ANALYSIS OF STOCK MARKET USING COMPLEX NETWORKS AND MACHINE LEARNING

A DISSERTATION SUBMITTED TO THE UNIVERSITY OF MANCHESTER
FOR THE DEGREE OF MASTER OF SCIENCE
IN THE FACULTY OF SCIENCE AND ENGINEERING

2018

By
Lingjie Zhang
School of Computer Science

Contents

Abstract **7**

Declaration **8**

Copyright **9**

Acknowledgements **10**

1 Introduction **12**

1.1 Motivation 12

1.2 Objectives 13

1.3 Proposed methodology 14

1.3.1 Construction of the stock network 14

1.3.2 Logarithmic return of stock prices 14

1.3.3 Directed Watts–Strogatz small-world network 14

1.3.4 Directed random network 14

1.3.5 Topological properties of directed networks 14

1.4 Summary of results 14

1.5 Outline of the thesis 14

2 Background **15**

3 Pre-processing of data **17**

3.1 Data source 17

3.2 EIO (Economic Input-Output) table 17

3.3 Logarithmic return of stock prices 18

3.4 Correlation coefficient 18

4 Network Topological properties **19**
4.1 Degree centrality and strength centrality 19
4.2 Degree distribution and strength distribution 20
4.3 Average shortest path length 20
4.4 Betweenness centrality 20
4.5 Clustering coefficient 21
4.6 Efficiency 21
4.6.1 Global efficiency 21
4.6.2 Local efficiency 22
4.7 Assortativity and Degree Correlations 22
4.8 Modularity 23

5 Community detection **24**

6 Benchmarking networks generation **27**
6.1 Directed Watts-Strogatz small world network 27
6.2 Directed Erdős–Rényi random network 28

7 Empirical study and results **29**
7.1 Network construction 29
7.2 Analysis of the directed unweighted network 33
7.2.1 Power-law distribution 33
7.2.2 Small-world property 34
7.2.3 Community structure of the directed unweighted stock network 34
7.3 Analysis of the directed weighted network 39
7.3.1 Topologies 39
7.3.2 Analysis on the relationship between price return and betweenness centrality 39

8 Conclusions **43**

Bibliography **44**

A Example of operation **47**
A.1 Example input and output 47
A.1.1 Input 47
A.1.2 Output 48

A.1.3 Another way to include code 48

Word Count: 1626

List of Tables

7.1 Main properties of stock network, small-world network, and random network 33
7.2 Main topologies of conventional stock price network 35
7.3 Main topologies of weighted stock network 39

List of Figures

7.1 Transaction densities 30
7.2 Stock price return correlation coefficient distribution 30
7.3 31
7.4 Numbers of directed edges per EIO-threshold and correlation-coefficient-threshold 31
7.5 Transaction densities 32
7.6 Community structure of the 2016 US stock price return network. Five distinct communities are detected represented by different colours of nodes. The direction of edge is clockwise. The size of nodes and thickness of edges are related to the value of degrees and weights. The grey nodes do not belong to any communities and most of them have zero degree. 36
7.7 Community sole views of the directed stock network. Stock tickers are displayed for the sparsely distributed communities. 37
7.8 Stacked bar chart about the distribution of communities upon industrial sectors. Colours of stacks correspond to the colours of communities in figure 7.6 and figure 7.7, except the black stack indicating the nodes not belong to any communities. Sectors are arranged alphabetically. . 38
7.9 Out-degree distribution of stock network 39
7.10 Out-degree distribution and P-P plot of small-world network 40
7.11 Out-degree distribution and P-P plot of random network 41
7.12 Bivariate distributions with betweenness centrality 42

Abstract

ANALYSIS OF STOCK MARKET
USING COMPLEX NETWORKS
AND MACHINE LEARNING

Lingjie Zhang

A dissertation submitted to the University of Manchester
for the degree of Master of Science, 2018

Due to the complexity of financial market and the interconnect- edness and interdependencies of industry sectors in the economy, the price returns of each coupling stocks might have certain underlying economic link. Such behaviours can hardly be explained by tradi- tional financial models and theories. This project combines machine learning techniques, individual stock features, and empirical data of Industry Economic Accounts (IEAs) from Bureau of Economic Anal- ysis (BEA) in the US to predict Granger causality of coupling US stocks. Limited Granger causalities are calculated as a small sample set compared to the target date set. A directed weighted complex network (DWCN) is constructed by considering companies as nodes, correlations of abnormal stock returns (alpha) as weights of links, and predicted Granger causalities indicate directions of links. The generated DWCN is visualised and its topological properties, sta- bility, and effects on individual stocks and industries are researched in this paper. Suggestions towards financial market investment are provided based on the results of this research.

Declaration

No portion of the work referred to in this dissertation has been submitted in support of an application for another degree or qualification of this or any other university or other institute of learning.

Copyright

Acknowledgements

I would like to thank...

Contents

Chapter 1

Introduction

1.1 Motivation

Financial markets are complex systems, the interconnectedness and interdependencies of industry sectors in the economy are highly inter-coupled with strong correlations with stock price fluctuations, i.e., the price returns of each coupling stocks underlying certain economic link, e.g. two companies that manufacture similar products, or both in one supply chain. Such behaviours can hardly be explained by traditional financial models and theories.

During recent times, weighted but undirected complex network models have been applied to study the correlations of stock prices. Prevailing approach is using companies as nodes, and correlations between each pair of stock price series, return series, or fluctuation patterns as links, e.g., much of the previous researches have proved the represented complex networks of worldwide stock markets are scale-free and small-world [LLH07, CLL10].

Theoretically, directed complex network for stock market can be achieved therefore more potential information can be produced which is helpful for investment decision and financial market supervision. Conducting Granger causality test between stock pairs is straightforward but not feasible due to the heavy- precondition and time-complexity in programme. Compared to the large order of magnitude of total stock pairs, manually calculated Granger causalities are too less to be used as training samples. Enlightened by the recent work by Jean et al. [3] which uses transfer learning and noisy proxy information per- formed very well at predicting poverties, demonstrating that machine learning techniques is powerful to be applied in a setting with limited training data, so an exploration towards the directed network of stock market

is motivated in this project, combines machine learning techniques, transfer learning, individual stock features, and empirical data of Industry Economic Accounts (IEAs) from Bureau of Economic Analysis (BEA) in the US to predict Granger causality of coupling US stocks. Therefore, a directed weighted complex network (DWCN) is constructed by considering companies as nodes, correlations of abnormal stock returns (alpha) as weights of links, and predicted Granger causalities as the in- dications of directions of links.

The goal of this project is to reveal the Granger causality of price return series and utilise them into the topological analysis and visualisation of DWCN as so far no previous work has attempted to construct a directed network about stock price. In addition, suggestions for stock market are provided according to the results and findings.

The outline of this document is as follows. In Section 2, specific objectives and deliverables are listed. In Section 3, relevant previous researches on financial market, complex networks, and machine learning are reviewed. Section 4 de- scribes the analytical methods. Ethics and professional considerations and risk considerations are respectively discussed in Section 5 and 6. Section 7 describes project evaluation approaches and finally, planning Gantt chart is presented in Section 8.

1.2 Objectives

The goal of this project is to construct a directed complex network using economical industrial transaction data and stock price data to depict the US stock market by means of topological properties analysis, community detection and visualisation. Same-sized directed Watt-Strogats small-world network and random networks are generated for the purpose of comparison. This paper will explore whether the conclusions are consistent with the undirected complex network researches.

1.3 Proposed methodology

1.3.1 Construction of the stock network

EIO (Economic Input-Output) table

1.3.2 Logarithmic return of stock prices

Correlation coefficient

1.3.3 Directed Watts–Strogatz small-world network

1.3.4 Directed random network

1.3.5 Topological properties of directed networks

1.4 Summary of results

1.5 Outline of the thesis

The rest of this paper is organised as follows. Chapter 2 discusses the development of quantified financial analysis for stock markets and the application of complex theories of complex networks towards stock markets. The subsequent chapters of methodologies introduce the critical analytical methods implemented in the research by this paper. Detailed outcomes are then illustrated in Chapter 7. Finally, the findings and conclusions are discussed in Chapter 8.

Chapter 2

Background and state of the art

The revolutionary who pioneered the theory of portfolio, Markowitz [Mar52] proposed the mean-variance model and established a clear mathematical definition of the two vague concepts of risk and return. Sharp [Sha64] and Linter [Lin65] added two key assumptions based on the mean-variance model to enable the portfolio mean-variance valid, forming a capital asset pricing model (CAPM) with the support of economic theories. The CAPM believes that only non-dispersible systemic risks can be compensated off, while non-systematic risks can be eliminated by effectively decentralised investments. Investors could only assume systemic risks through decentralised investment. The systematic risk of a single security or portfolio can be characterised by beta, which represents the extent to which a single security or portfolio is affected by the overall market volatility.

Worldwide scholars had been actively conducting empirical tests towards the practicality of the CAPM then, while early results show that beta is able to explain return movements of stocks. However, in the late 1970s, some empirical studies upon the CAPM began to show that a large part of the changes in the stock returns cannot be explained by beta, with increasingly market return anomalies were found.

Researchers had proposed models and theories considering the individual stock features to explain. For instance, Fama and French [FF93, FF96] proposed a three-factor model based on the inter-temporal capital asset model (ICAPM) [Mer73] and the arbitrage pricing theory (APT) [CR76], which reveals a large part of the cross-section of the stocks' average return that cannot be explained by CAPM, can be explained using firm size, book-to-market equity ratio, and overall market return.

While traditional stock pricing models still capture limited forms of financial behaviour, the premises of standard financial theory contradict the modern notion of financial markets are complex systems [JJH+03], by which many statistical niceties such as stationarity no longer can be taken for granted. Recent researches have implemented the network theory to reveal the underlying factors of price movements. Huang et al. [HZY09] implemented the threshold method to build 's correlation network in China's A-Share stock market and studied the networks topological properties and topological stability. Namaki et al. [NSRJ11] utilised Random Matrix Theory (RMT) to specify the biggest eigenvector in the complex network of price correlations. Yu [Lon13] studied the evolution of gold price from a network perspective using the visibility network approach. Chopra and Khanna [CK15] developed a framework which associates the economic input–output (EIO) model with techniques for understanding interdependencies and interconnectedness in the economy of US, based on complex networks theory. Boginski et al. [BBP05] identified cliques and independent groups among stock networks. Chen et al. [CLSW15] studied the inter-stock and inter-industry effects towards stock returns based on the topological properties of a complex network of correlations.

In a nutshell, prevailing complex network approaches to analyse stock markets are almost all about investigating weighted or unweight but undirected networks. To the best knowledge of mine, no previous work has attempted to construct a directed network so far.

Chapter 3

Pre-processing of stock market and industrial data

3.1 Data source

This paper considers 1,418 stocks of listing US companies that were traded consecutively in the NYSE and NASDAQ stock market of US on the trading days from January 4, 2016 to December 30, 2016 and uses daily closing price during this period and the economical use table data from the Industry Economic Accounts (IEAs) of year 2016 in a summary-level of industrial sectors are collected from the official website of Bureau of Economic Analysis, US Department of Commerce [oEA18].

3.2 EIO (Economic Input-Output) table

The Bureau of Economic Analysis (BEA) in the US publishes Economic Input-Output (EIO) tables each year, which are the transaction matrices of all purchases and sales between sectors in a certain industry group level of a year, i.e. depict how industries provide input to, and use output from, each other to produce Gross Domestic Product (GDP).

This paper uses the use table of 2016. Among the transaction matrix **Z** there are Total Industry Input row **I** at the bottom and the Total Industry Output column **O** at the right are the statistics of total purchase by each sectors and total sales from each sectors respectively. Below are the equations for generating the matrices of normalised direct demand **A** and direct requirement **B**:

$$a_{i,j} = -log_{10}(z_{i,j}/I_j)^{-1} \tag{3.1}$$

$$b_{i,j} = -log_{10}(z_{i,j}/O_i)^{-1} \tag{3.2}$$

Moreover, certain threshold values θ_{DD} and θ_{DR} are specified and a directed edge can be added between stock i and stock j if the value of $a_{i,j}$ is greater than θ_{DD} or the value of $c_{i,j}$ is greater than θ_{DR}.

3.3 Logarithmic return of stock prices

Logarithmic return of a stock in this paper is calculated as the log of the close price of one day divided by the close price of the previous day, which is obtained from the following formula:

$$r_i(\tau) = lnP_i(\tau) - lnP_i(\tau - \Delta t) \tag{3.3}$$

As a proxy for the percentage change in the price, logarithmic return is symmetric and has mathematical conveniences for adding up or subtracting values on the log scale, which are useful for mathematical finance. Therefore, logarithmic return is the measure of price changes in this paper.

3.4 Correlation coefficient

The correlation coefficient between two stocks is considered in terms of the matrix $\mathbf{C}$, as the following equation shows:

$$c_{i,j} = \frac{\langle r_i r_j \rangle - \langle r_i \rangle \langle r_j \rangle}{\sqrt{(\langle r_i^2 \rangle - \langle r_i \rangle^2)(\langle r_j^2 \rangle - \langle r_j \rangle^2)}} \tag{3.4}$$

where r denotes the return and the bracket indicates a temporal average over the period. Additionally, a certain threshold value θ_{corr}, $0 \leq \theta_{corr} \leq 1$ is specified, and a directed edge is qualified to be linked between stock i and stock j if the value of $c_{i,j}$ is greater than or equal to θ_{corr}.

Chapter 4

Topological properties of directed complex networks

4.1 Degree centrality and strength centrality

The degree of a node k represents the number of its neighbours. In directed network, out-degree k_{out} is the number of edges which start from the given node and end at others, while in-degree k_{in} is the number of edges which end at the given node and start from others. Thus, there is relationship between k_{in} and k_{out}:

$$k = k_{in} + k_{out}. \tag{4.1}$$

As one of the most widespread measures to calculate network centrality, degree centrality of a node can be described as the number of direct links that relate to a specific node [**?**]. In terms of the directed stock price return network, this paper mainly focuses on the out-degree analysis on the nodes. Moreover, the strength centrality has generally been accumulated to the sum of weights of out-degrees to form the weighted networks. The equation of this measure is shown as bellow:

$$C_D^W(i) = \sum_j^N w_{ij} \tag{4.2}$$

where W represents the matrix of weighed adjacencies, and w_{ij} represents the weight of the link between node i and j.

4.2 Degree distribution and strength distribution

The degree distribution of stock price return network $p(k)$ can be defined as:

$$p_d(k) = \frac{N_k}{N}, \tag{4.3}$$

while N_k represents the number of nodes whose out-degree value is k. The distribution of strength has a similar definition:

$$p_s(w) = \frac{N_w}{N}, \tag{4.4}$$

while N_w represents the number of nodes whose strength value is w.

4.3 Average shortest path length

The average shortest path length of a directed network G is defined as the following equation:

$$l_G = \frac{1}{n(n-1)} \sum_{i,j \in V} d(i,j) \tag{4.5}$$

where V is the set of nodes of G.

4.4 Betweenness centrality

Other than strength, betweenness centrality [Fre77] can be used to determine the critical nodes among the entire network and to recognise the most associated firms in the chosen stock market. When it comes to weighted networks, betweenness centrality of a node is the sum of the weights in the fraction of all-pairs shortest paths that pass through this node, which can be described as the following equation:

$$C_B(v) = \sum_{s,t \in V} \frac{\sigma(s,t|v)}{\sigma(s,t)} \tag{4.6}$$

where V is the set of nodes, $\sigma(s,t)$ is the sum of weights of all-pairs shortest (s,t)-paths, and $\sigma(s,t|v)$ is the sum of weights of those paths passing through some node v other than s,t. If $s = t$, $\sigma(s,t) = 1$, and if $v \in s,t$, $\sigma(s,t|v) = 0$.

4.5 Clustering coefficient

Clustering coefficient is a measure of the degree to which nodes in a network tend to cluster together. Concerning the clustering coefficient of the complex networks, it is defined as:

$$C_i = \frac{2E_i}{(k_i(k_i-1))}, \tag{4.7}$$

where k_i is the degree of a given node v_i, E_i is the real existing edges among the nearest neighbour nodes of the given node v_i, and $k_i(k_i-1)/2$ means the maximum possible edges existing between its nearest neighbours of the node vi. Besides, the clustering coefficient of a node accounts for the extent to which the transmission relationship between the given node and its neighbours also exists between its neighbours, and the clustering coefficient may be given by:

$$C = \frac{3 \times number\ of\ triangles\ in\ the\ networks}{number\ of\ connected\ triples\ of\ nodes}. \tag{4.8}$$

This measure gives an indication of the clustering in the whole network, and can be applied to both undirected and directed networks.

4.6 Efficiency

Network efficiency measures how efficient for information being conducted and exchanged in the network, which can help to determine whether the objective network shows small-world property. There are global and local efficiencies that on the different scale sizes [LM01].

4.6.1 Global efficiency

Global efficiency quantifies the conduction and exchange of information through out the entire network. The global efficiency of network $\mathbf{G}$ is defined as:

$$E_{glob}(\mathbf{G}) = \frac{\sum_{i \neq j \in \mathbf{G}} \varepsilon_{ij}}{N(N-1)} = \frac{1}{N(N-1)} \sum_{i \neq j \in \mathbf{G}} \frac{1}{d_{ij}} \tag{4.9}$$

4.6.2 Local efficiency

The local efficiency evaluates the resistance of a network towards node i and quantifies the conduction and exchange of information among its neighbours. The local efficiency of node i in network $\mathbf{G}$ is defined as:

$$E_{loc}(G,i) = \frac{1}{N}\sum_{i\in\mathbf{G}} E_{glob}(\mathbf{G}_i) \tag{4.10}$$

4.7 Assortativity and Degree Correlations

The phenomenon of assortative [New02] mixing can be quantified by means of an assortative coefficient. Let E_{ij} be the number of edges in the network that connect a vertex of type i to one of type j, with $i, j = 1, ..., n$, then similar in spirit to the adjacency matrix for vertices, these edges can be represented in the form of an edge incidence matrix $\mathbf{E}$, with elements E_{ij}. A normalized mixing matrix is defined as follows:

$$\mathbf{e} = \frac{\mathbf{E}}{\|\mathbf{E}\|}, \tag{4.11}$$

where $\|\mathbf{E}\|$ refers to the sum of the elements of the matrix $\mathbf{E}$. The entries e_{ij} in the normalized matrix represent the fraction of edges that connect vertices of types i and j, and satisfies the normalization condition,

$$\sum_{ij} e_{ij} = 1. \tag{4.12}$$

The assortativity coefficient r is then defined thus,

$$r = \frac{Tr(\mathbf{e}) - \|\mathbf{e}\|^2}{1 - \|\mathbf{e}\|^2}, \tag{4.13}$$

where $Tr(\mathbf{e})$ is the standard matrix trace—the sum of the diagonal elements e_{ii}. The value of the coefficient r lies in the range $-1 \leq r \leq 1$, where 1 represents a perfectly assortative network, 0 a randomly mixed one and -1 a perfectly disassortative network.

Since the degree is an important topological measure, degree correlations assume a significant amount of relevance as they can give rise to complicated network structural

effects. The degree correlation can be computed using Eqn. 4.13, where the elements e_{ij} represent the fraction of edges that connect a vertex of degree i to that with degree j.

4.8 Modularity

Modularity stands for the difference between fraction of links that fall within communities and the expected fraction if links are randomly distributed [NG04]. This project introduces modularity as a measure to evaluate the connection strength between node pair within a group. Regarding to the industry where the stocks belong to, these stocks are divided into different groups hence modularity is used to measure the closeness of intra- and inter-group.

Two groups are combined to generate the modularity value while computing the closeness of two groups, as formula below shows:

$$Q = \frac{1}{2m} \sum_{j} \left[w_{ij} - \frac{k_i k_j}{2m} \right] \delta \left(c_i, c_j \right) \tag{4.14}$$

where c_i is the community to which node i is assigned, and k_i represents the degree of node i. The δ-function $\delta(u,v)$ is 1 if $u = v$ and 0 otherwise and $m = 0.5 \sum_{ij} w_{ij}$ is the sum of weights in the whole network.

Chapter 5

Community detection

This paper considers a theoretic concept of modularity-based community detection method for directed graphs to recognise natural faults occur in the stock network along which it partitions. Community detection is applied for further understanding to the overall pattern of economical and stock price relations of listed companies.

While there are many methods to identify communities in undirected graphs, the community detection method used in this paper is for the directed graphs proposed by Leicht and Newman [LN08], which based on modularity optimisation method. Modularity optimisation method identifies communities by maximizing the modularity Q, which is defined as:

$$Q = (\textit{fraction of intra-community edges}) - (\textit{expected fraction of such edges}) \quad (5.1)$$

It signifies that a community is figured when the number of edges inside the community is more than the expected number on the basis of chance. As a result, modularity-based community detection maximised intra-community density and minimised inter-community density. While the complexity of modularity optimisation is NP-complete problem, this paper uses the spectral optimisation methodology, which finds the best partition of the directed US stock network by the following expression of Q:

$$Q = \frac{1}{m}\sum_{ij}\left[A_{ij} - \frac{k_i^{\text{in}} k_j^{out}}{m}\right]\delta_{c_i,c_j} \quad (5.2)$$

where A_{ij} is defined to be 1 if there is an edge from j to i and 0 otherwise. k_i is the in-degree for node i, k_j is the out-degree for node j, m is the total number of edges in the network. δ is the Kronecker delta symbol that is 1 if nodes i and j are in

the same community, i.e., $C_i = C_j$, and 0 otherwise. Spectral optimisation technique for modularity maximisation assigns nodes to different communities based on the sign of the eigenvector, corresponding to the largest positive eigenvalue of the modularity matrix **B**, whose elements are:

$$B_{ij} = A_{ij} - \frac{k_i^{in} k_j^{out}}{m} \tag{5.3}$$

This paper applies the repeated bisection graph-partitioning algorithm in the cause of community detection according to Leicht and Newman [LN08]. This approach begins with partition the network in two and then repeating it while optimising for the maximum modularity score of the communities. A preferred partition of a network results in a higher modularity score, therefore the modularity Q is maximised over all possible partitions of the stock network to detect communities of listed companies.

The following algorithm describes the details about the partitioning process and maximising modularity score in community detection. The functions of calculating modularity and subdividing node group are called repeatedly over iterations until no further increment of the overall modularity score.

Algorithm 1 Community detection

procedure COMMUNITY(*G, nNode, nEdge, EntireModMat, EntireNodeSpace*)
  procedure CALDELTAQ(**s**, **B**)
    return $Q \leftarrow 1/(4 * nNode) * \boldsymbol{s}^T(\boldsymbol{B}+\boldsymbol{B}^T)\boldsymbol{s}$
  procedure UPDCOMMUNITYASSIGNMENT(*NodeSpace*, *UpdAssign*)
    $Mark1, Mark2 \leftarrow \max(Assignment)+1, \max(Assignment)+2$
    for each $node \in NodeSpace$ **do**
      if $node \in UpdAssign > 0$ **then** node of $Assignment \leftarrow Mark1$
      if $node \in UpdAssign < 0$ **then** node of $Assignment \leftarrow Mark2$
    return *Assignment*
  procedure SUBDIVIDECOMMUNITY(**B**)
    $SymmetricMatrix \leftarrow \mathbf{B}+\mathbf{B}^T$
    $eigv \leftarrow$ eigenvector as $\max(eigenvalues)$ in *SymmetricMatrix*
    return $\text{sign}(eigv)$
  procedure CALMODULARITY(*assignment*)
    for each $node1 \in Nodes\ of\ G$ **do**
      for each $node2 \in Nodes\ of\ G$ **do**
        if *assignment* of *node1* $\leftarrow$ *assignment* of *node2* **then**
          $Q \leftarrow Q + HasEdge - (\text{nIn}(node1)) * (\text{nOut}(node2))/(\text{nEdge})$
    return $Q/(nEdge)$
  procedure GENMODULARITYMATRIX(*NodeSpace*, *ModMat*)
    for each $node1 \in NodeSpace$ **do**
      for each $node2 \in NodeSpace$ **do**
        $B \leftarrow HasEdge - (\text{nIn}(node1)) * (\text{nOut}(node2)/\text{nEdge})$
        if Assignment of *node1* = Assignment of *node2* **then**
          for each $node \in NodeSpace$ **do** $C \leftarrow C + HasEdge1 + HasEdge2 - (\text{nIn}(node1) * \text{nOut}(node) + \text{nIn}(node) * \text{nOut}(node1))/\text{nEdge}$
  procedure INTERATEBISECTION(*ModMat, NodeSpace*)
    $UpdAssign \leftarrow \text{SubdivideCommunity}(ModMat)$
    $DeltaQ \leftarrow \text{CalDeltaQ}(UpdAssign, ModMat)$
    if $DeltaQ > 0$ **then**
      $Assignment \leftarrow \text{UpdCommunityAssignment}(NodeSpace, UpdAssign)$
      for each $side \in UpdCommunityAssignment$ **do**
        $ModMat \leftarrow \text{GenModularityMatrix}(NodeSpace)$
        $\text{InterateBisection}(ModMat, NodeSpace)$
  $\text{InterateBisection}(EntireModMat, EntireNodeSpace)$
  return *Assignment*

Chapter 6

Benchmarking networks generation

Numerous literatures support the idea that undirected stock networks have small-world features. It is helpful to examine whether the directed one is a small-world network or not by comparing it with known and conventional small-world networks like Watts-Strogatz model [WS98], as well as conventional random networks like Erdős–Rényi model [ER59].

6.1 Directed Watts-Strogatz small world network

The Watts-Strogatz (WS) model is a randomly generated graph with small world network properties such as high clustering coefficient and small average short path lengths.

In the Complex Networks science, a network with both a small average path length and a large average clustering coefficient feature is called a small world network. In the WS model, when the random reconnection probability p of the connected nodes is gradually increased from 0 to 1, it can be observed that the initial regular network will go through the following three phases: regular network, small-world network, and eventually random network.

This paper uses an alternative method based on WS model [SW14]. Specify the number of nodes N, the mean degree K (assumed to be an even integer), and a special parameter β, satisfying $0 \leq \beta \leq 1$ and $N \gg K \gg \ln N \gg 1$, the model constructs an undirected graph with N nodes and $NK/2$ edges as the following algorithm depicts:

Algorithm 2 WattsStrogatzSmallWroldNetwork

1: **procedure** GENERATESMALLWORLDNETWORK(*nNodes, p0, beta*)
2: $Dmax \leftarrow nNodes \ \% \ 2$
3: $R \leftarrow range from 1 to Dmax$
4: $D \leftarrow$ circulant matrix of $R/Dmax$
5: $p \leftarrow beta * p0 + ((D \leq p0) * (1 - beta))$
6: $A \leftarrow 1 * (\text{randomised matrix } p < p)$
7: fill diagonal of matrix A
8: **return** A

6.2 Directed Erdős–Rényi random network

The Erdös–Rényi (ER) model generates a graph that winded randomly between N nodes in the network with probability p. The degrees of nodes comply with a Poisson distribution, indicating that most nodes have approximately same number of edges.

Erdős and Rényi has found that as the number of edges M gradually increases from a small value, the random graph will evolve from a fragmented graph with many independent components to a fully connected one [Str01].

Chapter 7

Empirical study and results

7.1 Network construction

This paper first calculated normalised direct requirement and normalised direct demand values for every sectors using formula 3.1 and formula 3.2 and cross-correlation coefficients between every stock pairs using formula 3.4.

According to the figure 7.1, the transaction densities decrease as threshold of normalised direct requirement and normalised direct demand increase, and their patterns are exactly similar with the same inflection point at around $threshold = 0.136$ where the densities begin to decline. Therefore, the values of thresholds for normalised direct requirement and normalised direct demand are set to be equal, i.e., $\theta_{EIO} = \theta_{DD} = \theta_{DR}$, to filter the directed edges among the stock network.

Figure 7.2 shows the distribution of stock price correlation coefficients has a shape complies to the normal distribution with long tails. The correlation coefficients are vary from -0.687 to 0.977 with the mean of 0.265. It implies that the prices of most stocks traded in NYSE and NASDAQ usually fluctuate to the same direction, but the patterns are less similar.

Figure 7.4 shows the number of directed edges remain at the conditions of different value combinations of $\{\theta_{EIO}, \theta_{corr}\}$. When both of the thresholds set to be minimal at their own value range, i.e., $\theta_{EIO} = 0$ and $\theta_{corr} = -1$, the number of directed edges is $N \times (N-1) = 2,009,306$, while N indicates the total number of nodes, which is 1418. According to the figure 7.4, the number of edges will be less than $100,000$, in which case the network has a density of lower than 5%, if $\theta_{EIO} \geq 0.3545$ or $\theta_{corr} \geq 0.5020$.

It is obvious that the larger values assigned to θ_{EIO} and θ_{corr}, the more significant will be for the weights and directions of the remaining edges. But if the network

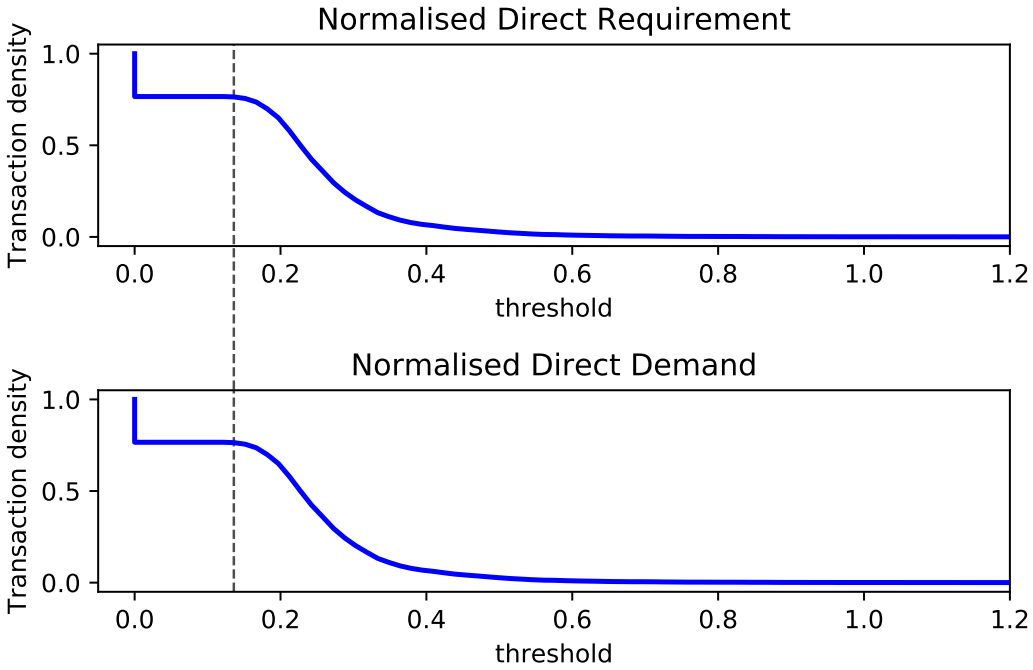


Figure 7.1: Transaction densities

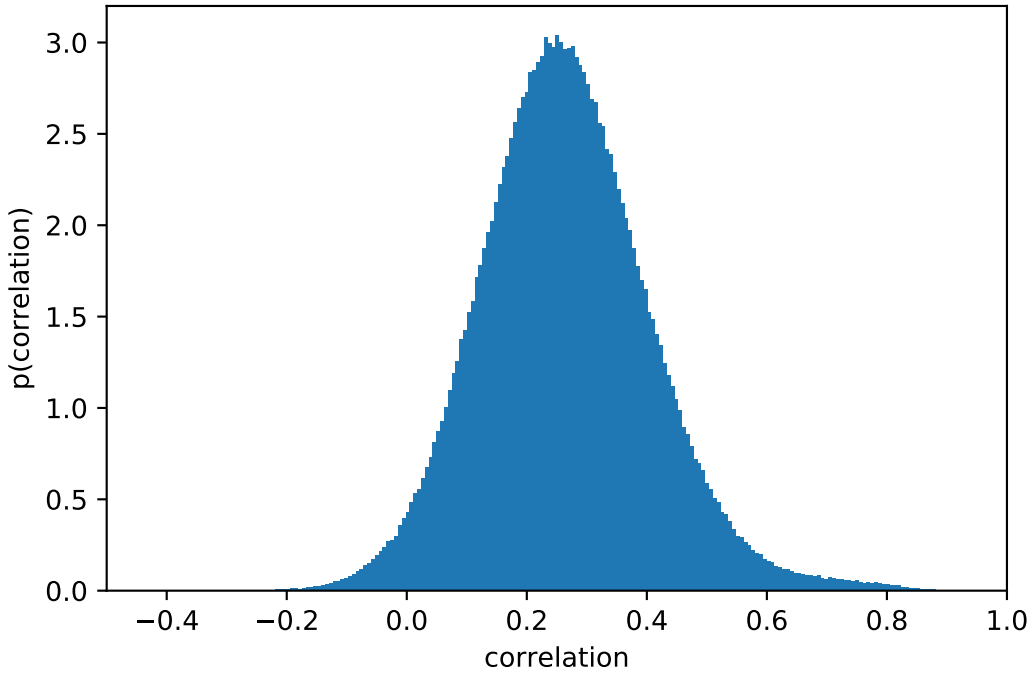


Figure 7.2: Stock price return correlation coefficient distribution

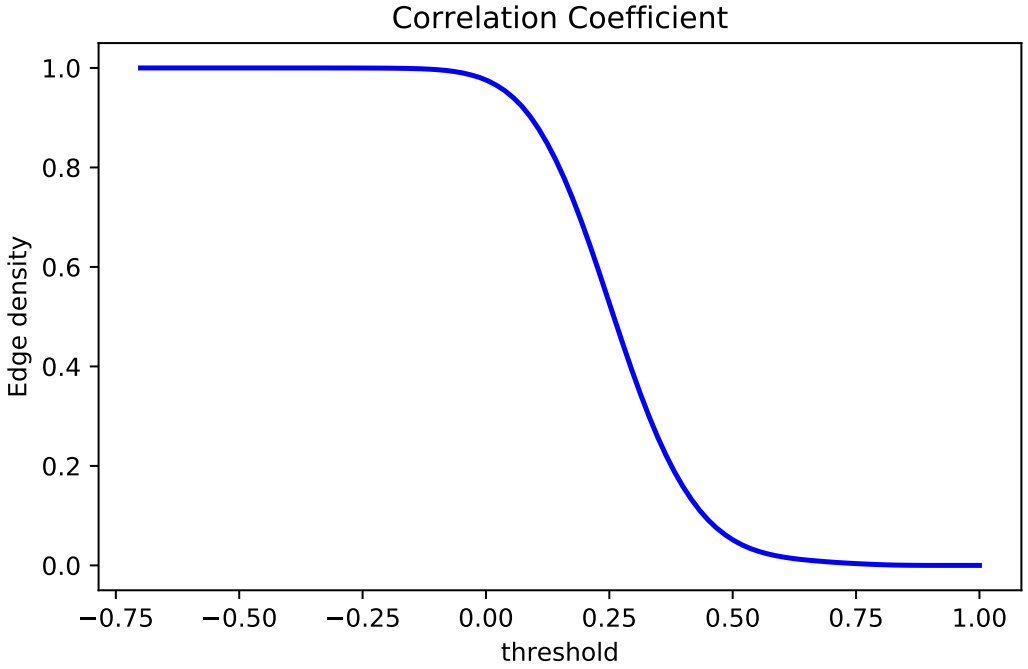


Figure 7.3

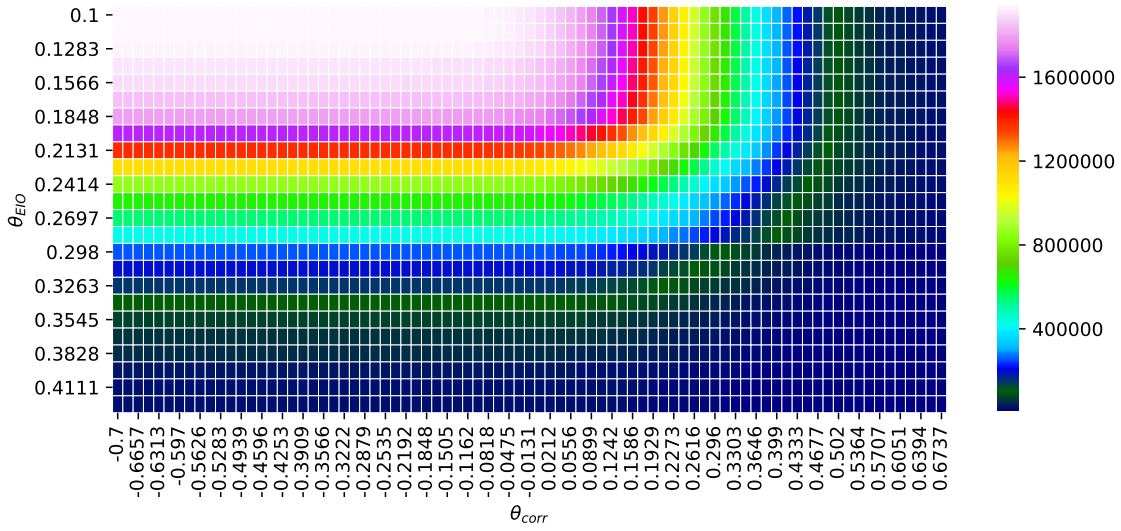


Figure 7.4: Numbers of directed edges per EIO-threshold and correlation-coefficient-threshold

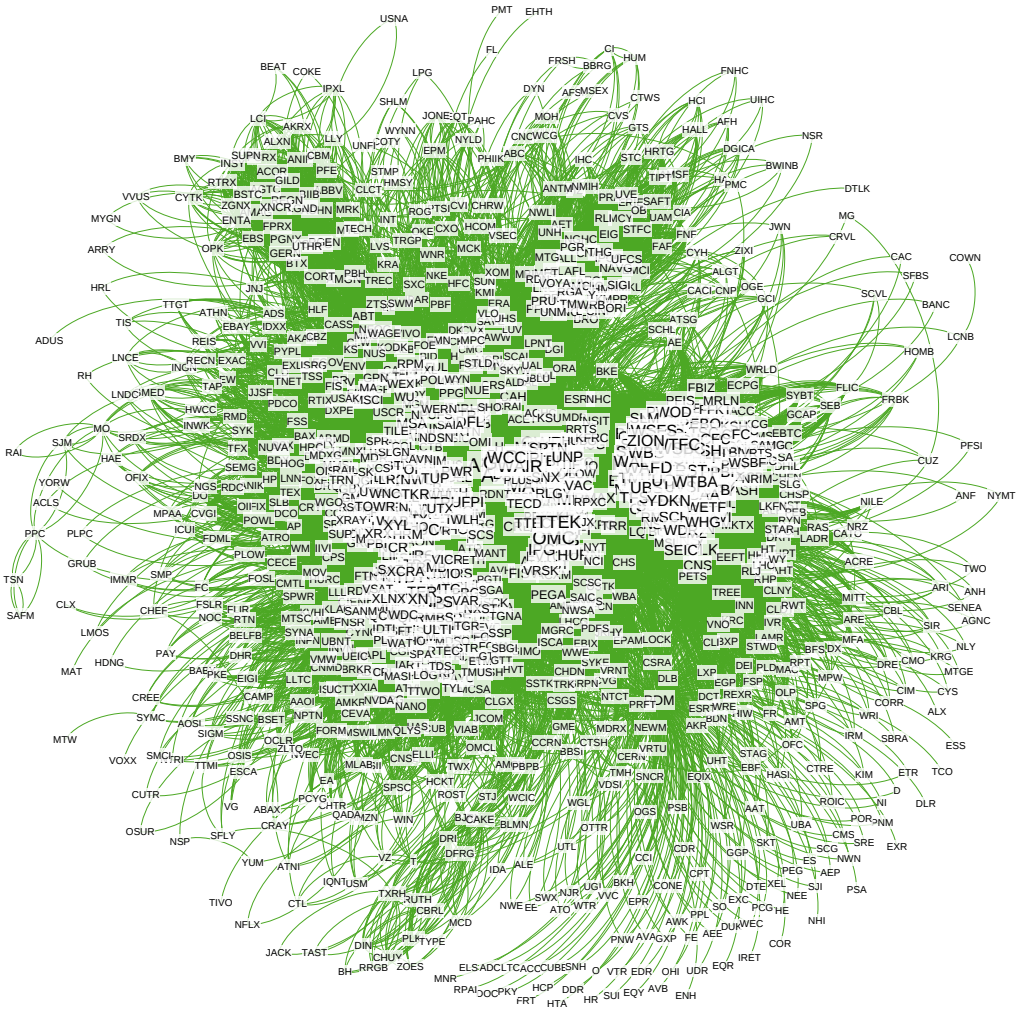


Figure 7.5: Transaction densities

Directed networks	Stock network	Small-world network	Random network
Number of nodes	1418	1418	1418
Number of edges	102051	102088	102097
Out-degree distribution	Power-law	Normal	Normal
Average out-degrees	143.94	143.99	144.00
Average path length	2.775	2.005	1.973
Clustering coefficient	0.4675	0.1367	0.05105
Global efficiency	0.2563	0.5161	0.5216
Local efficiency	0.6276	0.5027	0.4456
Assortativity	0.02004	-0.002180	0.001452

Table 7.1: Main properties of stock network, small-world network, and random network

becomes too sparse, it can not be strongly or even weakly connected and there would be many independent cliques, hence the network becomes too inefficient to be a senseful network. As a result, this paper selects the threshold-value-pair $\{\theta_{EIO} = 0.292, \theta_{corr} = 0.379\}$ to construct a directed unweighted network and a directed weighted network for the 2016 US stock market.

7.2 Analysis of the directed unweighted network

A directed Watts Strogatz small-world (WS) network and a random network with the same number of nodes and similar number of edges with the stock directed unweighted network are generated. Table 7.1 compares the main topologies of the three networks.

7.2.1 Power-law distribution

According to the table 7.1, the values of average out-degrees of stock network, small-world network and random network are almost the same due to the identical numbers of nodes and edges, but in terms of the distributions of out-degrees, stock network is totally different from the others.

Figure 7.10 and 7.11 show clearly that the out-degree distributions of small-world network and random network follow the normal distribution, i.e, most degree numbers of nodes fall in the middle range, while figure 7.9a illustrates that for the stock network, only a few number of nodes show higher out-degree, while most nodes are in the positions of low out-degree level. The distribution of the directed stock price return

networks follows power-law distribution with the exponent of 4.057.

7.2.2 Small-world property

The average path length of small-world network and random network are both close to 2, indicates that taking any node in the network, it can expect to reach any other nodes just through one node as medium. For stock network, the expectation number of medium nodes is 1.775, which is also a small number for connection, so like small-world and random networks, stock network also has the small-world property.

Although the assortativity values for the three networks are all non-significant,

7.2.3 Community structure of the directed unweighted stock network

The larger value of clustering coefficient for stock network than the other two networks indicating that the nodes in stock network tend to cluster together. Therefore, communities of stock network will be identified implementing the *algorithm 1* for directed networks in this section. According to the composition of industrial sectors of each community, as figure 7.8 shows, the following five communities are identified: (1) Production (2) Finance (3) Livelihood (4) Insurance and chemical products (5) Utilities and financial vehicles.

The communities of production (purple) and livelihood (blue) are sparsely distributed while there are some large-sized nodes acting as hubs of the overall network. The hubs not only connect to the nodes of same communities, but also the externals. These two communities are partially intertwined due to the high relevancy of production industry and livelihood industry.

Unlike the above two communities, it can be seen from figure 7.6 that the community of finance (green) is decentralised, i.e., there is no obvious hubs and the degrees of each node distribute evenly. It also has a very dense structure, connected closely inside and completely exclusive from other nodes or communities. This means the co-movements among financial stocks are incredibly strong and economically they rely tightly to each other.

The other two communities are more interesting because of their peculiar structural features. Every industrial sectors of individual stocks in community are identified to investigate the properties of the community of insurance and chemical products (yellow). As figure 7.7d illustrates and through the investigation, almost all firms in

	Undirected stock network
Degree distribution	Power-law
Average out-degrees	Power-law
Average path length	2.775
Clustering coefficient	0.4675
Global efficiency	0.2563
Local efficiency	0.6276
Assortativity	0.02004

Table 7.2: Main topologies of conventional stock price network

the upper and lower clusters are in the sectors of ”chemical products” and ”insurance carriers and related activities” respectively, while firms between the two big clusters, like ”MCK” (McKesson) and ”CAH” (Cardinal Health), are large medical supplier, pharmaceutical and heathcare service companies with high out-degrees to both of the two clusters. Apart from that, there are also a considerable number of links from the nodes in upper cluster to the hubs of chemical companies. Thus, it is reasonable to infer that the prices of medicines have significant influence to medical insurance industry, additionally the purchases of chemical products of pharmaceutical firms and the sales of chemical products have made pharmaceutical and chemical companies influence to each other.

Another investigation towards the community of utilities and financial vehicles (orange) is conducted by the same measure. As figure 7.7e illustrates, there is only one huge hub (PDM) which is the company ”Piedmont Office Realty Trust” among the whole community while all the others are one-degree nodes located remotely. There are more links from the hub to the rest than the opposite direction, and also the weights of the former links are generally higher. The hub, ”Piedmont Office Realty Trust”, is a real estate investment trust company, and the rest in the community contains 59 ”funds, trusts, and other financial vehicles” firms and 44 ”utilities” firms. For a big realty trust enterprise, demand for financial trust business is extremely high, and its successes of investments upon real estates will promote the development of utilities companies. It depicts that the major realty trust enterprise alone has significant influence to all of these financial trust and utilities companies.

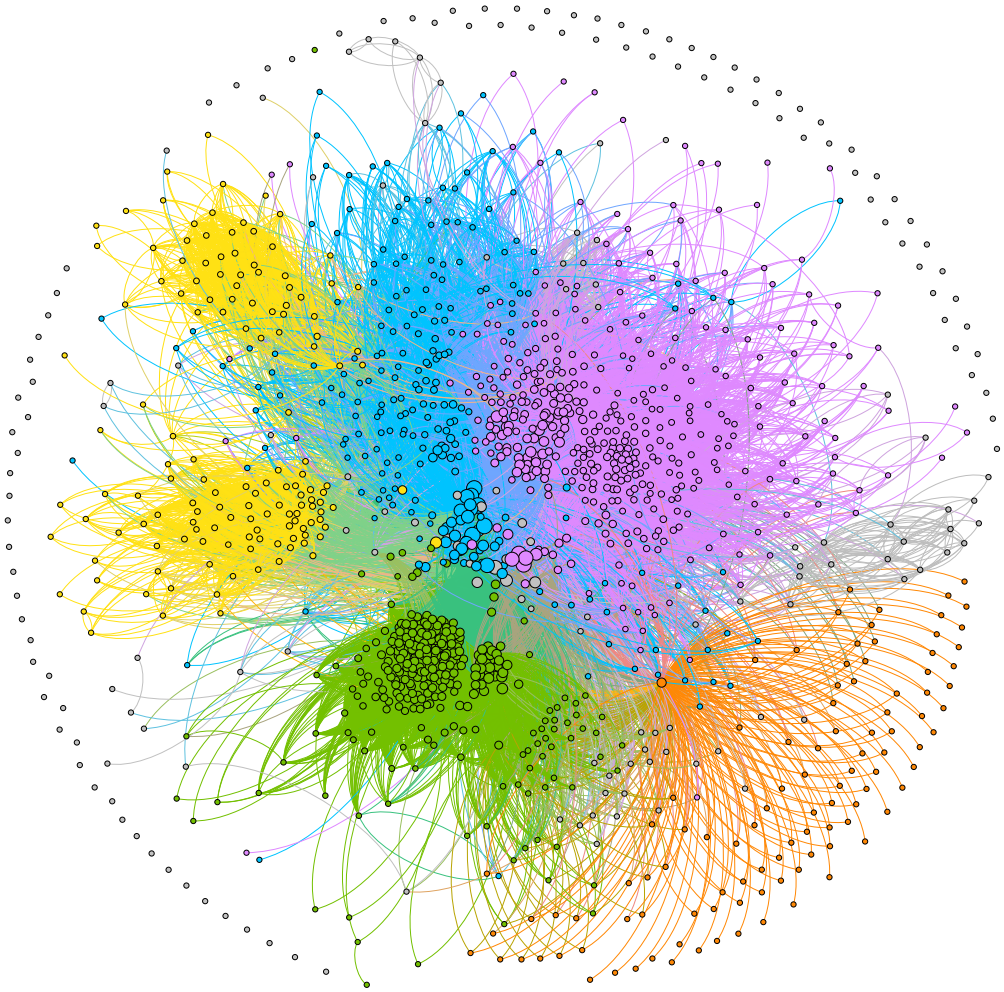
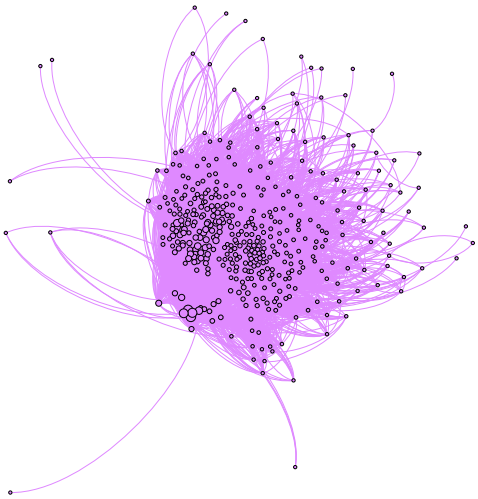
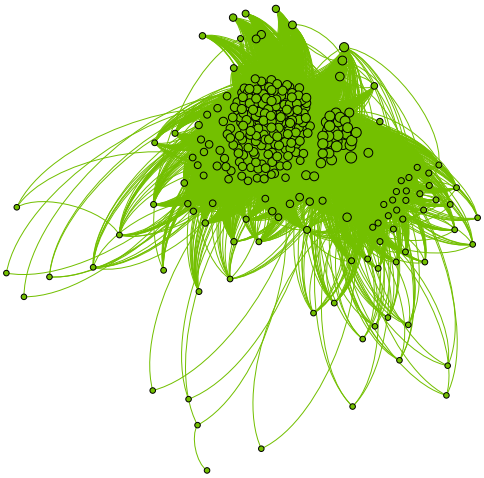


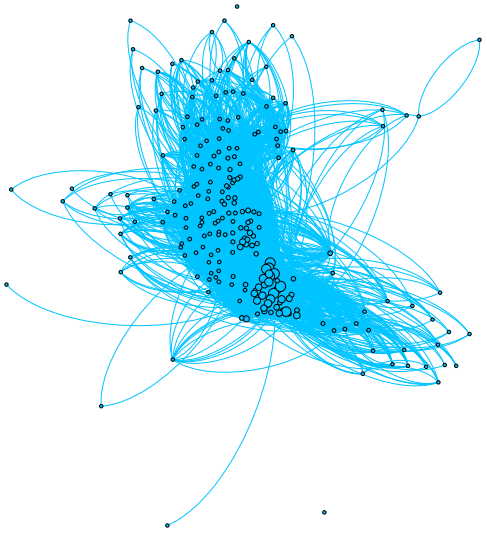
Figure 7.6: Community structure of the 2016 US stock price return network. Five distinct communities are detected represented by different colours of nodes. The direction of edge is clockwise. The size of nodes and thickness of edges are related to the value of degrees and weights. The grey nodes do not belong to any communities and most of them have zero degree.



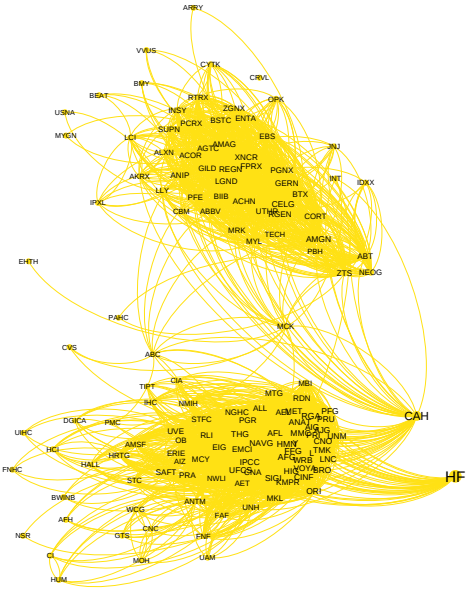
(a) Production



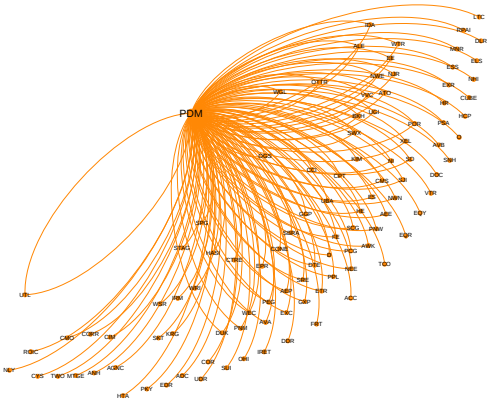
(b) Finance



(c) Livelihood



(d) Insurance and chemical products



(e) Utilities and financial vehicles

Figure 7.7: Community sole views of the directed stock network. Stock tickers are displayed for the sparsely distributed communities.

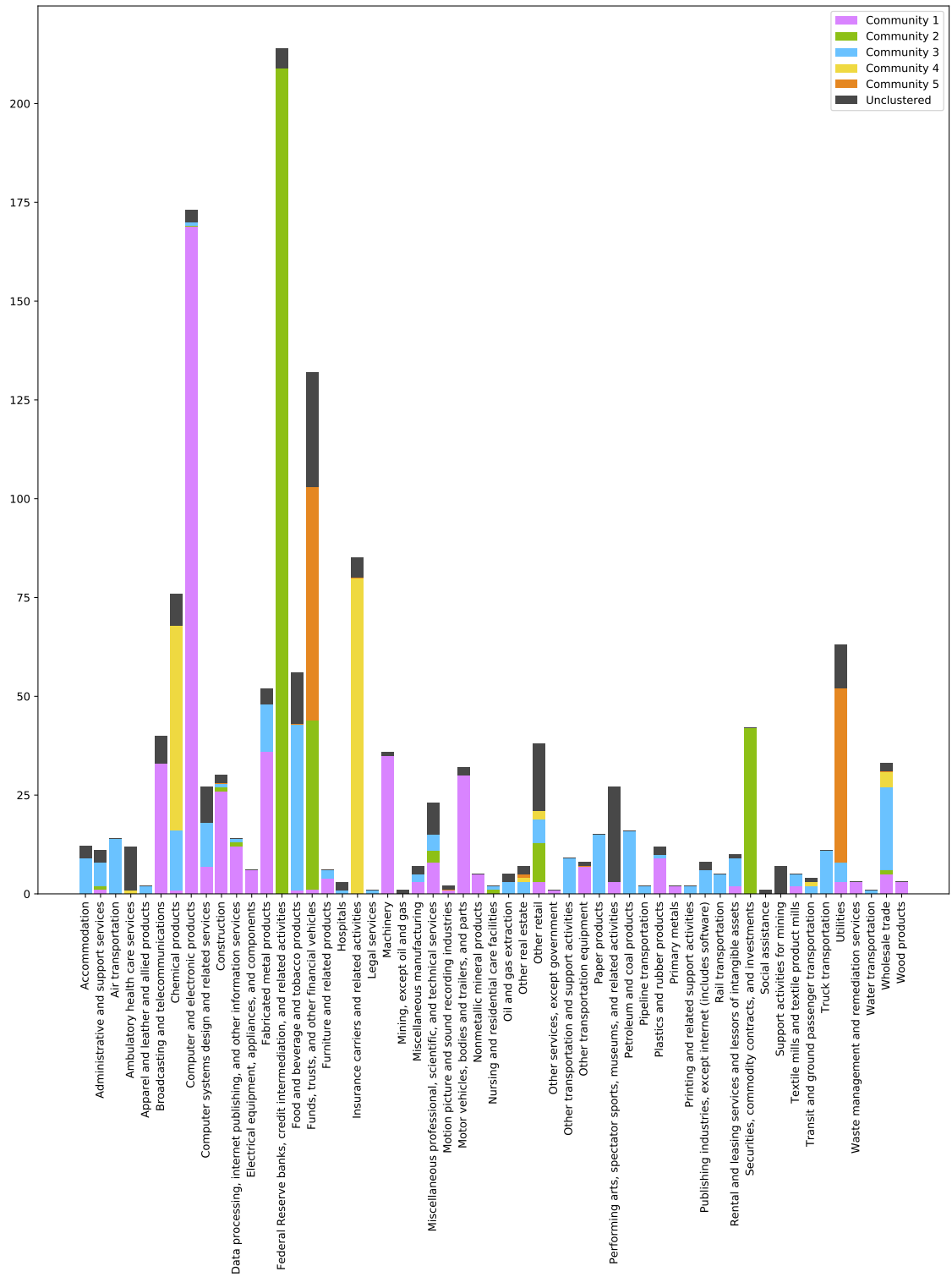
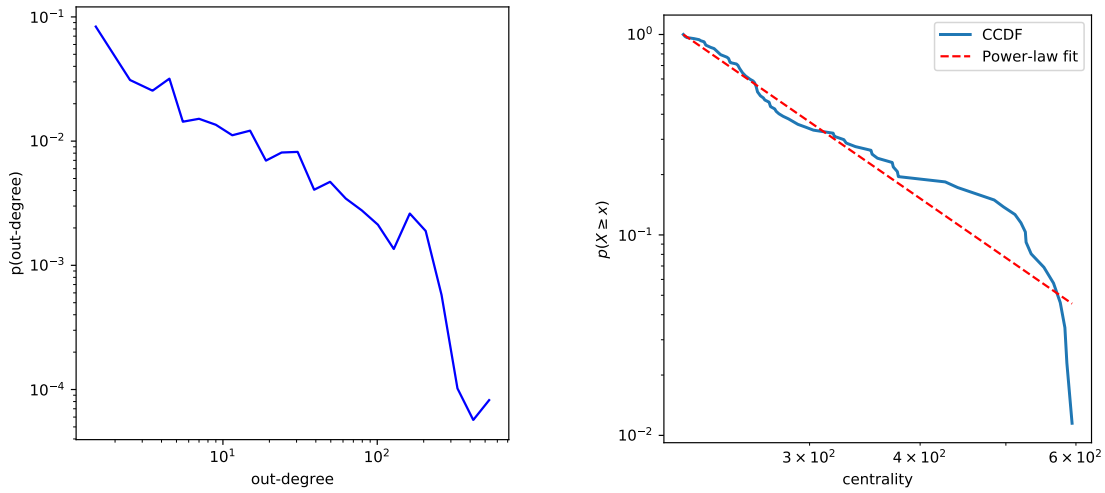


Figure 7.8: Stacked bar chart about the distribution of communities upon industrial sectors. Colours of stacks correspond to the colours of communities in figure 7.6 and figure 7.7, except the black stack indicating the nodes not belong to any communities. Sectors are arranged alphabetically.



(a) Empirical out-degree distribution

(b) CCDF and power-law fit

Figure 7.9: Out-degree distribution of stock network

	Weighted stock network
Strength distribution	Power-law
Average strength	40.89
Average betweenness centrality	0.0007440
Weighted assortativity	0.1244

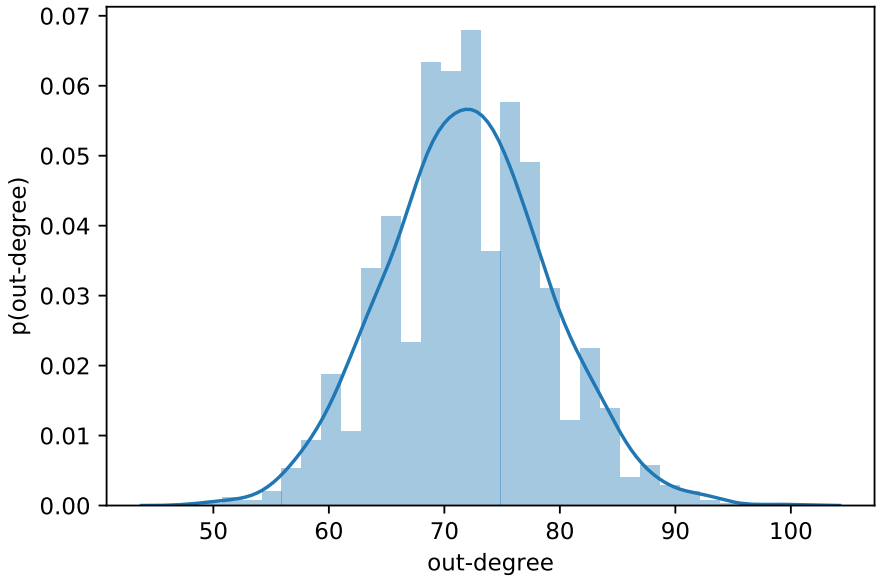
Table 7.3: Main topologies of weighted stock network

7.3 Analysis of the directed weighted network

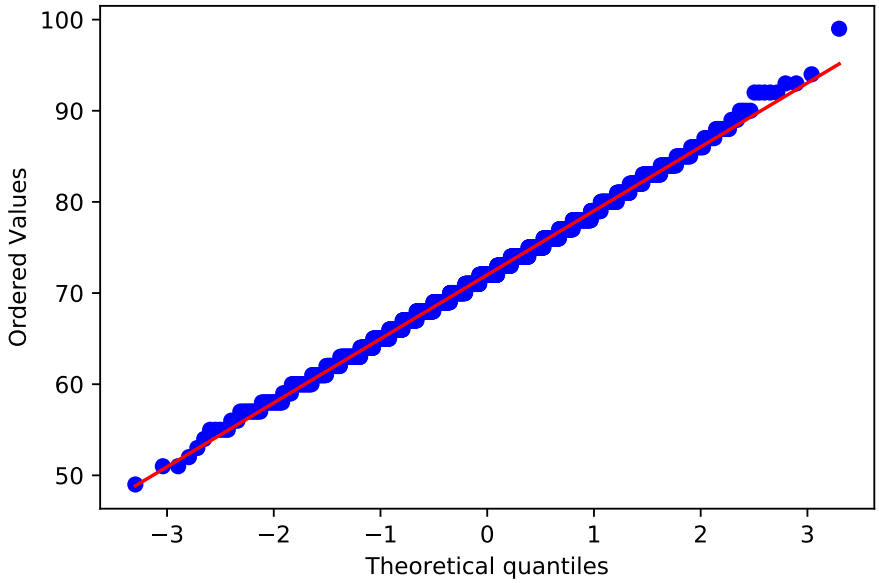
7.3.1 Topologies

7.3.2 Analysis on the relationship between price return and betweenness centrality

Figures 7.12 reveal the relationships between betweenness centralities and price returns of stocks. First, in figure 7.12a as much more nodes have low betweenness centrality values (lower than 0.001), and the cumulative sum of returns fall intensively in the range of $(-0.1, 0.6)$, while that of the nodes with high betweenness centrality values (higher than 0.001) also fall evenly in the same range Therefore, there is no significant difference between the expected return for stocks with different betweenness centrality values.

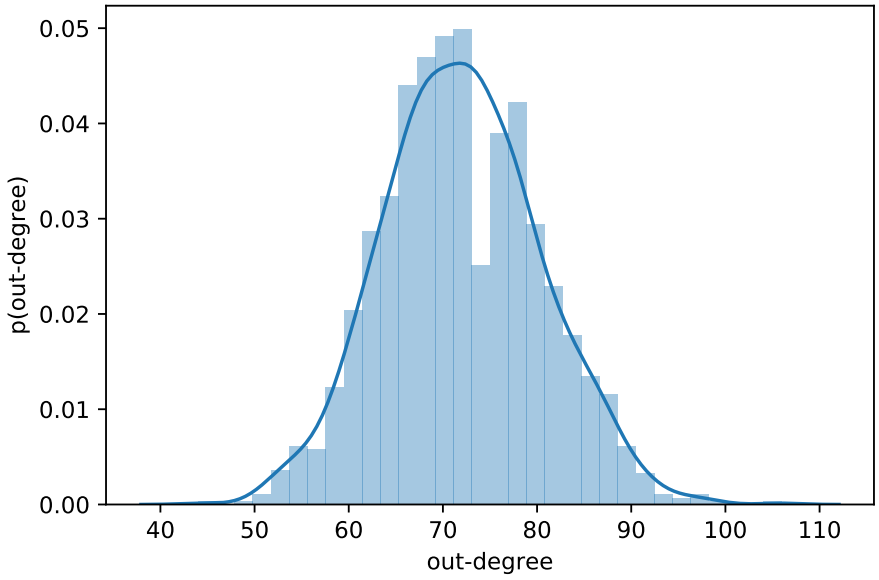


(a) Empirical out-degree distribution

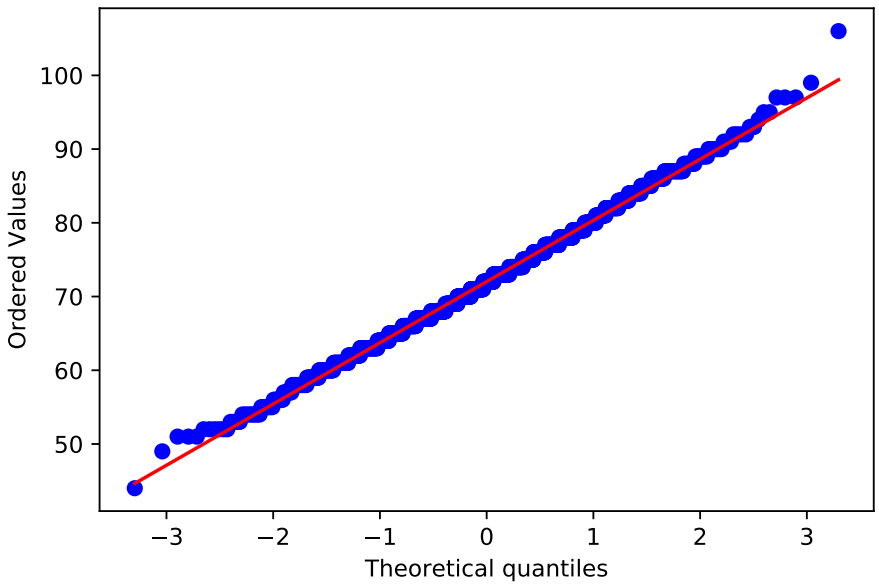


(b) P-P plot

Figure 7.10: Out-degree distribution and P-P plot of small-world network

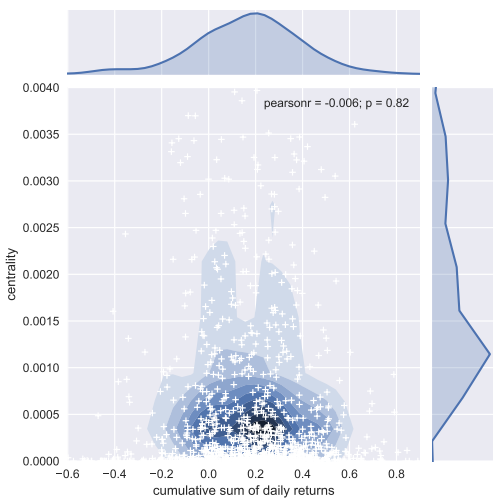


(a) Empirical out-degree distribution

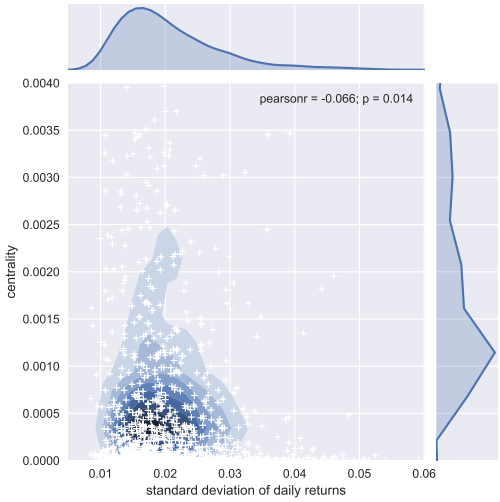


(b) P-P plot

Figure 7.11: Out-degree distribution and P-P plot of random network



(a) Bivariate distribution beween betweenness centralities of nodes and cumulative sums of stock daily return. The returns are actually logarithmic returns therefore the accumulation of all daily logarithmic returns in an entire year equals to a corresponding yearly logarithmic return.



(b) Bivariate distribution beween betweenness centralities of nodes and standard deviations of stock daily return.

Figure 7.12: Bivariate distributions with betweenness centrality

Second, according to the figure 7.12b, in spite of several outliers, as the betweenness centrality of nodes becomes higher, there will be a higher possibility of nodes tend to have low standard deviation of stock daily return. This indicates the hubs in the network have considerably stable return during the specified researched year among the whole stock market. The average standard deviation for all values of betweenness centrality remain similar because of the more frequent occurrences of outliers with higher betweenness centralities from the figure.

As a result, although choosing stocks with high centralities possibly will not bring a higher expected return for a portfolio, they have the functionality of decreasing the overall risks and generating more stable returns, which is also a vital feature for stock investment.

Chapter 8

Conclusions and future watch

This paper studied the directed complex networks of US stock market of 2016. The directions and weights of edges are determined by the economical transaction relations and stock price correlation coefficients respectively. Overall, the characteristics of topology properties have not changed significantly from undirected weighted stock networks from other literatures which used correlation coefficients of stock prices as the weights of edges. However, from the new horizon, this paper is able to analyse on a higher dimensionality – topological property research and community detection with methods for directed networks which utilised the feature of edge directions. The resulting features of power-law and small-world for directed stock complex networks show continuity with the results in undirected stock complex networks researches. The partitioned communities are highly related with the economical activities among industries and indicate the potential cascading impact from a collapse of a specific firm or sector.

In terms of future work, stock complex networks in more years can be generated and compared in together, the periods correspond to bull, bear and stable market can be recognised and analysed separately and accordingly. New methods for determining the directions of edges to generate directed complex networks are expected to be proposed.

Bibliography

[BBP05] Vladimir Boginski, Sergiy Butenko, and Panos M Pardalos. Statistical analysis of financial networks. *Computational statistics & data analysis*, 48(2):431–443, 2005.

[CK15] Shauhrat S Chopra and Vikas Khanna. Interconnectedness and interdependencies of critical infrastructures in the us economy: Implications for resilience. *Physica A: Statistical Mechanics and its Applications*, 436:865–877, 2015.

[CLL10] K Tse Chi, Jing Liu, and Francis CM Lau. A network perspective of the stock market. *Journal of Empirical Finance*, 17(4):659–667, 2010.

[CLSW15] Kun Chen, Peng Luo, Bianxia Sun, and Huaiqing Wang. Which stocks are profitable? a network method to investigate the effects of network structure on stock returns. *Physica A: Statistical Mechanics and its Applications*, 436:224 – 235, 2015.

[CR76] John C Cox and Stephen A Ross. The valuation of options for alternative stochastic processes. *Journal of financial economics*, 3(1-2):145–166, 1976.

[ER59] Paul Erdős and Alfréd Rényi. On random graphs i. *Publicationes Mathematicae (Debrecen)*, 6:290–297, 1959.

[FF93] Eugene F Fama and Kenneth R French. Common risk factors in the returns on stocks and bonds. *Journal of financial economics*, 33(1):3–56, 1993.

[FF96] Eugene F Fama and Kenneth R French. Multifactor explanations of asset pricing anomalies. *The journal of finance*, 51(1):55–84, 1996.

[Fre77] Linton C Freeman. A set of measures of centrality based on betweenness. *Sociometry*, pages 35–41, 1977.

[HZY09] Wei-Qiang Huang, Xin-Tian Zhuang, and Shuang Yao. A network analysis of the chinese stock market. *Physica A: Statistical Mechanics and its Applications*, 388(14):2956–2964, 2009.

[JJH+03] Neil F Johnson, Paul Jefferies, Pak Ming Hui, et al. Financial market complexity. *OUP Catalogue*, 2003.

[Lin65] John Lintner. Security prices, risk, and maximal gains from diversification. *The journal of finance*, 20(4):587–615, 1965.

[LLH07] Kyoung Eun Lee, Jae Woo Lee, and Byoung Hee Hong. Complex networks in a stock market. *Computer Physics Communications*, 177(1-2):186, 2007.

[LM01] Vito Latora and Massimo Marchiori. Efficient behavior of small-world networks. *Physical review letters*, 87(19):198701, 2001.

[LN08] E. A. Leicht and M. E. J. Newman. Community structure in directed networks. *Phys. Rev. Lett.*, 100:118703, Mar 2008.

[Lon13] Yu Long. Visibility graph network analysis of gold price time series. *Physica A: Statistical Mechanics and its Applications*, 392(16):3374–3384, 2013.

[Mar52] Harry Markowitz. Portfolio selection. *The journal of finance*, 7(1):77–91, 1952.

[Mer73] Robert C Merton. An intertemporal capital asset pricing model. *Econometrica: Journal of the Econometric Society*, pages 867–887, 1973.

[New02] Mark EJ Newman. Assortative mixing in networks. *Physical review letters*, 89(20):208701, 2002.

[NG04] Mark EJ Newman and Michelle Girvan. Finding and evaluating community structure in networks. *Physical review E*, 69(2):026113, 2004.

[NSRJ11] A Namaki, AH Shirazi, R Raei, and GR Jafari. Network analysis of a financial market based on genuine correlation and threshold method. *Physica A: Statistical Mechanics and its Applications*, 390(21-22):3835–3841, 2011.

[oEA18] U.S. Bureau of Economic Analysis. Use tables/before redefinitions/producer value. https://www.bea.gov/industry/io_annual.htm, 2018.

[Sha64] William F Sharpe. Capital asset prices: A theory of market equilibrium under conditions of risk. *The journal of finance*, 19(3):425–442, 1964.

[Str01] Steven H Strogatz. Exploring complex networks. *nature*, 410(6825):268, 2001.

[SW14] H Francis Song and Xiao-Jing Wang. Simple, distance-dependent formulation of the watts-strogatz model for directed and undirected small-world networks. *Physical Review E*, 90(6):062801, 2014.

[WS98] Duncan J Watts and Steven H Strogatz. Collective dynamics of 'small-world'networks. *nature*, 393(6684):440, 1998.

Appendix A

Example of operation

An appendix is just like any other chapter, except that it comes after the appendix command in the master file.

One use of an appendix is to include an example of input to the system and the corresponding output.

One way to do this is to include, unformatted, an existing input file. You can do this using `\verbatiminput`. In this appendix we include a copy of the C file hello.c and its output file hello.out. If you use this facility you should make sure that the file which you input does not contain `TAB` characters, since LaTeX treats each `TAB` as a single space; you can use the Unix command `expand` (see manual page) to expand tabs into the appropriate number of spaces.

A.1 Example input and output

A.1.1 Input

(Actually, this isn't input, it's the source code, but it will do as an example)

```
/* Hello world program */

#include <stdio.h>

int main(void)
{
    printf("Hello World!\n") ;
    return 0 ;
```

```
}
```

A.1.2 Output

```
Hello World!
```

A.1.3 Another way to include code

You can also use the capabilities of the `listings` package to include sections of code, it does some keyword highlighting.

```
/* Hello world program */

#include <stdio.h>

int main(void)
{
    printf("Hello World!\n");
    return 0;
}
```