Chapter 1

Financial market – a network of capital flows

J.-P. Onnela[1], K. Kaski[1], and J. Kertesz[1,2]

[1] Laboratory of Computational Engineering, Helsinki University of Technology, P.O. Box 9203, FIN-02015 HUT, Finland

[2] Dept of Theoretical Physics, Budapest University of Technology and Economics, Budafoki ut 8, H-1111, Budapest, Hungary

Network theory provides an approach to complex financial systems containing many interacting units, in which the interactions are reflected in temporal correlations based on price, volume, flow of capital, or their relative changes. We construct correlation based networks for a subset of NYSE traded stocks, in which the nodes correspond to stocks, and edges to distances between them. According to random matrix theory, the correlation matrix underlying the graph is dominated by noise and, therefore, needs to be pruned. For this we use two approaches, by constructing (i) a minimum spanning tree of edges (asset tree) or (ii) a graph based on agglomerative clustering, i.e., adding edges to the graph based on their rank (asset graph). Asset trees are found to be scale-free networks with almost equal scaling exponents for all input quantities. Asset graphs lead to more robust topologies but, in this case, the scale-free topology is not as clear.

1.1 Introduction

Choosing the most appropriate variables to study the properties of financial markets is not a trivial task. Mantegna and Stanley characterize this difficulty in [1] by stating: "The scales used are often given in units (currencies) that are themselves fluctuating in time and transactions occur at random times with random intensities." Price and volume are the two most basic quantities used

to characterize stock performance. In what follows, we define these quantities and introduce some others derived from them that are potentially interesting in a financial network context.

Here the **price** of a stock is the last reported transaction price of the day, or the daily closing price, on a stock exchange, and for stock i at time τ, we denote it by $P_i(\tau)$. Since investors obviously work in terms of relative returns, i.e., their potential gain in proportion to the invested sum, absolute prices are seldom used as such. Instead, **logarithmic returns** are used commonly, since they incorporate an average correction of scale changes, and inflation is simply lumped together with other sources of steady compound growth. The daily logarithmic return of stock i is denoted by $r_i(\tau) = \ln P_i(\tau) - \ln P_i(\tau - 1)$.

In stock market language, the trading volume or just **volume** is the number of shares traded during a given time period. As it reflects on trading activity, sometimes 'activity' is used interchangeably with volume. In this study, we use $V_i(\tau)$ to denote the volume of stock i at time τ. However, it is not always a suitable variable because, just like price, its absolute level is somewhat arbitrary. The term market capitalization refers to a company's market price and is calculated by multiplying the number of shares outstanding by the price per share. Although price and volume are related through market capitalization, their absolute values can be arbitrary, in this sense. Therefore, it may be more suitable to use **logarithmic relative volume**, defined as $\dot{V}_i(\tau) = \ln V_i(\tau) - \ln V_i(\tau - 1)$. The use of logarithm is motivated by its tendency to smoothen large fluctuations of volumes. Studying relative volume can reveal whether trading activities of different stocks are correlated.

The term dollar volume or monetary volume is used to indicate the dollar amount of shares traded during a given time period. Since market prices fluctuate throughout the day, to obtain a daily monetary volume one multiplies the price of a share in a given transaction by the transaction volume and then sums over all the transactions within the day. Instead we approximate monetary volume by calculating the product of the daily closing price and daily volume $F_i(\tau) = P_i(\tau)V_i(\tau)$, which we call **flow of capital**, or simply flow. Based on the discussion above, it seems desirable to render the patterns of flows independent of scale, and the most obvious way of doing so is to study their relative change. Analogously to the logarithmic relative volume, **logarithmic relative flow** is defined as $\dot{F}_i(\tau) = \ln F_i(\tau) - \ln F_i(\tau - 1) = \ln P_i(\tau)V_i(\tau) - \ln P_i(\tau - 1)V_i(\tau - 1) = r_i(\tau) + \dot{V}_i(\tau)$.

1.2 Data

In this study, the financial market refers to a subset of daily price and volume data for a set of New York Stock Exchange (NYSE) traded stocks. The stocks are identified by index $i = 1, \ldots, 5128$, and time is indicated by $\tau = 1, 2, \ldots, 5056$, corresponding to a 20-year-period from 2-Jan-1980 to 31-Dec-1999. In general, when studying price, logarithmic return, volume, or logarithmic relative volume,

one needs to use split-adjusted data[1], because splitting causes a sudden fall (jump) in price (volume). For flow or logarithmic relative flow, stock splits do not cause problems, but for consistency we use split adjusted data in all cases. The data are divided time-wise into M windows, where $t = 1, 2, ..., M$ of width $T = 1000$, the number of quotes in the window, which corresponds to a four year period, assuming 250 trading days a year. Several consecutive windows overlap, the extent of which is dictated by the window step length $\delta T = 250/12 \approx 20.8$, the displacement of the window in trading days, resulting in $M = 219$ windows.

We have earlier studied the same set of data, but accepted only stocks that were traded throughout the 20 year period and, hence, the number of stocks in a window was constant N. This naturally includes only the most persistent companies and thus the system loses some of the dynamics caused by introduction and departure of stocks to and from the market. We now follow an alternative approach by including all stocks that exist inside the window, regardless of their status outside. This leads to a more realistic view of the market, increases the number of stocks, and results in $N(t)$ varying between 835 and 1609.

Since the data contains some errors, a choice needs to be made as to what degree of error is tolerated and how errors are dealt with. The results presented in this paper were calculated from one dataset obtained by pruning the raw data. The prerequisites for selection were: (1) a price for the first entry within a window existed, (2) stocks with any number of zero volumes were left out[2], and (3) for each stock, neither price nor volume were missing in more than 1% of the window entries, thus allowing 10/1000 quotes to be missing. A missing price quote was substituted by the previous day's quote, since stock prices on consecutive days are very highly correlated. As consecutive trading volumes can be very different, it is more appropriate to replace a missing volume with an average volume of the stock calculated over existing volume quotes in the window (mean imputation). The method works well if price and volume are not strongly correlated.

1.3 Asset trees and asset graphs

This section is formulated in terms of logarithmic returns $r_i(\tau)$, but this is for notational convenience only, since $r_i(\tau)$ could equally well be substituted by $P_i(\tau)$, $V_i(\tau)$, $\dot{V}_i(\tau)$, $F_i(\tau)$, or $\dot{F}_i(\tau)$. In order to investigate correlations between the different measures, we use equal time correlation coefficients between assets i and j defined as

$$\rho_{ij}^t = \frac{\langle r_i^t r_j^t \rangle - \langle r_i^t \rangle \langle r_j^t \rangle}{\sqrt{[\langle {r_i^t}^2 \rangle - \langle r_i^t \rangle^2][\langle {r_j^t}^2 \rangle - \langle r_j^t \rangle^2]}}, \tag{1.1}$$

[1] A stock split is an increase in the number of outstanding shares of a company's stock, such that proportionate equity of each shareholder and the total capitalization of the company remain the same. Split-adjusted data means that price and volume quotes in the past are adjusted to reflect subsequent splits.

[2] These are problematic because they cause $\dot{V}_i(\tau)$ and $\dot{F}_i(\tau)$ to tend to infinity.

where $\langle ... \rangle$ indicates a time average over the return vectors. These correlation coefficients between $N(t)$ assets form a symmetric $N(t) \times N(t)$ correlation matrix $\mathbf{C}^t$. We define a distance between each pair of stocks as $d_{ij}^t = [2(1-\rho_{ij}^t)]^{1/2}$, motivated by considerations of ultrametricity [1]. We use this definition for reasons of compatibility with earlier work, although for our purposes any monotonically decreasing distance function of the correlation coefficient ρ_{ij}^t would do. With the chosen transformation, the individual correlation coefficients are mapped from $[-1, 1]$ to $[2, 0]$, and the correlation matrix is mapped into a symmetric distance matrix $\mathbf{D}^t$. As the different time windows are displaced by δT, the sequence of distance matrices for $t = 1, 2, ..., M$ can be interpreted to give rise to a sequence of time evolutionary steps of a single asset tree or asset graph, the two types of financial networks studied here.

Asset trees are constructed according to Mantegna [1] by determining the minimum spanning tree (MST) of the distances, denoted $\mathbf{T}^t$. The spanning tree is a simply connected acyclic graph that connects all $N(t)$ nodes (stocks); its size (number of edges) is fixed at $N(t) - 1$ such that the sum of all edge weights, $\sum_{d_{ij}^t \in \mathbf{T}^t} d_{ij}^t$, is minimum. The spanning tree, by definition, spans all $N(t)$ vertices in all time windows t and is, thus, time independent, but the set of edges $E(t)$ is time dependent. The main motivation behind studying asset trees is to learn about market taxonomy, dynamics, and their relation to portfolio optimization, as we have done in the past [4]. **Asset graphs**, in contrast, are created by inserting the $N(t) - 1$ shortest elements of the $\mathbf{D}^t$ matrix, and now $V(t)$, the set of *spanned* vertices, is also time dependent. The choice of the number of edges to add is motivated by the tree also having $N(t) - 1$ edges, and this renders the two approaches comparable in terms of network size. There is no acyclicity condition for asset graphs, nor do they need to be connected, i.e., they may consist of several disconnected components. The main motivation behind studying asset graphs is to learn about stock market clustering, and earlier we have hypothesized a connection between asset graphs and the information content of its edges, i.e., the fraction of edges needed to represent clustering of the market [4].

1.4 Simple network dynamics

In order to characterize network robustness and evolution, we define *multi-step survival ratio* at time t as the fraction of edges found common in the last k steps of the network as

$$\sigma(t, k) = \frac{1}{|E(t-k)|} |E(t) \cap E(t-1) ... E(t-k+1) \cap E(t-k)|, \qquad (1.2)$$

Here $E(t)$ refers to the set of edges present at time t, $\cap$ is the intersection operator and $|...|$ gives the number of elements in the set. The fraction is normalized by $|E(t-k)|$, the maximum number of edges that can survive. We define a

characteristic time for the network, the so called *half-life* of the survival ratio $t_{1/2}$ as the (interpolated) time interval in which half of the initial connections have decayed, i.e., $\sigma(t, t_{1/2}/\delta T) = 0.5$. As an important special case for $k = 1$ we obtain *single-step survival ratio* $\sigma(t, 1) = \sigma(t)$, giving the fraction of edges found common in two consecutive steps. With these measures it is expected that while some of the differences can reflect real changes in the network topology, others may simply be due to noise. Some results are shown in Fig. 1.1 and Table 1.1.

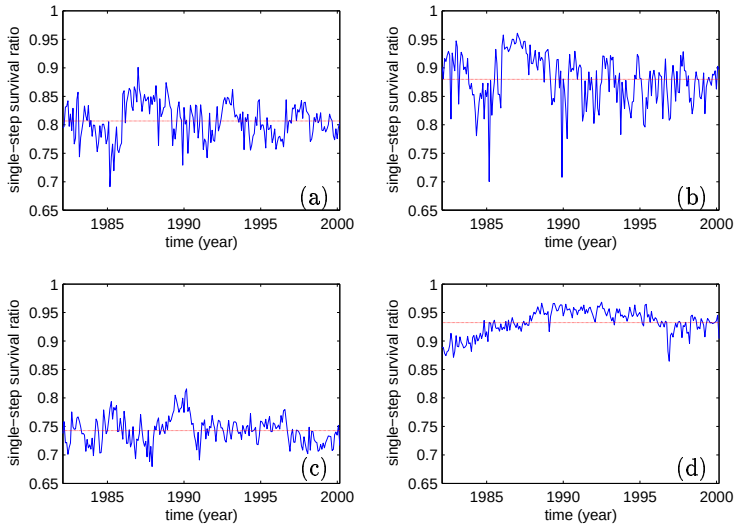

Figure 1.1: Single-step survival ratio $\sigma(t)$ as a function of time: (a) asset tree and (b) asset graph for flow $F(t)$, (c) asset tree and (d) asset graph for logarithmic relative flow $\dot{F}(t)$. The dashed lines indicate the corresponding average values $\bar{\sigma}$.

Earlier we have studied single-step survival ratios for logarithmic returns for a fixed sized $N = 477$ NYSE dataset over the same period, and obtained for the tree $\bar{\sigma} = 82.6\%$ and graph $\bar{\sigma} = 94.8\%$, which are well compatible with the present results. Thus, the measure generalizes well for the larger dataset, and the slightly smaller values obtained here result, most likely, from allowing stocks to enter and leave the market. The infamous 1987 Black Monday crash can be seen clearly only in the return based tree and graph (not shown, see [4]) and in the capital flow graph in Fig. 1.1(b), as evidenced by the two prominent dips located symmetrically around the crash, window width T apart. Also, the single-step survival plot for logarithmic relative volumes (not shown) resembles very closely that of logarithmic relative flows. This is to be expected from the definition of the latter, given that fluctuations in volume dominate over fluctuations in price.

To establish a reference, we have repeated the study for simulated data. From the definition of logarithmic return $r_i(\tau) = \ln P_i(\tau) - \ln P_i(\tau - 1)$, we obtain a recursive formula for the simulated price as $P_i'(\tau) = P_i'(\tau - 1)e^{r_i^*}$, where r_i^* denotes a sample from the empirical return distribution for stock i, determined

	TREE					GRAPH				
	$\bar{\sigma}^e$	$\bar{\sigma}^s$	std σ^e	$t^e_{1/2}$	$t^s_{1/2}$	$\bar{\sigma}^e$	$\bar{\sigma}^s$	std σ^e	$t^e_{1/2}$	$t^s_{1/2}$
P	78.0%	74.7%	2.7%	0.25	0.21	87.0%	83.8%	2.7%	0.43	0.33
r	79.9%	71.6%	6.8%	0.37	0.21	94.5%	75.3%	5.5%	1.59	0.25
V	82.3%	81.7%	2.6%	0.39	0.37	89.3%	86.8%	4.8%	0.68	0.52
$\dot{V}$	74.1%	70.5%	2.5%	0.25	0.20	93.2%	73.6%	2.3%	1.99	0.23
F	80.7%	75.5%	3.2%	0.34	0.25	88.0%	83.8%	4.3%	0.56	0.39
$\dot{F}$	74.2%	70.7%	2.4%	0.25	0.20	93.2%	73.9%	2.2%	1.99	0.23

Table 1.1: Mean $\bar{\sigma}^e$ and standard deviation std σ^e of the single-step survival ratio for empirical data, and the mean $\bar{\sigma}^s$ for simulated data. Empirical and simulated half-lives $t^e_{1/2}$ and $t^s_{1/2}$, respectively, are measured in years.

from the price quotes within the window. For 'new' stocks the simulation starts from the first actual price quote in the window so that $P_i^{'}(1) = P_i(1)$, but for 'old' ones only the additional δT prices are simulated. This approach is loyal to the underlying return distributions, but removes the correlation structure present in the market. Assuming price and volume to be statistically independent, i.e., $\Pr[V_i(\tau) = V | P_i(\tau) = P] = \Pr[V_i(\tau) = V]$, volumes can be simulated by sampling directly from the empirical distribution, giving $V_i^{'}(\tau) = V_i^*$ for all τ. The simulated results are also shown in Table 1.1.

1.5 Scale-free topology

During the last few years, considerable attention has been devoted to studying different complex networks describing a wide range of systems in nature. In particular, the degree distributions of complex networks have come under scrutiny, and studies have indicated that scale-free networks, i.e., networks where the distribution follows a power law, occur very frequently in various fields, from human relationships through cell metabolism to the Internet [2]. Recently, examples of scale-free networks have also been found in economic and financial systems. For example, Vandewalle et al. [3] found scale-free behavior for an asset tree based on logarithmic returns in a one year (1999) time window for stocks traded at the NYSE, NASDAQ and AMEX, and proposed that the distribution of the vertex degrees $f(k)$ follows a power law

$$f(k) \sim k^{-\alpha}, \tag{1.3}$$

with the exponent $\alpha \approx 2.2$. We studied the dynamics of scaling properties of a return based asset tree for the fixed sized $N = 477$ NYSE dataset mentioned earlier. The exponent was found to fluctuate, to some extent, over time, and the goodness of power law fits, as measured by the R^2 *coefficient of determination*, also varied, sometimes masking the scale-free behavior, especially if no outliers were removed.[3] When we analyzed the time period more carefully, we observed

[3] In the tree, the vertex that is most clearly an outlier corresponds to the central node, or the root of the tree [4]. The reported results refer to the case where the outlier is included.

scale-free behavior with a rather robust exponent $\alpha \approx 2.1 \pm 0.1$ ($R^2 \approx 0.86$) for business as usual times, whereas for the period containing Black Monday we obtained $\alpha \approx 1.8 \pm 0.1$, reflecting on a tighter tree configuration caused by the crash. Still using log-returns, for the asset graph we obtained a significantly lower value of $\alpha \approx 0.9 \pm 0.1$ ($R^2 \approx 0.75$), but in this case the evidence for scale-free behavior was not conclusive, and we did not observe distinctively different topologies for normal and crash markets [4].

Adopting the approach proposed in this paper, using log-returns for the asset tree, we now obtain $\alpha \approx 1.8 \pm 0.2$ ($R^2 \approx 0.87$), and for the graph $\alpha \approx 1.0 \pm 0.1$ ($R^2 \approx 0.78$). These findings are well in line with our earlier work, given the different approach taken here with larger and non-constant dataset size $N(t)$. Working with asset tree topology resulting from capital flows F we obtain $\alpha \approx 2.0 \pm 0.2$ ($R^2 \approx 0.88$) so, rather surprisingly, the system's topology is, on average, marginally closer to 'pure' scale-freeness than for the returns, with a similar value for the exponent. For the capital flow graph, we obtain $\alpha \approx 1.1 \pm 0.2$ ($R^2 \approx 0.82$), where the exponent is clearly lower in comparison to the related tree. Using logarithmic relative flow $\dot{F}_i(\tau)$, for the tree and graph we get $\alpha \approx 1.7 \pm 0.2$ ($R^2 \approx 0.80$) and $\alpha \approx 0.9 \pm 0.1$ ($R^2 \approx 0.77$), respectively. Degree distributions for F and $\dot{F}$ in selected time windows are shown in Fig. 1.2.

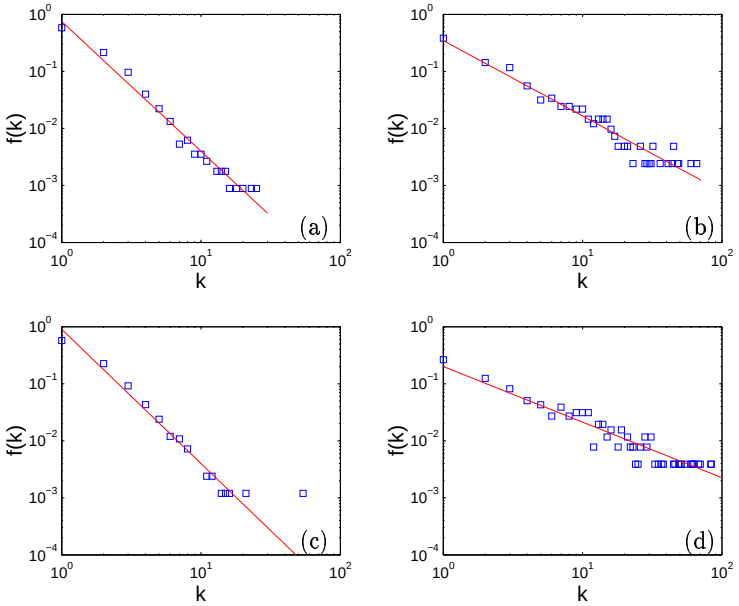

Figure 1.2: Examples of vertex degree distribution plots: (a) asset tree ($t = 117$, $\alpha \approx 2.28$, $R^2 \approx 0.98$) and (b) asset graph ($t = 119$, $\alpha \approx 1.32$, $R^2 \approx 0.94$) for flow $F(t)$, (c) asset tree ($t = 4$, excluding the outlier $\alpha \approx 2.35$, $R^2 \approx 0.98$) and (d) asset graph ($t = 152$, $\alpha \approx 0.97$, $R^2 \approx 0.89$) for logarithmic relative flow $\dot{F}(t)$.

1.6 Concluding remarks

We have studied correlation based financial networks and found network robustness to depend on the input quantity used. In general, asset graphs are topologically stronger than the related asset trees. We have also studied degree distributions for networks based on logarithmic returns, capital flows, and logarithmic capital flows. Asset trees were found to display scale-free behavior, with exponents close to 2. For asset graphs, the evidence is not as clear, although there are indications of scale-freeness, and in this case exponents are close to 1.

Acknowledgments

We are thankful to Professor Marcel Ausloos for encouraging us to involve trading volumes in our studies. J.-P. O. is grateful to the Graduate School in Computational Methods of Information Technology (ComMIT), Finland. This research was partially supported by the Academy of Finland, project no. 44897 (Finnish Center of Excellence Program 2000-2005).

Bibliography

[1] Mantegna, R. N., and H. E. Stanley, *An Introduction to Econophysics; Correlations and Complexity in Finance*, Cambridge University Press, (2000). Mantegna, R. N., "Hierarchical structure in financial markets", *Eur. Phys. J. B* **11** (1999), 193–197.

[2] Albert, R. and A.-L. Barabasi, "Statistical mechanics of complex networks", *Rev. Mod. Phys.* **74** (2002), 47–97. Dorogovtsev S.N. and J.F.F. Mendes, *Evolution of Networks: From Biological Nets to the Internet and WWW*, Oxford University Press, (2003). Dorogovtsev S. N. and J. F. F. Mendes, "Evolution of networks", *Advances in Physics* **51** (2002), 1079–1187.

[3] Vandewalle, N., F. Brisbois, and X. Tordoir, "Non-random topology of stock markets", *Quantitative Finance* **1** (2001), 372–374. Marsili,M. "Dissecting financial markets: Sectors and states", preprint available at cond-mat/0207156 (2002). Yang I., H. Jeong, B. Kahng, and A.-L. Barabasi, "Emerging behavior in electronic bidding", preprint available at cond-mat/0301513 (2003).

[4] Onnela, J.-P., A. Chakraborti, K. Kaski, and J. Kertesz, "Dynamic asset trees and portfolio analysis", *Eur. Phys. J. B* **30** (2002), 285–288, and references therein. Onnela, A. Chakraborti, K. Kaski, J. Kertesz, and A. Kanto, "Dynamics of market correlations: Taxonomy and portfolio analysis", *Phys. Rev. E* **68** (2003), 056110. Onnela, J.-P., K. Kaski, and J. Kertesz, "Clustering and information in correlation based financial networks", *Eur. Phys. J. B* (2004), in press.