ with no relative time shift. The cross correlation between x_i and x_j with no time shift is given by [19]

$$c_{ij} = \frac{\sum_t [(x_i(t) - \overline{x_i})(x_j(t) - \overline{x_j})]}{\sqrt{\sum_t (x_i(t) - \overline{x_i})^2}\sqrt{\sum_t (x_j - \overline{x_j})^2}} \tag{2}$$

where $\overline{x_i}$ and $\overline{x_j}$ are the means of the time series and the summations are taken over $t = 0$ to $N-1$.

In defining our criterion for connecting a pair of nodes, we need a threshold value for the cross correlation. Since cross correlation is a measure of similarity and its value is between 0 and 1, we simply choose a positive fractional value as the threshold. Suppose the threshold is ρ. Then, the connection criterion for stock i and stock j is

$$c_{ij} > \rho. \tag{3}$$

IV. Measured Network Parameters

We begin with relatively large values of ρ as our objective is to construct stock networks that reflect connections of highly correlated stock price time series. The total duration of the data is 564 trading days (from July 1, 2005 to August 30,2007). It is found that the degree distributions display scalefree characteristics when ρ is sufficiently large. Applying least mean square method with data points in the straight line segment of the log-log degree distribution plots, the power-law exponent is found to vary between 1 and 3. We also calculate the mean fitting error to examine the fitness of the power-law distribution over the data points. In addition, we have calculated the number of connections L, average shortest length s, average clustering coefficient C, average degree K, and the power-law exponents. Tables I, II and III show the results for networks based on closing prices, price returns and trading volumes, respectively. Fig. 2 illustrates the power-law degree distribution for $\rho = 0.9$.

For ρ below about a certain value, the power law distribution becomes blur, i.e., the mean fitting error increases. The networks thus constructed do not show clear scalefree

TABLE I

NETWORK PARAMETERS FROM US STOCK NETWORKS CONSTRUCTED FROM DAILY CLOSING PRICES USING A WINNER-TAKE-ALL CONNECTION CRITERION.

Parameters	$\rho = 0.85$	$\rho = 0.90$	$\rho = 0.95$
Number of Nodes N	19807	19807	19807
Number of connections L	4652650	1495250	143181
Average shortest length s	3.375	3.954	4.995
Diameter D	16	18	30
Average clustering coefficient C	0.421	0.302	0.148
Average degree K	469.80	150.98	14.46
Power-law exponent γ	0.778	1.075	0.992
Mean fitting error	6.26e-7	4.26e-7	1.65e-7

TABLE II

NETWORK PARAMETERS FROM US STOCK NETWORKS CONSTRUCTED FROM DAILY PRICE RETURNS USING A WINNER-TAKE-ALL CONNECTION CRITERION.

Parameters	$\rho = 0.70$	$\rho = 0.80$	$\rho = 0.90$
Number of Nodes N	19807	19807	19807
Number of connections L	15785	6675	2359
Average shortest length s	2.946	2.290	2.043
Diameter D	20	7	8
Average clustering coefficient C	0.104	0.058	0.238
Average degree K	1.594	0.674	0.238
Power-law exponent γ	2.019	3.067	2.920
Mean fitting error	16.07e-5	8.29e-5	2.78e-6

TABLE III

NETWORK PARAMETERS FROM US STOCK NETWORKS CONSTRUCTED FROM DAILY TRADING VOLUMES USING A WINNER-TAKE-ALL CONNECTION CRITERION.

Parameters	$\rho = 0.70$	$\rho = 0.80$	$\rho = 0.90$
Number of Nodes N	19807	19807	19807
Number of connections L	256046	167340	96203
Average shortest length s	4.542	4.927	7.165
Diameter D	21	19	19
Average clustering coefficient C	0.260	0.194	0.140
Average degree K	25.854	16.897	9.714
Power-law exponent γ	1.374	1.285	1.5933
Mean fitting error	1.33e-5	2.56e-6	2.50e-7

characteristics. This result should be expected since with small ρ, the network tends to be randomly connected, and in the extreme case of $\rho = 0$, the network is fully connected.

V. DISCUSSIONS

The properties of the stock networks constructed on the basis of cross correlation of stock prices are dependent upon the choice of the threshold ρ. We generally observe that the total number of connections increases with decreasing ρ, and as ρ approaches 0, the network becomes fully connected, as expected. The average shortest distance and the diameter decrease with ρ, while the clustering coefficient increases with decreasing ρ. The power-law degree distribution holds for large ρ, and becomes blur as ρ decreases, which is again consistent with the fact that the network becomes effectively more fully connected as ρ decreases.

We are particularly interested in the case where ρ is high as the network so formed would connect stocks of closely

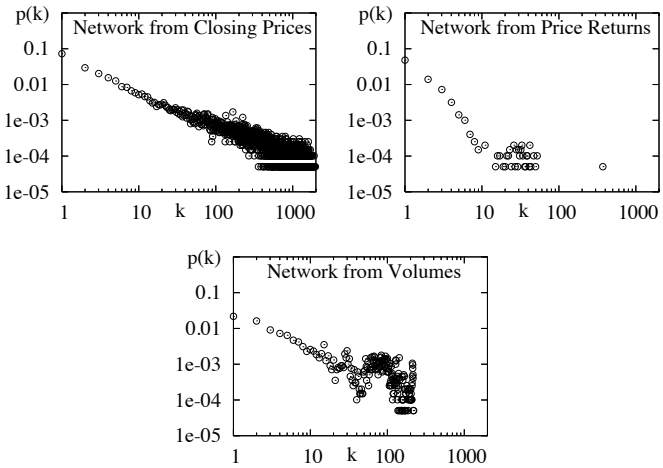

Fig. 2. Scalefree degree distribution of networks formed by a winner-take-all connection criterion with $\rho = 0.9$ applied to cross correlation of daily closing prices (upper left panel); daily price returns (upper right panel); and daily trading volumes (lower panel).

TABLE IV

CROSS CORRELATIONS BETWEEN NEW DEGREE-BASED INDEXES AND OTHER INDEXES.

	Degree-based Indexes (Closing Price Network)	(Price Return Network)
Dow Jones	0.9849	0.9753
S&P500	0.9771	0.9774
Nasdaq Composite	0.8985	0.8998

resembling daily price fluctuations. As we have shown earlier, the stock network is scalefree and displays clear power-law degree distributions. Thus, we may conclude that stocks having close resemblance with a large number of other stocks are relatively few. This transpires that the stock market is essentially influenced by a relatively small number of stocks, and hence we may introduce an index that reflects on the performance of the stock market based on a small number of stocks that have a relatively high number of connections. In other words, an index can be defined by the stocks of high degrees.

VI. DEGREE-BASED INDEXES

From the above winner-take-all correlation-based networks, we can identify stocks that have the highest degrees. These stocks have the largest numbers of connections with themselves and other stocks in the market. On the basis of the top 10% most highly connected stocks, we select those whose share information is fully available for the period from July 1, 2005 to August 30, 2007. New indexes are computed using the market capitalization formula [10], i.e.,

$$\text{Index} = \frac{\sum_i [\text{price}_i \times \text{number of shares}_i]}{\text{total market value of stocks during base period}} \tag{4}$$

In selecting the network, we choose ρ that gives about 500 stocks out of the top 10%. For the network based on

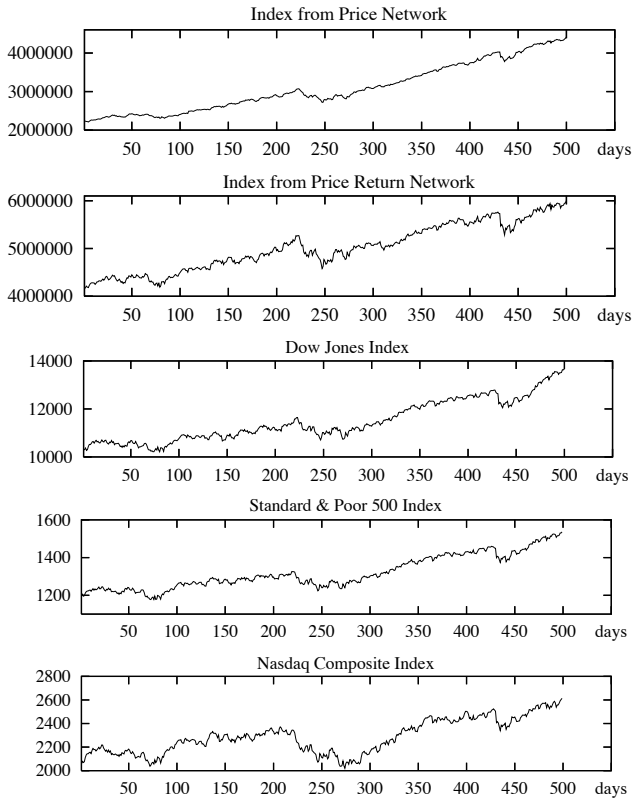

Fig. 3. Comparison of new degree-based indexes based on the closing price network with (first upper panel) and price return network (second upper panel), with Dow Jones index, S&P500 index and Nasdaq Composite index. The new indexes are computed using (4) and from the top 10% of most highly "connected" stocks.

closing prices, we choose $\rho = 0.9$, and for the network based on price returns, we choose $\rho = 0.5$. From the top 10% highly connected stocks, we get 330 stocks with full share information for the closing price network, and 486 stocks with full share information for the price return network. These stocks will be used to compute indexes, as mentioned above. Moreover, to make the new indexes fall within comparable range of other indexes, we use a different normalizing divisor. Fig. 3 shows the time series of the new degree-based index, Dow Jones index, Standard & Poor 500 index, and Nasdaq Composite index for the 564 trading days from July 1, 2005 to August 30, 2007. The cross correlations between the new index and other existing index are shown in Table IV.

It is worth noting that the new indexes defined in terms of highly connected stocks are fundamentally different from the commonly used ones which reflect performance of stock markets on the basis of stocks selected from different business and industrial sectors. For instance, among the 500 stocks used in the S&P500 index, only 16 overlap with those used in our degree-based index from the closing price network, and only 64 overlap with those used in our degree-based index from the price return network.

VII. Conclusion

Complex networks have been constructed for 19807 US stocks (all the US stocks that were traded each trading day from July 1, 2005 to August 30, 3007). The construction procedure is based on connecting any two stocks whose daily price time series are "similar" in terms of cross correlation evaluated over a period of time. For the first time, full network data of all US stocks traded each trading day over a 2-year period have been reported. It has been found that the networks formed using high cross correlation as the connection criterion are scalefree. Some network parameters have been calculated. The results suggest that a relatively small number of stocks are exerting much of the influence over the majority of stocks. New indexes may be defined based on market capitalization of a relatively small number of highly connected stocks.

Acknowledgment

This work was supported by Hong Kong Polytechnic University Research Project 1-BBZA.

References

[1] R. N. Mantegna, "Hierarchical structure in financial markets," *Euro. Phys. J. B,* vol. 11, pp. 193–197, 1999.
[2] N. Vandewalle, F. Brisbois, and X. Tordoir, "Self-organized critical topology of stock markets," *Quantit. Finan.,* vol. 1, pp. 372–375, 2001.
[3] G. Bonanno, G. Caldarelli, F. Lillo, S. Micciché, N. Vandewalle, and R. N. Mantegna, "Networks of equities in financial markets," *Euro. Phys. J. B,* vol. 38, pp. 363–371, 2004.
[4] G. Bonanno, F. Lillo, and R. N. Mantegna, "High-frequency cross-correlation in a set of stocks," *Quantit. Finan.,* vol. 1, pp. 96–104, 2001.
[5] G. Bonnanno, 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, 046103, 2003.
[6] J.-P. Onnela, A. Chakraborti, and K. Kaski, "Dynamics of market correlations: taxonomy and portfolio analysis," *Phys. Rev. E,* vol. 68, 056110, 2003.
[7] M. Tumminello, T. Aste, T. di Matteo, and R. N. Mantegna, "A tool for filtering information in complex systems," *Proc. National Academy of Sciences USA,* vol. 102, no. 3, pp. 10421–10426, July 2005.
[8] J.-P. Onnela, K. Kaski, and J. Kertesz, "Clustering and information in correlation based financial networks," *Euro. Phys. J. B,* vol. 38, pp. 353–362, 2004.
[9] Standard and Poor's, http://www2.standardandpoors.com/
[10] "How is the value of S&P 500 calculated?" *Investopedia,* http://www.investopedia.com/ask/answers/05/sp500calculation.asp
[11] Dow Jones Indexes, http://www.djindexes.com/
[12] Nasdaq Composite Index, http://dynamic.nasdaq.com/dynamic/composite_0.stm
[13] A.-L. Barabasi and R. Albert, "Emergence of scaling in random networks," *Science,* vol. 286, pp. 509–512, October 1999.
[14] S. H. Strogatz, "Exploring complex networks," *Nature,* vol. 410, pp. 268–276, March 2001.
[15] M. E. J. Newman, "Scientific collaboration networks I: Network construction and fundamental results," *Physical Review E,* vol. 64, pp. 016131-1-8, 2001.
[16] G. Csanyi and B. Szendroi, "Structure of a large social network," *Physical Review E,* vol. 69, pp. 036131-1-5, 2004.
[17] G. Ravid and S. Rafaeli, "Asynchronous discussion groups as small world and scale free networks," *Peer-Reviewed Journal on the Internet,* vol. 9, no. 9, 2004.
[18] X. Zheng, F. C. M. Lau, and C. K. Tse, "Study of LPDC codes built from scale-free networks," *Proc. Int. Symp. Nonlinear Theory and Its Applications,* Bologna, Italy, pp. 563–566, September 2006.
[19] J. Cohen, P. Cohen, S. G. West, and L. S. Aiken, *Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences,* (3rd ed.) Hillsdale, NJ: Lawrence Erlbaum Associates, 2003.