Network Analysis of the Stock Market

Wenyue Sun, Chuan Tian, Guang Yang

Abstract

In this study, we built a network for the US stock market based on the correlation of different stock returns. Community detection techniques were then applied to the constructed correlation network. The resulting communities were consistent with the identified market sections using Standard Industrial Classification code, which demonstrates that performances of public stocks within the same sector tend to have similar patterns. Furthermore, we used an open-source network analysis and visualization software, Gephi, to generate visualizations of the return correlations among various public stocks. The visualization results offer a very intuitive way to look at the overall correlation structure of different public stocks and identify key market segments, which could be very useful for real practice such as market monitoring.

For network applications, We first looked at the US credit crisis spreading represented by the spread of negative stock performances between Jul/2007 and Feb/2009. Stock performance were classified into three categories based on the rate of return (-20%, 0%, and 20%), then a sequence of snapshots of the stock network, colored by the three defined categories, show the cascading behavior of stock performances. We observed that the cascading starts from some stocks in different communities, and then spread from the initially infected (negative return) stocks. This observation indicates that the established stock network could potentially be very useful for stock market performance prediction from macroeconomic factors.

Then we investigated the application of network in portfolio management. Traditional approaches of portfolio management often rely on certain statistical properties, such as expected return and price variance. These properties, however, generally represent the local behavior of the stocks and are thus not able to represent the stock characteristics in terms of whole stock market. One advantage of creating a network characterizing different stocks within a market is that some important global properties of the stock within the network can be extracted, such as degree centrality, betweenness centrality, and closeness centrality. In this work, we did some preliminary studies of creating investment portfolios using the network properties of different stocks, and then compare the performance with some standard market index such as NASDAQ and S&P 500. Results show that with proper weights given to the top centrality nodes (i.e. stocks), we could outperform the S &P 500 for certain periods.

1 Introduction

Network analysis is popular to describe the characteristics or behaviors of complex networks. Recently there has been some research conducted to model the stock market using networks. The motivation is that the performances of certain stocks are often correlated, either because of the general market direction or the cyclicity of the same segments of the market.

To model the stock market using network analysis, different stocks are represented as different nodes. However, defining the interaction, or creating edges, between different nodes is rather non-intuitive, unlike some physical networks, such as friendship network, in which interaction between different nodes can be defined explicitly. A traditional way to create edges between different nodes for stock market is to look at the correlations of some defined attributes (e.g., trading volume, net return.) over a selected time frame. When the correlation is larger than some predefined threshold value (e.g., 0.7), then we think this is an edge connecting the two stocks (nodes).

Previous work, will be described in detail in Sect. 2, primarily studied some standard characteristics of constructed stock market correlation network, such as degree distribution, some centrality distributions, and average shortest length. However, to the best of our knowledge, the investigation of using stock market network analysis as a practical tool is very limited. In our work, the primary focuses will be on the practical side. One advantage of network analysis is that it enables people to understand the whole network via multiple ways of visualizing the constructed network. With the increased popularity of network analysis, many software were wrote that enables nice network visualization.

In this work, we will show two different applications of stock market network visualization. We first apply some community detection techniques to group stocks into different network sectors, and then visualize the network after grouping with different node colors representing the stock sectors classified with Standard Industrial Classification (SIC) code. This type of visualization offers an intuitive way of looking at whether the stock SIC code, which is determined from the physical functionality of the stock, is indicative of stock performance (for example, stocks with same SIC code might tend to behave similarly in the stock market). Another application investigated of visualization is to look at the stock market behavior before and during the market crisis. We studied this by looking at a dynamic evolution of the network with node colored by the stock returns over fixed range of time period, which allows to looked at how the stock crisis spread over the network. A proper visualization of the crisis spreading

behavior might potentially be very applicable in real practice to predict the market crisis and help reduce the risk of investment.

One main driving reason to study stock market is to optimize investment portfolio. Traditional approaches were primarily based on the calculation of variance and expected return of selected portfolio. A pareto front is normally generated with expected return against variance (which indicates the portfolio risk in general), then a portfolio is selected based on the risk preference of the investor. The performance of this type of approach heavily depends on how consistent the stock performance (e.g., return, volatility) behaves chronologically. Thus for the time period when there is no clear consistency, the "optimized" portfolio could end up being inferior. Network analysis, on the other hand, enables to provide some unique attributes for each node (stock) (i.e., degree centrality, betweenness centrality, and closeness centrality in this work). These attributes represent the "importance", defined by different metrics, of stock throughout the whole network. By assuming that the network structure is more likely to be consistent chronologically, these attributes are thus very likely to be consistent over time. In this work, we investigated several portfolio selection based on these attributes and compare the performance with the most well-known market index (i.e., NASDAQ and S&P500).

In the following section, we will provide a very brief literature review that summarize works been done related to network analysis in stock market. Due to the limit of space, discussions of each work will not be in much detail.

2 Literature Review

Past studies about network analysis for stock market can be classified into three categories: (1) applying network analysis techniques for different markets and analyze the topological characteristics of each market[2, 4]; (2) propose different correlation metric analysis among various stock markets to suggest different definitions of edges between stocks and study the impact on the network using different edge definitions[7, 11]; (3) stocks selection for portfolio management using the information from the network analysis and benchmark the portfolio performance against indexes[14].

2.1 Edge Definition

Approach to construct the edges of stock market network is not unique. In the current literature, multiple measures were investigated to construct the edges between nodes: zero-lag correlation[7], detrended covariance [10], time-lag correlations of prices changes over a certain period of time [11]. Besides price changes, there have been studies extending the measure to price-volume cross correlation [9] or only volume changes, as transaction volume is a good liquidity measure of the company.

2.2 Network Properties

Studies have covered both emerging and mature markets. Authors claim that understanding the topological properties can help to understand correlation patterns among stocks, thus providing guidance for risk management [7]. Topological properties often of interest include degree distribution, clustering and component structure. In this subcategory study, usually only one correlation measure is proposed to establish the connections between nodes. In the introduction session of [7], the author covered a wide range of previous studies in this category.

2.3 Network Dynamics

One interesting study covers the spreading of terrible stock performance due to the credit crisis between August 2007 to October 2008 [13]. The study is based on the a minimal spanning tree of the stock network for stocks in the S&P 500 and the NASDAQ-100 indices. The author concludes that, based on a short empirical investigation, the losses in US stock markets, following a cascade or epidemic flow-like model along the correlations built upon historical prices of various stocks. In this study, we propose to test its idea and use the dynamic behavior to examine the validity of our stock network. We would also extend its applicability to predict future stock performance or possible cascading among energy stocks caused by the recent big drop of oil price.

2.4 Portfolio Management

Network analysis was also applied for stock market portfolio management. In [14], the selection of stocks for inclusion in a stock index was studied. They compared the selected index from network analysis with existing indexes (i.e., Dow Jones Index, Standard & Poor 500 Index, Nasdaq Composite Index). Strong correlations between the selected index with existing indexes were found, even though the resulting composition of the indexes are very different. This work inspires us to look at the application network analysis for portfolio management. Many works studied the portfolio management by analyzing the cross-correlation matrix of stock returns [3, 12], however, direct application of networks analysis, as far as we know, still represents a very new and exciting research area.

2.5 Discussion

Although some promising results have been achieved for stock network analysis, the existing works have certain limitations.

The first limitation is the limited work on useful visualization of the constructed network. The primary focuses of previous work were normally on looking at some basis characteristics of constructed network, for instance, correlation distribution, degree distribution (e.g., whether it follows a power-law distribution), and clustering coefficient. These characteristics, however, do not offer people an intuitive way of improving understanding of the stock market, and are also not very helpful in providing direct guidance of market performance and investment.

The second limitation exists in the strategy of portfolio management. The existing work only mentions about diversifying the investment portfolio by choosing less correlated stocks, but does not provide a quantitative approach to achieve better portfolio. We think it would be very interesting to establish such kind of quantitative analysis by using the characteristics of the network (e.g. centrality). In particular, we can optimize the return function using network features in a machine learning framework.

3 Problem Definition

Based on the background information and literature review, the problem to be explored in this work is defined as the following:

Network visualization: community detection. We will apply community detection algorithm on the stock network, and generated corresponding visualization results. The nodes will be colored based on the market sector it belongs to, and then we'll look at the consistency between detected communities and classified market sectors.

Network visualization: cascading of market crisis. We will construct a network using stock data during 2006-Jul-1 to 2007-Jun-30. We classify the node color into three different groups based on the rate of return, and then create a dynamic visualization of the constructed network with node color changes with time, but not any other properties (such as node size and network structure). This allows us to visualize and investigate the spreading behavior of market crisis.

Network utilization: portfolio management. We will construct a network using stock data during 2012-Jan-01 to 2012-Dec-31, and then compute the degree centrality, betweenness centrality, and closeness centrality for each node. These centrality attributes will be used to generate multiple portfolios. The performance of generated portfolios will be compared with S&P500 index.

4 Methodologies

In this section, the source, treatment, and summary of stock data are described. The the approach to construct the network is introduced.

4.1 Data Summary

The data used in this paper are obtained from Center for Research in Security Prices (CSRP), WRDS [1]. We looked at the ticker, header standard industrial classification code (HSICCD), price, share volume, holding period return, number of shares outstanding and delisting code for the all available stocks in U.S.. HSICCD is used in this work to classify the stocks into different sectors which will be used in the visualization part. The table below listed the summary of the daily stock data during 2007-Jul-01 and 2008-Jul-1, which is the main data set we looked at in this work. The calculation of market capital (Mkt cap) in this work is introduced in Sect. 4.3.

Table 1: Daily stock data during 2007-Jul-01 and 2008-Jul-01

Parameter	Number
All stocks	7580
Delisted stocks	547
IPO stocks[1]	777
Stocks with Mkt cap over 1B$	2440
Stocks with Mkt cap over 10B$	494
Daily records	253

Because delisted and IPO stocks do not full span the whole time period, the cross-correlation computation of the stock returns, which will be described in the following section, can only rely on incomplete daily records during the time range considered. Thus it might create bias to compare the cross-correlation of those stocks with those cross-correlation that are calculated based on a full daily record basis. Thus for the correlation network in this work, we do not consider delisted and IPO stocks, which results in the two networks we looked at. One with 465 nodes (non-delisted and non-IPO stocks with market capital over 10B$) and another 2189 (non-delisted and non-IPO stocks with market capital over 1B$). In the next section, we will describe the criteria to construct the edges and thus the correlation network.

4.2 Network Construction

The basic methodology is similar as described in [14]. Some further definitions are introduced.

Let $x_i(t)$ be the time series data we are interested in for stock i at time t. The cross-correlation between the time series data for stock i and j is calculated as:

$$c_{ij} = \frac{\sum_t \left[\left(x_i(t) - \bar{x_i}\right)\left(x_j(t) - \bar{x_j}\right)\right]}{\sqrt{\sum_t \left(x_i(t) - \bar{x_i}\right)^2}\sqrt{\sum_t \left(x_j(t) - \bar{x_j}\right)^2}}, \tag{1}$$

where the summation acts on the selected time range to compute the correlation, and in this case it is from 2007-Jul-01 to 2008-Jul-01. The time t is discrete with step size of one day. $\bar{x_i}$ represents the mean value of $x_i(t)$ over the time range computing the correlation.

We consider constructing the correlation stock network, and an edge between two nodes (stocks) exist if and only if the cross-correlation between the two nodes are higher then a specified threshold value θ. Then the adjacent matrix A for the network is constructed as:

$$A_{ij} = \begin{cases} 1 & \text{if } c_{ij} >= \theta \text{ and } i \neq j \\ 0 & \text{otherwise} \end{cases}$$

Let $p_{c,i}(t)$ be the closing price of stock i on day t, then the logarithmic change ("returns") of stock i at day t is defined as [14, 2]:

$$r_i(t) = \ln\left[\frac{p_{c,i}(t)}{p_{c,i}(t-1)}\right]. \tag{2}$$

In current work, the price return defined by Eq. 2 is used as the time series data of interest to compute the cross-correlation, and thus serve to construct the network.

For future work, we would like to consider the time lag effect of stock prices [11], which will result in a different way to constructing the cross-correlation network. The general form of the cross-correlation with time lag effect can be written as:

$$\tilde{c}_{ij}(\Delta t) = \frac{\sum_t \left[(x_i(t) - \bar{x_i})\,(x_j(t+\Delta t) - \bar{x_j})\right]}{\sqrt{\sum_t (x_i(t) - \bar{x_i})^2}\sqrt{\sum_t (x_j(t+\Delta t) - \bar{x_j})^2}}. \quad (3)$$

Considering the time lag effect makes sense because it is very likely that the influence of one stock to anther does not happen instantaneously. With this definition, we can construct a direct graph, in which the direct represents the direction of the influence. A proper use of this direct graph might help us to better understand and predict the cascading between of stock market, and thus help to make better portfolio management. Note that this time-lag cross-correlation defined by Eq. 3 will lead to an nonsymmetric matrix.

4.3 Calculation of Some Quantities

The construction and analysis of correlation network in this work involved the definition of stock market capital over a time rage (for filtering stocks noted in Sect. 4.1) and stock performance (or called as cumulative return as in Sect: 5.2).

Denote the market capital of stock i at time t as $V_{i,t}$, the market capital V_i over the time range considered (2007-Jul-01 to 2008-Jul-01) is defined as the maximum daily market capital (defined as the product of total share volume and stock price) ever been achieved, which can be written as:

$$V_i = \underset{t}{\operatorname{argmin}}(V_{i,t}) \quad t_s \le t \le t_e, \quad (4)$$

where t_s and t_e represent the starting and ending time of computing the cross-correlation shown in Eq. 1.

The performance of a stock at a particular time is used to look at the network cascading behavior in Sect. 5.2. The stock performance for stock i at a particular time t, $W_{i,t}$, in this work is calculated based on the capital return of that stock at that time, which can be expressed as:

$$W_{i,t} = \frac{V_{i,t} - V_{i,t_s}}{V_{i,t_s}}, \quad (5)$$

5 Results

5.1 Community Detection

The Standard Industrial Classification SIC is a system for classifying industries by a four-digit code, which used by government to classify industry areas. The four-digit code can be grouped into ten ranges, with each representing a different division, see **??** for detailed description of the different divisions. Figure 1 shows the visualization of the correlation network using data during 2007-Jan-01 and 2007-Jun-30. We applied the Fruchterman Reingold algorithm enabled in the Gephi network analysis software to generate the layout of the network, in which strongly connected stocks tend to form cliques , or in other words, network sections. This layout allows us to see the global correlation pattern of the stock market.

We then color different nodes using the associated SIC code (in total of ten different colors). The purpose of doing this is to see whether the classification of stocks based on the physical functionality is still representative in terms of looking at the stock performance correlation. We can see that, in a general sense, stocks within the same division based on the SIC code also tend to be correlated from the stock performance point of view. For example, BAC (Bank of America), C (Citi Bank), and JPM (J.P. Morgan) are known to within the finance division (and thus shown as the same color in Fig. 1), in the network, we can see that they also cluster into the same group (lower in the middle of the circle). However, there are also cases when the SIC code classify stocks into different divisions, but their market performance are shown to be in the same group from the network analysis results. For example, CVX (Chevron), XOM (ExxonMobile) and COP (ConocoPhillips) are often perceived as large oil producers by the general public. Their strong stock performance correlation also indicates that the market perceives these companies as the same type regardless of the different SIC code they possessed. According to SIC code, chevron is categorized as manufacturing company and ConocoPhillips is attributed to mining category.

5.2 Visualizing Market Crisis

We modeled a dynamic network showing the spread of bearish performance of stocks among the market during the financial crisis in 2008 (Fig. 2). Price return from Jan. 2007 to June. 2007 was used to established the network and color the node of stocks. Stocks with cumulative return larger than 0% are given green color, and stocks with cumulative return less than -20 % are given red color. The cumulative return is calculated using Eq. 5. Then the established network was tested on data from July 2007 to Jan 2009. It is observed that the negative return first started from some famous insurance companies, financial services companies and cyclical consumer goods companies, such as AIG(American International Group Inc), JPM(JPMorgan Chase & Co.), DFS(Discover Financial Services), MER (Merrill Lynch & Co.), C (Citigroup Inc), HD (Homedepot), M(Macy's) and SHLD(Sears Holdings Corp). The bearish performance then get quickly spread in these sectors. By the time of the end of Jan 2008, stock performances of a lot of department store companies has become negative more than -20 % (turning red in the network), such as LOW (Lowe's), KSS (Kohl's Corporation), JCP (J C Penny). It further influenced other cyclical consumer goods companies such as HOT (Starwood Hotels & Resorts Worldwide Inc),MAS(Masco Corp), and even GE(General Electric Company) because of the limited consumer spending power at the time of credit crisis. By the time of Apr 2008, almost every big company in the financial and cyclical consumer goods sector is in big trouble, resulting in less than -20 % stock performance. Because of the credit issue, companies have to sell

their stocks to raise cash to maintain operation of the company. Based on this observation, we could confer that the established network has correctly captured the performance connections between companies and is indicative of the dynamics of the industrial group that the companies lie within. For grouping purposes, we used stocks above 10 billion because we want to decrease the number of nodes and edges in the network so that we could clearly see the spreading process.

5.3 Portfolio Management

The investment philosophy for most funds is to catch up with and even beat the market. In order to achieve that goal, they need to select the stocks that are representative of the market. A common choice would be the index stocks, since their high market caps make them potentially impactful in the market. If we apply the similar logic to a stock network, an intuitive try would be to select the stocks with high centralities, which measure different aspects of the importance of a stock in the market. To start off, we assigned equal weights to degree centrality (C_d), betweenness centrality (C_b) and closeness centrality (C_c), and selected the top xx stocks with highest average centrality (C_{avg}):

$$C_{avg} = \frac{1}{3}C_d + \frac{1}{3}C_b + \frac{1}{3}C_c \tag{6}$$

Next we formulated the portfolio management into an optimization problem, where the weights on centralities are the decision variables and the return is the target:

$$max \quad r(\alpha_1 C_d + \alpha_2 C_b + \alpha_3 C_c) \tag{7}$$

From the existing work, we learn that network characteristics such as centrality are of great importance to well represent the stock market. Thus we propose to use centrality along with other network characteristics as the features to model the capital gain in a machine learning framework. Then a portfolio management can be implemented by optimizing the capital gain by selecting the stocks with the desired network characteristics.

6 Conclusions

There are mainly three conclusions of this work: 1. We constructed a network to model the stock market behaviors based on time-lag correlation of the stock return. 2. The network was visualized to compare the detected communities with industrial classification code, and to simulate the network cascading during financial crisis. 3. The constructed network was utilized for portfolio management, by selecting the stocks with high centralities.

References

[1] Center for research in security prices (crsp). wrds.

[2] V. Boginski, S. Butenko, and P. M. Pardalos. Statistical analysis of financial networks. *Computational statistics and data analysis*, 48(2):431–443, 2005.

[3] P. Gopikrishnan, B. Rosenow, V. Plerouand, and H. E. Stanley. Quantifying and interpreting collective behavior in financial markets. *Physical Review E*, 64(3):035–106, 2001.

[4] W. Q. Huang, X. T. Zhuang, and S. Yao. A network analysis of the chinese stock market. *Physica A: Statistical Mechanics and its Applications*, 2009.

[5] P. Potters J. Bouchaud L. Laloux, P. Cizeau. Random matrix theory and financial correlation. *International Journal of Theoretical and Applied Finance*, 3(3), 2000.

[6] Renaud Lambiotte, J-C Delvenne, and Mauricio Barahona. Laplacian dynamics and multiscale modular structure in networks. *arXiv preprint arXiv:0812.1770*, 2008.

[7] A. Namaki, A.H. Shirazi, R. Raei, and G.R. Jafari. Network analysis of a financial market based on genuine correlation and threshold method. *Physica A: Statistical Mechanics and its Applications*, 2011.

[8] M. E. Newman. Power laws, pareto distributions and Zipf's law. *Contemporary physics*, 46(5):323–351, 2005.

[9] B. Podobnik, D. Horvatic, A. M. Petersen, and H. E. Stanley. Cross-correlations between volume change and price change. *PNAS*, 106(52):22079–22084, 2009.

[10] B. Podobnik and H. E. Stanley. Detrended corss-correlation analysis: A new method for analyzing two non-stationary time series. *Physical review letters*, 100(8):084–102, 2008.

[11] B. Podobnik, D. Wang, D. Horvatic, I. Grosse, and H. E. Stanley. Time-lag cross-correlations in collective phenomena. *Europhysics Letters*, 2010.

[12] B. Rosenow, V. Plerou, P. Gopikrishnan, and H. E. Stanley. Portfolio optimization and the random magnet problem. *Europhysics Letters*, 59(4):500, 2002.

[13] R. D. Smith. The Spread of the Credit Crisis: View from a Stock Correlation Network. *Journal of Korean Physical Society*, 54:2460, June 2009.

[14] C. K. Tse, J. Liu, and C. M. Lau. A network perspective of the stock market. *Journal of Empirical Finance*, 2010.

[15] Wikipedia. Plagiarism — Wikipedia, the free encyclopedia, 2004.

Appendix

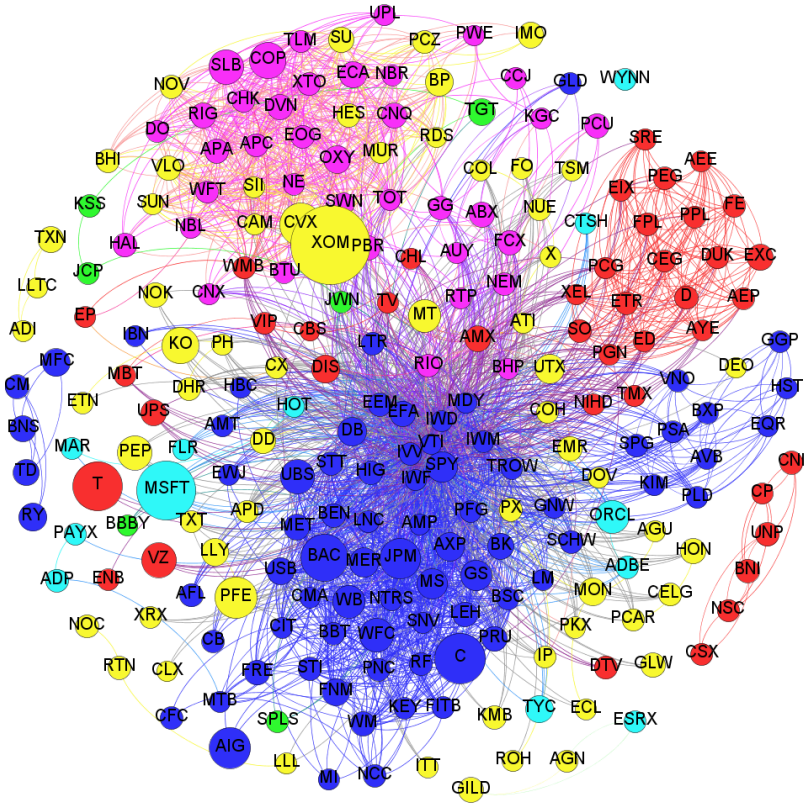

Figure 1: Visualization of the network of stocks based on correlation threshold of 0.6 using daily price return (log) during 01/01/2007 to 06/30/2007. Nodes are colored by SIC code (in total of ten different colors) commonly used in practice. Size of the nodes represent the maximum market cap during the period. Only stocks above 10 billion market capital are shown in this network.

(a) July 2007

(b) Jan 2008

(c) Apr 2008

(d) July 2008

(e) Oct 2008

(f) Jan 2009

Figure 2: Visualization of market crisis during 2007-Jul-01 to 2009-Feb-1. The network is constructed using stock data during 2007-Jan-01 to 2007-Jun-30. Node color are red (capital return less than -20%), white (capital return less than 0%,) and green (capital return larger than 20%). Only stocks above 1 billion market capital are shown in this network.

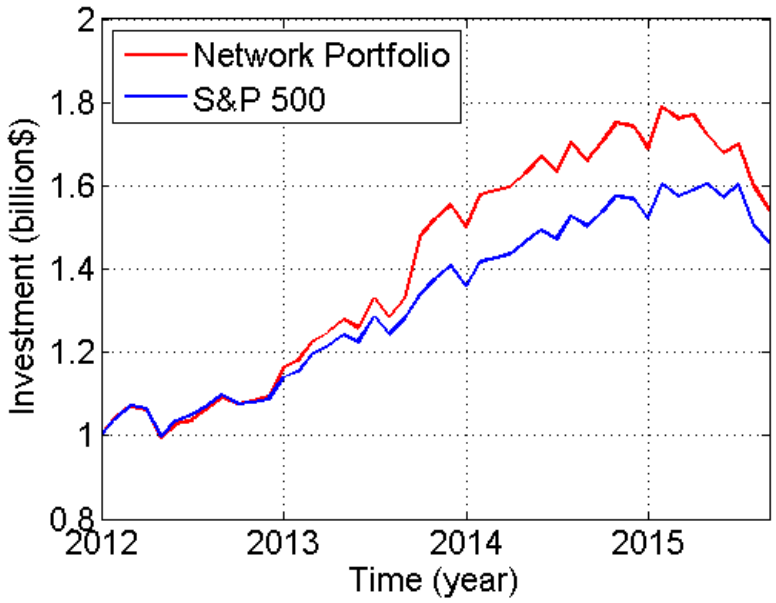

Figure 3: Comparing between selected portfolio performance and S&P 500 benchmark.