Quantitative Finance

ISSN: (Print) (Online) Journal homepage: www.tandfonline.com/journals/rquf20

Network analysis of aggregated money flows in stock markets

Joonas Karaila, Kestutis Baltakys, Henri Hansen, Anubha Goel & Juho Kanniainen

To cite this article: Joonas Karaila, Kestutis Baltakys, Henri Hansen, Anubha Goel & Juho Kanniainen (22 Oct 2024): Network analysis of aggregated money flows in stock markets, Quantitative Finance, DOI: 10.1080/14697688.2024.2409272

To link to this article: https://doi.org/10.1080/14697688.2024.2409272

Published online: 22 Oct 2024.

Quantitative Finance, 2024
https://doi.org/10.1080/14697688.2024.2409272

Network analysis of aggregated money flows in stock markets

JOONAS KARAILA†, KESTUTIS BALTAKYS‡, HENRI HANSEN‡, ANUBHA GOEL‡ and JUHO KANNIAINEN *‡

†Nordea Bank Abp, Helsinki, Finland
‡Research Group of Financial Computing and Data Analytics, Tampere University/Computing Sciences, Tampere, Finland

(*Received 26 April 2023; accepted 19 September 2024*)

We introduce the formation of a network of money flows between assets in stock markets, which captures directed relations between assets in terms of how investors have re-allocated money in the stock exchange. Our approach is based on identifying a directed link, or money flow, that occurs when an investor funds a purchase of an asset by selling another asset(s). We extract investor-level money flow networks on daily basis from shareholder registration data, which are then aggregated for both financial institutional and retail investors. Overall, we have a time series of 877 daily networks from 2006 to 2009, which is exceptionally long data on temporal networks. Through our analysis of non-reciprocated triadic patterns in the aggregated money-flow networks, we find that these patterns are both recurrent and significant. However, they are not related to the 2008 financial crisis. Additionally, we observe that the counts on different triadic motifs exhibit not only an autoregressive process but are also interconnected contemporaneously and dynamically. These findings suggest the need for further research using sophisticated network models to provide a comprehensive representation of money flows in stock markets.

Keywords: Money flow network; Asset allocation; Graphs; Complex networks; Motifs

JEL Classifications: C00, C40, G10

1. Introduction

The application of network science to study various financial phenomena has increased in recent years. For example, networks have been used to analyze the World Trade Web (Squartini *et al.* 2013), systemic risk and financial contagion (Haldane and May 2011, Acemoglu *et al.* 2015), investor networks in stock markets (Ozsoylev *et al.* 2014, Baltakiene *et al.* 2021, Baltakys 2023), and investors portfolio strategies (Gualdi *et al.* 2016). In this paper, we introduce the formation of a network of money flows, which represents a completely new kind of network analysis in stock market research.

The net money flow network describes how investors transfer money from one set of assets to another set of assets. Practically, one can say that an investor has funded purchased assets with the assets they sold on the same day. At the same time, an investor can adjust their stock market portfolio by injecting or extracting money to/from it, which can also be analyzed using our approach. For a given day, the money flow network with directed edges is created based on transactions, which are gleaned from investor-level transaction data. The net amount of money that flows between the assets is proportional to the euro-volumes of the assets traded. If an investor has bought more assets than sold in terms of euro volume during a given time period, she has injected money into the portfolio. Conversely, if she has sold more assets than bought, she has extracted money from the portfolio. The difference between the buy and sell euro volumes is calculated and added to an additional *balance node* to the system. The balance node represents the injections and withdrawals of money to stock markets. Once the investor-level networks are obtained on a given day, they are aggregated over all investors. The resulting aggregated networks, which are observed on daily basis, show the directed relations between assets in the stock exchange in terms of the net money flows.

*Corresponding author. Email: juho.kanniainen@tuni.fi

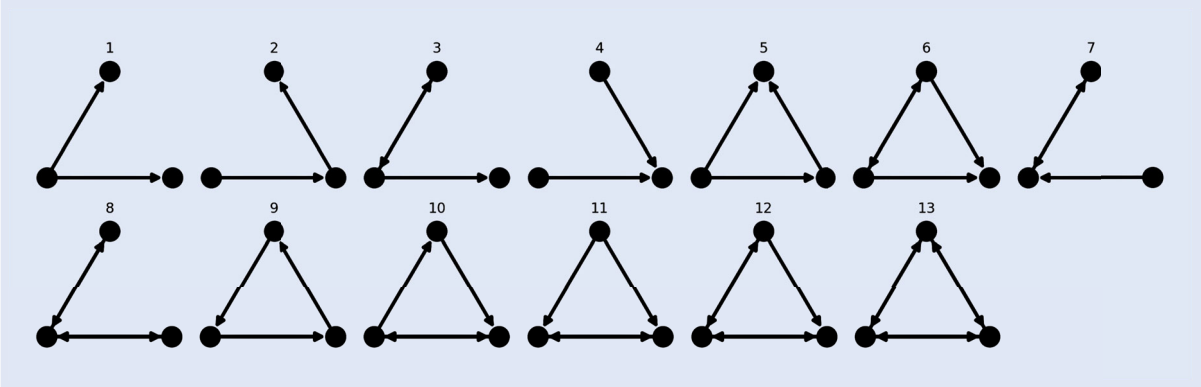

Figure 1. The 13 possible non-isomorphic triadic subgraphs (motifs). Only motifs 1, 2, 4, 5, and 9 are possible in the money flow networks because there cannot be money flows in both directions between assets by the definition of the networks.

Our investor-level transaction data from Helsinki Stock Exchange spans from 2006 to 2009 and includes 877 trading days. For each day, we extract an aggregated money flow network, which then results in an empirical series of networks. Consequently, we actually analyze a temporal network, where the edges are not continuously active (for temporal networks, see, for example, Holme and Saramäki 2012). Temporal networks have recently been analyzed with so-called *motif counts*. In the previous literature, network motif analysis has been used to study cash flows between banks during financial crises (Squartini *et al.* 2013), problem-solving networks (Braha 2020) and social networks (Holland and Leinhardt 1974), as well as biological networks (Ingram *et al.* 2006, Alon 2007). Motifs can be thought of as small connected subnetworks which appear in an observed network significantly more often than expected in a null hypothesis model of random networks. In this paper, we focus on five particular motifs that include no two-way directed links, i.e. non-reciprocated triadic motifs. This restriction comes from the way the money flow networks are formed: two assets (nodes) can have a directed money flow (edge) only in one direction because a directed edge represents the aggregated *net*-euro-volume from one asset to another.

The aggregated network can be analyzed in terms of motifs (or sub-graphs). Figure 1 illustrates 13 variations of triadic motifs, of which the aggregated net money flow network can represent five (Nos. 1, 2, 4, 5, and 9). In the context of this paper, the motifs, as further discussed in the paper, describe the local aggregated money flow structures between the securities. For example, the motif counts for motif No. 4 describe the prevalence of days where, on an aggregated level, investors have decided to concentrate money into one security from two or more securities. It is important to note that this does not mean that individual investors concentrate from multiple securities into one, but this occurs at the aggregated level, indeed. In other words, a group of investors, such as households, can collectively concentrate ownership, which we believe could be related to the general stock market dynamics. Similarly, Motif 1 describes the situation where investors, at the aggregated level, have exchanged one security for two (or more) securities; Motif 2 represents a chain-type structure, where money flows via one security to a third; Motif 5 depicts a case where both diversification and concentration, as well as chain-type motifs, are combined. Finally, Motif No. 9 represents a money loop.

The main question of this paper is whether motifs (the five non-reciprocated triadic) in the aggregated money flow network are overexpressed, which would indicate that the money flows do not form randomly. This paper finds that almost all the motifs are significantly overexpressed compared to two null models and that they repeat themselves in the aggregated money flow networks. This finding is robust with respect to both approaches 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 (directed random graph and directed configuration model, respectively). This indicates the underlying nonrandom structural principles that might have been involved in investors' capital allocation decisions in the stock market.

The second question is whether the motif counts in the aggregated money flow reacted to the 2008 financial crisis. This paper finds no evidence of this, which we consider a bit surprising, especially since, for example, Squartini *et al.* (2013) shows an abrupt change in 2008 in the interbank networks in terms of the motif counts. Of course, the interpretation of the money flow network, and thus also the motifs, is very different from those in interbank networks. While nodes in an interbank network are banks, which are the actors, in our case the nodes are securities, which are the objectives of transactions. Therefore, the links also signify different things. Here, they represent money flows, aggregated over investors, from one security to another, whereas in the interbank networks, the links represent exposures between the banks. However, as there is existing evidence that the 2008 financial crisis affected portfolio rebalancing strategies (Vermeulen 2013), one could expect that the money flows between securities could change significantly and long-term, especially at the aggregated level, around the crisis. Even though this finding is negative, we believe that it is valuable because it adds to the understanding that the money flow network did not change significantly in structure (from the motifs' point of view) around the financial crisis.

Third, one could postulate that motif count dynamics can follow complicated autoregressive processes, where they interact with each other. For example, if a certain kind of motif is common today, it can predict some other motif type

to become prevalent tomorrow. Our analysis shows that the dynamics of different motifs are not only non-random but also cannot be considered independent. More specifically, we find that motif counts on the five motifs are both contemporaneously and dynamically associated. By impulse response analysis built on VAR model on motif counts, we identify two motifs whose shocks have an immediate impact on the counts of all the other motifs. The count of each motif is driven by its previous values, exhibiting AR(1) type dynamics, and moreover, there are cross-sectional, lagged associations between the counts of several motifs. This indicates that the dynamics of the money flow network structure are non-trivial as its subgraphs are driven by each other.

This paper is also related to the existing research on flows in the financial markets research domain. To highlight the most relevant and important paper, Gabaix and Maggiori (2015) study the exchange rate determination based on capital flows in imperfect financial markets. Moreover, Vayanos and Woolley (2013) find that flows between investment funds are triggered by changes in fund managers' efficiency, which an investor can observe. Recently, Gabaix and Koijen (2021) published a paper on inelastic markets, which analyzes capital flows in and out of the aggregate stock markets, finding that the flows have a significant impact on prices and risk premia. However, to the best of our knowledge, in contrast to the present paper, there is no research that examine money flows between individual assets in stock markets.

Overall, the methodological contribution of this paper is that we formalize the money flow networks in stock markets. Moreover, we provide empirical contribution by analyzing the network motifs (subgraphs) and their dynamic relations. By analyzing the network structure of these money flows, we can identify patterns and interconnections that may not be immediately apparent from other types of data. These findings can be useful for investors, financial analysts, and policymakers interested in understanding the behavior of investors in stock markets and the factors that influence the dynamics of securities.

2. Money flow network formation

In this section, we describe how money flow networks are formed. First, money flows are approximated between assets for individual investors. Second, multiple networks of individual investors are aggregated together, and the net money flows are calculated between assets. The resulting money flow representation of the transactions during an observation period can be further studied using network analysis methods.

Let $s_{i,l}$ be the total amount of money investor l, $l = 1, 2, \ldots, L$, received from the sell transactions and $b_{i,l}$ the total amount of money they used to purchase asset i in a given day, or more generally in a given time period. The net-flow is $v_{i,l} = s_{i,l} - b_{i,l}$. If $v_{i,l} > 0$, an investor is a net-seller and thus receives more money than pays. If $v_{i,l} < 0$, an investor is a net-buyer, thus paying more money than receiving on trading asset i.

Let us assume that there are $M - 1$ distinct assets available for investors to trade during the day. It is now possible that $\sum_{k=1}^{M-1} v_{k,l} \neq 0$, which would mean that an investor has either extracted from (> 0) or injected money into (< 0) the stock market. We solve this by creating a synthetic asset, called 'Balance', defined by the residual euro volume, $v_{M,l} = -\sum_{k=1}^{M-1} v_{k,l}$. By including the Balance Node, by definition, we have $\sum_{k=1}^{M} v_{k,l} = 0$, and thus an investor always has a balanced network between different assets.

Suppose that investor l has net-sold at least one asset and net-bought at least one other asset, i.e. $v_{i,l} > 0$ for some i and $v_{j,l} < 0$ for some j, $i \neq j$. The idea is to distribute the money flow from the asset i an investor sold to the assets s/he bought on the day in terms of proportional euro-volumes. Formally, a money flow (the weight of an edge) from stock i to asset j in investor l's money flow network is defined by

$$e_{i,j}^{(l)} = v_{i,l}^{+} \frac{v_{j,l}^{-}}{\sum_{k=1}^{M} v_{k,l}^{-}}, \tag{1}$$

where $v_{i,l}^{+} = \max(v_{i,l}, 0)$ is the positive part and $v_{i,l}^{-} = \min(v_{i,l}, 0)$ negative part of $v_{i,l}$. Investor l's adjacency matrix $\boldsymbol{E}^{(l)} \in \mathbb{R}^{M \times M}$ is constructed from edges $e_{i,j}^{(l)}$ by setting $\boldsymbol{E}^{(l)}(i,j) = e_{i,j}^{(l)}$. Some remarks on the definition:

- Given that investor l net-sold asset i, the ith row of the adjacency matrix $\boldsymbol{E}^{(l)}$ shows the distribution of the money flow from asset i to the other assets.
- Correspondingly, given that asset j was net-bought, column j shows the money flows from other assets to jth asset.
- The money flows always have a non-negative sign.
- There is a money flow between assets i and j if and only if i is sold and j is bought. Otherwise, the link value is zero.
- In an investor-level network, there are no nodes with both incoming and outgoing edges.
- There are no self-loops, i.e. $e_{i,i} = 0$ for all $i = 1, 2, \ldots, M_l$.
- The total money flow from node i to all the nodes is $\sum_{j=1}^{M_l} e_{i,j}^{(l)} = v_{i,l}^{+}$.
- Moreover, the total money flow to node j from all the other nodes is

$$\sum_{i=1}^{M} e_{i,j}^{(l)} = v_{j,l}^{-} \frac{\sum_{i=1}^{M} v_{i,l}^{+}}{\sum_{k=1}^{M} v_{k,l}^{-}} = v_{j,l}^{-},$$

  because $\sum_{k=1}^{M} v_{k,l} = 0$ when the Balance Node is included.
- Consequently, there is no other net inflow or outflow for a single asset than the amount the investor has bought or sold/bought.

Overall, the way the cash flows are weighted in (1) is based on the source and target asset net euro cash flows.

To illustrate (1), assume there are 4 asset. Let asset 1 be bought with 400 EUR and sold for 300 EUR, asset 2 bought with 200 EUR, asset 3 bought with 100 EUR and sold for 150 EUR, and asset 4 sold for 250 EUR in total by a given investor during a given day. Now investor's net euro money flows for the assets would be -100, -200, 50, and 250 EUR,

respectively. As the money flows over the four assets are in balance, the 'Balance' node has no in or out flow. Using (1) to calculate edge weights yields

$$\boldsymbol{E}^{(l)} = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 50 \times \frac{100}{300} & 50 \times \frac{200}{300} & 0 & 0 & 0 \\ 250 \times \frac{100}{300} & 250 \times \frac{200}{300} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix} = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 50/3 & 100/3 & 0 & 0 & 0 \\ 250/3 & 500/3 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}.$$

The money flow network aggregated over all the $1, 2, \ldots, L$ investors is denoted by $\overline{\boldsymbol{E}}$ with

$$\bar{e}_{i,j} = \left(e_{i,j} - e_{j,i}\right)^{+}, \tag{2}$$

where

$$e_{i,j} = \sum_{l=1}^{L} e_{i,j}^{(l)}.$$

We choose to make all negative aggregated edges zero, as otherwise there would be double counting in the net money flows from asset i to j (one positive from i to j and another one negative, with with the same magnitude, from j to i).

To provide an example of aggregating investor networks together, let

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

$$\boldsymbol{E}^{(B)} = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 200 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 100 & 0 & 0 \\ 0 & 0 & 50 & 0 & 0 \end{bmatrix}.$$

The aggregation results in

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

In this example, investor A and B money inflow and outflow for asset 2 are in opposite directions, which leads to no net money flows for asset 2. Money did still flow through asset 2 from asset 4 to asset 3, which is important when studying the system as a whole later on. The network is visualized in figure 2.

A couple of remarks on the aggregated networks:

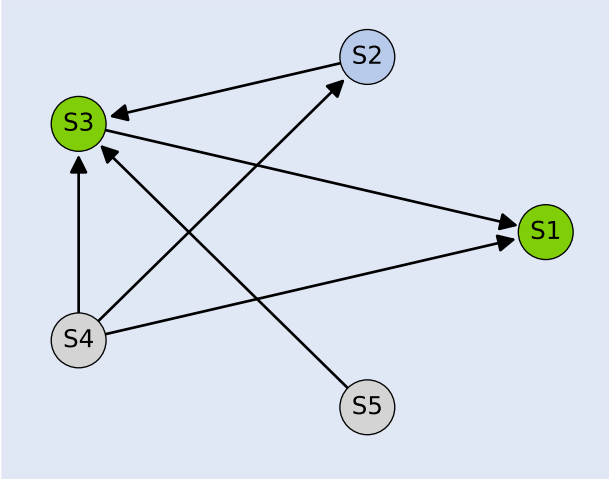

Figure 2. Aggregated money flow network presented in (3). Green nodes have net positive money flows, grey nodes have net negative money flows and white nodes have 0 net money flows. The size of the arrows illustrates the size of the money flow.

- The size of the aggregated network is $M \times M$, where M is the total number of assets the investors traded, including the balance node.
- Whereas a node in the investor-level networks can have either incoming or outgoing links in a given time-period, in the aggregated network a node can have both types of links (e.g. nodes 2 and 3 in the above example).
- If the money flow networks were aggregated over all the possible investors who have traded the assets under the analysis, there would be no residual money flow for any asset (node). That is, for any i, $\overline{\boldsymbol{E}}_{i*}\mathbf{1} = (\overline{\boldsymbol{E}}_{*i})'\mathbf{1}$, i.e. the sum of the elements in the ith row is equal to that of the ith column, because every asset bought would be an asset sold by another investor.
- However, given that a subset of investors is analyzed, such as retail investors or financial institutions, then an asset (node) can have residual money flow, either positive or negative (see figure 2). This is the case in the empirical analysis of our paper.

In this paper, the length of the time-period from which the aggregated networks are observed is one (trading) day. Consequently, we have daily observations on the aggregated money flow network, $\{\boldsymbol{E}(t); t = 1, 2, \ldots, T\}$, where $T = 877$ is the total number of days in our data set. The link weights are used to aggregate the investor-level networks, but once the aggregated network is made observable, then the links are digitalized, i.e. only binary links are considered. More specifically, a directed link from node i to node j is considered to exist on day t if $\bar{e}_{i,j}(t) > 0$

3. Network motifs

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 *et al.* 2004). In this paper,

the focus is on connected triads, or so-called triadic motifs, which are visualized in figure 1 up to isomorphism. For example, motif 5 illustrated in figure 1 means there is at least one asset which is seen as a source of money, and at least one node, which is seen as a sink of money, while for motif 9, there is no clear consensus of sources and sinks. As the edges are defined by (1), there are no reciprocated edges, meaning only motifs 1, 2, 4, 5, and 9 are possible in the networks. Hence the problem is limited to only a few graphs with 3 nodes, and the problem can be solved easily. It is possible to count motifs analytically from an adjacency matrix of a network (see Squartini *et al.* 2013).

Motif discovery from a network contains three different steps: (Patra and Mohapatra 2020)

(1) Different size motifs are extracted from the network.
(2) Frequencies of the found motifs are counted.
(3) Determining the statistical significance of the motifs by comparing the found frequencies to frequencies found in random networks.

In the first step of the process, there is also a need to know the different isomorphic forms of the motif, which is known to be an NP-complete problem (McKay and Piperno 2014). The third step multiplies the computational cost by sampling frequencies from randomly generated networks.

We use two approaches in motif discovery: the first uses a grand canonical ensemble of directed random graphs, while the second uses a sampling of directed graphs with a given degree sequence. By a grand canonical ensemble, it is possible to derive an analytical first-order approximation for the standard deviation for the number of motifs in a directed random graph model. In the case of random graphs with a given degree sequence, it is possible also to derive a first-order approximation of the number of motifs, but it does not scale with the number of nodes in the network as well as with the directed random graph model. Therefore, to speed up the computation, a sample of graphs is used in the second approach.

Different statistical measures can be used to determine the statistical significance of motif counts. Milo *et al.* (2002) uses frequency, p-value, and Z-score. The frequency of a motif is a simple cut-off value at which the motif is said to be statistically significant. This can be useful for larger motifs as the expected abundance of them is lower. For example, Kashani *et al.* (2009) uses a frequency of 4 to determine that a motif is significant. For a given motif, the p-value is the cut-off probability value for observing a particular number, or more, of the motif in a randomized network. Milo *et al.* (2002) uses a probability cut-off value of 0.01. The last measure (Milo *et al.* 2002) uses is a standard score which is calculated by

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

where N_m is the number of Motif m and $\langle , \rangle$ are the expected values for the number of motif count and standard deviation in the set $R(G) \subseteq \Omega(G)$ of random graphs. Kashani *et al.* (2009) uses a 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.

This paper uses Z-score as the measure of how dissimilar the observed networks are from the randomized networks. The problems of underestimating the count of motifs and comparing different-sized networks should have almost no impact on the analysis. 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 cash flows between assets are modeled by the random network, and thus the cash flow directions are random.

A basic null model for a binary directed network is the *directed random graph*. Another null model used in this paper is the *directed configuration model*, which is used to generate random networks with a given degree sequence. The upside of 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 into heads and tails, depending on whether the link is directed to or from the vertex. The calculation of the expected number and the standard deviation of motif counts under the random graph and directed configuration model are addressed in the Appendix.

4. Data description

Networks analyzed in this paper are constructed from transaction data from Euroclear Finland. The networks are extracted for each trading day from February 2006 to July 2009 (877 days in total). Transactions are grouped based on whether they involve financial institutions or retail investors. In total, transaction data includes transactions for 608 financial institutions and 316 567 retail investors during the period. Financial institutions had 340 809 transactions, while retail investors had 5 331 006 transactions. Daily money flow networks are extracted for each individual investor and then aggregated using (2). Only assets that were listed in the stock market during the period were included. This means we have 134 nodes (assets) in the network plus the balance node.

Table 1 presents the basic network statistics. The values are averages and standard deviations of the values from the time series. The 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 the transitivity of a network by giving higher weight to nodes with low degree (Newman 2018) and by measuring the cliquishness of a typical neighborhood (Watts and Strogatz 1998). The average shortest path is not provided because the networks are not always connected.

Table 1. Average and standard deviations of links counts L, average degree k, and clustering coefficient c in addition to the node count for all different network time series. M is the number of assets (nodes).

	M	$(L, \sigma(L))$	$(k, \sigma(k))$	$(c, \sigma(c))$
Financial institutions	134	(875, 245)	(3.05, 1.14)	(0.15, 0.02)
Retail investors	134	(780, 194)	(3.70, 1.36)	(0.24, 0.03)

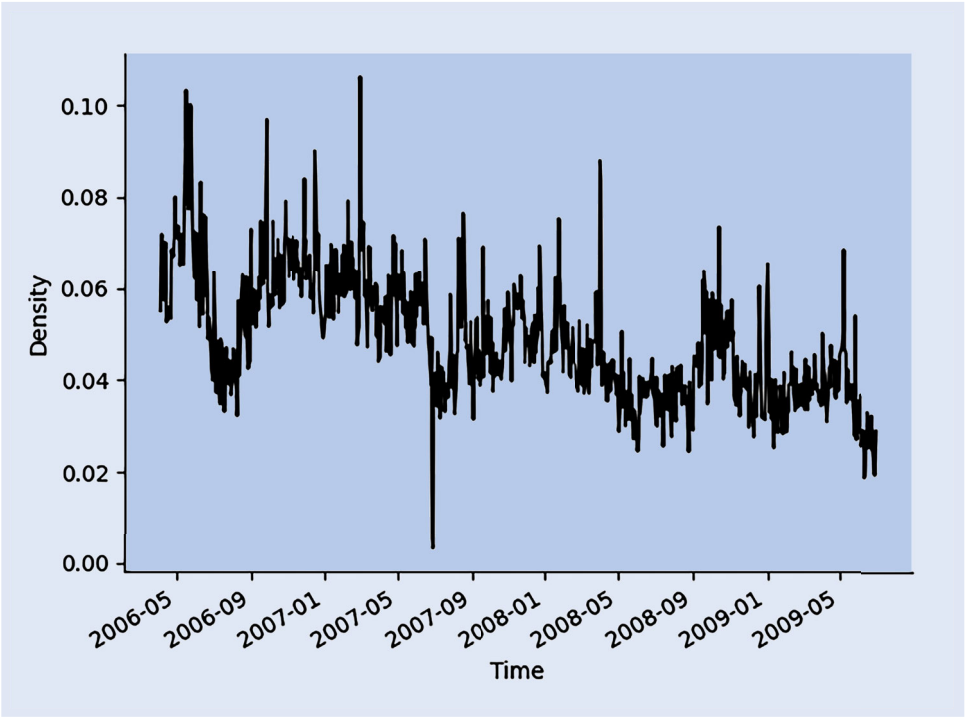

Figure 3. Network density dynamics based on the transactions of financial institutions on daily basis.

In table 1, all of the networks have a rather low number of edges, and the average clustering coefficients are quite high when the low number of edges is considered. A configuration model with similar degree distribution would give a clustering coefficient of $8.059 \cdot 10^{-4}$, a difference of nearly three orders of magnitude. The maximum number of edges in the networks is 8 911. Standard deviations of the clustering coefficients are low, which means there are no large changes in 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.

Network density affects the degree centrality measures and motif counts, as higher density leads to more edges. The maximum density is 0.5, because there cannot be edges in both directions between two nodes. In figures 3 and 4, daily network time series have their densities plotted from 2006 to 2009 for institutions and retail investors, respectively. Networks extracted from the transactions of financial institutions show densities with a negative trend in comparison to the networks based on retail investors' transactions. However, in both cases, we can conclude that the networks are rather stable over time. There is one downward peak in the summer of 2007, and a couple of upward peaks in 2007, 2008, and 2009, which appear to be statistical outliers. We attempted to find specific macro and micro events related to these peaks but could not identify any, suggesting that some unexplainable events may be associated with them.

In the Appendix, we provide additional descriptive statistics to describe quantile statistics of the numbers of investors trading a given asset and conversely the number of assets traded by an investor in a day.

Table 2 shows the assets that appeared to be the most central nodes in the money flow networks over 877 trading days in our data sample. It can be seen that these assets are either large companies or the balance node. Large companies are traded more and thus they have more edges which give them high values for centrality measures.

5. Results

5.1. *Z-scores of motif counts*

As demonstrated by figures A1 and A2 in the Appendix, the counts of motifs 1, 2, 4, and 5 are approximately on the same level, while the appearance of motif 9 (loop) is much more infrequent. However, from motif counts alone, we cannot directly determine if they are overrepresented; we have to analyze their Z-scores. In Appendix, figures A3 and A4 show the dynamics of Z-scores under the directed random graph model with data coming from institutional and retail

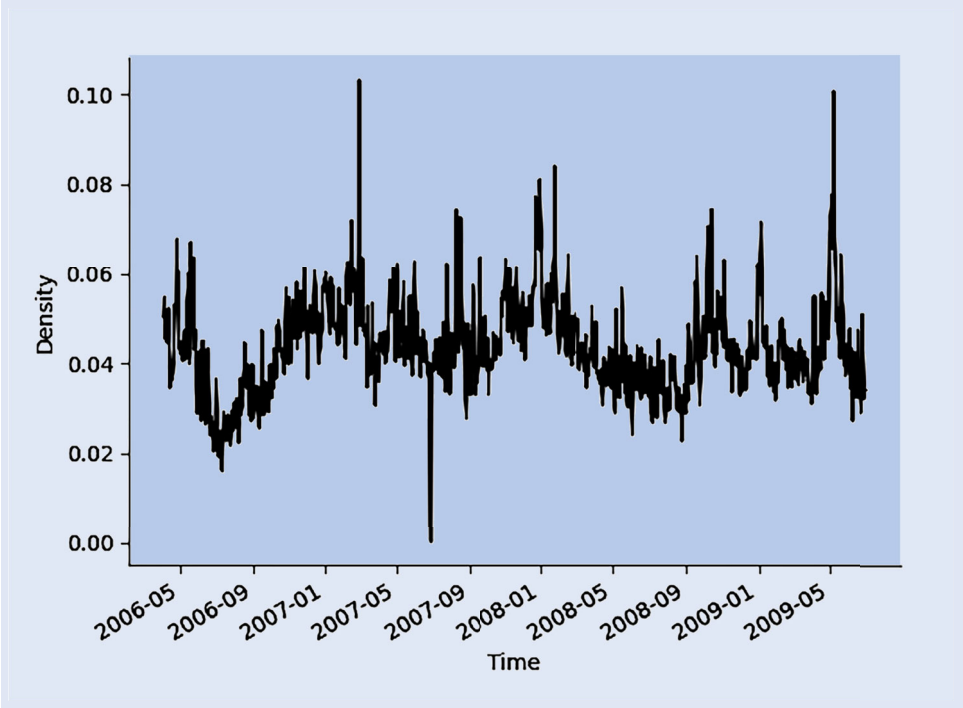

Figure 4. Network density dynamics based on the transactions of retail investors on daily basis.

Table 2. The appearance of assets in the top 0.1 percentile in daily money flow networks according to in-degree, out-degree, and betweenness centralities (% of total trading days). The total number of trading days is 877.

In-degree		Out-degere		Betweenness	
		Panel A: Financial institutions			
Balance	75.48%	Balance	71.95%	Balance	92.70%
Nokia	54.05%	Nokia	50.29%	Nokia	79.70%
Neste	46.86%	Fortum	48.12%	UPM Kymmene	63.40%
UPM Kymmene	44.81%	UPM Kymmene	47.66%	Fortum	61.12%
Fortum	40.48%	Outokumpu	44.70%	Neste	57.13%
		Panel B: Retail investors			
Balance	92.13%	Balance	92.47%	Balance	92.70%
Neste	53.36%	Nokia	64.88%	Nokia	77.65%
Metso	51.31%	Outokumpu	50.63%	Neste	57.81%
Nokia	50.29%	Rautaruukki	46.41%	Metso	52.22%
Telia	47.43%	Fortum	45.95%	Outokumpu	51.54%

investor traders, respectively. Moreover, figures A5 and A6 show the corresponding dynamics under the directed configuration model. Importantly, in contrast to interbank networks (Squartini *et al.* 2013), there are no observable changes in the Z-scores of motif counts of money flow networks around the 2008 financial crisis.

Table 3 shows summary statistics for the Z-scores calculated using both Directed Random Graph (DRG) and Directed Configuration Model (DCM). Moreover, results are reported separately for daily networks extracted from the transactions of financial institutions and retail investors. Panel A, which reports the statistics for the Z-scores, shows that compared to DCM, DRG assigns very high values for motifs 4 and 5, and from this point of view, DCM is more conservative. With DCM, the standard deviations show some time variation in Z-scores over the trading days, but they are clearly lower than the mean values. Overall, we find that the Z-scores are statistically significant, which is further demonstrated by figure 5, which shows the empirical cumulative distributions of motifs' Z-scores with different null models and for data observed on institutions and retail investors. The figure also plots the vertical lines $z \pm 2$ for the region within two standard deviations from the model's expectations. Several observations can be drawn. First, the DCM, which we consider a better null model as it is based on the actual degrees of the nodes, yields Z-scores that are greater than zero with statistical significance.† Overall, we provide strong evidence that the five

† The use of DRG implies significant values for the Z-scores of the motifs with data coming from retail investors' transactions, while the results are partially mixed with data on financial institutions (yet, just on motifs 2 and 9). On the other hand, DRG can generate very high Z-scores: the highest Z-score is 160 for motif 6 based on DRG with the retail investor data (see also table 3). This can be an unrealistically high value as DRG does not retail the actual degree distributions. Overall, the motif counts of money flow networks show surprisingly high Z-scores.

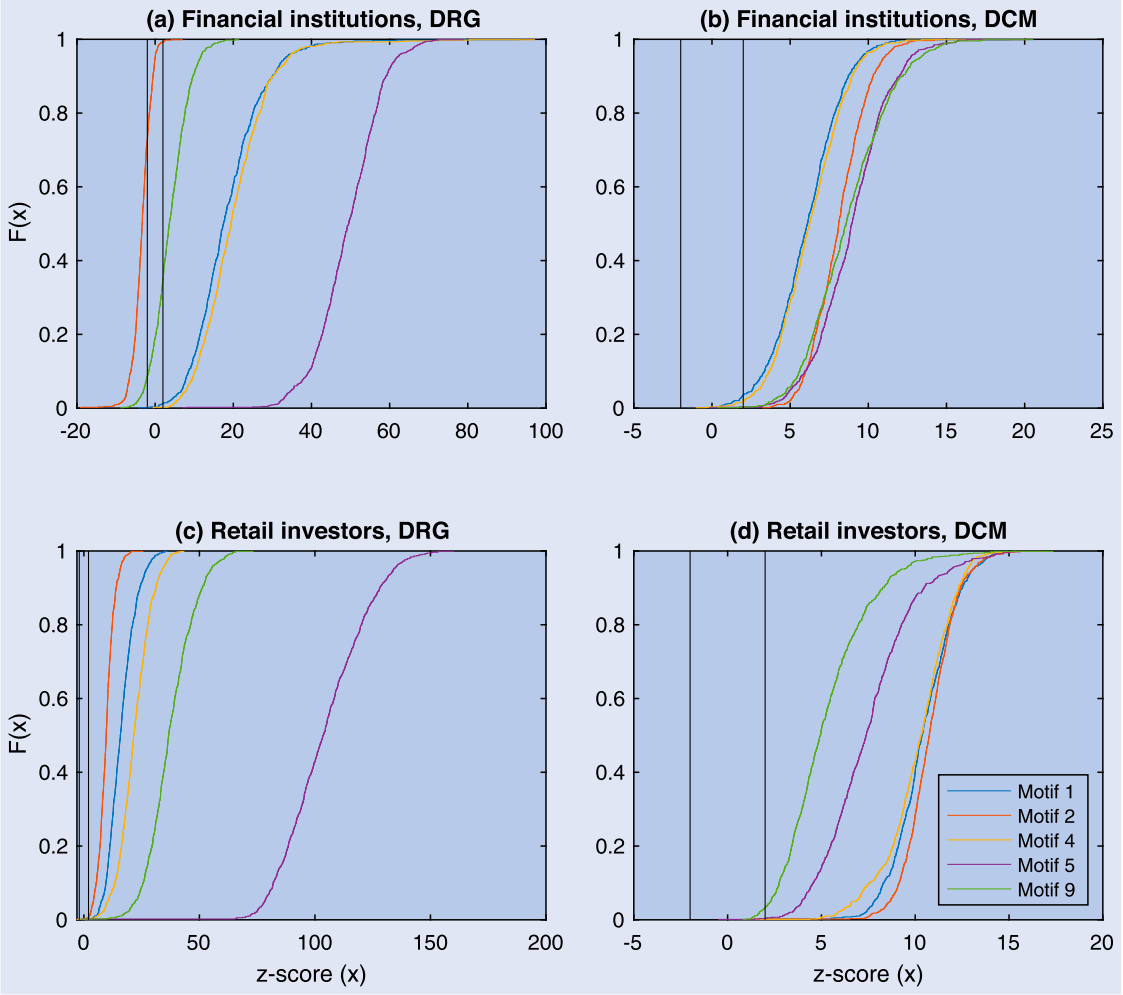

Figure 5. The empirical cumulative distributions of Z-scores on motif counts of money flow network extracted from transactions of financial institutions and retail (household) investors based on Directed Random Graph and Directed Configuration Model as a null model. The vertical lines represent two standard deviations from the models' expectations.

motives analyzed in this paper are clearly over-represented in money-flow networks.

Panel B in Table 3, which reports the differences between two consecutive Z-scores, i.e. ΔZ-scores, shows that in comparison to DCM, DRG yields more time variation in ΔZ-scores. That is, the Z-scores vary more over consecutive days under the DRG than DCM. Under both models, ΔZ-scores are quite symmetrically distributed in terms of minimum and maximum values.

5.2. *Dynamics of motif counts*

To analyze the contemporaneous and lagged associations between the motif counts, we implement a Vector Autoregression (VAR) model. Intuitively, different motifs can have contemporaneous interactions: for example, given that motifs 1 and 4 appear, then also motif 5 exists, and, on the other hand, there must be at least as many motif 2 as motif 9. However, it is not clear how the motif counts are dynamically dependent on each other, which we study by the VAR model

With the Augmented Dickey-Fuller and Phillips-Perro tests, we find that the null hypothesis that a unit root is present in a time series sample is rejected for the counts of all the five motifs. For that reason, the motif counts are not differentiated, but the VAR model is estimated directly for the counts.

The results in table 4 show that for each motif, its count is driven by its previous values, which can be observed from the main diagonals of the tables. This holds with transaction data from both financial institutions (Panel A) and retail investors (Panel B). Therefore we can safely argue that the motif counts follow AR(1)-type dynamics. Moreover, the table shows that with the money flow networks constructed with data of institutions' (retail investors) transactions, motif 2's previous values drive the current values of all the motifs (motifs 4 and 9). Moreover, with data from financial institutions (retail investors), motif 1 is a driving force for motifs 4, 5, and 9 (5 and 9). These observations can partially be explained by the fact that motif 1 is nested in motif 5 and motif 2 in motif 9, which can create a dynamic interdependence between the motifs. On the other hand, we find that with data on financial institutions, motif 9 is driven by motif 5, even if they do not nest one of other.

Second, we analyze the variables' contemporaneous effects on each other with the orthogonal impulse response on the estimated VAR model. The impulse responses of each variable are plotted in figure 6. We observe that shocks to the counts of motifs 1 and 2 have an impact contemporaneously

Table 3. Summary statistics on Z-scores for money flow networks, calculated using Directed Random Graph (DRG) and Directed Configuration Model (DCM).

		Motifs				
		1	2	4	5	9
		Panel A: Z-scores				
Financial institutions						
DRG	Mean	18.69	− 3.38	20.20	49.35	3.95
	StDev	(9.06)	(2.37)	(9.05)	(8.27)	(4.48)
	Min	− 4.15	− 19.13	− 0.50	7.12	− 8.80
	Max	93.11	6.98	97.09	78.78	21.39
DCM	Mean	6.08	8.15	6.29	8.94	8.73
	StDev	(2.18)	(1.71)	(2.15)	(2.35)	(2.61)
	Min	− 0.27	3.25	− 1.01	− 0.14	0.37
	Max	19.06	14.62	17.43	20.02	20.52
Retail investors						
DRG	Mean	16.32	9.67	21.92	104.27	37.61
	StDev	(5.83)	(3.44)	(6.96)	(18.03)	(10.10)
	Min	2.31	− 0.57	− 3.00	− 1.22	− 1.22
	Max	36.02	25.55	43.26	159.99	73.20
DCM	Mean	10.43	10.76	10.14	7.49	5.27
	StDev	(1.65)	(1.41)	(1.82)	(2.39)	(2.23)
	Min	1.36	1.36	1.14	− 0.47	0.83
	Max	15.62	15.74	15.37	16.81	17.33
		Panel B: ΔZ-scores				
Financial institutions						
DRG	Mean	3.63E − 02	− 1.18E − 04	− 6.04E − 03	− 4.99E − 02	− 9.47E − 03
	StDev	(11.41)	(2.98)	(11.80)	(8.02)	(5.73)
	Min	− 61.62	− 16.42	− 64.15	− 43.34	− 18.66
	Max	75.09	14.31	83.71	37.76	20.21
DCM	Mean	5.79E − 03	− 3.70E − 03	3.22E − 04	− 1.03E − 02	− 3.51E − 03
	StDev	(2.59)	(1.71)	(2.51)	(2.35)	(2.68)
	Min	− 10.40	− 7.33	− 7.52	− 8.36	− 8.37
	Max	14.23	5.92	10.98	8.99	9.34
Retail investors						
DRG	Mean	− 1.54E − 02	− 5.90E − 03	1.03E − 02	3.64E − 02	− 5.92E − 03
	StDev	(6.03)	(3.27)	(7.61)	(14.40)	(12.96)
	Min	− 24.36	− 10.32	− 31.20	− 100.22	− 44.95
	Max	20.51	14.22	29.77	98.00	46.15
DCM	Mean	1.11E − 03	− 1.34E − 03	− 1.67E − 02	− 9.74E − 03	2.25E − 03
	StDev	(1.96)	(1.69)	(1.85)	(2.43)	(2.22)
	Min	− 11.01	− 8.92	− 9.55	− 9.41	− 6.94
	Max	9.33	9.86	6.41	7.14	6.08

Note: The results are reported for money flow networks extracted from the transactions of financial institutions and retail investors. The motifs are defined in figure 1. Panel A reports the mean, standard deviation, minimum, and maximum for daily Z-scores. Panel B reports statistics for the differences between two consecutive Z-scores (ΔZ-score).

(i.e. immediately) on counts on all the other motifs. These shocks die relatively slowly, lasting for more than five days. In fact, regarding a shock to motif 2, the only exception is that its impact on motif 1 is lagged (the two plots at the top in figure 6). Then, on the other hand, there are no major impacts of shocks to the counts of motif 9, which can probably be related to the fact that motif 9 is a 'loop', while other motifs are 'branches'. The impacts of shocks to motifs 4 and 5 to other motifs are mixed. Finally, the impact of shocks on the count of a given motif is always positive and immediate to the motif count itself.

6. Conclusions

The contribution of this paper is multi-fold. First, we introduced the formation of money flow networks in stock markets. In the network, the assets are the nodes (or vertex), and the links (or edges) represent the net-money flow between the assets that investors have sold and bought. The links of the money flow network can be extracted from investor-level data on buy and sell transactions of specific assets on a given day. Second, we analyze the empirical properties of money flow networks using data from the Helsinki stock exchange from February 2006 to July 2009. Our daily resolution data set covers investors' transactions over multiple years, and therefore we have time-series of networks over 877 trading days. Such a long temporal network data is unusually long in empirical network science.

The network has some specific properties, such as there are no reciprocal links between any two nodes because here, we consider the net-money flows between assets. In that way, we identified five commonly analyzed sub-graphs that, by construction, could be observed from our empirical networks.

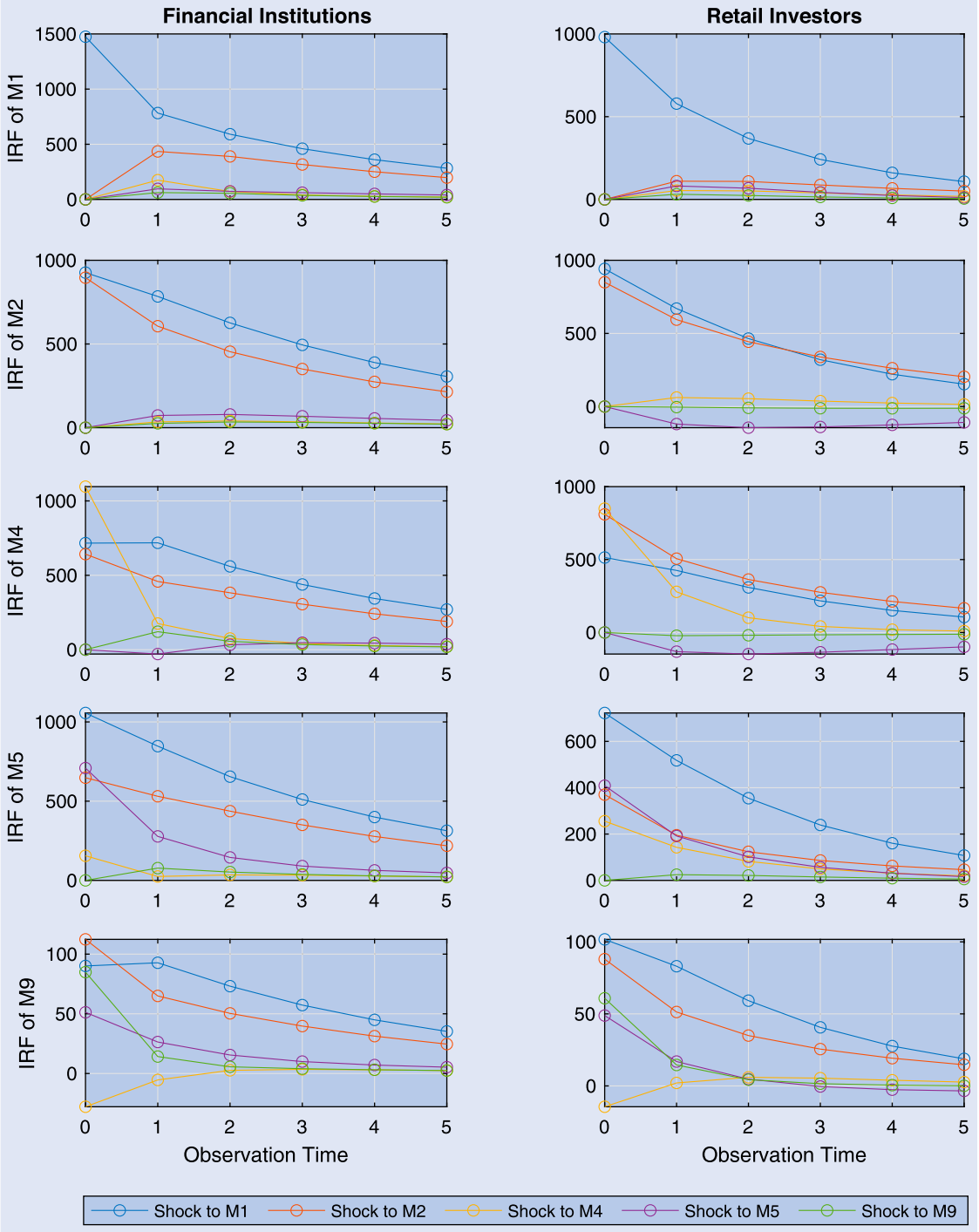

Figure 6. Impulse response of motif counts.

To understand the properties of money flow networks, we analyzed (i) the statistical over-representation of five subgraphs, i.e. motifs, in the money flow networks and (ii) the lagged and contemporaneous impacts of motifs on each other.

First, all of the five motifs are significant against the directed random graph and directed configuration model consistently almost every day, that is the money flow motif counts do not emerge from a random process. This indicates that in terms of investors' aggregated behavior, the sources and destinations of the money flows (between the securities) are chosen in a non-random way. This can, of course, be related to the properties of the securities, such as liquidity or other aspects, a question that must be examined in detail in future research.

Second, we find that the development of the Z-scores motif counts did not indicate any significant reactions in the structure of money flow networks around the 2008 financial crisis. At any moment, no large and lasting changes in any of the Z-score time series could b be observed, indicating a clear change of regime in the crisis. This is in contrast to Squartini *et al.* (2013), who found that in terms of motif counts, topological properties of interbank networks displayed an abrupt change in 2008, yet these two different financial networks

Table 4. Parameter estimates for single-lag VAR model estimated for counts of motifs 1, 2, 4, 5, and 9.

Panel A: Data from financial institutions' transactions					
m_1	**0.21*****	**0.21***	**0.17*****	0.08	0.73
	(9.54E − 05)	(0.02)	(7.26E − 04)	(0.29)	(0.20)
m_2	0.09	**0.56*****	**0.03**	0.08	0.30
	(0.06)	(3.94E − 12)	(0.52)	(0.24)	(0.55)
m_4	**0.22*****	0.28**	**0.22*****	− 0.14	**1.42***
	(3.71E − 05)	(2.02E − 03)	(8.21E − 06)	(0.06)	(0.01)
m_5	**0.13***	**0.24****	− 3.92E − 04	**0.33*****	0.91
	(0.01)	(6.97E − 03)	(0.99)	(2.04E − 05)	(0.10)
m_9	**0.01***	**0.04****	− 4.23E − 03	**0.03****	**0.17***
	(0.03)	(1.02E − 03)	(0.47)	(7.70E − 03)	(0.02)
Panel B: Data from retail investors' transactions					
m_1	**0.43*****	− 0.01	0.03	0.14	0.52
	(1.16E − 10)	(0.86)	(0.56)	(0.14)	(0.33)
m_2	0.16	**0.68*****	**0.16***	**− 0.29***	− 0.07
	(0.06)	(4.61E − 14)	(0.02)	(0.02)	(0.92)
m_4	0.11	**0.37*****	**0.41*****	**− 0.28***	− 0.36
	(0.21)	(5.44E − 05)	(2.86E − 09)	(0.02)	(0.60)
m_5	**0.19****	− 0.04	0.05	**0.42*****	0.41
	(3.01E − 03)	(0.53)	(0.33)	(3.14E − 06)	(0.42)
m_9	**0.02***	**0.03***	2.65E − 03	0.01	**0.24****
	(0.03)	(0.01)	(0.75)	(0.39)	(4.34E − 03)

Note: The model is of the form $m_{1,t} = c_1 + a_{1,1}m_{1,t-1} + a_{1,2}m_{2,t-1} + \ldots$, where $m_{i,t}$ is the count of ith motif at time t. Panel A shows the coefficient matrix with p-values in parentheses, where the estimate on ith row and jth column corresponds to the value of $a_{i,j}$. Statistically significant parameter estimates are in bold.

represent completely different aspects of financial markets research.

Third, our results show clear evidence that the motifs follow autoregressive processes, meaning that for each motif, its count is driven by its previous values. Moreover, we found partial evidence that motifs drive other motifs in which they are nested (for example, motif 1 is nested within motif 5). When analyzing contemporaneous effects, we found that counts of motifs 1 (diversification) and 2 (chain) have an immediate impact on the counts of all the other motifs, lasting for more than a week, while motif 9 (loop) has no impact on others. This indicates that the motif count dynamics are complex, as different types of motifs can interact, but the magnitude of interaction depends on the type of motif. Overall, this analysis shows that the dynamics of different motifs cannot be considered to be independent. For that reason, there is a need for further research to develop reliable temporal models that allow substructures to have causal and lagged relations. In our future research, we will focus on this question.

Disclosure statement

This paper expresses the private opinion of the authors and does not reflect the opinion, policy or research of Nordea.

ORCID

Juho Kanniainen http://orcid.org/0000-0001-7737-659X

References

Acemoglu, D., Ozdaglar, A. and Tahbaz-Salehi, A., Systemic risk and stability in financial networks. *Am. Econ. Rev.*, 2015, **105**(2), 564–608.

Alon, U., Network motifs: theory and experimental approaches. *Nature Rev. Genetics*, 2007, **8**(6), 450–461.

Baltakiene, M., Kanniainen, J. and Baltakys, K., Identification of information networks in stock markets. *J. Econ. Dyn. Control*, 2021, **131**, 104217.

Baltakys, K., Inference of monopartite networks from bipartite systems with different link types. *Sci. Rep.*, 2023, **13**(1), 1072.

Braha, D., Patterns of ties in problem-solving networks and their dynamic properties. *Sci. Rep.*, 2020, **10**(1), 18137.

Forsgren, A., Gill, P.E. and Wright, M.H., Interior methods for nonlinear optimization. *SIAM Rev.*, 2002, **44**(4), 525–597.

Gabaix, X. and Koijen, R.S.J., In search of the origins of financial fluctuations: The inelastic markets hypothesis. Technical Report. National Bureau of Economic Research, 2021.

Gabaix, X. and Maggiori., M., International liquidity and exchange rate dynamics. *Q. J. Econ.*, 2015, **130**(3), 1369–1420.

Gualdi, S., Cimini, G., Primicerio, K., Di Clemente, R. and Challet, D., Statistically validated network of portfolio overlaps and systemic risk. *Sci. Rep.*, 2016, **6**(1), 39467.

Haldane, A.G. and May, R.M., Systemic risk in banking ecosystems. *Nature*, 2011, **469**(7330), 351–355.

Holland, P.W. and Leinhardt, S., The statistical analysis of local structure in social networks, 1974.

Holme, P. and Saramäki, J., Temporal networks. *Phys. Rep.*, 2012, **519**(3), 97–125.

Ingram, P.J., Stumpf, M.P.H. and Stark, J., Network motifs: structure does not determine function. *BMC. Genomics.*, 2006, **7**(1), 1–12.

Kashani, Z.R.M., Ahrabian, H., Elahi, E., Nowzari-Dalini, A., Ansari, E.S., Asadi, S., Mohammadi, S., Schreiber, F. and Masoudi-Nejad, A., Kavosh: A new algorithm for finding network motifs. *BMC. Bioinformatics.*, 2009, **10**(1), 1–12.

McKay, B.D. and Piperno, A., Practical graph isomorphism, II. *J. Symb. Comput.*, 2014, **60**, 94–112.
Milo, R., Itzkovitz, S., Kashtan, N., Levitt, R., Shen-Orr, S., Ayzenshtat, I., Sheffer, M. and Alon, U., Superfamilies of evolved and designed networks. *Science*, 2004, **303**(5663), 1538–1542.
Milo, R., Shen-Orr, S., Itzkovitz, S., Kashtan, N., Chklovskii, D. and Alon, U., Network motifs: Simple building blocks of complex networks. *Science*, 2002, **298**(5594), 824–827.
Newman, M., *Networks*, 2018 (Oxford University Press: Oxford).
Ozsoylev, H.N., Walden, J., Yavuz, M.D. and Bildik, R., Investor networks in the stock market. *Rev. Financ. Stud.*, 2014, **27**(5), 1323–1366.
Patra, S. and Mohapatra, A., Review of tools and algorithms for network motif discovery in biological networks. *IET. Syst. Biol.*, 2020, **14**(4), 171–189.
Picard, F., Daudin, J.-J., Koskas, M., Schbath, S. and Robin, S., Assessing the exceptionality of network motifs. *J. Comput. Biol.*, 2008, **15**(1), 1–20.
Squartini, T. and Garlaschelli, D., Analytical maximum-likelihood method to detect patterns in real networks. *New. J. Phys.*, 2011, **13**(8), 083001.
Squartini, T., Van Lelyveld, I. and Garlaschelli, D., Early-warning signals of topological collapse in interbank networks. *Sci. Rep.*, 2013, **3**(1), 3357.
Vayanos, D. and Woolley, P., An institutional theory of momentum and reversal. *Rev. Financ. Stud.*, 2013, **26**(5), 1087–1145.
Vermeulen, R., International diversification during the financial crisis: A blessing for equity investors. *J. Int. Money. Finance.*, 2013, **35**, 104–123.
Watts, D.J. and Strogatz, S.H., Collective dynamics of 'small-world' networks. *Nature*, 1998, **393**(6684), 440–442.

Appendix

A.1. Additional descriptive statistics

For each day, the quantile statistics are calculated, and these numbers are then aggregated; the column indicates the aggregated quantile of the row quantile, i.e. the max column – mean row means the maximum daily mean over the whole dataset.

Table A1 shows the number of institutional investors per asset. In Panel A, which includes all the assets, we see that most assets have relatively few investors; on the quietest days, more than 75% of the assets have no institutional traders, and on most days the median is 1 trader per asset. In Panel B we restrict each day's analysis to the assets that had at least one trade, i.e. this is a table for assets that actually had some liquidity among institutions. Even so, the median asset is only traded by 5 institutions on the busiest day.

Table A2 Panel A shows the number of assets traded by all institutional investors in our data. Here the contrast is even more stark, on any given day, 75% of institutions do not trade at all. This is mostly because most institutions make transactions rather rarely. In Table A2 Panel B we restrict to those institutions on each day that actually were active. The median institution among those still makes rather few trades, and on the busiest day, the median is still only 4 assets per investor.

Table A3 Panel A takes into account all the assets. While we see a stark contrast between liquid and illiquid assets' trading patterns among institutional investors, the effect is far less pronounced among retail investors. Notice that the maximum is, by necessity, the same for both of these tables. In Table A3 Panel B We have similarly restricted to liquid assets, i.e. assets that were actually traded, but this time by retail investors. The most liquid assets were traded by 672.7 retail investors, but with significant variability.

Table A4 shows the numbers of assets traded by active retail investors. As can be guessed, most investors are inactive, so it would make very little sense to calculate the same over all retail investors (it would consist of zeros with simply the maximum remaining identical to this table).

A.2. Null models

Here we summarize the null models used to calculate the Z-scores of the motifs. For more information, see Squartini and Garlaschelli (2011) and Squartini *et al.* (2013).

A.2.1. Directed random graph model. Under the directed graph model, the expected number and standard deviation for the number of motifs are

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

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

$$\begin{aligned} \langle N_m \rangle &= \frac{1}{\alpha_m} \sum_{i \neq j \neq k} p^k (1-p)^{6-k} \\ &= \frac{1}{\alpha_m} {}^N P_3 p^k (1-p)^{6-k} \\ &= \frac{1}{\alpha_m} \frac{N!}{(N-3)!} p^k (1-p)^{6-k} \\ &= \frac{1}{\alpha_m} N(N-1)(N-2) p^k (1-p)^{6-k}, \end{aligned} \tag{A1}$$

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 in Squartini and Garlaschelli (2011).

Variance for any topological quantity X across a 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):

$$\begin{aligned} (\sigma^*(X))^2 = \sum_{i,j} \Bigg[&\left(\sigma^*[g_{ij}] \frac{\partial X}{\partial g_{ij}} \right)^2_{\boldsymbol{G}=\langle \boldsymbol{G} \rangle^*} \\ &+ \sigma^*[g_{ij}, g_{ij}] \left(\frac{\partial X}{\partial g_{ij}} \frac{\partial X}{\partial g_{ji}} \right)^2_{\boldsymbol{G}=\langle \boldsymbol{G} \rangle^*} \Bigg] + \ldots \end{aligned} \tag{A2}$$

In (A2) 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^* defined in Squartini *et al.* (2013) and $X = N_m$ which are defined in Table A5.

The motif counts can be generalized the same way as in (A1) to be

$$N_m = \frac{1}{\alpha_m} \sum_{i \neq j \neq k} p^k (1-p)^{6-k}, \tag{A3}$$

which leads to the partial derivative

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

Table A1. Institutional investors per asset.

	Mean	Std	Min	25%	50%	75%	Max
Panel A: All Assets							
Min	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Q1	0.00	0.07	0.00	0.00	0.00	0.00	1.00
Mean	2.92	0.70	0.50	2.45	2.89	3.39	5.94
Median	0.93	0.46	0.00	1.00	1.00	1.00	2.00
Q3	3.57	1.23	0.00	3.00	3.00	4.00	8.00
Max	23.74	6.48	5.00	20.00	23.00	26.00	65.00
Panel B: Liquid Assets							
Min	1.00	0.00	1.00	1.00	1.00	1.00	1.00
Q1	1.14	0.33	1.00	1.00	1.00	1.00	2.00
Mean	5.00	0.78	2.03	4.49	5.00	5.53	8.14
Median	2.76	0.64	1.00	2.00	3.00	3.00	5.00
Q3	7.54	1.48	2.00	6.75	7.75	8.00	13.00
Max	23.74	6.48	5.00	20.00	23.00	26.00	65.00

Table A2. Assets per institutional investor.

	Mean	Std	Min	25%	50%	75%	Max
Panel A: All Investors							
Min	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Q1	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Mean	0.64	0.15	0.11	0.54	0.63	0.74	1.30
Median	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Q3	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Max	35.51	6.58	12.00	31.00	34.00	38.00	70.00
Panel B: Active Investors							
Min	1.00	0.00	1.00	1.00	1.00	1.00	1.00
Q1	1.02	0.15	1.00	1.00	1.00	1.00	2.00
Mean	6.23	0.74	3.57	5.69	6.18	6.71	9.20
Median	2.22	0.52	1.00	2.00	2.00	2.50	4.00
Q3	6.53	1.99	3.00	5.00	6.00	7.50	17.00
Max	35.51	6.58	12.00	31.00	34.00	38.00	70.00

Table A3. Retail investors per asset.

	Mean	Std	Min	25%	50%	75%	Max
Panel A: All Assets							
Min	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Q1	4.65	2.16	0.00	3.00	4.00	6.00	20.00
Mean	45.67	16.72	0.02	35.42	43.02	51.53	151.31
Median	13.30	5.42	0.00	9.00	12.00	17.00	57.00
Q3	39.38	14.15	0.00	29.00	38.00	47.00	123.00
Max	672.71	476.45	1.00	391.00	522.00	781.00	4370.00
Panel B: Liquid Assets							
Min	1.03	0.17	1.0	1.0	1.0	1.0	2.0
Q1	6.68	2.34	1.0	5.0	6.0	8.0	23.0
Mean	50.64	17.90	1.0	39.51	47.10	56.73	157.22
Median	15.70	5.62	1.0	12.0	15.0	19.0	61.0
Q3	46.73	14.29	1.0	37.25	45.25	53.5	141.25
Max	672.70	476.45	1.0	391.0	522.0	781.0	4370.0

where α_m is a constant that depends on the symmetries of the particular motif. The sum in (A3) 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 motif 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 motif counts are defined by

$$(\sigma^*(N_m))^2 \approx \sum_{i,j} \left[\left(\sigma^*[p_{ij}] \frac{\partial N_m}{\partial p_{ij}} \right)^2_{\boldsymbol{A}=\langle \boldsymbol{A} \rangle^*} + \sigma^*[p_{ij}, p_{ij}] \left(\frac{\partial N_m}{\partial p_{ij}} \frac{\partial N_m}{\partial p_{ji}} \right)^2_{\boldsymbol{A}=\langle \boldsymbol{A} \rangle^*} \right] \tag{A4}$$

Table A4. Assets per daily active Retail investor.

	Mean	Std	Min	25%	50%	75%	Max
Min	1.00	0.00	1.00	1.00	1.00	1.00	1.00
Q1	1.00	0.00	1.00	1.00	1.00	1.00	1.00
Mean	1.50	0.08	1.00	1.45	1.49	1.54	1.90
Median	1.00	0.00	1.00	1.00	1.00	1.00	1.00
Q3	1.76	0.43	1.00	2.00	2.00	2.00	2.00
Max	18.12	5.36	1.00	15.00	17.00	20.00	96.00

Table A5. Equations for calculating the number of triadic motifs m in a network from an adjacency matrix A. The motifs are visualized in figure 1.

Triadic motif (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	$\sum_{i\neq j\neq k}a_{ij}(1-a_{ji})a_{jk}(1-a_{kj})(1-a_{ik})(1-a_{ki})$
4	$\sum_{i\neq j\neq k}(1-a_{ij})(1-a_{ji})a_{jk}(1-a_{kj})a_{ik}(1-a_{ki})$
5	$\sum_{i\neq j\neq k}(1-a_{ij})a_{ji}a_{jk}(1-a_{kj})a_{ik}(1-a_{ki})$
9	$\sum_{i\neq j\neq k}(1-a_{ij})a_{ji}(1-a_{jk})a_{kj}a_{ik}(1-a_{ki})$

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 (A1) and noting that self-edges are excluded from the sum

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

and finally

$$\sigma^*(N_m) = \frac{T_2}{\alpha_m}\left(kp^{k-1}(1-p)^{6-k} - (6-k)p^k(1-p)^{5-k}\right), \quad \text{(A5)}$$

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 A5 when calculating the Z-scores. Using notation

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

where equations for M_m are given in table A5 without dividing by α_m, then the Z-scores can be calculated as

$$\begin{aligned}\mathrm{Z-score}(G_m) &= \frac{N_m - \langle N_m\rangle}{\sigma[N_m]}\\ &= \frac{M_m/\alpha_m - T_1/\alpha_m p^k(1-p)^{6-k}}{T_2/\alpha_m\left(kp^{k-1}(1-p)^{6-k}\right.\\ &\quad\left.-(6-k)p^k(1-p)^{5-k}\right)}\\ &= \frac{M_m - T_1p^k(1-p)^{6-k}}{T_2\left(kp^{k-1}(1-p)^{6-k} - (6-k)p^k(1-p)^{5-k}\right)}.\end{aligned} \quad \text{(A6)}$$

Thus, there is a method for calculating the Z-scores analytically without needing to solve the automorphism group order. The equations in table A5 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 motifs (e.g. Kashani *et al.* 2009), and use (A1) and (A5) to calculate the Z-scores without needing to generate random graphs, but the networks are small enough that equations given in table A5 are fast enough to be used for counting the motifs in this paper.

A.2.2. Directed configuration model. Deriving equations for the expected number of motifs 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

$$\langle a_{ij}\rangle = p_{ij} = \frac{x_i y_j}{1 + x_i y_j}, \quad \text{(A7)}$$

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

$$\begin{aligned}\sum_{i\neq j}\frac{x_i^* y_j^*}{1 + x_i^* y_j^*} &= k_i^{out} \quad \forall\, i\\ \sum_{i\neq j}\frac{x_j^* y_i^*}{1 + x_j^* y_i^*} &= k_i^{in} \quad \forall\, i,\end{aligned} \quad \text{(A8)}$$

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

$$\begin{aligned}\min_{\boldsymbol{x}^*,\boldsymbol{y}^*}\quad & \frac{1}{2}\|F(\boldsymbol{x}^*,\boldsymbol{y}^*)\|^2\\ \text{s.t.}\quad & \boldsymbol{x}^*,\boldsymbol{y}^* > 0,\end{aligned} \quad \text{(A9)}$$

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 (A9) can be solved by using the interior-point method for nonlinear optimization provided in 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 count of triadic motifs is

$$\langle N_m \rangle = N_m(a_{ij}^*, a_{ji}^*, a_{jk}^*, a_{kj}^*, a_{ik}^*, a_{ki}^*), \tag{A10}$$

where $N_m(\boldsymbol{a}^*)$ are given in table A5 and the values for a_{ij}^* can be calculated using (A7).

The variance of the motif counts can be calculated using (A2). 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 a maximum of $N-2$ 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 motif 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

$$(\sigma^*(N_m))^2 = \sum_{i,j}\left[\frac{\sqrt{x_i^* y_j^*}}{1+x_i^* y_j^*}\frac{\partial}{\partial p_{ij}} N_m(\boldsymbol{a}^*)\right]^2, \tag{A11}$$

which has a time complexity $O(N^3)$.

Equation (A11) can also be structured so that first the motifs g_i are found from the graph and then for each motif the change in the weight of the motif is calculated if weight of an edge is changed. In mathematical notation

$$(\sigma^*(N_m))^2 = \sum_i \sigma^2(g_i), \tag{A12}$$

where i goes from 0 to the count of motifs,

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

where the i and j values go from 0 to the number of nodes in the motif, and $G(\boldsymbol{a})$ is the product of a_{ij} and $(1-a_{st})$ which define the motif. The function $d(a_{ij})$ is derived from the derivative of $G(\boldsymbol{a})$ and it is

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

This approach makes it possible to use other algorithms to find the motifs from the networks, so there is no need to loop through all possible node permutations if the network is not complete. Thus, the last approach has time complexity $O(n(n-1)f_m^*)$ in which n is the number of edges in the motif, and f_m^* is the function which limits the motif enumeration algorithm of the maximum-entropy approximation of the network.

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

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

where $\bar{N}_m$ is the expected count of motifs. On one hand, the time complexity for calculating the unbiased sample variance is $O(SN^3)$, which is worse than for the linear approximation. On the other hand, the unbiased sample variance is easy to implement. In this paper, the standard deviation of the count of motifs is calculated by generating 100 random networks in the case of a directed configuration model. The motifs could be counted by some other algorithm also, which would have time complexity $O(f_m)$, and for S samples the complexity would be $O(Sf_m)$. Using the approach defined by (A11) could be faster than sampling if the network is dense as in the sampling approach more of the edges would need to be enumerated every time compared to a sparse network. The maximum-entropy approach could be faster even though the approximated network is complete as the enumeration would only need to be done once.

A.3. Time-series of motif counts

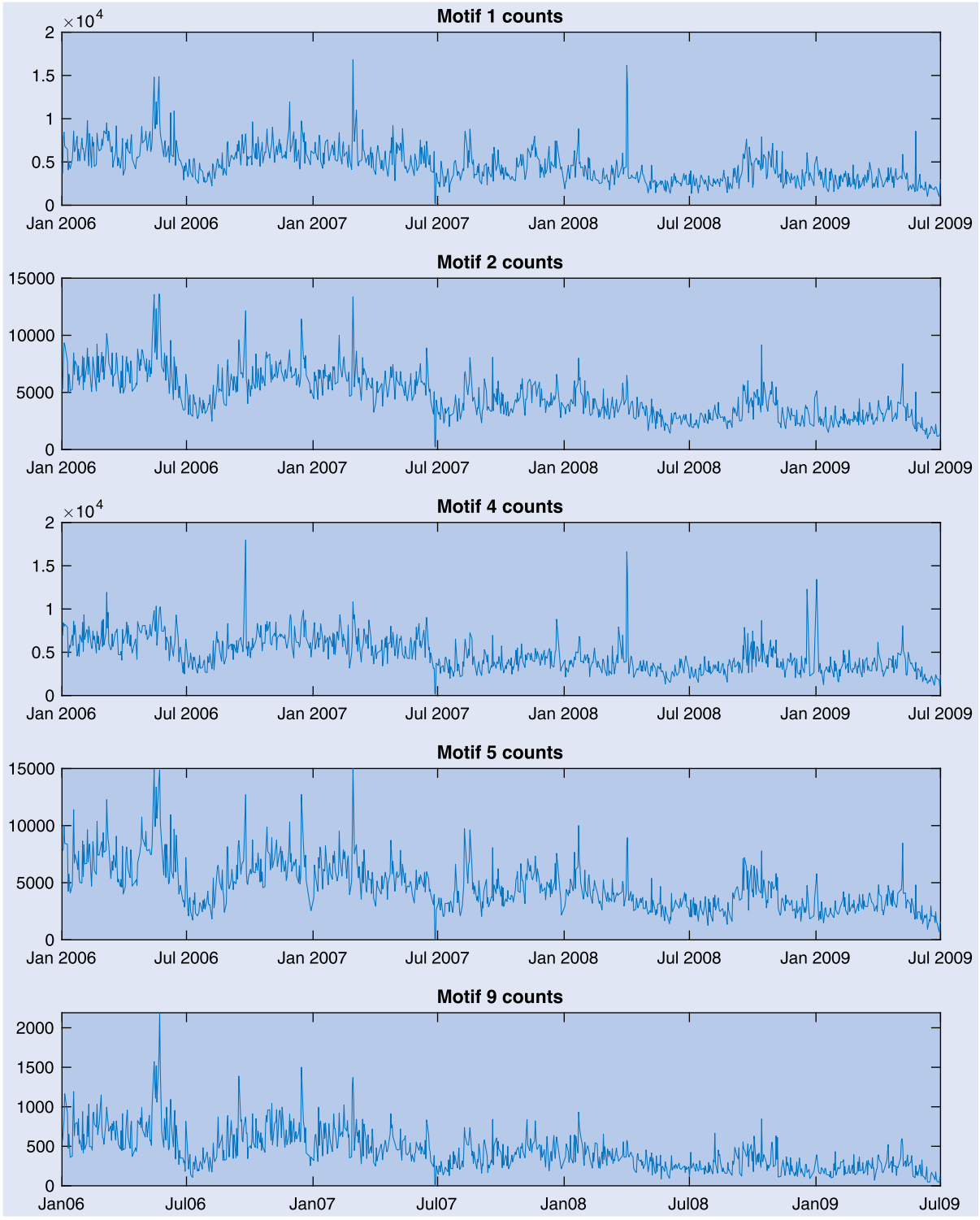

Figure A1. Motif counts of money flow network extracted from transactions of financial institutions.

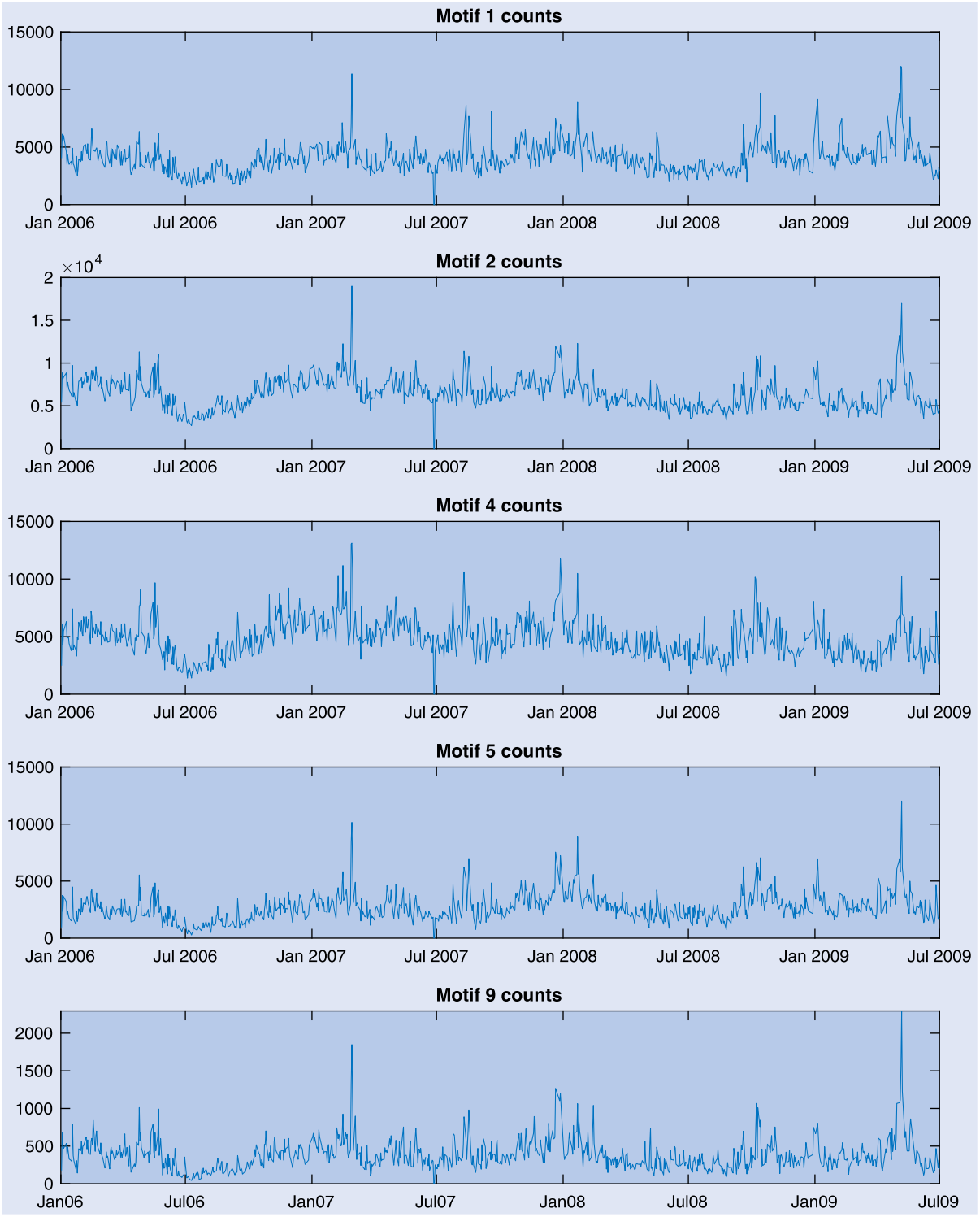

Figure A2. Motif counts of money flow network extracted from transactions of retail (household) investors.

A.4. Time-series of Z-scores

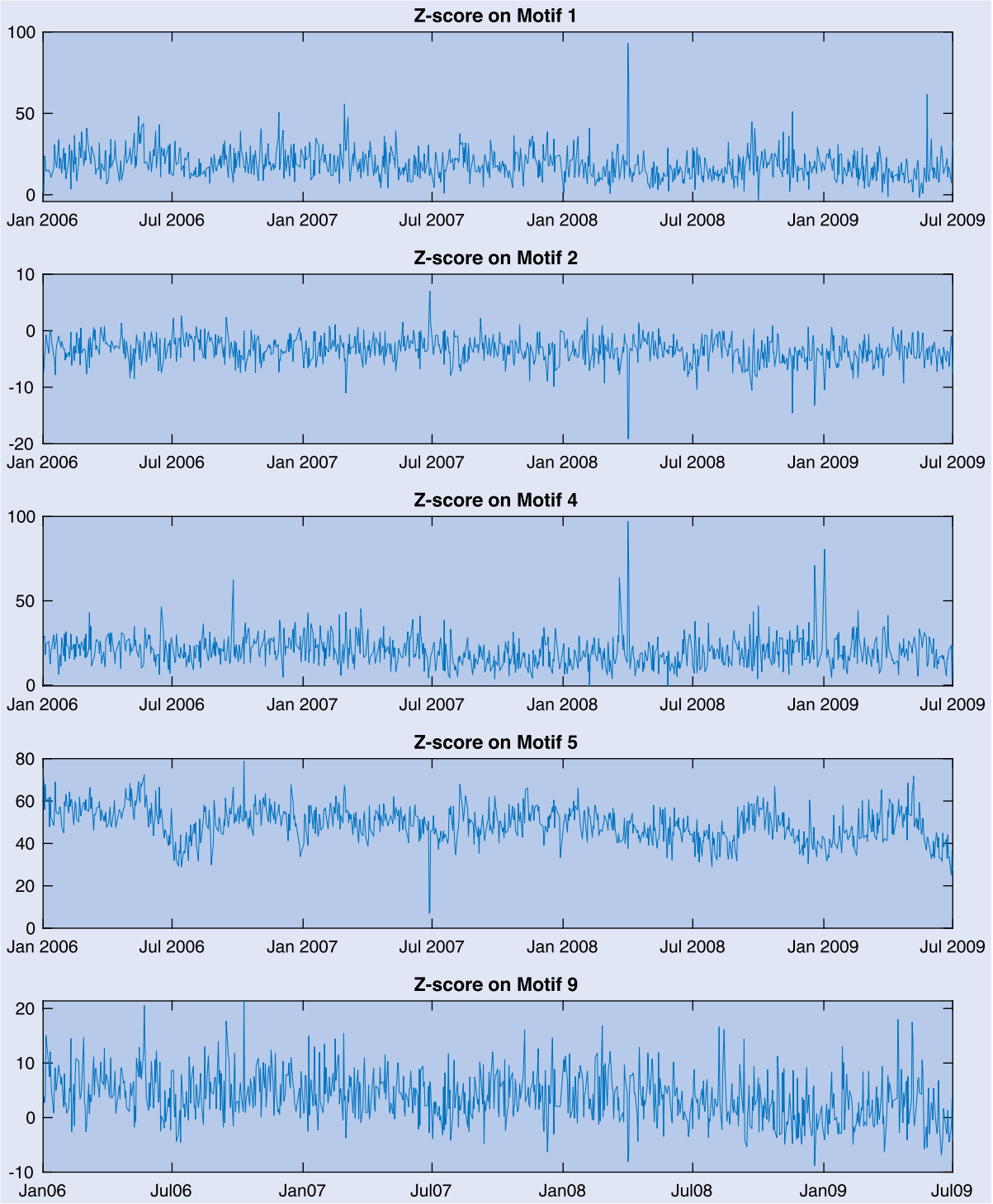

Figure A3. Z-scores on motif counts of money flow network extracted from transactions of financial institutions based on Directed Random Graph as a null model.

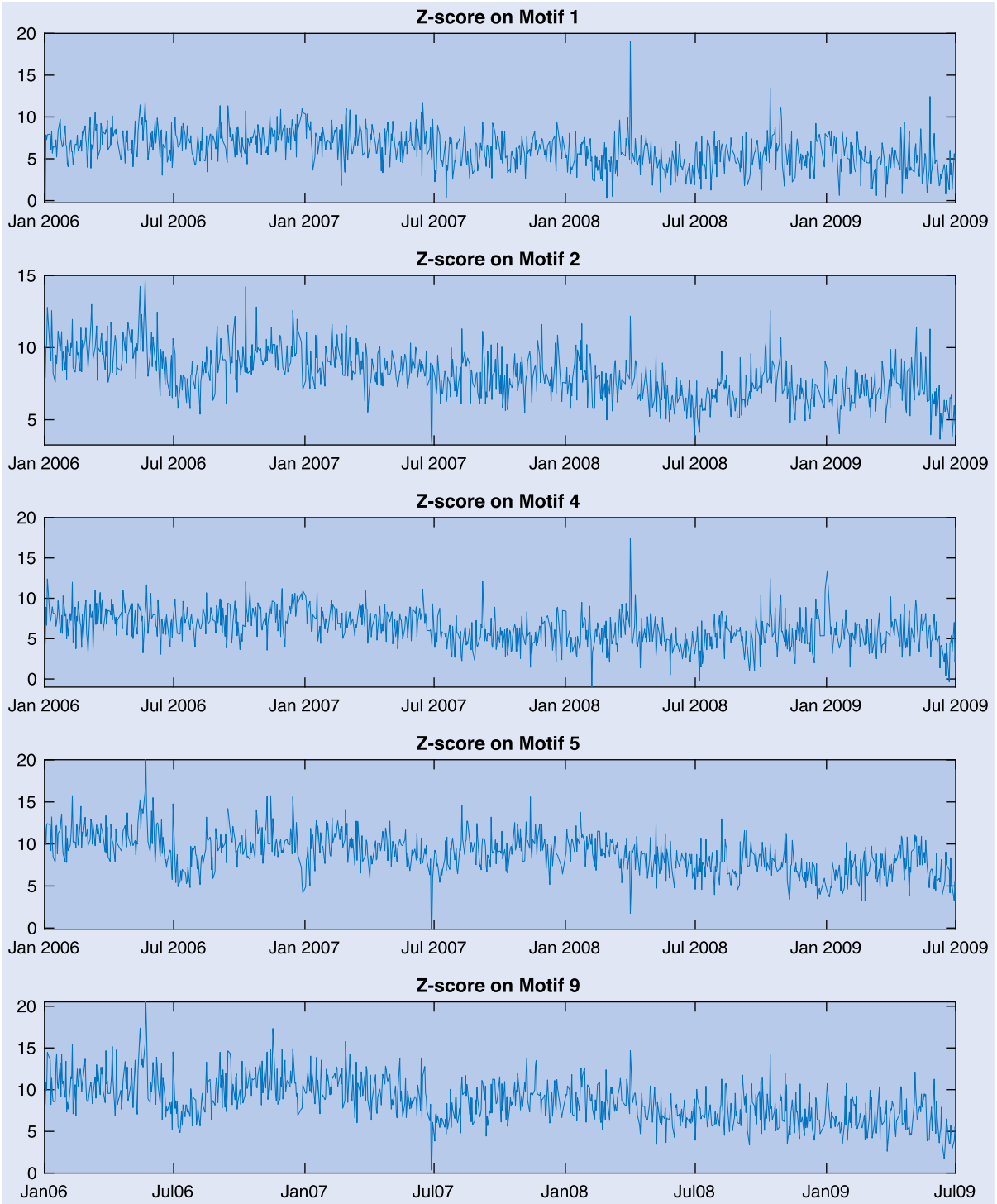

Figure A4. Z-scores on motif counts of money flow network extracted from transactions of retail (household) investors based on Directed Random Graph as a null model.

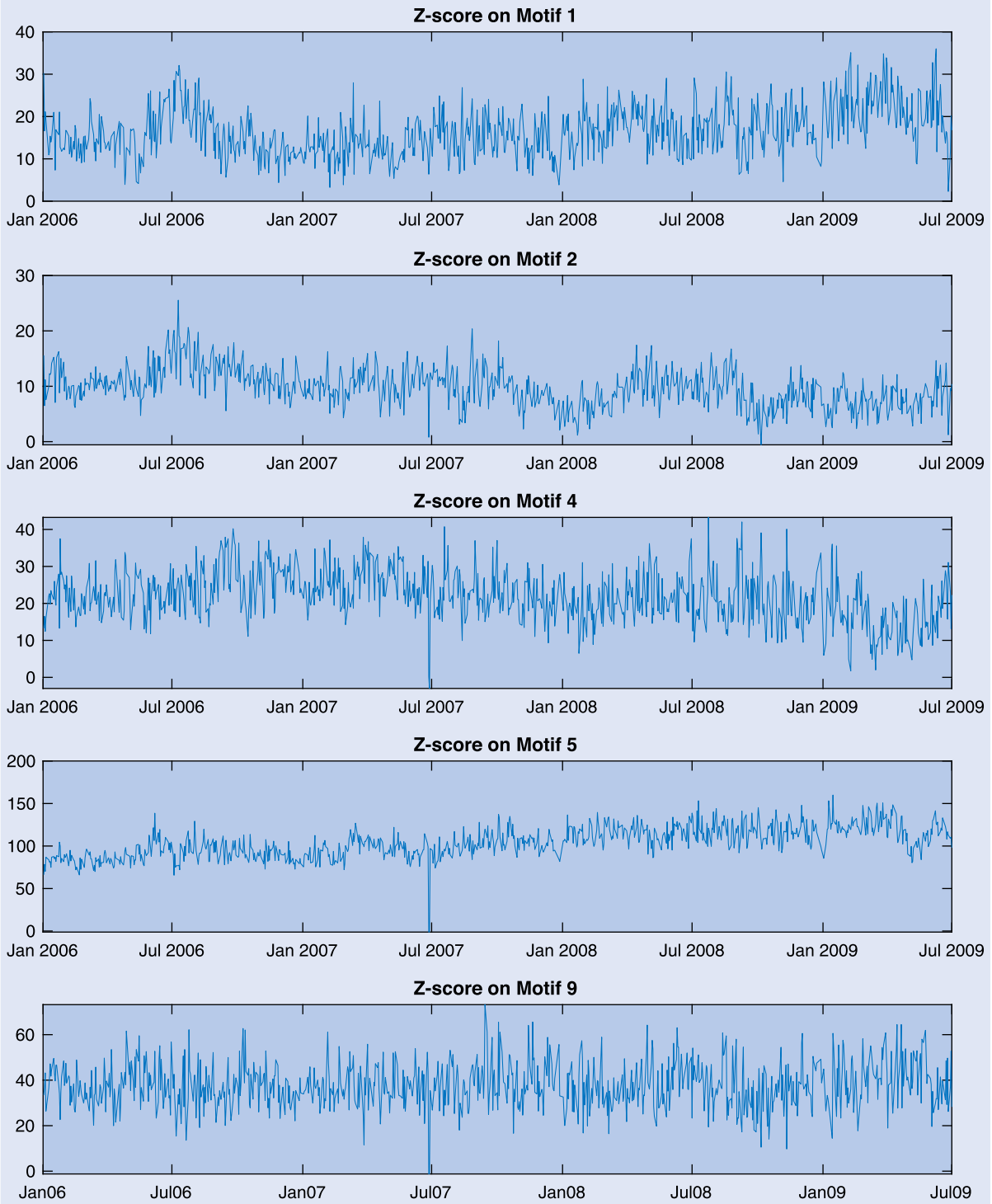

Figure A5. Z-scores on motif counts of money flow network extracted from transactions of financial institutions based on Directed Configuration Model as a null model.

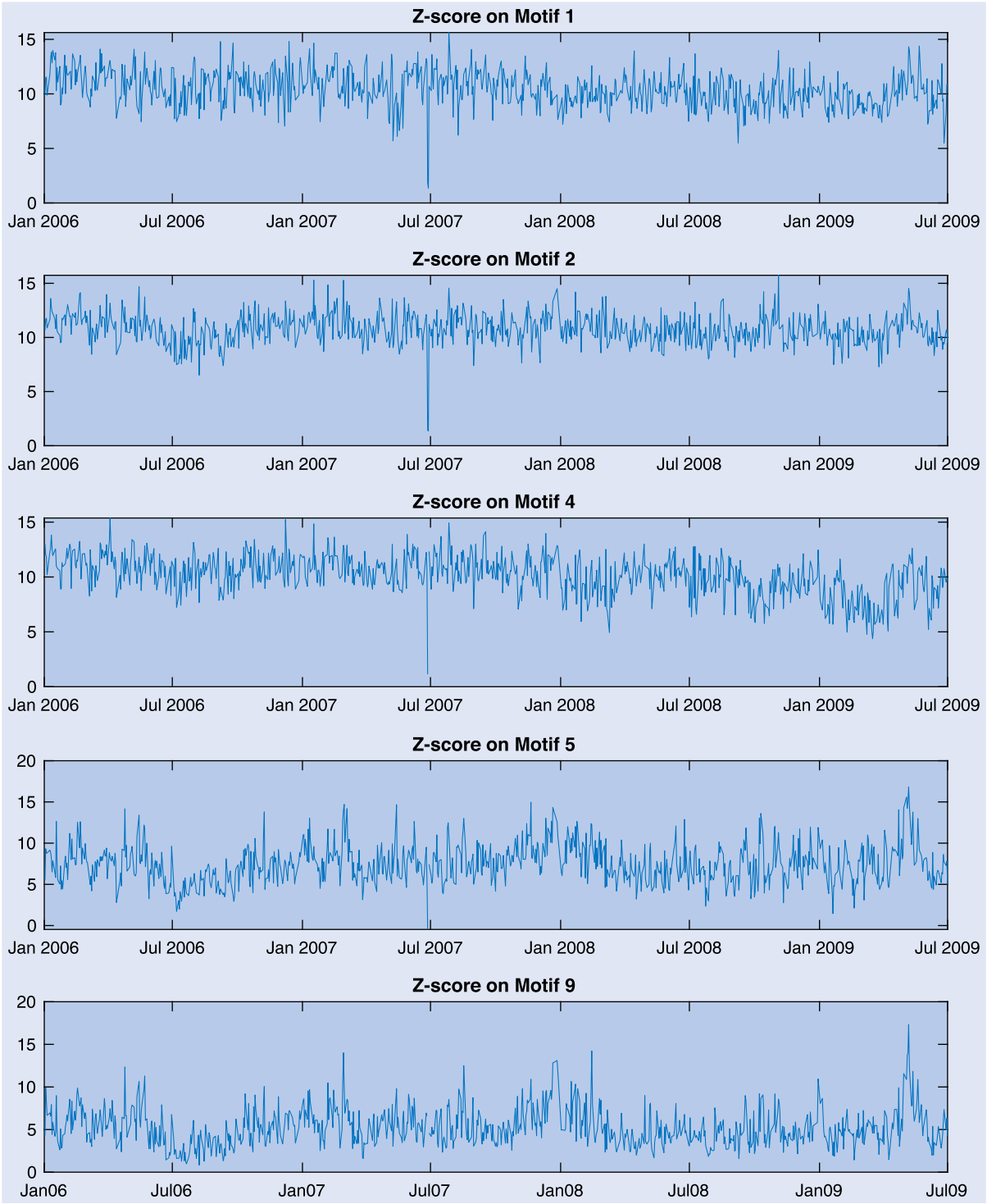

Figure A6. Z-scores on motif counts of money flow network extracted from transactions of retail (household) investors based on Directed Configuration Model as a null model.