공학박사학위논문

Complex Network Analysis of Financial Market using Link Prediction

링크 예측을 이용한 금융시장 복잡계 네트워크 분석

2020 년 8 월

서울대학교 대학원

산업공학과

박 지 환

Complex Network Analysis of Financial Market using Link Prediction

링크 예측을 이용한 금융시장 복잡계 네트워크 분석

지도교수 장 우 진

이 논문을 공학박사 학위논문으로 제출함

2020 년 6 월

서울대학교 대학원

산업공학과

박 지 환

박지환의 공학박사 학위논문을 인준함

2020 년 7 월

위 원 장 ______이 재 욱______ (인)

부위원장 ______장 우 진______ (인)

위 원 ______박 건 수______ (인)

위 원 ______박 우 진______ (인)

위 원 ______장 희 수______ (인)

Abstract

Complex Network Analysis of Financial Market using Link Prediction

Park, Ji Hwan

Department of Industrial Engineering

The Graduate School

Seoul National University

Financial risk sets off a chain reaction in the market and leads to a collapse of the system, called a domino effect. Since the U.S. subprime mortgage crisis in 2008 hit economies across the world, it has emerged important research fields to understand and analyze the financial system properly to deal with financial risk. Econophysics is an interdisciplinary research field to explain the stylized facts in financial systems that are unexplainable by traditional financial theories. In particular, the complex network models that represent a system by nodes and links are widely applied regardless of research areas. However, since the existing complex network models for financial markets usually end up in confirming empirical results such as a structural change in the network and diffusion paths of risk, based on historical data, it has limitations to suggest direct alternatives. To cope with these limitations, this dissertation proposes a link prediction model based on the real effective exchange rate (REER) that reveals the relationships clearly between the compositions. At first, it is confirmed that the network successfully mimics the market to ensure the validity of the network structure prediction. The results show that the return of REER has fat-tailed distributions whose tails are not exponentially bounded and follow a power-law. Also, for the analysis, the changes are focused on cross-sectional topology and time-varying properties of the network during the U.S. subprime mortgage crisis, the European debt crisis, and the Chinese stock market turbulence. The result implies that the network appropriately describes the market by showing the significant increments in out-degrees and in-degrees of the originating continents of the crises. Secondly, the Weighted Causality Link Prediction (WCLP) model is proposed to predict future possible links by measuring the similarities between different nodes. This model has differentiations that it measures the strength of directed Granger causality directions as effect sizes based on F-statistics, while the existing models are based on correlations. The experiment is conducted under the hypothesis that the intensity of connections is different from each other and maintains longer when the effect size is larger. The higher prediction accuracy is observed rather than that of unweighted or correlation-based weighted models by showing the statistical significance of higher Area Under Curve (AUC) in every aspects. Finally, a decision making model for investment is proposed based on the results of the link prediction. Once the portfolio is composed of stocks located in the periphery of the PMFG, it distributes the risk due to the low correlation between assets. However, the correlation does not represent the relationship by time lags since it implies only the extent of association between them. Therefore, this dissertation proposes the Weighted Causality Planar Graph (WCPG) that is improved from the Planar Correlation Planar Graph (PCPG) model. It differs from the existing models in that it considers directions and strength of links based on the similarity score between assets. As a result, the proposed model improves the performance in terms of risk-adjusted return compared to the benchmarks. Especially, it has an advantage in long-term investment for over 6 months. In conclusion, the contributions of this dissertation involve the development of an effective link prediction model based on the effect size and the attempt to suggest a decision-making model for investment.

Keywords: Real effective exchange rate, Granger causality network, Stylized facts, Effect size, Directed network, Link Prediction, Currency market, Portfolio selection, Planar Graph, Weighted causality planar graph

Student Number: 2014-21813

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Problem Description

Risk is defined as the chance of investors' gains will differ from the return and also the possibility of losing something of value to an individual or group. The risk always derives from uncertainty. In the modern financial market where many countries of the world are intimately interconnected, it is necessary to understand and analyze the market accurately since the uncertainty is more prominent and the collapse of the system has a huge ripple effect. Then, risk management in financial markets has been a constantly important agenda and become more indispensable than ever because of several financial crises. Since the financial markets have some empirical statistical regularities, called stylized facts (Cont, 2001), traditional financial risk management theories applied simple time series models cannot well explain the market (e.g. heavy-tailed distribution, volatility clustering, and et cetera). To overcome this problem, various econophysics theories appeared. Specifically, complex network theory has used to explain the relationship between market components.

A complex system exists between perfect order and disorder, which means that various components cause complex phenomena through the interactions among each other. Now, it is known that the world we live in is made up of complex systems (Chen et al., 2014). These systems can be expressed by nodes that are components of the system and edges (or links) that connect them to each other. It can be represented mathematically as an adjacency matrix and various studies can be done through it. Therefore, many researchers have devoted their efforts to analyze real-world systems by complex network analysis, such as computer networks, biological networks, social networks.

Particularly, a lot of studies are conducted on complex networks since the late 1990s and early 2000s. In the earliest days, research on small-world networks (Watts and Strogatz, 1998; Amaral et al., 2000; Comellas and Sampels, 2002; Latora and Marchiori, 2002) and scale-free networks (Barabási and Albert, 1999; Wang and Chen, 2002; Barabási and Bonabeau, 2003) are studied. Both indicate the unique characteristics of the network, which means that the structure is not a random network, but a few nodes are connected to many other nodes as hubs. Therefore, further researches are conducted on how the network affects and changes when an event occurs in the network. In the early and mid-2000s, research on the diffusion of networks along with random networks (Noh and Rieger, 2004; Derényi et al., 2004; Simonsen et al., 2004; Liu et al., 2004; Tadić and Thurner, 2004) is studied. In the second half of 2000, an empirical study is conducted on the problems to be solved in society such as traffic (Bagler, 2008; Wu et al., 2008; Zheng and Gao, 2008) and mobile communication (Lambiotte et al., 2008; Hidalgo and Rodríguez-Sickert, 2008). On the other hand, since the middle of 2010, researches on the information of the node (Konno, 2016) and the flow of the information (Liu et al., 2016; Ma et al., 2016b) emerge as a challenge.

In addition, many researchers analyze the financial market from the perspective of the complex network. The topological structure of a network composed of nodes that mean the active agent of a financial system and links that represent the interaction between them has the fundamental and important meaning in understanding a system. In many previous studies, the interaction is measured by correlation between different nodes for mapping a system to a network by representing as the topological structure of the network based on hierarchical clustering methods such as minimum spanning tree (MST) and planar maximally filtered graph (PMFG) (Mantegna, 1999; Onnela et al., 2003; Tumminello et al., 2005, 2007; Aste et al., 2010). If the complex network reflects the system well, it can be used to estimate or manage the market risk since an important node can be distinguished from other nodes.

In terms of the application of network analysis, the portfolio optimization is an important research area in finance and management science. Investors only concern with how to distribute their wealth among the various assets. They hope to have an appropriate balance of risk and profit in a trade-off relationship according to their risk tolerance. Since Markowitz's modern portfolio theory (MPT) is the most fundamental theory in terms of portfolio management, theoretical and mathematical models for more effective portfolio diversification have emerged (Markowitz, 1952), such as Black and Litterman's model (Black and Litterman, 1990) and the Post-Modern Portfolio Theory (Rom and Ferguson, 1994). Also, researches on risk evaluation and portfolio optimization based on complex networks have prominently increased (Onnela et al., 2003; Pozzi et al., 2013; Boginski et al., 2014; Ren et al., 2017; Tola et al., 2008; Nanda et al., 2010; Peralta and Zareei, 2016; Li et al., 2019). The detailed summary of related studies is reported in the last part of Chapter 2.

From the perspective of portfolio selection based on the network structure, it becoms an important issue to understand, analyze, and predict the structure of the network with the emergence of studies using the topological properties. One of the primary applications is predicting a link in a complex network, called Link Prediction (LP). Link prediction is a method to estimate the possible existence of links based on the observed links and the properties of the nodes that can be obtained in the topology structure of the network (Getoor and Diehl, 2005). Using this method, one can obtain hidden information from the network with incomplete data or possible links for the network in the future. The demand for link prediction appears in many research areas. For predicting the missing links, it can be used in the biological networks (Yu et al., 2008; Stumpf et al., 2008; Lei and Ruan, 2013), social networks (Schafer and Graham, 2002; Kossinets, 2006). And for predicting the future possible links, it can be used to recommend new friendship or evaluate the evolving mechanism of the given network (Zhou and Mondragón, 2004). Also, models that measure the proximity between nodes appear based on the structural properties (Newman, 2001; Adamic and Adar, 2003; Murata and Moriyasu, 2007; Lü et al., 2009). The detailed summary of related studies is reported in the second part of Chapter 2.

1.2 Motivations of Research

Despite the attempts to analyze financial markets, the common shortcomings of traditional network analysis is that it does not consider the future movement of the market. Many complex network theories for financial markets are developed from the empirical network mapping based on historical data. However, there is a disadvantage to manage financial risk since it is difficult to prepare for an oncoming recession. Then, investors have to take risk of uncertainty in the future, such as a financial crisis. Also, the Granger causality network is mostly utilized for revealing the boolean Granger causality directions without considering the intensity of connections. To cope with this limitation, this dissertation provides the resolution by proposing the link prediction model that estimates the likelihood of the existence of directional links between nodes based on the effect size in the Granger causality network. Once the models are mathematically organized, the validities of proposed models are evaluated for the global currency market. Also, an advanced portfolio selection method is proposed for managing financial risk in the U.S. stock market.

For studying network models, the concept of Granger causality network is selected for mapping the interactions of financial systems. The interactions visualized as directed edges in the network should imply the (de)coupling or dependency between the nodes. In the case of the exchange rate and the stock price, it is clear that cause-and-effect relationships exist between countries and institutions, respectively. In this dissertation, the Granger causality network is constructed based on the real effective exchange rate and total return index, which indicate the value of each currency and institution. By investigating the dependency among countries and stocks, predicting links that can be exist in the future enables to understand the time-varying financial market.

For studying link prediction, the concept of effect size is chosen to describe the contribution of the individual currency value of a country to that of other countries, which implies the strength of interconnectedness in the financial system. This model only requires the daily real effective exchange rate dynamics of each country. Since it already involves several features such as the relative trade balance and inflation rate, the relationship between the different currency values is clearly stated. The existing model of the complex network is limited in reflecting the directed interactions between different assets since it is based on the correlation. In this dissertation, the effect size is introduced to describe the strength of Granger causality directions and is obtained from the F-statistics in the Granger causality test.

For managing financial risk, the portfolio selection is chosen to describe the contribution of predicting links to the decision making of investors. This model only requires the daily total return index of individual stock. Since the bidirectional association between the individual institution represents more complicated relationships, the relationship between different financial firms can be clearly stated. However, the existing models of constructing optimal portfolios from a perspective of the complex networks are limited in revealing the directed interaction since they mostly utilize the correlation. Thus, this dissertation modifies the base concept of PMFG and proposes a model to construct portfolios using the predicting link based on the effect size in the directed network. It helps investors expect high returns against the risk they expose.

Then, the prediction of network structure is suggested based on newly proposed models using the effect size. In addition, this dissertation utilizes the prediction result for constructing an optimal portfolio. To do so, the interactions are represented via similarity-based methods.

1.3 Organization of the Thesis

The rest of this dissertation is organized as follows. In Chapter 2, the theoretical backgrounds are introduced for network models, link prediction and portfolio selection; In Chapter 3, the constitution of the exchange rate systems are explained in terms of countries, data description, considered experiment period, and sectoral analysis based on the financial crises. The return series of the real effective exchange rate (REER) is examined by statistical test and power-law fitting for identifying the shape of the distribution. Also, the change of network structure before and after financial crises is observed whether it reflects the market properly. Chapter 4 focuses on presenting the economic implications of the REER network and predicting links in the network with benchmarks and proposed methods. Specifically, the effect size is introduced based on the F-statistics that has statistic significance. In Chapter 5, the portfolio selection strategy is proposed based on the results of link prediction in the U.S. stock market. Lastly, the contributions and limitations of this dissertation are reviewed in Chapter 6 with possible future work for improvement. Also, the contents and mechanisms of Chapter 3 and Chapter 4 follow the same procedures stated in Park et al. (2020).

Chapter 2

Literature Review

2.1 Network models

Network science is a field of research focusing on modeling and analyzing various phenomena in the real world as a system. In this field, the complex network analysis is applied for real-world networks such as social networks, financial networks, computer networks, and biological networks. Now, it is a well-known fact that the world consists of complex systems (Chen et al., 2014). These systems can be represented by the nodes playing the active agents in the system and by the edges (or links) representing the interactions and interconnectedness among nodes. In this regard, the nodes and edges can be mathematically represented as an adjacency matrix, which then can be utilized for further analysis. One of the primary applications of such an adjacency matrix is predicting the possible link in a complex network.

In Econophysics, a lot of researches are conducted on complex networks since the late 1990s and early 2000s, especially about the small-world networks and the scale-free networks. For the small-world networks, Watts and Strogatz (1998) say that dynamical systems can be highly clustered, but have small characteristic path lengths, which are called 'small-world' networks. Also, Price (1965) first introduce the concept of a scale-free network and analyze the incidence of citations and showed that a degree distribution follows a power law. Amaral et al. (2000) analyze the statistical properties of real-world networks and supported to the occurrence of small-world networks in three classes: scale-free, broad-scale, and single-scale networks. Comellas and Sampels (2002) construct a small-world network in a deterministic way, which allowed to calculate the relevant network parameters directly. Latora and Marchiori (2002) propose a more sophisticated analysis that relies on transportation efficiency in real transportation networks and converted the complex problems into the underlying construction problem. And for the scale-free networks, Wang and Chen (2002) study a stabilization problem for the scale-free dynamic networks using local feedback pinning and explained why complex networks remain stable despite some nodes had unstable status frequently.

In the early and mid-2000s, Onnela et al. (2003) propose a model that construct a network based on correlations between asset returns, called a dynamic asset graph. From the perspective of diffusion in networks, Noh and Rieger (2004) investigate a random walk in scale-free networks. The Mean First Passage Time (MFPT) is introduced as an important property of a random network, which means that the time a random walker passes from node i to another node j. Simonsen et al. (2004) analyze a diffusion process on real-world networks by focusing on the slowest decaying diffusive modes. Liu et al. (2004) show that disease can spread throughout the network even if the local recovery rate is high in the household-structure network. In the second half of 2000, more practical studies are conducted than before. The related studies focus on analyzing the financial network based on correlation in various countries such as Brazil (Tabak et al., 2010), China (Zhuang et al., 2007; Huang et al., 2009), and Korea (Jung et al., 2006).

And since the early 2010, the empirical network is analyzed among the financial institutions with financial data such as stock prices, exchange rates and et cetera. In particular, many researches focus on how the evolution of the stock price affects the value of other companies since a stock price represents the value of a company. Song et al. (2011) convert the cross-correlation using industrial portfolios in the U.S. stock market to PMFG to measure of mutual information to detect the dynamics. Preis et al. (2012) observe a universal relationship between stock market portfolios and the normalized returns of them and introduce a concept of diversification breakdown that are needed for portfolio protection. In addition, Vŷrost et al. (2015) examine the structures of Granger causality networks using stock prices of developed markets and verify the role of the temporal proximity for return spillovers. In terms of the node information, Konno (2016) introduce the knowledge spillover model that represents externality in economic phenomena and found several characteristics of the long-run growth rate, the productivity level. In addition, Liu et al. (2016) propose a ranking method based on the degree value of nodes using five real networks as test data.

2.2 Link Prediction

Link prediction is the most fundamental problem of estimating the likelihood of the existence of a link in a given network (Getoor and Diehl, 2005). Kossinets (2006) analyze the effect of missing data based on the structural characteristics of the social network using three mechanisms: network boundary specification, survey non-response, and censoring by vertex degree. Clauset et al. (2008) show that the hierarchical structure of the network could be used to predict the missing connections of the network partially known with high accuracy. Liben-Nowell and Kleinberg (2007) define a link prediction problem as to whether new interactions would occur in the near future between members in a social network and proposed a method to analyze the proximity of nodes. Clauset et al. (2008) introduce a method of inferring hierarchical structures in a network and showed that the existence of a hierarchy could explain the topological properties of the network.

On the other hand, link prediction can also be used to predict future possible links. Zhou and Mondragón (2004) propose the positive-feedback preference (PFP) model and reproduced the degree distribution and centrality more accurately using internet-history data. Guns (2011) improve the prediction accuracy using neighbor-based predictors in bipartite networks compared to the unipartite counterparts. Bliss et al. (2014) propose a new link predictor by combining topological features and node attributes in the Twitter reciprocal reply network and observes high accuracy and fast convergence for the top twenty predicted links. Also, Castilho et al. (2019) propose determining the constants of the mean-variance analysis using machine learning and a new weighted link prediction in stock networks.

2.3 Portfolio optimization

Merton (1980) and Jobson and Korkie (1980) insist that the large estimation error for expected return and covariance in Markowitz's theory came from the poor performance of that. Against the Markowitz's modern portfolio theory, theoretical and mathematical models for more effective portfolio diversification are proposed. Black and Litterman (1990) introduce a model that reflect investors' specific views of asset returns. It aims to solve the problems that investors encounter in which the difficulty to estimate expected returns, unlike the covariances. Also, Rom and Ferguson (1994) introduce the post-modern theory (PMPT), using the risk of downward returns instead of the average variance of the investment returns. Meanwhile, DeMiguel et al. (2009) show that complicated portfolio strategies are not consistently better than equally-weighted portfolio rules by discovering that out of sample obtained from optimal diversification is more than offset by estimation error. Nevertheless, mathematical and theoretical models for well-diversified portfolios have emerged, some of which apply complex networks.

From the perspective of the complex network, Mantegna (1999) first propose a minimum spanning tree (MST), which is created by transforming the correlation to a distance and selecting among them. Onnela et al. (2003) makes additional hypotheses about the topology of the metric space named ultrametricity hypothesis from the trees and introduces a new asset graph. Tumminello et al. (2005) introduce filtering techniques that extract subgraphs from complex data by controlling the genus of the graph and proposed a graph, called PMFG, which contains more information because it maintains the hierarchical organization of MST, but has a large number of links. However, since it is not adequate to represent a directed network, Kenett et al. (2010) propose a Partial Correlation Planar Graph (PCPG) that deals with asymmetric interactions between system elements using partial correlation analysis.

Pozzi et al. (2013) investigate how MST or PMFG could be used to characterize the heterogeneous spreading of risk in financial markets and extract the dependency structure of financial equity and constructed a well-diversified portfolio that effectively reduced investment risk. The results showed that investing in stocks in the periphery of the filtered network has a diversified and improved ratio in terms of average return and standard deviation. Boginski et al. (2014) propose a method of clustering stocks based on correlation using a weighted market graph model to construct diversified portfolios. Also, Peralta and Zareei (2016) observe a negative correlation between asset centrality and optimal portfolio weighting in the financial market network. Then, he proposed to overweight low-central stocks to create an efficient portfolio since the portfolio composed of the strongly-connected stocks is inefficient, which is called a ρ-dependent strategy. Li et al. (2019) filter a network made up of stocks of CSI 300 in China and S&P 500 in the United States, constructing a portfolio with high-peripherality stocks, and found that the performance was improved.

Chapter 3

Time-varying Granger Causality Network

3.1 Overview

The aim of this chapter is to construct the Granger causality network that represents the financial market, especially the global currency market. Most of the previous financial networks are based on the correlation between different assets, which is not proper to reflect the cause-and-effect relations that have directionality. Instead, this chapter introduces the network based on the Granger causality of Real Effective Exchange Rate (REER) which includes directed interactions such as international trade. The causal relation is proposed by Granger (1969), which evaluates the significance of cause-and-effect between two time series.

In general, the interest rate policies and fluctuations of the exchange rate of one country affect the currency value of neighboring countries. Therefore, we use the REER data reflecting the interactions between countries from several countries around the world provided by the BIS (Bank for International Settlements). If a causality network constructed artificially based on a mathematical method can represent macroeconomic conditions reflecting information of the exchange rates or the financial market, the links between the nodes of the network will have important implications in other areas of research. Therefore, the economic implication of causality networks based on REER data is analyzed in detail and it provides meaningful results.

Once the Granger causality network is constructed, descriptive statistics of the return series is observed to identify whether the tails of the distribution are fatter than that of Gaussian distribution. Also, the cross-sectional topology of the network is examined. For the three financial crises in the experiment period, the structures of the network are studied before and after the crisis. Since the structural change of the network would have occurred significantly in the financial crisis, we confirm that the Granger causality network represents the global currency market well. Furthermore, the time-varying properties of the network are observed by moving window method to understand the structure. The size of the moving window is 504 trading days, roughly 2 years, which is the size that the statistical test has a robust result.

3.2 Architecture of Time-varying Granger Causality Network

3.2.1 Granger Causality Direction

Granger Causality Test is the most popular and general method of describing the causal relationship between random variables (Granger, 1969). The test can be simply expressed in the case of two variables. Consider the two stationary log-return time series X_t, Y_t and then the base model can be defined as,

$$X_t = \sum_{j=1}^{m} a_j X_{t-j} + \sum_{j=1}^{n} b_j Y_{t-j} + \epsilon_t^{'} \tag{3.1a}$$

$$Y_t = \sum_{j=1}^{m} c_j X_{t-j} + \sum_{j=1}^{n} d_j Y_{t-j} + \epsilon_t^{''} \tag{3.1b}$$

where a_j, b_j, c_j, d_j is the coefficient of the model; $\epsilon_t^{'}, \epsilon_t^{''}$ are standard normal random variables; and m and n are length of lags, respectively. Note that the above equation is a simple case and the definition of causality implies that Y_t granger cause X_t if one of b_j is not zero. Eq.(3.1a) and Eq.(3.1b) indicate how the past informations of X_t and Y_t affect one another. Therfore, when $b_j = 0$, Eq.(3.1a) implies the self-forecast by using only historical data of X_t, whereas Eq.(3.1b) forecasts by using the past information both X_t and Y_t. Then, the statistical inference on the existence of causality is performed by the F-test on these two equations based on the residual sum of squares (RSS) of both equations as follows.

$$F\text{−statistics} = \frac{(RSS(m) - RSS(m,n))(T - m - n - 1)}{RSS(m,n) \times n} \tag{3.2}$$

where m and n refer to the time lags. Note that the time lags are set of $m = n = 1$ by assuming an efficient market hypothesis (EMH) that the current exchange rate reflects all previous information. Then, two types of Granger causality networks are derived through F-statistics: *unweighted* and *weighted* networks.

3.2.2 Granger Causality Network

An *unweighted* network at time t can be represented by a diagraph, $G_t = (V_t, E_t)$, where V_t is a set of countries and E_t is a set of edges with Granger causality direction. That is, an element in the adjacency matrix, A_{ij}, has a value of 1 when node i granger cause j and vice versa. In general, a Granger causality network has a boolean edge from node i to j at time t when the F-test rejects the null hypothesis at 1% significance level such that,

$$E_t(i,j) = \begin{cases} 1, & \text{if } i \neq j \text{ and } H_0(i,j) \text{ is rejected} \\ 0, & \text{otherwise} \end{cases} \tag{3.3}$$

Note that the null hypothesis, $H_0(i,j)$, is accepted if and only if no lagged values of x are retained in the regression.

In this dissertation, a link prediction method is proposed using a *weighted* network based on the F-statistics and eta squared η^2. In particular, we confirm whether node x granger causes node y using the F-statistics obtained from the Granger causality test. Then, the strength of the causality can be described by the effect size, η^2, since the p-value based on the F-statistics in the conventional null hypothesis testing does not provide the evidence for the differences among the groups (Nakagawa and Cuthill, 2007; Sullivan and Feinn, 2012). The effect size is a quantitative measure to represent the differences between the two groups to be compared in the experiment.

In the General Linear Model (GLM), Pearson's correlation coefficient, r, and the coefficient of determination, R^2, are similar to η^2 in terms of calculation or interpretation (Maxwell et al., 1981; Hayes, 2009). For example, the value of η^2 of 0.2 means that the independent variable accounts for 20% for the change of the value in the dependent variable (Olejnik and Algina, 2003). Brown (2008) showed that the variables in the two-way ANOVA analysis of anxiety and tension help interpreting the results since the value of η^2 represents the relative degree of each main effect and their interactions. In addition, many studies have introduced and reported the effect sizes along with statistics or p-values when showing statistical significance test results. The value of η^2 can be calculated as follows. At first, in the Granger causality test, we compute the F-statistics such that,

$$F = \frac{MS_B}{MS_W} = \frac{SS_B/(k-1)}{SS_W/(N-k)} \tag{3.4}$$

where MS_B and MS_W refers to the mean squares between and within groups; SS_B and SS_W refers to the sum of squares between and within groups; N and k are the number of observations and the groups, respectively. In this case, given that $SS_B = F * MS_W * (k-1)$, $SS_W = MS_W * (N-k)$, η^2 can be expressed in terms of the F-statistics as follows.

$$\eta^2 = \frac{SS_B}{SS_T} = \frac{(k-1)F}{(k-1)F + (N-k)} \tag{3.5}$$

In this regard, the higher the value of η^2, the stronger the causality is. Also, as the effect size represents the similarity between the nodes x and y, the value of η^2 reflects how much the change in the currency value of one country can be explained by that of another country. Therefore, it can help predict future possible links in networks.

Hence, in the link prediction perspective, we compare and analyze the prediction methods using the weighted network as well as those of unweighted similarity measures that are frequently used in many previous studies Lü et al. (2009); Adamic and Adar (2003); Fouss et al. (2007). The larger the η^2-value is, the higher the statistical significance of the assumption that the current returns are better predicted when using historical data from different time series rather than self-forecast. Hence, we assume that high η^2-values contain relatively more significant and strong causality direction information than low η^2-values. In the case of weighted network, an edge is assigned in the same manner as in Eq.(3.3) with different values of weights such that,

$$E_t(i,j) = \begin{cases} \eta_t^2(i,j), & \text{if } i \neq j \text{ and } H_0(i,j) \text{ is rejected} \\ 0, & \text{otherwise} \end{cases} \tag{3.6}$$

3.2.3 Measures of Granger Causality Network

For these Granger causality networks, we can compute some topology metrics at time t which determine the network size, in terms of Total Degree (TD), Average Path Length (APL), and Diameter (DM) as follows.

$$TD = \frac{1}{|V_t|^2} \sum_{x \in V_t} (k_x^{out} + k_x^{in}) \tag{3.7a}$$

$$APL = \frac{1}{|V_t|(|V_t|-1)} \sum_{i \neq j} d(v_i, v_j) \tag{3.7b}$$

$$DM = \max_{i,j \in V_t} \{d(v_i, v_j)\} \tag{3.7c}$$

where $|V_t|$, k_x^{out}, k_x^{in}, and $d(v_i, v_j)$ refer to the number of nodes in the network at time t, node i's number of out-degrees, node i's number of in-degrees, and shortest path between node i and node j, respectively. The APL in Eq.(3.7b) measures the average length of the shortest path for all node pairs, which is one of the most robust measures of network topology. And DM in Eq.(3.7c) measures the longest length for all shortest paths. Therefore, the larger the TD and the smaller the APL and DM, the more nodes of the network are connected to each other, which means that the transfer of information is faster.

3.3 Data description

The nominal effective exchange rate (NEER) is calculated by the weighted average of the changes in the currency value of each major trading currency. In other words, the exchange rate of the local currency in terms of NEER is defined as the relative change in the exchange rate of the trading country as compared with the base year. However, the external value of the national currency is also can be changed by the difference in the inflation rate between countries. Therefore, the real effective exchange rate (REER) that incorporates the price level of trading partner countries is used in this dissertation. REER measures the real purchasing power of currency; it indicates that the local currency is overvalued against the leading trading country if it is above 100, and vice versa.

The data period is a total of 3633 days from 2003-01-29 to 2016-12-30 for 61 countries provided by The Bank for International Settlement (BIS) as summarized in Table 3.1. Then, it is divided into four groups for comparison: America, Europe, Asia, and PIIGS(Portugal, Italy, Ireland, Greece, Spain). As stated in Polanco-Martínez et al. (2018), the PIIGS countries had a major impact on the European Monetary Union(EMU) during the European debt crisis. For the same reason, they are separated from other European countries for further analysis.

REER is obtained from NEER by computing weights for import, w_i^m, and export, w_i^x as follows (Turner and vant Dack, 1993; Klau and Fung, 2006).

$$w_i^m = \frac{m_j^i}{m_j} \tag{3.8a}$$

$$w_i^x = \left(\frac{x_j^i}{x_j}\right)\left(\frac{y_i}{y_i + \sum_h x_h^i}\right) + \sum_{k \neq i}\left(\frac{x_j^k}{x_j}\right)\left(\frac{x_i^k}{y_k + \sum_h x_h^k}\right) \tag{3.8b}$$

$$W_i = \left(\frac{m_j}{x_j + m_j}\right) w_i^m + \left(\frac{x_j}{x_j + m_j}\right) w_i^x \tag{3.8c}$$

where $x_j^i(m_j^i)$ is the amount of export(import) from country j to i; $x_j(m_j)$ is the amount of total export(import) of country j; W_i is the weighted average of total export and import; y_i is the home supply of the gross domestic product in country i; and $\sum_h x_h^i$ is the summation of exports from all countries except j to country i. The weight for import in Eq.(3.8a) is relatively simple, whereas that for export in Eq.(3.8b) is somewhat complicated. In Eq.(3.8b), the first term on the right-hand side indicates the export weight from country j to country i and the openness of country i. If the country i is less open, then the exports to country i in other countries belonging to the basket of country j will be less, which eventually means that the weight of country i increases for country j. Furthermore, the second term on the right-hand side is the third market where country i and j compete simultaneously. If the amount of exports from country i to k is large, country j should compete more

Table 3.1: List of 61 countries in four categories

Continent	Symbol	Country	Continent	Symbol	Country
America	CA	Canada	Europe	AT	Austria
	US	United States		BE	Belgium
	AR	Argentina		BG	Bulgaria
	BR	Brazil		CH	Switzerland
	CL	Chile		CY	Cyprus
	CO	Colombia		CZ	Czech Republic
	MX	Mexico		DE	Germany
	PE	Peru		DK	Denmark
	VE	Venezuela		EE	Estonia
Asia Oceania Africa	AE	United Arab Emirates		FI	Finland
	CN	China		FR	France
	HK	Hong Kong SAR		GB	United Kingdom
	ID	Indonesia		HR	Croatia
	IL	Israel		HU	Hungary
	IN	India		IS	Iceland
	JP	Japan		LT	Lithuania
	KR	Korea		LU	Luxembourg
	MY	Malaysia		LV	Latvia
	PH	Philippines		MT	Malta
	SA	Saudi Arabia		NL	Netherlands
	SG	Singapore		NO	Norway
	TH	Thailand		PL	Poland
	TR	Turkey		RO	Romania
	TW	Chinese Taipei		RU	Russia
	DZ	Algeria		SE	Sweden
	ZA	South Africa		SI	Slovenia
	AU	Australia		SK	Slovak Republic
	NZ	New Zealand		XM	Euro area
			PIIGS	GR	Greece
				IE	Ireland
				IT	Italy
				PT	Portugal
				ES	Spain

with country i, which yields growing of the weight of country i. For instance, when the export ratio from country j to k is large, the export from country i to country k will have more impact on country j. In this context, NEER of the currency i is obtained as a weighted average of the nominal exchange rate as follows.

$$NEER_i = \prod_{j=1}^{N} \left(\frac{e_i}{e_j} \right)^{w_j} \tag{3.9}$$

where $j = 1, \ldots, N$; e_i is the nominal exchange rate of home country i to the numeraire(USD); e_j is the nominal exchange rate of foreign country j to the numeraire(USD); w_j is the bilateral trade weight between country i and trading parter j in Eq.(3.8c); and N is the number of countries in the group of trading partner. Lastly, REER considering the inflation rate of each country can be expressed as follows.

$$REER_i = \prod_{j=1}^{N} \left(\frac{e_i}{e_j} \times \frac{P_i}{P_j} \right)^{w_j} \tag{3.10}$$

where P_i is the inflation rate of country i.

Based on the definition of REER, the existence of Granger causality direction from a country to country implies the coupling or decoupling of their past economic conditions depending on the sign of F-statistic. Especially, the strength of (de)coupling is stronger when the η^2 in Eq.(3.2.2) is larger, which suggests longer survival of the edge in the future. Also, suppose a Granger causality direction exists from country A to country B. If there is a Granger causality direction from country C to country A and from country B to country D, then countries A, B, C, and D can be regarded as one economic cluster. Thus, if there is no Granger causality direction from country C to D in present, one can assume that it may occur in the future.

In this context, the information extracted from the REER is examined whether it helps predict the links in the future.

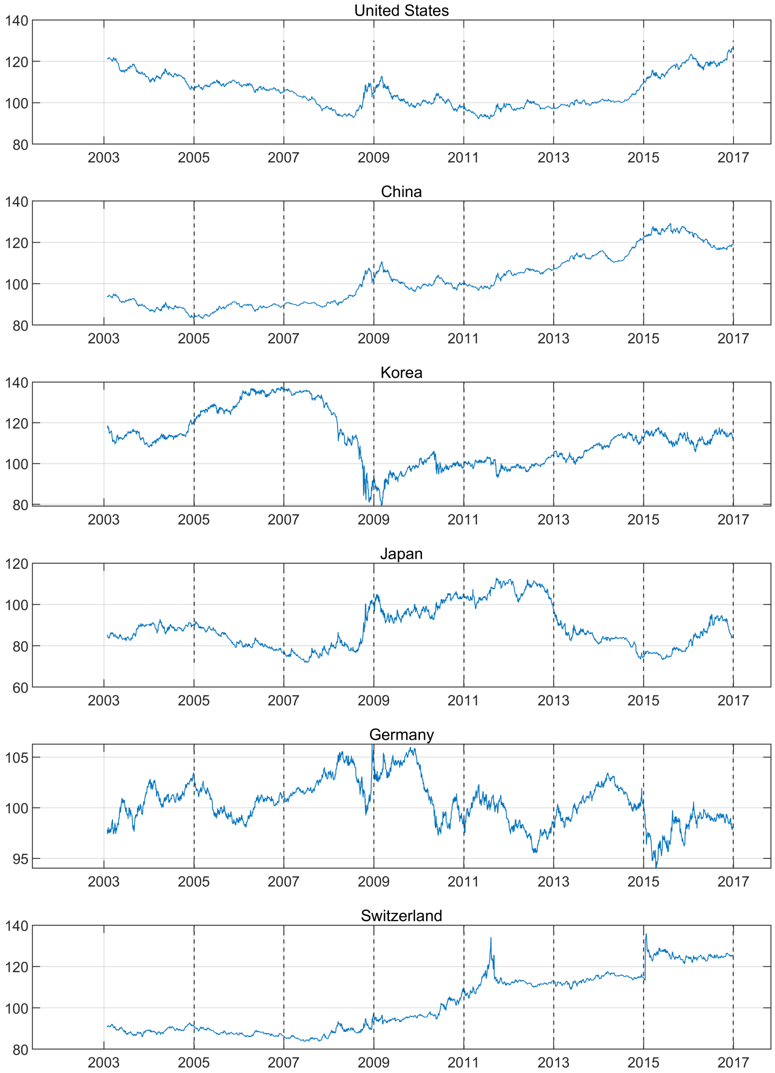
3.4 Results

3.4.1 Cross-sectional Topology of REER Networks

The Granger causality network constructed with REER data is considered as the mapping of the monetary value of countries according to the exchange rate fluctuation of the countries. This enables the identification of relationships between countries within the financial system. Therefore, the topology of the Granger causality network should follow that of the exchange rate network based on the REER data.

At first, the price series of REER for a few representative countries, including the United States, China, Korea, Japan, Germany, and Switzerland, is plotted in Figure 3.1. Note that the 12 years of data is divided into six separate periods so that each can represent approximately two years of REER information. The points of separation are plotted in black dash vertical lines throughout the analysis. As shown in Figure 3.1, the price series varies in each country. Notably, during the default of Lehman Brothers in September-2008, all the countries except Korea exhibited the increment in REER.

Since a currency with $REER > 100$ represents the relatively overvalued currency as stated in Chapter 3.2, the case of Korea indicates that the value of the currency has fallen significantly compared to other currencies during the financial


United States

China

Korea

Japan

Germany

Switzerland

Figure 3.1: Evolution of the price series of REER for six representative countries

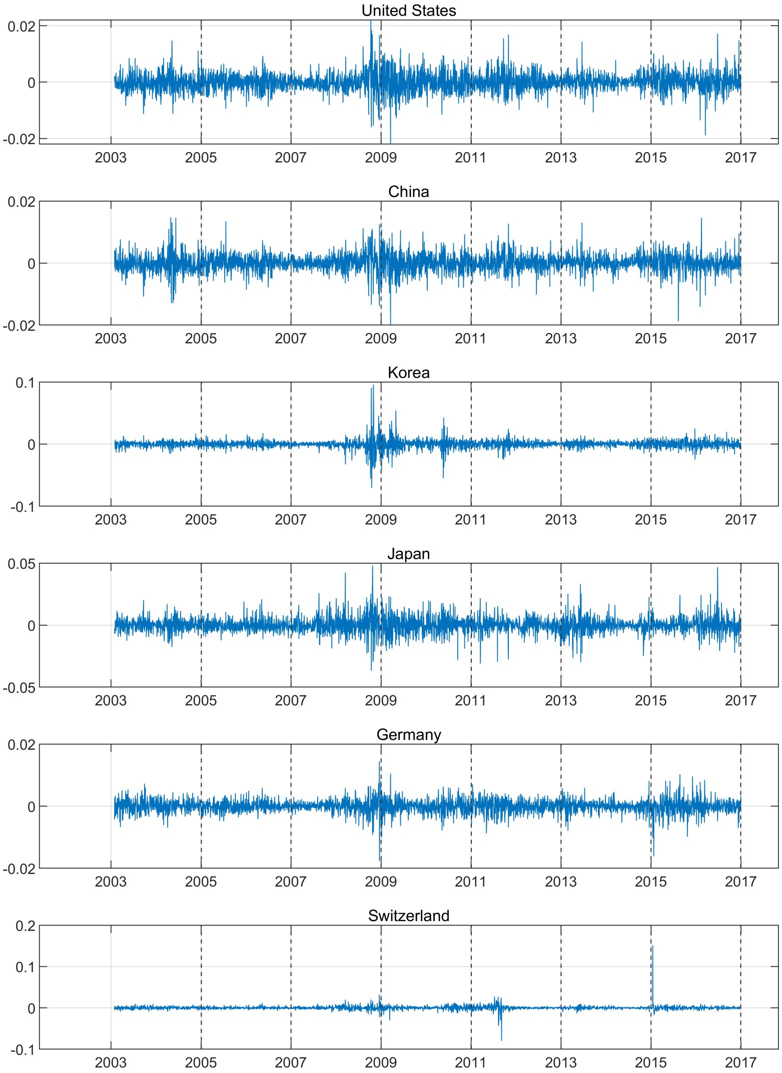


Figure 3.2: Evolution of the return series of REER for six representative countries

crisis. Such implication is reasonable given that Korea is an export-dependent country with a relatively smaller domestic economy than that of others. In fact, the Won to Dollar exchange rate(KRW/USD) at this time was the record high of 1,600 won per dollar. Besides, the REER of Germany shows a significant decrement during the European debt crisis in 2010. In contrast, as the investment capital sought a safe asset during the European debt crisis, the currency value of Switzerland shows a sharp rise. In this regard, the REER-driven network can be expected to produce an implicit result since its evolution well represents the relative value of the currency and the relationship between countries in different economic circumstances. Then, the return series of REER for the same representative countries is shown in Figure 3.2. Similar to the price series in Figure 3.1, the return series also varies over time in different countries. In particular, higher volatility of REER during the global financial crisis is observed. Also, since the return series determines the presence of edges in networks, the descriptive statistics of the return series for all countries is investigated and summarized in Table 3.2.

A leptokurtic distribution has positive excess kurtosis whose tails are fatter than that of Gaussian distribution (Mantegna and Stanley, 1999; Back, 2006). Specifically, the high kurtosis in return distribution is an indicator of the fat-tailed distribution or existence of outliers. Table 3.2 shows that the kurtoses of all countries are greater than 3, which implies the positive excess kurtosis in return distribution of REER. However, some countries show unusually high kurtosis, which generally has an extremely high positive or negative return in comparison to those of other countries. For instance, Argentina whose kurtosis is 684.11, Venezuela whose kurtosis is 1222.77, Switzerland whose kurtosis is 290.36, Iceland whose kurtosis is 206.80

Table 3.2: Descriptive statistics of REER return series

Symbol	Min	Mean	Max	Median	Std	Skewness	Kurtosis	Jarque-Bera
CA	−0.03082	0.00004	0.03966	0.00011	0.00564	−0.13	6.27	**1.6225E + 03*****
US	−0.02202	0.00001	0.02207	−0.00009	0.00331	0.18	6.5	**1.8714E + 03*****
AR	−0.32096	−0.00044	0.04624	−0.00035	0.00814	−17.99	684.11	**7.0400E + 07*****
BR	−0.10895	0.00009	0.08513	0.00024	0.01001	−0.26	14.07	**1.8574E + 04*****
CL	−0.05118	0.00003	0.03261	0	0.0063	−0.4	7.03	**2.5524E + 03*****
CO	−0.04504	0.00004	0.05695	0.0001	0.00718	−0.12	9.22	**5.8625E + 03*****
MX	−0.10801	−0.00017	0.0709	0.00011	0.00761	−1.12	26.75	**8.6120E + 04*****
PE	−0.06017	0.00003	0.05138	0	0.00365	−0.64	41.62	**2.2596E + 05*****
VE	−0.69367	−0.00045	0.18412	−0.00005	0.01615	−32.14	1222.77	**2.2578E + 08*****
AE	−0.02142	0.00002	0.01735	−0.00009	0.00297	0.02	5.87	**1.2470E + 03*****
CN	−0.01991	0.00007	0.01478	0.00009	0.00285	−0.12	6.71	**2.0912E + 03*****
HK	−0.01716	−0.00001	0.01449	0	0.00241	−0.1	6.25	**1.6030E + 03*****
ID	−0.06783	−0.00013	0.05042	−0.0001	0.00532	−0.4	20	**4.3851E + 04*****
IL	−0.03576	0.00008	0.02638	0.00017	0.00446	−0.27	8.5	**4.6182E + 03*****
IN	−0.03782	−0.0001	0.03761	−0.00004	0.00465	−0.02	11.47	**1.0857E + 04*****
JP	−0.03673	0	0.04807	−0.00023	0.00632	0.37	7.93	**3.7611E + 03*****
KR	−0.07047	−0.00001	0.09598	0.00023	0.00704	0.6	30.34	**1.1332E + 05*****
MY	−0.02019	−0.00005	0.02772	0	0.00379	0.16	7.46	**3.0214E + 03*****
PH	−0.01751	0.00002	0.02296	0.00009	0.00348	−0.01	5.46	**9.1875E + 02*****
SA	−0.02075	0	0.01722	0	0.00283	−0.06	6.07	**1.4318E + 03*****
SG	−0.01413	0.00005	0.01117	0	0.00224	−0.07	7.42	**2.9637E + 03*****
TH	−0.01907	0.00005	0.01846	0	0.00286	0.07	6.13	**1.4847E + 03*****
TR	−0.05235	−0.0002	0.0678	0.00016	0.00736	−0.51	10.19	**7.9776E + 03*****
TW	−0.01579	0.00001	0.01951	−0.00009	0.00257	0.17	7.44	**3.0038E + 03*****
DZ	−0.03353	−0.00008	0.02503	−0.00011	0.00378	−0.54	9.94	**7.4604E + 03*****
ZA	−0.07687	−0.00012	0.04881	0.0001	0.00953	−0.48	6.73	**2.2419E + 03*****
AU	−0.07027	0.00005	0.06551	0.00018	0.00675	−0.62	13.75	**1.7725E + 04*****
NZ	−0.04672	0.00005	0.03166	0.00041	0.00635	−0.51	6.53	**2.0359E + 03*****
AT	−0.00989	0	0.00785	0	0.00119	−0.39	8.64	**4.9078E + 03*****
BE	−0.01346	0	0.01133	0	0.00167	−0.28	7.48	**3.0905E + 03*****
BG	−0.01378	0.00002	0.00997	0	0.00167	−0.12	7.74	**3.4115E + 03*****
CH	−0.0798	0.00009	0.15135	0	0.00484	7.07	290.36	**1.2527E + 07*****
CY	−0.016	0.00001	0.0107	0.0001	0.00193	−0.26	6.96	**2.4104E + 03*****
CZ	−0.04465	0.00004	0.02647	0	0.00391	−0.63	13.05	**1.5516E + 04*****
DE	−0.01775	0	0.0144	0	0.00209	−0.31	8	**3.8378E + 03*****
DK	−0.01621	0.00001	0.01374	0	0.0018	−0.26	8.62	**4.8224E + 03*****
EE	−0.01295	0.00001	0.02	0	0.00163	0.14	14.88	**2.1384E + 04*****
ES	−0.01274	0.00001	0.01041	0	0.00151	−0.27	7.97	**3.7790E + 03*****
FI	−0.0174	0.00001	0.01648	0	0.00202	−0.17	8.9	**5.2781E + 03*****
FR	−0.01448	0	0.01178	0	0.00174	−0.3	7.53	**3.1652E + 03*****
GB	−0.06209	−0.00008	0.02175	0	0.00453	−1.22	17.59	**3.3131E + 04*****
GR	−0.01377	0.00001	0.00923	0	0.00159	−0.3	7.41	**2.9946E + 03*****
HR	−0.01195	0	0.01745	0	0.00199	0.21	8.64	**4.8458E + 03*****
HU	−0.05274	−0.00006	0.03401	0.00011	0.00614	−0.57	9.99	**7.5913E + 03*****
IE	−0.0213	0.00001	0.0193	0.0001	0.00264	−0.25	7.16	**2.6547E + 03*****
IS	−0.23654	−0.00009	0.1879	0.0001	0.00913	−1.07	206.8	**6.2861E + 06*****
IT	−0.01486	0	0.01178	0	0.0018	−0.28	7.68	**3.3656E + 03*****
LT	−0.01137	0.00002	0.02128	0	0.00154	0.46	18.46	**3.6298E + 04*****
LU	−0.00731	0	0.00568	0	0.00087	−0.32	7.67	**3.3554E + 03*****
LV	−0.0114	−0.00002	0.01894	0	0.00167	0.16	11.68	**1.1425E + 04*****
MT	−0.02454	0	0.01537	0.0001	0.00273	−0.46	9.1	**5.7519E + 03*****
NL	−0.01669	0	0.01419	0	0.00198	−0.26	7.45	**3.0389E + 03*****
NO	−0.03215	−0.00005	0.03285	0	0.00483	−0.1	7.42	**2.9634E + 03*****
PL	−0.04571	−0.00002	0.03487	0.00011	0.00584	−0.33	9.16	**5.8127E + 03*****
PT	−0.0074	0	0.00596	0	0.00092	−0.26	7.56	**3.1923E + 03*****
RO	−0.08209	−0.00005	0.04914	0.00009	0.00493	−1.49	35.82	**1.6439E + 05*****
RU	−0.19785	−0.00017	0.10443	0	0.00812	−3.3	121.81	**2.1426E + 06*****
SE	−0.02462	0	0.03047	0.00009	0.00428	−0.19	6.48	**1.8484E + 03*****
SI	−0.00881	−0.00001	0.00752	0	0.00104	−0.21	9.62	**6.6674E + 03*****
SK	−0.01675	0.00009	0.03043	0.0001	0.00223	0.64	18.3	**3.5646E + 04*****
XM	−0.03106	0.00001	0.02541	0.0001	0.00363	−0.28	7.75	**3.4632E + 03*****

Note: The star superscripts *,**,*** refer to 5%,1% and 0.1% statistical significants, respectively.

Table 3.3: Kurtoses of five countries after removal of outliers on both tails

	Country				
Percentile	Argentina	Venezuela	Switzerland	Iceland	Russia
99.99	66.01	1359.37	14.90	67.83	27.62
99.95	8.50	1797.08	10.63	47.86	22.40
99.9	7.34	449.07	8.01	12.53	16.09
99.5	4.27	3.96	4.41	7.67	9.04

and Russia whose kurtosis is 121.81 show the extreme daily return of -0.32096, -0.69367, 0.15135, -0.23654 and -0.19785, respectively. In general, the cases of extremely high kurtosis are consequences of unstable economic conditions or radical currency-related policies (e.g., abolitions of Euro-Franc peg in Switzerland and Dollar-Peso peg in Argentina).

For the five countries with abnormally high kurtosis, we re-calculate their kurtosis by removing outlier. Table 3.3 shows kurtoses when values above 99.99-, 99.95-, 99.9-, and 99.5-th percentile and values below 0.01-, 0.05-, 0.1-, and 0.05-th percentile are removed from the REER distribution. Even if outliers are excluded, the return distributions of REER return in five countries still show positive excess kurtosis, which consistently implies fat-tailed distributions. The last column of Table 3.2, the statistics of the Jarque-bera test, also provides evidence of non-Gaussian distributions. The results show that the null hypothesis of normality assumption is strongly rejected for the return series of all countries. In addition, traditional economic theories tend to ignore the extreme values of the data since they assume that a return distribution is a Gaussian distribution. However, in real-world markets, extreme events are not very rare and have important implications in terms of risk

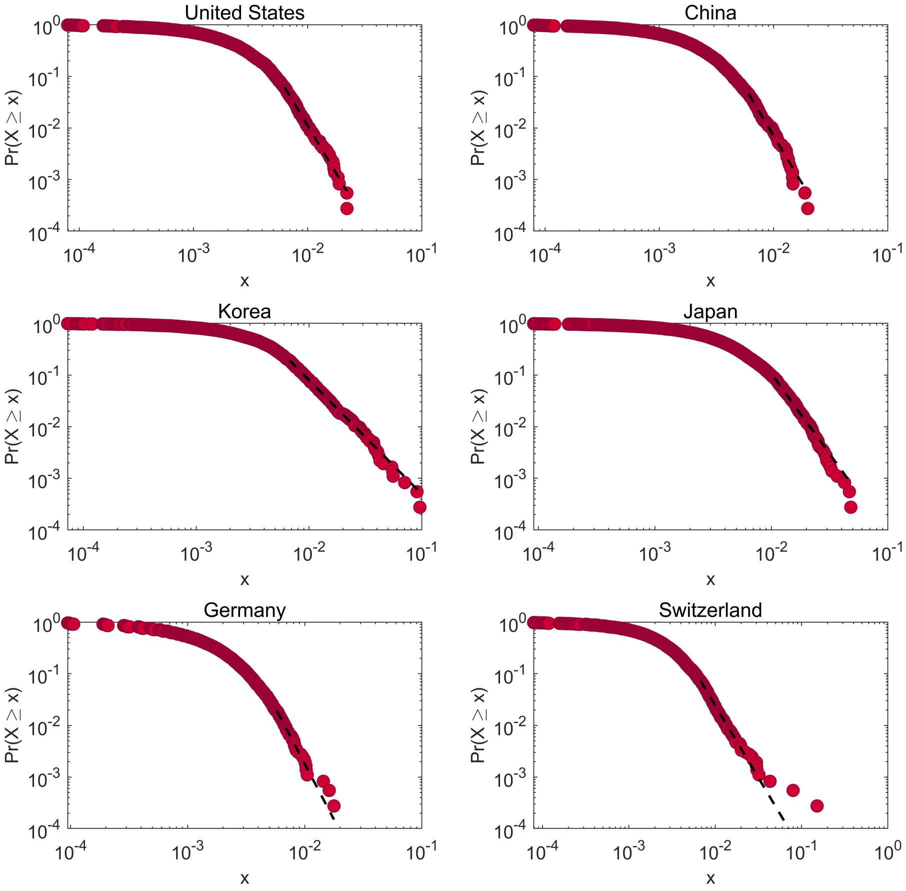


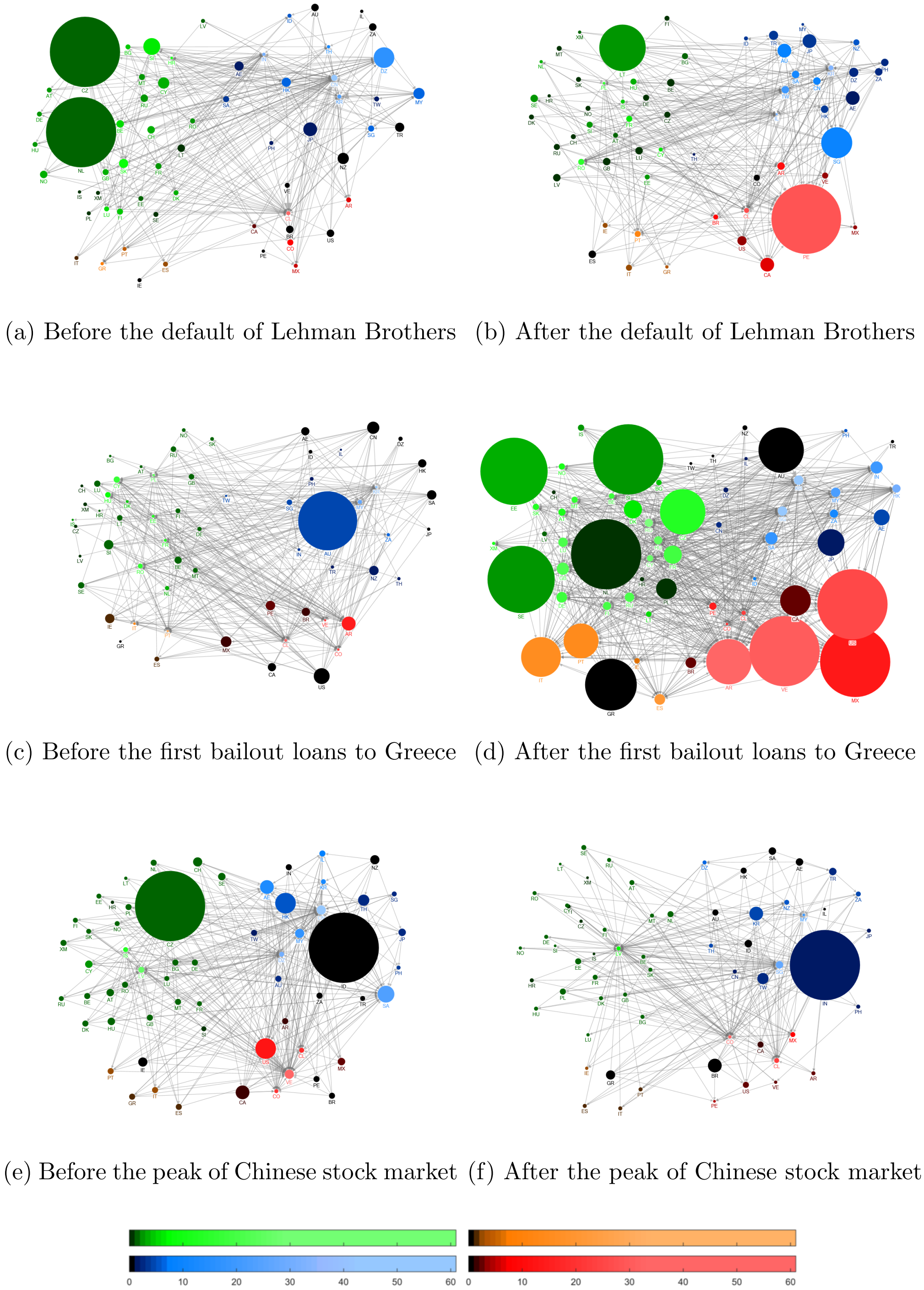
Figure 3.3: Decumulative distribution function of the REER return series for six representative countries

management. Therefore, the return series of REER is examined whether it follows the power-law distributions rather than the Gaussian distributions for understanding the extreme values of data. Figure 3.3 shows the log-log plot of the decumulative distributions of the six representative countries.

The plots in Figure 3.3 capture the tail distribution of the absolute REER return series and a black dashed line indicates the regression on the log tail probabilities, $P(|r_t| > x) \sim x^{-\gamma}$, where γ is the exponent of power-law function. And the results of power-law exponents are shown in Table 3.4. Note that γ, x_{min}, and L refer to the maximum likelihood estimate of the scaling exponent, an estimate of the lower bound of the power-law behavior, and log-likelihood of the data $x \geqq x_{min}$ under the fitted power law, respectively. The mean γ for all countries, America, Asia, Europe, and PIIGS are 4.58, 4.15, 4.65, 4.70, and 4.40, respectively. As a result, the probability that a return has an absolute value larger than x obeys the power law with $\gamma = 4.58 \pm 0.67$. Based on the return series, we construct the Granger causality networks for different cross-sectional periods of financial crises including the time-series of REER for one year before and after (252 days) the default of Lehman Brothers in 2008-09-15 (Sub-prime mortgage crisis), the day when Greece received the first tranche of the bailout loans in 2010-05-07 (European debt crisis), and the peak of Chinese stock market in 2015-06-15 (Chinese stock market turbulence). The configuration of each network is visualized in Figure 3.4. Note that the colors are assigned to each continent (America: red, Europe: green, Asia/Oceania/Africa: blue, PIIGS: orange). Also, the size of the nodes and the color intensity represent the out-degree and in-degree scales, respectively.

Table 3.4: Descriptive statistics of power-law fitting

Continent	Symbol	γ	x_{min}	L	Continent	Symbol	γ	x_{min}	L
America	CA	5.30	0.0144	429.9409	Europe	AT	5.26	0.0033	428.5657
	US	4.77	0.0072	685.0991		BE	5.22	0.0041	634.8575
	AR	3.73	0.0114	750.0709		BG	4.58	0.0036	879.0600
	BR	4.09	0.0202	616.5358		CH	3.82	0.0078	803.7272
	CL	4.64	0.0123	804.0700		CY	4.93	0.0041	897.4143
	CO	4.32	0.0152	745.0684		CZ	4.32	0.0090	595.7282
	MX	3.60	0.0143	651.0134		DE	5.25	0.0058	371.4364
	PE	3.68	0.0069	794.9846		DK	5.57	0.0056	264.5160
	VE	3.21	0.0051	734.7322		EE	4.44	0.0038	703.5593
	Mean	4.15	0.0119	690.1684		FI	5.26	0.0044	310.9507
	Stdev	0.63	0.0045	109.1621		FR	4.77	0.0051	475.8437
Asia Oceania Africa	AE	4.92	0.0063	793.7877		GB	5.16	0.0041	682.6556
	CN	4.68	0.0060	854.8238		HR	4.31	0.0098	634.0589
	HK	4.76	0.0050	884.1211		HU	4.69	0.0033	930.0639
	ID	3.95	0.0118	633.4507		IS	4.43	0.0043	816.5817
	IL	5.06	0.0138	236.3416		LT	4.40	0.0125	786.5423
	IN	3.82	0.0092	753.4560		LU	5.35	0.0068	435.0839
	JP	5.58	0.0216	150.1600		LV	3.42	0.0168	491.8695
	KR	3.22	0.0123	679.6180		MT	5.43	0.0057	218.1319
	MY	4.27	0.0078	838.3652		NL	3.90	0.0030	1007.2937
	PH	5.86	0.0085	504.8511		NO	5.11	0.0021	756.9530
	SA	5.51	0.0093	145.0381		PL	4.92	0.0041	679.1006
	SG	3.98	0.0045	941.5227		RO	4.58	0.0057	940.7701
	TH	5.36	0.0069	564.6942		RU	4.97	0.0044	766.1293
	TR	3.95	0.0145	724.4538		SE	4.47	0.0101	720.5069
	TW	4.43	0.0052	948.2262		SI	3.86	0.0119	703.8635
	DZ	4.28	0.0091	473.5391		SK	5.29	0.0024	564.3463
	ZA	4.63	0.0197	647.4073		XM	3.86	0.0098	763.1794
	AU	4.15	0.0169	357.1834		Mean	4.70	0.0062	652.2425
	NZ	5.90	0.0192	217.9915		Stdev	0.56	0.0035	208.1878
	Mean	4.65	0.0109	597.3174	PIIGS	GR	3.07	0.0141	618.4215
	Stdev	0.73	0.0052	261.5051		IE	5.31	0.0106	554.9303
						IT	4.66	0.0027	584.5022
						PT	4.25	0.0057	490.5519
						ES	4.69	0.0075	898.2967
						Mean	4.40	0.0081	629.3405
						Stdev	0.74	0.0040	140.9016



(a) Before the default of Lehman Brothers (b) After the default of Lehman Brothers

(c) Before the first bailout loans to Greece (d) After the first bailout loans to Greece

(e) Before the peak of Chinese stock market (f) After the peak of Chinese stock market

Figure 3.4: Structural changes of the REER network between one year before and after the financial crises

Higher the out-degree, the larger the size of the node where the relative size is scaled within the same crisis period for comparison. Higher the in-degree, the brighter the color as shown in the color bar in Figure 3.4. Moreover, we further analyze the changes in out-degrees of countries since a node with the highest out-degree indicates a leading country in the currency market with severe influences on other countries. The changes in out-degrees are summarized in Table 3.5c by showing the top 10 highest out-degree countries for each cross-section.

In the case of the Sub-prime mortgage crisis, the countries in America show the high out-degrees after the default of Lehman Brothers as illustrated in Figure 3.4b, whose size is relatively small in Figure 3.4a. More specifically, as shown in Table 3.5a, the countries in Europe and Asia including Czech, Lithuania, Algeria, Poland, and Japan show the top 10 highest out-degrees before the default, while any country in America does not exist. After the default, however, we could detect the top 10 highest out-degrees from countries in America, including Peru, Canada, and the United States. Based on the emergence of countries in America, it seems that the Sub-prime mortgage crisis, originated from the United States, influences the global currency market, and such phenomenon can be observed through the changes in the structure of causality network.

Similar results can be found in the case of the European debt crisis where the out-degrees and in-degrees of countries in Europe and America are substantially increased after the bailout loans for Greece as illustrated in Figure 3.4c and 3.4d. Furthermore, the increment of the size of PIIGS countries, orange nodes, is observed. More specifically, as shown in Table 3.5b, the countries in Asia/Oceania and America

Table 3.5: Top 10 countries with the highest out-degrees in one year before and after the beginning of each financial crises

(a) Default of Lehman Brothers

	Before					After			
Rank	indegree	outdegree	country	continent	Rank	indegree	outdegree	country	continent
1	1	22	CZ	Europe	1	28	17	PE	America
2	1	19	LT	Europe	2	2	15	HU	Europe
3	16	11	DZ	Africa	3	9	13	SG	Asia
4	6	10	PL	Europe	4	3	9	JP	Asia
5	1	9	JP	Asia	4	1	9	AE	Asia
6	6	8	MY	Asia	4	8	9	AU	Oceania
6	0	8	NZ	Oceania	4	6	9	CA	America
6	3	8	CY	Europe	8	2	7	DZ	Africa
9	1	7	AE	Asia	8	3	7	TR	Asia
9	6	7	HK	Asia	8	3	7	US	America
9	0	7	TR	Asia					
9	8	7	PT	PIIGS					

(b) First bailout loans to Greece

	Before					After			
Rank	indegree	outdegree	country	continent	Rank	indegree	outdegree	country	continent
1	4	25	AU	Oceania	1	13	32	MX	America
2	0	15	US	America	2	0	31	LT	Europe
3	14	14	AR	America	3	25	29	US	America
4	0	13	CN	Asia	4	2	28	PL	Europe
5	1	12	MX	America	4	30	28	VE	America
6	2	11	PE	America	6	2	26	NO	Europe
6	0	11	SA	Asia	6	3	26	EE	Europe
6	1	11	PL	Europe	8	0	24	RU	Europe
6	1	11	NZ	Oceania	9	0	23	AU	Oceania
6	1	11	SE	Europe	9	60	23	AR	America
					9	13	23	CZ	Europe

(c) Peak of Chinese stock market

	Before					After			
Rank	indegree	outdegree	country	continent	Rank	indegree	outdegree	country	continent
1	1	26	CZ	Europe	1	1	30	IN	Asia
2	0	18	ID	Asia	2	0	9	BR	America
3	12	11	US	America	2	4	9	KR	Asia
3	5	11	HK	Asia	4	2	8	TW	Asia
5	23	10	SA	Asia	5	0	7	RU	Europe
6	1	9	CA	America	6	0	6	ID	Asia
6	13	9	AE	Asia	6	0	6	SA	Asia
8	2	8	TH	Asia	6	0	6	AE	Asia
9	60	7	VE	America	6	27	6	SG	Asia
9	17	7	MY	Asia	6	2	6	TR	Asia
9	60	7	CN	Asia					
9	0	7	SE	Europe					
9	0	7	NZ	Oceania					
9	1	7	CH	Europe					

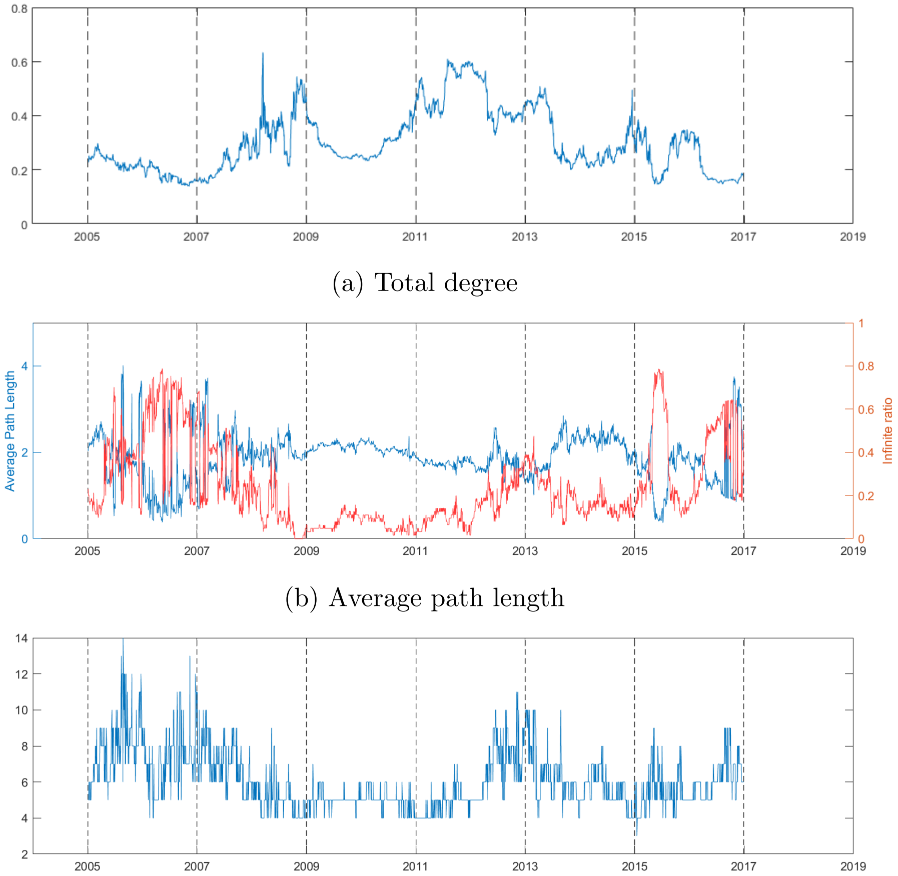
such as Australia, the United States, Argentina, and China show the highest out-degrees prior to the bailout. However, we could observe the highest out-degrees from newly entered countries in Europe, including Lithuania, Norway, Estonia, Russia, and Czech after the bailout.

The case of the Chinese stock market turbulence shows somewhat different results. As shown in Figure 3.4e and 3.4f, we could not observe any particular increment of the out-degrees of countries in Asia or changes in the structure of the network between before and after the peak of Chinese stock market. Unlike the notable increment of the out-degrees of the United States and PIIGS countries during the Sub-prime mortgage and European debt crisis, respectively, the expected increment of China during the Chinese stock market turbulence is not detected in Table 3.5c, which indicates no influence of event to the global currency market. Interestingly, the increment of the out-degrees is observed in India, Korea, and Chinese Taipei whose locations and trade are close to China. Such results are understandable in a sense that the Chinese stock market turbulence is a relatively local event, which yields a limited effect on the global financial market, in comparison to two other major financial crises. In conclusion, the Granger causality network based on REER can implicitly serve as a representation of the global currency market by presenting the changing influences of countries and continents with respect to the different financial crises.

3.4.2 Time-varying Properties of REER Networks

The time-varying properties of REER networks is analyzed by sliding 504 days (approximately two years) moving window from the beginning to the end of the experiment periods. At first, the network topology based on the total degree, average path length, and diameter is illustrated in Figure 3.5. Figure 3.5a shows the evolution of total degree defined in Eq.(3.7a). Interestingly, the total degree is noticeably decreased at the beginning of 2008, mid-2010, and the beginning of 2015, and each point is on the verge of an outbreak of the financial crisis in this dissertation. Specifically, the number of connections among the network is decreased before such a large financial crisis, whereas many connections are quickly reproduced after the financial crisis. Therefore, the results suggest that the evolution of the network can be used as a proxy to monitor the emergence of the financial crisis that may affect the global currency market.

Figure 3.5b shows the evolution of average path length defined in Eq.(3.7b) and infinite ratio, the rate of the infinite shortest path between two nodes in the total edge. The results show that the average path length is not short or vice versa when the total degree is high. In general, the studies in the financial network often utilize an undirected network based on correlation. In this case, the relationship between the total degree and the shortest path calculating average path length becomes more apparent. However, since this study deals with a directed network, the increased number of edges does not directly imply the decreased shortest path between nodes. In fact, since the path between many nodes is disappeared due to a disconnected directed edge, the infinite length of the shortest path is frequently observed. The



(a) Total degree

(b) Average path length

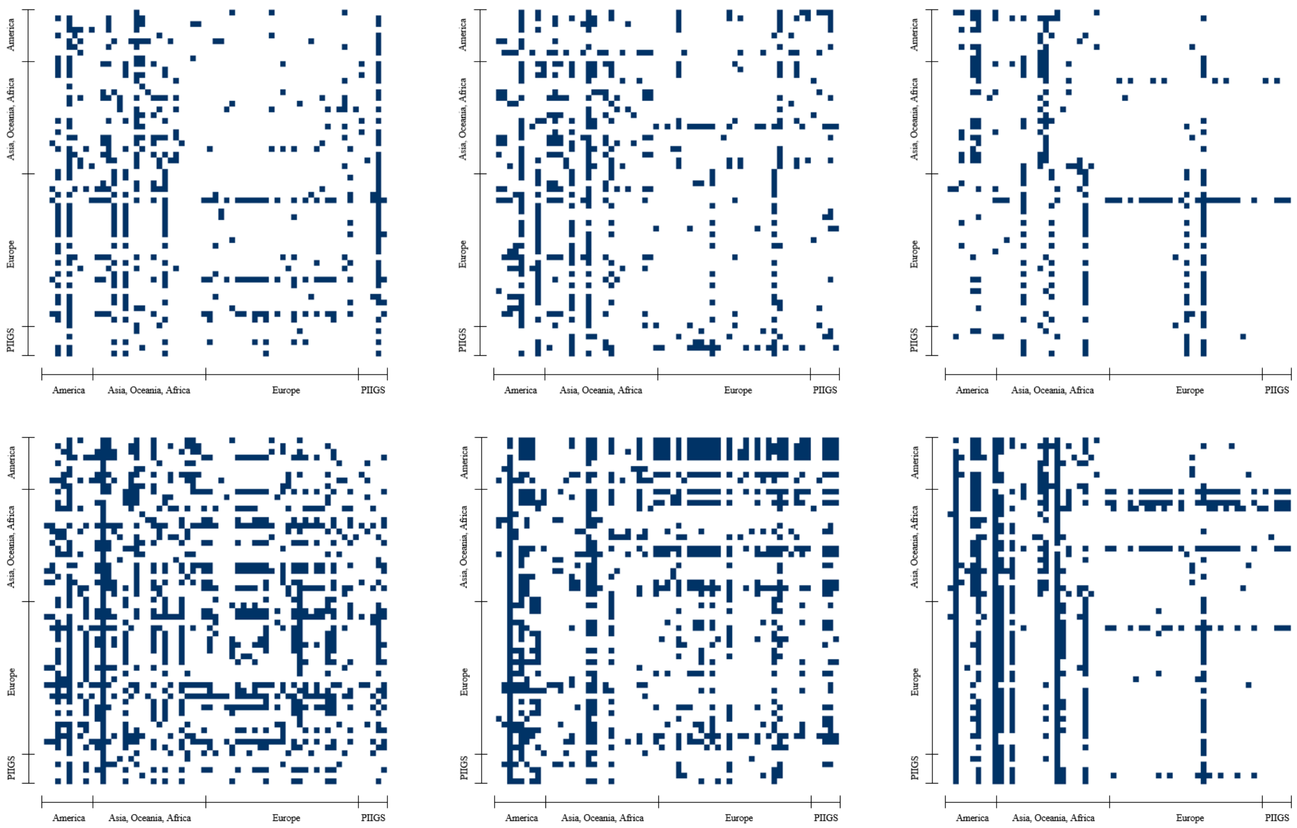
(c) Diameter

Figure 3.5: Time-vaying properties of REER networks

evolution of diameter in Figure 3.5c also does not reveal a clear correlation with the evolution of the total degree.

Then, it is important to identify the cause of the sudden increase or decrease in the total degree since it has implications for the financial markets. For instance, the sudden decrease may infer to the disconnection of hub nodes connected to many other nodes or that of entire nodes across the networks as a whole. Therefore, we analyze the connections between nodes in Figure 3.6 and 3.7. In Figure 3.6, note that the heatmaps in top and bottom refer to the start dates of the financial crises and the date with the highest total degree in its associated cross-section, respectively. In Figure 3.6a, it seems that the financial crisis originated in the United States does not affect the edge connectivity of certain countries, but rather seems to increase the overall edge of all countries. Therefore, it suggests that the increase in edge connections during the sub-prime mortgage crisis is a global phenomenon in the currency market. We can see similar results in Figure 3.6b, which shows a significant increase of in-degrees in PIIGS countries. In contrast, Figure 3.6c shows an increase of out- and in-degrees of several countries in America and Asia/Oceania/Africa while the change of Europe is very limited. And it supports that Chinese stock market turbulence is a local financial crisis with a limited impact on the global currency market, as shown in Figure 3.4e and 3.4f.

While Figure 3.6 represents only snapshots at one time point, Figure 3.7 shows the time-varying degrees of each group before and after each financial crisis and the values refer to the number of edges compared to 21 days before each time point. The black dash line is the date of the financial crisis. Also, the figures at the top, middle,

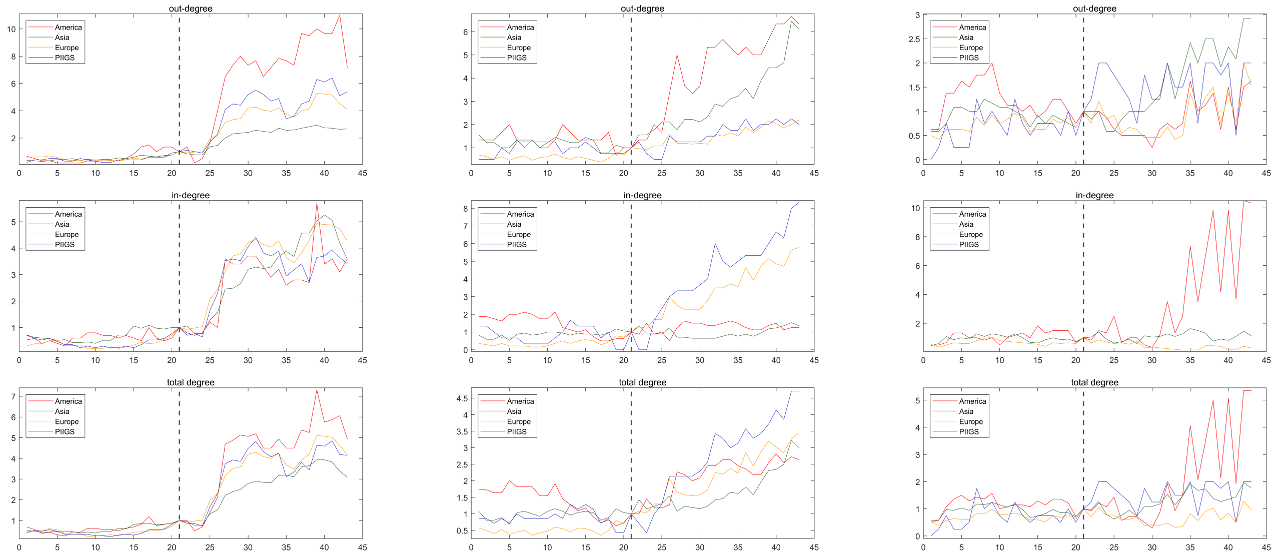


(a) Sub-prime mortgage crisis: (top) 2008-09-15 / (bottom) 2008-10-30

(b) European debt crisis: (top) 2010-5-7 / (bottom) 2011-01-03

(c) Chinese stock market turbulence: (top) 2015-6-15 / (bottom) 2015-12-17

Figure 3.6: Heatmaps of causality connections for the beginning of financial crisis and and highest total degree afterward



(a) Sub-prime mortgage crisis

(b) European debt crisis

(c) Chinese stock market turbulence

Figure 3.7: Evolutions of out-degree (top), in-degree (middle), and total degrees (bottom) of 21 days before and after the crisis

Table 3.6: Descriptive statistics: Stay in same, Inflow from others, Outflow to others

		2005 − 2006	2007 − 2008	2009 − 2010	2011 − 2012	2013 − 2014	2015 − 2016
America	No. of countries	9	9	9	9	9	9
	Stay in same	0.1976	0.3342	0.2609	0.2496	0.3037	0.3561
	Inflow from others	0.1996	0.2525	0.2291	0.2347	0.2750	0.2462
	Outflow to others	0.0518	0.1182	0.1637	0.2825	0.1333	0.0732
Asia Oceania Africa	No. of countries	19	19	19	19	19	19
	Stay in same	0.1356	0.2375	0.2376	0.1744	0.1674	0.1569
	Inflow from others	0.1020	0.2105	0.1681	0.1844	0.1230	0.1425
	Outflow to others	0.0965	0.1079	0.1310	0.2332	0.1654	0.1323
PIIGS	No. of countries	5	5	5	5	5	5
	Stay in same	0.2313	0.1080	0.0178	0.1568	0.0843	0.0334
	Inflow from others	0.0851	0.0803	0.0841	0.3638	0.1196	0.0472
	Outflow to others	0.1417	0.1474	0.1272	0.2661	0.1932	0.1098
Europe	No. of countries	28	28	28	28	28	28
	Stay in same	0.0732	0.0883	0.1014	0.2812	0.1313	0.0604
	Inflow from others	0.0654	0.0680	0.1194	0.2531	0.1377	0.0651
	Outflow to others	0.1278	0.2043	0.1716	0.2164	0.1506	0.1426

and bottom indicate the out-degrees, in-degrees, and total degree, respectively. In Figure 3.7a, we find that the Sub-prime mortgage crisis, originated in the United States, increases the number of edges in all countries, not in particular countries. That is, it implies that the increased connections in the network during the crisis are a global phenomenon across the currency market. Similar results also can be found in Figure 3.7b. At this time, the increases of in-degrees in PIIGS countries are remarkable, which reflects the properties of the European debt crisis. In contrast, we find that the increases of out- and in-degrees of several countries in America and Asia/Oceania/Africa, while the countries in Europe have little change. It supports the results that the Chinese stock market turbulence is a regional financial crisis with only a limited impact on the global currency market in Figure 3.4e and 3.4f. We summarize the results of connectivity between different continents in Table 3.6.

Note that Stay in same, Inflow from others and Outflow to others refer to the ratios of the edges within the possible connections for the same continent, incoming edges from other continents, and outgoing edges to other continents to all possible connections, respectively. For the cross-section of 2007-2008 including the Sub-prime mortgage, the values of America increases significantly in all three measures. Especially, the significant increase of edges within America is observed. Similar results can be seen in Asia/Oceania/Africa except for small increases in outflow to others. On the other hand, the values of Europe show a significant increase in outflows in the same period but little changes in other measures. Interestingly, we find the opposite results in the cross-section of 2011-2012 including the European debt crisis. There is only a small increase in outflows, while dramatic increases are observed within the same continent and inflows from others. In contrast, the values of America and Asia/Oceania/Africa show the increases in outflows but limited changes or even decreases within the same continent and inflows from others. In summary, the financial crisis that affects the structure of the global currency market increases the in-degrees of the continent whose originator associated. Finally, the cross-section of 2015-2016 including the Chinese stock market turbulence shows no particular pattern for other continents.

3.5 Summary and Discussion

In this chapter, a Granger causality network is constructed using REER data for 61 countries, provided by The Bank for International Settlement (BIS). REER refers to the weighted average of the changes in the currency value of each major trading currency, reflecting the share of exports and imports between countries and the inflation rate, while the nominal effective exchange rate does not. That is, the currency value is determined in the economy group in which one country belongs and implies the real purchasing power.

At first, the descriptive statistics of each country's return to REER are analyzed. The results show that they are leptokurtic distributions with positive excess kurtosis and have fatter tails than that of Gaussian distribution. At this time, the outliers in some countries are observed that resulted from the country's unstable economic conditions or radical currency-related policies. However, after removing the outliers, the distributions of REER still have fatter tails than that of Gaussian distribution. Secondly, the cross-sectional topology is analyzed. When we observe the network of data for a year (252 days) before and after the outbreak of financial crises, the results show that Sub-prime mortgage and European debt crisis change the network structure significantly, whereas Chinese stock market turbulence has just impact on the network partially. It means that the impact on the network depends on the properties of the crisis. Lastly, the time-varying properties of REER network are studied. The total degree decreases significantly at the beginning of 2008, mid-2010, and the beginning of 2015, which are just before the outbreak of the Sub-prime mortgage, European debt crisis, and Chinese stock market turbulence, respectively. The results indicate that the characteristics of the financial crisis on the market, depending on whether the decrement in connectivity occurred in the whole network or in parts. Also, since the number of connections is quickly reproduced after the crisis, it can be used as a proxy to monitor the emergence of the financial crisis. However, in the case of average path length, the result shows that the increased number of edges does not directly imply the decreased shortest path since the movement between nodes is limited compared to an undirected network. In other words, even if just one directed edge is disconnected, many paths between nodes disappear. Furthermore, the proportions of active flows within the possible connections are investigated. The results show that the significant increment in three measures on the continent whose financial crisis started.

In conclusion, the Granger causality network of REER favorably generates the theoretical mapping of the global currency market based on real interactions like international trades. By reminding that the purpose of this chapter is to construct the directed network that represents the financial market with exchange rate data, the results of a cross-sectional, time-varying Granger causality network provide good implications which can be applied to predict the future structure. In this context, methods for predicting future links in the network will be discussed in Chapter 4.

Chapter 4

Link Prediction

4.1 Overview

The aim of this chapter is to predict links in the future network, called Link Prediction, based on the Granger causality network constructed using REER data in Section 3.3. Link Prediction in a network has two implications. First, it is to predict missing links with existing data and the other is to predict possible links in the future based on present information. Although link prediction has been used in many areas of network analysis, there is little research on link prediction for financial markets. In financial networks, identifying visible links between two nodes, such as biological networks or social networks, is not easy unless the nodes that make up the network are institutions, banks, and countries actually deal with each other. Therefore, this chapter introduces the link prediction in Granger causality network of the financial market.

Once the causality network is constructed, we perform the link prediction. Unlike previous studies where the causality network has a boolean edge of 0 or 1, there will be differences between the edges, so proper weights should be used to accurately reflect the properties. To this end, a link prediction method is proposed called Weighted Causality Link Prediction(WCLP) using eta squared based on F-statistics in the Granger causality test. It refers to the effect size between nodes, which can be used to explain how the REER series of one country affects the other. Then, we evaluate the results by comparing the results with other benchmarks.

4.2 Benchmarks for Link Prediction

Similarity-based methods are most commonly used in the link prediction. The term, similarity, refers to how analogous a pair of nodes are on a given network structure. That is, the higher the similarity score, the more likely two nodes are connected to each other in the network. Hence, many similarity measures, used as benchmarks in this research, have been developed in previous researches as summarized in Table 4.1. Obviously, benchmarks can be divided into unweighted and weighted measures.

4.2.1 Unweighted Measures

Common Neighbors(CN) (Lü et al., 2009) is defined as,

$$s_{xy}^{CN} = \mid \Gamma(x) \cap \Gamma(y) \mid \tag{4.1}$$

where $\Gamma(x)$ denotes the set of neighbors of node x. CN considers the higher probability of having a link between two nodes when they share more common neighbors. A common neighbor refers to a node which connect with two nodes x and y simulta-

Table 4.1: Benchmarks for link prediction

Method	**Indices**	**Equation**
Common Neighbor	Common Neighbors	$s_{xy}^{CN} =\mid \Gamma(x) \cap \Gamma(y) \mid$
	Salton Index	$s_{xy}^{Salton} = \frac{\mid \Gamma(x) \cap \Gamma(y) \mid}{\sqrt{k_x \times k_y}}$
	Jaccard coefficient	$s_{xy}^{Jaccard} = \frac{\mid \Gamma(x) \cap \Gamma(y) \mid}{\mid \Gamma(x) \cup \Gamma(y) \mid}$
	Sϕrensen Index	$s_{xy}^{Sorensen} = \frac{2 \mid \Gamma(x) \cap \Gamma(y) \mid}{k_x + k_y}$
	Hub Promoted Index	$s_{xy}^{HPI} = \frac{\mid \Gamma(x) \cap \Gamma(y) \mid}{min(k_x, k_y)}$
	Hub Depressed Index	$s_{xy}^{HDI} = \frac{\mid \Gamma(x) \cap \Gamma(y) \mid}{max(k_x, k_y)}$
	Leicht-Holme-Newman Index	$s_{xy}^{LHN1} = \frac{\mid \Gamma(x) \cap \Gamma(y) \mid}{k_x \times k_y}$
	Adamic-Adar Index	$s_{xy}^{AA} = \sum_{z \in \Gamma(x) \cap \Gamma(y)} \frac{1}{log(k_z)}$
	Resource Allocation Index	$s_{xy}^{RA} = \sum_{z \in \Gamma(x) \cap \Gamma(y)} \frac{1}{k_z}$
	Preferential Attachment Index	$s_{xy}^{PA} = k_x \times k_y$
Path	Local Path(LP)	$s_{xy}^{LP} = A^2 + \epsilon A^3$
	Katz Index	$s_{xy}^{Katz} = \sum_{l=1}^{\infty} \beta^l \times \mid paths_{xy}^l \mid= \beta A_{xy} + \beta^2 (A^2)_{xy} + \ldots$
	Leicht-Holme-Newman Index	$s_{xy}^{LHN2} = \delta_{xy} + \frac{2M}{k_x k_y} \sum_{l=0}^{\infty} \phi^l \lambda^{1-l} (A^l)_{xy}$
Random Walk	Average Commute Time	$s_{xy}^{ACT} = \frac{1}{l_{xx}^{+} + l_{yy}^{-} - 2l_{xy}^{+}}$
	Cosine based on L^+	$s_{xy}^{cos^+} = cos(x, y)^+ = \frac{v_x^T v_y}{\mid v_x \mid \cdot \mid v_y \mid} = \frac{l_{xy}^{+}}{\sqrt{l_{xx}^{+} \cdot l_{yy}^{-}}}$
	Random Walk with Restart	$s_{xy}^{RWR} = q_{xy} + q_{yx}$
	SimRank	$s_{xy}^{SimRank} = C \cdot \frac{\sum_{z \in \Gamma(x)} \sum_{z \in \Gamma(y)} s_{zz}^{SimRank}}{k_x \cdot k_y}$
	Local Random Walk	$s_{xy}^{LRW} = q_x \pi_{xy}(t) + q_y \pi_{yx}(t)$
	Superposed Random Walk	$s_{xy}^{SRW} = \sum_{\gamma=1}^{t} s_{xy}^{LRW}(\gamma) = \sum_{\gamma=1}^{t} [q_x \pi_{xy}(\gamma) + q_y \pi_{yx}(\gamma)]$
	Matrix Forest Index	$S = (I + L)^{-1}$
Others	H-index	$s_{xy}^{H} = \beta A_{xy}^{H} + \beta^2 (A_{xy}^{H})^2 + \cdots + \beta^n (A_{xy}^{H})^n,\ A_{xy}^{H} = \frac{h_y + 1}{\sum_{N_x} (h_n + 1)}$
	Weighted Common Neighbor	$s_{xy}^{WCN} = \sum_{z \in \Gamma(x) \cap \Gamma(y)} w(x, z) + w(y, z)$
	Weighted Adamic Adar	$s_{xy}^{WAA} = \sum_{z \in \Gamma(x) \cap \Gamma(y)} \frac{w(x, z) + w(y, z)}{log(1 + s(z))},\ s(x) = \sum_{z \in \Gamma(x)} w(x, z)$
	Weighted Resource Allocation Index	$s_{xy}^{WRA} = \sum_{z \in \Gamma(x) \cap \Gamma(y)} \frac{w(x, z) + w(y, z)}{s(z)}$
	Generalized Degree	$s_{xy}^{GD} = q_x \pi_{xy}(t) + q_y \pi_{yx}(t)$

neously. Since the Granger causality network is a diagraph, the common neighbor, specifically, is the node that receives the common causality direction from nodes x and y. Also, nodes x and y are connected with path length of 2, based on $(A^2)_{xy}$ in Eq.(4.1).

Salton index (Chowdhury, 2010), also called as cosine similarity, measures the similarity of the inner space between two non-zero vectors as the cosine of their inter-angle as follows,

$$s_{xy}^{Salton} = \frac{\mid \Gamma(x) \cap \Gamma(y) \mid}{\sqrt{k_x \times k_y}} \tag{4.2}$$

where k_x is the degree of node x. Note that we only consider the number of out-degrees due to the directionality of diagraph. Therefore, all k_x quoted afterward refer to the out-degree.

Jaccard coefficient (Jaccard, 1901) yields one if the two sets, $\Gamma(x)$ and $\Gamma(y)$ are indentical, or zero if the two sets do not have a common element where

$$s_{xy}^{Jaccard} = \frac{\mid \Gamma(x) \cap \Gamma(y) \mid}{\mid \Gamma(x) \cup \Gamma(y) \mid}. \tag{4.3}$$

Sϕrensen Index (Sørensen, 1948), also known as F1 score, is similar to Jaccard index with binary classification where

$$s_{xy}^{S\phi rensen} = \frac{2 \mid \Gamma(x) \cap \Gamma(y) \mid}{k_x + k_y}. \tag{4.4}$$

Hub Promoted Index (HPI) (Ravasz et al., 2002) is defined as the ratio of the common neighbors of nodes x and y to the minimum number of degrees of them.

Since the denominator is the minimum of the degree, the closer the link is to the hub which has a larger outdegree, the higher the score will be as follows.

$$s_{xy}^{HPI} = \frac{\mid \Gamma(x) \cap \Gamma(y) \mid}{min(k_x, k_y)} \tag{4.5}$$

Hub Depressed Index (HDI) (Lü and Zhou, 2011; Zhou et al., 2009) is defined as the ratio of common neighbors of nodes x and y to the maximum of degrees of nodes either x or y as follows.

$$s_{xy}^{HDI} = \frac{\mid \Gamma(x) \cap \Gamma(y) \mid}{max(k_x, k_y)} \tag{4.6}$$

Leicht-Holme-Newman Index (LHN1) (Leicht et al., 2006) assigns a higher score to a pair of nodes that has more common neighbors in terms of the possible maximum number of common neighbors between them as follows.

$$s_{xy}^{LHN1} = \frac{\mid \Gamma(x) \cap \Gamma(y) \mid}{k_x \times k_y} \tag{4.7}$$

Adamic-Adar Index (AA) (Adamic and Adar, 2003) assigns a higher score to a pair of nodes as the outdegree of the common neighbor z, a common neighbor of nodes x and y, is smaller.

$$s_{xy}^{AA} = \sum_{z \in \Gamma(x) \cap \Gamma(y)} \frac{1}{log(k_z)} \tag{4.8}$$

Resource Allocation Index (RA) (Zhou et al., 2009) is motivated by resource allocation dynamics. For nodes x and y, a common neighbor z acts as a transmitter to transfer a resource between them. In case of equal allocation, the resource from node x to y can be expressed as,

$$s_{xy}^{RA} = \sum_{z \in \Gamma(x) \cap \Gamma(y)} \frac{1}{k_z} \tag{4.9}$$

Preferential Attachment Index (PA) (Barabási and Albert, 1999; Chen et al., 2005; Xie et al., 2008), one of the most renowned measure, assumes that the newly connected link for node x is proportional to k_x. In other words, the new link is likely to occur on a node already connected to many nodes such that,

$$s_{xy}^{PA} = k_x \times k_y \tag{4.10}$$

Local Path (LP) (Zhou et al., 2009; Lü et al., 2009) is defined by the equation below and gives more model fleximility into CN.

$$s_{xy}^{LP} = A_{xy}^2 + \epsilon A_{xy}^3 \tag{4.11}$$

where ϵ is a free parameter. Clearly, it is equivalent to CN when $\epsilon = 0$. Similarly, A^3 is an adjacency matrix connected to a path is length 3.

Katz Index (Katz, 1953) considers paths of all lengths for all node pairs.

$$s_{xy}^{Katz} = \sum_{l=1}^{\infty} \beta^l \times \mid paths_{xy}^l \mid = \beta A_{xy} + \beta^2 A_{xy}^2 + ... \tag{4.12}$$

where $paths^l_{xy}$ is the set of all paths with length l connecting x and y. Since β is a free parameter as weight, the smaller β, the smaller the value for the long path, which finally becomes similar to CN.

Leicht-Holme-Newman Index (LHN2) (Leicht et al., 2006) is modified from Katz index. If neighbors connected directly, called as immediate neighbors, to node x and y are similar, then node x and y are considered to be similar.

$$S_{xy} = \phi \sum_v A_{xv} S_{vy} + \psi \delta_{xy} \tag{4.13}$$

where ϕ and ψ are free parameters. Note that ϕ reduces the degree of contribution of long paths. δ_{xy} is *Kronecker function* where the value is one if $x = y$ and zero otherwise. In matrix form, the above equation is expressed as follows.

$$\begin{aligned} S &= \phi AS + \psi I \\ &= \psi (I - \phi A)^{-1} \\ &= \psi (I + \phi A + \phi^2 A^2 + \cdots) \end{aligned} \tag{4.14}$$

The expected values of $(A^l)_{xy}$, $E[(A^l)_{xy}]$ is $(\frac{k_x k_y}{2M})\lambda_1^{l-1}$ where λ_1 is the largest eigenvalue of A. Also, M is the total number of edges in a network. Finally, by replacing A with $\frac{A_{xy}}{E[A_{xy}]}$, Eq.(4.14) can be expressed as follows.

$$\begin{aligned} s^{LHN2}_{xy} &= \delta_{xy} + \frac{2M}{k_x k_y} \sum_{l=0}^{\infty} \phi^l \lambda^{1-l} (A^l)_{xy} \\ &= [1 - \frac{2M\lambda_1}{k_x k_y}]\delta_{xy} + \frac{2M\lambda_1}{k_x k_y}[(I - \frac{\phi}{\lambda_1}A)^{-1}]_{xy} \end{aligned} \tag{4.15}$$

Average Commute Time (ACT) (Göbel and Jagers, 1974) is calculated by defining that $m(x, y)$ is the average number of steps for a random walker travels from node x to y. In this case, the ACT between nodes x and y is,

$$n(x, y) = m(x, y) + m(y, x) \tag{4.16}$$

Note that the above equation also can be expressed with Moore-Penrose pseudoinverse (Klein and Randić, 1993; Fouss et al., 2007), L^+, such that,

$$n(x, y) = M(l^+_{xx} + l^-_{yy} - 2l^+_{xy}) \tag{4.17}$$

where l^+_{xy} refers to entry of L^+. If the smaller ACT means that the two nodes are similar, the similarity of the nodes x and y can be expressed as,

$$s^{ACT}_{xy} = \frac{1}{l^+_{xx} + l^-_{yy} - 2l^+_{xy}} \tag{4.18}$$

Cosine based on L^+ (Fouss et al., 2007) is based on dot product. The cosine similarity is defined as the cosine of node vectors.

$$s^{cos^+}_{xy} = cos(x, y)^+ = \frac{v^T_x v_y}{\mid v_x \mid \cdot \mid v_y \mid} = \frac{l^+_{xy}}{\sqrt{l^+_{xx} \cdot l^-_{yy}}} \tag{4.19}$$

where $l^+_{xy} = v^T_x v_y$ and $v_x = \Lambda^{1/2} U^T \vec{e}_x$. Note that U is an orthonormal matrix of eigenvectors of L^+ arranged in descending order of eigenvalue λ_x. Also, Λ and $\vec{e}$ denote diagonal matrix of λ_x and $N \times 1$ vector where one for x^{th} element and zero otherwise, respectively.

Random Walk with Restart (RWR) (Pan et al., 2004) is calculated as the probability that a random walker starting at node x will enter a steady state at node y rather than itself given that the probability of moving to random neighbor is q and the probability of returing to its position is $1 - q$.

$$\begin{aligned} \vec{q}_x &= qP^T\vec{q}_x + (1-q)\vec{e}_x \\ &= (1-q)(I - qP^T)^{-1}\vec{e}_x \end{aligned} \tag{4.20}$$

Note that P is a transition matrix where $Q_{xy} = 1/k_x$ if nodes x and y are connected and zero otherwise. Therefore, RWR can be defined as,

$$s_{xy}^{RWR} = q_{xy} + q_{yx} \tag{4.21}$$

where q_{xy} is y-th element of $\vec{q}_x$.

SimRank (Jeh and Widom, 2002) assumes that nodes x and y are similar if they are connected to similar nodes, which is similar to $LHN2$.

$$s_{xy}^{SimRank} = C \cdot \frac{\sum\limits_{z\in\Gamma(x)} \sum\limits_{z'\in\Gamma(y)} s_{zz'}^{SimRank}}{k_x \cdot k_y} \tag{4.22}$$

where $s_{xx} = 1$ and $C \in [0,1]$ is decay factor. This is calculated by how fast two random walkers are expected to meet at a particular node when they start at diffrent node x and y.

Local Random Walk (LRW) (Liu and Lü, 2010) considers a random walker with an initial density vector, $\vec{\pi}_x(0) = \vec{e}_x$, at the starting node x. The initial vector satisfies $\vec{\pi}_x(t+1) = P^T \vec{\pi}_x(t)$ for $t \geq 0$. That is, at time step t, LRW is as follows

$$s_{xy}^{LRW} = q_x \pi_{xy}(t) + q_y \pi_{yx}(t) \tag{4.23}$$

where q is initial configuration function. This index focuses on a few-step random walk instead of being stationary state.

Superposed Random Walk (SRW) (Liu and Lü, 2010) is designed to overcome the shortcomings of the random walk-based methods. The shortcoming is the dependency on parts of the network that are too far from the target node. That is, for example, when going from node x to y, there is a possibility that the random walker falls too far even though the two nodes are adjacent. Since real-world networks often have high clustering coefficients, the random walker circulates locally rather than going further to other parts of the network. In order to overcome this shortcoming, random walker is continuously released at the starting node to increase the similarity between the target node and the neighbor nodes such that,

$$s_{xy}^{SRW} = \sum_{\gamma=1}^{t} s_{xy}^{LRW}(\gamma) = \sum_{\gamma=1}^{t} [q_x \pi_{xy}(\gamma) + q_y \pi_{yx}(\gamma)] \tag{4.24}$$

Matrix Forest Index (MFI) (Chebotarev and Shamis, 2006) is defined as the ratio of all spanning rooted forests in the network and the number of spanning rooted forests that the two nodes belong together in the tree extending from the starting node.

$$s^{MFI} = (I + L)^{-1} \tag{4.25}$$

Zhou and Jia (2017) proposed H-index and in this study, it is defined as the knowledge quantity of each node. Therefore, H-index of a node is determined to be the maximum value h such that there exists at least h neighbors of degree no less then h. Thus, H-index of node i is calculated as follows.

$$h_i = H(k_{j1}, k_{j2}, ..., k_{jk_i}) \tag{4.26}$$

where $k_{j1}, k_{j2}, ..., k_{jk_i}$ refer to degrees of neighbor nodes of node i. Note that detailed step-by-step description is introduced in Appendix A.2 with examples. From this, we calculate the ratio of the amount of information from node x to any other node y to calculate the score of the adjacency matrix A_t at time t. In other words, using the H-index to find the adjacency matrix consisting of the edge from node x to y is as follows.

$$A^H_{xy} = \frac{h_y + 1}{\sum_{N_x}(h_n + 1)} \tag{4.27}$$

where N_x is the neighbor node pointed from node x. Note that one is added to the denominator for convenience of calculation since it becomes zero if the H-index is zero. Computing all node pairs with Eq.(4.27) yields the adjacency matrix A^H_t at time t. Then, as in Eq.(4.12), s^H_t is obtained as follows.

$$s^H_{xy} = \beta A^H_{xy} + \beta^2(A^H_{xy})^2 + \beta^3(A^H_{xy})^3 + \cdots + \beta^n(A^H_{xy})^n \tag{4.28}$$

4.2.2 Weighted Measures

Weighted Common Neighbor (WCN) (Murata and Moriyasu, 2007) is similar to CN, but it considers average edge weight of the arbitrary nodes x and y with common neighbor node z. Since we examine a directed graph, the neighbor node receives the arrows common to nodes x and y as follows.

$$s_{xy}^{WCN} = \sum_{z \in \Gamma(x) \cap \Gamma(y)} w(x,z) + w(y,z) \tag{4.29}$$

Weighted Adamic Adar (WAA) (Murata and Moriyasu, 2007) is similar to WCN, but calculated by dividing the average edge weight by the logarithmic of the strength of the node.

$$s^{WAA} = \sum_{z \in \Gamma(x) \cap \Gamma(y)} \frac{w(x,z) + w(y,z)}{\log(1 + s(z))} \tag{4.30}$$

where $s(x) = \sum_{z \in \Gamma(x)} w(x,z)$ which indicates the strength of node x. Note that $\log(1 + s(z))$ is used to prevent negative condition.

Weighted Resource Allocation (WRA) (Murata and Moriyasu, 2007) is also similar to WCN but uses the strength of node in the denominator.

$$s^{WRA} = \sum_{z \in \Gamma(x) \cap \Gamma(y)} \frac{w(x,z) + w(y,z)}{s(z)} \tag{4.31}$$

Generalized Degree (GD) deals with the node degree, most widely used method to calculate node centrality in network analysis. Unlike degrees in an unweighted network, the concept of node strength is proposed (Barrat et al., 2007; Newman, 2004; Opsahl et al., 2008) for a weighted network and used to measure the power of a node (Opsahl et al., 2010). As mentioned above, node strength can be expressed as the sum of weights, η^2 in this dissertation, such that,

$$k_i = C_D(i) = \sum_j^N x_{ij} \tag{4.32a}$$

$$s_i = C_D^W(i) = \sum_j^N \eta_{ij}^2 \tag{4.32b}$$

$$C_D^{W\alpha}(i) = k_i \times (\frac{s_i}{k_i})^\alpha = k_i^{1-\alpha} \times s_i^\alpha \tag{4.32c}$$

where $N, \alpha, k_i, s_i, x_{ij}$, and η_{ij}^2 are the total number of nodes, the tuning parameter, the node degree, the node strength, the unweighted adjacency matrix, and the weighted adjacency matrix, respectively. The adjacency matrix can be calculated using Eq.(4.32c) as follows.

$$A_{xy}^{GD} = \frac{C_y^{W\alpha} + 1}{\sum_{N_x}(C_n^{W\alpha} + 1)} \tag{4.33}$$

where the score matrix at time t, s_t^{GD}, also can be computed as follows.

$$s_{xy}^{GD} = \beta A_{xy}^{GD} + \beta^2(A_{xy}^{GD})^2 + \beta^3(A_{xy}^{GD})^3 + \cdots + \beta^n(A_{xy}^{GD})^n \tag{4.34}$$

4.3 Proposed measures

In this dissertation, two simple indices are proposed for weighted digraph, called Sum of eta squared (SE) and weighted causality (WC). SE uses only η^2 as weight while other indices use a mixture of degree and η^2. Therefore, the power of each node can be computed as the sum of eta squared in outflow directions such that,

$$P = \begin{pmatrix} \sum_{N_1}^{j} \eta_{1j}^2 \\ \sum_{N_2}^{j} \eta_{2j}^2 \\ \cdots \\ \sum_{N_n}^{j} \eta_{nj}^2 \end{pmatrix} \tag{4.35}$$

where P, η^2, $N_i(i = 1, 2, \dots, n)$, and j refer to power of each node, the value of F-statistics, the neighbor of each node, and nodes belonging to N_i, respectively.

If each element of P_t is called $P_{t,1}, P_{t,2}, \dots, P_{t,n}$ at time t, the adjacency matrix for nodes x and y can be calculated as in Eq.(4.27) such that,

$$A_{xy}^{SE} = \frac{P_y + 1}{\sum_{N_x}(P_n + 1)} \tag{4.36}$$

where N_x and n are the outflow neighbors of node x and nodes belonging to N_x, respectively. Therefore, the score matrix at time t, s^{SE} is,

$$s_{xy}^{SE} = \beta A_{xy}^{SE} + \beta^2(A_{xy}^{SE})^2 + \beta^3(A_{xy}^{SE})^3 + \cdots + \beta^n(A_{xy}^{SE})^n \tag{4.37}$$

Note that an example of SE is described in Appendix A.4.

To overcome the drawback of the other methods mentioned above, which consider only the ratio of the row of the adjacency matrix (i.e. outflow from one node), Weighted Causality (WC) is proposed. Therefore, in order to represent the power of each edge with respect to the entire edge of the network as a ratio in WC, the new adjacency matrix $A_{i,j}$ is computed by dividing by the largest F-statistics for the entire edge such that,

$$A_{xy}^{WC} = \frac{\eta_{xy}^2}{max(\eta_{xy}^2)} \tag{4.38}$$

For each node x and y, we calculate the similarity score between nodes using the weighted adjacency matrix. Then, the summation of paths from node x to y with length n path is defined as follows.

$$s_{xy}^{WC} = \beta A_{xy}^{WC} + \beta^2 (A_{xy}^{WC})^2 + \beta^3 (A_{xy}^{WC})^3 + \cdots + \beta^n (A_{xy}^{WC})^n \tag{4.39}$$

where β is a free parameter set to be 0.1, which should be less than the largest eigenvalue of adjacency matrix. The smaller β means more weight for short path. Finally, s^{WC} is obtained by increasing n until Eq.(4.39) converges.

4.4 Results

4.4.1 Evaluation of Link Prediction

There are two standard metrics for measuring the accuracy of the prediction algorithm: area under the receiver operating characteristic curve (AUC) and precision (Geisser, 1993; Herlocker et al., 2004). Let the true positive (TP), false negative (FN), false positive (FN), and true negative (TN) refer to the samples correctly identified as positive, incorrectly identified as negative, incorrectly identified as positive, and correctly identified as negative, respectively. The receiver operating characteristic (ROC) curve is a threshold curve plotted the ratio of true positive rate (TPR) and false positive rate (FPR), which represents trade-offs of performance.

In addition, the TPR measures the percentage of actual positives that are correctly identified while the FPR is the percentage of negatives that are incorrectly identified among the total actual negatives. They can be expressed as $TP/(TP + FN)$ and $(TN + FP)$, respectively. Therefore, the AUC of link prediction implies the probability that a correctly predicted selected link has a higher similarity score than that of a randomly selected link that is not connected (Meghanathan, 2016).

If U is a set of all possible links in the network at time $t + \tau$, $G_{t+\tau}$, it can be classified into two categories: existent link ($E^e_{t+\tau}$) and non-existent link ($E^n_{t+\tau}$). The existent links represent that the connected links whose Granger causality direction exists and the non-existent links are vice versa. At this time, the missing links, a subset of the existent links, refer to link randomly sampled from $E^e_{t+\tau}$ at a certain ratio α set to be 0.5 in this dissertation and can be expressed as $E^m_{t+\tau}$. In summary, $U = E^e_{t+\tau} \cup E^n_{t+\tau}$ and $E^m_{t+\tau} \subset E^e_{t+\tau}$. Then, we compare the similarity scores for every pair of these two sets of edges and count n' if the similarity score of the missing link is higher than that of the non-existent link, whereas we count n'' if the similarity scores are equal. In other words, when n independent comparisons are performed on the ranked list of all observed links, we count n' if the true positive(missing link) one has a higher similarity score than that of false positive one(non-existent link) and count n'' if both scores are equal. As suggested by Lü and Zhou (2011), the AUC can be calculated as follows.

$$AUC = \frac{n' + 0.5n''}{n} \tag{4.40}$$

Thus, AUC becomes one if the scores of all missing links are greater than those of

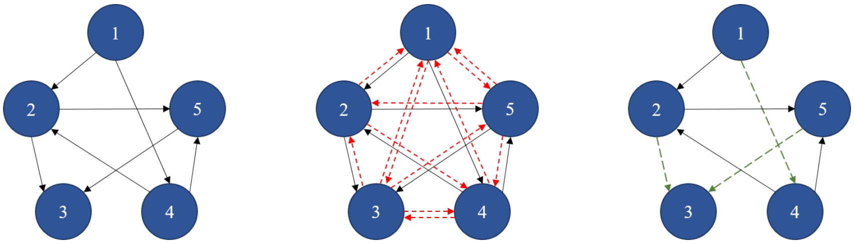


Figure 4.1: Granger causality network at time $t+\tau$

non-existent link ($n' = n, n'' = 0$). This can be illustrated as an example.

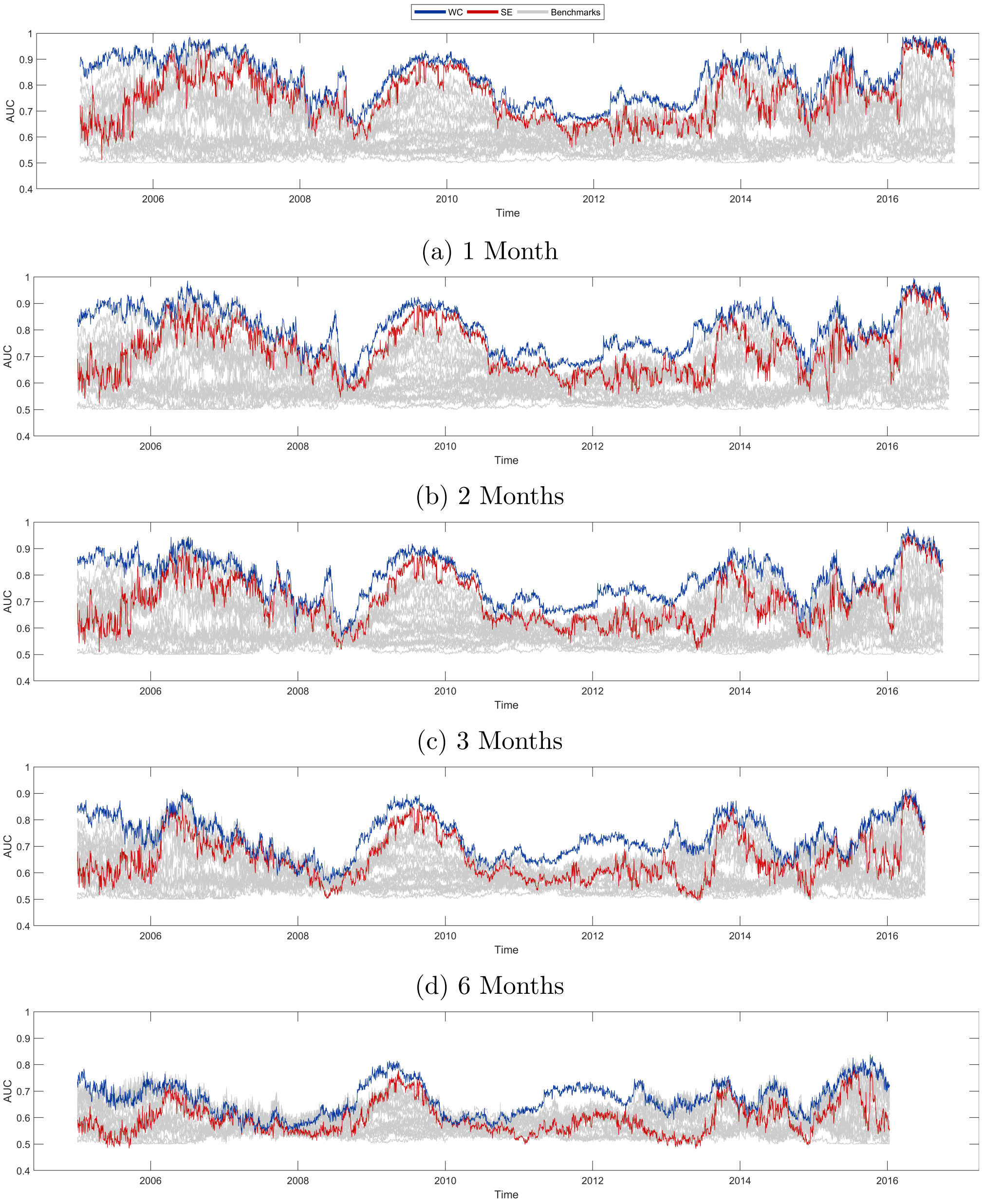
The actual Granger causality network at time $t+\tau$ is represented in Figure 4.1. The figure on the left refers to the basic network structure, and the two figures on the right show 13 unconnected(non-existent) links in red and 3 randomly sampled(missing) links in green. Then, we compare the similarity scores at time t, s_t^{WC}, for 39 pairs of red and green links. The AUC evaluates the performance of the algorithm for the whole links, whereas the precision is only interested in the L links with the highest scores. That is, the precision equals L_r/L if L_r links among the L top-ranked links are correctly predicted. Therefore, the AUC can evaluate the performance of the prediction algorithm better such as a link prediction problem in which the number of existent links is significantly less than that of non-existent links (Menon and Elkan, 2011). Hence, in this study, we decide to utilize the AUC as the performance measure.

4.4.2 Result of Link Prediction

As a result of analyzing the time-varying properties and topologies of the REER networks, we find that the structural change of the Granger causality network of the global currency market has implications for the evolution of the financial market. In particular, we observe that newly connected or disconnected links between different nodes from the changes in the global economy and the financial market. In this context, predicting possible links of the Granger causality network in the future is useful for understanding the changes in the financial market. Therefore, we perform the link prediction with the benchmarks in 4.2 and the proposed methods, SE and WC. The results are summarized in Figure 4.2, 4.3, and 4.4.

Figure 4.2 shows the AUCs of the proposed methods, SE and WC, and other benchmarks in red, blue, and gray solid lines, respectively, for the prediction of the links of the REER network after 1, 2, 3, 6 and 12 months. We find that the AUC decreases when prediction lag is increased for all of the 27 methods, and the trends of prediction according to the experimental period are similar. However, we also find that the AUC drops significantly during the financial crises mentioned above. Except for the extreme structural break in the network, the results of link predictions have consistency for the entire experimental period. In addition, WC outperforms SE and the other benchmarks in most cases for five prediction lags. Figure 4.3 shows the boxplots of AUCs, the prediction performance for every method. The green dash line refers to the mean AUC of the WC, which is superior to those of SE and benchmarks. On the other hand, the AUC of SE is greater than those of most benchmarks but equal or even lower in some cases. It means that it is more appropriate to represent the power of each edge by the proportion within the whole network rather than that in the only within the row of an adjacency matrix.

Therefore, more details about the result of WC that has the highest AUC is investigated. Figure 4.4 represents the average of AUCs for different cross-sections



(a) 1 Month

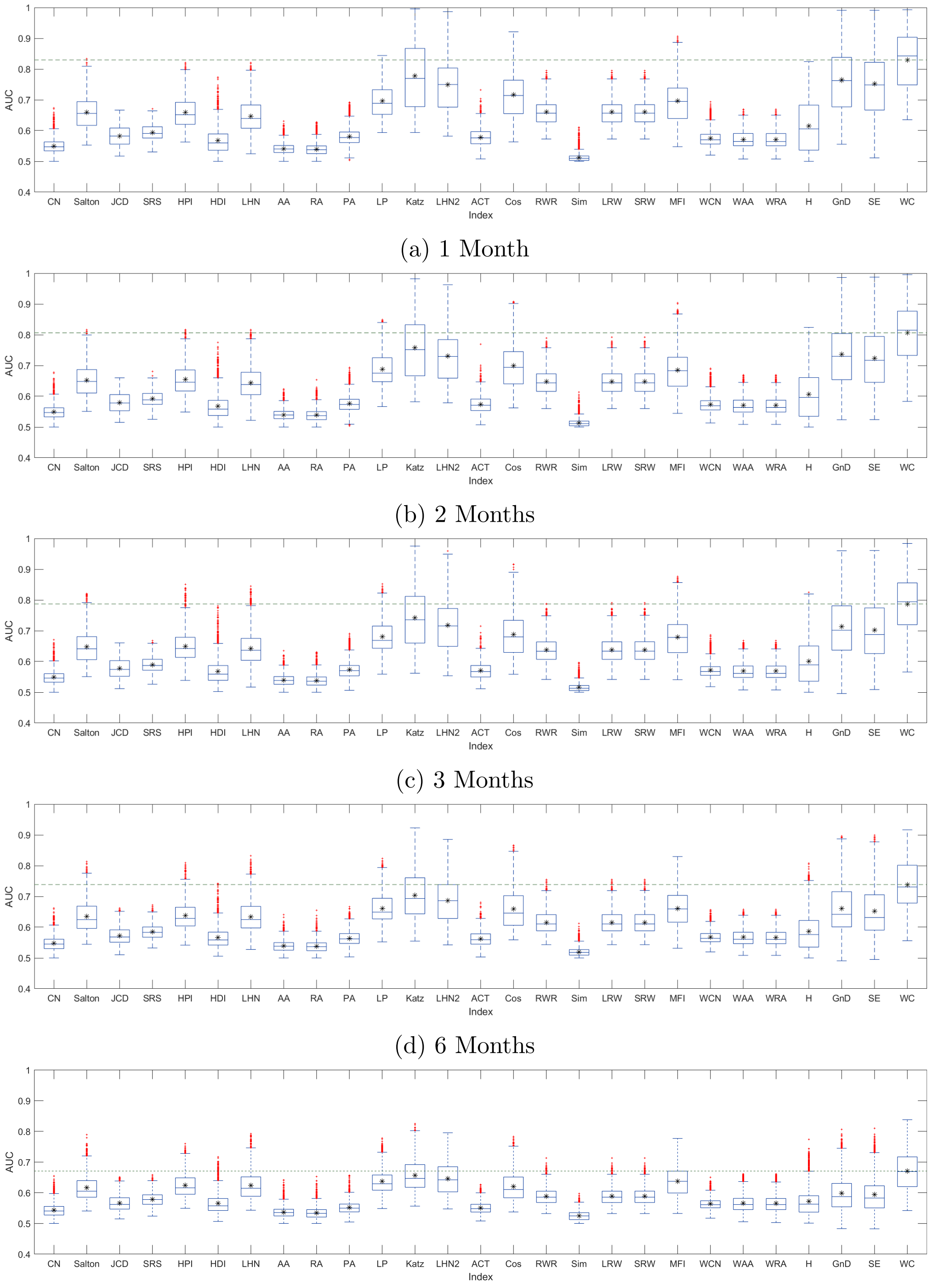
(b) 2 Months

(c) 3 Months

(d) 6 Months

(e) 12 Months

Figure 4.2: Time-varying AUCs of the proposed method and benchmarks in different prediciton lags



(a) 1 Month

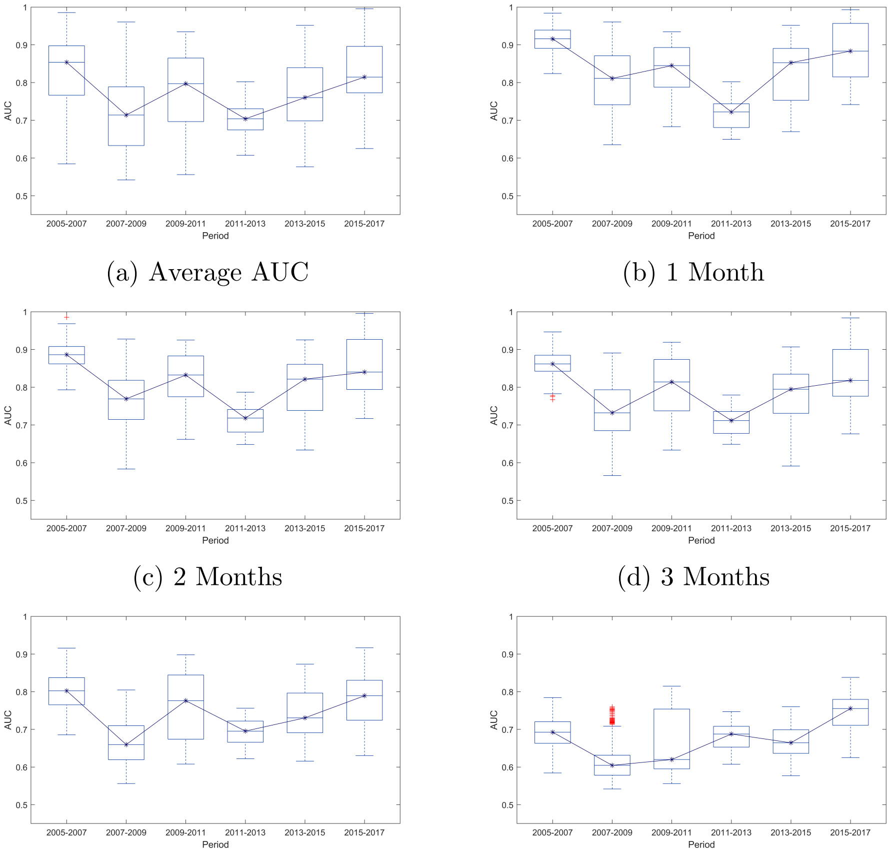
(b) 2 Months

(c) 3 Months

(d) 6 Months

(e) 12 Months

Figure 4.3: Boxplots for AUCS of the proposed method and benchmarks in different prediciton lags



(a) Average AUC

(b) 1 Month

(c) 2 Months

(d) 3 Months

(e) 6 Months

(f) 12 Months

Figure 4.4: Boxplots for AUCs of the proposed method in different prediction lags

during the experimental periods and shows that the prediction accuracy is relatively low in 2007-2009 and 2011-2013 when the financial crisis occurred. The black solid line implies the trend of the median of AUCs and has W shape. Figure 4.4b, 4.4c, 4.4d, 4.4e, and 4.4f shows the prediction accuracies for different lags. As shown in Figure 4.2 and 4.3, the overall accuracy decreases as the lag increases, and the accuracy of 2007-2009 and 2011-2013 are lower than those of other cross-sections.

Lastly, a statistical test is performed to verify that there is a significant difference between the proposed method, especially WC, and benchmarks. First, the Kolmogorov-Smirnov test, a non-parametric test, is conducted to decide if the AUCs follow a Gaussian distribution. It can be used to compare a sample with a reference probability distribution, in this case, the Gaussian distribution. The results show that the p-values converge to zero for all methods and prediction lags, which means that the distributions of AUCs are non-Gaussian. Therefore, the Wilcoxon signed-rank test, a non-parametric test of the paired t-test, is chosen to verify that there is a statistical difference between the AUC of the WC and that of other methods. Note that the null hypothesis is that the medians of the two samples are equal. The results show that the p-values converge to zero for all methods and prediction lags and provide the evidence that the prediction performance of the WC is higher than that of other methods. And it proves the superiority of the WC in the link prediction of the REER network.

4.5 Summary and Discussion

In this chapter, we propose link prediction methods for a Granger causality network using REER data. The prediction accuracy is measured for a total of 27 methods, including benchmarks commonly used and proposed methods.

At first, the prediction accuracy is evaluated with AUC method. The set of all links is divided into existent links and non-existent links, and the sampled links from the existent links at a rate of 0.5 are named missing links. When similarity scores calculated by each method are compared for every pair of the missing links and non-existent links, the higher the score of the missing link, the higher the prediction accuracy. The results show that more decreased AUCs for all 27 methods when prediction lag increased given that the trends of the prediction according to the experiment period. Especially, the AUC drops significantly during the financial crises but is consistent for the entire period. Furthermore, WC outperforms SE and the other benchmarks, which means that the ratio within the whole network is more suitable to reflect the power of each edge than the ratio only within a node. Note that the performance of SE is better than most benchmarks, but is similar or even worse than some benchmarks.

For identifying the statistical difference between the proposed method WC whose AUC is the highest and other methods, statistical test is performed. At first, the result of the Kolmogorov-Smirnov test shows the AUCs have non-Gaussian distributions. Since the distributions of AUCs does not satisfy the assumption of normality, the Wilcoxon signed-rank test is performed. Similarly, the results suggest that the prediction accuracy of WC shows a statistical difference with other prediction methods since the p-values converge to zero.

In conclusion, by reminding that the purpose of this chapter is to predict future possible links in Granger causality network, the results provide good implications that eta squared based on F-statistics is useful for understanding the future structure of the Granger causality network. Then, knowing which links to appear or disappear from the network in advance helps investors make decisions. In this context, methods for constructing portfolios based on the result of link prediction will be discussed in Chapter 5.

Chapter 5

Application of Link Prediction

5.1 Overview

As mentioned in the end of Chapter 4, this chapter aims to utilize the proposed WC-based result of link prediction to construct optimal portfolios. The implications of the Granger causality network and the link prediction suggest that knowledge about the network structure in advance can be applied to manage financial risk. However, private investors have difficulty in investing in REER directly. Therefore, this chapter focuses on constructing investment portfolios in the U.S. stock market, which is the largest in the world and is comfortable for investors to invest directly. Specifically, the performance of the portfolio applied WC-based link prediction is compared to several benchmarks.

In this chapter, we utilize the total return index of institutions belonging to the S&P 500 index, which consists of about 500 stocks and divide into 11 sectors by the Global Industry Classification Standard (GICS). As of December 31, 2018, a total of 110 stocks are selected, top 10 stocks in each sector by market capitalization. The proposed method uses a Weighted Causality Planar Graph (WCPG) devised from

PMFG. It represents the intensity of interactions between nodes as the effect size of Granger causality direction, not as correlation or distance. Then, the peripherality of the stocks in WCPG is measured by the method introduced in Pozzi et al. (2013) and the portfolio consists of stocks with high peripherality.

Once the portfolio is constructed, the performance is evaluated by the Sharpe ratio and Signal-to-noise ratio (Information ratio) for comparing to several benchmarks. In this dissertation, a total of four benchmarks are introduced: one naive model, two classical models, including a market index, and a network approach based on PMFG.

5.2 Benchmark models

5.2.1 Classical models

In this chapter, one naive model and two classical models are chosen as benchmarks. The first is the $1/N$ strategy, which is an equally weighted strategy for all assets. There are many mathematical and theoretical models for portfolio optimization, but many actual investors tend to rely on rough and instinctive common-sense. They assume that the mean return or covariance of an asset is equal to that of other assets since there is a lack of information about the future risks and returns of the individual asset. In this regard, Samuelson (1967) show that it is optimal to buy each asset uniformly if the return is identically distributed. Also, Duchin and Levy (2009) suggest that investors tend to choose the $1/N$ strategy despite having a theoretical approach like the classical mean-variance diversification theory of Markowitz and show that it is optimal to buy each asset uniformly. This is a passive investment strategy usually used as a benchmark for evaluating portfolio strategies (Shilling, 1992; Najafi and Pourahmadi, 2016). The second is a capitalization-weighted strategy that assigns weights according to the total market value of firms' outstanding shares. Since the value of stocks changes every day and the impact is proportional to the company's overall market value, companies with large market capitalization, which contribute more reliably to the growth of the index, are assigned larger weights. Conversely, since companies with low market capitalization have relatively large volatility, it has the effect of reducing risk by assigning low weights to them. The third is the S&P 500 market index. It is also a capitalization-weighted index, which measures the stock performance of 500 large companies listed on the stock exchange in the United States. In other words, it is equivalent to investing in all assets that make up the index and is the simplest way to invest without portfolio selection.

5.2.2 Planar Maximally Filtered Graph(PMFG)

Since the correlation between two assets can be measured at any time regardless of the strength of connection in a network, all assets are connected to each other which is called a *complete graph.* However, it is more common to analyze proper subgraphs with the same number of nodes and fewer but more relevant edges instead of complete graphs in the financial market, such as minimum spanning tree (MST) and planar maximally filtered graph(PMFG) (Vỳrost et al., 2019).

Spanning tree is an acyclic graph that satisfies the properties of a tree, including all nodes in the graph and all nodes connected to each other. Mantegna (1999) introduce the minimum spanning tree, which has the smallest sum of edge weights among the set of possible spanning trees. However, since MST has only $N-1$ edges, whereas the maximum number of edges that a complete graph on N nodes can have is $N(N-1)/2$, information loss may occur. Therefore, Tumminello et al. (2005) propose a planar maximally filtered graph(PMFG) to filter out complex networks while retaining more edges. PMFG has $3(N-2)$ edges and MST is a subgraph of a PMFG. The correlation between the two assets can be expressed as follows.

$$\rho_{i,j} = \frac{E[r_i r_j] - E[r_i]E[r_j]}{\sqrt{(E[r_i^2] - E[r_i]^2)(E[r_j^2] - E[r_j]^2)}} \tag{5.1}$$

where E refers to the mathmatical expectation of the sequence over time t and r_i, r_j refers to the rate of return of node i and j. Mantegna (1999) also proposed a nonlinear decreasing transformation, which means that the larger the correlation, the shorter the distance. Then, $\rho_{i,j}$ can be transformed as the distance between nodes i and j as follows.

$$d_{i,j} = \sqrt{2(1-\rho_{i,j})} \tag{5.2}$$

The PMFG algorithm starts by sorting the edges in ascending order in weights and adds edges one by to the graph until the number of edges is 3(N-2). The edges that violate the planarity are discarded. Then, the peripherality of a stock is defined as a combination of several centrality measures to distinguish between central and peripheral regions in a filtered network. The portfolio consisted of the peripheral stocks from the PMFG is proposed by Pozzi et al. (2008, 2013). Peripherality measure P is obtained from the following five measures.

- **Degree Centrality(DC)**: the number of edges connected to a node.
- **Betweenness Centrality(BC)**: the number of shortest paths that pass through a node.
- **Eccentricity(E)**: the maximum distance of the shortest paths from node x to any other node y.
- **Closeness(C)**: the average distance of all shortest paths from node x to any other node y.
- **Eigenvector Centrality(EC)**: i-th component of the eigenvector corresponding to the largest eigenvalue of the adjacency matrix.

The value of DC, BC, and EC refers to the centrality measures, while that of E and C refers to peripherality measures. The peripheral nodes have small DC, BC, and EC and large E and C. They are sorted in ascending and descending order, respectively, and the tied rank value is obtained. Then, P can be calculated as follows.

$$\begin{aligned} P = &\frac{DC^w + DC^u + BC^w + BC^u}{4(N-1)} \\ &+ \frac{E^w + E^u + C^w + C^u + EC^w + EC^u}{6(N-1)} \end{aligned} \tag{5.3}$$

where the subscripts w and u refer to weighted and unweighted PMFG, respectively. The edge weight is $1 + \rho_{i,j}$ for DC and EC, and $d_{i,j}$ for others. A node with a large value of P belongs to the peripheral region of the network and vice versa. So, we select the top 10, 20, and 30 stocks with the largest P values and construct a portfolio by Markowitz weights.

5.3 Weighted Causality Planar Graph(WCPG)

5.3.1 Realization of WCPG

Since PMFG has undirected edges based on correlation, it is not appropriate for representing cause-and-effect relationships. Kenett et al. (2010) propose a Partial Correlation Planar Graph(PCPG) to deal with asymmetric interactions among the components of a system. It is a variation of the PMFG and is obtained by starting from correlation influence $d(X, Y : Z)$. At this time, the average influence $d(X : Z)$ of an element Z that affects the correlation between element X and all the other elements are defined as follows.

$$d(X : Z) =< d(X, Y : Z) >_{Y \neq X,Z} \tag{5.4}$$

Note that $d(X : Z) \neq d(Z : X)$ in general since PCPG is a directed network. And if $d(X : Z) > d(Z : X)$, which means that the influence of Z on X and that of X on Z, only the link $Z \rightarrow X$ is carded in the PCPG. In the same way, a graph is newly constructed using the eta-squared obtained in Eq.(3.2.2), which is named a Weighted Causality Planar Graph (WCPG). Note that the edge weights are assigned as similarity scores between nodes obtained in Eq.(4.39). Since the WCPG is a directed network, the P-measure in Eq.(5.3) is calculated based on the Pagerank,

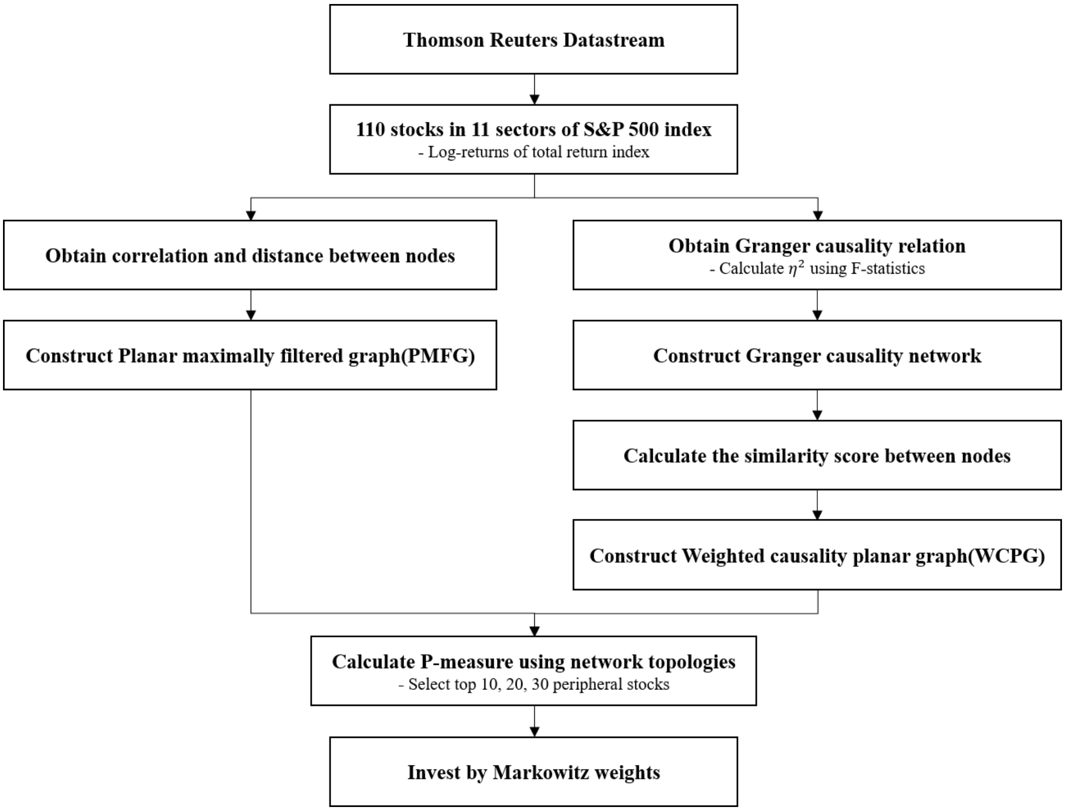

Thomson Reuters Datastream
110 stocks in 11 sectors of S&P 500 index
- Log-returns of total return index
Obtain correlation and distance between nodes
Obtain Granger causality relation
- Calculate η^2 using F-statistics
Construct Planar maximally filtered graph(PMFG)
Construct Granger causality network
Calculate the similarity score between nodes
Construct Weighted causality planar graph(WCPG)
Calculate P-measure using network topologies
- Select top 10, 20, 30 peripheral stocks
Invest by Markowitz weights

Figure 5.1: Procedure of link prediction and portfolio selection

which is a variant of eigenvector centrality. It is based on normalized eigenvector centrality and is combined with a random jump assumption (Zaki et al., 2014). Then, the number of 10, 20, and 30 stocks having large P-measure is involved in the portfolio and is assigned Markowitz weights. Figure 5.1 summarizes the graph-based portfolio selection algorithms. In summary, the bidirectional relationships between different assets are expressed as Granger causality directions and the effect sizes. They are utilized to construct a proposed network, WCPG. And we observe the impact of considering asymmetric interactions on the portfolio. Table 5.1 shows the portfolio strategies used in this dissertation and the abbreviations are used in the rest

Table 5.1: List of benchmarks and proposed model

#	Model	Abbreviation
Naive		
0.	$1/N$ strategy	ew or $1/N$
Classical approach		
1.	Market capitalization weighted strategy	cw or cap-weighted
2.	S&P 500 index	sp or market index
Network approach		
3.	Planar Maximally Filtered Graph	pmfg or PMFG
4.	Weighted Causality Planar Graph	wcpg or WCPG

of this chapter. The five models are categorized according to the method that select or weight to stocks. Investing in S&P 500 index refers to invest for all stocks, a $1/N$ strategy, and a cap-weighted strategy are methods of assigning weights uniformly and by market capitalization, respectively. Besides, two strategies are used to select stocks based on the network topology, which are the methods of selecting from the PMFG based on correlation, and from the WCPG based on the causal relationship between different assets, respectively. In this case, the weights of investment are the Markowitz weights.

5.4 Data description

In this dissertation, the classification of the industry sector is based on GICS (Global Industry Classification Standard) developed by MSCI (Morgan Stanley Capital International). It divides the S&P 500 index into 11 sectors and the stocks belonging to each sector as of December 31, 2018. A total of 110 stocks are selected, which means that the top 10 by market capitalization in each sector based on the companies that continuously exist within the experimental period. The dataset is a daily Total Return Index (TRI) of 4,027 days from January 2, 2003, to December 31, 2018 and is extracted from Thomson Reuters Datastream. Table 5.2 summarizes the sectors and the stocks. Since TRI includes the dividend, interest, rights offerings and other distributions realized over a given period of time unlike the price index only considers price movements, it has the advantage of more accurately representing the performance of an individual stock.

5.5 Results

5.5.1 Evaluation measures

In this dissertation, the performance of the models is evaluated for 25 holding periods, τ without setting a rebalancing point. In other words, the models that show consistent results regardless of the period of dataset or the time point of the investment are considered as useful strategies. The Sharpe ratio is the most widely used measure of calculating the risk-adjusted return, which is a ratio as an indicator of how much the profit investors can get for a given risk (Sharpe, 1966). It can be calculated by the return and standard deviation of a portfolio strategy during the entire experiment period regardless of whether considering the rebalancing (Dichtl et al., 2016; Bessler et al., 2017; Dai and Wang, 2019; Yu et al., 2020). It is determined such that,

$$Sharpe\ ratio = \frac{E(R_p) - R_f}{\sigma_p} \tag{5.5}$$

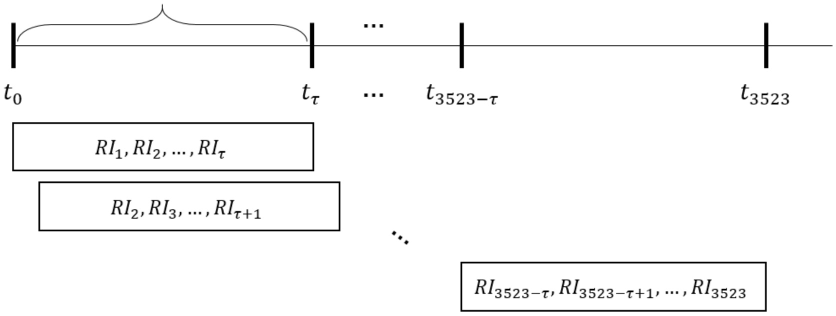
where $E(R)$, R_f and σ_p refer to the expected return of the portfolio, the risk-free

Table 5.2: List of top 10 stocks by market capitalization in each sector

Sector	Company	Code	Sector	Company	Code	Sector	Company	Code
Communication Services	AT&T	T	Financials	JP MORGAN CHASE & CO.	JPM	Materials	LINDE	LIN
	VERIZON COMMUNICATIONS	VZ		BANK OF AMERICA	BAC		ECOLAB	ECL
	WALT DISNEY	DIS		WELLS FARGO & CO	WFC		SHERWIN-WILLIAMS	SHW
	COMCAST A	CMCSA		CITIGROUP	C		AIR PRDS.& CHEMS.	APD
	NETFLIX	NFLX		AMERICAN EXPRESS	AXP		NEWMONT GOLDCORP	NEM
	ACTIVISION BLIZZARD	ATVI		US BANCORP	USB		PPG INDUSTRIES	PPG
	ELECTRONIC ARTS	EA		CME GROUP	CME		BALL	BLL
	OMNICOM GROUP	OMC		MORGAN STANLEY	MS		FREEPORT-MCMORAN	FCX
	CBS 'B'	CBS		BLACKROCK	BLK		INTERTIOL PAPER	IP
	CENTURYLINK	CTL		BERKSHIRE HATHAWAY 'B'	BRK.B		NUCOR	NUE
Consumer Discretionary	AMAZON.COM	AMZN	Health Care	JOHNSON & JOHNSON	JNJ	Real Estate	AMERICAN TOWER	AMT
	HOME DEPOT	HD		MERCK & COMPANY	MRK		CROWN CASTLE INTL.	CCI
	MCDONALDS	MCD		UNITEDHEALTH GROUP	UNH		PROLOGIS	PLD
	NIKE 'B'	NKE		PFIZER	PFE		EQUINIX REIT	EQIX
	STARBUCKS	SBUX		ABBOTT LABORATORIES	ABT		SIMON PROPERTY GROUP	SPG
	BOOKING HOLDINGS	BKNG		MEDTRONIC	MDT		PUBLIC STORAGE	PSA
	LOWE'S COMPANIES	LOW		AMGEN	AMGN		WELLTOWER	WELL
	TJX	TJX		THERMO FISHER SCIENTIFIC	TMO		EQUITY RESD.TST.PROPS. SHBI	EQR
	TARGET	TGT		ELI LILLY	LLY		AVALONBAY COMMNS.	AVB
	MARRIOTT INTL.'A'	MAR		BRISTOL MYERS SQUIBB	BMY		VENTAS	VTR
Consumer Staples	WALMART	WMT	Industrials	BOEING	BA	Utilities	NEXTERA ENERGY	NEE
	PROCTER & GAMBLE	PG		HONEYWELL INTL.	HON		DUKE ENERGY	DUK
	COCA COLA	KO		UNITED TECHNOLOGIES	UTX		DOMINION ENERGY	D
	PEPSICO	PEP		UNION PACIFIC	UNP		SOUTHERN	SO
	COSTCO WHOLESALE	COST		LOCKHEED MARTIN	LMT		EXELON	EXC
	MONDELEZ INTERNATIONAL CL.A	MDLZ		3M	MMM		AMER.ELEC.PWR.	AEP
	ALTRIA GROUP	MO		UNITED PARCEL SER.'B'	UPS		SEMPRA EN.	SRE
	COLGATE-PALM.	CL		GENERAL ELECTRIC	GE		XCEL ENERGY	XEL
	WALGREENS BOOTS ALLIANCE	WBA		CATERPILLAR	CAT		CONSOLIDATED EDISON	ED
	KIMBERLY-CLARK	KMB		CSX	CSX		PUB.SER.ENTER.GP.	PEG
Energy	EXXON MOBIL	XOM	Information Technology	MICROSOFT	MSFT			
	CHEVRON	CVX		APPLE	AAPL			
	CONOCOPHILLIPS	COP		INTEL	INTC			
	SCHLUMBERGER	SLB		CISCO SYSTEMS	CSCO			
	EOG RES.	EOG		ORACLE	ORCL			
	OCCIDENTAL PTL.	OXY		INTERNATIONAL BUS.MCHS.	IBM			
	VALERO ENERGY	VLO		ADOBE (NAS)	ADBE			
	WILLIAMS	WMB		ACCENTURE CLASS A	ACN			
	PIONEER NTRL.RES.	PXD		TEXAS INSTRUMENTS	TXN			
	HALLIBURTON	HAL		NVIDIA	NVDA			

rate and the volatility of the portfolio, respectively. Note that the risk-free rate is the U.S. 10 year treasury yield and the Sharpe ratio is annualized for convenience of comparison.

It is known that a good investment strategy has low volatility, high return, and consistency. Holding a portfolio constructed at time t for the holding period τ, the return is $r_{t,\tau}$, the average return for all time point t is $\bar{r}(\tau)$. Pozzi et al. (2013) introduce the signal-to-noise ratio (or information ratio) whose the ratio between $\bar{r}(\tau)$ and the standard deviation $s(\tau)$. Figure 5.2 shows a scenario of calculating the information ratio.


...
t_0
t_τ
...
$t_{3523-\tau}$
t_{3523}
$RI_1, RI_2, \dots, RI_\tau$
$RI_2, RI_3, \dots, RI_{\tau+1}$
$\ddots$
$RI_{3523-\tau}, RI_{3523-\tau+1}, \dots, RI_{3523}$

τ	**Example of mean return**	IR_τ
10	$r_{1,10} = \frac{RI_1 + RI_2 + \cdots + RI_{10}}{10}$ $\vdots$ $r_{3513,10} = \frac{RI_{3513} + RI_{3514} + \cdots + RI_{3523}}{10}$	$\frac{mean(r_1, r_2, \dots, r_{3513})}{std(r_1, r_2, \dots, r_{3513})}$

Figure 5.2: A scenario of the signal-to-noise ratio

Table 5.3: Annualized Sharpe ratio for each holding period

τ	**ew**	**cw**	**sp**	**pmfg**$_{10}$	**pmfg**$_{20}$	**pmfg**$_{30}$	**wcpg**$_{10}$	**wcpg**$_{20}$	**wcpg**$_{30}$
10	0.7052	0.4953	0.4385	0.4715	0.7320	0.8242	0.7582	0.7234	0.6869
20	0.6750	0.4685	0.3910	0.1993	0.5028	0.5804	0.7628	0.7936	0.6631
30	0.6894	0.5018	0.4126	0.5670	0.7790	0.7606	0.6542	0.7128	0.6661
40	0.7150	0.5127	0.4259	0.5419	0.5815	0.5886	0.5861	0.7057	0.6366
50	0.6797	0.4738	0.4024	0.4600	0.5024	0.8263	0.4959	0.4094	0.5679
60	0.6717	0.4930	0.3907	0.5336	0.7146	0.6343	0.5653	0.7841	0.7876
70	0.6259	0.4179	0.3431	0.3193	0.5260	0.6804	0.5962	0.7183	0.6885
80	0.6634	0.4599	0.3845	0.3301	0.6348	0.5089	0.6636	0.7897	0.7014
90	0.6573	0.4829	0.3815	0.4516	0.6400	0.7462	0.6156	0.6638	0.5765
100	0.6420	0.4638	0.3638	0.4913	0.5106	0.7970	0.5914	0.5158	0.6976
110	0.5610	0.3785	0.2817	0.2896	0.4378	0.6872	0.2745	0.3131	0.3099
120	0.6859	0.5089	0.4062	0.4453	0.5159	0.4242	0.6899	0.8474	0.8932
130	0.6321	0.4414	0.3533	0.5103	0.7589	0.8520	0.6909	0.7030	0.7327
140	0.6066	0.4179	0.3278	0.2981	0.3678	0.6861	0.6914	0.5083	0.5868
150	0.6664	0.4666	0.3949	0.6433	0.3196	0.5582	0.5914	0.6958	0.6960
160	0.6367	0.4617	0.3587	0.3570	0.6022	0.6514	0.6724	0.7549	0.8135
170	0.6471	0.4420	0.3721	0.5036	0.3770	0.7290	0.5193	0.6424	0.7027
180	0.6919	0.5198	0.4166	0.5457	0.6494	0.6634	0.6226	0.8653	0.8836
190	0.6643	0.4863	0.3955	0.4098	0.5833	0.6799	0.7114	0.6521	0.5775
200	0.6762	0.4909	0.3979	0.6806	0.5995	0.9367	0.7052	0.7188	0.8010
210	0.6406	0.4247	0.3604	0.5629	0.5730	0.6664	0.6426	0.9436	0.9675
220	0.5879	0.4083	0.3078	0.3644	0.5280	0.5706	0.5935	0.6922	0.7485
230	0.6493	0.4846	0.3766	0.7708	0.8811	0.8455	0.7138	0.8531	0.9220
240	0.6634	0.4853	0.3889	0.4532	0.4302	0.3939	0.5223	0.7346	0.7395
250	0.6319	0.4524	0.3534	0.6620	0.4452	0.9919	0.6324	0.6982	0.7875
mean	0.6546	0.4655	0.3770	0.4745	0.5677	0.6913	0.6225	0.6976	0.7134

5.5.2 Evaluation of portfolio strategies

Table 5.3 shows the annualized Sharpe ratio by each holding period for the five strategies. The holding period means that the duration of keeping a portfolio composed of the stocks obtained by each strategy at time t. At first, the market index (sp) has the smallest value of 0.3770, showing the worst performance. It means that any other portfolio strategy outperforms the market, and demonstrates the necessity of investment strategies. In particular, the cap-weighted strategy, which is most similar to the market index, outperforms the market and it suggests the efficiency of investing in specific stocks. Interestingly, the $1/N$ strategy (ew) has an average Sharpe ratio of 0.6546, which is about 42.40% and 28.89% higher than that of the market index and cap-weighted strategies, respectively. Since this strategy assigns equal weights to small companies, it is more distributed than the cap-weighted strategy in terms of the number of stocks in the portfolio. In other words, Since the dataset consists of major companies in each sector by market capitalization as of December 31, 2018, they can be overvalued than the actual firm value in the past.

In network approaches, the PMFG strategy has higher Sharpe ratios as the number of stocks increases. It outperforms the naive and classical models when the number of selected stocks is 30, but when it is less than 30, the performance is worse than $1/N$ strategy. That is, it shows better performance to invest in Markowitz weights on stocks that have high peripherality obtained from Eq.(5.3), especially 30 stocks. The proposed method, WCPG strategy also has higher Sharpe ratios as the number of stocks increases and it outperforms all other strategies when more than 20 stocks are selected. If the number of stocks is 10, it tends to invest in too few stocks as a result of Markowitz optimization, resulting in bad performance. In addition, it always outperforms the PMFG strategy regardless of the number of stocks. On the other hand, the Sharpe ratios of the WCPG strategy are similar to or lower than that of several strategies when the holding period τ is shorter than 6 months. But it outperforms most strategies when τ is longer than 6 months.

The information ratio indicates the fluctuation between returns on investment at every time point with the same holding period. Table 5.4 shows the information ratio by each holding period for the five strategies. The results show that the ratio increases as the holding period is longer. It means that the growth of the U.S. stock

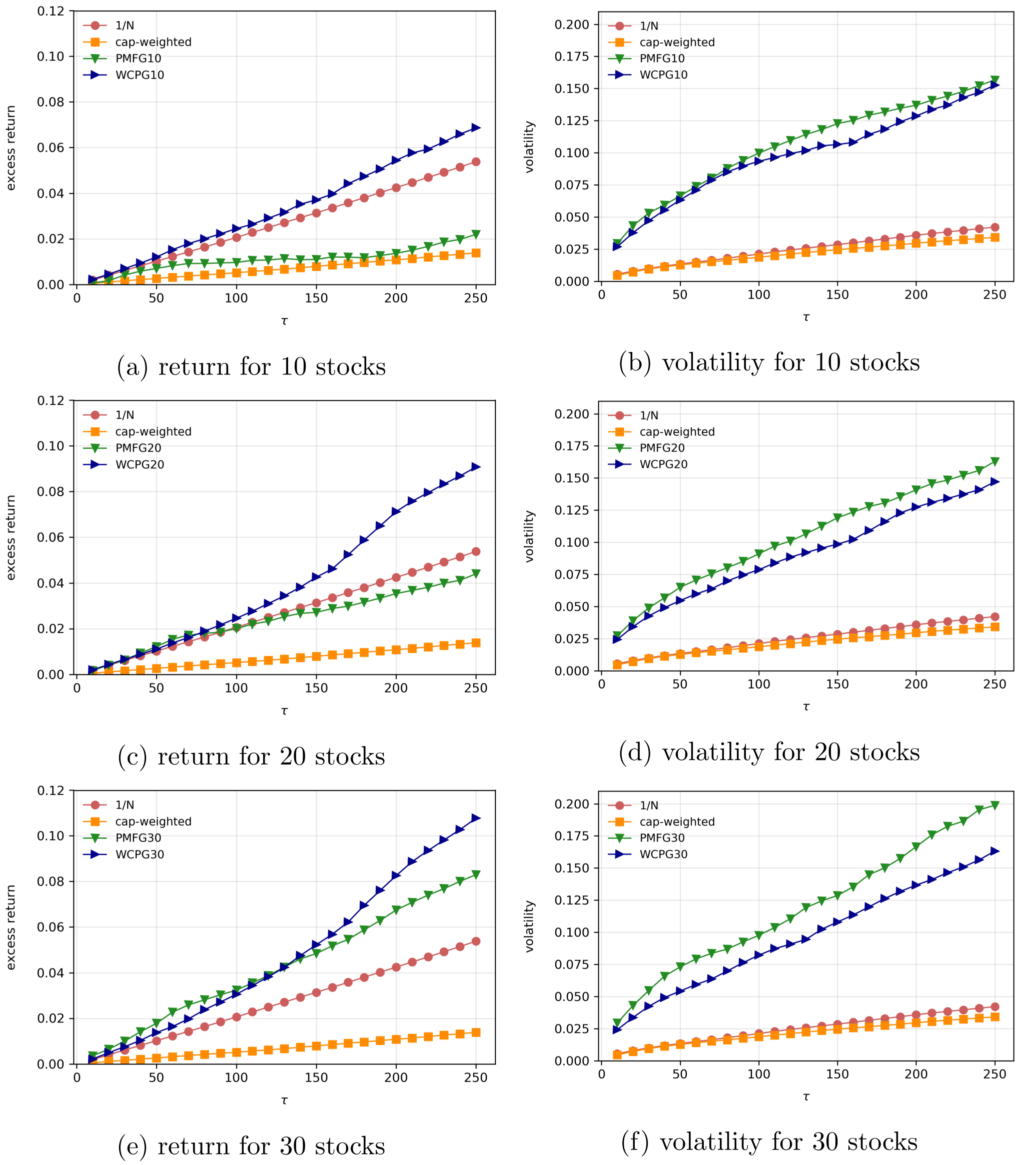
Table 5.4: Information ratio of benchmarks and network approaches

τ	**ew**	**cw**	**sp**	$\mathbf{pmfg}_{10}$	$\mathbf{pmfg}_{20}$	$\mathbf{pmfg}_{30}$	$\mathbf{wcpg}_{10}$	$\mathbf{wcpg}_{20}$	$\mathbf{wcpg}_{30}$
10	0.1670	0.1286	0.1038	0.1067	0.1389	0.1808	0.1361	0.1403	0.1493
20	0.2582	0.2042	0.1621	0.1650	0.2277	0.2589	0.2144	0.2217	0.2363
30	0.3235	0.2577	0.2035	0.2354	0.2824	0.3163	0.2640	0.2737	0.2882
40	0.3828	0.3054	0.2404	0.2948	0.3428	0.3743	0.3052	0.3178	0.3339
50	0.4270	0.3420	0.2678	0.3236	0.3755	0.4150	0.3395	0.3583	0.3811
60	0.4792	0.3860	0.3008	0.3458	0.4150	0.4632	0.3779	0.3982	0.4156
70	0.5294	0.4283	0.3321	0.3712	0.4498	0.5005	0.4032	0.4338	0.4533
80	0.5713	0.4622	0.3591	0.3779	0.4674	0.5270	0.4244	0.4622	0.4911
90	0.6027	0.4906	0.3808	0.3953	0.4839	0.5422	0.4520	0.4966	0.5187
100	0.6280	0.5158	0.3995	0.4146	0.5067	0.5609	0.4780	0.5311	0.5486
110	0.6511	0.5396	0.4167	0.4332	0.5275	0.5807	0.5047	0.5611	0.5802
120	0.6754	0.5624	0.4330	0.4412	0.5496	0.5989	0.5355	0.5946	0.6211
130	0.6981	0.5833	0.4481	0.4574	0.5655	0.6110	0.5623	0.6272	0.6600
140	0.7237	0.6066	0.4642	0.4643	0.5790	0.6278	0.5932	0.6606	0.6819
150	0.7524	0.6319	0.4810	0.4762	0.5819	0.6422	0.6206	0.6934	0.7031
160	0.7762	0.6547	0.4959	0.4965	0.5953	0.6508	0.6501	0.7242	0.7223
170	0.7998	0.6771	0.5110	0.5049	0.6046	0.6501	0.6719	0.7466	0.7414
180	0.8210	0.6966	0.5245	0.5208	0.6282	0.6702	0.6968	0.7717	0.7659
190	0.8412	0.7138	0.5373	0.5393	0.6414	0.6813	0.7167	0.7904	0.7953
200	0.8638	0.7348	0.5520	0.5607	0.6566	0.6917	0.7449	0.8307	0.8317
210	0.8866	0.7553	0.5667	0.5781	0.6698	0.6980	0.7623	0.8596	0.8669
220	0.9140	0.7781	0.5824	0.6005	0.6888	0.7066	0.7751	0.8901	0.8930
230	0.9407	0.7995	0.5973	0.6176	0.7055	0.7135	0.7931	0.9172	0.9208
240	0.9682	0.8229	0.6131	0.6352	0.7273	0.7142	0.8182	0.9482	0.9438
250	0.9892	0.8419	0.6256	0.6567	0.7463	0.7300	0.8333	0.9632	0.9618
mean	0.6668	0.5568	0.4240	0.4405	0.5263	0.5642	0.5469	0.6085	0.6202

market during this period provides better results for long-term investors. In addition, the $1/N$ strategy outperforms the cap-weighted strategy and the market index. It implies that companies with small market-caps show better growth under the given risk rather than companies with large market-caps, which is called *small-cap premium* introduced in Fama and French (1993). For example, the market-cap of NETFLIX on January 3, 2005 is \$624.14 million, which is the lowest among 110 stocks. However, on December 31, 2018, it grows approximately 187 times to $\$116,859.90$ million. Similarly, BOOKING HOLDINGS and AMAZON.COM increase the market-cap by about 85 times and 40 times in the same period, respectively. On the other hand, GENERAL ELECTRIC and EXXON MOBIL, which have the largest market-caps in 2005, shows negative growth of -83% and -11%. For this reason, the $1/N$ strategy is superior to the cap-weighted strategy. However, since both strategies invest in too many assets, it is difficult for investors to understand them correctly, which can be a disadvantage in terms of risk management. Also, holding a large number of assets regardless of the market-cap generates a lot of transaction costs since it requires more frequent trading of smaller stocks, which have less liquidity. In addition, holding more proportion of small-cap companies that generally pay out smaller dividends than large-cap companies is another disadvantage of strategies with a large number of stocks.

Furthermore, the result shows that the performance of the PMFG strategy according to holding periods and the number of stocks. Noticeably, it shows an improvement in performance when the number of selected stocks increased from 10 to 20, and from 20 and 30. The longer the holding period, the larger the information ratio, but the performance improved slowly when the holding period is longer than 3 months. In the WCPG strategy, the performance improves as the number of stocks increases, there is no noticeable change when it increases from 20 to 30, unlike the PMFG strategy. However, the difference with the PMFG strategy is that it continues to improve even if the holding period is longer than 3 months. Then, further analysis is conducted about the difference between the strategies, and the results are shown in Figure 5.3.

Figure 5.3a, 5.3c, 5.3e show the mean excess return to market index for each holding period when the number of selected stocks is 10, 20 and 30, respectively. In the network approaches, the excess returns increase as the number of stocks



(a) return for 10 stocks

(b) volatility for 10 stocks

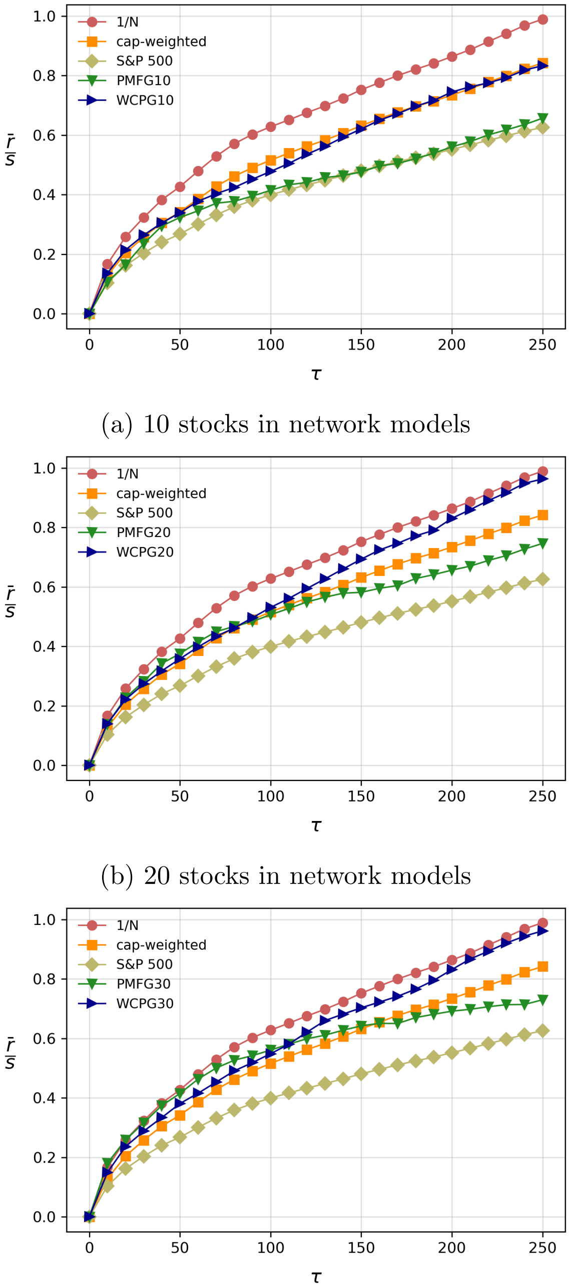
(c) return for 20 stocks

(d) volatility for 20 stocks

(e) return for 30 stocks

(f) volatility for 30 stocks

Figure 5.3: Excess return and standard deviation of investment strategies



(a) 10 stocks in network models

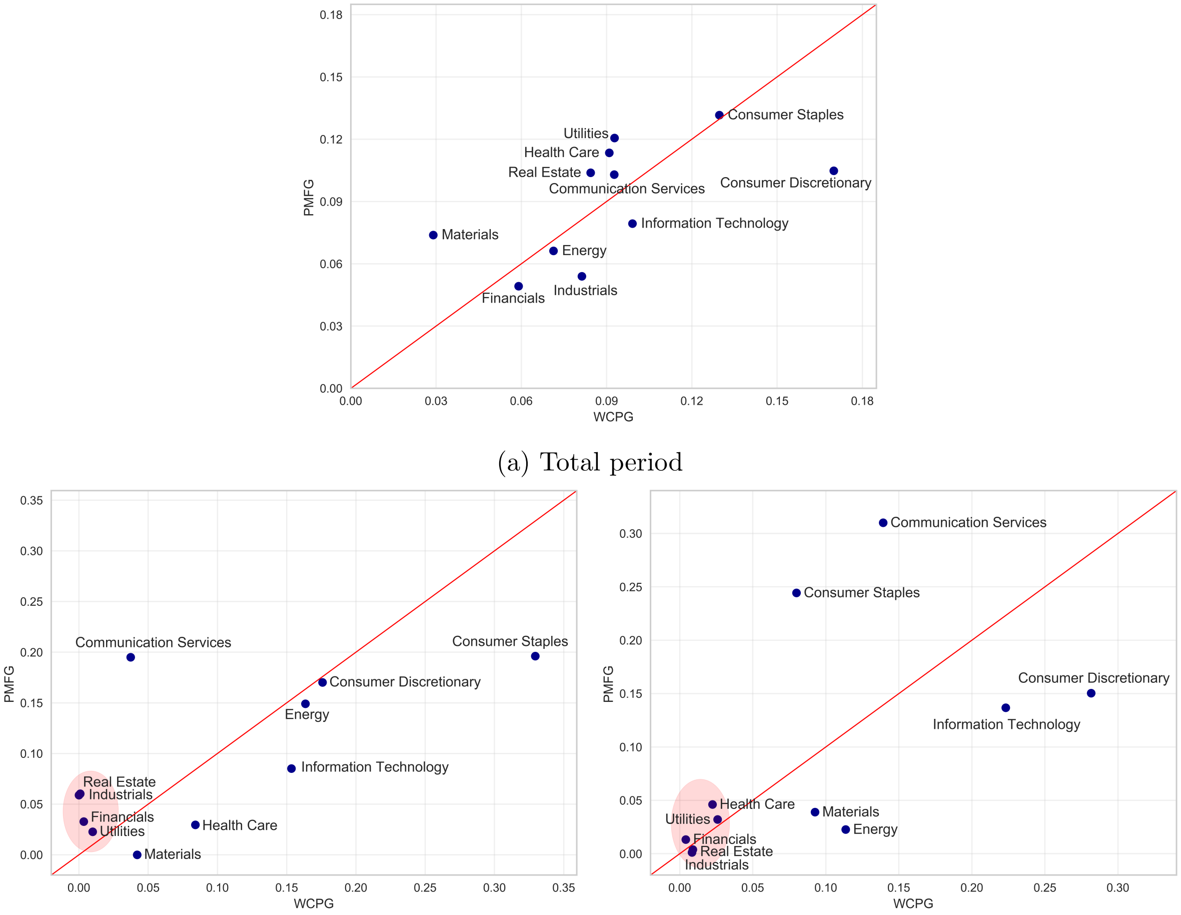
(b) 20 stocks in network models

(c) 30 stocks in network models

Figure 5.4: Information ratio for naive, classical, and network models

increase. For 30 stocks, the WCPG and PMFG strategy have much higher mean excess return rather than the $1/N$ and the cap-weighted strategy. Especially, the difference is noticeable when the holding period is over 6 months. Figure 5.3b, 5.3d, 5.3f show the average volatility for each holding period when the number of selected stocks is 10, 20 and 30, respectively. In this case, the PMFG strategy has a higher volatility than that of the WCPG strategy. It increases as the number of stocks increases, while that of the WCPG strategy remains almost the same. Note that the $1/N$ and the cap-weighted strategy have much lower volatility since the number of stocks is much larger. In summary, the WCPG strategy has a much higher excess return compared to that of naive and classical models, while the volatility is higher than that of those models.

Figure 5.4 summarizes the results and is divided into three subfigures according to the number of selected stocks of the network-based strategies. In Figure 5.4a, the performances of the WCPG and PMFG strategy are similar to that of the cap-weighted and S&P 500 index, respectively. Also, the performance of the $1/N$ strategy outperforms other methods. However, as shown in Figure 5.4b and 5.4c, the WCPG strategy outperforms other strategies except for the $1/N$ strategy. Especially, the slope of the information rate is relatively steep for long-term investments longer than 6 months whose performance is almost the same as the $1/N$ strategy. As summarized in Table 5.4, for all holding periods, the WCPG strategy outperforms about 11.39%, 46.29%, 9.92% compared to the cap-weighted, the market index, and the PMFG strategy, respectively. Even though it has lower information ratio rather than $1/N$ strategy, it is competitive since the $1/N$ strategy deals with too many stocks to operate.

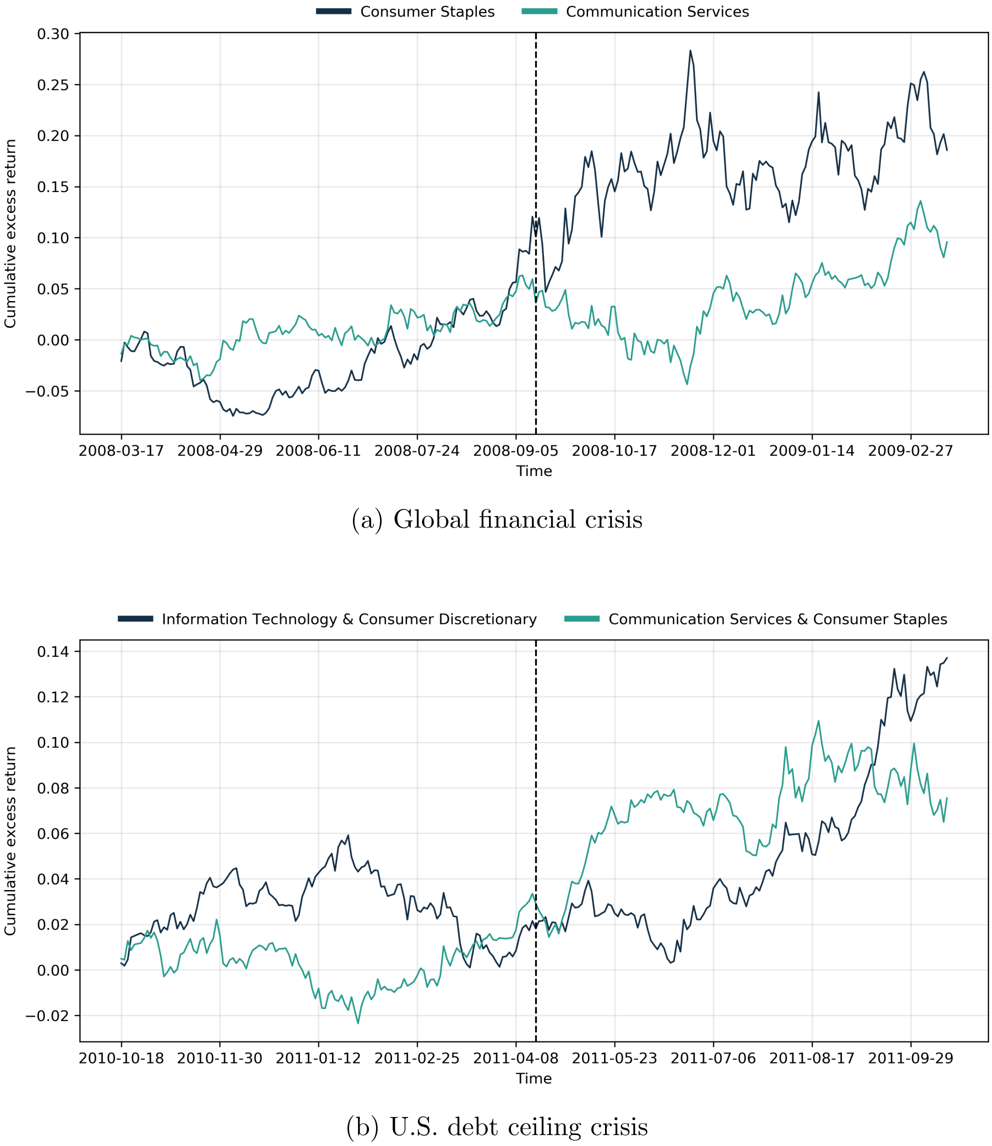


(a) Total period

(b) Global financial crisis

(c) U.S. debt ceiling crisis

Figure 5.5: Average portfolio composition during 6 months before the outbreak of financial crises



(a) Global financial crisis

(b) U.S. debt ceiling crisis

Figure 5.6: Cumulative excess return 6 months before and after the outbreak of financial crises

Table 5.5: Volatility increments 6 months before and after the financial crisis

	Global financial crisis	U.S. debt ceiling crisis
Information Technology	0.9042	0.3615
Consumer Discretionary	0.7146	0.4001
Financials	**1.3202**	**0.9011**
Health Care	1.2249	0.5305
Consumer Staples	1.0553	0.4107
Energy	1.3400	0.6083
Communication Services	1.2696	0.4144
Industrials	1.0863	**0.7294**
Utilities	**1.5125**	0.5369
Materials	1.1786	0.5246
Real Estate	**1.7892**	**0.6349**

From the perspective of portfolio composition, both strategies with 30 stocks showing the best performance are further analyzed. Figure 5.5 describes the average proportion of each sector that compose the portfolio according to two strategies. In Figure 5.5a, the result shows that the WCPG strategy is characterized by consisting of high proportions in three sectors during the entire experimental period: Consumer Discretionary whose includes AMAZON.COM and MACDONALDS, Consumer Staples whose includes WALMART and COCA-COLA, and Information Technology whose includes APPLE and MICROSOFT. Interestingly, all these sectors are closely related to consumers. Figure 5.5b and 5.5c show the average proportions of the portfolio composition during the 6 months before the outbreak of each crisis that has a direct impact on the U.S. stock market: Global financial crisis in 2008 and the U.S. debt ceiling crisis in 2011. Note that Table 5.5 indicates the increments of volatility 6 months before and after both crises. The result shows that the volatilities of sectors such as Real Estate, Financials, and Utilities notably increase after the global financial crisis. In addition, the volatilities of sectors such as Financials, Industrials, and Real Estate sector increase after the U.S. debt ceiling crisis. In this regard, the

WCPG strategy has very little proportion of high volatility sectors, which is marked by red circles.

Furthermore, it is known that most of the excess return by Consumer Staples sector is generated around the global financial crisis since the companies produce products that consumers need regardless of the condition of the economy. Figure 5.6 shows the cumulative excess return (CER) of sectors that has a high proportion in the portfolio for the market index. The black dash line refers to the date of the outbreak of the crisis. In Figure 5.6a, the CERs of two sectors that have the highest weight in each strategy is plotted. The result shows that the CER of Consumer Staples soars whereas that of Communication Services goes sideways. In Figure 5.5c, Information and Consumer Discretionary, and Communication Services and Consumer Staples have the largest weights in the WCPG and PMFG strategies, respectively. The sum of the CERs of two sectors in each strategy is described in Figure 5.6b. In this case, it is not clear which sector is superior as shown in Figure 5.6a. However, both have positive CER after the crisis. In summary, the result shows that WCPG strategy can be less exposed to the oncoming risk by reducing the proportion of sectors affected significantly by the financial crisis in advance.

5.6 Summary and Discussion

In this chapter, a framework for constructing the portfolio is proposed based on the link prediction method in the Granger causality network. By comparing with the performance of each benchmark using two portfolio evaluation measures, it is discovered that the proposed method achieves competitive outcomes.

At first, it is discovered that network-based strategies outperform the classical models in terms of Sharpe ratio. Portfolios are constructed at every time point during the experimental period and maintained for holding periods, which ranges from 10 to 250 days. The results show that all strategies have better performance than the market index, which implicates the necessity of portfolio selection or asset allocation. Interestingly, although the $1/N$ rule strategy is the most naive approach, it has much higher Sharpe ratio than the cap-weighted strategy and the market index. Besides, from the perspective of the network approach, the performance is improved as the number of stocks increased. Notably, the proposed method has an advantage in a long-term investment of more than 6 months.

Secondly, the value of the information ratio for each holding period is used to offset the deviation in portfolio performance according to the rebalancing time point. The proposed model outperforms most of the strategies except the $1/N$ strategy. However, it should be stressed that the $1/N$ strategy has disadvantages of dealing with a large number of stocks and being more costly to manage. In addition, one of the advantages of the WCPG strategy is that it has a much higher excess return compared to that of naive and classical models even though the volatility is higher than that of those models. From the perspective of the network approaches, the excess return of both strategies gradually increases as the number of selected stocks and the holding period increases. Noticeably, the WCPG strategy outperforms the PMFG strategy in investments of more than 6 months, which is beneficial to long-term investors.

In conclusion, it is discovered that the link prediction considering the cause-and-effect relationship between financial institutions is useful to construct portfolios. As a result, the application of link prediction is suitable for managing financial risk.

Chapter 6

Concluding Remarks

6.1 Contributions and Limitations

The disturbance of financial crises has resulted in the collapse of the interconnected financial system, leading to bankruptcy in many countries or institutions. For this reason, the world has realized to prevent another possible crisis. Consequently, many researchers have tried to develop models for mapping the system and manage the financial risk. However, the previous studies using complex networks on financial markets are mostly about observing the properties of past events based on historical data rather than coping with future events. Therefore, this dissertation proposes a more improved model based on various financial data. To achieve the purpose, this dissertation deals with two domains. The first is the network model, and the other is the investment model. They are applied to analyze the global currency market and the U.S. stock market in macroscopic and microscopic perspectives, respectively.

For developing the network model, the Granger causality network is considered to represent the dependency within the financial system. For the purpose of predicting links that can be appeared shortly, it is necessary to ensure the market is properly mapped to the network. Once the network is constructed, it is analyzed based on the cross-sectional topology to observe whether the system is mapped appropriately. The mapping of directed networks proves that it has similar topological properties to other empirical financial markets; the return of data has a non-Gaussian distribution and follows the power-law distribution. Especially, it is confirmed that the emergence of financial crises is well reflected in the network structure. In addition, it is analyzed based on the moving window method for generating time-varying measures. The results show that the number of connections within the network provides evidence of an approaching crisis and produces the proxy of the topological importance of each institution.

For predicting links within the network, the advanced method is considered to represent the strength of links between different nodes. The contribution of this dissertation is focused on describing the interactions between countries based on the effect size, which is obtained from F-statistics in the Granger causality test. Once the models are derived, the performance is evaluated for each model based on AUC for five prediction lags: 21, 42, 63, 126, and 252 days. Indeed, the results of the proposed models show the improved prediction accuracy and it falls at the stages of two major financial crises. Based on this model, the link prediction model is suggested for the application in investment for managing financial risk.

For the application of the proposed link prediction model, the variation of PMFG network and the concept of peripherality is utilized to construct the portfolio. At first, the performance of benchmarks and the proposed model is evaluated by the Sharpe ratio. It is turned out that network-based models have advantages of ensuring better returns under the given risk. However, the $1/N$ strategy, cap-weighted strategy, and the market index have little deviation in the Sharpe ratio when the holding period changes, while the network-based models have larger ones. Secondly, in the Information ratio, the $1/N$ strategy is further improved in performance than others. However, it has obvious disadvantages that operate a large number of stocks, which increases management costs. On the other hand, even though the volatility of the proposed model is higher than the naive and classical model, it has a much higher mean excess return. It is attractive for investors who have higher risk tolerance. In addition, the WCPG strategy becomes more prominent in long-term investments for more than 6 months, especially in terms of returns. Lastly, the portfolio compositions before the financial crises are analyzed. Interestingly, the result demonstrates the clear distinction between the WCPG strategy and the PMFG strategy by decreasing the proportions of vulnerable sectors to the crises in advance.

In conclusion, conventional network analysis is well upgraded so that they represent cause-and-effect relationships between assets. The proposed method of this dissertation provides a strategy for managing financial risk based on the prediction of the network evolution. Derived from the Granger causality test, the eta-squared-based measure provides an inference that the connection strength is different even at the same connected edge in the network. Also, the proposed portfolio strategy shows better performance for a given risk. Despite the excellent performance and perception, the proposed link prediction model and the portfolio strategy also have limitations. At first, in the perspective of the link prediction, the Granger causality test can incur misinterpretations without prior knowledge of the phenomenon due to many reasons for rejecting the null hypothesis (Maziarz, 2015). This dissertation focuses on the link prediction in the causality network as a theoretical representation of the financial market rather than finding the economic implications in each Granger causality test. Consequently, only circumstantial evidence of the currency market is provided based on the changes in the topology of the network and its time-varying properties due to the financial crises. Furthermore, in the case of developing countries, the currency market is unstable, and the movement of REER is volatile. Also, the degree of government's involvement in the currency market varies in countries, which may dilute the meaning of the causality network. Secondly, from the perspective of portfolio optimization, the benchmark model has a problem with data selection. In this dissertation, only the companies with the largest market capitalization on the most recent date during the experimental period are considered. They might be invested more than the actual value even though the market capitalization was extremely small in the past. It may overestimate the strategy and leads to misinterpretation. But still, the advantage of the proposed models are satisfactory.

6.2 Future Work

This dissertation also has the part to further develop that should be addressed in future work. At first, Granger causality in the network model only considers the simple bivariate interdependency. Instead, the concept of the modern auto-regressive approach can be considered to overcome this, such as the multivariate autoregressive conditional heteroskedasticity (ARCH), the generalized ARCH (GARCH), and vine copula. They are expected to provide rich economic implications. Secondly, in perspective of link prediction, all the methods fail to predict the structural changes incurred from the tail events. That is, the prediction accuracy relatively decreases when the network structure changes significantly due to the financial crisis. Instead, machine learning algorithms with external variables should be developed to predict extreme structural changes. It requires the idea of how to reflect the abnormal signals to training sets and extract the parameters from them. Thirdly, the time-varying properties of networks can be further analyzed whether it can be utilized as an alarm index for the financial crisis. Since the connectivity or centrality measures contain information about the market structure, it can provide useful inferences about the changes, especially sudden big changes. Lastly, the portfolio selection models are only focused on the network approaches that depend on the metrics such as the correlation and the effect size. Also, the only variables that influence investors' decision-making are the expected return and the variance of the return. Instead, it can be developed by using other technical indicators of stocks and by applying other weighting techniques rather than Markowitz weights. In addition, it requires further analysis of what characteristics the portfolio usually has, and how it differs from other strategies.

Bibliography

Adamic, L. A. and Adar, E. (2003). Friends and neighbors on the web. *Social networks*, 25(3):211–230.

Amaral, L. A. N., Scala, A., Barthelemy, M., and Stanley, H. E. (2000). Classes of small-world networks. *Proceedings of the national academy of sciences*, 97(21):11149–11152.

Aste, T., Shaw, W., and Di Matteo, T. (2010). Correlation structure and dynamics in volatile markets. *New Journal of Physics*, 12(8):085009.

Back, K. (2006). *A course in derivative securities: Introduction to theory and computation.* Springer Science & Business Media.

Bagler, G. (2008). Analysis of the airport network of india as a complex weighted network. *Physica A: Statistical Mechanics and its Applications*, 387(12):2972–2980.

Barabási, A.-L. and Albert, R. (1999). Emergence of scaling in random networks. *science*, 286(5439):509–512.

Barabási, A.-L. and Bonabeau, E. (2003). Scale-free networks. *Scientific american*, 288(5):60–69.

Barrat, A., Barthelemy, M., and Vespignani, A. (2007). The architecture of complex weighted networks: Measurements and models. In *Large Scale Structure And Dynamics Of Complex Networks: From Information Technology to Finance and Natural Science*, pages 67–92. World Scientific.

Bessler, W., Opfer, H., and Wolff, D. (2017). Multi-asset portfolio optimization and out-of-sample performance: an evaluation of black–litterman, mean-variance, and naïve diversification approaches. *The European Journal of Finance*, 23(1):1–30.

Black, F. and Litterman, R. (1990). Asset allocation: combining investor views with market equilibrium. *Goldman Sachs Fixed Income Research*, 115.

Bliss, C. A., Frank, M. R., Danforth, C. M., and Dodds, P. S. (2014). An evolutionary algorithm approach to link prediction in dynamic social networks. *Journal of Computational Science*, 5(5):750–764.

Boginski, V., Butenko, S., Shirokikh, O., Trukhanov, S., and Lafuente, J. G. (2014). A network-based data mining approach to portfolio selection via weighted clique relaxations. *Annals of Operations Research*, 216(1):23–34.

Brown, J. D. (2008). Effect size and eta squared. *JALT Testing & Evaluation SIG News*.

Castilho, D., Gama, J., Mundim, L. R., and de Carvalho, A. C. (2019). Improving portfolio optimization using weighted link prediction in dynamic stock networks. In *International Conference on Computational Science*, pages 340–353. Springer.

Chebotarev, P. and Shamis, E. (2006). The matrix-forest theorem and measuring relations in small social groups. *arXiv preprint math/0602070*.

Chen, G., Wang, X., and Li, X. (2014). *Fundamentals of complex networks: models, structures and dynamics.* John Wiley & Sons.

Chen, H., Li, X., and Huang, Z. (2005). Link prediction approach to collaborative filtering. In *Digital Libraries, 2005. JCDL'05. Proceedings of the 5th ACM/IEEE-CS Joint Conference on*, pages 141–142. IEEE.

Chowdhury, G. G. (2010). *Introduction to modern information retrieval.* Facet publishing.

Clauset, A., Moore, C., and Newman, M. E. (2008). Hierarchical structure and the prediction of missing links in networks. *Nature*, 453(7191):98–101.

Comellas, F. and Sampels, M. (2002). Deterministic small-world networks. *Physica A: Statistical Mechanics and its Applications*, 309(1-2):231–235.

Cont, R. (2001). Empirical properties of asset returns: stylized facts and statistical issues.

Dai, Z. and Wang, F. (2019). Sparse and robust mean–variance portfolio optimization problems. *Physica A: Statistical Mechanics and its Applications*, 523:1371–1378.

DeMiguel, V., Garlappi, L., and Uppal, R. (2009). Optimal versus naive diversification: How inefficient is the 1/n portfolio strategy? *The review of Financial studies*, 22(5):1915–1953.

Derényi, I., Farkas, I., Palla, G., and Vicsek, T. (2004). Topological phase transitions of random networks. *Physica A: Statistical Mechanics and its Applications*, 334(3-4):583–590.

Dichtl, H., Drobetz, W., and Wambach, M. (2016). Testing rebalancing strategies for stock-bond portfolios across different asset allocations. *Applied Economics*, 48(9):772–788.

Duchin, R. and Levy, H. (2009). Markowitz versus the talmudic portfolio diversification strategies. *The Journal of Portfolio Management*, 35(2):71–74.

Fama, E. F. and French, K. R. (1993). Common risk factors in the returns on stocks and bonds. *Journal of.*

Fouss, F., Pirotte, A., Renders, J.-M., and Saerens, M. (2007). Random-walk computation of similarities between nodes of a graph with application to collaborative recommendation. *IEEE Transactions on knowledge and data engineering*, 19(3):355–369.

Geisser, S. (1993). *Predictive inference*, volume 55. CRC press.

Getoor, L. and Diehl, C. P. (2005). Link mining: a survey. *Acm Sigkdd Explorations Newsletter*, 7(2):3–12.

Göbel, F. and Jagers, A. (1974). Random walks on graphs. *Stochastic processes and their applications*, 2(4):311–336.

Granger, C. W. (1969). Investigating causal relations by econometric models and cross-spectral methods. *Econometrica: Journal of the Econometric Society*, pages 424–438.

Guns, R. (2011). Bipartite networks for link prediction: Can they improve prediction performance. In *Proceedings of ISSI*, volume 13, pages 249–260.

Hayes, A. F. (2009). *Statistical methods for communication science*. Routledge.

Herlocker, J. L., Konstan, J. A., Terveen, L. G., and Riedl, J. T. (2004). Evaluating collaborative filtering recommender systems. *ACM Transactions on Information Systems (TOIS)*, 22(1):5–53.

Hidalgo, C. A. and Rodríguez-Sickert, C. (2008). The dynamics of a mobile phone network. *Physica A: Statistical Mechanics and its Applications*, 387(12):3017–3024.

Huang, W.-Q., Zhuang, X.-T., and Yao, S. (2009). A network analysis of the chinese stock market. *Physica A: Statistical Mechanics and its Applications*, 388(14):2956–2964.

Jaccard, P. (1901). Étude comparative de la distribution florale dans une portion des alpes et des jura. *Bull Soc Vaudoise Sci Nat*, 37:547–579.

Jeh, G. and Widom, J. (2002). Simrank: a measure of structural-context similarity. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 538–543. ACM.

Jobson, J. D. and Korkie, B. (1980). Estimation for markowitz efficient portfolios. *Journal of the American Statistical Association*, 75(371):544–554.

Jung, W.-S., Chae, S., Yang, J.-S., and Moon, H.-T. (2006). Characteristics of the korean stock market correlations. *Physica A: Statistical Mechanics and its Applications*, 361(1):263–271.

Katz, L. (1953). A new status index derived from sociometric analysis. *Psychometrika*, 18(1):39–43.

Kenett, D. Y., Tumminello, M., Madi, A., Gur-Gershgoren, G., Mantegna, R. N., and Ben-Jacob, E. (2010). Dominating clasp of the financial sector revealed by partial correlation analysis of the stock market. *PloS one*, 5(12).

Klau, M. and Fung, S. S. (2006). The new bis effective exchange rate indices.

Klein, D. J. and Randić, M. (1993). Resistance distance. *Journal of mathematical chemistry*, 12(1):81–95.

Konno, T. (2016). Knowledge spillover processes as complex networks. *Physica A: Statistical Mechanics and its Applications*, 462:1207–1214.

Kossinets, G. (2006). Effects of missing data in social networks. *Social networks*, 28(3):247–268.

Lambiotte, R., Blondel, V. D., De Kerchove, C., Huens, E., Prieur, C., Smoreda, Z., and Van Dooren, P. (2008). Geographical dispersal of mobile communication networks. *Physica A: Statistical Mechanics and its Applications*, 387(21):5317–5325.

Latora, V. and Marchiori, M. (2002). Is the boston subway a small-world network? *Physica A: Statistical Mechanics and its Applications*, 314(1-4):109–113.

Lei, C. and Ruan, J. (2012). A novel link prediction algorithm for reconstructing protein–protein interaction networks by topological similarity. *Bioinformatics*, 29(3):355–364.

Lei, C. and Ruan, J. (2013). A novel link prediction algorithm for reconstructing protein–protein interaction networks by topological similarity. *Bioinformatics*, 29(3):355–364.

Leicht, E. A., Holme, P., and Newman, M. E. (2006). Vertex similarity in networks. *Physical Review E*, 73(2):026120.

Li, Y., Jiang, X.-F., Tian, Y., Li, S.-P., and Zheng, B. (2019). Portfolio optimization based on network topology. *Physica A: Statistical Mechanics and its Applications*, 515:671–681.

Liben-Nowell, D. and Kleinberg, J. (2007). The link-prediction problem for social networks. *Journal of the American society for information science and technology*, 58(7):1019–1031.

Liu, J., Wu, J., and Yang, Z. (2004). The spread of infectious disease on complex networks with household-structure. *Physica A: Statistical Mechanics and its Applications*, 341:273–280.

Liu, J., Xiong, Q., Shi, W., Shi, X., and Wang, K. (2016). Evaluating the importance of nodes in complex networks. *Physica A: Statistical Mechanics and its Applications*, 452:209–219.

Liu, W. and Lü, L. (2010). Link prediction based on local random walk. *EPL (Europhysics Letters)*, 89(5):58007.

Lü, L., Jin, C.-H., and Zhou, T. (2009). Similarity index based on local paths for link prediction of complex networks. *Physical Review E*, 80(4):046122.

Lü, L. and Zhou, T. (2011). Link prediction in complex networks: A survey. *Physica A: statistical mechanics and its applications*, 390(6):1150–1170.

Ma, C., Zhou, T., and Zhang, H.-F. (2016a). Playing the role of weak clique property in link prediction: A friend recommendation model. *Scientific reports*, 6:30098.

Ma, L.-l., Ma, C., Zhang, H.-F., and Wang, B.-H. (2016b). Identifying influential spreaders in complex networks based on gravity formula. *Physica A: Statistical Mechanics and its Applications*, 451:205–212.

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

Mantegna, R. N. and Stanley, H. E. (1999). *Introduction to econophysics: correlations and complexity in finance.* Cambridge university press.

Markowitz, H. (1952). Portfolio selection. *The Journal of Finance*, 7:77–91.

Maxwell, S. E., Camp, C. J., and Arvey, R. D. (1981). Measures of strength of association: A comparative examination. *Journal of Applied Psychology*, 66(5):525.

Maziarz, M. (2015). A review of the granger-causality fallacy. *The journal of philosophical economics: Reflections on economic and social issues*, 8(2):86–105.

Meghanathan, N. (2016). *Advanced methods for complex network analysis.* IGI Global.

Menon, A. K. and Elkan, C. (2011). Link prediction via matrix factorization. In *Joint european conference on machine learning and knowledge discovery in databases*, pages 437–452. Springer.

Merton, R. C. (1980). On estimating the expected return on the market: An exploratory investigation. Technical report, National Bureau of Economic Research.

Murata, T. and Moriyasu, S. (2007). Link prediction of social networks based on weighted proximity measures. In *Proceedings of the IEEE/WIC/ACM international conference on web intelligence*, pages 85–88. IEEE Computer Society.

Murata, T. and Moriyasu, S. (2008). Link prediction based on structural properties of online social networks. *New Generation Computing*, 26(3):245–257.

Najafi, A. A. and Pourahmadi, Z. (2016). An efficient heuristic method for dynamic portfolio selection problem under transaction costs and uncertain conditions. *Physica A: Statistical Mechanics and Its Applications*, 448:154–162.

Nakagawa, S. and Cuthill, I. C. (2007). Effect size, confidence interval and statistical significance: a practical guide for biologists. *Biological reviews*, 82(4):591–605.

Nanda, S., Mahanty, B., and Tiwari, M. (2010). Clustering indian stock market data for portfolio management. *Expert Systems with Applications*, 37(12):8793–8798.

Newman, M. E. (2001). Clustering and preferential attachment in growing networks. *Physical review E*, 64(2):025102.

Newman, M. E. (2004). Analysis of weighted networks. *Physical review E*, 70(5):056131.

Noh, J. D. and Rieger, H. (2004). Random walks on complex networks. *Physical review letters*, 92(11):118701.

Olejnik, S. and Algina, J. (2003). Generalized eta and omega squared statistics: measures of effect size for some common research designs. *Psychological methods*, 8(4):434.

Onnela, J.-P., Chakraborti, A., Kaski, K., Kertesz, J., and Kanto, A. (2003). Asset trees and asset graphs in financial markets. *Physica Scripta*, 2003(T106):48.

Opsahl, T., Agneessens, F., and Skvoretz, J. (2010). Node centrality in weighted networks: Generalizing degree and shortest paths. *Social networks*, 32(3):245–251.

Opsahl, T., Colizza, V., Panzarasa, P., and Ramasco, J. J. (2008). Prominence and control: the weighted rich-club effect. *Physical review letters*, 101(16):168702.

Ou, Q., Jin, Y.-D., Zhou, T., Wang, B.-H., and Yin, B.-Q. (2007). Power-law strength-degree correlation from resource-allocation dynamics on weighted networks. *Physical Review E*, 75(2):021102.

Pan, J.-Y., Yang, H.-J., Faloutsos, C., and Duygulu, P. (2004). Automatic multimedia cross-modal correlation discovery. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 653–658. ACM.

Papadimitriou, A., Symeonidis, P., and Manolopoulos, Y. (2012). Fast and accurate link prediction in social networking systems. *Journal of Systems and Software*, 85(9):2119–2132.

Park, J. H., Chang, W., and Song, J. W. (2020). Link prediction in the granger causality network of the global currency market. *Physica A: Statistical Mechanics and its Applications*, page 124668.

Peralta, G. and Zareei, A. (2016). A network approach to portfolio selection. *Journal of empirical finance*, 38:157–180.

Polanco-Martínez, J. M., Fernández-Macho, J., Neumann, M., and Faria, S. H. (2018). A pre-crisis vs. crisis analysis of peripheral eu stock markets by means of wavelet transform and a nonlinear causality test. *Physica A: Statistical Mechanics and its Applications*, 490:1211–1227.

Pozzi, F., Di Matteo, T., and Aste, T. (2008). Centrality and peripherality in filtered graphs from dynamical financial correlations. *Advances in Complex Systems*, 11(06):927–950.

Pozzi, F., Di Matteo, T., and Aste, T. (2013). Spread of risk across financial markets: better to invest in the peripheries. *Scientific reports*, 3:1665.

Preis, T., Kenett, D. Y., Stanley, H. E., Helbing, D., and Ben-Jacob, E. (2012). Quantifying the behavior of stock correlations under market stress. *Scientific reports*, 2:752.

Price, D. J. D. S. (1965). Networks of scientific papers. *Science*, pages 510–515.

Ravasz, E., Somera, A. L., Mongru, D. A., Oltvai, Z. N., and Barabási, A.-L. (2002). Hierarchical organization of modularity in metabolic networks. *science*, 297(5586):1551–1555.

Ren, F., Lu, Y.-N., Li, S.-P., Jiang, X.-F., Zhong, L.-X., and Qiu, T. (2017). Dynamic portfolio strategy using clustering approach. *PloS one*, 12(1).

Rom, B. M. and Ferguson, K. W. (1994). Post-modern portfolio theory comes of age. *Journal of Investing*, 3(3):11–17.

Samuelson, P. A. (1967). General proof that diversification pays. *Journal of Financial and Quantitative Analysis*, 2(1):1–13.

Schafer, J. L. and Graham, J. W. (2002). Missing data: our view of the state of the art. *Psychological methods*, 7(2):147.

Sharpe, W. F. (1966). Mutual fund performance. *The Journal of business*, 39(1):119–138.

Shilling, A. G. (1992). Market timing: Better than a buy-and-hold strategy. *Financial Analysts Journal*, 48(2):46–50.

Simonsen, I., Eriksen, K. A., Maslov, S., and Sneppen, K. (2004). Diffusion on complex networks: a way to probe their large-scale topological structures. *Physica A: Statistical Mechanics and its Applications*, 336(1-2):163–173.

Soekarno, S. and Damayanti, S. M. (2012). Asset allocation based investment strategy to improve profitability and sustainability of the smes. *Procedia Economics and Finance*, 4:177–192.

Song, D.-M., Tumminello, M., Zhou, W.-X., and Mantegna, R. N. (2011). Evolution of worldwide stock markets, correlation structure, and correlation-based graphs. *Physical Review E*, 84(2):026108.

Sørensen, T. A. (1948). A method of establishing groups of equal amplitude in plant sociology based on similarity of species content and its application to analyses of the vegetation on danish commons. *Biol. Skar.*, 5:1–34.

Stumpf, M. P., Thorne, T., de Silva, E., Stewart, R., An, H. J., Lappe, M., and Wiuf, C. (2008). Estimating the size of the human interactome. *Proceedings of the National Academy of Sciences*, 105(19):6959–6964.

Sullivan, G. M. and Feinn, R. (2012). Using effect size - or why the p value is not enough. *Journal of graduate medical education*, 4(3):279–282.

Tabak, B. M., Serra, T. R., and Cajueiro, D. O. (2010). Topological properties of stock market networks: The case of brazil. *Physica A: Statistical Mechanics and its Applications*, 389(16):3240–3249.

Tadić, B. and Thurner, S. (2004). Information super-diffusion on structured networks. *Physica A: Statistical Mechanics and its Applications*, 332:566–584.

Tola, V., Lillo, F., Gallegati, M., and Mantegna, R. N. (2008). Cluster analysis for portfolio optimization. *Journal of Economic Dynamics and Control*, 32(1):235–258.

Tumminello, M., Aste, T., Di Matteo, T., and Mantegna, R. N. (2005). A tool for filtering information in complex systems. *Proceedings of the National Academy of Sciences*, 102(30):10421–10426.

Tumminello, M., Di Matteo, T., Aste, T., and Mantegna, R. N. (2007). Correlation based networks of equity returns sampled at different time horizons. *The European Physical Journal B*, 55(2):209–217.

Turner, P. and vant Dack, J. (1993). *Measuring international price and cost competitiveness*. Number 39. Bank for International Settlements, Monetary and Economic Department.

Vỳrost, T., Lyócsa, Š., and Baumöhl, E. (2015). Granger causality stock market networks: Temporal proximity and preferential attachment. *Physica A: Statistical Mechanics and its Applications*, 427:262–276.

Vỳrost, T., Lyócsa, Š., and Baumöhl, E. (2019). Network-based asset allocation strategies. *The North American Journal of Economics and Finance*, 47:516–536.

Wang, X. F. and Chen, G. (2002). Pinning control of scale-free dynamical networks. *Physica A: Statistical Mechanics and its Applications*, 310(3-4):521–531.

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

Wu, J.-J., Gao, Z.-Y., and Sun, H.-j. (2008). Optimal traffic networks topology: A complex networks perspective. *Physica A: Statistical Mechanics and its Applications*, 387(4):1025–1032.

Xie, Y.-B., Zhou, T., and Wang, B.-H. (2008). Scale-free networks without growth. *Physica A: Statistical Mechanics and its Applications*, 387(7):1683–1688.

Yu, H., Braun, P., Yıldırım, M. A., Lemmens, I., Venkatesan, K., Sahalie, J., Hirozane-Kishikawa, T., Gebreab, F., Li, N., and Simonis, N. (2008). High-quality binary protein interaction map of the yeast interactome network. *Science*, 322(5898):104–110.

Yu, J.-R., Chiou, W.-J. P., Lee, W.-Y., and Lin, S.-J. (2020). Portfolio models with return forecasting and transaction costs. *International Review of Economics & Finance*, 66:118–130.

Zaki, M. J., Meira Jr, W., and Meira, W. (2014). *Data mining and analysis: fundamental concepts and algorithms.* Cambridge University Press.

Zhang, M. and Chen, Y. (2017). Weisfeiler-lehman neural machine for link prediction. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 575–583. ACM.

Zhang, Q.-M., Lü, L., Wang, W.-Q., and Zhou, T. (2013). Potential theory for directed networks. *PloS one*, 8(2):e55437.

Zhang, Q.-M., Xu, X.-K., Zhu, Y.-X., and Zhou, T. (2015). Measuring multiple evolution mechanisms of complex networks. *Scientific reports*, 5:10350.

Zheng, J.-F. and Gao, Z.-Y. (2008). A weighted network evolution with traffic flow. *Physica A: Statistical Mechanics and its Applications*, 387(24):6177–6182.

Zhou, S. and Mondragón, R. J. (2004). Accurately modeling the internet topology. *Physical Review E*, 70(6):066108.

Zhou, T., Lü, L., and Zhang, Y.-C. (2009). Predicting missing links via local information. *The European Physical Journal B*, 71(4):623–630.

Zhou, W. and Jia, Y. (2017). Predicting links based on knowledge dissemination in complex network. *Physica A: Statistical Mechanics and its Applications*, 471:561–568.

Zhuang, X.-t., Min, Z.-f., and Chen, S.-y. (2007). Characteristic analysis of complex network for shanghai stock market. *JOURNAL-NORTHEASTERN UNIVERSITY NATURAL SCIENCE*, 28(7):1053.

Appendix A

EXAMPLES OF LINK PREDICTION METHODS

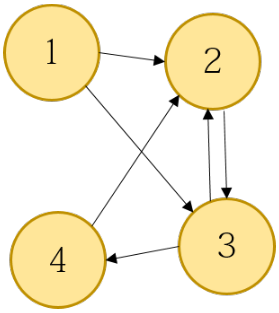


Figure A.1: **Network Example**

A.1 General Procedures of Causality Link Prediction

The adjacency matrix A in Figure A.1 above can be expressed as:

$$A = \begin{pmatrix} 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \\ 0 & 1 & 0 & 0 \end{pmatrix}$$

Using Eq.(4.12), we can obtain s^{CLP},

$$s^{CLP} = 0.001 \times \begin{pmatrix} 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \\ 0 & 1 & 0 & 0 \end{pmatrix} + \cdots + 0.001^n \times \begin{pmatrix} 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \\ 0 & 1 & 0 & 0 \end{pmatrix}^n$$

where n is 9 when this converges and the finally converged score matrix, s^{CLP}, can be obtained as follows.

$$s^{CLP} = \begin{pmatrix} 0 & 0.001 & 0.001 & 1.001 \times 10^{-6} \\ 0 & 1.001 \times 10^{-6} & 0.001 & 1.001 \times 10^{-6} \\ 0 & 0.001 & 1.001 \times 10^{-6} & 0.001 \\ 0 & 0.001 & 1.001 \times 10^{-6} & 1.001 \times 10^{-9} \end{pmatrix}$$

A.2 H-index Link Prediction

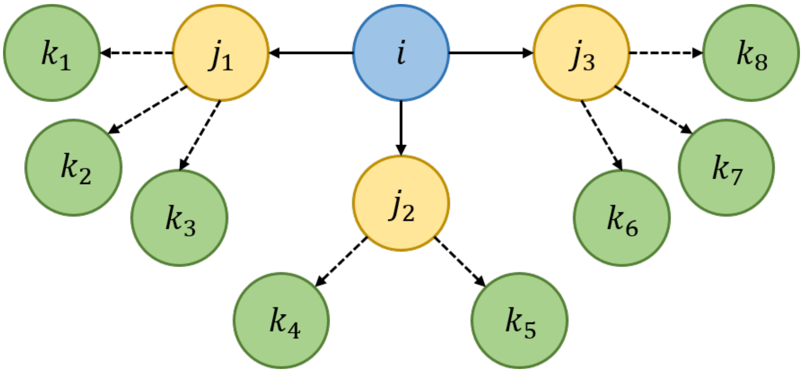


Figure A.2: **H-index example**

At first, a simple example of H-index is as follows. Since the directed edge is used, we can divide edges into inflow and outflow. The nodes pointed by node i are the outflow neighbors, and the nodes pointing to node i are the inflow neighbors. In Figure A.2, the outflow neighbors of node i are j_1, j_2 and j_3, which again have outflow neighbors of (k_1, k_2, k_3), (k_4, k_5) and (k_6, k_7, k_8), respectively. Let the neighbors with path length 2 from the node i be the second neighbors. Outflow neighbors with one or more second neighbors are j_1, j_2 and j_3, and with three or more second neighbors are j_1, j_3. In this case, the H-index of node i is 2. Now we can compute the H-index of all nodes in Figure A.1. For example, since node 1 points to nodes 2 and 3 without any node pointing to node 1, the outflow neighbor of node 1 is node 2 and node 3, whereas the inflow neighbor does not exist. In this regard, the neighbor matrix can be expressed for other nodes as follows.

$$Neighbor = \begin{pmatrix} [\,] & [2,3] \\ [1,3,4] & [3] \\ [1,2] & [2,4] \\ [3] & [2] \end{pmatrix}$$

Using Eq.(4.26) to obtain the H-index of each node, the H-index of node 1 is 1 because outflow neighbor nodes 2 and 3 of node 1 have node 3 and node [2, 4] as second outflow neighbors, respectively. Similarily, the H-index can be calculated for other nodes and for the inflows. Finally, the H-index matrix of this network is as follows.

$$H - index = \begin{bmatrix} 0 & 1 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \end{bmatrix}$$

Now we can find the revised adjacency matrix using Eq.(4.27).

$$A^H = \begin{pmatrix} 0.3333 & 0.3333 & 0.6667 & 0.3333 \\ 0.5 & 0.5 & 1 & 0.5 \\ 0.3333 & 0.3333 & 0.3333 & 0.6667 \\ 1 & 1 & 1 & 1 \end{pmatrix}$$

The final score, s^H using Eq.(4.12) is as follows.

$$s^H = \begin{pmatrix} 0.000334 & 0.000334 & 0.000668 & 0.000334 \\ 0.000501 & 0.000501 & 0.001002 & 0.000502 \\ 0.000334 & 0.000334 & 0.000335 & 0.000668 \\ 0.001002 & 0.001002 & 0.001003 & 0.001003 \end{pmatrix}$$

A.3 Generalized Degree Link Prediction

We consider both degrees and weights, the most basic properties of causality networks. The out-degree of each node in Figure A.1 is calculated by Eq.(4.32a) as follows.

$$A = \begin{pmatrix} 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \\ 0 & 1 & 0 & 0 \end{pmatrix} \rightarrow k_i = \begin{pmatrix} 2 \\ 1 \\ 2 \\ 1 \end{pmatrix}$$

Now we can calculate the strength of each node using Eq.(4.32b).

$$A^w = \begin{pmatrix} 0 & 8.82 & 13.04 & 0 \\ 0 & 0 & 25.32 & 0 \\ 0 & 9.59 & 0 & 14.56 \\ 0 & 8.46 & 0 & 0 \end{pmatrix} \rightarrow s_i = \begin{pmatrix} 21.86 \\ 25.32 \\ 24.15 \\ 8.46 \end{pmatrix}$$

Then, the generalized degree of each node with Eq.(4.32c) can be obtained as follows.

$$A_{1,2} = 2^{1-\alpha} \times 21.86^{\alpha} = 6.61211$$

$$G - degree = \begin{pmatrix} 0 & 6.61211 \\ 8.978307 & 5.031898 \\ 8.758995 & 6.94982 \\ 3.815757 & 2.908608 \end{pmatrix}$$

Now, we find the revised adjacency matrix, A^{GD}, using Eq.(4.27). For instance, the value from node 1 to node 2 can be calculated as follows.

$$A_{1,2} = \frac{GD_2 + 1}{\sum_{N_1}(GD_n + 1)} = \frac{5.031898 + 1}{5.031898 + 6.94982 + 2} = 0.431413$$

By computing the values for all the node pairs, we obtain the following adjacency matrix, A^{GD}, as follows.

$$A^{GD} = \begin{pmatrix} 0.071522 & 0.431413 & 0.568587 & 0.071522 \\ 0.125789 & 0.125789 & 1 & 0.125789 \\ 0.100598 & 0.6068 & 0.100598 & 0.3932 \\ 0.165785 & 1 & 0.165785 & 0.165785 \end{pmatrix}$$

A.4 Sum-eta Link Prediction

If each edge has a value of 0 or 1 based on the threshold of F-statistics and if we let the power of each node is the sum of the outflow η^2 values, then the weighted adjacency matrix A^w can be computed as follows.

$$A^w = \begin{pmatrix} 0 & 8.82 & 13.04 & 0 \\ 0 & 0 & 25.32 & 0 \\ 0 & 9.59 & 0 & 14.56 \\ 0 & 8.46 & 0 & 0 \end{pmatrix} \rightarrow \begin{pmatrix} 0 & 0.017266 & 0.025318 & 0 \\ 0 & 0 & 0.048016 & 0 \\ 0 & 0.018745 & 0 & 0.028186 \\ 0 & 0.016573 & 0 & 0 \end{pmatrix}$$

$$sum - \eta^2 = \begin{pmatrix} 0.0426 \\ 0.0480 \\ 0.0469 \\ 0.0166 \end{pmatrix}$$

Note that the given F-statistics are arbitrary values for an example. Similarly, using Eq.(4.27), the revised adjacency matrix, A^{SE}, is obtained. Then, the likelihood value from node 1 to node 2 is as follows.

$$A_{1,2} = \frac{f_2 + 1}{\sum_{N_1}(f_n + 1)} = \frac{0.0480 + 1}{0.0480 + 0.0469 + 2} = 0.5003$$

By computing the values for all the node pairs, we obtain the following adjacency matrix, A^{SE}.

$$A^{SE} = \begin{pmatrix} 0 & 0.5003 & 0.4997 & 0.4852 \\ 0.9958 & 0 & 1 & 0.9710 \\ 0.5050 & 0.5076 & 0 & 0.4924 \\ 0.9948 & 1 & 0.9990 & 0 \end{pmatrix}$$

A.5 Weighted Causality Prediction

In the weighted adjacency matrix, A^w, obtained from Appendix A.4, the ratio of all element values to the largest η^2 value is calculated. Then, we can obtain a new adjacency matrix, A^{WC}.

$$
\begin{aligned}
A^{WC} &= \begin{pmatrix} 0 & 8.82 & 13.04 & 0 \\ 0 & 0 & 25.32 & 0 \\ 0 & 9.59 & 0 & 14.56 \\ 0 & 8.46 & 0 & 0 \end{pmatrix} \rightarrow \begin{pmatrix} 0 & 0.0173 & 0.0253 & 0 \\ 0 & 0 & 0.0480 & 0 \\ 0 & 0.0187 & 0 & 0.0282 \\ 0 & 0.0166 & 0 & 0 \end{pmatrix} \times \frac{1}{0.048016} \\
&= \begin{pmatrix} 0 & 0.3596 & 0.5273 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0.3904 & 0 & 0.5870 \\ 0 & 0.3452 & 0 & 0 \end{pmatrix}
\end{aligned}
$$

국문초록

금융 시장에서 발생하는 위험은 하나의 금융 체계(System)에서 연쇄 작용으로 이어지며 이것은 곧 시스템의 붕괴로 이어진다. 세계 경제에 큰 타격을 주었던 미국의 서브프라임 모기지 사태 이후 위기 대처 능력 제고를 위해 금융 체계를 올바르게 이해하고 분석하는 것이 매우 중요한 과제로 떠올랐다. 전통적인 금융 위험 관리 이론으로 설명되지 않는 정형화된 사실(stylized facts)들의 발견으로 새롭게 등장한 연구 분야가 경제물리학(Econophysics)이다. 특히, 점(노드)과 선(링크)으로 하나의 체계를 나타내는 복잡계 네트워크 모형은 분야를 막론하고 다양하게 응용되고 있다. 하지만 금융 시장에 대한 기존의 복잡계 네트워크 모형은 대부분 과거 데이터를 기반으로 네트워크 구조 변화, 위험의 확산 경로와 같은 실증적 연구결과를 확인하는 데 그쳐 위험에 대비한 능동적인 대안을 제시하는데 제약이 존재한다. 본 학위논문은 이러한 결점을 보완하고자 네트워크 구성 요소 간 관계가 명확하게 드러나는 환율 데이터 기반의 네트워크 링크 예측 모형을 제시하였다. 먼저, 네트워크 구조 예측의 타당성을 확보하기 위해 네트워크가 시장을 성공적으로 모방하는지 확인하였다. 그 결과 실질실효환율 데이터는 두꺼운 꼬리(Fat-tailed) 분포를 가지며 꼬리 분포가 멱함수(Power-law) 분포를 따르는 것을 확인하였다. 또한, 미국의 서브프라임 모기지 사태, 유럽 부채 위기, 중국 주식 시장 위기 동안 네트워크의 단면(cross-sectional) 토폴로지와 시간에 따라 변화하는 성질을 관찰하였다. 위기 발생 대륙에서 증가하는 링크의 수량을 봤을 때 제시된 그레인저-인과관계(Granger causality) 네트워크가 시장을 적절히 나타내고 있었다. 두 번째로, 네트워크에서 새롭게 생겨날 수 있는 링크를 예측하기 위해 구성 요소 간 유사도를 측정하는 Weighted Causality Link Prediction (WCLP) 모형을 제시하였다.

기존의 많은 네트워크 모형이 구성 요소 간 상관관계에 기반하였다면, 본 모형은 그레인저 인과관계를 측정하여 네트워크의 방향성을 함께 고려하고, 연결 강도를 통계량에 기반한 효과 크기(Effect size)로 나타내었다는 점에서 그 차별성이 있다. 네트워크의 링크는 서로 다른 연결 강도를 가지며 효과 크기가 클 수록 오래 유지된다는 가설 하에 실험을 진행하였다. 그 결과, 높은 수신자 조작 특성 곡선의 면적 (Area Under the receiver operating characteristic Curve, AUC) 값을 가져 비가중치(Unweighted) 또는 상관관계 기반 유클리드 거리(Euclidean distance)를 가중치를 이용한 기존 모형들에 비해 통계적으로 개선된 예측 성능을 보였다. 마지막으로, 네트워크 링크 예측 결과를 기반으로 미국 금융 시장에서의 투자 의사 결정 모형을 제시하였다. PMFG의 주변부에 위치하는 종목으로 포트폴리오가 구성되면, 자산 간의 낮은 상관관계는 포트폴리오 위험의 분산화를 가능하게 한다. 하지만 상관관계는 두 변수 간 연관된 정도만을 나타내므로 시차를 두고 나타나는 인과관계를 나타내지 못한다는 단점이 있다. 따라서 본 학위논문에서는 기존의 Partial Correlation Planar Graph (PCPG) 모형에서 개선 된 새로운 그래프를 제시하고, Weighted Causality Planar Graph (WCPG)라고 명명한다. WCPG는 링크 예측을 통해 얻은 자산 간 유사도를 이용하여 만들어지며 방향성과 세기가 함께 고려된다는 점에서 기존 모형과 차별성이 있다. 그 결과, 위험 조정 수익률 측면에서 제시된 모형이 기존의 네트워크 모형 대비 개선된 성능을 보이며 특히 6개월 이상의 장기 투자에서 강점을 가졌다. 결론적으로 본 학위논문은 효과적인 링크 예측 모형을 효과 크기와 결부하여 개선된 모형을 제시하고 투자 의사 결정을 위한 모형에 응용하였다는 점에서 그 의의를 찾을 수 있다.

주요어: 실질실효환율, 인과관계 네트워크, 정형화된 사실, 효과 크기, 방향성 있는 네트워크, 링크 예측, 포트폴리오 최적화, 평면 그래프, 가중치 있는 인과관계 평면 그래프

학번: 2014-21813