Tampere University

Joonas Karaila

MONEY FLOWS BETWEEN SECURITIES: NETWORK ANALYSIS IN A STOCK MARKET

Master of Science Thesis
Faculty of Management and Business
Examiners: Professor Juho Kanniainen
University Lecturer Henri Hansen
November 2021

ABSTRACT

Joonas Karaila: Money Flows Between Securities: Network Analysis in a Stock Market
Master of Science Thesis
Tampere University
Master's Degree Programme in Industrial Engineering and Management
November 2021

Application of network science to study various phenomena has increased in the recent years as for example networks have been used to model the effects of age to the spread of COVID-19, how the World Trade Web changed during the financial crisis of 2008 and to rank web pages. Networks have been used to study financial markets although research has focused mainly on interbank lending. The financial crisis of 2008 has been shown to have happened partly because of the financial industry, and the interbank lending networks showed early-warning signals of the crisis. In 2008 stock markets crashed and investors changed their allocations to different assets based on their views of the future. Therefore, money flowed between securities which can be modelled using networks, and network science offers tools to study the topological features.

The goals of the thesis were to define money flow networks and to study possible changes in the networks during the financial crisis of 2008 in Finland. Money flows to securities have been used before as a technical indicator but the sources of the flows have not been incorporated to the analyses. Thus, a method for approximating money flows between securities from transaction data is defined which forms a network of money flows between securities. To compare the money flows, the absolute money flows are scaled using the largest money flow during the last 90 days between two securities.

The networks of financial institutions and households are analyzed by ranking the securities based on their centralities and by analyzing changes in the z-scores of subgraph abundancies. The ranking of securities based on their centralities is used to find out if some securities are favoured or neglected during the crisis, and different centrality measures are used with statistical testing to ensure the results. The z-scores of subgraph abundancies are used to find general changes in the structure of the networks during the crisis. Since the subgraph counts are affected by the number of links and degrees of nodes, two different random graph models are used in calculating the deviations from the expected subgraph counts.

The centrality rankings showed that large companies are more often in the top percentiles of the rankings while the bottom percentiles mostly do not have certain securities in them more often than expected. Finance institutes had more random ranking than households as in the daily networks households did not invest in smaller companies as often. The z-scores of subgraph abundancies had multiple observations and interpretations, but further analyses are needed to understand the changes better. Household networks had more complex structure than finance institute networks which may mean households had less similar opinions about the securities. Based on the analyses, the centrality rankings failed in finding securities with unusual money flows, even though differences between investing of finance institutes and households were found, and there were changes in the structure of the networks during the crisis. In the end, money flow networks should be studied further, and future analyses should incorporate additional market data.

Keywords: quantitative finance, complex networks, financial crisis, structural analysis

The originality of this thesis has been checked using the Turnitin OriginalityCheck service.

TIIVISTELMÄ

Joonas Karaila: Rahavirrat osakkeiden välillä: verkostoanalyysi osakemarkkinoilla
Diplomityö
Tampereen yliopisto
Tuotantotalouden diplomi-insinöörin tutkinto-ohjelma
Marraskuu 2021

Verkostotieteen hyödyntäminen eri ilmiöiden tutkimisessa on yleistynyt viime vuosien aikana. Verkostoja on käytetty esimerkiksi tutkittaessa miten ikä vaikuttaa COVID-19 leviämiseen, miten maailmankauppaverkosto muuttui vuoden 2008 finanssikriisin aikana ja järjestämään verkkosivuja. Verkostoja on käytetty finanssimarkkinoiden tutkimiseen, mutta tutkimus on keskittynyt pääasiassa pankkien välisiin lainaverkostoihin. On osoitettu, että vuoden 2008 finanssikriisi oli osittain finanssialan syytä, ja pankkien väliset lainaverkostot sisälsivät vaaran merkkejä jo ennen kriisiä. Vuonna 2008 osakemarkkinat romahtivat, minkä seurauksena sijoittajat muuttivat heidän allokaatioitaan vastaamaan heidän tulevaisuuden näkymiään. Tämän seurauksena raha liikkui arvopapereiden välillä, mitä on mahdollista mallintaa verkostoilla, ja verkostotiede mahdollistaa topologisten ominaisuuksien tutkimisen.

Tutkimuksen tavoitteina oli määritellä rahavirtaverkostot ja analysoida niiden mahdollisia muutoksia Suomessa finanssikriisin aikana. Rahavirtoja on käytetty aiemmin teknisenä indikaattorina, mutta rahavirtojen lähteitä ei ole hyödynnetty analyyseissä. Täten työssä määritellään metodi arvopaperien välisten rahavirtojen approksimointiin, mikä mahdollistaa rahavirtaverkostojen luomisen. Jotta rahavirtoja voi vertailla arvopapereiden välillä, absoluuttiset rahavirrat skaalataan viimeisimpien 90 päivän suurimmalla rahavirralla kyseisten arvopapereiden välillä.

Finanssi-instituutioiden ja kotitalouksien verkostoja analysoidaan järjestämällä arvopaperit niiden keskeisyyden mukaan ja tarkastelemalla aligraafien määrien z-arvojen vaihteluita. Arvopaperien keskeisyyksien mukaisten järjestyksien avulla on mahdollista tunnistaa ovatko jotkin arvopaperit suosittuja tai hyljeksittyjä kriisin aikana. Järjestyksien tulokset vahvistetaan käyttämällä useita eri keskeisyysmittoja ja merkitsevyystasoja. Aligraafien määrien z-arvojen avulla tarkastellaan verkostojen yleisiä rakenteellisia muutoksia kriisin aikana. Aligraafien määriin vaikuttaa linkkien määrä verkoissa ja solmujen asteluvut, minkä vuoksi poikkeavuudet odotusarvoista lasketaan kahden eri satunnaisgraafimallin avulla.

Keskeisyyksien mukaiset järjestykset osoittivat suurempien yrityksen olevan useammin järjestyksien huipulla, ja järjestyksien häntäpäissä ei ole tiettyjä arvopapereita suurimmassa osassa analysoituja verkkosarjoja useammin kuin odotettua. Finanssi-instituuttien järjestykset olivat enemmän satunnaisia kuin kotitalouksien, koska kotitaloudet eivät sijoittaneet pieniin yrityksiin yhtä useasti. Aligraafien z-arvoista voi tehdä useita erilaisia päätelmiä, mutta lisätutkimuksia tarvitaan muutoksien parempaan ymmärtämiseen. Kotitalouksien verkot olivat rakenteeltaan monimutkaisempia kuin finanssi-instituuttien, mikä voi tarkoittaa suurempia erimielisyyksiä arvopapereista. Analyysien perusteella keskeisyysmitat eivät onnistuneet tunnistamaan arvopapereita, joilla oli epätavallisia rahavirtoja, mutta rahavirtaverkostojen rakenteissa oli muutoksia kriisin aikana. Finanssi-instituutioiden ja kotitalouksien verkkojen välillä pystyttiin tunnistamaan eroavaisuuksia. Rahavirtaverkostoja pitää tutkia lisää ja jatkotutkimuksien pitää hyödyntää muuta markkinadataa.

Avainsanat: matemaattinen rahoitus, kompleksiset verkostot, finanssikriisi, rakenneanalyysi

Tämän julkaisun alkuperäisyys on tarkastettu Turnitin OriginalityCheck -ohjelmalla.

PREFACE

All the hard work paid off in the end and this thesis marks the end of my studies at least for now. The process of writing this thesis has been long, but it has been interesting and rewarding. I've learned a lot about multiple different mathematical fields, how imagination is needed in addition to mathematical knowledge to form different hypotheses, and how hard work conducting research is. The subject was challenging, and it tested what I learned during my studies. In my studies I tried to not take the easy way out but to challenge myself and reach higher highs in knowledge. I'm happy about the outcome of the thesis as I wanted it to show mastery in what I've learned. Facing challenges head-on did not go to waste. My respect for researchers has increased as behind new ideas there is surprisingly large amount of work and some of the work behind the scenes is never seen.

I'd like to thank Juho Kanniainen and Kęstutis Baltakys for all their help and advice during the thesis. I'm thankful for you trusting me with this subject. Thanks should also go to Henri Hansen for reading the thesis and giving feedback on it. I also want to thank my family for their support and understanding during the writing process. Lastly, I'd like to thank *SCPM* for supporting each other through the university years and making sure we have fun in addition to all the hard work.

Too many times I have said to my friends and family that the thesis would be finished next month, but the next month is finally here and the thesis is finished. Let curiosity lead the way and may the adventures continue.

Tampere, 18th November 2021

Joonas Karaila

CONTENTS

1. INTRODUCTION

Network science is an interdisciplinary field which studies social (Wasserman 1994), biological (Barabási, Gulbahce et al. 2011) and information (Maslov, Sneppen and Zaliznyak 2004) networks for example. The networks contain nodes which may be people, genes or computers, and edges which connect the nodes together. In social networks an edge may mean a friendship between two persons, who calls whom or how often the persons call each other. The first edge is undirected as friendships go both ways, the second is directed because the edge is directed from the caller to the receiver and the third is weighted by the number of calls. It is possible to have networks which are dependant on time (Holme and Saramäki 2012) and multiple layers of networks in which layers contain different information (Kivela et al. 2014) among other network models which makes it possible to model various phenomena with networks. For example, Hâncean et al. (2021) used network science to study how age affected the spread of COVID-19 in Bucharest and they found out that the transmissions happen more likely within an age group, the infection risk increases as the age of the contacts decreases and adults between the ages 35 to 44 are critical in the transmission of the virus between age groups.

The study of networks may focus on for example finding the most important nodes, shortest paths, communities or finding how assortative the nodes are. Assortativity quantifies how similar nodes connect to similar nodes or dissimilar nodes (Noldus and Van Mieghem 2015). Network properties are also studied with different random network models to understand how different properties affect other properties of the networks and to compare the properties of real-world networks to random models (Newman 2018, pp. 342-343). Real-world networks can have properties such as the power-law degree distribution (Clauset et al. 2009) or the small-world property (Watts and Strogatz 1998). Networks with a power-law degree distribution means that the networks are dominated by relatively small number of nodes that are connect to many others. The networks are called scale-free as the networks contain hubs, i.e. nodes, with large number of links and there is not a typical node in the network. (Barabási and Bonabeau 2003) Small-world property means that nodes in a network are not neighbours of most other nodes but nodes can be reached from other nodes with relatively small number of steps, and the clustering in the network is high (Watts and Strogatz 1998). These properties of networks were found by using Erdös–Rényi model, in which a graph with n nodes and M edges is chosen

uniformly at random (Erdös and Rényi 1959), and by comparing the properties of the real-world networks to the mathematically defined graphs (Barabási 2013).

In this thesis network science is used to study financial markets of Finland during the financial crisis of 2008. A new way of studying the markets is formed by creating a network from transactions of households and finance institutes which forms a network of money flows. The used data gives a unique possibility to use the real transaction of finance institutes and households which makes it possible to approximate how money generated from a sale of security is used to buy another security. The goal of the thesis is to study changes in the networks during the crisis and form a basis for further research. The focal point of the thesis is on changes in the rankings of securities based on their centralities and the subgraph counts compared to the expected counts of two different random graph models. Centrality measures are used to identify the most or the least important securities which may help with identifying unusual activity, and the subgraph counts may help with identifying general changes in how investments are made. A hypothesis for the centrality measures is that the relative importances of the securities change from time to time, and a hypothesis for the subgraphs is that their abundancies change relative to the null models. There could be changes in relative importances of the securities in the networks because the prospects of the securities are affected differently by the crisis. The subgraph counts relative to the null models could change because of higher activity in the markets during the crisis or more similar actions by investors for example. It would be surprising if there were no changes in the subgraph counts relative to the null models as there would not be relatively more good and bad investments in the market during the crisis. If there are more good and bad investments in the market, there should be more money flows from bad investments to good investments. The thesis does not relate the generated network data to other market data as the goal is to define the markets in a new way, study basic properties of the networks and possibly find changes in the networks during the crisis. Therefore, the main goal is to understand money flow networks better and do exploratory research.

Networks have been used in studying financial markets and economics before by studying for example Dutch interbank lending before and during the 2007–2008 crisis (Squartini, Van Lelyveld et al. 2013), bipartite World Trade Web before and during the 2007–2008 crisis (Saracco et al. 2016), equilibrium financial flows and prices with three different types of agents in the economy (Nagurney and K. Ke 2001), how buyer and seller agents form networks and what the structure of those networks are (Kranton and Minehart 2001), and how risks are shared across communities (Bramoulle and Kranton 2007). Therefore, multiple different financial networks have been studied in which the nodes have been institutions, countries, individuals, securities or communities and the links are dependencies between the nodes which may be transactions, loans or exposure to the same entity. A lot of the financial network research before 2008 focused on interbank markets but research

has spread to other phenomena in finance and there are multiple different research directions (Allen and Babus 2008). Money flows have been used before in networks as links (Fujiwara et al. 2021; Nagurney and K. Ke 2001; Saracco et al. 2016) but the nodes have not been securities. The nodes were firms' bank accounts for Fujiwara et al. (2021), sources of funds (households and firms), intermediaries (banks and investment companies) and consumers (household and business loans) for Nagurney and K. Ke (2001), and countries and products for Saracco et al. (2016). The reason for choosing securities as nodes is to observe how money flows between securities which may mean preference of some security over the other securities. Money flows have been used as a technical indicator for returns before (Brown and Brooke 1993) but the source of the money and the network structure have not been studied to the best of knowledge at the time of writing this thesis. Money flows, which were defined to be the difference between uptick and downtick dollar trading volume, have been shown to correlate with stock returns (Bennett and Sias 2001).

The financial crisis of 2008 is used as the time of study to maximize the possibility of finding changes in the networks and to possibly find early signals of the crisis which could be used for forecasting a possible crisis. The reasons which contributed to the financial crisis of 2008 have been studied to be the bursting of the United States housing bubble, low interest rates, excessive risk-taking by financial institutions, subprime loans (Holt 2009), predatory lending, mortgage-backed securities (Fligstein and Roehrkasse 2016) and complex over-the-counter derivatives (Bullard et al. 2009). Squartini, Van Lelyveld et al. (2013) found that the structure of the credit network of Dutch banks played a role in the crisis which shows that the whole credit network had higher systematic risk before the collapse. The crisis was global as the effects spread from United States to other countries with varying outcomes (McKibbin and Stoeckel 2010). In Finland the effects on real economy were substantial as the production was heavily focused on capital-intensive goods but the effects on private consumption and unemployment were smaller than expected as compared to the large negative change in gross domestic product (Valtioneuvoston kanslia 2011, p. 7). House prices in Finland only decreased by 5% at the end of 2008 but the effects on the stock market were still large in 2008-2009 (Valtioneuvoston kanslia 2011, p. 24).

Additionally, the money flow networks are divided to two different groups of investors: financial institutions and households. The goal is to find out if the groups invest in different ways that are noticeable in the networks. In general, humans have limited cognitive resources (Kahneman 1973, pp. 7-11) and are prone to use heuristics in more complex tasks which can lead to systematic errors (Tversky and Kahneman 1974). These heuristics and biases include anchoring, which means adjusting an initial value to get the final answer (Tversky and Kahneman 1974), confirmation bias, which means looking for information which agrees with the existing hypothesis (Nickerson 1998), and probability

neglect, which means neglecting probabilities when making a decision (Rottenstreich and Hsee 2001; Sunstein 2002) for example. In finance these biases play a role, an investor may value the security based on past prices, look for information which agrees with the investment hypothesis or use a single point estimate for an investment instead of variety of outcomes. Individual investors have been shown for example to trade too much (Barber and Odean 2000), maintain undiversified portfolios (Horne et al. 1975), hold losing positions for too long (Ferris et al. 1988), overinvest in their own companies' stock (Huberman 2001), make irrational index fund choices (Elton et al. 2004) and avoid risky investment if they, their neighbours or their relatives have had adverse effects of a depression before (Knüpfer et al. 2017). Whereas for institutional investors, hedge funds correct mispricing (Cao et al. 2018), market makers provide liquidity to the market (Grossman and Miller 1988), mutual funds have been shown to have superior stock-picking skill compared to mechanical strategies, which earns an excess return of approximately the size of the management fee, (Daniel et al. 1997), institutional investors that trade to maximize short term profit exploit the post-earnings announcement drift (B. Ke and Ramalingegowda 2005) and high-frequency traders trade in the direction of permanent price changes (Brogaard et al. 2014). Overall, institutional and retail investors are subject to biases but institutional investors are seen more often as more rational, even though they may act irrationally as well (Zeng 2016), and they deploy more complex strategies. In the networks these effects may be seen for example as institutions having more flows between securities.

The remainder of this thesis is organized as follows. Chapter 2 defines the money flow networks used in this thesis. Chapter 3 introduces the methods which are used to analyze the networks. Chapter 4 describes the basic properties of the data and the networks. Chapter 5 presents the found results when the methods are applied to the networks. Chapter 6 discusses the results and Chapter 7 concludes the thesis.

2. DEFINING MONEY FLOW NETWORKS

In this chapter money flow networks are defined. First, money flows are approximated between securities for single investors. Second, multiple simple investor networks are aggregated together, and the net money flows are calculated between securities. Hence, a network representation of the transactions during an observation period is created which can be studied further using network analysis methods. Before the analysis, in this thesis the edge weights are scaled using maximum value of the edge during a period so that the edge weights between different securities can be compared to each other.

2.1 Network definition

In this thesis the money flows between securities are defined by having money flows go from all sold securities to all bought securities for a single investor. The amount of money that flows between the securities is proportional to the amount of the security sold and the amount of the security bought. In the case in which an investor has bought or sold more securities in a euro volume during a time period, the difference between the two euro volumes is calculated and added as an additional security, which is called *balance* in this thesis. The euro volume difference between the bought and sold securities is calculated as $-\sum_{i=1}^{N^{(P)}(t)} v_i$, where $v_i(t)$ is net money flow for stock i and $N^{(P)}(t)$ is the number of securities the investor bought and sold during the time period. The euro volume difference is then added as an euro volume for the balance node, and thus the node can be handled like any other security. The net money flow for a stock is positive if the stock has been sold and negative if the stock has been bought. In this way, an investor always has a balanced network where the amount of value stays the same, but money flows between different securities and the balance node. To keep the money flows intact, a percentage of money flow from every sold security is directed to every bought security. This is done by dividing the product of the sold and bought security euro volumes by the total money flow

$$M^{(P)}(t) = \sum_{k=1}^{N^{(P)}(t)} (f_k(t) - 1)v_k(t),$$

where

$$f_i(t) = \begin{cases} 1, & \text{if } v_i > 0 \text{ (sold)} \\ 0, & \text{if } v_i \leq 0 \text{ (bought).} \end{cases}$$

The total money flow could also be calculated by having $f_k(t)$ in the definition instead of $f_k(t) - 1$ because the total euro volume of sold and bought securities are equal. Finally, edges for an investor money flow network are defined as

$$e_{i,j}^{(P)}(t) = f_i(t)v_i(t)\frac{(f_j(t) - 1)v_j(t)}{M^{(P)}(t)}, \tag{2.1}$$

where $e_{i,j}^{(P)}(t)$ is a positive money flow from stock i to j. Investor P's adjacency matrix $\boldsymbol{E}_s^{(P)}(t)$ is constructed from edges $e_{i,j}^{(P)}(t)$ by placing the edges to coordinates (i, j).

To illustrate Eq. 2.1 let stock 1 be bought by $\$100$, stock 2 bought by $\$200$, stock 3 sold by $\$50$ and stock 4 sold by $\$250$ in total by an investor during some time period. Now the net euro money flows for the stocks would be $-\$100$, $-\$200$, $\$50$ and $\$250$. Using Eq. 2.1 to calculate edge weight from stock 3 to 1 yields

$$\begin{aligned} e_{3,1} &= \frac{1 \cdot 50}{0 \cdot (-100) + 0 \cdot (-200) + 1 \cdot (50) + 1 \cdot (250)} \cdot (0 - 1) \cdot (-100) \\ &= \frac{50 \cdot 100}{300} \\ &= \frac{50}{3}. \end{aligned}$$

Furthermore, in Table 2.1 all combinations of edges without self-loops are calculated. The investor's adjacency matrix is then

$$\boldsymbol{E}_s^{(P)}(t) = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 50/3 & 100/3 & 0 & 0 \\ 250/3 & 500/3 & 0 & 0 \end{bmatrix},$$

which can be used for different analyses and visualizations.

Table 2.1. *Edges of an investor network.*

Stock i	Stock j	$e_{i,j}$
1	2	0
1	3	0
1	4	0
2	1	0
2	3	0
2	4	0
3	1	$50/3$
3	2	$100/3$
3	4	0
4	1	$250/3$
4	2	$500/3$
4	3	0

Figure 2.1 shows a visualization of the example network.

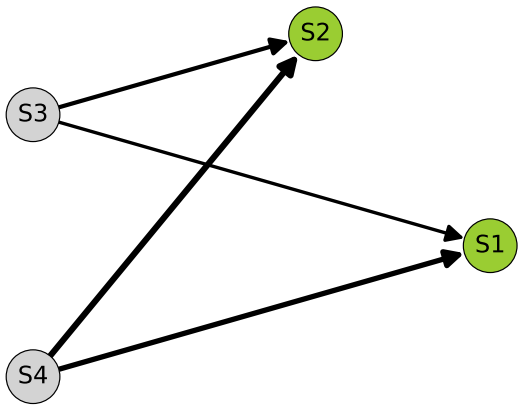


Figure 2.1. *Investor network for 4 stocks, where the money flow from stock 4 to 2 is the largest.*

In Figure 2.1 there are no nodes for which there are both incoming and outgoing edges. This is a result from Eq. 2.1 as if $v_i > 0$ and $v_j \leq 0$, then there can only be outgoing edges,

$$e_{i,j}^{(P)}(t) = \frac{-v_i(t)v_j(t)}{M^{(P)}(t)},$$

and

$$e_{j,i}^{(P)}(t) = 0$$

because $f_j = 0$ in Eq. 2.1. Other combinations of $v_i(t)$ and $v_j(t)$ can be shown to result in only outgoing or incoming edges for a node. However, aggregating multiple investor networks together can have nodes for which there are both incoming and outgoing edges.

In addition, Eq. 2.1 has important properties when $v_i^{(P)} > 0$ and $v_j^{(P)} \leq 0$,

$$\sum_{i=1}^{N^{(P)}(t)} e_{i,j}^{(P)}(t) = (f_j(t) - 1)v_j^{(P)}(t)\frac{\sum_{i=1}^{N^{(P)}(t)} f_i(t)v_i^{(P)}(t)}{M^{(P)}(t)} = -v_j^{(P)}(t)$$

which is the money flow to stock j, and

$$\sum_{j=1}^{N^{(P)}(t)} e_{i,j}^{(P)}(t) = f_i(t)v_i^{(P)}(t)\frac{\sum_{j=1}^{N^{(P)}(t)}(f_j(t) - 1)v_j^{(P)}(t)}{M^{(P)}(t)} = v_i^{(P)}(t)$$

which is the total money flow from stock i. These two properties show that Eq. 2.1 keeps the money flows intact as there is no net inflow or outflow of money to a single security other than the amount the investor has bought or sold. In the case the money flows would be randomized between securities the amount the investor has bought or sold a security would not be reflected in the money flows between securities. Randomizing could be done by taking the money flow from A to B and using it for the money flow from B to C for example. When the money flows reflect the amount an investor has bought or sold a security it is possible to study for example if the money flows are random in size or not.

Equation 2.1 also weights money flows based on the source and target security net euro money flows, which can be seen as good approximation of the real money flows. If an investor has sold a large amount of security A and bought a large amount of security B then it can be argued that the investor likely used more money generated from the sale of A than from a small sale of security C. Equation 2.1 only approximates the real underlying thought process of an investor as an investor could think of trades as pairwise exchanges while the equation assumes that all sales contribute to all buys.

Multiple investor networks with $I(t)$ investors are aggregated together by

$$\boldsymbol{E}(t) = \sum_{s=1}^{I(t)} \boldsymbol{E}^{(s)}(t). \tag{2.2}$$

The net money flow networks are calculated by using

$$\overline{e}_{i,j}(t) = e_{i,j}(t) - e_{j,i}(t), \tag{2.3}$$

and then changing all $\overline{e}_{i,j}(t) \leq 0$ into 0. The same operation can be calculated in matrix form by $\boldsymbol{E}_n(t) = \boldsymbol{E}(t) - \boldsymbol{E}^T(t)$ and changing values less than 0 to 0 again. All negative edges are removed as otherwise there would be double counting in the net money flows from stock i to j. For example, assume investor A had a money flow of $\$50$ from stock 1 to 2 and investor B had a money flow of $\$100$ from stock 2 to 1. Over both investors there was a net money flow of $\$50$ from stock 2 to 1 but by using Eq. 2.3 without removing negative weight the result is a money flow of $-\$50$ from 1 to 2 and $\$50$ from 2 to 1, which is incorrect as in total there would be no flow of money. Thus, removing negative weights is mandatory to have a network containing positive money flows between securities.

Moreover, in the aggregated networks it is possible to calculate the net inflow or outflow of money during a time period as

$$C_i(t) = k_i^{in} - k_i^{out},$$

where k_i^{in} is the in-degree of a node and k_i^{out} is the out-degree of a node. If $C_i(t) > 0$ then there was a net inflow of money to the stock and if $C_i(t)$ is negative, then there was a net outflow of money from the stock.

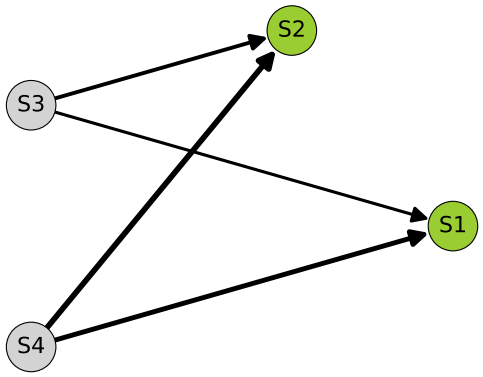
An example of aggregating investor networks together can be done using the example network from before and a new one. Let

$$\boldsymbol{E}^{(A)} = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 50/3 & 100/3 & 0 & 0 & 0 \\ 250/3 & 500/3 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix} \text{ and } \boldsymbol{E}^{(B)} = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 200 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 100 & 0 & 0 \\ 0 & 0 & 50 & 0 & 0 \end{bmatrix}.$$

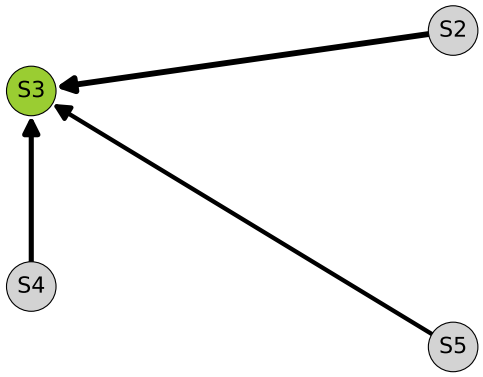
Notice that the network for A has the same money flows as before but now it has the fifth stock in its adjacency matrix also so that the matrices are the same size. The aggregation is done using Equations 2.2 and 2.3 which results in

$$\boldsymbol{E} = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 500/3 & 0 & 0 \\ 50/3 & 0 & 0 & 0 & 0 \\ 250/3 & 500/3 & 100 & 0 & 0 \\ 0 & 0 & 50 & 0 & 0 \end{bmatrix}.$$

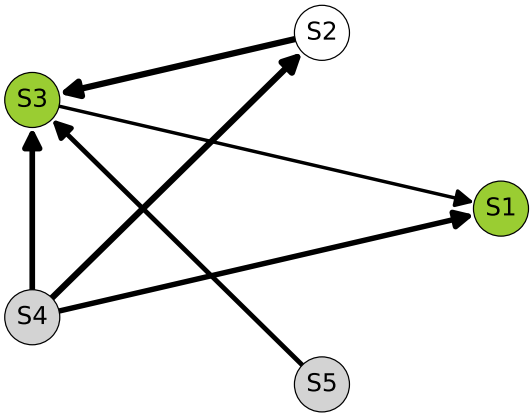
Visualization of the example aggregated network is shown in Figure 2.2.



(a) *Investor A money flow network.*



(b) *Investor B money flow network.*



(c) *Aggregated money flow network.*

Figure 2.2. *Investor A and B money flow networks with four mutual stocks and one nonmutual stock. In the figure green nodes have net positive money flows, grey nodes have net negative money flows and white nodes have 0 net money flows. The size of the arrows illustrate the size of the money flow.*

In Figure 2.2 investor A and B inflow and outflow of money from stock 2 cancel each other out which leads to no net money flows for stock 2. Money did still flow through stock 2 from stock 4 to stock 3, which is important when studying the system as a whole later.

2.2 Edge weight scaling

The idea behind edge weight scaling is to have all edges comparable to each other no matter what the target and source nodes are. Different securities are different in for example their market capitalization, volatility, and future prospects, which leads to different sizes of money flows to and from the securities. Scaling the edge weights with some function makes the relative sizes of the money flows comparable to each other, and thus it is possible to identify relatively exceptional money flows. These relatively exceptional money flows can carry additional information about the relative performance of the securities for example.

Scaling can be done using many different methods, but often a rolling time window is used in the process. Rolling window means that a function is always applied to a series of k values $\{v_{i-k+1}, v_{i-k+2}, \ldots, v_i\}$, which is then moved to time step $t+1$. The function in use is then applied to all k-size series in the data.

The k-size time series of an edge $\bar{e}_{i,j}(t)$ is

$$\bar{\boldsymbol{e}}_{i,j}(t) = \{\bar{e}_{i,j}(t-k+1), \bar{e}_{i,j}(t-k+2), \ldots, \bar{e}_{i,j}(t)\}, \tag{2.4}$$

where k is the number of days used in the sliding window. In Figure 2.3 an example time series of an edge is shown. There is a cutoff in the time series as the first $k-1$ values are taken away from the series because there are not enough prior values to use for the scaling.

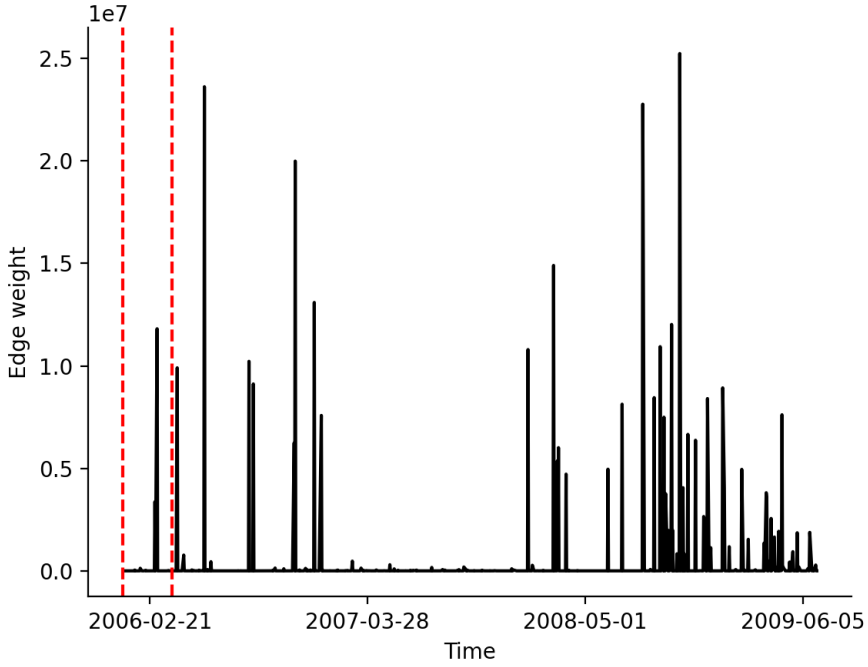


Figure 2.3. *Balance to Seligson & Co OMX Helsinki* 25 *edge value time series for finance institutes. The first* 90 *day period is shown with red lines. The series between the red lines is cut off as it is used for the first scaled value at the second red line.*

In Figure 2.3 there is no money flow between the nodes each day. This means that there was not a single investor in the data set that had both a negative balance and bought Seligson & Co OMX Helsinki 25 during that day. Figure 2.4 shows the histogram for edge weights of Figure 2.3. Using a histogram helps with understanding how the edge weights are distributed, and it helps with determining which scaling methods are useful for the time series. Figure 2.4 shows that most of the weights are close to or equal to 0 and there are weights, which have very large values compared to rest of the values.

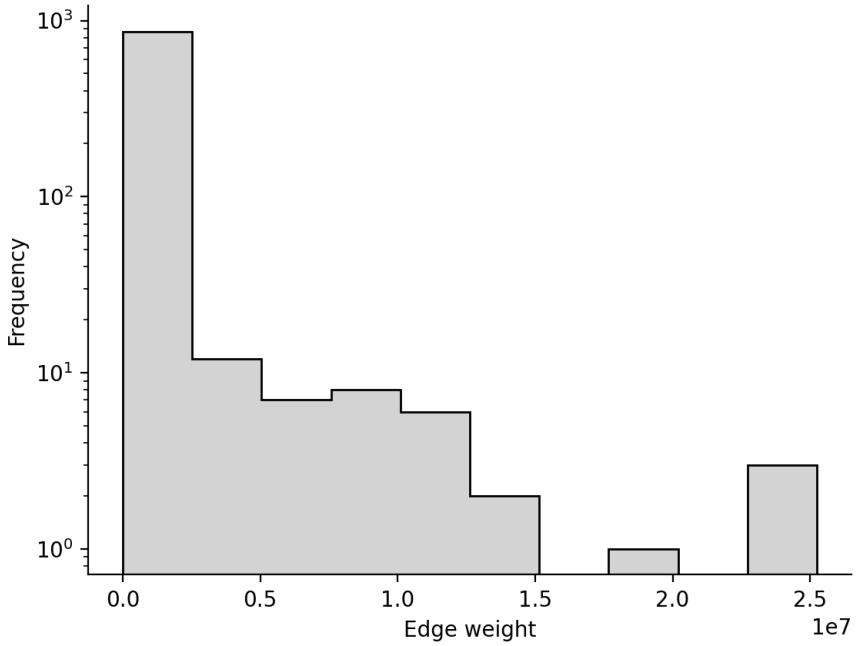


Figure 2.4. *Histogram of balance to Seligson & Co OMX Helsinki* 25 *edge weights of financial institutions. It can be seen that most of the weights are close to or equal to* 0 *which means that there are rarely money flows.*

The large number of zeros in the edge weight time series means that the data is sparse, which makes finding the right scaling method harder. Sparseness of the data means that centering all of the data using naive techniques would destroy the sparseness, and hence have a large effect on the results. In data preprocessing centering means subtracting a constant from the data, for example the mean or the median of the data. In machine learning feature scaling is a common method as it can enhance the performance of the models. (Scikit-Learn 2021)

Choosing the right scaling method is a difficult task as the interpretability of the weights weakens with more complex methods that would otherwise do well. For example, Yeo-Johnson transformation (Yeo 2000) maps the data as closely to Gaussian distribution as possible while stabilizing variance and minimizing skewness. The problem with the Yeo-Johnson transformation with this edge weight time series data is that the relative sizes of the money flows disappear, and it introduces negative weights, which are problematic. Negative weights are problematic as their interpretability in this context is not straightforward and in network analysis the usual methods do not work with negative weights. It is possible to get rid of the negative weights by simply mapping them to 0 or using some other more sophisticated method but the use cases for the network model would drop as the amount of information in the network would be reduced. Thus, keeping the scaling method simple is desirable. In an ideal situation the scaling would be done using some

security-based metric, which could be market capitalization or valuation of the security for example. The problem with these two example metrics is that they are not applicable to the balance node that cannot be removed from the network as it contains information about unobserved transactions.

All in all, choosing the right scaling method is not simple as it affects the network model in many ways. Choosing the method should be done based on the use case of the model as then the effects on the analysis can be understood the best.

Scaling of edge weight time series in this thesis is done by using maximum value scaling

$$\tilde{w}_{i,j}(t) = \frac{\bar{e}_{i,j}(t)}{\max\{\bar{e}_{i,j}(t-k+1), \bar{e}_{i,j}(t-k+2), \ldots, \bar{e}_{i,j}(t)\}}. \tag{2.5}$$

In this way all the edge values are in a range $[0, 1]$ and for all security pairs the maximum weight is 1. The maximum term in Eq. 2.5 can be seen as an approximation for the maximum money flow between the two securities. The length of the k-size moving window needs to be large enough so that the used maximum value does not change too often.

In a situation that there are not enough values in the moving window there would be high amount of 1's in the time series because of more variance in the maximum term. This same problem arises also if there is some period when the values in the time series increase in size and every new value is a new maximum value. Thus, the edge weight would be 1 for multiple following values in the edge weight time series. However, having multiple 1's in a row in the time series is not the worst problem since the values tell that there is something significant happening between the securities. Maximum scaling removes the relative importance of the largest value edges but that is the goal of using scaling.

In Figure 2.5 the same example edge series from Figure 2.3 is scaled using the maximum value scaling.

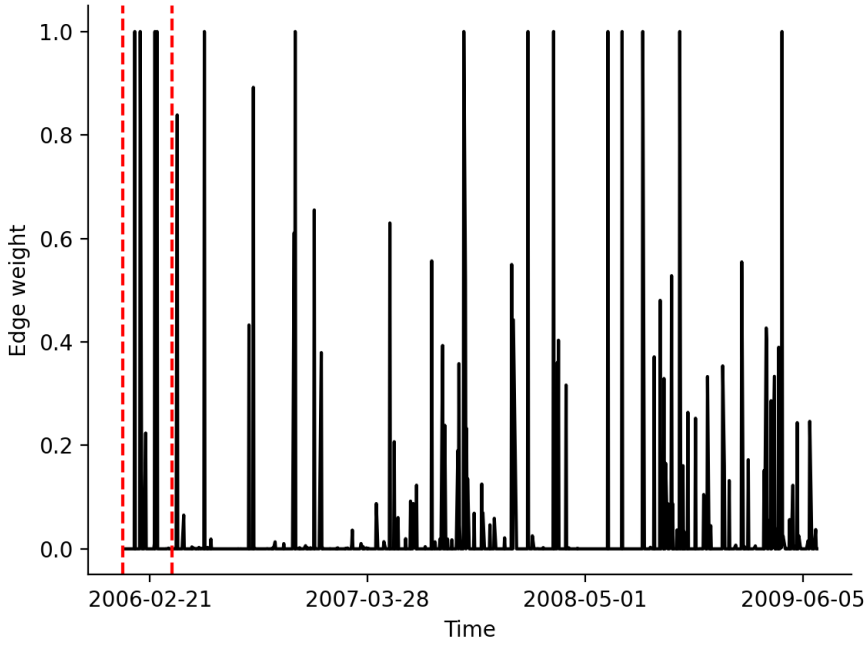


Figure 2.5. *Time series of the scaled edge weights between balance and Seligson & Co OMX Helsinki* 25*. The first* 90 *days are cut off from the time series as there are not enough dates to determine the maximum edge weight.*

Figure 2.5 shows that during the first 90 day period there is a large number of edges with weight 1 which was expected as there are not enough observations from which to choose the maximum value. The weights from 28.3.2007 to 1.5.2008 have relatively higher importance than before because the unscaled weights had lower values during the period as compared to before and after the interval. In Figure 2.6 the distribution of the example edge weights is shown after the scaling.

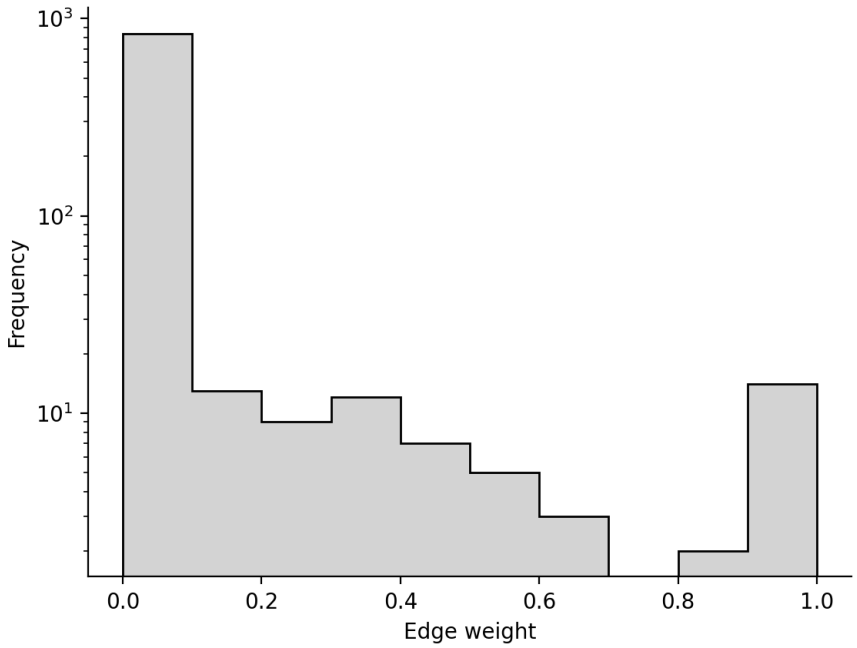


Figure 2.6. *Histogram of the scaled edge between Balance and Seligson & Co OMX Helsinki* 25*. The shape of the histogram is visually close to the unscaled weights in Figure 2.4.*

Figure 2.6 shows that the scaling works well based on the shape of the distribution as it is close to the unscaled distribution in Figure 2.4 visually. The number of 1's and high values in the series increases because if the unscaled weights have low values for over 90 days in a row, the maximum value is a low value which leads to increase in 1's. If the security is actively trading during the whole crisis there should be only small relative changes in the scaled and the unscaled distributions. The maximum value scaling can therefore be used to compare edge weights between different node pairs.

3. METHODS

In this thesis the main methods used are centrality-based rankings and subgraph counts compared to null models. Centrality measures and rankings based on them help with understanding how important different securities are in the network and how stable the possible roles of securities are in the networks. Subgraph counts and their comparisons to null models help with identifying structural changes in the network that are not explained by chance. By creating time series of the comparisons to the null models, the structural changes over time are possible to be analyzed. This can help in identifying a time when the crisis began or possible warning-signs, which Squartini, Van Lelyveld et al. (2013) found using interbank networks.

3.1 Centrality measures

Centrality measures are used for understanding how important nodes are in a network. Nodes that are connected to many other nodes can be seen as more central to nodes that are not connected to as many. Furthermore, nodes that are connected to more central nodes are themselves more central than nodes that are connected to lone nodes. Nodes can also work as bridges between different communities, which leads to those nodes being important in the network structure.

In the next subsections different centrality measures are introduced formally and more in-depth. Degree centrality is the most basic centrality measure as it is calculated as the weighted sum of the connected edges to a node. Betweenness centrality calculates the number of shortest paths that go through a node, and it can help with understanding how much information flows through the node. Eigenvector centrality tells how the node influences the network by considering how central the connected nodes are in addition to degree centrality. PageRank centrality expands eigenvector centrality by giving all nodes some set amount of centrality while also dampening the effect high valued nodes have on values of neighboring nodes, and thus PageRank works well in finding the most central nodes in general.

3.1.1 Degree centrality

Degree centrality is a straightforward centrality measure as the centrality of a node has a simple interpretation. For example, in a simple social network of friends the degree centrality of a node tells how many friends a person has if the edges are binary, and the edges tell if two persons are friends. Degree centrality can be calculated for undirected and directed networks as weighted or not. In the undirected case the degree of a node is $k_i = \sum_{j=1}^{n} a_{ij}$, where a_{ij} is an element of the adjacency matrix and n is the number of elements. The same equation can be used for the weighted case also by using the weighted edges w_{ij} instead of the binary links. For a directed network, the degree centrality splits into in-degree and out-degree that tell the weight incoming to the node or outgoing from the node. The in-degree is calculated by $k_i^{in} = \sum_{j=1}^{n} a_{ji}$, and the out-degree by $k_i^{out} = \sum_{j=1}^{n} a_{ij}$. In matrix notation the in-degrees can be written $\boldsymbol{k}^{in} = \boldsymbol{A}^T\mathbf{1}$, and the out-degrees $\boldsymbol{k}^{out} = \boldsymbol{A}\mathbf{1}$, where in both $\mathbf{1}$ is a column vector with all values equal to 1.

As an example, let a directed graph be defined by adjacency matrix

$$\boldsymbol{A} = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix},$$

which may be written as an edge list given in Table 3.1.

Table 3.1. *Edge list of the graph used for centrality measure examples.*

Node i	Node j
2	1
2	4
3	2
4	2
4	6
5	6
6	4
6	8
7	6

The example graph defined by edge list in Table 3.1 is plotted in Figure 3.1.

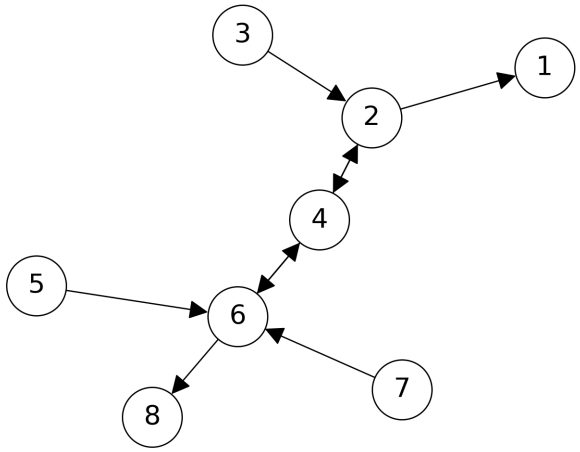
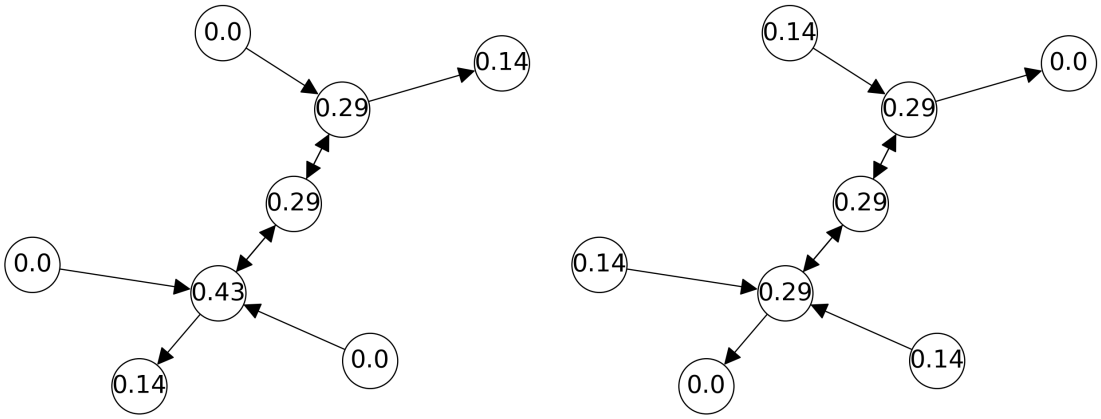


Figure 3.1. *Example graph for centrality measures.*

In Figure 3.2 the normalized in-degrees and out-degrees have been written in the corresponding nodes.



(a) *Example graph in-degrees.* ***(b)*** *Example graph out-degrees.*

Figure 3.2. *In-degrees and out-degrees of the example graph defined in Table 3.1.*

In Figure 3.2 nodes with the same number of incoming nodes have the same in-degree and nodes with the same number of outgoing edges have the same out-degree. Node 6 has the highest in-degree and nodes 2, 4 and 6 have the same out-degree, which is the highest value. Node 6 is central according to both measures. Nodes with 0 incoming links have 0 in-degree and nodes with 0 outgoing links have 0 out-degree.

3.1.2 PageRank centrality

Degree centrality is a simple metric which does not account for how central other connected nodes are and while eigenvector centrality would account for this, the eigenvector centrality gives high centrality scores to nodes which may only be pointed to from central nodes which point to many other nodes. Newman (2018, p. 165) gives an example from the Internet: the web page of Amazon may be linked by many other web pages, which makes the web page of Amazon important, but the web pages that Amazon links to are likely not as important as Amazon points to many pages. Eigenvector centrality would give high centrality scores to all web pages which are linked to from Amazon but PageRank solves this by dividing the given centrality by the out-degree of the node from which the link comes from (Page et al. 1999). Eigenvector centrality also has a problem in which nodes which do not have ingoing edges have a centrality score 0 which cannot be distributed to nodes which are connected to the node with only outgoing edges. PageRank centrality solves this by giving all nodes a predefined amount of centrality (Page et al. 1999).

PageRank of a node can be formulated mathematically as

$$x_i = \alpha \sum_j a_{ji} \frac{x_j}{k_j^{out}} + \beta, \tag{3.1}$$

where α and β are free parameters (Page et al. 1999). The definition is indeterminate if $k_i^{out} = 0$ but this can be fixed by changing the values k_i^{out} to be equal to 1 in the special cases. The problem in 3.1 is in the undefined division but even when the division is defined after the change of k_j^{out}, $a_{ji} = 0$ and the node with no outgoing edges contributes 0 centrality to other nodes. (Newman 2018) Note that in Eq. 3.1 the indices i and j have been swapped in a_{ji} as compared to (Newman 2018, p. 165) because the adjacency matrix is defined as a transpose in this thesis in contrast to the definition of the adjacency matrix by Newman (2018).

PageRank centrality in Eq. 3.1 can be defined in matrix terms as

$$\boldsymbol{x} = \alpha \boldsymbol{A}^T \boldsymbol{D}^{-1} \boldsymbol{x} + \beta \mathbf{1}, \tag{3.2}$$

where $\mathbf{1} = [1, 1, \ldots, 1]^T$ and $\boldsymbol{D}$ is a diagonal matrix with elements $D_{ii} = max(k_i^{out}, 1)$ (Newman 2018, p. 165). Now $\boldsymbol{x}$ can be solved to be

$$\begin{aligned} & \boldsymbol{x} = \alpha \boldsymbol{A}^T \boldsymbol{D}^{-1} \boldsymbol{x} + \beta \mathbf{1} \\ \iff & \boldsymbol{x} - \alpha \boldsymbol{A}^T \boldsymbol{D}^{-1} \boldsymbol{x} = \beta \mathbf{1} \\ \iff & (\boldsymbol{I} - \alpha \boldsymbol{A}^T \boldsymbol{D}^{-1}) \boldsymbol{x} = \beta \mathbf{1} \\ & \boldsymbol{x} = \beta (\boldsymbol{I} - \alpha \boldsymbol{A}^T \boldsymbol{D}^{-1})^{-1} \mathbf{1}. \end{aligned}$$

Thus, the value of β plays no role and it can be set as 1 because it only scales the results. The value of α should be less than the inverse of the largest eigenvalue of $\boldsymbol{A}^T \boldsymbol{D}^{-1}$ and it is conventionally set as 0.85 (Newman 2018, p. 166).

Figure 3.3 shows PageRank values of the example graph defined in Table 3.1.

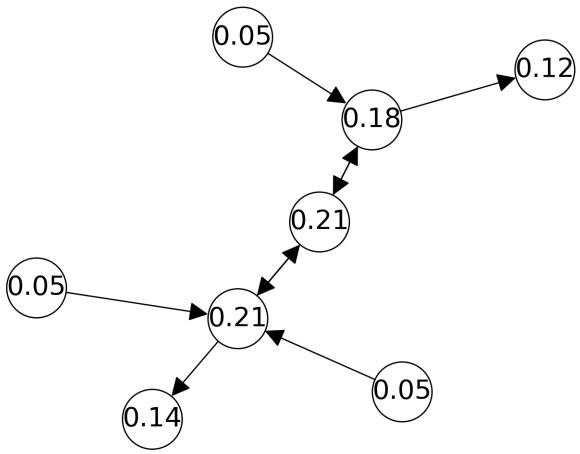


Figure 3.3. *PageRank centrality scores of the example graph.*

In Figure 3.3 the nodes 4 and 6 have the highest centrality scores, and they are therefore

central nodes in the network. In contrast to in-degrees and out-degrees in Figure 3.2, none of the nodes have 0 centrality score. The way PageRank distributes centrality can be seen from the way nodes 1 and 8 have different centralities even though they both have 1 linked edge. Nodes 4 and 6 have the highest centrality scores and are therefore the most important nodes in the network according to PageRank.

3.1.3 Betweenness centrality

Betweenness centrality helps with understanding which nodes affect flows in a network the most. Betweenness centrality of a node is calculated as the number of shortest paths between all nodes that go through the node in question. For example, in an information network where edges are possible ways that information can go through and nodes assign information to these edges, betweenness centrality helps with understanding which node is used the most in distributing information. Removing the node with the highest betweenness centrality would disturb the flow of information the most and in an information network the node could have the highest workload.

In the calculation of betweenness centrality there may be more than one shortest path between two nodes. This is the case when the shortest path may be ABC or ADC between nodes A and C. Thus, betweenness centrality is calculated so that a node gets the fraction of shortest paths that go through it out of all the shortest paths that are between two nodes. So, betweenness centrality is calculated as

$$c_b(i) = \sum_{st} \frac{n_{st}^i}{g_{st}},$$

where the sum is taken over all node pairs that do not include i, n_{st}^i is the number of shortest paths that go through node i and g_{st} is the number of shortest paths between nodes s and t. (Newman 2018, pp. 173-177) Betweenness centrality is defined the same way for directed and undirected networks as in the undirected case the same path is simply counted twice. Double counting does not matter because when betweenness centrality is used the relative importance of nodes is the most interesting part and not the absolute values. Therefore, betweenness centrality can be scaled to be a fraction of the paths that go through a given node. In an undirected network this is done by dividing the centrality by $(N-1)(N-2)/2$ and in a directed network by $(N-1)(N-2)$ where in both N is the number of nodes.

Betweenness centralities of the graph defined in Table 3.1 have been calculated in Figure 3.4.

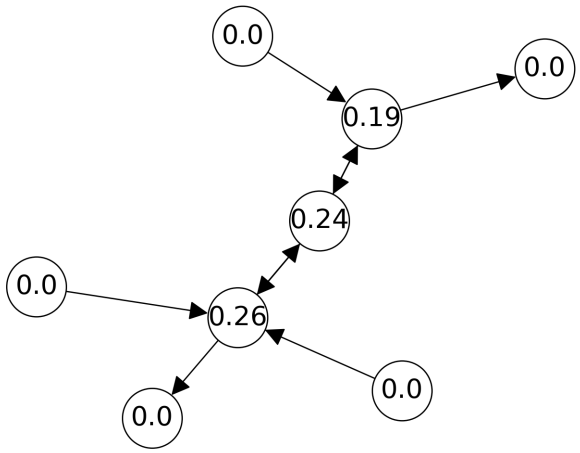


Figure 3.4. *Betweenness centrality scores of the example graph.*

The betweenness scores show that nodes 2, 4 and 6 have shortest paths that go through them and the other nodes have 0 shortest paths that go through them. The other nodes have 0 betweenness scores as they do not have both ingoing and outgoing edges which are required for there to be a path between other nodes which goes through the node in question. As before, node 6 has the highest centrality score and it can be deemed that node 6 is the most central node in the network and its removal would change the network the most.

3.2 Centrality-based ranking and statistical testing

In this section a method is explained for testing if some security is more often among the highest or lowest centrality securities than if the securities had uniformly random order. The main goal is to find out if a security has higher number of observations n_i among the extreme values than would be expected. After calculating centrality measures $c_i(t)$ at a time t for all securities i, the securities are sorted in ascending order based on the centrality values in a sequence

$$\boldsymbol{r}(t) = \{c_3(t), c_1(t), \ldots, c_i(t)\}.$$

The indices in the sequence are mixed on purpose because the index i tells which security the centrality value belong to, and the important part is the order $r_{i-1}(t) \leq r_i(t) \leq r_{i+1}(t)$ of the values. Based on the sorting $\boldsymbol{r}(t)$, the highest q-percentile of securities have their number of observations n_i increased by 1. The number of observations n_i is used to track how often a security is among the highest or lowest centrality securities. This means that if q has a value of 0.1 and there are 100 securities then the 10 securities with the highest centralities in ranking $\boldsymbol{r}(t)$ have their number of observations n_i increased by 1. This ranking and calculation of observations is done for all time steps $t \in T$ when there are

M time steps in total.

Furthermore, the probability of security i belonging to q-percentile of the ranking is q if the ranking is seen as a uniform distribution. This hypothesis can be tested with a binomial test for all securities as the ranking is repeated M times. So, for stock i the p-value of observing the stock being in q-percentile n_i times is

$$\Pr(X \geq n_i) = \sum_{n_i}^{M} \binom{M}{n_i} q^{n_i} (1-q)^{M-n_i}, \tag{3.3}$$

which is then compared to a critical value α. The null hypothesis that the ranking is uniformly distributed H_0 holds if $\Pr(X \geq n_i) \geq \alpha$ and the alternative hypothesis H_1 holds if $\Pr(X \geq n_i) < \alpha$.

If the alternative hypothesis holds for a security, the number of times the hypothesis has been debunked R is increased by 1. Once the binomial test has been done for all securities, it is possible to deduct from the value of R if there are securities that are among the q-percentile more often than expected. For instance, if the value of R is 3 and the value of q is 0.1, then there were 3 securities more often in the top 10% than would be expected. In other words, high value of R means that there are multiple securities that are in the q-percentile more often than expected if they were uniformly distributed.

It is also interesting to ask how large the value R can be because it could possible that some securities have high centralities when returns are positive and some securities have high centralities when returns are negative or there could be some other way the centralities are affected inequally at different times. This can lead to higher values of R than would be expected by simply assuming that the maximum value of R is qN, which would be 10 if $q = 0.1$ and $N = 100$. It is possible that more than 10 securities are among the highest 10% of securities than expected if the securities were uniformly distributed according to the binomial test during the time period. The maximum value of R can be calculated by finding $n_{i,min}(\alpha)$, which is the lowest number of observations n_i that has p-value below the threshold α, and dividing the total number of observations awarded during the observation period by the minimum n_i. The total number of observations during the period is always qNM, which is the number of securities that are seen to belong to the highest values qN multiplied by the number of time steps M. For instance, if $q = 0.1$ and $N = 100$ there are 10 securities which belong to the highest values at every time step, and if there are 10 time steps then the n_i values are increased at total 100 times. Thus, the maximum value of R is

$$R_{max}(\alpha) = \frac{qNM}{n_{i,min}(\alpha)}. \tag{3.4}$$

The value of $n_{i,min}$ can be found by lowering the value of n_i in Eq. 3.3 until the p-value is smaller than α.

The calculation of R can be repeated for different values of q and α to study the rankings more in-depth and the analysis can be repeated for the top and bottom percentile of securities in the same way. In this thesis the q-values for the top and bottom are 10%, 20%, 30%, 40% and 50% and they are studied with significance levels of 10%, 1%, 0.1% and $0.01/N$. The last significance level is gotten by using Bonferroni correction $\alpha_{new} = \alpha/n_{tests}$ at the 1% significance level, and it balances the equation 3.4 by taking the number of securities into account as higher number of securities increases R_{max}.

3.3 Network motifs

Real-world complex networks often share similar global properties such as a scale-free degree distribution (Barabási and Albert 1999), small-world property (Watts and Strogatz 1998) and community structure (Newman 2006). Even though real-world complex networks share many global properties they can differ on the local level (Maslov and Sneppen 2002) and it has been proposed that these local level differences are studied by counting subgraphs from the networks (Leinhardt and Holland 1974). In biological networks different subgraphs have varying roles in the network, for example negative autoregulation subgraph has been found to speed up response to signals in *E. Coli* SOS DNA repair system (Camas et al. 2006). Network subgraph analysis has been used to study lending between banks during the financial crisis (Squartini, Van Lelyveld et al. 2013), problem-solving networks (Braha 2020) and social networks (Leinhardt and Holland 1974) in addition to biological networks. The study of subgraphs in networks is therefore interesting and it helps with understanding local network topology better.

Network motifs are subgraphs of a larger graph that appear more frequently than expected in a randomized graph the same size as the larger graph (Milo, Itzkovitz et al. 2004). Motif discovery from a network contains three different steps (Patra and Mohapatra 2020)

1. Different size subgraphs are extracted from the network.
2. Frequencies of the found subgraphs are counted.
3. Expected frequencies of random graphs are compared to the counted frequencies, and statistical significance is determined.

The process is computationally expensive as the number of possible subgraphs grows exponentially with the size of the subgraph (integer sequence A000273 (OEIS Foundation Inc., The On-Line Encyclopedia of Integer Sequences 2021)) likewise the maximum number of subgraphs grows exponentially in the network with the size of the network. In the first step of the process there is also a need to know the different isomorphic forms of

the subgraph which is known to be a NP-complete problem (McKay and Piperno 2014). The third step multiplies the computational cost by sampling frequencies from randomly generated networks. Alternative way of determining the statistical significance has been proposed by Squartini and Garlaschelli (2011), and they use maximum-entropy ensemble to approximate the expected topological properties and their variances.

Many different algorithms have been developed for motif discovery that may use exact, sampling, or analytical approaches accompanied by different methods that speed up the process. One of the first algorithms proposed by Milo et al. (Milo, Shen-Orr et al. 2002) used an exact approach to calculate all of the subgraphs. Kashtan et al. (Kashtan et al. 2004) proposed an algorithm that uses an edge sampling technique to only count a fraction of the subgraphs which can be used to show that there are more of those subgraphs than expected. One of the fastest concurrent methods is Kavosh, which uses exact enumeration and it may be used for subraphs with more than 8 nodes while other algorithms may start to have problems with larger subgraphs (Kashani et al. 2009). Subgraph enumeration is computationally expensive, and it also needs to be repeated multiple times for randomly generated graphs. For example (Kashani et al. 2009; Milo, Shen-Orr et al. 2002) used 1000 randomized networks to determine the statistically significant subgraphs. Analytical methods also need to count the subgraphs from the network but they use the network properties, such as density or degree distribution, to determine the statistical significance of the subgraph (Picard et al. 2008; Squartini and Garlaschelli 2011).

In this thesis two approaches are used in motif discovery: the first one uses a grand canonical ensemble of directed random graphs, while the second one uses sampling of directed graphs with a given degree sequence. A grand canonical ensemble is used in the first approach because it is possible to derive an analytical first-order approximation for the standard deviation for the number of subgraphs in a directed random graph model. In the case of random graphs with a given degree sequence it is possible to also derive first-order approximation of the number of subgraphs and their standard deviations, but the implementation is not as straightforward and therefore a sampling technique is used. The approximation would need to be implemented if the studied networks were large as the method would be faster for small subgraphs.

3.3.1 Network subgraphs and their statistical significance

Formally a graph $G' = (V', E')$ is a subgraph of a graph $G = (V, E)$ if $V' \subseteq V$ and $E' \subseteq E \cap (V' \times V')$. Graphs G and G' are isomorphic if there is a bijection $f\colon V \to V'$ so that $\langle u, v\rangle \in E \iff \langle f(u), f(v)\rangle \in E'$ for all $u, v \in V$. An appearance of G' in G happens when G'' is isomorphic to the graph G' and $G'' \subset G$. By this definition the number of appearances N_m of a particular subgraph counts all the appearances of the isomorphism class of graphs. In this thesis the focus is on connected triads which are visualized in

Figure 3.5 up to isomorphism. It is possible to count subgraphs analytically from an adjacency matrix A of a network and the equations for counting triadic subgraphs are given in Table 3.2 up to a constant α_m that depends on the symmetries of the subgraph (Squartini, Van Lelyveld et al. 2013). The value of α_m is the order of the automorphism group of the subgraph which is discussed more in detail later.

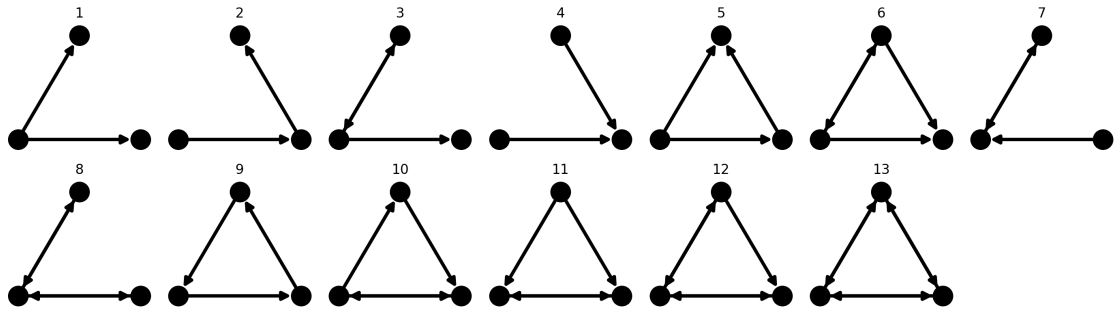


Figure 3.5. *The 13 possible non-isomorphic triadic subgraphs. Only the subgraphs 1, 2, 4, 5 and 9 are possible in the money flow networks because there cannot be money flows in both directions between securities by the definition of the networks.*

Table 3.2. *Equations for calculating the number of triadic subgraphs m in a network from an adjacency matrix A. The subgraphs are visualized in Figure 3.5.*

Triadic subgraph (m)	Count (N_m) up to a constant α_m
1	$\sum_{i\neq j\neq k}(1-a_{ij})a_{ji}a_{jk}(1-a_{kj})(1-a_{ik})(1-a_{ki})$
2	$\sum_{i\neq j\neq k}a_{ij}(1-a_{ji})a_{jk}(1-a_{kj})(1-a_{ik})(1-a_{ki})$
3	$\sum_{i\neq j\neq k}a_{ij}a_{ji}a_{jk}(1-a_{kj})(1-a_{ik})(1-a_{ki})$
4	$\sum_{i\neq j\neq k}(1-a_{ij})(1-a_{ji})a_{jk}(1-a_{kj})a_{ik}(1-a_{ki})$
5	$\sum_{i\neq j\neq k}(1-a_{ij})a_{ji}a_{jk}(1-a_{kj})a_{ik}(1-a_{ki})$
6	$\sum_{i\neq j\neq k}a_{ij}a_{ji}a_{jk}(1-a_{kj})a_{ik}(1-a_{ki})$
7	$\sum_{i\neq j\neq k}a_{ij}a_{ji}(1-a_{jk})a_{kj}(1-a_{ik})(1-a_{ki})$
8	$\sum_{i\neq j\neq k}a_{ij}a_{ji}a_{jk}a_{kj}(1-a_{ik})(1-a_{ki})$
9	$\sum_{i\neq j\neq k}(1-a_{ij})a_{ji}(1-a_{jk})a_{kj}a_{ik}(1-a_{ki})$
10	$\sum_{i\neq j\neq k}(1-a_{ij})a_{ji}a_{jk}a_{kj}a_{ik}(1-a_{ki})$
11	$\sum_{i\neq j\neq k}a_{ij}(1-a_{ji})a_{jk}a_{kj}a_{ik}(1-a_{ki})$
12	$\sum_{i\neq j\neq k}a_{ij}a_{ji}a_{jk}a_{kj}a_{ik}(1-a_{ki})$
13	$\sum_{i\neq j\neq k}a_{ij}a_{ji}a_{jk}a_{kj}a_{ik}a_{ki}$

In the Table 3.2 the logic in the equations is straightforward. For example, in the case of the triadic subgraph 1 there must be edges from j to i and j to k and not any other edges between any other nodes. The equations are not unique in the sense that they could be defined in other ways also, for example triadic subgraphs 1 could be counted by equation

$$\sum_{i \neq j \neq k} a_{ij}(1 - a_{ji})(1 - a_{jk})(1 - a_{kj})a_{ik}(1 - a_{ki}).$$

The two equations give the same answer as the graphs defined by the equations are isomorphic to each other and the triple sums in the equations count all of the permutations, which means the triple sums count different automorphisms of the subgraph separately. Therefore, by dividing the counts N_m by the order of the automorphism group of the subgraphs α_m, the equations yield the counts of the subgraphs up to the isomorphism classes which is made more clear next.

To show that the count N_m in equations given in Table 3.2 are multiplied by the order of the automorphism group of the graph, the first step is to notice how the triple sums count all the permutations of the nodes separately. If i, j and k are the indices in the sums then the sums count $\{i, j, k\}$ and $\{i, k, j\}$ separately. In general, the indices can be ordered in $n!$ different ways, where n is the number of nodes in the subgraph. The operation of swapping two indices corresponds to mapping the edges of the subgraph to new coordinates in the adjacency matrix A of the subgraph. Using the subgraph 1 as an example, its adjacency matrix A is

$$\begin{matrix} i & j & k & \\ \begin{bmatrix} 0 & 0 & 0 \\ 1 & 0 & 1 \\ 0 & 0 & 0 \end{bmatrix} & & & \begin{matrix} i \\ j \\ k \end{matrix} \end{matrix}$$

and swapping indices i and j changes the adjacency matrix to

$$\begin{matrix} j & i & k & \\ \begin{bmatrix} 0 & 1 & 1 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} & & & \begin{matrix} j \\ i \\ k \end{matrix} \end{matrix}$$

while swapping indices i and k changes the adjacency matrix to

$$\begin{array}{c} \begin{matrix} j & i & k \end{matrix} \\ \begin{bmatrix} 0 & 0 & 0 \\ 1 & 0 & 1 \\ 0 & 0 & 0 \end{bmatrix} \begin{matrix} j \\ i \\ k \end{matrix} \end{array} .$$

The operation of swapping indices i and j yielded a graph which is isomorphic to the original graph while swapping indices j and k yielded a graph that is automorphic to the original graph. Graphs are automorphic to each other if they are isomorphic and additionally their edge sets are the same. An automorphism is by definition a permutation σ of the vertex set V so that if (u, v) is an edge then $(\sigma(u), \sigma(v))$ is also an edge. Thus, by the construction of the equations and the definition of graph automorphism, subgraphs are calculated $|\mathrm{Aut}(G)|$ times in the equations as the sums run over all permutations of the nodes. Note that $|\mathrm{Aut}(G)|$ and $|\mathrm{Iso}(G)|$ are connected by equation $|\mathrm{Iso}(G)| = n!/|\mathrm{Aut}(G)|$ (Picard et al. 2008), which is a direct result from the Orbit-Stabilizer theorem.

It might be possible to derive an equation for the orders of automorphism groups of the graphs, but in this thesis only a few triadic subgraphs are analyzed and thus a brute-force approach is adequate. As the edges are defined by Eq. 2.1 there are no reciprocated edges which means that only subgraphs 1, 2, 4, 5 and 9 are possible in the networks. Hence, the problem is limited to only few graphs with 3 nodes and the problem can be solved easily. Table 3.3 contains all the automorphisms of the triadic graphs that are under study in this thesis. In the table, the automorphisms are given in one-line notation which means a permutation

$$\sigma = \begin{pmatrix} x_1 & x_2 & x_3 & \cdots & x_{n-1} & x_n \\ \sigma(x_1) & \sigma(x_2) & \sigma(x_3) & \cdots & \sigma(x_{n-1}) & \sigma(x_n) \end{pmatrix}$$

is written as

$$\sigma = \big(\sigma(x_1)\ \sigma(x_2)\ \sigma(x_3)\ \cdots\ \sigma(x_{n-1})\ \sigma(x_n)\big)$$

to save space.

Table 3.3. *Triadic subgraph automorphism group orders* (α_m)*.*

Triadic subgraph	Automorphisms	$\alpha_m = \lvert\text{Aut}(G_m)\rvert$
1	$(1\ 2\ 3), (1\ 3\ 2)$	2
2	$(1\ 2\ 3)$	1
4	$(1\ 2\ 3), (2\ 1\ 3)$	2
5	$(1\ 2\ 3)$	1
9	$(1\ 2\ 3), (3, 1, 2), (2, 3, 1)$	3

Now it is possible to count the triadic subgraphs by dividing the values given in Table 3.2 by α_m.

To determine the statistically significant subgraphs there are different statistical measures that can be used, Milo, Shen-Orr et al. (2002) used frequency, p-value and z-score. Frequency of the subgraph is a simple cut-off value at which the subgraph is said to be statistically significant. This can be useful for larger subgraphs as the expected abundance of them is lower, for example Kashani et al. (2009) used frequency of 4 to determine that a subgraph is significant. For a subgraph, p-value is the cut-off probability value for observing the number of subgraphs or a higher count in a randomized network. Milo, Shen-Orr et al. (2002) used a probability cut-off value of 0.01. The last measure Milo et al. used was a standard score which is calculated by

$$\text{z-score}(G_m) = \frac{N_m - \langle N_m \rangle}{\sigma[N_m]},$$

where N_m is the observed number of subgraphs m, $\langle N_m \rangle$ is the expected value of the number of subgraphs m and $\sigma[N_m]$ is the standard deviation in the set $R(G) \subseteq \Omega(G)$ of random graphs from an ensemble $\Omega(G)$. Kashani et al. (2009) used z-score of 1 as a cut-off value which means that the observed count is one standard deviation away from the expected count. The z-score calculation assumes a Gaussian distribution which may yield false positives as the distribution underestimates the tail probabilities (Picard et al. 2008). Thus, a stricter cut-off z-score value could be used to lower the chance of having false positives.

It is also possible to study the relative significance of the subgraphs by creating significance profiles from the z-scores. This is done by normalizing the vector of the subgraph z-scores to 1:

$$\text{SP}_i = \frac{z_i}{\sqrt{\sum z_j^2}},$$

where z_i is the z-score of graph i and the sum is taken over all the z-scores. (Milo, Itzkovitz et al. 2004) Milo et al. note that the significance profiles are useful when motifs of different networks are compared because larger networks are inclined to have motifs with higher z-scores than smaller networks.

This thesis uses the z-score as the measure for how dissimilar the observed networks are from the randomized networks. The problems of underestimating the count of subgraphs and comparing different sized networks should have almost no impact on the analysis. In this thesis, the possible trends in the dissimilarities are the most important aspect as changes in the networks would show that something happened in the networks during the financial crisis of 2007-2008. If the z-scores are close to 0 and there are no clear trends then it may be concluded that the money flows between securities are modelled by the random network, and thus the money flow directions are random.

In all of the measures an ensemble $\Omega(G)$ of random graphs associated to the graph G is used to determine the statistical significance. In the next subsections two different null models are introduced: directed random graph model and directed configuration model which are used to determine the z-scores of the subgraphs in this thesis.

3.3.2 Directed random graph model

A basic null model for binary directed network is the directed random graph. In Erdős-Rényi-Gilbert model probability p defines how likely it is for the edge a_{ij} to be present in the random graph (Fienberg 2012). Squartini, Van Lelyveld et al. (2013) show that the observed connectance is

$$c(A) = \frac{L(A)}{N(N-1)} = p, \tag{3.5}$$

where A is the adjacency matrix, L is the number of directed links and N is the number of nodes in the network. Thus, p can be estimated from an observed network and the probability can be used to create random graphs whose attributes can be compared to the observed network.

Below I show that under the directed graph model, the expected number and standard deviation for the number of motifs are

$$\langle N_m \rangle = \frac{T_1}{\alpha_m} p^k (1-p)^{6-k},\ \sigma_{N_m} = \frac{T_2}{\alpha_m} [k p^{k-1}(1-p)^{6-k} - (6-k)p^k(1-p)^{5-k}],$$

where k is the number of links in the subgraph, $T_1 = N(N-1)(N-2)$ and $T_2 = (N-2)\sqrt{N(N-1)p(1-p)}$. The expected number of triadic subgraphs in directed random graph model is easy to derive as

$$\begin{aligned} \langle N_m \rangle &= \frac{1}{\alpha_m} \sum_{i \neq j \neq k} p^k (1-p)^{6-k} \\ &= \frac{1}{\alpha_m} {}^N P_3 p^k (1-p)^{6-k} \\ &= \frac{1}{\alpha_m} \frac{N!}{(N-3)!} p^k (1-p)^{6-k} \\ &= \frac{1}{\alpha_m} N(N-1)(N-2) p^k (1-p)^{6-k}, \end{aligned} \tag{3.6}$$

where ${}^N P_3$ are the 3-permutations of N nodes, p^k is the probability of the motif having k edges and $(1-p)^{6-k}$ is the probability of the motif not having $6-k$ edges. The standard deviation is not as straightforward to derive but it can be done using the equations provided by Squartini and Garlaschelli (2011).

Variance for any topological quantity X across an maximum-entropy ensemble of random graphs with only local constraints can be calculated by using a linear approximation for the variance by equation $B.16$ from (Squartini and Garlaschelli 2011)

$$\begin{aligned} (\sigma^*(X))^2 = \sum_{i,j} \Bigg[& \left(\sigma^*[g_{ij}] \frac{\partial X}{\partial g_{ij}} \right)^2_{\boldsymbol{G} = \langle \boldsymbol{G} \rangle^*} \\ & + \sigma^*[g_{ij}, g_{ij}] \left(\frac{\partial X}{\partial g_{ij}} \frac{\partial X}{\partial g_{ji}} \right)^2_{\boldsymbol{G} = \langle \boldsymbol{G} \rangle^*} \Bigg] + \dots \end{aligned} \tag{3.7}$$

In Eq. 3.7 the indices i and j run from 1 to N, $\sigma^*[g_{ij}]$ is the standard deviation of the probability of edge ij in the maximum-entropy ensemble, $\boldsymbol{G} = \langle \boldsymbol{G} \rangle^*$ means that the adjacency matrix $\boldsymbol{G}$ is replaced by the expected maximum-entropy ensemble adjacency matrix $\langle \boldsymbol{G} \rangle^*$ and $\sigma^*[g_{ij}, g_{ji}]$ is the covariance of two edges ij and ji. Now in the directed random graph model all the probabilities g_{ij} are p, which is defined by Eq. 3.5, and $X = N_m$, which are defined in Table 3.2.

The motif counts can be generalized the same way as in Eq. 3.6 to be

$$N_m = \frac{1}{\alpha_m} \sum_{i \neq j \neq k} p^k (1-p)^{6-k}, \tag{3.8}$$

which leads to the partial derivative

$$\frac{\partial N_m}{\partial p_{ij}} = \frac{1}{\alpha_m}(N-2)(kp^{k-1}(1-p)^{6-k} - (6-k)p^k(1-p)^{5-k}),$$

where α_m is a constant that depends on the symmetries of the particular subgraph. The sum in Eq. 3.8 runs over $N(N-1)(N-2)$ permutations but taking the partial derivative with respect to p_{ij} leaves only $N-2$ values to choose the last node from and then the symmetries of the subgraph define how the nodes can be ordered. To show this let $\{i, j, k\}$ be a permutation of the values i, j and k. Now if i and j are not equal to the indices defined in p_{ij} the partial derivative $\partial N_m / \partial p_{ij} = 0$. Thus in $\{i, j, k\}$ the values of i and j must stay the same and only k can be chosen from $N-2$ values.

Therefore, for the directed random graph model variances for number of subgraphs are defined by

$$\begin{aligned}(\sigma^*(N_m))^2 \approx \sum_{i,j} \Bigg[& \left(\sigma^*[p_{ij}] \frac{\partial N_m}{\partial p_{ij}} \right)^2_{\boldsymbol{A}=\langle \boldsymbol{A} \rangle^*} \\ & + \sigma^*[p_{ij}, p_{ij}] \left(\frac{\partial N_m}{\partial p_{ij}} \frac{\partial N_m}{\partial p_{ji}} \right)^2_{\boldsymbol{A}=\langle \boldsymbol{A} \rangle^*} \Bigg].\end{aligned} \tag{3.9}$$

The probabilities p_{ij} and p_{ji} are independent of each other and thus $\sigma^*[p_{ij}, p_{ij}] = 0$. For the probability p_{ij}

$$\sigma^*[p_{ij}] = \sqrt{\langle p_{ij}^2 \rangle - \langle p_{ij} \rangle^2} = \sqrt{p - p^2} = \sqrt{p(1-p)}.$$

Substituting all calculated values to Eq. 3.6 and noting that self-edges are excluded from the sum

$$\begin{aligned}(\sigma^*(N_m))^2 &\approx \sum_{i \neq j} p(1-p) \left(\frac{1}{\alpha_m}(N-2)(kp^{k-1}(1-p)^{6-k} - (6-k)p^k(1-p)^{5-k}) \right)^2 \\ &= N(N-1)p(1-p) \left(\frac{1}{\alpha_m}(N-2)(kp^{k-1}(1-p)^{6-k} - (6-k)p^k(1-p)^{5-k}) \right)^2\end{aligned}$$

and finally

$$\sigma^*(N_m) = \frac{T_2}{\alpha_m}\left(kp^{k-1}(1-p)^{6-k} - (6-k)p^k(1-p)^{5-k}\right), \tag{3.10}$$

where $T_2 = (N-2)\sqrt{N(N-1)p(1-p)}$ which is the wanted result.

Note that solving α_m is not mandatory when using the equations given in Table 3.2 when calculating the z-scores. Using notation

$$N_m = \frac{M_m}{\alpha_m},$$

where equations for M_m are given in Table 3.2 without dividing by α_m, then the z-scores can be calculated as

$$\begin{aligned}\text{z-score}(G_m) &= \frac{N_m - \langle N_m \rangle}{\sigma[N_m]} \\ &= \frac{M_m/\alpha_m - T_1/\alpha_m p^k(1-p)^{6-k}}{T_2/\alpha_m\left(kp^{k-1}(1-p)^{6-k} - (6-k)p^k(1-p)^{5-k}\right)} \\ &= \frac{M_m - T_1 p^k(1-p)^{6-k}}{T_2\left(kp^{k-1}(1-p)^{6-k} - (6-k)p^k(1-p)^{5-k}\right)}.\end{aligned} \tag{3.11}$$

Thus, there is a method for calculating the z-scores analytically without needing to solve the automorphism group order. The equations in Table 3.2 have a time complexity $O(N^3)$ where N is the number of nodes in the supergraph. It is possible to combine a smarter way to count the subgraphs, e.g. Kavosh (Kashani et al. 2009), and use the equations 3.6 and 3.10 to calculate the z-scores without needing to generate random graphs, but the networks are small enough that equations given in Table 3.2 are fast enough to be used for counting the subgraphs in this thesis.

3.3.3 Directed configuration model

The directed configuration model is used to generate random networks with a given degree sequence. The upside in this is that the degree distribution does not need to be known and it can be arbitrary. The way to construct a random network with a given degree sequence is by giving each vertex k_i half-links and then connecting those half-links to each other with uniform probability. In the directed case the half-links are divided to two groups: heads and tails depending on if the link directed to or from the vertex. The number of directed edges, which can be seen as a probability for a directed edge if $k_i^{out}k_j^{in}/m \ll 1$, from vertex i to vertex j is

$$p_{ij} = \frac{k_i^{out} k_j^{in}}{m},$$

where k_i^{out} is the out-degree of vertex i, k_j^{in} is the in-degree of vertex j, and m is the number of edges in the graph (Newman 2018, p. 418). In this case, the vertices may be connected to themselves forming self-loops, and there may be multi-edges or cycles. As $N \to \infty$, the expected number of self-edges and multi-edges tends to zero (Barabási and Pósfai 2016).

Deriving equations for the expected number of subgraphs and the standard deviations is possible but they are not constant time like the ones for directed random graph model. This follows from the fact that the probabilities for edges are not the same for all of the edges as they depend on both nodes i and j. The probability of an edge being present in a maximum-entropy ensemble of a binary directed graph can be shown to be

$$\langle a_{ij} \rangle = p_{ij} = \frac{x_i y_j}{1 + x_i y_j}, \tag{3.12}$$

where x_i and y_j can be solved from $2N$ coupled equations

$$\begin{aligned} \sum_{i \neq j} \frac{x_i^* y_j^*}{1 + x_i^* y_j^*} &= k_i^{out} \quad \forall i \\ \sum_{i \neq j} \frac{x_j^* y_i^*}{1 + x_j^* y_i^*} &= k_i^{in} \quad \forall i, \end{aligned} \tag{3.13}$$

in which k_i is the in- or out-degree of the node i and $x_i^*, y_i^* > 0 \ \forall i$ (Squartini and Garlaschelli 2011). The notation x^* means that the value of x is estimated from a maximum-entropy ensemble. The values for x_i^* and y_i^* can be solved by using nonlinear optimization algorithms. The optimization problem can be formulated as

$$\begin{aligned} \min_{\boldsymbol{x}^*, \boldsymbol{y}^*} \quad & \frac{1}{2} ||F(\boldsymbol{x}^*, \boldsymbol{y}^*)||^2 \\ \text{s.t.} \quad & \boldsymbol{x}^*, \boldsymbol{y}^* > 0, \end{aligned} \tag{3.14}$$

where $|| \cdot ||$ is the Euclidian norm and

$$F(\boldsymbol{x}^*, \boldsymbol{y}^*) = \begin{bmatrix} \vdots \\ f_{i,out} \\ f_{i,in} \\ \vdots \end{bmatrix}$$
$$= \begin{bmatrix} \vdots \\ \sum_{i \neq j} \frac{x_i^* y_j^*}{1 + x_i^* y_j^*} - k_i^{out} \\ \sum_{i \neq j} \frac{x_j^* y_i^*}{1 + x_j^* y_i^*} - k_i^{in} \\ \vdots \end{bmatrix},$$

where i and j take values from 1 to N. The optimization problem 3.14 can be solved by using interior-point method for nonlinear optimization provided by (Forsgren et al. 2002) by adding a small error term ϵ to the greater than 0 constraint to change the problem to a standard form. The $\boldsymbol{x}^*, \boldsymbol{y}^* > 0$ constraint cannot hold the case $\boldsymbol{x}^*, \boldsymbol{y}^* \geq 0$ as the x_i and y_i values are originally parameters which were changed for ease of notation $x_i = e^{-\alpha_i}$ and $y_i = e^{-\beta_i}$ (Squartini and Garlaschelli 2011) and thus the equality would mean that the original parameters $\alpha_i = \beta_i = \infty$.

The expected number of triadic subgraphs is

$$\langle N_m \rangle = N_m(a_{ij}^*, a_{ji}^*, a_{jk}^*, a_{kj}^*, a_{ik}^*, a_{ki}^*), \tag{3.15}$$

where $N_m(\boldsymbol{a}^*)$ are given in Table 3.2 and the values for a_{ij}^* can be calculated using Eq. 3.12.

Variance of the subgraph counts can be calculated by using Eq. 3.7. Now

$$\sigma^*[p_{ij}^*] = \frac{\sqrt{x_i^* y_j^*}}{1 + x_i^* y_j^*},$$

$$\frac{\partial N_m}{\partial p_{ij}^*} = \frac{\partial}{\partial p_{ij}^*} N_m(\boldsymbol{a}^*),$$

which has at maximum $N - 2$ times non-zero values, and

$$\sigma^*[p_{ij}^*, p_{ji}^*] = \langle p_{ij}^* p_{ji}^* \rangle - \langle p_{ij}^* \rangle \langle p_{ji}^* \rangle = 0$$

as

$$\langle p^*_{ij} p^*_{ji} \rangle = \langle p^*_{ij} \rangle \langle p^*_{ji} \rangle.$$

The partial derivative of N_m can be calculated without looping the 3 indices in the definition of N_m by taking two indices whose values are locked to the same values as i and j in p_{ij}, letting the third index roll through the $N-2$ values, which are not equal to i or j, and then going through all the $3! = 6$ permutations of the indices in the triple sum of N_m. For example, the partial derivative of subgraph 1 with respect to p_{21} with indices $\{i, j, k\} = \{1, 2, 3\}$ in the triple sum is

$$\frac{\partial}{\partial p_{21}}((1-p_{12})p_{21}p_{23}(1-p_{32})(1-p_{13})(1-p_{31})) = \\ (1-p_{12})p_{23}(1-p_{32})(1-p_{13})(1-p_{31}).$$

In the end, for the directed configuration model

$$(\sigma^*(N_m))^2 = \sum_{i,j} \left[\frac{\sqrt{x^*_i y^*_j}}{1 + x^*_i y^*_j} \frac{\partial}{\partial a^*_{ij}} N_m(\boldsymbol{a}^*) \right]^2, \tag{3.16}$$

where p^*_{ij} has been changed to a^*_{ij} in the partial derivative to make the notation clearer. The calculation has a time complexity $O(N^3)$. Eq. 3.16 can also be structured so that first the subgraphs g_i are found from the graph and then for each subgraph the change in the weight of the subgraph is calculated if weight of an edge is changed. In mathematical notation

$$(\sigma^*(N_m))^2 = \sum_i \sigma^2(g_i), \tag{3.17}$$

where i goes from 0 to the number of subgraphs,

$$\sigma^2(g_i) = \sum_{i \neq j} \left[\frac{\sqrt{x^*_i y^*_j}}{1 + x^*_i y^*_j} \frac{G(\boldsymbol{a}^*)}{d(a_{ij})} \right]^2,$$

where the i and j values go from 0 to the number of nodes in the subgraph, and $G(\boldsymbol{a})$ is the product of a_{ij} and $(1 - a_{st})$ which define the equation for the subgraph like in the subgraph count equations in Table 3.2. The function $d(a_{ij})$ is derived to help in the calculation of the partial derivate in Equation 3.16 of $N_m(\boldsymbol{a})$, and

$$d(a_{ij}) = \begin{cases} a_{ij}, \text{ if } a_{ij} \text{ in definition of } G(\boldsymbol{a}) \\ -(1 - a_{ij}), \text{ if } (1 - a_{ij}) \text{ in definition of } G(\boldsymbol{a}). \end{cases}$$

This approach makes it possible to use other algorithms to find the subgraphs from the networks so there is no need to loop through all possible node permutations if the network is not complete, which means the network does not have all edges between all nodes. Thus, the last approach has time complexity $O(n(n-1)f_m^*)$ in which n is the number of edges in the subgraph and f_m^* is the function which limits the subgraph enumeration algorithm of the maximum-entropy approximation of the network. The $n(n-1)$ scaling follows from the fact that if a subgraph is found, then all of its edges need to be evaluated to find the variance of the whole subgraph, and in the maximum-entropy ensemble the subgraphs have $n(n-1)$ edges, which follows from the probabilistic approach. Unfortunately, the maximum-entropy digraph is complete, which means in the directed graph there are edges both ways between all nodes. This follows from the fact that $p_{ij} > 0$ and $x_i^*, y_j^* > 0$ in Equation 3.12. Therefore, $f_m^* = N^n$ as all of the node permutations in the subgraphs need to be accounted for.

Another approach to calculating the standard deviation is by generating S random networks and then calculating the unbiased sample variance

$$s_m^2 = \frac{1}{S-1} \sum_{i=1}^{S} (N_m(i) - \bar{N}_m)^2,$$

where $\bar{N}_m$ is the expected number of subgraphs. On the one hand, time complexity for calculating the unbiased sample variance is $O(SN^n)$ which is possibly worse than for the linear approximation. On the other hand, the unbiased sample variance is easy to implement. The subgraphs could be counted by some other algorithm, which would have time complexity $O(f_m)$ and for S samples the complexity would be $O(Sf_m)$. Using the approach defined by Eq. 3.16 could be faster than sampling if for sufficiently large number of nodes in the network $\beta n(n-1)N^n < \gamma S f_m$, in which for clarity n is the number of nodes in the subgraph, N is the number of nodes in the network, and the values β and γ are scaling factors. Thus, the maximum-entropy approach is possibly faster for subgraphs with small number of nodes n, if the network is dense as f_m would be closer to N^n. In a sparse network it is possible to speed up the enumeration by only going through the edges which are in the network, but in a dense network there are more edges which means the speed up is not as large as compared to going through all the possible edges. In this thesis the standard deviation of the number of subgraphs is calculated by generating 100 random networks in the case of the directed configuration model.

4. DATA DESCRIPTION

To illustrate the unusuality of the time period from the beginning of 2006 to halfway of 2009, a look at an index from Finland is justified. OMX Helsinki 25 index, which contains the 25 most actively traded stocks in the Helsinki exchange (Nasdaq 2021), is plotted in Figure 4.1 during the financial crisis. The index drops from values of over $3\ 000$ in 2007 to values under $1\ 500$ in 2009.

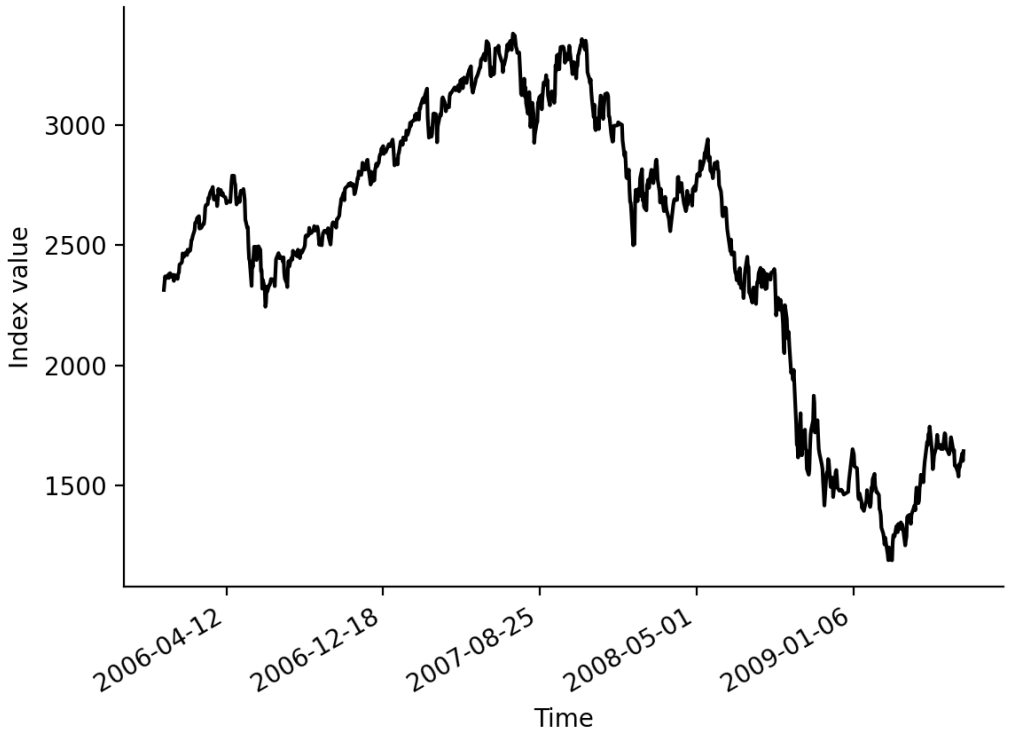


Figure 4.1. *OMX Helsinki 25 index values during the financial crisis.*

As seen from Figure 4.1 the index mainly grew steadily during the first half and during the second half the index had large drawdowns and some periods of stable values. Figure 4.2 plots the logarithmic returns of the index. Logarithmic returns are calculated as $r_t = \ln V_t - \ln V_{t-1}$, where V_t is the value of the index at time t, and they are the same as continuously compounded returns.

The logarithmic returns of the index in Figure 4.2 make it clearer that the second half of

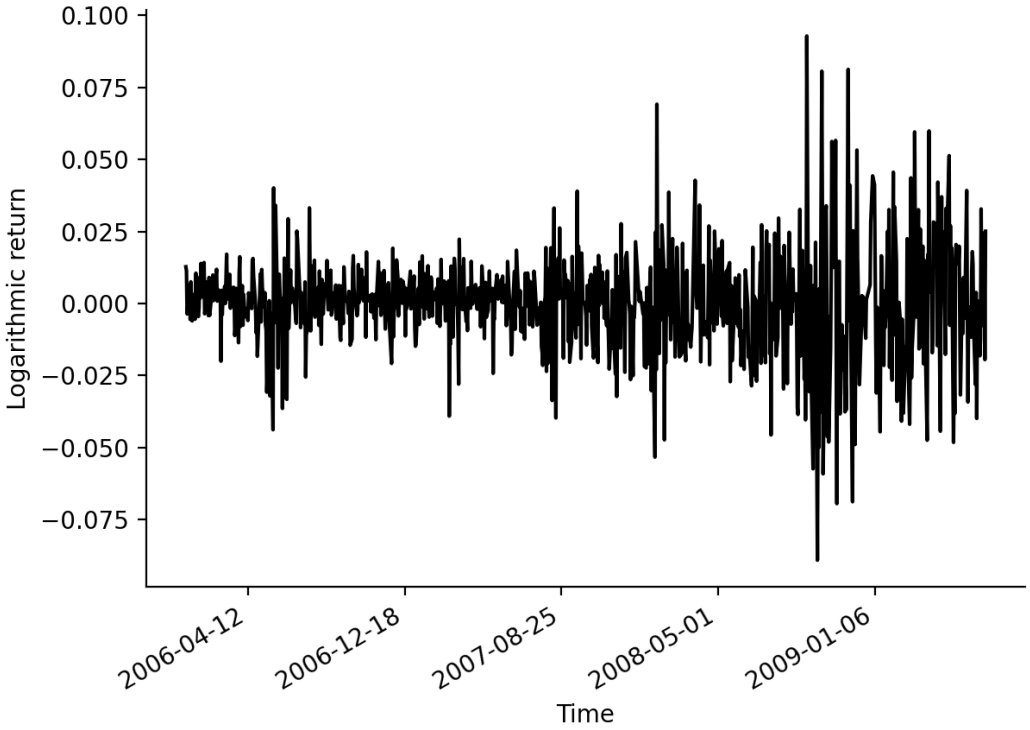


Figure 4.2. *OMX Helsinki* 25 *index logarithmic returns during the financial crisis.*

the time frame had larger relative changes. There was a shorter period of instability in 2006 but since late 2007 the returns grew in both directions. Figure 4.3 plots the 30-day rolling volatility of the index. Volatility is standard deviation and it is calculated by

$$\sigma = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(r_i - \bar{r})^2},$$

where n is the number of observations, r_i is the logarithmic return at time i and $\bar{r}$ is the mean of the logarithmic returns.

Figure 4.3 shows that the volatility increased during the time frame and had short periods when it deviated from the usual range which may be because of outliers during the 30-day rolling windows. Overall, the stock market in Finland had substantial volatility during the crisis.

Networks analyzed in this thesis are constructed from transaction data from Euroclear Finland. Transactions are grouped based on if there was a financial institution or a household in the transaction and by the transaction date, daily edge values are calculated using Eq. 2.1 for individual investors and then daily networks are aggregated using Eq. 2.2. These daily aggregated networks are studied in addition to weekly networks, which are

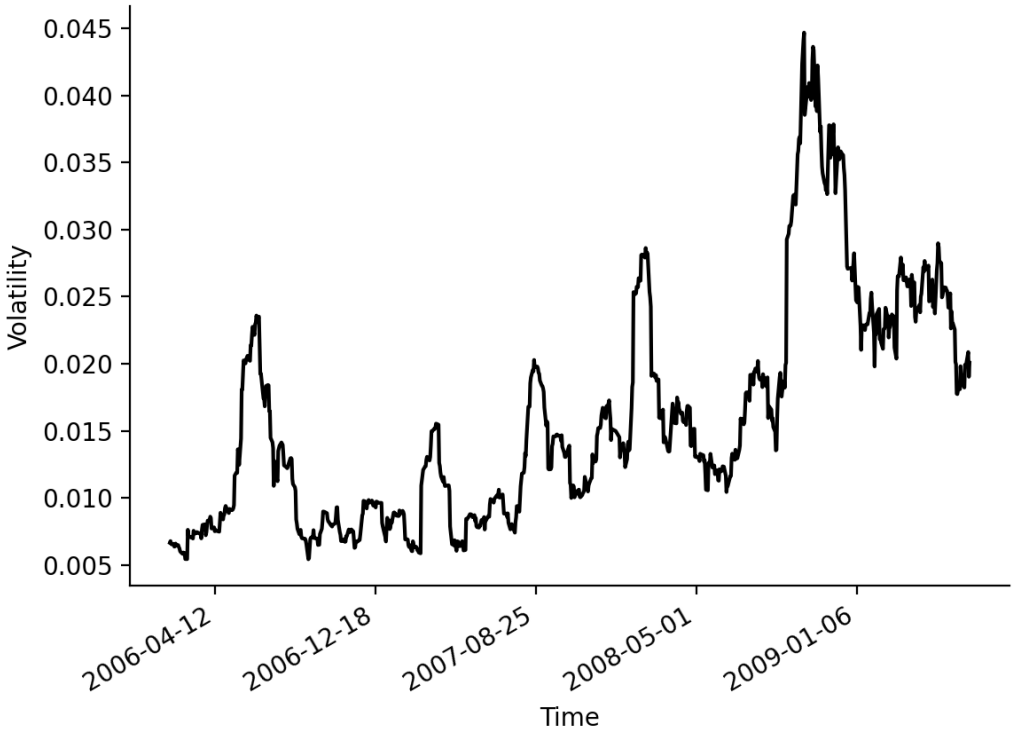
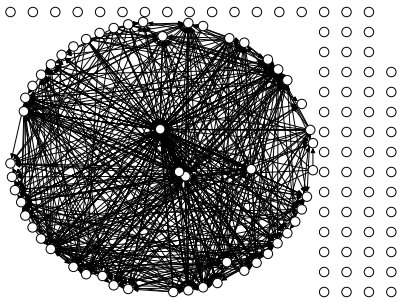
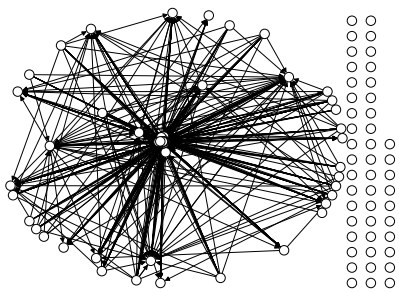


Figure 4.3. *OMX Helsinki* 25 *index* 30*-day rolling volatility values during the financial crisis.*

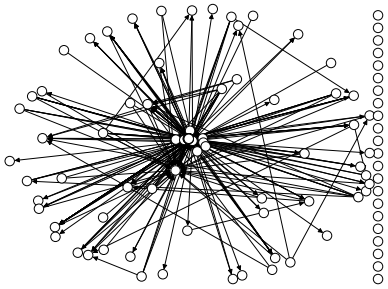
aggregated from the daily investor networks by applying Eq. 2.2 to 7 successive days, which include weekends. Weekends do not have transaction data but they are included to make the aggregation easier. In the raw data some holidays had transactions even when the market was closed, so only days when Nokia had trading data on Nasdaq were analyzed. Finance institutes had 22 holidays in the data and households had 3. Additionally, only companies, which were in listed in the stock market for the whole time were included. In total there was transaction data from 608 finance institutes and $316\ 567$ households from the time period. Finance institutes had $340\ 809$ transactions while households had $5\ 331\ 006$. Example networks are drawn in Figure 4.4.



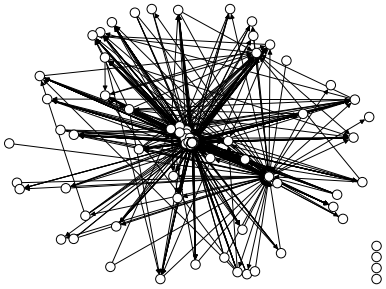
(a) *Example daily finance institute network from 25.7.2008.*



(b) *Example weekly finance institute network from 21.7.2008.*



(c) *Example daily household network from 25.7.2008.*



(d) *Example weekly household network from 21.7.2008.*

Figure 4.4. *Example networks from all network time series. The networks are drawn in weakly connected components which makes it possible to keep disconnected components in the same figure while also using edge weight-based drawing methods. In all of the networks there are* 134 *nodes which can be tightly together in the figures. The weekly networks have their first day of data reported in the captions.*

Figure 4.4 shows that the structure of the networks is hard to understand visually and mathematical methods are needed to study the networks.

Basic network statistics have been calculated in Table 4.1. The values are averages and standard deviations of the values from the time series, except for the number of securities N, which is the same for all networks. Average degree is calculated as if the network was undirected, which means that in-degree and out-degree are summed for the nodes, and the average clustering coefficient is calculated by

$$c = \frac{1}{n} \sum_i c_i,$$

where c_i is the fraction of directed triangles that go through node i out of all possible triangles (Watts and Strogatz 1998). This average local clustering coefficient gives different values than transitivity of a network by giving higher weight to nodes with low degree (Newman 2018, p. 188) and by measuring the cliquishness of a typical neighborhood (Watts and Strogatz 1998). Average shortest path has not been provided because the networks are not connected which means that there is no shortest path between some nodes.

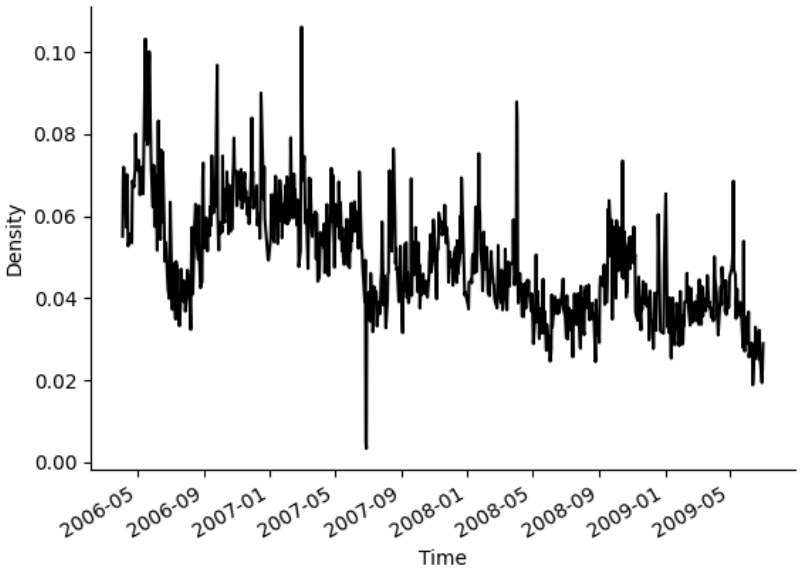
In Table 4.1 all of the networks have a low number of edges, the values of daily degree centralities are lower than weekly degree centralities, and the average clustering coefficients are quite high when the low number of edges is considered. The maximum number of edges in the networks is $8\ 911$. The degree values of weekly networks may be explained simply by the number of edges, which is higher for weekly networks when the number of nodes stays the same. Standard deviations of the clustering coefficients

Table 4.1. *Average and standard deviations of links counts, average degree and clustering coefficient in addition to the node count for all different network time series.*

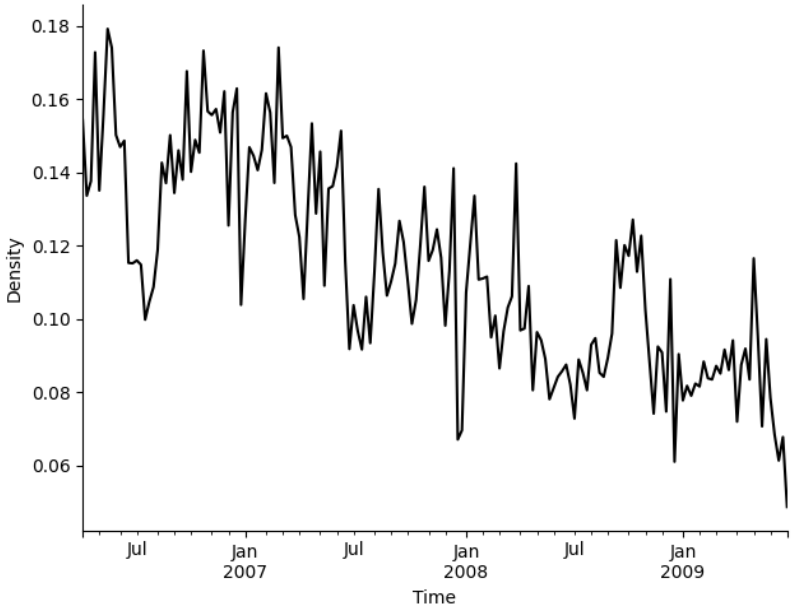
		N	$(L, \sigma(L))$	$(k, \sigma(k))$	$(c, \sigma(c))$
Finance institutes	Daily	134	$(875, 245)$	$(3.05, 1.14)$	$(0.15, 0.02)$
	Weekly	134	$(2020, 506)$	$(12.1, 3.63)$	$(0.24, 0.03)$
Households	Daily	134	$(780, 194)$	$(3.70, 1.36)$	$(0.24, 0.03)$
	Weekly	134	$(2480, 538)$	$(17.4, 5.2)$	$(0.32, 0.02)$

are low which means there are no large changes in the clustering during the time period according to this definition of clustering. Standard deviations of the number of edges and degrees are relatively the same when compared to the means of the values: standard deviations of the number of edges are approximately $1/4$ of the means and standard deviation of the degrees are roughly $1/3$ of the means. The standard deviation of number of edges of weekly household networks is relatively a little lower than for the other networks as the standard deviation is $1/5$ of the mean.

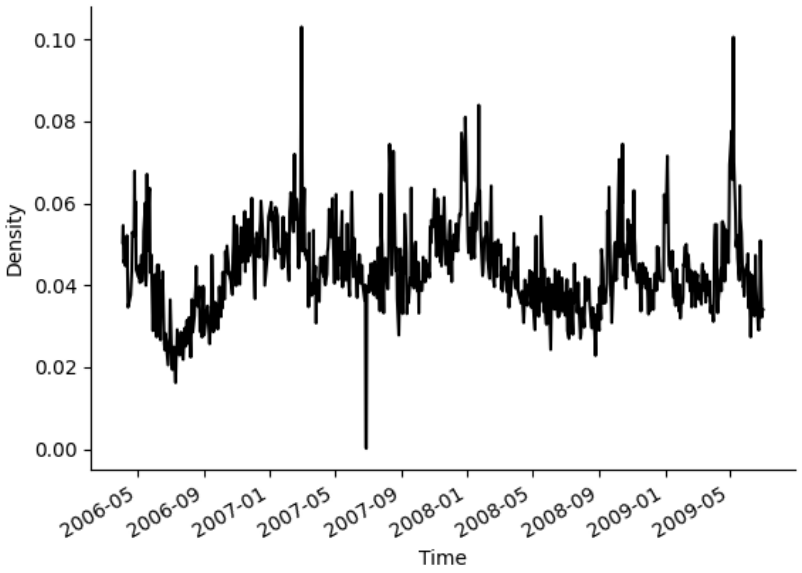
Network density affects the degree centrality measures and subgraph counts as higher density leads to more edges. Density is the same as connectance, which means it can be calculated with Eq. 3.5. Because of how the networks are defined, the maximum density is 0.5 because there cannot be edges both ways between two nodes at the same time. In Figure 4.5 all 4 network time series have their densities plotted over the crisis.



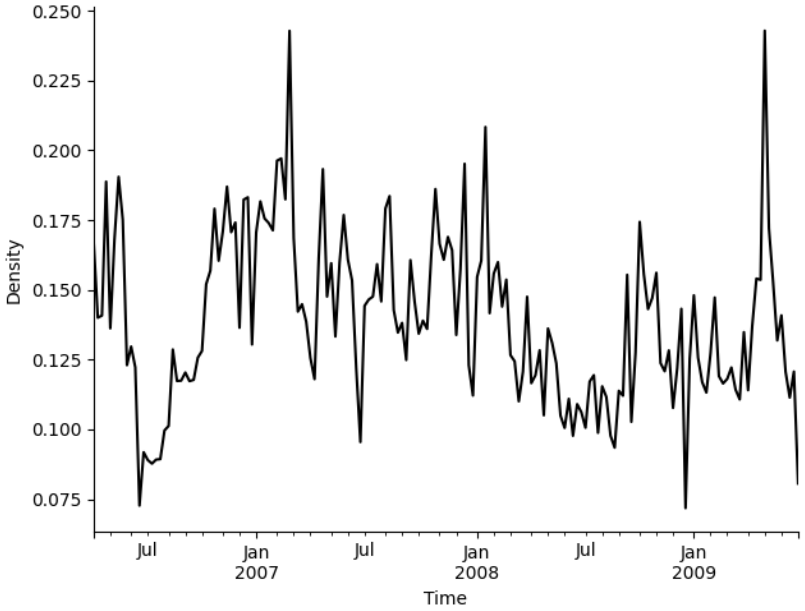
(a) *Daily finance institute network densities which show a clear negative trend.*



(b) *Weekly finance institute network densities which show larger negative trend than the daily network densities.*



(c) *Daily household network densities. The values are more stable than finance institute network densities.*



(d) *Weekly household network densities. The values don't show similar negative trend as in the case of finance institutes, but variance of the values is large.*

Figure 4.5. *Density plots of all network time series studied. All of the networks show similar short time trends in the densities, even though finance institute network densities have larger negative trends than household densities during the whole time period. Weekly household networks have the highest density values, which is logical.*

In Figure 4.5 the density plots show similar short time trends and in general a negative trend. The daily household networks have their densities closest to their mean over the

crisis and they don't show as clear negative trends as the other series. Daily household networks have the lowest density values which means there are overall less transactions, or the transactions are concentrated to certain securities during the days. Concentration to certain securities would be supported by the high density values of the weekly household network density values, because changing the securities from day to day would show as higher density in a weekly network. The daily finance institute network densities have higher values than the daily household networks during the first half of the crisis but in the second half of the crisis the densities are closer to each other. Curiously, the weekly density values are closer to each other between the groups, though the weekly household networks show more extreme values. These changes may be explained by finance institutes having consistent strategies over short time periods and them changing portfolios by reweighting the whole portfolios while households don't follow strategies for long times and do trades in pairs.

Degree distributions of networks can be used to determine structural properties of networks (Newman 2018, p. 313-319). Degree distribution gives the probability of choosing a node with degree k randomly from the network. In an undirected network the probability is

$$P(k) = \frac{n_k}{n},$$

where n_k is the number of nodes with degree k and n is the number of nodes in the network. In a directed network it is possible to calculate the in- and out-degrees of nodes, which can be used to create two different degree distributions or a two-dimensional joint distribution where the probabilities are defined as a node having a specified in- and out-degree at the same time. (Newman 2018, p. 316) The in- and out-degree distributions are plotted in Appendix A. Because a figure cannot be used to determine which theoretical distribution the observed distribution follows, a statistical test is needed.

Finding a theoretical distribution which generates a given empirical distribution is non-trivial, and in a sense, it is not a good question to ask. The process of finding the theoretical distribution should start from theory which gives a hypothesis that can be tested using statistical tools. There are many known theoretical distributions, which have been named (SciPy 2021), but there are infinite number of distributions which cannot be tested by going through the known distributions. In addition, some statistical tools do not work as one would expect if the parameters of the theoretical distribution are fitted to the empirical data. For example, Kolmogorov-Smirnov test does not produce correct results without modifications to the test (Massey 1951) and Pearson's χ^2 test rejects the true hypothesis too often (type I error) when the parameters are estimated using maximum likelihood estimation (Chase 1972; Chernoff and Lehmann 1954). Thus, a theory is needed to find a possible theoretical distribution.

Some degree distributions of networks follow a power-law distribution, and the networks are called scale-free if they do (Barabási and Bonabeau 2003). Based on the Appendix A, power-law distribution could be a good fit for the networks as the distributions have negative trends which could be linear in the log-log plots. The scale-free property comes from the fact that for the network

$$P(k) \sim k^{-\gamma},$$

where $P(k)$ is the fraction of nodes with degree k and γ is a scaling parameter, which means that there is not a typical node and there are more nodes with large deviations from the average. For example, the Pareto distribution

$$p(x, b) = \frac{b}{x^{b+1}}$$

with $x \geq 1$ and $b > 0$, is scale invariant as

$$p(cx, b) = \frac{b}{(cx)^{b+1}} = \frac{b}{c^{b+1}} p(x, b) \propto p(x, b).$$

Thus, the underlying network would be scale-free if its degree distribution followed the Pareto distribution.

According to Clauset et al. (2009) testing if a distribution follows a power-law distribution is not straightforward and common methods such as the least-squares fitting can produce inaccurate estimates of the parameters. They detail a procedure for fitting a power-law distribution to empirical data using maximum-likelihood methods and goodness-of-fit testing using Kolmogorov-Smirnov statistic and likelihood ratios. Alstott et al. (2014) have created a Python package implementing the methods of Clauset et al. (2009) which is used here for testing if the distributions follow a power-law. In Table 4.2 the value x_{min} tells the point at which the power-law fit starts and γ is the scaling parameter in addition to tests if an exponential or a log-normal distribution better explain the data at significance level 0.05.

Table 4.2 shows that for daily networks the x_{min} and γ values are smaller than for weekly networks. The estimated γ values are large when compared to the values Clauset et al. (2009) found as they estimated power-law fits to multiple data sets and found the values to be in range $[2, 3]$ for most the data sets. The x_{min} values show that the power-law fits work only for the tails of the distributions and for some distributions the exponential and log-normal distributions fit the tails better. For household networks the γ values for k^{in} are smaller than for k^{out} which means the values of k^{in} have more larger values of k^{in}. Households concentrate their buys more to some securities than their sales. For finance

Table 4.2. *Power-law fits of the networks. The notation * means that exponential distribution fits the data better than the power-law distribution, and † means the log-normal distribution fits the data better. The condition for a better fit is the logarithm of the likelihood ratio $\mathcal{R} < 0$ in addition to $p \leq 0.05$.*

		(x_{min}, γ) (k^{in})	(x_{min}, γ) (k^{out})
Finance institutes	Daily	$(15, 6.1)$	$(15, 6.2)$
	Weekly	$(28, 7, 5)$	$(25, 7, 5)^\star$
Households	Daily	$(15, 4.6)$	$(13, 5.5)^\dagger$
	Weekly	$(29, 4.8)^{\star\dagger}$	$(35, 7.3)$

institutes the γ values are close to each other when comparing the values of k^{in} and k^{out} to each other. Overall, the networks seem to have scale invariant degree distribution tails.

5. RESULTS

In this chapter the methods described earlier in Chapter 3 are applied to the finance institute and household networks. There are 4 different time series of networks: daily and weekly finance institute and household networks. First centrality rankings of securities in the networks are studied which helps with understanding if there are important nodes in the networks and if the centralities hold over time. After this, the network subgraphs are studied to understand the structure of the networks further. The study of the subgraphs helps to identify statistically significant changes in the networks over time.

5.1 Centrality ranking

In this section there are tables containing R values, which tell how many securities are more often in the percentile than expected, for different values of α and q. The α notation in this section is the significance level of the statistical test, and it should not be confused with the order of the automorphism group of a subgraph used in the network motif analysis. The notation $q = 0.1$ means the top or bottom 10% percentile of the ranking and $q = 0.5$ means the top or bottom 50% percentile of the ranking. The R-values range from 0 to R_{max} as there may be 0 securities which are statistically more often in the percentile than expected and at most there can be R_{max} securities which are more often in the percentile according to Eq. 3.4. The R_{max} depends on the significance level α, percentile q, number of observations M and number of securities N, so the R_{max} values for daily and weekly networks can be tabulated. The R_{max} values for daily networks are calculated in Table 5.1 and for weekly networks the R_{max} values are calculated in Table 5.2 according to the Equation 3.4. The R_{max} scores are reported only for the top 10% to top 50% percentiles as the values are the same for bottom 10% to bottom 50% percentiles.

Table 5.1. *R_{max} values of daily networks which had 813 observations.*

	α			
q	0.1	0.01	0.001	$7.45 \cdot 10^{-5}$
0.1	117	105	99	93
0.2	122	114	108	104
0.3	124	118	114	110
0.4	126	121	117	114
0.5	127	123	120	118

Table 5.2. *R_{max} values of weekly networks which had 170 observations.*

	α			
q	0.1	0.01	0.001	$7.45 \cdot 10^{-5}$
0.1	99	81	73	67
0.2	108	94	87	81
0.3	113	103	96	91
0.4	118	108	102	96
0.5	121	112	107	102

Tables 5.1 and 5.2 show that many values of R_{max} for both daily and weekly networks are near 100 for all values of α and q. Some of the values are close to the total 134 of securities in the networks which means almost all of the securities may be more often in the top q percentile than expected if the ranking was uniformly distributed. The closer R value of one tail of the ranking is to 134, the more it affects the R values of the other tail. This follows from the fact that if the same securities are in q percentile then the other securities are in the $1 - q$ percentile of the other tail. The R_{max} values of the weekly networks are lower because there are less observations. The lowest R_{max} value is 67 for weekly networks with $q = 0.1$ and $\alpha = 7.45 \cdot 10^{-5}$.

In Table 5.3 the R values have been calculated for daily finance institute networks using in-degree, out-degree, PageRank and betweenness based centrality scores.

Table 5.3. *R values of daily finance institute networks. There are in total 813 networks.*

		α			
In-degree	q	0.1	0.01	0.001	$7.45 \cdot 10^{-5}$

Top	0.1	30	30	30	30
	0.2	33	32	32	32
	0.3	42	40	39	38
	0.4	47	45	42	42
	0.5	44	41	40	39
Bottom	0.5	82	81	78	76
	0.4	95	93	92	90
	0.3	67	66	63	61
	0.2	0	0	0	0
	0.1	0	0	0	0
Out-degree					
Top	0.1	32	32	30	30
	0.2	33	33	33	33
	0.3	41	38	38	37
	0.4	48	45	45	43
	0.5	45	43	41	40
Bottom	0.5	80	79	77	77
	0.4	93	91	90	89
	0.3	78	76	73	70
	0.2	0	0	0	0
	0.1	0	0	0	0
PageRank					
Top	0.1	36	31	29	28
	0.2	39	35	34	33
	0.3	51	45	42	41
	0.4	50	49	44	43
	0.5	44	42	40	39
Bottom	0.5	82	80	77	76
	0.4	95	93	92	90
	0.3	67	66	63	61
	0.2	0	0	0	0

Betweenness	0.1	0	0	0	0
Top	0.1	31	31	30	29
	0.2	34	32	31	31
	0.3	38	35	35	34
	0.4	34	34	34	34
	0.5	34	32	31	31
Bottom	0.5	100	100	100	100
	0.4	103	101	100	100
	0.3	0	0	0	0
	0.2	0	0	0	0
	0.1	0	0	0	0

Table 5.3 shows that the bottom 20% of securities are not the same always according to all of the centrality measures and for betweenness centrality the same is true for the bottom 30%. Betweenness centrality may have a different ranking for the bottom 30% because there are many nodes with only 1 or 0 edges which means betweenness centrality of 0 as the nodes are not between any other nodes. Otherwise, all of the R values follow about the same structure of 30 securities in the top 10% to 50% and 90 securities in the bottom 50%. The R values change only a little bit with changes in the significance level, and thus the R values can be trusted. Therefore, there is structure in the ranking of the securities with some securities being more often among the highest values than others and no clear bottom ranked securities for the whole time period. Securities with the highest n_i values with $q = 0.1$ are reported in Table 5.4.

Table 5.4. *Securities which were the most often in top 0.1 percentile in daily finance institute networks according to in-degree, out-degree, PageRank and betweenness centralities. The n_i value is the number of times the security was in the top 10% of securities.*

In-degree		Out-degree		PageRank		Betweenness	
	n_i		n_i		n_i		n_i
Balance	662	Balance	631	Balance	756	Balance	813
Nokia	474	Nokia	441	Nokia	291	Nokia	699
Neste	411	Fortum	422	Fortum	189	UPM Kymmene	556
UPM Kymmene	393	UPM Kymmene	418	Stora Enso	186	Fortum	536
Fortum	355	Outokumpu	392	Neste	186	Neste	501

From Table 5.4 it can be seen that the securities with the highest n_i values are large companies or the balance node. Large companies are traded more and thus they have more edges which gives them high centralities. Betweenness centrality has the largest n_i values while PageRank has the lowest values. Thus, the nodes act as bridges between groups often while they may have low centrality according to PageRank. At the same time in-degree and out-degree n_i values are higher than the values of PageRank, so the nodes are often connected to each other as PageRank lowers the effect of nodes distributing centrality to each other if the nodes have high out-degree.

The networks have been changed to weekly finance institute networks in Table 5.5.

Table 5.5. *R values of weekly finance institute networks. There are in total 170 networks.*

		α			
In-degree	q	0.1	0.01	0.001	$7.45 \cdot 10^{-5}$
Top	0.1	34	31	28	28
	0.2	38	34	32	32
	0.3	44	40	36	34
	0.4	47	43	41	38
	0.5	54	51	48	45
Bottom	0.5	62	57	52	48
	0.4	60	58	58	49
	0.3	67	62	61	57
	0.2	47	38	36	34
	0.1	0	0	0	0

Out-degree					
Top	0.1	33	31	30	29
	0.2	36	33	33	33
	0.3	41	38	35	35
	0.4	49	47	41	37
	0.5	54	53	48	44
Bottom	0.5	61	56	49	49
	0.4	60	51	50	49
	0.3	64	60	57	53
	0.2	43	39	36	35
	0.1	0	0	0	0
PageRank					
Top	0.1	28	11	6	3
	0.2	38	29	25	17
	0.3	44	35	31	30
	0.4	54	42	35	32
	0.5	60	54	49	44
Bottom	0.5	60	52	47	45
	0.4	58	56	56	50
	0.3	67	62	61	57
	0.2	47	38	36	34
	0.1	0	0	0	0
Betweenness					
Top	0.1	31	30	30	29
	0.2	33	32	32	31
	0.3	40	37	35	35
	0.4	49	48	46	44
	0.5	54	48	46	45
Bottom	0.5	71	70	66	63
	0.4	77	74	71	69
	0.3	70	66	61	59

0.2	0	0	0	0
0.1	0	0	0	0

Table 5.5 shows that the R values of weekly networks are generally lower than those of daily networks but the values in the top 10% to 50% are close to each other and in weekly networks only the bottom 10% of securities are random. For PageRank there is a large change in the R values when $q = 0.1$ and the significance level is lowered. According to PageRank in the strictest case only 3 securities would be in the top 10% more than expected and 30 securities would be reached at 30%. Thus, the most central securities change more than in daily networks according to PageRank. This change in the top 0.1 percentile from daily to weekly networks could be explained by higher connectivity among the securities as during a week more combinations of securities can be traded. The close R values in the top 10% to 50% values show that there are 30 securities which are traded more often, and this property does not change across days when excluding the strictest cases according to PageRank. It could be that for one day a security is traded a lot while on another day the security is traded a lot but in the other direction which would cancel the other direction and there would not be as big money flow in the weekly networks. Because the R values in weekly and daily networks for the highest ranks are close to each other, the ranking can be seen as time invariant. This could simply be explained by the companies being large and many institutions trading the securities at all times. The bottom R values are interesting because there is a more extreme structure when compared to the daily networks. There are at least 34 securities which are ranked in the bottom 30% of the securities more often than expected. These securities may be small market capitalization securities, or they may belong to a certain unattractive industry, for example.

Table 5.6 lists the securities with the highest n_i values in weekly finance institute networks.

Table 5.6. *Securities which were the most often in top* 0.1 *percentile in weekly finance institute networks according to in-degree, out-degree, PageRank and betweenness centralities.*

In-degree		Out-degree		PageRank		Betweenness	
	n_i		n_i		n_i		n_i
Balance	108	Balance	86	Balance	149	Balance	170
Nokia	79	Rautaruukki	72	Nokia	35	Nokia	118
Neste	73	Fortum	72	Neste	32	Fortum	93
UPM Kymmene	72	Sampo	72	Fortum	31	Neste	87
TietoEVRY	67	Kone	68	Wärtsilä	31	Outokumpu	86

In Table 5.6 the balance node has the highest n_i values like it did in the daily networks. Additionally, the securities are large companies like in the daily networks. Some of the companies are different than in the daily networks but the other companies may be outside the top 5 nodes while still having high n_i values. In Table 5.5 the R value with $q = 0.1$ and $\alpha = 7.45 \cdot 10^{-5}$ is 3 which means that balance, Nokia and Neste are still more often than expected in the top 10% but Fortum and Wärtsilä are not. The difference happens to be a single observation of n_i.

R values of daily household networks are reported in Table 5.7.

Table 5.7. R *values of daily household networks. There are in total* 813 *networks*

		α			
In-degree	q	0.1	0.01	0.001	$7.45 \cdot 10^{-5}$
Top	0.1	31	31	31	30
	0.2	34	33	32	32
	0.3	43	43	42	41
	0.4	44	44	44	42
	0.5	48	46	45	45
Bottom	0.5	67	65	65	65
	0.4	61	58	54	53
	0.3	50	45	44	42
	0.2	59	57	53	51
	0.1	0	0	0	0

Out-degree					
Top	0.1	28	28	28	28
	0.2	33	33	33	33
	0.3	42	41	38	38
	0.4	47	45	44	42
	0.5	50	50	50	47
Bottom	0.5	70	67	67	65
	0.4	58	55	54	53
	0.3	51	48	46	46
	0.2	51	51	46	46
	0.1	21	17	17	16
PageRank					
Top	0.1	39	37	33	29
	0.2	43	40	39	39
	0.3	52	43	41	40
	0.4	55	48	45	44
	0.5	60	53	49	49
Bottom	0.5	62	59	56	54
	0.4	60	55	51	47
	0.3	54	47	44	44
	0.2	61	58	54	51
	0.1	0	0	0	0
Betweenness					
Top	0.1	28	28	28	28
	0.2	36	34	33	33
	0.3	39	39	37	37
	0.4	47	45	44	42
	0.5	54	51	47	47
Bottom	0.5	69	67	64	64
	0.4	70	67	66	63
	0.3	81	75	73	71

0.2	32	29	28	21
0.1	0	0	0	0

In Table 5.7 the daily bottom R values of households are lower than the same values of finance institutes. The difference may follow from the fact that the household networks have R values even when $q = 0.2$ which was not the case for finance institutes. Households have more defined structure because of this as there are securities which are favored less often than in the case of finance institutes. Therefore, households trade smaller number of securities or there are securities which are traded only rarely. In addition to the lower R values, there are about 28 securities in the top 10% more often than expected according to all of the centrality measures which is close to the value of 30 for finance institutes. Also, when $q = 0.2$ for the bottom R values all of the R values are non-zero which was not the case for finance institutes. This tells that there are nodes which have low betweenness scores more often than expected and their centrality scores are possibly 0 more often. This would support the idea of households trading smaller number of securities.

Table 5.8 contains securities with the highest n_i scores according to different centrality measures when $q = 0.1$.

Table 5.8. *Securities which were the most often in top* 0.1 *percentile in daily household networks according to in-degree, out-degree, PageRank and betweenness centralities.*

In-degree		Out-degree		PageRank		Betweenness	
	n_i		n_i		n_i		n_i
Balance	808	Balance	811	Balance	804	Balance	813
Neste	468	Nokia	569	Neste	300	Nokia	681
Metso	450	Outokumpu	444	Nokia	271	Neste	507
Nokia	441	Rautaruukki	407	Telia	258	Metso	458
Telia	416	Fortum	403	Outokumpu	237	Outokumpu	452

Table 5.8 shows the same as the n_i score tables before: the balance node and large companies have the top positions in the table.

Table 5.9 contains R values of the weekly household networks.

Table 5.9. R *values of weekly household networks. There are in total* 170 *networks.*

		α			
In-degree	q	0.1	0.01	0.001	$7.45 \cdot 10^{-5}$
Top	0.1	30	30	26	24
	0.2	38	36	34	32
	0.3	46	42	39	37
	0.4	53	45	43	42
	0.5	60	5	47	44
Bottom	0.5	62	55	49	47
	0.4	53	49	45	43
	0.3	46	43	42	38
	0.2	40	37	35	34
	0.1	33	31	27	24
Out-degree					
Top	0.1	31	28	27	27
	0.2	40	35	35	34
	0.3	49	45	42	42
	0.4	53	51	49	46
	0.5	59	56	53	51
Bottom	0.5	61	57	55	50
	0.4	50	48	48	44
	0.3	45	41	41	39
	0.2	37	34	32	30
	0.1	27	27	25	21
PageRank					
Top	0.1	37	22	19	16
	0.2	41	34	27	22
	0.3	50	41	36	30
	0.4	55	46	39	35
	0.5	55	50	44	38
Bottom	0.5	46	44	38	35
	0.4	46	43	38	34

	0.3	43	40	37	34
	0.2	37	36	33	29
	0.1	32	31	27	26
Betweenness					
	0.1	28	23	23	18
	0.2	33	32	32	32
Top	0.3	40	39	39	36
	0.4	49	44	44	43
	0.5	60	54	51	49
	0.5	65	59	55	50
	0.4	56	52	47	46
Bottom	0.3	46	45	41	41
	0.2	39	37	34	33
	0.1	35	32	29	25

In Table 5.9 the R values are the lowest of all network series and the bottom 10% R values are non-zero according to all of the centrality measures for only this network series. The top 10% R values are between 16 and 37 which both are based on PageRank. The top 10% R values are the lowest in this series when compared to the other network series. Notably, the bottom 10% has R value of at least 21 which indicates the clearest structure in the ranking yet. Households have about 24 securities they trade more often than expected and at least 21 securities which they do not trade as often as expected. The tails leave 89 securities to be in the middle 80% of the ranking if it is assumed that none of the securities are in the top and bottom values. Securities could be in both extremes if they had active and non-active weeks, but it is not likely in weekly networks. The middle 80% show signs of a stable ranking because the R values grow when the q values are increased from 0.1 to 0.5 in steps from both tails.

Table 5.10 shows the n_i values when $q = 0.1$ for weekly household networks.

Table 5.10. *Securities which were the most often in top* 0.1 *percentile in weekly household networks according to in-degree, out-degree, PageRank and betweenness centralities.*

In-degree		Out-degree		PageRank		Betweenness	
	n_i		n_i		n_i		n_i
Balance	139	Balance	160	Balance	144	Balance	170
Neste	102	Nokia	107	Telia	56	Nokia	153
Telia	92	Nokian Renkaat	84	Neste	51	Neste	132
Metso	82	Outokumpu	83	Metsä Board	49	Telia	119
Sampo	74	Telia	72	Nokia	46	Outokumpu	103

Table 5.10 shows that the balance node and large companies have the highest centrality scores often as their n_i values are the highest. Overall, to extract meaningful signals from the centrality scores other than a ranking based structure, large securities need to be balanced with small securities because many edges lead to proportionally larger centrality measures even if the edge weights are scaled.

The weekly household networks have securities in the bottom 10% of values more often than expected unlike the other network series. The bottom securities are listed in Table 5.11.

Table 5.11. *Securities which were the most often in bottom* 0.1 *percentile in weekly household networks according to in-degree, out-degree, PageRank and betweenness centralities.*

In-degree		Out-degree		PageRank		Betweenness	
	n_i		n_i		n_i		n_i
Vaahto Group	142	Vaahto Group	132	Vaahto Group	137	Vaahto Group	136
Ericsson	118	Investors House	127	Investors House	114	Investors House	132
Investors House	117	Interavanti	124	Ericsson	106	Ericsson	131
Plc Uutechnic	105	Ericsson	111	Plc Uutechnic	104	Plc Uutechnic	119
Interavanti	97	Soprano	106	Interavanti	101	Tamfelt	107

Table 5.11 shows that the bottom securities are mainly the same based on the different centrality measures and their n_i values are close to each other. On the one hand, Vaahto Group, Investors House (before 2015 known as SSK Suomen Säästäjien Kiinteistöt Oyj), Plc Uutechnic Group, Interavanti and Soprano had relatively lower market capitalizations than the securities in the top 10% of the ranking. These securities had market capitalization under 50 million euros in 2008 (Interavanti Oyj 2009; Soprano Oyj 2010; SSK Suomen

Säästäjien Kiinteistöt Oyj 2009; Vaahto Group Plc Oyj 2009). On the other hand, Tamfelt and Ericsson had relatively larger market capitalizations and them being in the bottom 10% of the ranking does not fit in the idea that the rankings are solely based on the size of the security. Tamfelt had market capitalization of 148 million euros in 2009 (Metso Oyj 2009) and Ericsson had market capitalization of 18 billion euros in 2009 (Ericsson 2009). Tamfelt is not as large company as Nokia or Neste but it is larger than the other companies in the bottom 10%. Ericsson is a clear outlier in the bottom securities based on its size. The market capitalizations are based on annual reports and company transactions, and they are therefore approximations for the year.

5.2 Network motif analysis

In this section the methods shown before for motif discovery are applied to the observed networks. Finance institute and household networks are studied separately to identify their individual characteristics which can then be used to conclude if the characteristics are present in both groups or not. Additionally, subgraphs are first studied by their z-score profiles to identify motifs which are meaningful to study more in-depth over the chosen time period. As a reminder, the links are analyzed as binary which means their weights are not studied.

Because the graphs had lone nodes and many disconnected components at many times, the largest weakly connected components of the graphs are used for the analysis. Weakly connected component is a subgraph of the original graph in which all the nodes are connected to each other by some undirected path. Only the weakly connected component is used because the effects of the disconnected components can be modelled separately, and they would have a large effect on the analysis otherwise. Increasing the number of nodes without increasing the number of links decreases connectance by a large amount, and connectance affects the directed random graph model. The effect of lone nodes could be modelled separately by first choosing some set of nodes from the network, and then forming the links between the nodes. The probability of choosing the node to be part of the model could be how many times the node has been part of the model before as a fraction of the observed number of networks.

5.2.1 Finance institute networks

Subgraph counts of daily finance institute networks are shown in Figure 5.1.

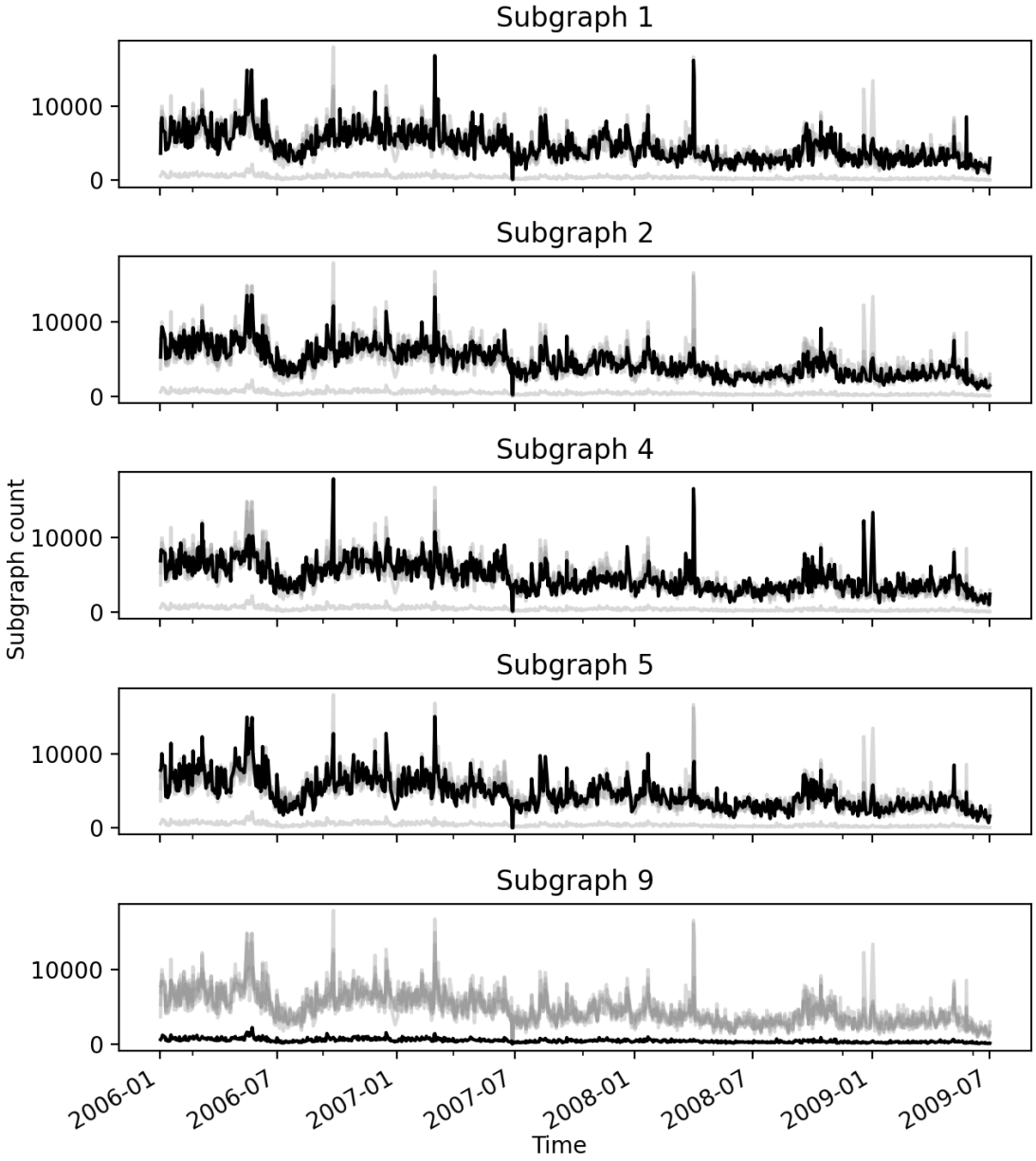


Figure 5.1. *Subgraph counts of daily finance institute networks.*

From Figure 5.1 it is possible to see that the absolute counts seem to have a negative trend over time and there are some days when the counts spike higher. Also, it is possible to see that subgraph 9 has a lot lower count than the other subgraphs. However, these absolute values may be explained by the number of nodes, the connectance of the networks and the expected number of subgraphs in the networks. Hence, the absolute values do not tell if the structure of the networks change over time, and the values need to be compared to the expected number of subgraphs by using z-scores.

In Figure 5.2 z-scores have been calculated for the daily finance institute networks under directed random graph model and profiles have been created for the subgraphs. The profiles help with seeing the statistically significant subgraphs and their deviations.

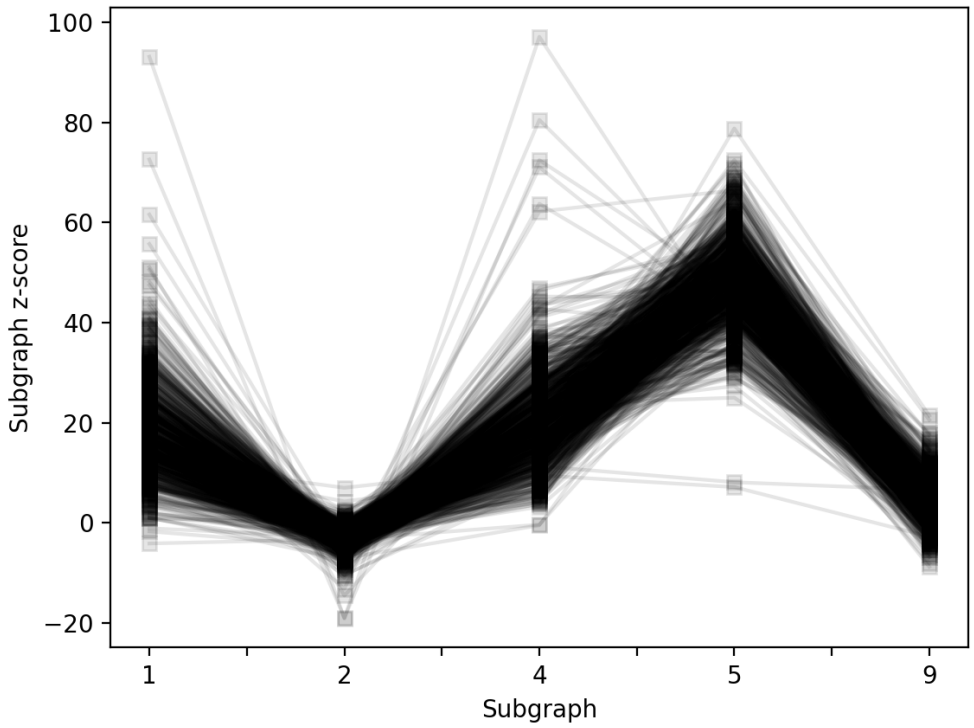


Figure 5.2. *Subgraph z-score profiles under directed random graph model of daily finance institute networks.*

From Figure 5.2 it is possible to see that all of the subgraphs 1, 2, 4, 5 and 9 are statistically significant because their z-scores are not close to 0. Subgraph 5 has the highest z-scores on average, and subgraph 2 has the lowest z-scores. Subgraphs 1 and 4 have higher z-scores than subgraph 2 which means there are more cases in which money is clearly invested to or divested from a security than a combination of some investors investing in the security and some divesting. The more complex combination of some investors investing and some divesting is usually accompanied by a third edge, which makes it clearer which security is preferred from the 3 securities and which is not. This complex combination can be observed from the z-scores of subgraph 5. Money flowing to a security from 2 securities means the target security is preferred as money flows to the security from the 2 securities, and the opposite is true for the security from which money flows to the 2 securities. The third security, which has inflow and outflow, does not have as clear future as the two other securities as investors are not agreeing if they should invest to or divest from the security. It is possible that the higher z-scores of subgraph 5 are explained by there being many combinations of 2 large securities which have large out- and inflows, and a third security which has more unclear future. Subgraph 2 has z-scores closest to 0, but at times the values are close to -20. Therefore, subgraph 2 should also be analyzed during the crisis.

Earlier the number of subgraphs 9 seemed low relative to other subgraphs in Figure 5.1, but Figure 5.2 shows that the expected number is simply lower than for the other subgraphs. Additionally, the equation for calculating the z-scores of subgraphs 4 and 9 are the same except for the number of subgraphs. In Figure 5.1 the subgraphs have been calculated only once which means for subgraph 9 the comparative subgraph count to subgraph 5 in the z-score equation is 3 times the count shown in Figure 5.1. This follows from the fact that $|\text{Aut}(G_9)| = 3$. Overall, the directed random graph model expects a different number of subgraphs than observed, and the z-scores have multiple different values during the crisis.

In Figure 5.3 the z-scores of the subgraphs have been plotted over time.

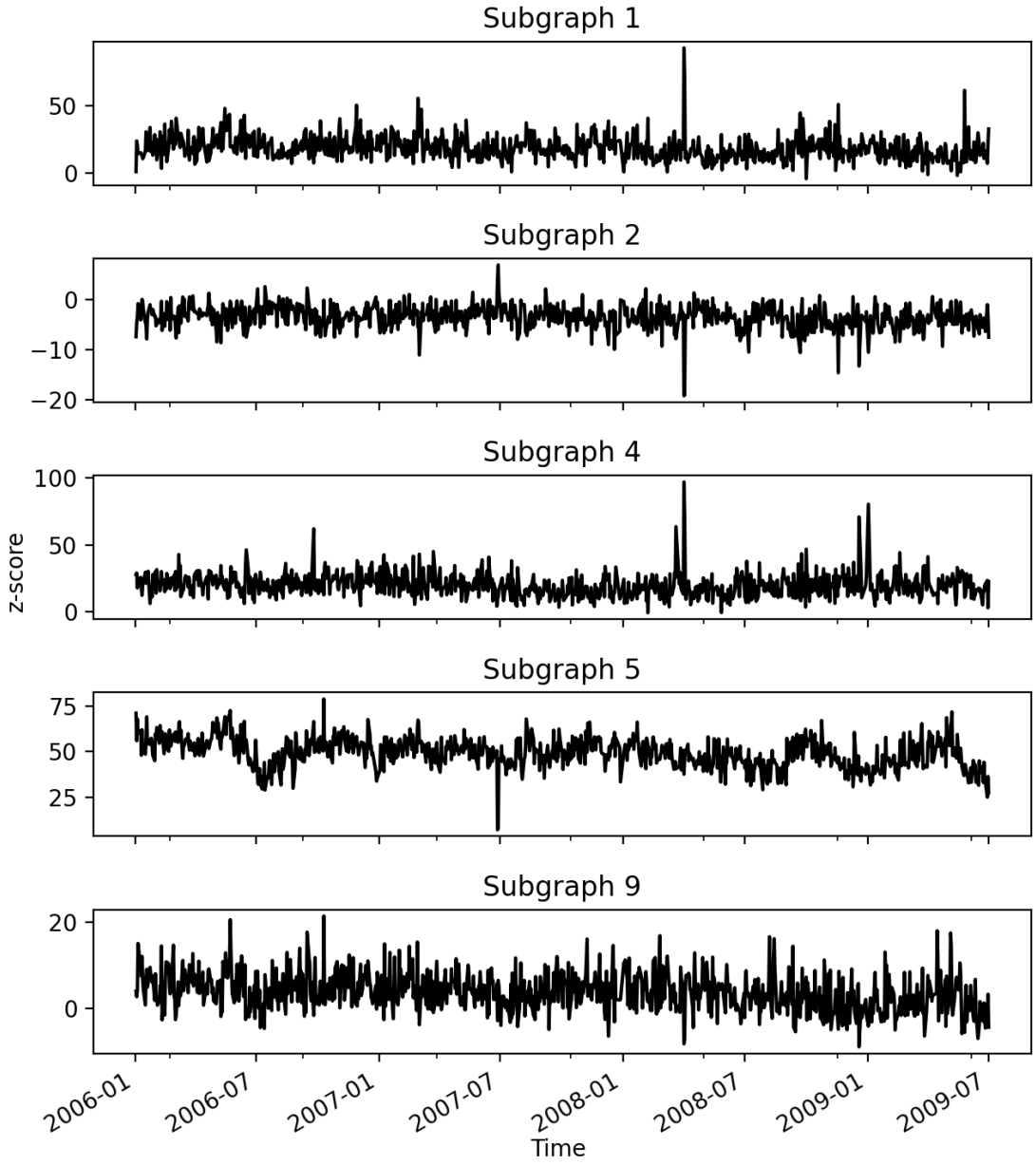


Figure 5.3. *Subgraph z-scores under directed random graph model for daily finance institute networks.*

In Figure 5.3 the z-scores of the subgraphs have short-term trends, times when the values spike higher or lower and at times change values differently than the other subgraphs. Subgraph 5 shows the clearest short-term trends in its z-scores as for example halfway through the year 2006 the values have downward trend around the same time the OMX Helsinki 25 index lowers also. Subgraph 5 z-scores can possibly correlate with the returns of the index even though the values do not have as large changes as the index. The other subgraphs do not have as clear short-term trends. There are times when the z-scores of different subgraphs have large changes in their values at different times. It is possible that these are errors in the data, but it is also possible that there is something else, which triggers the changes. For instance, subgraph 4 has relatively larger values in the second half of 2006, first half of 2008 and second half of 2008. These times may indicate moments when some securities were clearly preferred over others as subgraph 4 has two edges to a single security. In the first half of 2008 there is a time when subgraphs 1 and 4 have spikes up in their values, while subgraph 2 has a spike down. This could indicate a moment when finance institutes had similar strategies at the same time, and they reallocated their portfolios similarly from some securities to others. There is not a long lasting change at any time in the z-scores which would indicate a clear change in the way finance institutes invest.

Likewise under the directed random graph model, all of the subgraphs are significant under the directed configuration model as shown in Figure 5.4.

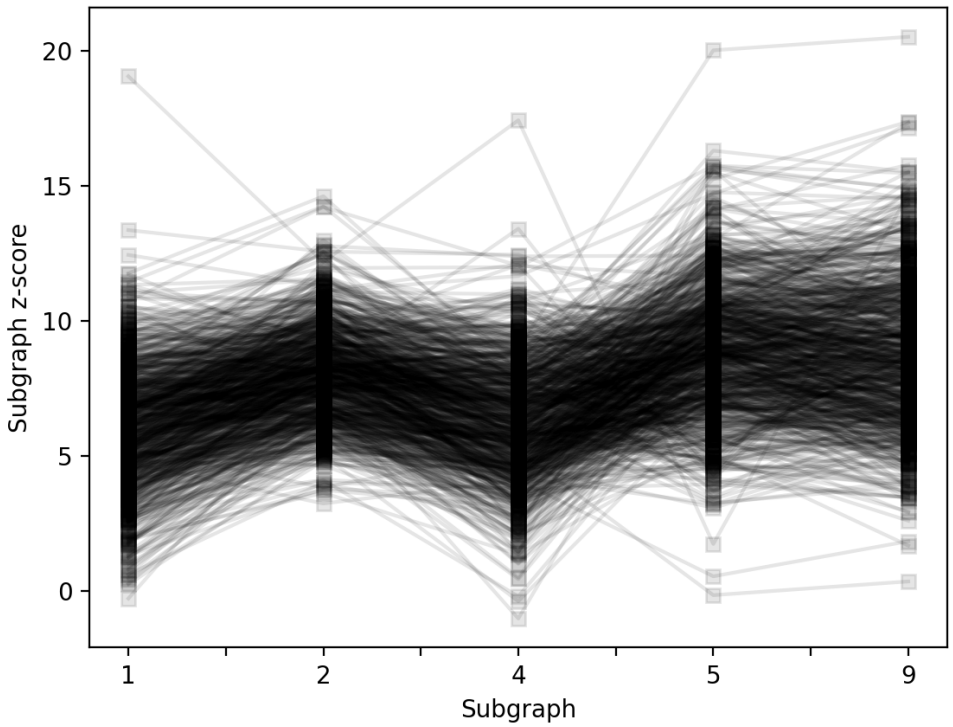


Figure 5.4. *Motif z-score profiles under directed configuration model for daily finance institute networks.*

Figure 5.4 shows that at almost all times the z-scores of all the subgraphs are over 2. If the subgraph counts were normally distributed, z-scores with absolute values over 2 would be unlikely. The model therefore underestimates the number of the subgraphs, and the in-degrees and out-degree sequences do not contain all the information about the subgraph counts. However, the directed configuration model has lower z-scores than the directed random graph model, so the degree sequences have relevant information.

In Figure 5.5 the z-scores under the directed configuration model have been plotted over time.

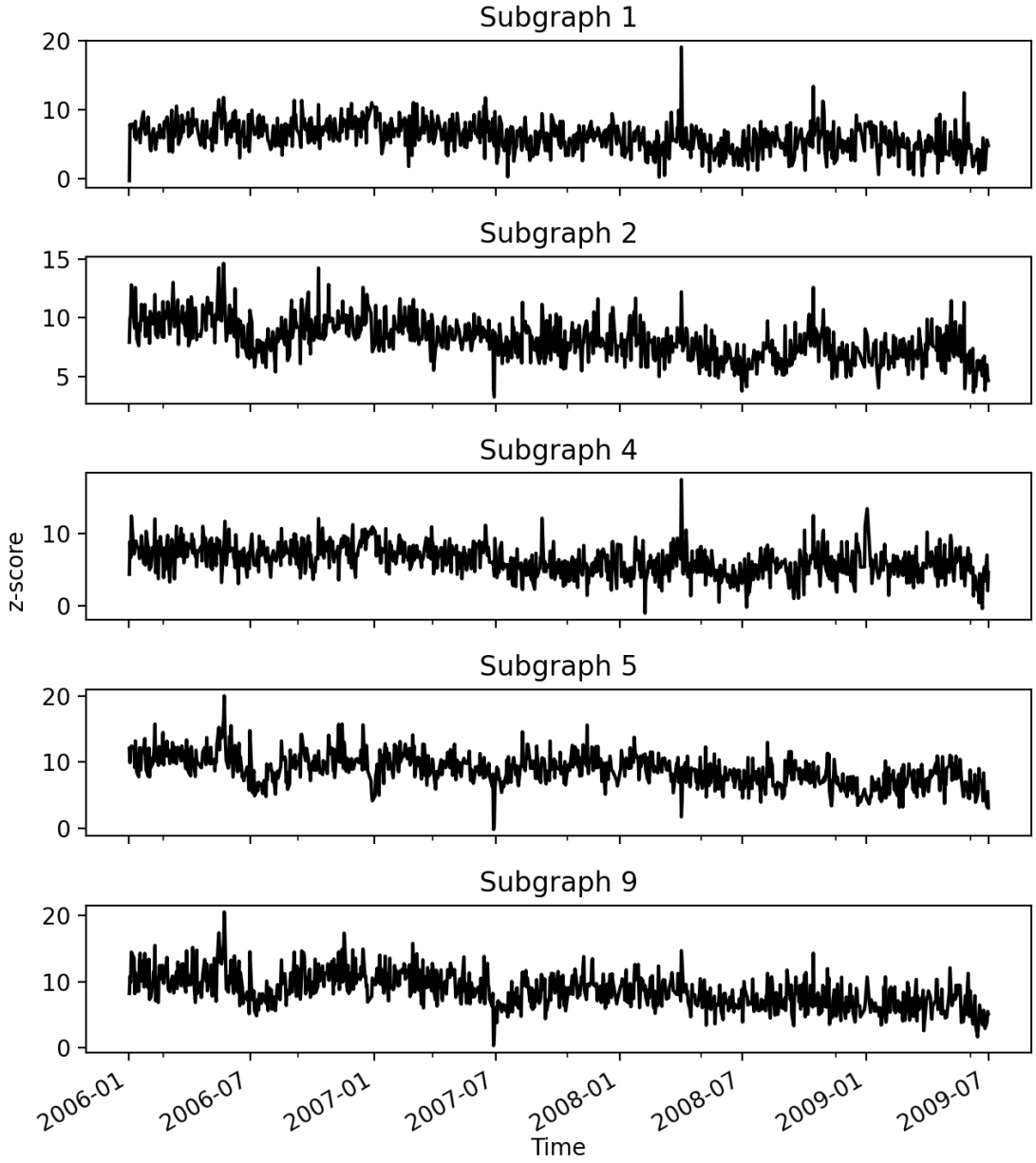


Figure 5.5. *Motif z-scores under directed configuration model for daily finance institute networks.*

Figure 5.5 shows more short trends in the z-scores and less spikes in the values than under the other model. Subgraphs 2, 5 and 9 have similar short-term trends at different times during the crisis. The short trend in the second half of 2006 is the clearest one. In the first half of 2008 there is a time when subgraphs 1 and 4 have upwards spikes in the z-scores while subgraphs 2 and 9 have relatively smaller upward spikes and subgraph 5 has a drop in its z-score. This could mean a clear time when finance institutions had similar strategies to invest in certain securities by divesting from others. It is interesting that the subgraph 2 has a spike upwards as it had spike downwards when using the other model. It could be that the upwards spike follows from finance institutes investing in more securities, but the observation is hard to explain. Moreover, the z-scores have

similar small negative long-term trends over the time period which means the total number of subgraphs is lower than earlier during the crisis which could mean less transactions. When the values decrease and stay over 0, the subgraph counts of the observed network are closer to the subgraph counts than if the connections were randomized but the in-degrees and out-degrees were kept the same. When the values increase, the observed network has a configuration of the links so that there are more subgraphs than the in-degree and out-degree sequences would explain. Therefore, when the z-scores are high the observed connections are more special than when the z-scores are low.

Weekly network motif counts are visualized in Figure 5.6

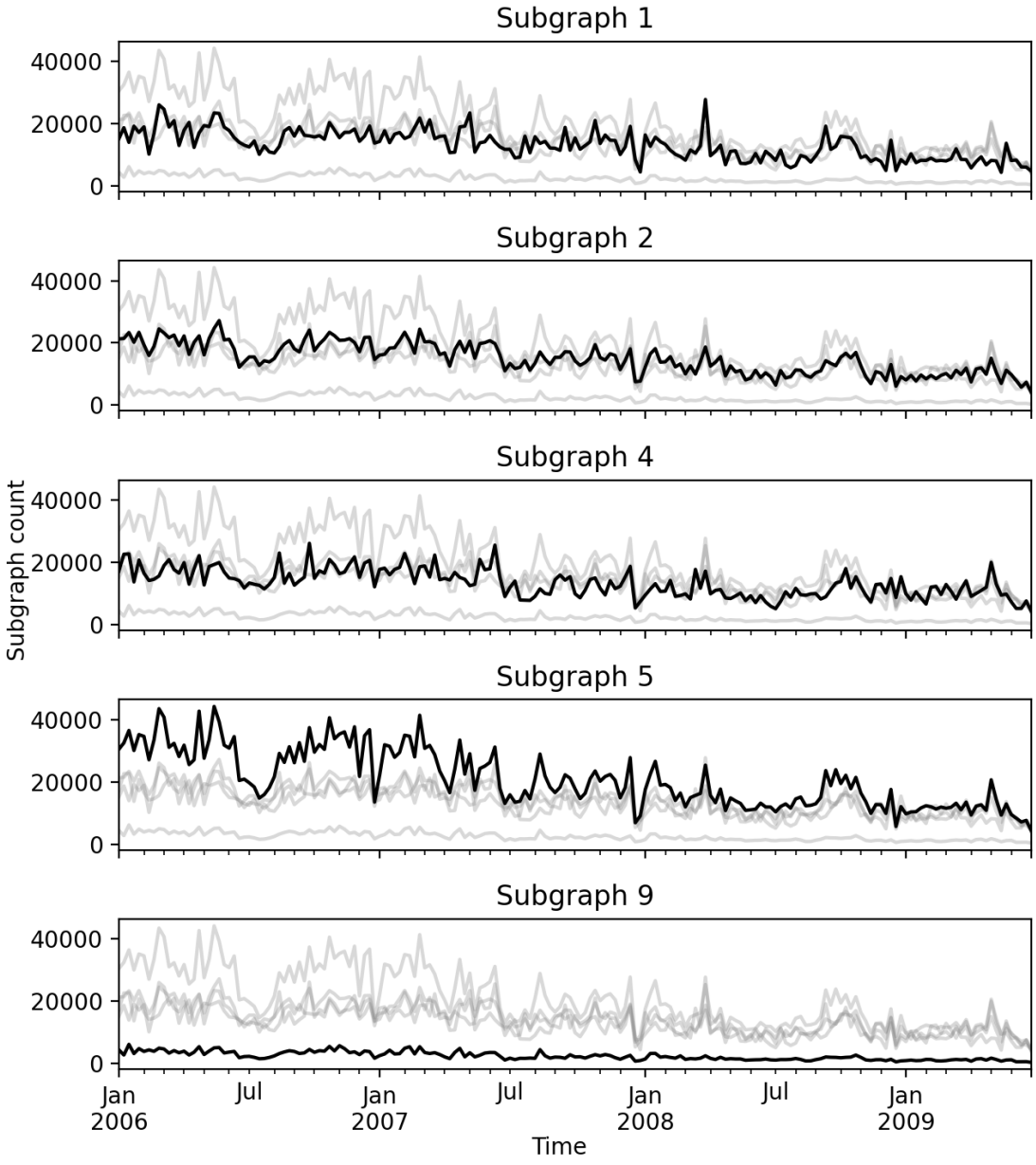


Figure 5.6. *Subgraph counts of weekly finance institute networks.*

The weekly subgraph counts of finance institutes are more stable in Figure 5.6 than the

daily counts in Figure 5.1 by a small amount. Subgraph 5 shows the largest changes and the largest downward trend. There is also a negative trend in the counts as the levels are lower in the second half of the time series than the first half. Calculating the z-scores under directed random graph model shows that the weekly values have similar z-score profile shape as the daily z-scores, but the z-scores are larger. The z-score profiles are shown in Figure 5.7.

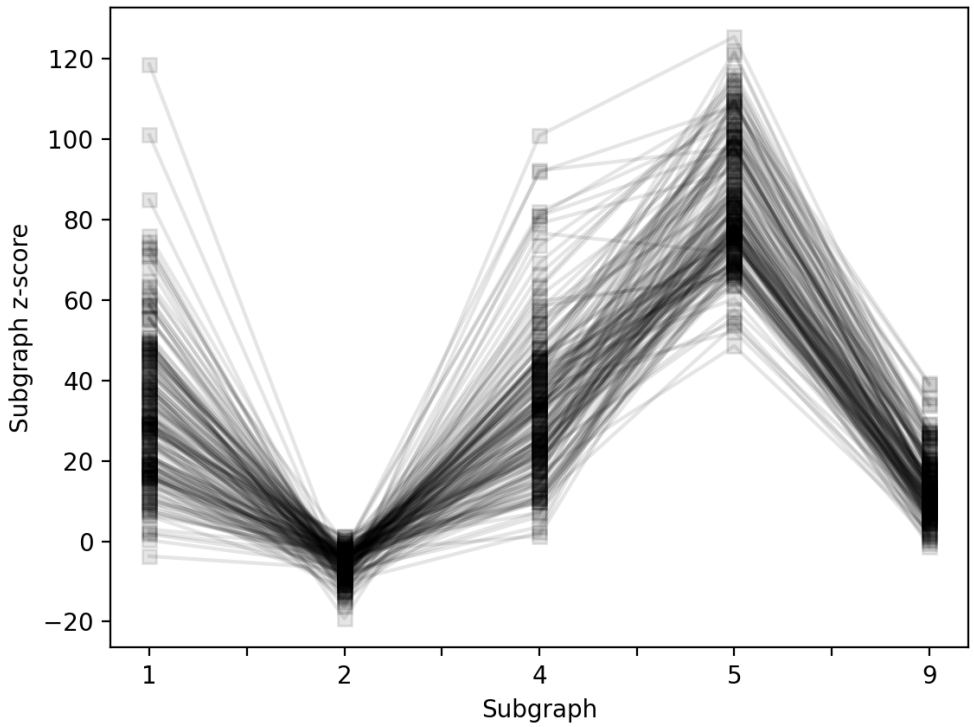


Figure 5.7. *Subgraph z-scores under directed random graph model for weekly finance institute networks.*

All of the subgraphs have large z-scores which means the subgraphs are statistically significant. Subgraph 2 is closest to being statistically insignificant, but its values are not within reasonable distance from 0 to determine the subgraph count to be insignificant. The z-scores of the subgraphs under the directed random graph model are shown in Figure 5.8.

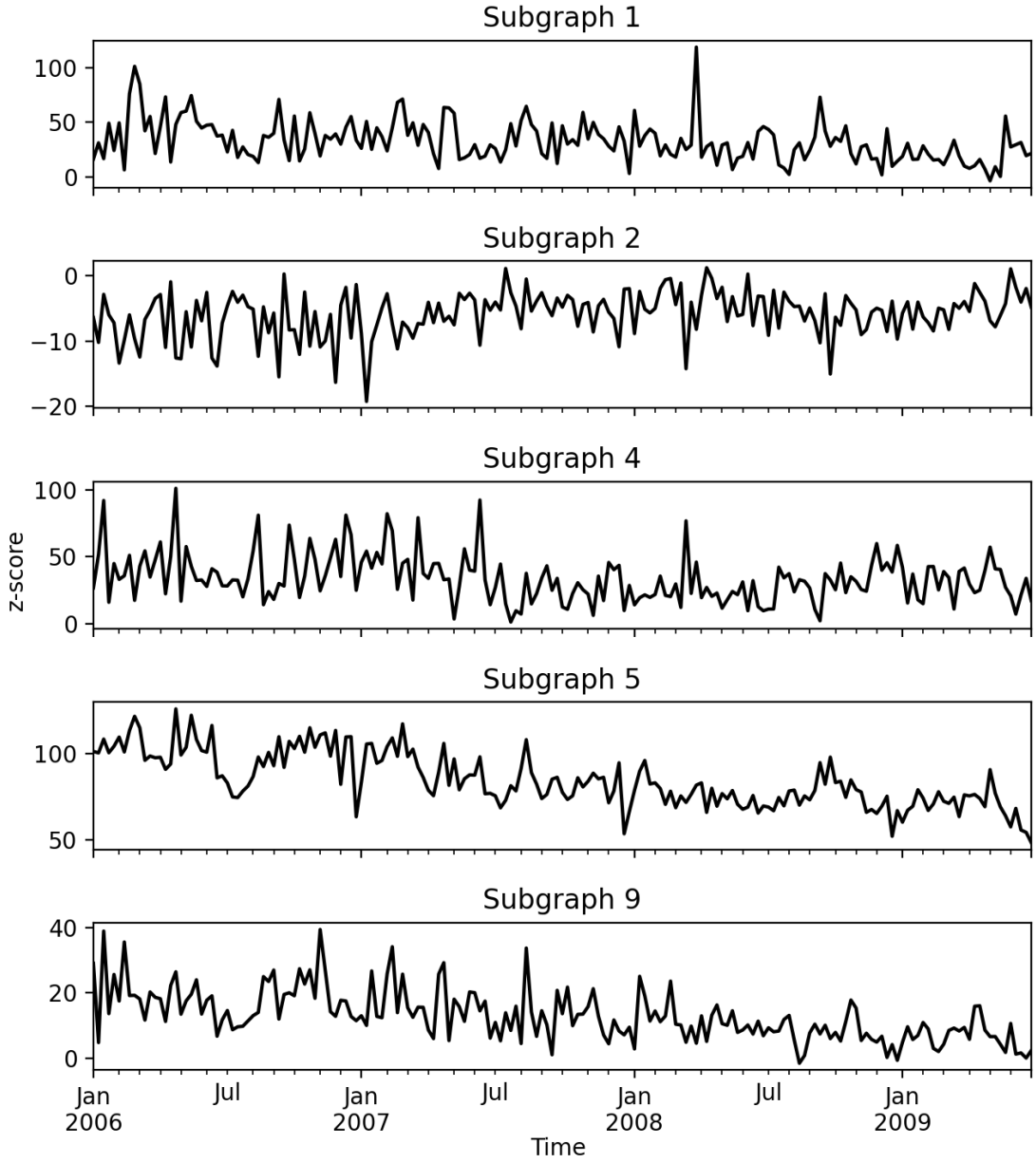


Figure 5.8. *Motif z-scores under directed random graph model for weekly finance institute networks.*

Subgraphs 5 and 9 have downward trends in their z-scores during the crisis, there are larger deviations in the z-scores in general during the first half of the crisis, and there are times when the values have large changes at the same time in Figure 5.8. Subgraphs 5 and 9 have downward trends during the crisis but the other subgraphs have no trends in their z-scores which could indicate similar strategies among the finance institutes as there are less 3 edge subgraphs. The z-scores of subgraphs 1, 2 and 4 do not have clear trends in their values, but z-scores of subgraphs 2 and 4 have large deviations in the first half of the crisis. Large changes in the z-scores can mean that the market is changing a lot from week to week, and it is hard to define a regime during large deviations. There are also times when the z-scores have similar large changes in them at the same time. For

example, in the first half of 2008 subgraphs 1, 2 and 4 have large changes in them at the same time, and that time can indicate a big reallocation event. In general, the z-scores seem to correlate with each other, but at times the values change differently.

Changing the model to the directed configuration model for weekly finance institute networks has the same effect as for daily networks. Figure 5.9 shows the z-score profiles of the subgraphs in weekly finance institute networks.

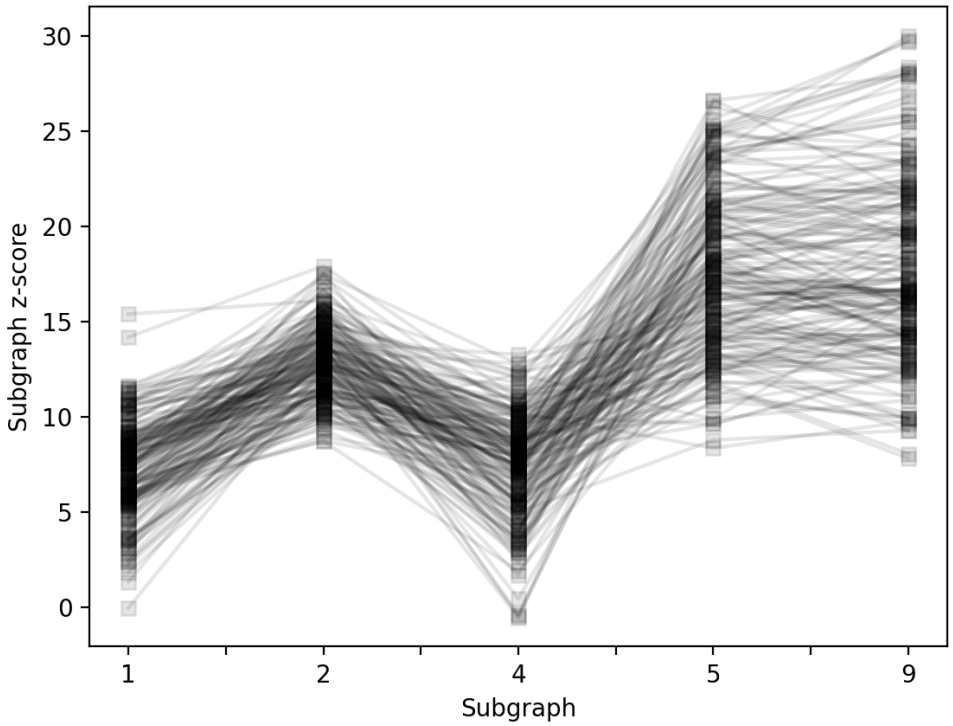


Figure 5.9. *Motif z-scores under directed configuration model for weekly finance institute networks.*

In Figure 5.9 all of the subgraphs are statistically significant. Compared to the daily networks the subgraphs 5 and 9 have larger z-scores which means that over a week finance institutes have more different transactions. The z-scores are lower for the subgraphs except for the subgraph 2 when using the directed configuration model as compared to the directed random graph model. Therefore, the degree sequences do contain important information about the subgraph counts, and a single probability for all of the edges does not work as well as using node specific probabilities. In other words, the weekly networks seem to not have a totally random structure, and there are local structures, which may be preferences of some securities over others for example. However, the directed configuration model has more parameters, which can fit the data better, and thus lower the z-scores. The z-scores of the subgraphs have been plotted over time in Figure 5.9.

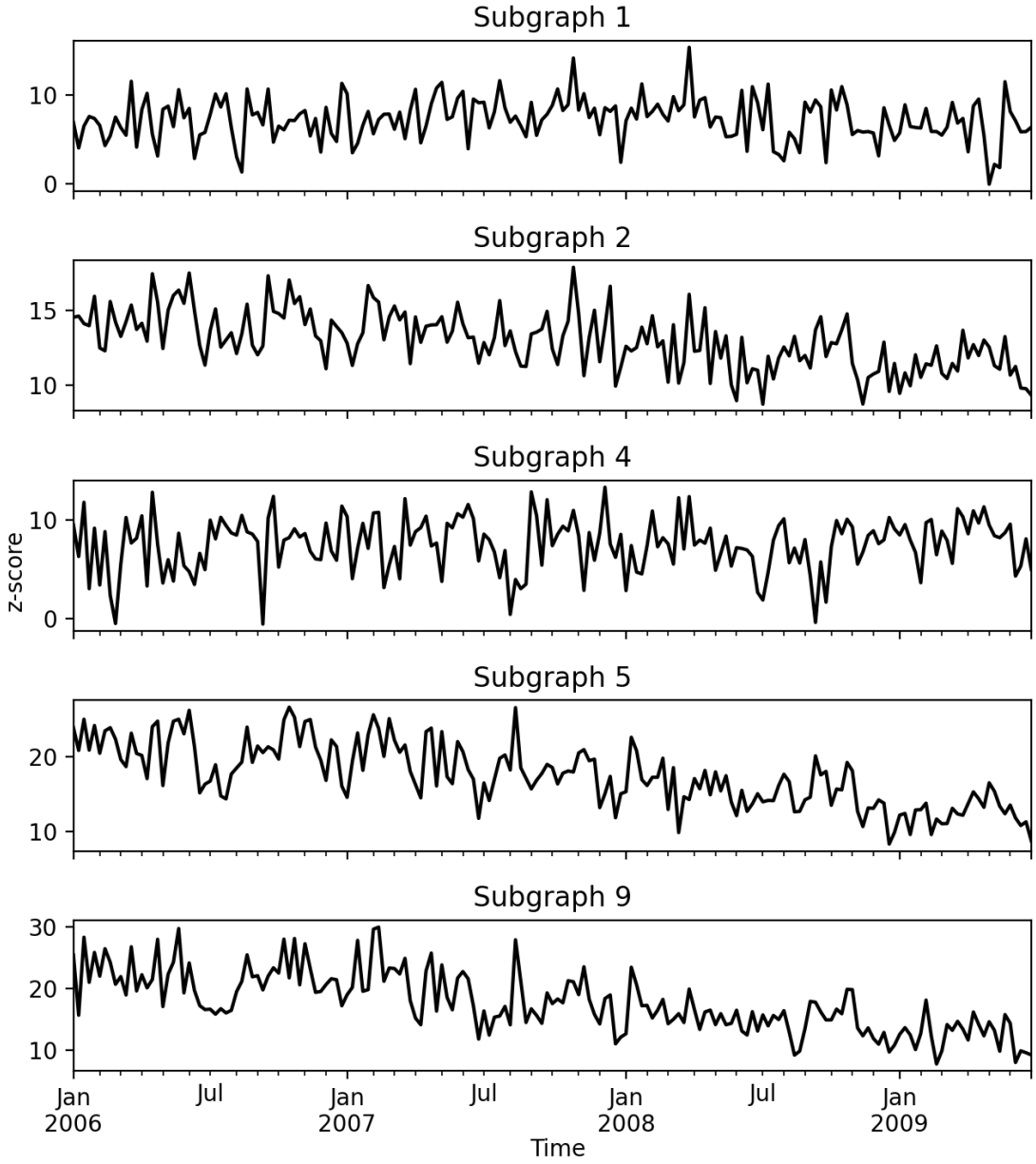


Figure 5.10. *Subgraph z-scores under directed configuration model for weekly finance institute networks.*

In Figure 5.9 the z-scores of subgraphs 2, 5 and 9 have negative trends while the z-scores of subgraphs 1 and 4 oscillate about a central value. This means that the level of agreement about some securities stays the same while there is less disagreement. This reinforces the prior possibility that the finance institutes have more similar strategies when the crisis advances because subgraphs 1 and 4 are clear decisions to divest from an asset or to invest in an asset while subgraphs 2, 5 and 9 show disagreement about an asset. Activity in the markets may also have been the reason for the negative trends or trading focused on larger companies while small companies had less activity which means less 3 edge subgraphs. Surprisingly there are not as large sudden changes in the z-scores as before even though there are some changes. The sudden large change in

the first half of 2008 is present, but the changes are not as large as before, and without prior information the change would be hard to notice.

5.2.2 Household networks

Subgraph counts for daily household networks are shown in Figure 5.11.

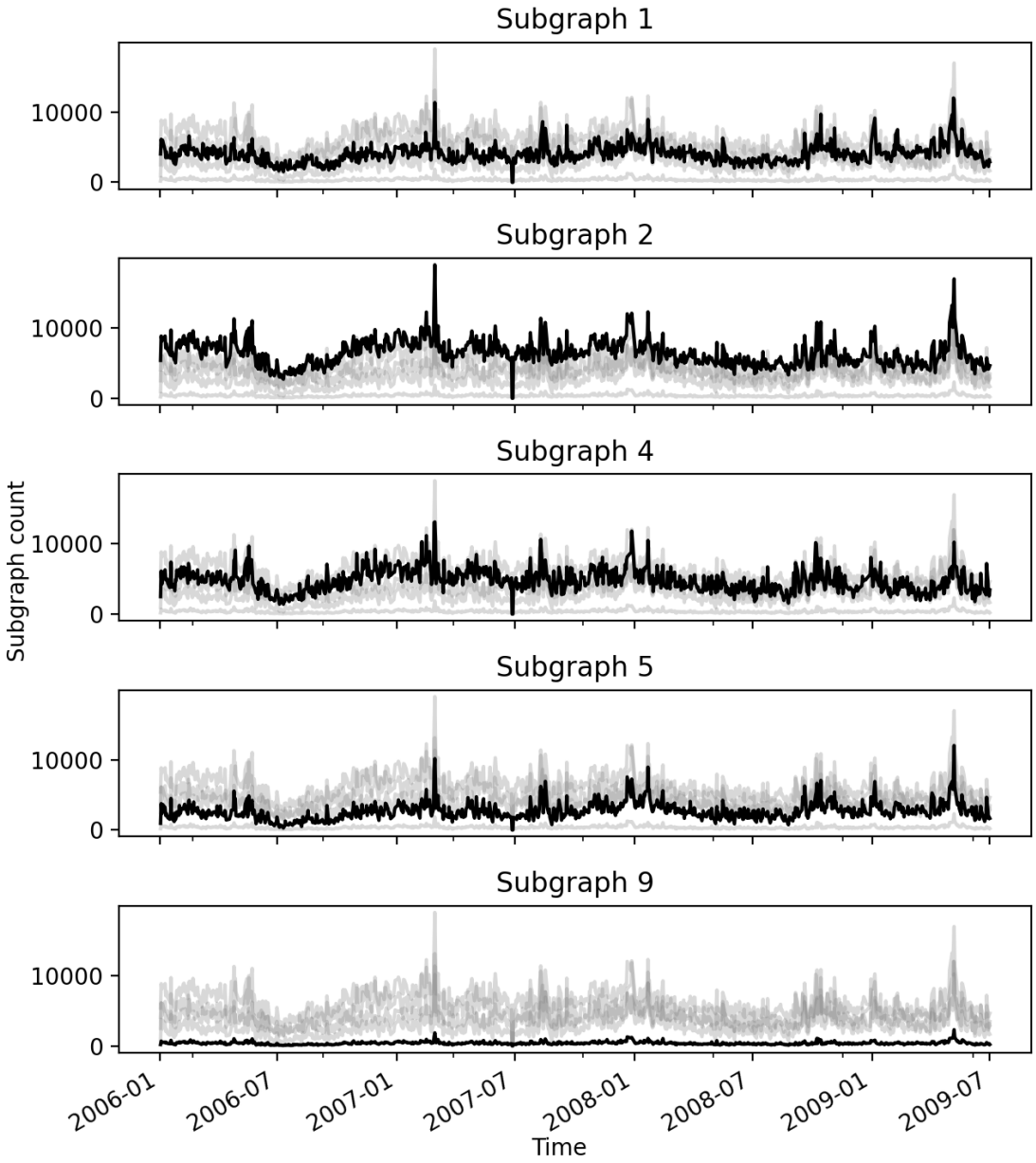


Figure 5.11. *Subgraph counts of daily household networks.*

In Figure 5.11 the counts have short trends over the time period. There are days when the counts spike upwards by a large number and then come down quickly. These spikes mean higher activity by the households in the market which could be because of volatility in the markets. There are two large spikes: one in the first half of 2007 and one in the first

half of 2009. Contrary to the finance institutes, there are no large spikes upward in first half of 2006 and first half of 2008, and some of the trends upward are at different times. This could mean that finance institutes and households react to different information, and they adjust their positions at different times while sometimes they act at the same time.

To analyze the changes in the daily household network structure further, z-score profiles have been calculated for the subgraphs under a directed random graph model in Figure 5.12.

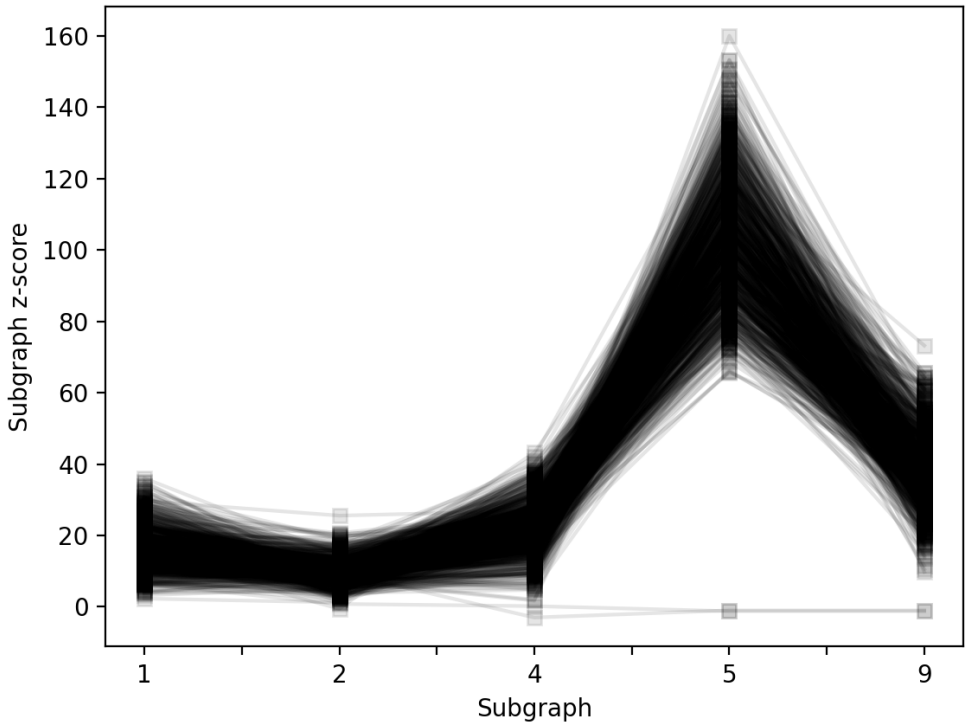


Figure 5.12. *Subgraph z-score profiles under directed random graph model for daily household networks.*

Figure 5.12 shows that the subgraphs 5 and 9 have the largest z-scores while subgraphs 1, 2 and 4 have lower z-scores. Subgraphs 2, 5 and 9 of daily household networks have higher z-scores than daily finance institute networks. Daily finance institute networks had at times larger z-scores for subgraphs 1 and 4 which means there are days when finance institutes concentrate their actions around certain securities, and the actions are more similar between different finance institutes as there are less subgraphs 5 and 9.

The z-scores of the subgraphs are plotted over time in Figure 5.13.

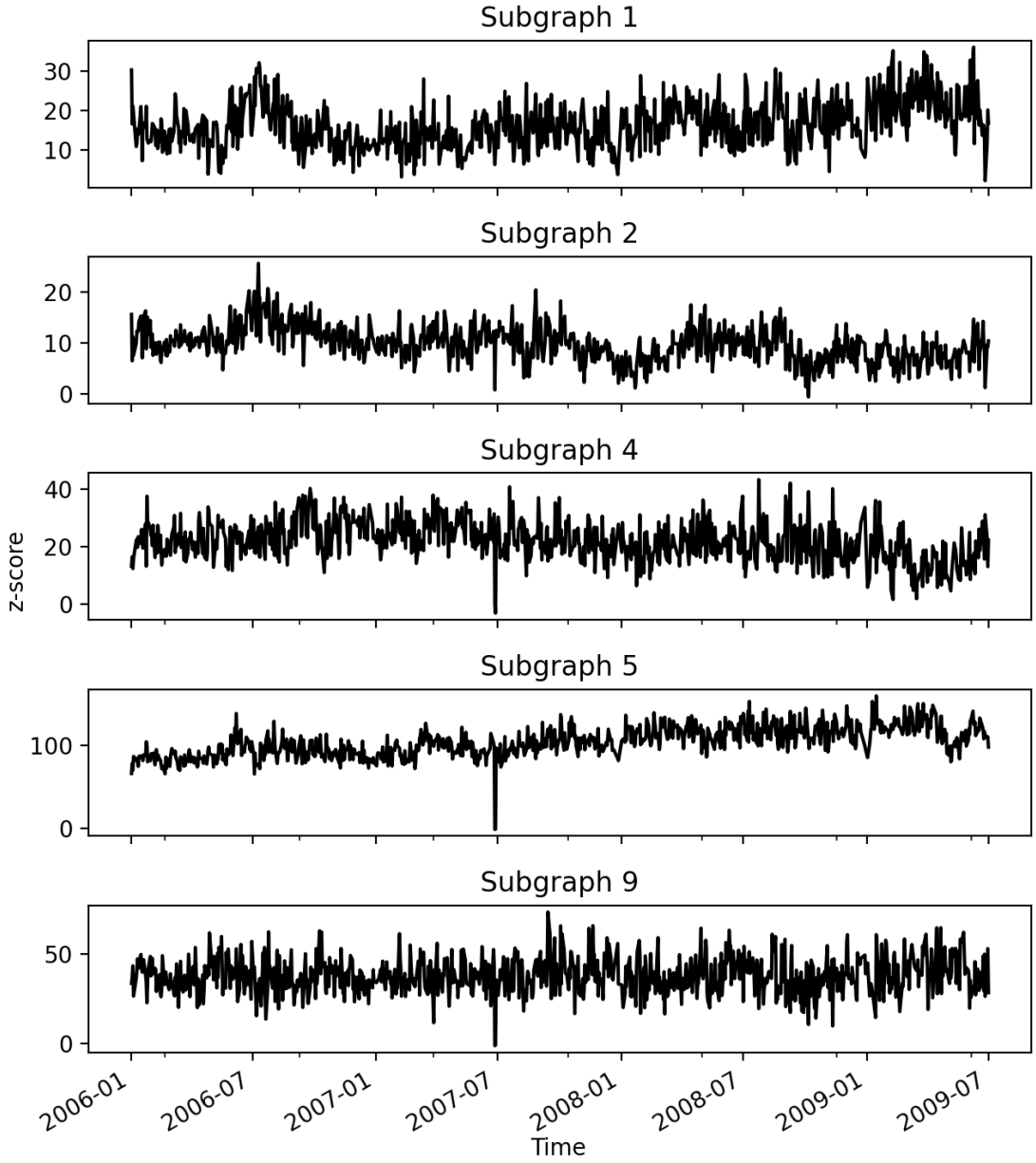


Figure 5.13. *Subgraph z-scores under directed random graph model for daily household networks.*

Figure 5.13 shows that subgraph 5 has an upwards trend during the whole time period, subgraphs 1, 2 and 4 have short-term trends in their z-scores, subgraph 9 does not have either clear long-term or short-term trend, and the z-scores of subgraph 4 may follow the z-scores of subgraph 1 with a lag. The increase in the z-scores of subgraph 5 during the crisis indicate that there are more cases in which 1 security is seen as a bad investment, 1 is seen as a good investment, and for the last 1 there is no clear consensus. This could mean that in the second half of the crisis households had more similar ideas about which securities are good and bad investments than in the first half of the crisis. It could also be that there were simply more transactions or some other explanation for the long-term trend, more in-depth analysis would be needed to draw better conclusions. Interestingly,

the z-scores of subgraph 4 seem to follow the values of subgraph 1 to some degree. It could be that households sell securities but they do not know where to invest yet, so they invest in some familiar securities or draw the money out of the markets and reallocate later. For daily finance institute networks the z-scores of subgraphs 1 and 4 had changes during the same days most of the time, so finance institutes are better in knowing where to invest the money when they sell securities. In second half of 2007 there is a spike down in all of the z-scores, but it is likely an error in the data as there are no other days with changes as large as this one. Based on these observations daily household networks have different structure than daily finance institute networks.

In Figure 5.14 the model has been changed to the directed configuration model and the z-score profiles have been calculated.

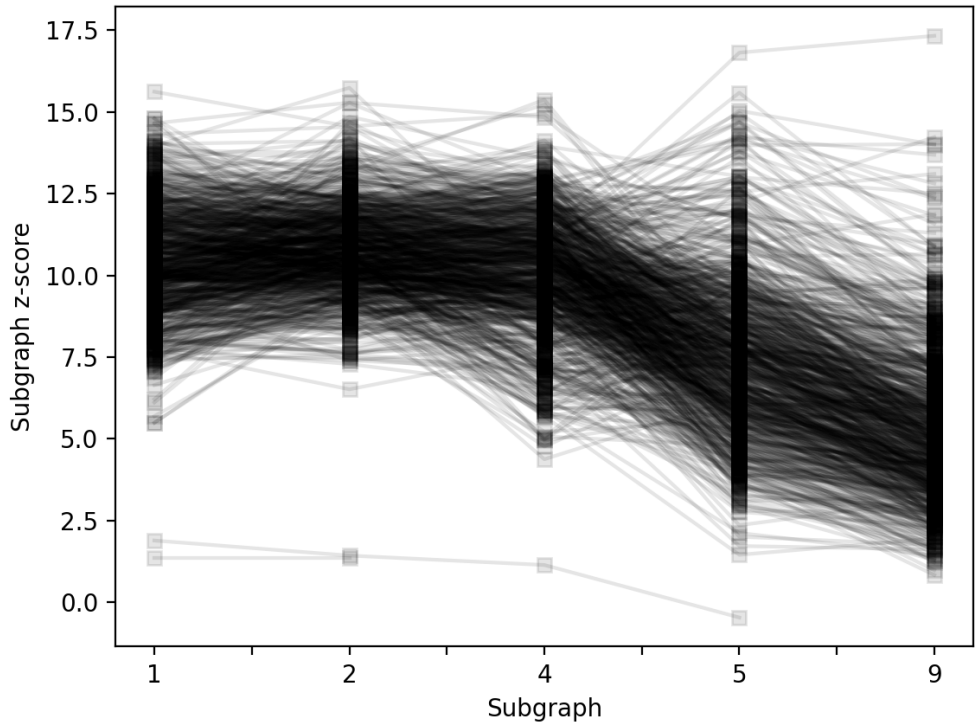


Figure 5.14. *Subgraph z-score profiles under directed configuration model for daily household networks.*

In Figure 5.14 the z-scores of the subgraphs are lower in absolute values which means the configuration model expects the subgraph counts to be closer to the observed values than the random graph model. In contrast to daily finance institute networks, the z-scores of subgraphs 5 and 9 are lower than the z-scores of subgraphs 1, 2 and 4. This means that the degree sequences of household networks are more important for the counts of subgraphs 5 and 9 than in the case of daily finance institute networks. It could be that for households there are more regular money flows between some securities, but finding an

explanation for this observation is hard and it can simply be a feature of the household networks.

Daily household network z-scores under the directed configuration model are visualized over time in Figure 5.15.

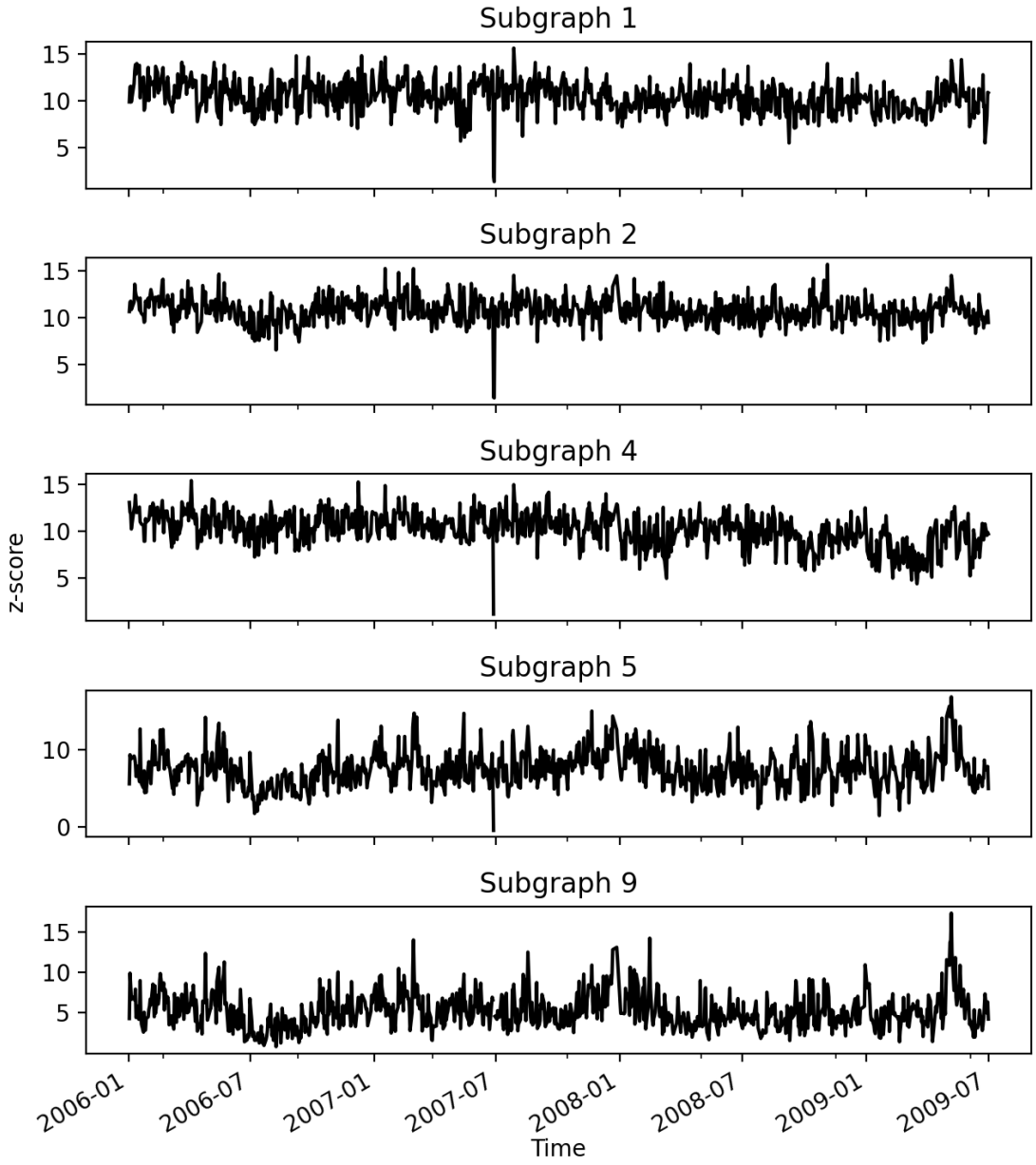


Figure 5.15. *Subgraph z-scores under directed configuration model for daily household networks.*

Figure 5.5 shows that subgraphs 1 and 2 do not have as clear short-term trends as with the other model, there is no long-term trend in z-scores of subgraph 5, there are short-term trends in the z-scores of subgraphs 5 and 9, and the importance of the degree sequences starts to change in the first half of 2009. The change of the null model from directed random graph model to directed configuration model had clear effects on the

z-scores as some trends vanished and new ones appeared. This means that the degree sequences of the daily household networks contain important information about the subgraph counts. The importance of the degree sequences start to change in the beginning of 2009 as all of the z-scores start to have larger changes in them during that time.

The subgraph counts of weekly household networks are plotted in Figure 5.16.

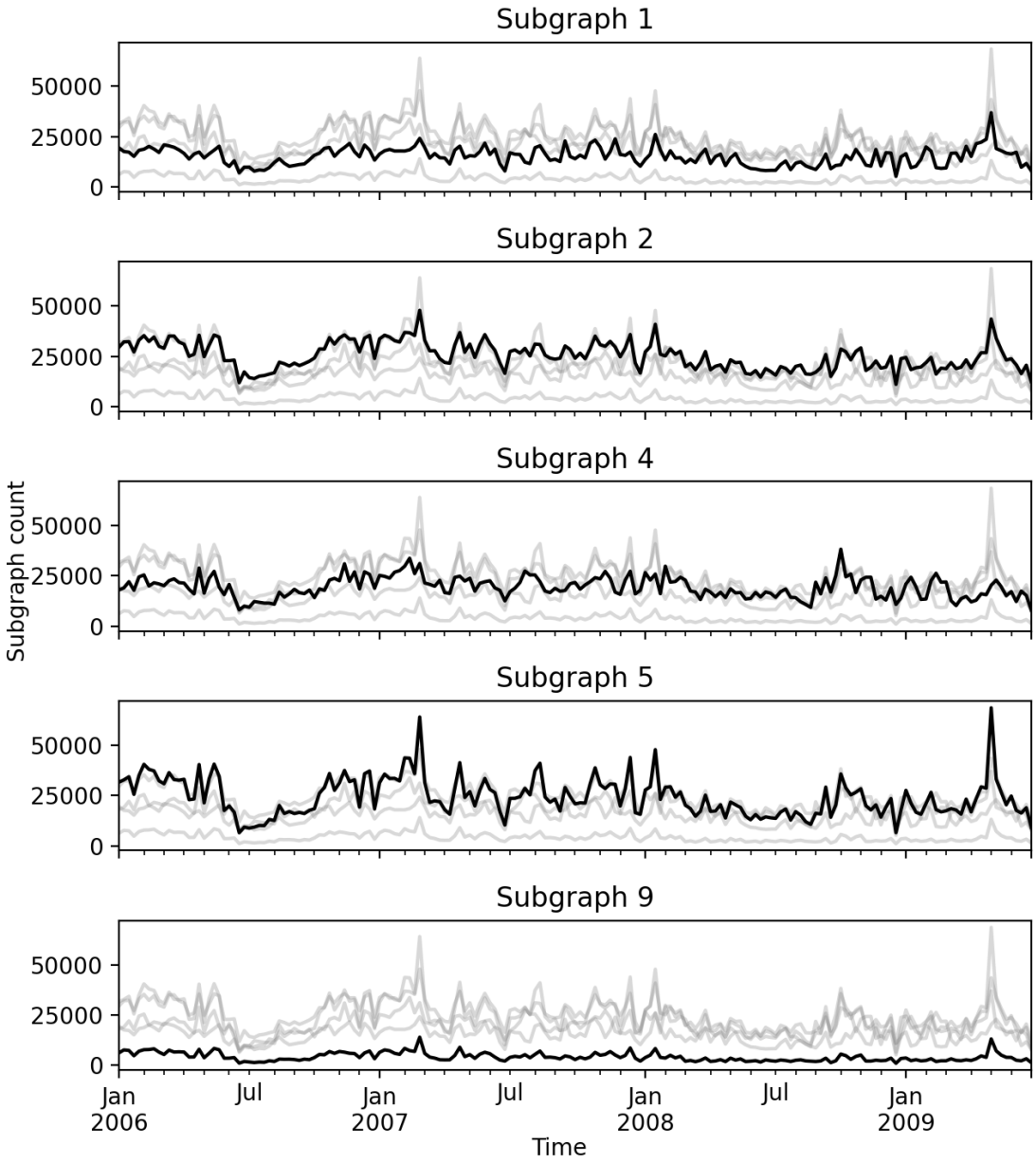


Figure 5.16. *Subgraph counts of weekly household networks.*

In Figure 5.16 there are clear highs and lows for all of the subgraphs. The counts spike in the first half of 2007, first half of 2008 and first half of 2009 while at other times the counts have high and low values at different times. The z-score profiles under the directed random graph model are plotted in Figure 5.17.

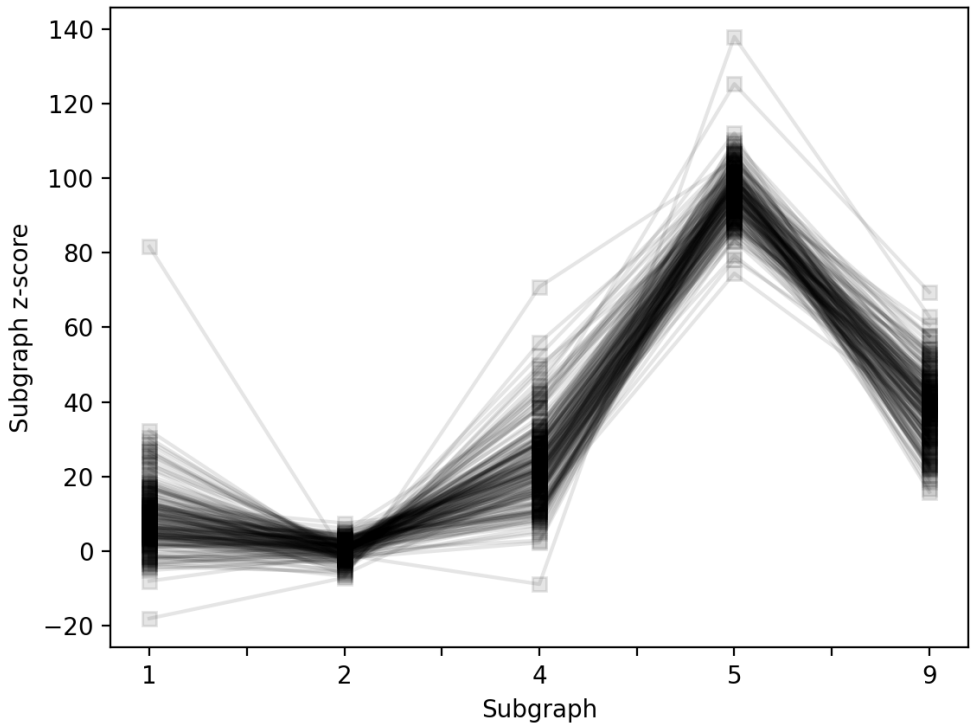


Figure 5.17. *Subgraph z-score profiles under directed random graph model for weekly household networks.*

Weekly household network profiles in Figure 5.17 have the same shape as the other three network series when using the directed random graph model: subgraph 5 has the highest z-scores, subgraph 2 has the lowest z-scores, and the other three are between these two values. Contrary to the daily household networks, there are times when subgraphs 1 and 4 have larger values than normally. The z-scores of the subgraphs during the crisis are plotted in Figure 5.18.

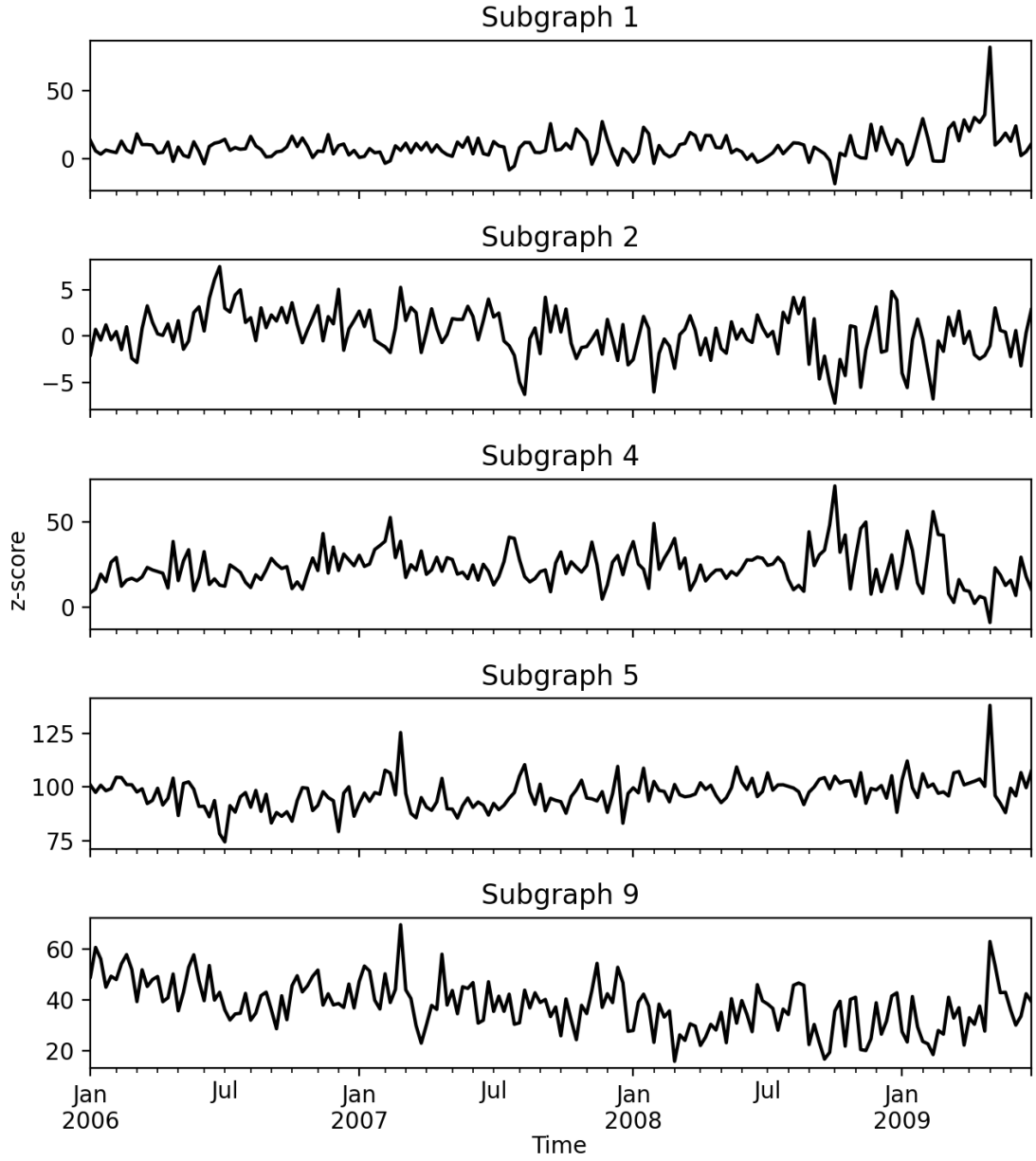


Figure 5.18. *Motif z-scores under directed random graph model for weekly household networks.*

In Figure 5.18 the z-scores of weekly household networks do not have as clear short-term or long-term trends as in other cases, but there are a few times when the z-scores have larger or lower z-scores than usually. The z-scores of subgraphs 1, 2, 4 and 5 seem to have constant values with random noise added in. Subgraph 9 has some short-term trends in its values as the values decrease in the years 2006 and 2007. The z-scores of subgraph 9 are larger in the first half of the time series than in the second half. Additionally, at a time in the first half of 2006 subgraph 2 has larger than usual z-scores, and in the second half of 2007 the z-scores are lower. In 2006 subgraph 5 has also lower values at the same time, but other z-scores do not seem to have large changes at the same time. For some reason in 2006 there were more 2 edge interactions than 3 as subgraph 2 has

the same structure as subgraph 5, but with one less edge. Other large changes in other subgraphs happened in first half of 2007, second half of 2008 and first half of 2009. In the first half of 2007 subgraphs 2, 4, 5 and 9 had large increases in their z-scores which could mean more activity in the markets because of volatility. Households could have thought some securities were good opportunities to buy as the z-scores of 4 were larger, but the z-scores of subgraph 1 stayed relatively constant. In the second half of 2008 the z-scores of subgraph 2 had a large decrease and the z-scores of subgraph 4 had a large increase. This could mean households had more similar ideas about which securities were good buying opportunities, and less disagreement as subgraph 2 has a security, which is being bought and sold by households. It is also possible that households took money away from the markets as the edges of subgraphs 4 could point at the balance node, and the OMX Helsinki 25 index had a large decrease at the same time. The first half of 2009 has the largest increase during the crisis in the z-scores of subgraphs 1, 5 and 9 which could mean households bought a lot of securities and reallocated their portfolios to better match future prospects of the markets. Around the same time signs about the end of the crisis had started to appear (Federal Reserve 2009a; Federal Reserve 2009b; McIntyre 2009) and the OMX Helsinki 25 index had increased.

The weekly household subgraph z-score profiles under the directed configuration model have been plotted in Figure 5.19.

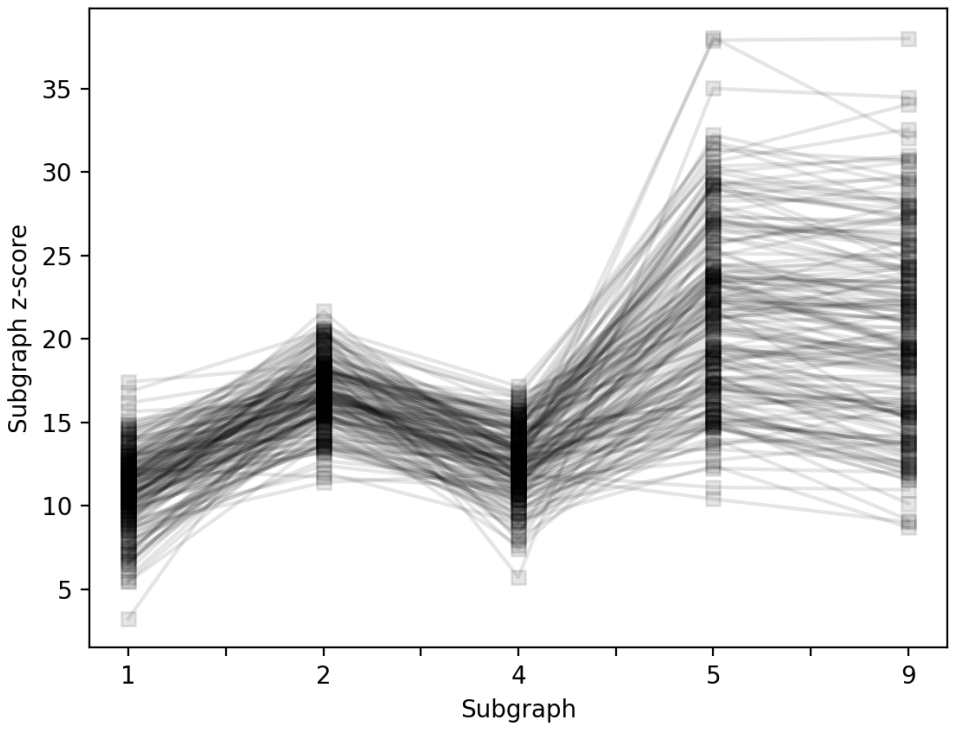


Figure 5.19. *Subgraph z-score profiles under directed configuration model for weekly household networks.*

All of the weekly household subgraphs are significant according to the directed configuration model in Figure 5.19 because the values are over 5 at almost all times. In contrast to the daily household z-score profiles under the same model, the subgraphs 5 and 9 have the highest z-scores. Thus, the timescale matters, and it affects the structure of the observed networks. The z-scores have been plotted over time in Figure 5.20.

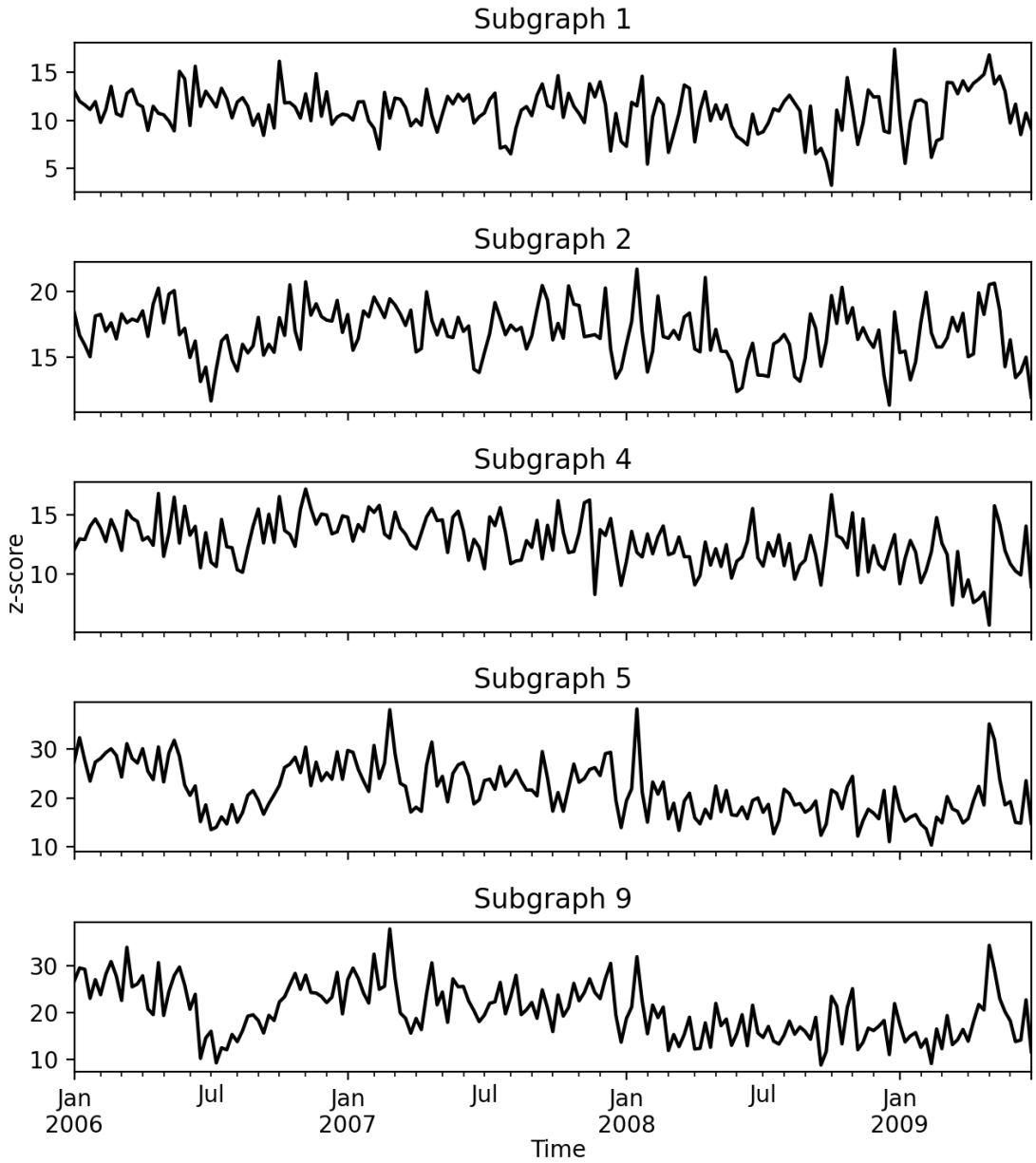


Figure 5.20. *Subgraph z-scores under directed configuration model for weekly household networks.*

In Figure 5.20 subgraphs 2, 5 and 9 have the largest changes during the crisis, whereas subgraphs 1 and 4 have the most stable values during the crisis. Subgraphs 2, 5 and 9 have similar short-term trends in their z-scores as they have a low values from June to July in 2006, while later from the first half of 2008 to April 2009 the values decrease,

but subgraph 2 does not follow the trend as closely. The decrease in the z-scores of subgraphs 2, 5 and 9 while the z-scores of subgraphs 1 and 4 stay relatively the same may mean more similar actions between households. The large increases in z-scores of subgraphs 5 and 9 indicate times when there is less consensus in the markets.

5.2.3 Summary of the motif analysis

To conclude the network motif analysis, the means and the standard deviations of all the z-scores have been gathered in Tables 5.12 and 5.13 under both null models.

Table 5.12. *Means and standard deviations (μ, σ) of subgraph z-scores under directed random graph model.*

	Finance institutes		Households	
Subgraph	Daily	Weekly	Daily	Weekly
1	$(19, 9.1)$	$(33, 19)$	$(16, 5.8)$	$(8.9, 9.5)$
2	$(-3.4, 2.4)$	$(-5.9, 3.5)$	$(9.7, 3.4)$	$(0.30, 2.5)$
4	$(20, 9.0)$	$(34, 18)$	$(22, 7.0)$	$(23, 11)$
5	$(49, 8.3)$	$(85, 16)$	$(100, 18)$	$(97, 7.3)$
9	$(4.0, 4.5)$	$(13, 7.8)$	$(38, 10)$	$(38, 9.7)$

Table 5.13. *Means and standard deviations (μ, σ) of subgraph z-scores under directed configuration model.*

	Finance institutes		Households	
Subgraph	Daily	Weekly	Daily	Weekly
1	$(6.1, 2.2)$	$(7.3, 2.4)$	$(10, 1.6)$	$(11, 2.3)$
2	$(8.1, 1.7)$	$(13, 1.9)$	$(11, 1.4)$	$(16, 2.0)$
4	$(6.3, 2.2)$	$(7.5, 2.7)$	$(10, 1.8)$	$(13, 2.1)$
5	$(8.9, 2.4)$	$(18, 4.2)$	$(7.5, 2.4)$	$(22, 5.4)$
9	$(8.7, 2.6)$	$(18, 4.9)$	$(5.3, 2.2)$	$(20, 5.8)$

Table 5.12 shows that all the networks have similar z-scores relative to each other: subgraphs 5 have the highest z-scores, subgraphs 2 have the lowest z-scores, subgraphs 1 and 4 are in between the two, and subgraphs 9 have lower values than 1 or 4. Only in the case of daily finance institute networks the number of subgraphs 2 is lower than what the directed random graph model would expect. It is possible that finance institutes have more similar opinions about which securities are good and bad as the z-scores of subgraphs 2, 5 and 9 are higher for household networks than finance institutes, and the z-scores of subgraphs 1 and 4 are higher for finance institutes. In some cases, there

could also be a small difference between the subgraphs 5 and 9 as they differ by only the direction of one edge, which may have small weight and the direction is by chance. This effect of a small weight could also contribute to the overrepresentation of the subgraph 5 as small and large money flows are analyzed as equal, and thus small transactions change the structure. Additionally, highly connected nodes may proportionally a large number of subgraphs, but that is a possibility for the network structure. Overall, finance institute networks and household networks have similar general structure, but households have more complex structure based on the z-scores under directed random graph model.

The z-scores under the directed configuration model are lower in general even though there are a few exceptions in Table 5.13. The average z-scores of the other network series follow the same shape: subgraphs 1 and 4 have the lowest values, subgraph 2 has the second lowest value and subgraphs 5 and 9 have the highest values. For daily household networks this is not true as the z-scores of subgraphs 5 and 9 have the lowest values. Therefore, the daily household networks have a different structure than the other networks. The directed configuration model uses the same degree distributions as the given network which means the importance of degree distributions of daily household networks are different than for the other networks. The degree distributions of daily household networks contain information, which explains the number of subgraphs 5 and 9 better.

6. DISCUSSION

In the centrality ranking analysis of the money flow networks the most central nodes were the same more often than expected, the bottom securities were often random and in higher percentiles there was structure in the rankings. The nodes with the largest centrality scores were large companies based on their market capitalization in addition to the balance node which can be thought to belong among large companies. The balance node contains cash and all other securities, which cannot be observed from the data, which means it is likely the largest node. The bottom securities based on the centrality measures were random in all network series except for the weekly household networks, where the bottom securities were mainly small in market capitalization. The results are not surprising, large securities are traded more and therefore they have more edges because of the way the networks are constructed, which leads to higher centrality measures. For a net sold security during a day or a week, depending on if the networks are daily or weekly networks, an edge goes to every net bought security during the same period which means more edges for securities, which are traded actively by many investors as there are more net sold and bought security combinations. The rankings also had structure in between the tails because there were more securities in the larger percentiles of the rankings. Overall, these results indicate that there are securities which are more central than others, but there are no certain securities which are always among the lowest ranked securities except for the daily household networks.

Usefulness of the centrality rankings based on the used way of constructing the networks are limited because the centrality scores do not exhibit a lot of additional information which could be used for further analysis as compared to simply using the market capitalization, trading activity, some other size metric of the securities or the money flow technical indicator defined by Bennett and Sias (2001). It could be possible to use the individual centrality score time series of the securities to look for a relationship with the returns or volatilities of the securities, but the results may be poor in predictive power or close to the results obtained by using market capitalization or the money flows as defined by Bennett and Sias (2001). It could be possible that the current top securities performed better than the bottom securities, but this thesis cannot answer that. A better approach would be to use a different way to construct the networks by defining the edges in a totally different way or by scaling the edges in a way, which would account for the size of the security

while working for the balance node at the same time. The balance node is important because it accounts for cash and all other securities which are not in the data, so the node is needed to account for these unobserved securities. One alternative way to scale the edges could be to use all of the ingoing edges to a node during the scaling period to find the maximum value which is used for the scaling. The goal of the centrality ranking would be to have securities with irregular inflows or outflows at the top of the ranking depending on the used centrality measure. Using the current network definition, the size of the security affects the centrality too much to differentiate unusual activity from usual activity.

The network motif study had a range of results depending on the network series, the subgraph, and the null model. Overall, directed configuration model explained the subgraph counts better than the directed random graph model. Because the directed configuration model had lower z-scores, the degree distributions contain important information about the networks, and the subgraph counts are not explained by a single probability for all of the edges. The exception was subgraph 2 as it had lower z-scores when using the directed random graph model in all cases. Subgraphs 1, 2 and 4 have 2 edges in different configurations and subgraphs 5 and 9 have 3 edges in different configurations. Subgraph 2 counts may have been better explained by the directed random graph model as in the subgraph 2 money flows to one security and out of it to another which should not be common if investors have similar opinion about the securities. Only the subgraph 2 had lower than expected count for daily finance institute networks when using either model or group of investors. In other cases there were more subgraphs than expected by either null model. For household networks the z-scores of subgraphs 2, 5 and 9 were higher than for finance institutes under the directed random graph model, and the z-scores of subgraph 1 and 4 were higher for finance institute networks. This may indicate that finance institutes have more similar opinions about the securities as the subgraph 1 and 4 are clearer decisions to invest to a security or to divest from a security. In the case of directed configuration model the z-scores of subgraphs 5 and 9 were lower than for other subgraphs for daily household networks. For other network series under the directed configuration model the z-scores of subgraphs 5 and 9 were higher than for other subgraphs. This indicates that the degree distributions have different importance to the structure of the daily household networks than for the other networks. In the end, the household networks had different results compared to the other networks, and doing regression analysis with the z-scores and other market data could yield interesting results.

Analyzing development of the z-scores over the crisis showed that there were no clear changes of regime in individual z-scores, there were large spikes in the z-scores and the z-scores of different subgraphs are not always changing values in the same way. At any moment there was not a large lasting change in any of the z-score time series which would indicate a clear change of regime in the crisis like Squartini, Van Lelyveld et al. (2013) found. The analysis of regime switching is not conclusive because the analysis is

based on interpretation of plots. Further analysis could show there are regimes, which have different means or variances that are not easily observable in the plots. It could be possible to find regimes in the time series by using all of the time series at the same time because the subgraphs contain different information about the market and their relative changes are important. For example, subgraph 5 means there is at least 1 security which is seen as a source of money and at least 1 node which is seen as a sink of money while for subgraph 9 there is no clear consensus of sources and sinks. The relative changes in these time series have additional information compared to the individual time series. At multiple times for all of the networks the z-scores of the subgraphs moved in opposite directions or had relatively different sized changes even though generally the z-scores changed in similar ways. At times when z-score values of some subgraphs increased a lot for a short period, other subgraphs had smaller changes. These unusual times can be during more volatile markets or there may be some other reason for the unusual period. Therefore, combining the z-score time series together can yield additional results and is a possible future research area. The multivariate study of the z-scores may include regime switching, forecasting the z-score values, and studying the possible connections to returns and volatilities of different securities, indexes, or possibly macroeconomic data.

In conclusion, the centrality measures of the securities correlated with the size of the securities and the motif analysis yielded values, which can be used better as a basis for further research. The centrality ranking method may be a viable way to identify signals about the individual securities using alternative way of constructing the networks and the current centrality values may be used in regression analysis with security return and volatility data. The motif analysis yielded various results, which can have multiple different interpretations, but further analysis is needed to understand the reasons for the changes in the significance levels better or to find what happens in the market at the same time as the values change.

The defined networks may be analyzed using different techniques also. The money flow networks add information about the source of the money as compared to the simpler technical indicator of money flowing either in or out of the security, and this more complex information should be studied further. The act of choosing to sell a security to buy another is a clear indicator of preferring the other security, and the money flow networks quantify how large the preference is among a group of investors. Edges between the balance node and other securities could contain additional information as money is brought to the network from the outside, or the edges from the balance node could simply mean investors are increasing already existing positions as they have gained additional money from their investors or wages. As the networks are defined for groups of investor, they could be compared to each other. For example, a possible analysis could be if institutions and households have similar or dissimilar edges by their directions. All in all, the study of money flow networks should be continued.

7. CONCLUSION

In this thesis money flow networks were introduced, and their structure was studied during the Financial Crisis of 2008. The goal of the thesis was to study the financial markets using a new method during unusual time period, which could show how the method identifies changes in the markets. Additionally, household and financial institutions were separated which added one more aspect to the thesis. The networks were defined by aggregating together investor networks, which contained directed money flows from sold securities to bought securities. Thus, the aggregated networks contained edges, which were weighted by the amount of money flowing from one security to another, and the weights were scaled by the highest edge value during the last 90 days.

The networks were studied over the crisis by using a ranking based on centrality scores and network motif analysis. The centrality-based ranking was used to identify securities, which were central more often than expected than if the securities were uniformly distributed. Securities with large market capitalization had the highest centrality scores during the crisis and were more often among the highest ranking securities than expected. Therefore, the ranking did not work for identifying securities with exceptional inflows or outflows of money relative to the usual money flows. Daily finance institute networks had larger percentile for which the bottom ranked securities followed the null hypothesis while weekly household networks had a large number of securities even in the bottom 10% percentile. Thus, there were differences between how finance institutes and households invested during the crisis as households invested less in smaller securities.

The network motif analysis had a range of results which need further analysis to tie the results to other market data. The network motif analysis consisted of studying the z-scores of the number of connected 3 node subgraphs in the weakly connected components of the networks. The number of subgraphs in general were explained better by the directed configuration model as the z-scores were higher when using the directed random graph model as the null model. The degree distributions therefore contain important information about the network subgraphs, and a single probability for all flows between securities does not work as well. There were also differences between the z-scores of finance institute networks and household networks. Subgraphs, which had clear structure of investing to or from a security, were more common in finance institute networks while more complex structures were more common in household networks. Therefore, finance institutes

may have more similar opinions about which securities are good or bad investments as compared to households. In addition, the directed configuration model explained the 3 edge subgraphs in the daily household networks better than for any other network series. Therefore, the structure of daily household networks is different from the other 3 network series, which are the weekly household and finance institute networks in addition to daily finance institute networks. On the one hand, during the crisis there was not a single time when the z-scores had a large lasting change which could have indicated a clear change in the way investing was done during the crisis. On the other hand, there were times when the z-scores had large changes up and down, and the z-scores moved differently relative to each other. These large changes and relative changes can have multiple different interpretations, and further analysis is needed to understand the changes better.

In this thesis both analyses focused on changes in the networks without relating the data to other market data or explaining the changes. The centrality rankings were used to identify if some securities were more often in the top or bottom percentiles, but the analysis does not tell if individual centrality values of the securities relate to other data about the securities. The analysis of the rankings is also inconclusive because only the 5 securities with the highest occurrences in the percentile were listed. Market capitalization data or some other size metric of all the securities would need to be used with the rankings to conclude if the size of the security explains the placement in the ranking. In the motif analysis analytical method was used for calculating the z-scores for the networks using the directed random graph model while sampling was used for directed configuration model. This difference may affect the results a little. Also, only the weakly connected components were used for the motif analysis which means the whole networks should be used in addition in future research to account for the disconnected nodes. The disconnected nodes are securities which did not have any trades for the group during the period which is significant information.

This thesis works as a starting point for further analysis using the money flow networks and the methods used in the thesis. Future research possibilities contain relating the network data to different market data and defining the networks in alternative ways. Relating the centrality values of individual securities to returns data can be a possible analysis for example. A hypothesis to test might be that the returns of a security have a high correlation with $k^{net} = k^{in} - k^{out}$. The subgraph z-score data may also be used in regression analysis with index or macroeconomic data to see if there is a correlation or a possibility to forecast the values for example. There are many possible ways to interpret the z-scores and testing different hypotheses is an interesting possibility. The hypotheses may include if higher volatility increases the z-scores of 3 edge subgraphs, if negative returns increases triadic subgraphs with nodes that have 2 edges directed at them or if the volatility of the z-scores correlates with the volatility of the market for example. In addition, money flows from the balance node may indicate excess returns as money flows

in from outside the network. Overall, the defined money flow networks introduced a new way to study the financial markets and they possess interesting properties, which should be studied further.

REFERENCES

Allen, F. and Babus, A. (2008). Networks in Finance. *The Network Challenge: Strategy, Profit, and Risk in an Interlinked World*. DOI: 10.2139/ssrn.1094883.

Alstott, J., Bullmore, E. and Plenz, D. (2014). Powerlaw: a Python Package for Analysis of Heavy-Tailed Distributions. eng. *PloS one* 9.1, e85777–e85777. ISSN: 1932-6203.

Barabási, A.-L. (2013). Network Science. eng. *Philosophical transactions of the Royal Society of London. Series A: Mathematical, physical, and engineering sciences* 371.1987, pp. 20120375–20120375. ISSN: 1364-503X.

Barabási, A.-L. and Albert, R. (1999). Emergence of Scaling in Random Networks. eng. *Science (American Association for the Advancement of Science)* 286.5439, pp. 509–512. ISSN: 0036-8075.

Barabási, A.-L. and Bonabeau, E. (2003). Scale-Free Networks. *Scientific american* 288.5, pp. 60–69.

Barabási, A.-L., Gulbahce, N. and Loscalzo, J. (2011). Network Medicine: a Network-Based Approach to Human Disease. eng. *Nature reviews. Genetics* 12.1, pp. 56–68. ISSN: 1471-0056.

Barabási, A.-L. and Pósfai, M. (2016). *Network Science*. Cambridge: Cambridge University Press. ISBN: 978-1107076266. URL: http://barabasi.com/networksciencebook/.

Barber, B. and Odean, T. (2000). Trading Is Hazardous to Your Wealth: The Common Stock Investment Performance of Individual Investors. eng. *The Journal of Finance (New York)* 55.2, pp. 773–806. ISSN: 0022-1082.

Bennett, J. and Sias, R. (2001). Can Money Flows Predict Stock Returns? eng. *Financial analysts journal* 57.6, pp. 64–77. ISSN: 0015-198X.

Braha, D. (2020). Patterns of Ties in Problem-Solving Networks and Their Dynamic Properties. eng. *Scientific reports* 10.1, pp. 18137–18137. ISSN: 2045-2322.

Bramoulle, Y. and Kranton, R. (2007). Risk Sharing across Communities. eng. *The American economic review* 97.2, pp. 70–74. ISSN: 0002-8282.

Brogaard, J., Hendershott, T. and Riordan, R. (2014). High-Frequency Trading and Price Discovery. eng. *The Review of financial studies* 27.8, pp. 2267–2306. ISSN: 0893-9454.

Brown, K. and Brooke, B. (1993). Institutional Demand and Security Price Pressure: The Case of Corporate Spinoffs. eng. *Financial analysts journal* 49.5, pp. 53–62. ISSN: 0015-198X.

Bullard, J., Neely, C. and Wheelock, D. (2009). Systemic Risk and the Financial Crisis: A Primer. eng. *Review - Federal Reserve Bank of St. Louis* 91.5, pp. 403–. ISSN: 0014-9187.

Camas, F., Blázquez, J. and Poyatos, J. (2006). Autogenous and Nonautogenous Control of Response in a Genetic Network. eng. *Proceedings of the National Academy of Sciences - PNAS* 103.34, pp. 12718–12723. ISSN: 0027-8424.

Cao, C., Chen, Y., Goetzmann, W. and Liang, B. (2018). Hedge Funds and Stock Price Formation. eng. *Financial analysts journal* 74.3, pp. 54–68. ISSN: 0015-198X.

Chase, G. (1972). On the Chi-Square Test When the Parameters are Estimated Independently of the Sample. eng. *Journal of the American Statistical Association* 67.339, pp. 609–611. ISSN: 0162-1459.

Chernoff, H. and Lehmann, E. (1954). The Use of Maximum Likelihood Estimates in χ^2 Tests for Goodness of Fit. eng. *The Annals of mathematical statistics* 25.3, pp. 579–586. ISSN: 0003-4851.

Clauset, A., Rohilla Shalizi, C. and Newman, M. (2009). Power-Law Distributions in Empirical Data. eng. *SIAM review* 51.4, pp. 661–703. ISSN: 0036-1445.

Daniel, K., Grinblatt, M., Titman, S. and Wermers, R. (1997). Measuring Mutual Fund Performance with Characteristic-Based Benchmarks. eng. *The Journal of Finance (New York)* 52.3, pp. 1035–1058. ISSN: 0022-1082.

Elton, E., Gruber, M. and Busse, J. (2004). Are Investors Rational? Choices among Index Funds. eng. *The Journal of Finance (New York)* 59.1, pp. 261–288. ISSN: 0022-1082.

Erdös, P. and Rényi, A. (1959). On Random Graphs I. *Publicationes Mathematicae Debrecen* 6, pp. 290–297.

Ericsson (2009). *Ericsson Annual Report 2008*. URL: https://www.ericsson.com/48fa08/assets/local/investors/documents/financial-reports-and-filings/annual-reports/annual-report-2008-complete-en.pdf (visited on 09/04/2021).

Federal Reserve (2009a). *FOMC statement*. URL: https://www.federalreserve.gov/newsevents/pressreleases/monetary20090429a.htm (visited on 10/27/2021).

– (2009b). *FOMC statement*. URL: https://www.federalreserve.gov/newsevents/pressreleases/monetary20090624a.htm (visited on 10/27/2021).

Ferris, S., Haugen, R. and Makhija, A. (1988). Predicting Contemporary Volume with Historic Volume at Differential Price Levels: Evidence Supporting the Disposition Effect. eng. *The Journal of Finance (New York)* 43.3, pp. 677–697. ISSN: 0022-1082.

Fienberg, S. (2012). A Brief History of Statistical Models for Network Analysis and Open Challenges. eng. *Journal of computational and graphical statistics* 21.4, pp. 825–839. ISSN: 1061-8600.

Fligstein, N. and Roehrkasse, A. (2016). The Causes of Fraud in the Financial Crisis of 2007 to 2009: Evidence from the Mortgage-Backed Securities Industry. eng. *American sociological review* 81.4, pp. 617–643. ISSN: 0003-1224.

Forsgren, A., Gill, P. and Wright, M. (2002). Interior Methods for Nonlinear Optimization. eng. *SIAM review* 44.4, pp. 525–597. ISSN: 0036-1445.

Fujiwara, Y., Inoue, H., Yamaguchi, T., Aoyama, H., Tanaka, T. and Kikuchi, K. (2021). Money Flow Network Among Firms' Accounts in a Regional Bank of Japan. eng. *EPJ data science* 10.1, pp. 19–19. ISSN: 2193-1127.

Grossman, S. and Miller, M. (1988). Liquidity and Market Structure. eng. *The Journal of Finance (New York)* 43.3, pp. 617–633. ISSN: 0022-1082.

Hâncean, M.-G., Lerner, J., Perc, M., Ghiţă, M., Bunaciu, D.-A., Stoica, A. A. and Mihăilă, B.-E. (Sept. 2021). The Role of Age in the Spreading of COVID-19 Across a Social Network in Bucharest. *Journal of Complex Networks* 9.4. cnab026. ISSN: 2051-1329. DOI: 10.1093/comnet/cnab026.

Holme, P. and Saramäki, J. (2012). Temporal Networks. eng. *Physics reports* 519.3, pp. 97–125. ISSN: 0370-1573.

Holt, J. (2009). A Summary of the Primary Causes of the Housing Bubble and the Resulting Credit Crisis: A Non-Technical Paper. *The Journal of Business Inquiry* 8, pp. 120–129.

Horne, J., Blume, M. and Friend, I. (1975). The Asset Structure of Individual Portfolios and Some Implications for Utility Functions. eng. *The Journal of Finance (New York)* 30.2, pp. 585–603. ISSN: 0022-1082.

Huberman, G. (2001). Familiarity Breeds Investment. eng. *The Review of financial studies* 14.3, pp. 659–680. ISSN: 0893-9454.

Interavanti Oyj (2009). *Vuosikertomus 2008*. URL: http://interavanti.fi/files/interavanti_vuosikertomus2008.pdf (visited on 09/04/2021).

Kahneman, D. (1973). *Attention and Effort*. eng. Prentice-Hall series in experimental psychology. Englewood Cliffs (N.J.): Prentice-Hall. ISBN: 0-13-050518-8.

Kashani, Z., Ahrabian, H., Elahi, E., Nowzari-Dalini, A., Ansari, E., Asadi, S., Mohammadi, S., Schreiber, F. and Masoudi-Nejad, A. (2009). Kavosh: a New Algorithm for Finding Network Motifs. eng. *BMC bioinformatics* 10.1, pp. 318–318. ISSN: 1471-2105.

Kashtan, N., Itzkovitz, S., Milo, R. and Alon, U. (2004). Efficient Sampling Algorithm for Estimating Subgraph Concentrations and Detecting Network Motifs. eng. *Bioinformatics* 20.11, pp. 1746–1758. ISSN: 1367-4803.

Ke, B. and Ramalingegowda, S. (2005). Do Institutional Investors Exploit the Post-earnings Announcement Drift? eng. *Journal of accounting & economics*. Journal of Accounting and Economics 39.1, pp. 25–53. ISSN: 0165-4101.

Kivela, M., Arenas, A., Barthelemy, M., Gleeson, J., Moreno, Y. and Porter, M. (2014). Multilayer Networks. eng. *Journal of complex networks* 2.3, pp. 203–271. ISSN: 2051-1310.

Knüpfer, S., Rantapuska, E. and Sarvimäki, M. (2017). Formative Experiences and Portfolio Choice: Evidence from the Finnish Great Depression. *The Journal of Finance* 72.1, pp. 133–166. DOI: https://doi.org/10.1111/jofi.12469.

Kranton, R. and Minehart, D. (2001). A Theory of Buyer-Seller Networks. eng. *The American economic review* 91.3, pp. 485–508. ISSN: 0002-8282.

Leinhardt, S. and Holland, P. (1974). The Statistical Analysis of Local Structure in Social Networks. eng. *[Online]*. NBER Working Paper Series. URL: https://www.nber.org/system/files/working_papers/w0044/w0044.pdf (visited on 06/11/2021).

Maslov, S. and Sneppen, K. (2002). Specificity and Stability in Topology of Protein Networks. eng. *Science (American Association for the Advancement of Science)* 296.5569, pp. 910–913. ISSN: 0036-8075.

Maslov, S., Sneppen, K. and Zaliznyak, A. (2004). Detection of Topological Patterns in Complex Networks: Correlation Profile of the Internet. *Physica A: Statistical Mechanics and its Applications* 333, pp. 529–540. ISSN: 0378-4371. DOI: https://doi.org/10.1016/j.physa.2003.06.002.

Massey, F. (1951). The Kolmogorov-Smirnov Test for Goodness of Fit. *Journal of the American Statistical Association* 46.253, pp. 68–78. ISSN: 01621459. URL: http://www.jstor.org/stable/2280095.

McIntyre, D. (2009). *More Quickly Than It Began, The Banking Crisis Is Over*. URL: http://content.time.com/time/business/article/0,8599,1890560,00.html (visited on 10/27/2021).

McKay, B. and Piperno, A. (2014). Practical Graph Isomorphism, II. eng. *Journal of symbolic computation* 60, pp. 94–112. ISSN: 0747-7171.

McKibbin, W. and Stoeckel, A. (2010). The Global Financial Crisis: Causes and Consequences. eng. *Asian economic papers*. Asian Economic Papers 9.1, pp. 54–86. ISSN: 1535-3516.

Metso Oyj (2009). *Metso and Tamfelt enter into a Combination Agreement; Share Exchange Offer for all of Tamfelt's shares*. URL: https://www.globenewswire.com/news-release/2009/11/05/142438/0/en/Metso-and-Tamfelt-enter-into-a-Combination-Agreement-Share-Exchange-Offer-for-all-of-Tamfelt-s-shares.html (visited on 09/04/2021).

Milo, R., Shen-Orr, S., Itzkovitz, S., Kashtan, N., Chklovskii, D. and Alon, U. (2002). Network Motifs: Simple Building Blocks of Complex Networks. eng. *Science (American Association for the Advancement of Science)* 298.5594, pp. 824–827. ISSN: 0036-8075.

Milo, R., Itzkovitz, S., Kashtan, N., Levitt, R., Shen-Orr, S., Ayzenshtat, I., Sheffer, M. and Alon, U. (2004). Superfamilies of Evolved and Designed Networks. eng. *Science (American Association for the Advancement of Science)* 303.5663, pp. 1538–1542. ISSN: 0036-8075.

Nagurney, A. and Ke, K. (2001). Financial Networks with Intermediation. eng. *Quantitative finance*. Quantitative Finance 1.4, pp. 441–451. ISSN: 1469-7688.

Nasdaq (2021). *OMX Helsinki 25 (OMXH25)*. URL: https://indexes.nasdaqomx.com/Index/Overview/OMXH25 (visited on 09/12/2021).

Newman, M. (2006). Modularity and Community Structure in Networks. eng. *Proceedings of the National Academy of Sciences - PNAS* 103.23, pp. 8577–8582. ISSN: 0027-8424.

Newman, M. (2018). *Networks*. eng. Oxford: Oxford University Press. ISBN: 0198805098.

Nickerson, R. (1998). Confirmation Bias: A Ubiquitous Phenomenon in Many Guises. eng. *Review of general psychology* 2.2, pp. 175–220. ISSN: 1089-2680.

Noldus, R. and Van Mieghem, P. (2015). Assortativity in complex networks. eng. *Journal of complex networks* 3.4, pp. 507–542. ISSN: 2051-1310.

OEIS Foundation Inc., The On-Line Encyclopedia of Integer Sequences (2021). *A000273*. URL: http://oeis.org/A000273 (visited on 07/01/2021).

Page, L., Brin, S., Motwani, R. and Winograd, T. (1999). The PageRank Citation Ranking: Bringing Order to the Web. 1999-66. Previous number = SIDL-WP-1999-0120. URL: http://ilpubs.stanford.edu:8090/422/.

Patra, S. and Mohapatra, A. (2020). Review of Tools and Algorithms for Network Motif Discovery in Biological Networks. eng. *IET systems biology* 14.4, pp. 171–189. ISSN: 1751-8849.

Picard, F., Daudin, J.-J., Koskas, M., Schbath, S. and Robin, S. (2008). Assessing the Exceptionality of Network Motifs. eng. *Journal of computational biology* 15.1, pp. 1–20. ISSN: 1066-5277.

Rottenstreich, Y. and Hsee, C. K. (2001). Money, Kisses, and Electric Shocks: On the Affective Psychology of Risk. eng. *Psychological science* 12.3, pp. 185–190. ISSN: 0956-7976.

Saracco, F., Di Clemente, R., Gabrielli, A. and Squartini, T. (2016). Detecting Early Signs of the 2007-2008 Crisis in the World Trade. eng. *Scientific reports* 6.1. ISSN: 2045-2322.

Scikit-Learn (2021). *Preprocessing Data*. URL: https://scikit-learn.org/stable/modules/preprocessing.html (visited on 09/09/2021).

SciPy (2021). *Statistical Functions*. URL: https://docs.scipy.org/doc/scipy/reference/stats.html (visited on 08/30/2021).

Soprano Oyj (2010). *Soprano Oyj Konserni tilinpäätös*. URL: https://www.soprano.fi/wp-content/uploads/2019/05/Konsernin-tilinpaatos-2009-1.pdf (visited on 09/04/2021).

Squartini, T. and Garlaschelli, D. (2011). Analytical Maximum-Likelihood Method to Detect Patterns in Real Networks. eng. *New journal of physics* 13.8, pp. 83001–. ISSN: 1367-2630.

Squartini, T., Van Lelyveld, I. and Garlaschelli, D. (2013). Early-Warning Signals of Topological Collapse in Interbank Networks. eng. *Scientific reports* 3.1, pp. 3357–3357. ISSN: 2045-2322.

SSK Suomen Säästäjien Kiinteistöt Oyj (2009). *Tilinpäätös 2008*. URL: https://vuosikertomukset.net/resources/Investors_house/fin/vuosikertomukset/SSK%5C_tilinpaatos%5C_2008.pdf (visited on 09/04/2021).

Sunstein, C. R. (2002). Probability Neglect: Emotions, Worst Cases, and Law. eng. *The Yale law journal* 112.1, pp. 61–. ISSN: 0044-0094.

Tversky, A. and Kahneman, D. (1974). Judgment under Uncertainty: Heuristics and Biases. *science* 185.4157, pp. 1124–1131.

Vaahto Group Plc Oyj (2009). *Vuosikertomus 2007-2008*. URL: https://web.lib.aalto.fi/fi/old/yrityspalvelin/pdf/2008/Fvaahtogroup2008.pdf (visited on 09/04/2021).

Valtioneuvoston kanslia (2011). Finanssikriisin reaalitaloudelliset vaikutukset Suomessa: alustava kokonaisarvio. *Valtioneuvoston kanslian raporttisarja* 7/2011.

Wasserman, S. (1994). *Social network analysis : methods and applications*. eng. Structural analysis in the social sciences ; 8. Cambridge: Cambridge University Press. ISBN: 0-521-38269-6.

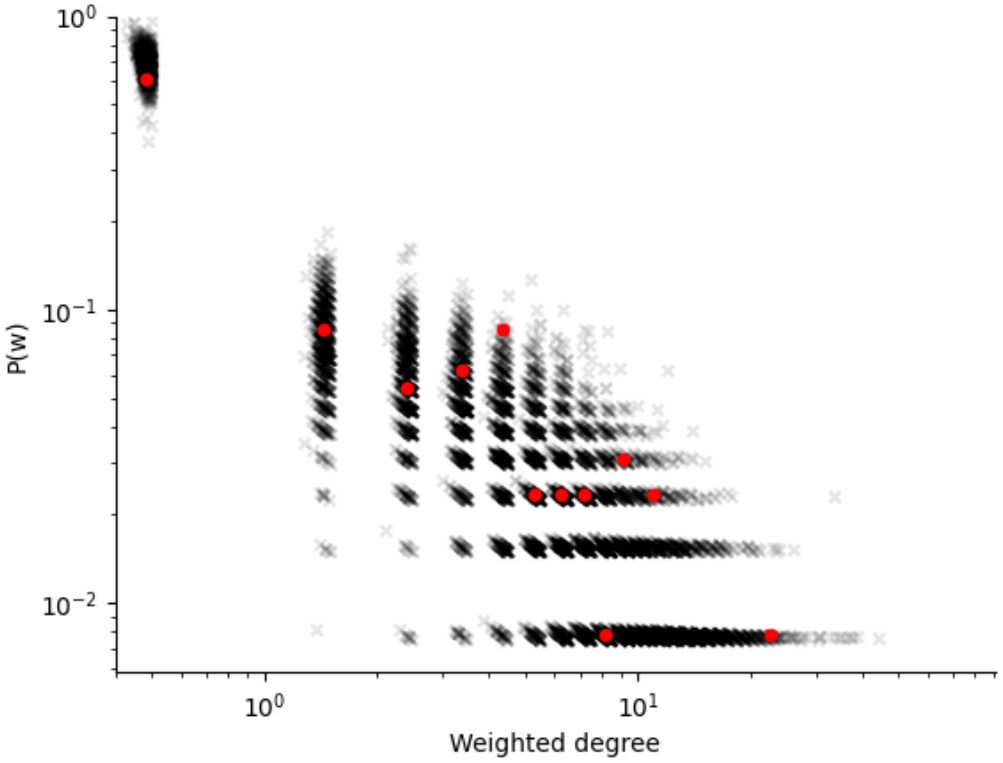
Watts, D. and Strogatz, S. (1998). Collective Dynamics of 'Small-World' Networks. eng. *Nature (London)* 393.6684, pp. 440–442. ISSN: 0028-0836.

Yeo, I.-K. (2000). A New Family of Power Transformations to Improve Normality or Symmetry. eng. *Biometrika* 87.4, pp. 954–959. ISSN: 1464-3510.

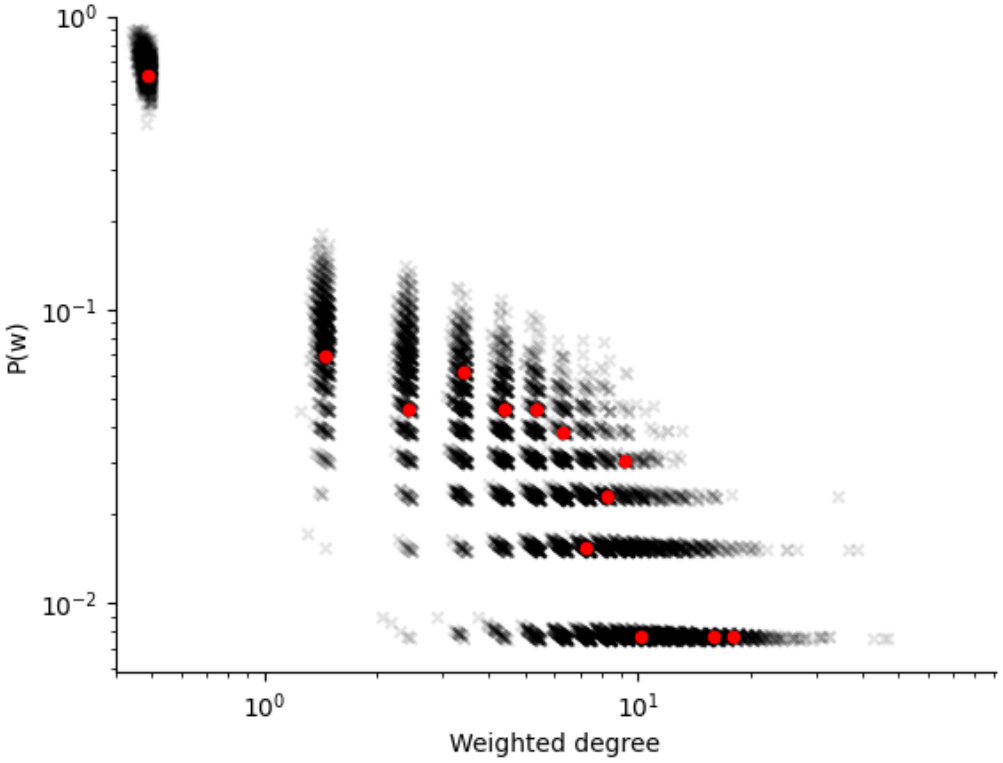
Zeng, Y. (2016). Institutional Investors: Arbitrageurs or Rational Trend Chasers. eng. *International review of financial analysis* 45, pp. 240–262. ISSN: 1057-5219.

APPENDIX A: DEGREE DISTRIBUTIONS OF THE NETWORKS

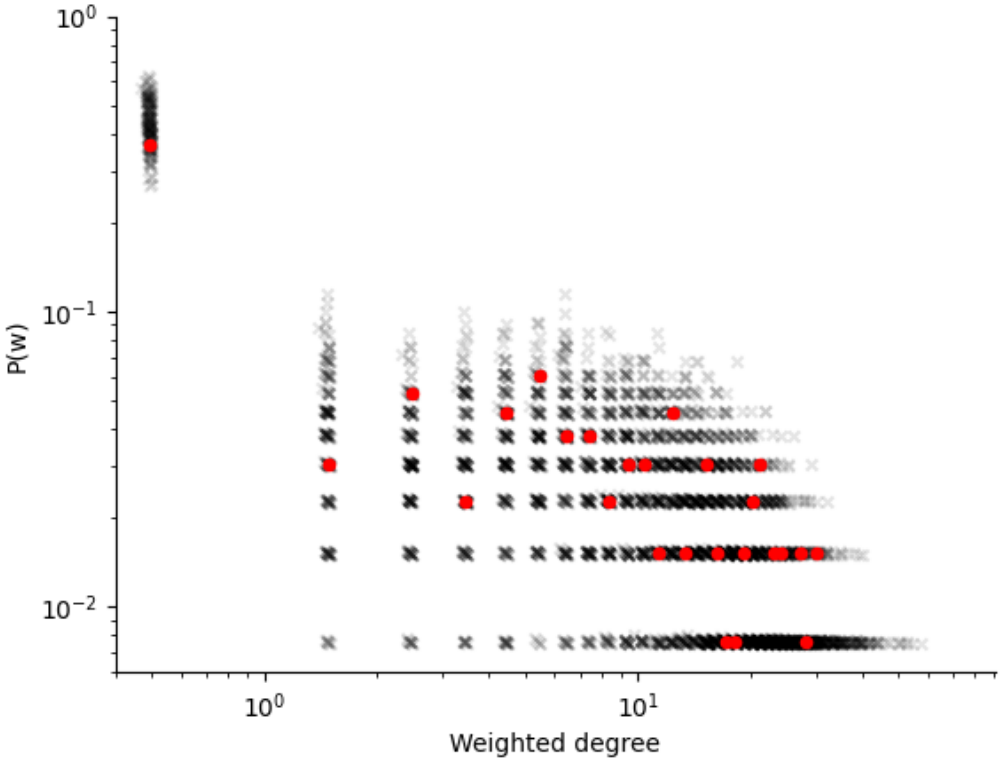
The one dimensional degree distributions can be visualized in histograms by counting the observed degrees and plotting them with k on the x-axis. Generalizing the degrees to include continuous weights poses a problem for plotting the degree histograms because two values may be close to each other, but the probabilities would be $1/n$. Thus, intervals are used to group values together. In Figure A.1 the in- and out-degree distributions of all network series are presented. The distributions are plotted with $\lceil k_{max} \rceil$ intervals and only the middle point of the intervals is plotted. The use of $\lceil k_{max} \rceil$ intervals adds a hue to the plots as the intervals have width $k_{max} / \lceil k_{max} \rceil$, which is not the same for all distributions.



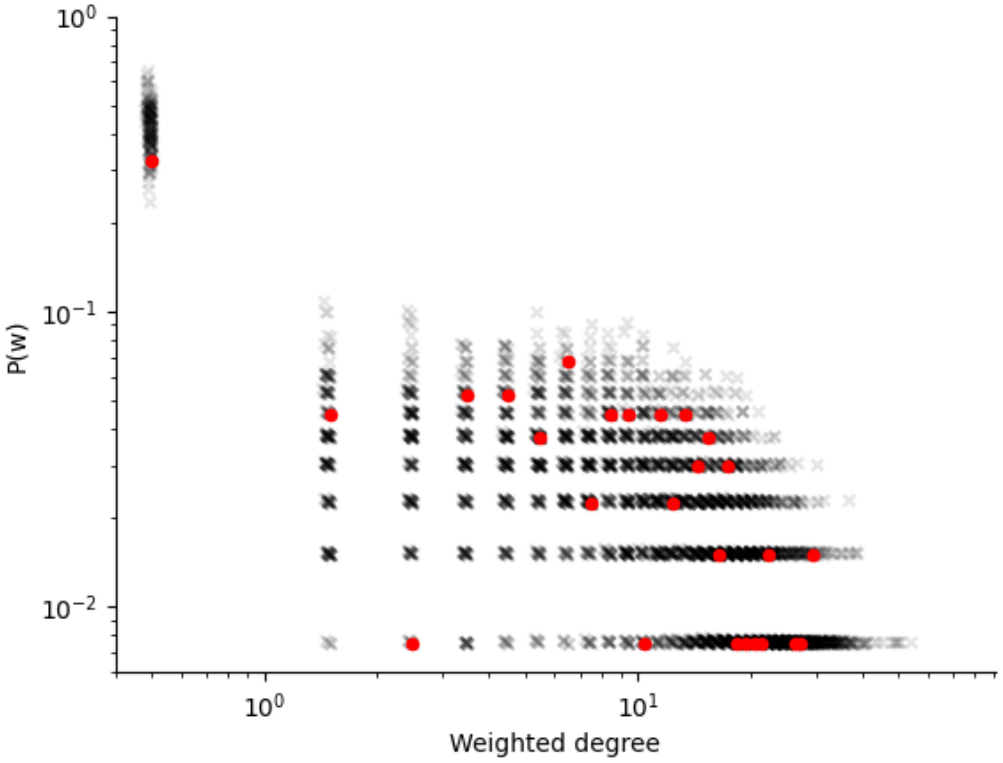
(a) *In-degrees of daily finance institute networks.*



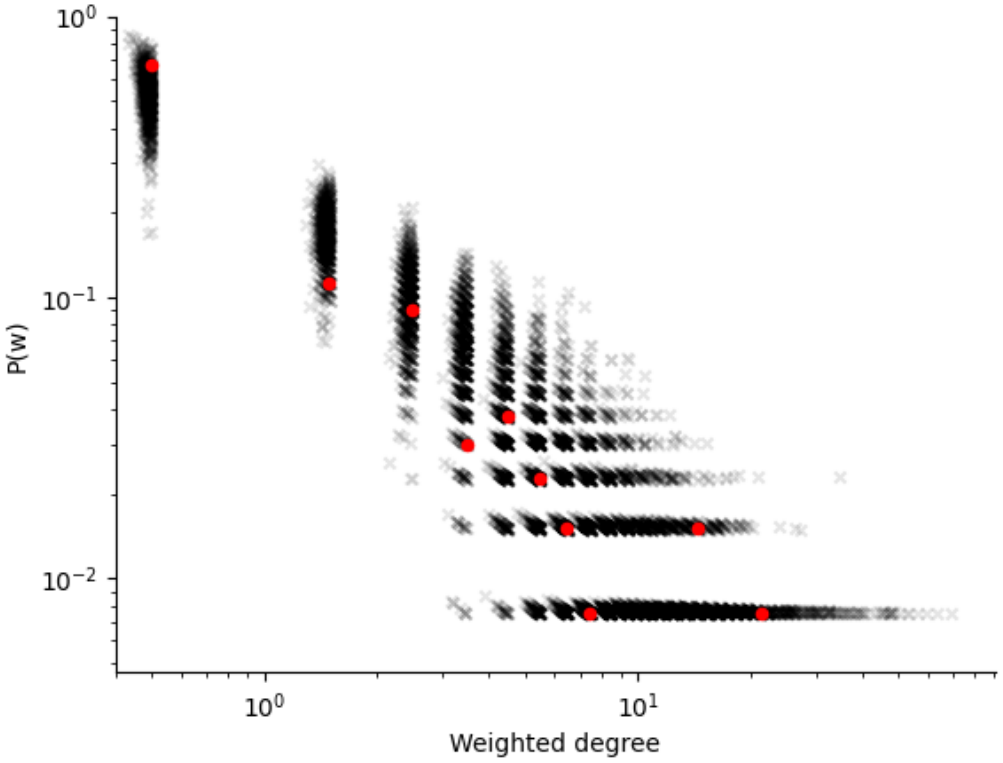
(b) *Out-degrees of daily finance institute networks.*



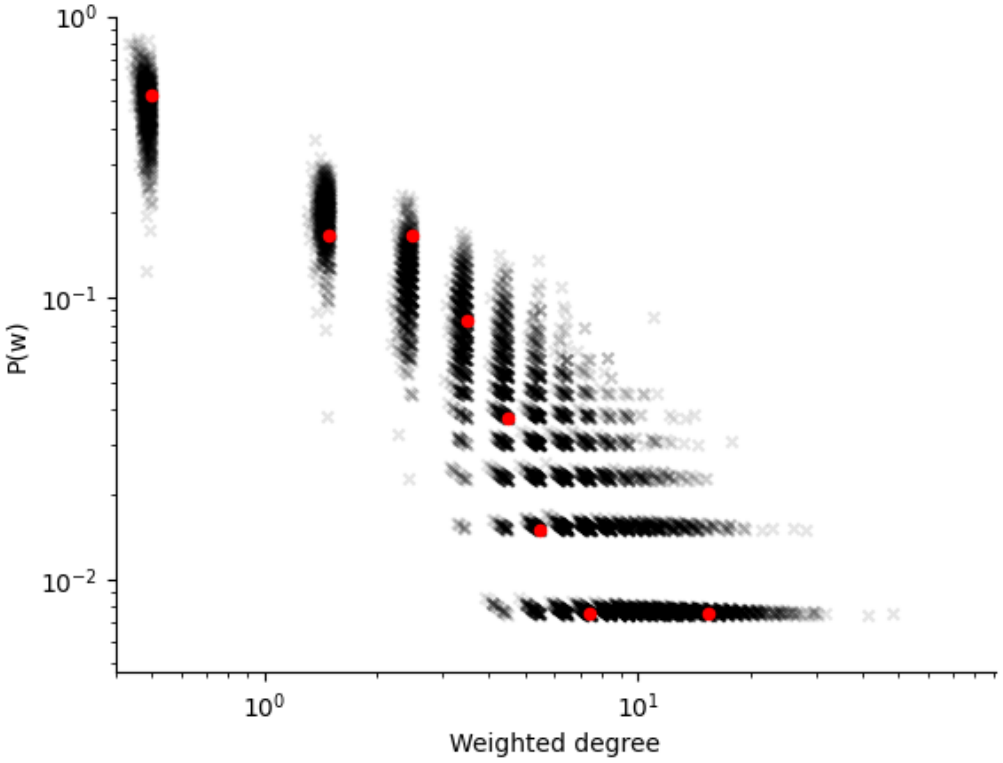
(c) *In-degrees of weekly finance institute networks.*



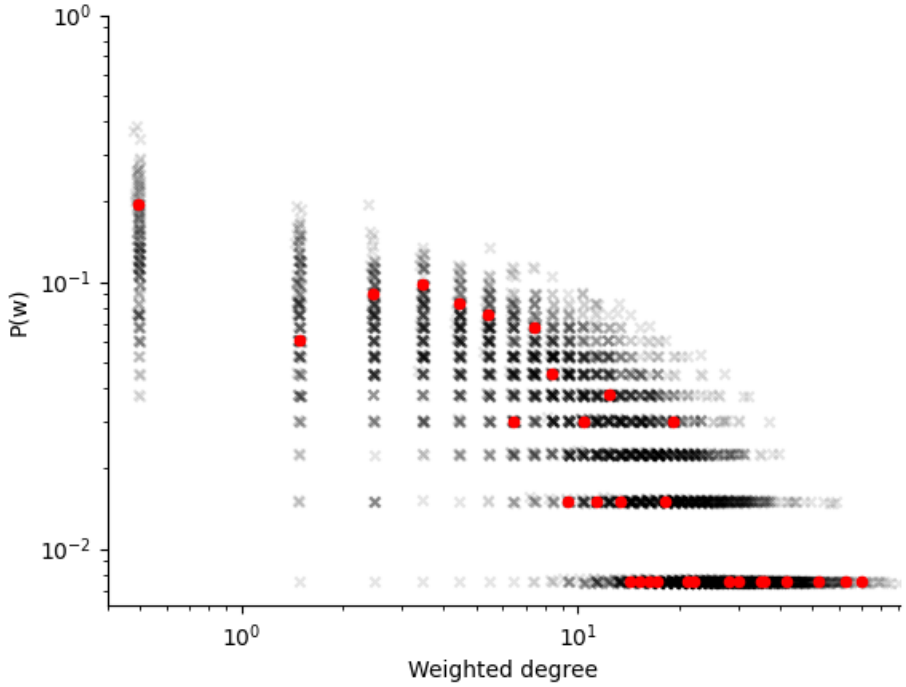
(d) *Out-degrees of weekly finance institute networks.*



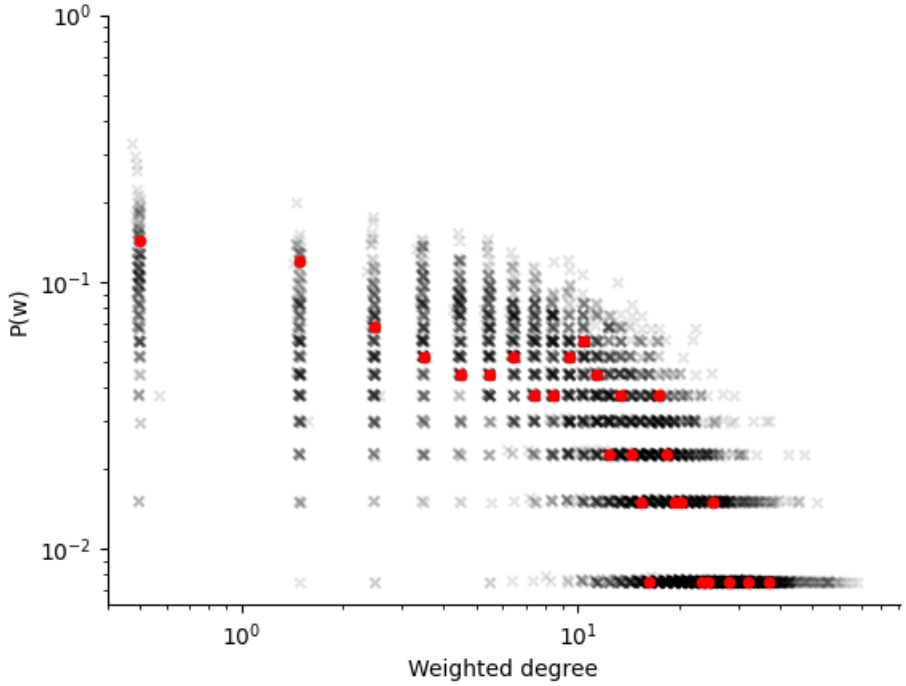
(e) *In-degrees of daily household networks.*



(f) *Out-degrees of daily household networks.*



(g) *In-degrees of weekly household networks.*



(h) *Out-degrees of weekly household networks.*

Figure A.1. *Log-log plots of in- and out-degree distributions of all networks with the last observed network highlighted as red balls. Most of the distributions show a downward trending curve in the values while daily household networks are closer to a straight line with a negative slope. Notation $P(w)$ is used for the weighted degrees to emphasize that the degrees are weighted.*

The degree distributions in Figure A.1 are mainly similar to each other except for the daily household networks. Most of the networks have large number of nodes with degree under 1 and some nodes with higher degrees. In addition, the distributions seem to have the same general shape over time, but there is still some variability in the values. The distributions do not seem to follow a distribution of a random directed graph as there is a large number of nodes with degree under 1 and there are nodes with high degrees. At the same time the distributions are not following a power-law strictly, which would show as a straight line on the log-log plot (Newman 2018, p. 317), because of the curve downwards. The curve could be because of large variability in the distributions and therefore outliers obscuring the linear trend. Daily household degree distributions could follow a power-law the best, but it cannot be said from a plot. The distributions show large deviations as for a degree k there are many possible probabilities $P(w)$ or in the other direction there are many possible degrees k for a given probability $P(w)$. The daily household degree distributions have the least variability of the distributions as the values are grouped closer together.