Information transmission among multiple investors: a micro-perspective revealed by motifs

Jie Li · Yongjie Zhang · Lidan Wang

Received: 8 July 2021 / Accepted: 13 February 2022 / Published online: 24 February 2022
© The Author(s), under exclusive licence to Springer Nature B.V. 2022

Abstract The concept of motifs provides a new perspective for studying local patterns, which is useful for understanding the nature of a network structure. In this study, the types and evolution of the motifs of the shareholder co-ownership network, constructed based on common shareholding data from 2007 to 2017, are explored from a micro-perspective. Although with a low proportion, the closed motifs were found to be important motifs with statistical significance in the network. Furthermore, the motifs containing financial investment company shareholders tend to disappear on both short (quarterly) and long (annual) time scales. In contrast, the motifs containing general corporate shareholders tend to remain unchanged. Finally, the abnormal abrupt changes in the proportions of important motifs in the real network relative to the random network before and after the financial crisis are calculated. The number of Motif 4 containing state-owned companies, general companies, and individual investors decreases abnormally during the financial crisis. This research is of great significance for understanding information interaction among multiple investors.

Keywords Shareholder co-holding network · Motif identification · Motif evolution · Information transmission · Random network model

1 Introduction

Previous studies have shown that investors with similar attributes, such as geographical location [1], education experience [2] and work experience [3], tend to have information interactions. Restricted by the availability of data and the difficulty in comprehensively considering all social relations, scholars have tended to extract the social relations among shareholders from shareholders’ shareholding behaviours in recent years [4, 5]. The logic of extracting social relations from shareholding behaviour is that the private information endowments of shareholders with a large proportion of investment in the same stock may be correlated with each other [6]. Milo et al. [7] proposed the concept of network motifs and defined
them as small-scale isomorphic subgraphs with statistical significance that repeatedly occur in real networks. The application of motifs in a shareholder co-ownership network (SCN) provides an effective means to analyse the information interaction between shareholders at the micro-level. As the basic building blocks of the network, the motifs in the SCN refer to the miniature “circle of friends” of specific shareholders who tend to cluster together. By observing the behaviour of other people in motifs, shareholders may change their investment decisions after rational reasoning and psychological preferences [8].

There are profound economic implications behind the formation and evolution of specific motifs. Consider the 3-node motifs formed by two fund companies and a state-owned enterprise as an example. During the stock prices rise, two fund companies and a state-owned enterprise form a stable motif structure. Subsequently, as prices fall, the state-owned enterprise disappears, and the motif structure is broken. There may be violations of financial institutions “lifting the sedan chair” for the state-owned enterprise in this process. In another practical case, four individual shareholders appeared in the list of the top ten shareholders of Jinyun Laser (stock code: 300,220) in the 2019 annual report. Subsequently, these four shareholders were exposed to close family relationships and were suspected of manipulating stock prices. Based on the analysis of network motifs, the number of 3-node motifs formed by individual shareholders increased significantly during the price rise process, suggesting that there may be violations of family joint action. At the theoretical level, Battiston et al. [9] pointed out in their review article published in Science that “major financial risk events are often caused by unpredictable random events that trigger minority shareholders to sell stocks. This behaviour spread throughout the network and ultimately led to the collapse of the entire stock market.”

Therefore, it is necessary to explore the motif structures of the SCN and their evolutionary laws, regardless of the practical level or the theoretical level.

Then, what types of important motif structures exist in SCN? How do they evolve? Is it possible to provide early warning of financial crises based on the evolutionary laws of abnormal motif structures? These are the questions to be studied in this article.

First, this research explores the important motifs in the SCN. The Rand Enumerating Subgraph (Rand-ESU) algorithm is adopted to efficiently identify the motif structures in the SCN and its corresponding random network. This research must introduce an analytical and unbiased maximum-entropy technique, which uses the knowledge of both strengths and degrees to reconstruct the unbiased ensemble of random networks. Considering the link weights, the closed motifs are identified as important with statistical significance in the network. Further considering the node attributes, the motifs with all three nodes being financial investment companies have the highest proportion.

Second, the evolution of different types of motifs is investigated on two-time scales: short cycle (quarterly) and long cycle (annual). This research distinguishes three evolutionary states: motif invariant, motif vanishing, and motif transforming. All motif structures were found to have a higher tendency to vanish. The longer the period, the greater the possibility of vanishment. Furthermore, considering the node attributes, motifs containing financial investment companies nodes have a higher tendency to disappear, while motifs containing general companies tend to keep the structure unchanged.

Finally, important motifs are being explored to determine whether there are abnormal quantitative mutations before and after the financial crisis.

---

1 Lifting the sedan chair, a proprietary term in the stock market, refers to the fact that the buyer anticipates that a certain stock will rise, and therefore buys a large amount of stock, which leads to an increase in the stock price. But the result is that other people (generally referred to as the banker) sell at a high position and make a profit, and eventually the buyer is locked up.

2 The 3-node motif structure is the most common subgraph form in the network [7]. If the number of nodes is less than 3, there are too few subgraph types, which is not conducive to mining local information of the network. If there are too many nodes, the number of possible motif structures increases exponentially, which increases the complexity of motif recognition. At the same time, the motif structure of more than 4 nodes can be generated by the superposition and combination of different 3-node motif structures. In a word, the 3-node motif structure can satisfy the mining of local features of the network. The “motif structure” in this study refers to the “3-node motif structure” unless otherwise specified.

3 The important motif structure is a subset of the motif structure, which refers to the motif structure whose proportion in the real network is significantly higher than that in the random network.
Compared with other important motifs, Motif 4 shows a significantly abnormal downward trend during the financial crisis. Furthermore, the node attributes of Motif 4 are distinguished to explore the evolution of specific motif categories. Specifically, the Motif 4 category, which contains nodes of state-owned companies, general companies, and individual investors, is exceptionally low during the financial crisis. In contrast, the Motif 4 category, which consists of nodes of general companies and individual investors, is unusually high around the financial crisis.

This study makes several contributions to the field. First, the maturity of motif recognition algorithms has prompted a large number of studies to explore the motif structure of various networks, including biological system networks [10], social networks [11], electronic communication networks [12], investor transaction networks [13], interbank networks [14, 15], listed company networks [16] and stock time series networks [17]. However, few studies have explored the motif structures of investor information networks. Considering the link weights, node attributes and different time scales, this research explores the motif structures of the SCN and their dynamic evolution.

Second, the analysis of economic and financial networks as channels of crisis transmission has received extensive attention. Much effort has been devoted to the search for regularities in the overall structure of financial networks, i.e., looking for community structure [18], core-periphery structure [19], network connectedness [21], network centrality [22], and network structure [23]. However, a major financial crisis usually develops from a change in local network characteristics to the collapse of the overall network. Squartini et al. [14] and Saracco et al. [24] studied the abnormal motif mutations of interbank networks and world trade networks before and after the financial crisis, respectively. To most of our knowledge, few studies have conducted early warnings of financial crises from the perspective of shareholder information network motifs. This study explores the abnormal changes in the specific motif structures of the SCN before and after the financial crisis.

Third, this study also makes some theoretical contributions. The process of motif recognition needs to compare the relative values of the real network and the corresponding random network. The simplest and most widely used random network model in weighted networks is the weighted configuration model (WCM), which is defined as an ensemble of random graphs with a given strength sequence [25–28]. It was recently shown that, despite its conceptual simplicity, the WCM has brought significant bias problems [29–31]. This study introduces an analytical and unbiased maximum-entropy technique, which uses both strengths and degrees knowledge to reconstruct the unbiased ensemble of weighted networks. In the shortest possible time, this method directly provides the expected value of the desired reconstructed properties in such a way that no explicit sampling of reconstructed graphs is required. Moreover, being based on maximum-entropy distributions, this method is unbiased by construction.

In applying our enhanced method to several networks of different nature, the results show that it leads to a significantly improved reconstruction. Moreover, rigorous information-theoretic criteria are introduced to confirm that the joint specification of the strengths and degrees cannot be reduced to that of the strengths alone. In other words, the reconstruction of the weighted network can be greatly enhanced by exploiting the irreducible set of joint degrees and strengths.

The remainder of this paper is organized as follows. Section 2 describes data and modelling. Section 3 reports results and analysis classified into three parts: motif identification and motif evolution and risk warning of important motifs. Section 4 concludes the paper.

## 2 Data and modelling

### 2.1 Data

The data of the top ten shareholders of tradable stocks comes from the RESSET database. The time is from the first quarter of 2007 to the fourth quarter of 2017. As of the end of 2006, 1301 companies had completed the split share reform, accounting for 97% of the total number of listed companies. In order to exclude the influence of shareholder structure and equity nature on the results, the starting point of the data in this paper is the first quarter of 2007.

The samples are the constituent stocks of the CSI 300 index and are updated every quarter. Shareholder names in different files are manually unified. The CSI 300 index consists of 300 of the largest and most liquid A-shares. The index is designed to reflect the overall performance of China’s A-share market.
Subsidiaries of the same parent company are considered to be the same shareholder due to concerted action. In addition, index funds and index-enhanced funds have been eliminated because of their passive ownership of stocks and lack of information interaction with other shareholders. In addition, controlling shareholders are eliminated because their holdings are to gain control of the company, which is different from ordinary shareholders’ use of private information for profits. According to Li et al. [32], shareholders are divided into five types: financial investment companies (FICs), state-owned companies (SOCs), general companies (GCs), qualified foreign institutional investors (QFIIs), and individual investors (IIs). The classification standard for shareholder categories comes from the RESSET database.

2.2 The SCN modelling

According to the network construction method in Newman [33], the shareholders is taken as the node, and the number of stocks together held by two shareholders is the edge weight. SCN is a weighted but undirected network used to describe collective holding behaviour among investors. Using \( H = \{ H_1, H_2, \ldots, H_n \} \) to represent shareholders, matrix \( W \) describes the SCN as follows:

\[
W = \begin{cases} 
    a_{ij} = w_{ij} & \text{if } H_i \text{ and } H_j \text{ jointly hold } \text{ listed companies} \\
    a_{ij} = 0 & \text{if } H_i \text{ and } H_j \text{ jointly hold no listed companies} 
\end{cases}
\]

where \( w_{ij} \) is the number of listed companies that shareholders \( H_i \) and \( H_j \) jointly hold. Referring to Khan et al. [34, 35], the nomenclature is provided in Appendix Table 5.

2.3 Motif modelling

2.3.1 Motif categories in SCN

In undirected and unweighted networks, only two forms of 3-node motifs exist, shown as Motif 1 and Motif 4 in Fig. 1. The links between shareholders in the SCN are divided into strong and weak links. Shareholders jointly holding one stock is a weak link, while jointly holding more than one stock is a strong link. When considering the link weights, the 3-node motifs have seven forms, as shown in Fig. 1.

Figure 1 shows seven weighted 3-node motifs. The thick line is the strong co-holding relationship, and the thin line is the weak co-holding relationship.

Motifs 1, 2, and 3 have two nodes that are not directly connected but indirectly related through intermediate nodes to form a chain structure (called the chain motifs). In contrast, Motifs 4, 5, 6, and 7 are connected to form a closed structure (called the closed motifs). As the shape of the motif changes from chain to closed and the link weight changes from weak to strong, the speed of information transmission of the motif increasing progressively [8]. Furthermore, node attributes are considered and classified into five types: FIC, SOC, GC, QFII, and II. Figure 2 shows the schematic figure of motif identification when considering both link weights and node attributes.

Figure 2 shows the schematic figure of motif identification. The node colour represents the shareholder category, where the red node represents FIC, the black node represents SOC, and the blue node represents QFII. The red edges represent the process of motif identification.

The Rand-ESU algorithm was adopted to identify the motif structures in the SCN efficiently. Compared with traditional algorithms such as edge sampling (ESA) and enumerating subgraph (ESU), the Rand-
ESU algorithm judges whether a subgraph needs to be extended by probability in the process of subgraph enumeration, thus greatly improving detection efficiency. The motif identification steps of the Rand-ESU algorithm are as follows. First, the motif identification process shown in Fig. 2 on the real network is performed to calculate the proportion of various motifs. Second, 1000 simulations on random networks of the same scale are performed to calculate the proportion of various motifs. Finally, the motif statistical significance is calculated.

2.3.2 Statistical indicators of motifs

Three common types of indicators measure the importance of motifs in networks, including the frequency of motifs, the $p$ value of motifs, and the $Z$-score of motifs.

1) The frequency of the motif

For a 3-node Motif $M$ with $n$ nodes, the number of occurrences in the real network is $n(M)$. The total number of occurrences of all 3-node motifs is $N$; then, the frequency of the motif $M$ is as follows.

$$f(M) = \frac{n(M)}{N}$$  \hspace{1cm} (2)

2) The $p$ value of the motif

The $p$ value of the motif refers to the number of times that a specific motif appears more frequently in the random network than in the real network divided by the total number of random networks.\(^6\) Therefore, the $p$ value of the motif is between 0 and 1. The smaller the $p$ value, the more significant the motif, and the more critical it is in the real network.

3) The $Z$-score of the motif

For Motif $M$, the number of times it appears in the real network is $N_{\text{real}_M}$, the number of times it appears in the random network is $N_{\text{rand}_M}$, and the standard deviation is $\sigma_{\text{rand}_M}$. The average value of $N_{\text{rand}_M}$ is $N_{\text{rand}_M}$. The $Z$-score of Motif $M$ in the real network is as follows.

$$Z_M = \frac{N_{\text{real}_M} - N_{\text{rand}_M}}{\sigma_{\text{rand}_M}}$$  \hspace{1cm} (3)

\(^6\) The number of random network in this paper is 1000.

The construction logic of the $Z$-score and $p$ value is similar. They both compare the frequency of the motif in the real and random networks and use the difference to measure motif importance.

2.3.3 Random network model

The most widely used random network model in weighted networks is the weighted configuration model (WCM), which is defined as an ensemble of random graphs with a given strength sequence [25–28]. The expected weight of the link between nodes $i$ and $j$ predicted by the WCM is routinely written in the form

$$w_{ij} = \frac{s_i s_j}{\sum_{m=1}^{N} s_m}$$  \hspace{1cm} (4)

where $s_i$ denotes the strength of node $i$ and $N$ is the number of nodes.

Unfortunately, despite its widespread use, the WCM has brought significant problems of bias. Although Eq. (4) is treated as an expected value, there is no indication of which probability distribution it is derived from. Therefore, it is impossible to derive the expected value of topological properties, which are nonlinear functions of the weights. This research now develops a maximum-entropy approach to generate an unbiased “enhanced configuration model” (ECM) with given strengths and degrees as the random network model.

Formally, an ensemble of weighted networks with $N$ nodes can be characterized by a collection $\{W\}$ of $N \times N$ matrices and by an appropriate probability $P(W)$. On each network $W$, the strength is defined as $s_i(W) \equiv \sum_{i \neq j} w_{ij}$, and the degree is defined as $k_i(W) \equiv \sum_{i \neq j} w_{ij}^{0}$, $w_{ij}$ is defined as a non-negative integer number. The normalized probability needs to be found to ensure that each node’s the expected degree and strength are both constrained while leaving the ensemble as random as possible. This is achieved by requiring that $P(W)$ maximizes Shannon’s entropy $S \equiv - \sum_{W} P(W) \ln P(W)$ with a constraint on the expected degree and strength sequences $\bar{k}$, $\bar{s}$. The fundamental result of this constrained maximization is the probability
where $\bar{x}$ and $\bar{y}$ are two N-dimensional Lagrange multipliers controlling for the expected degrees and strengths, respectively (with $x_i \geq 0$ and $0 \leq y_i < 1$, $\forall i$), and

$$q_{ij}(w|x, y) = \frac{(x_i x_j)^{y_{ij}} (y_{ij})^{1 - y_{ij}}}{1 - y_{ij} + x_i x_j y_{ij}}$$

is the probability that a link of weight $w_{ij}$ exists between nodes $i$ and $j$. In the above expression, $(x) = 1$ if $x > 0$ and $(x) = 0$ otherwise. Note that $\sum_{w=0}^{+\infty} q_{ij}(w|x, y) = 1$, $\forall i, j$.

Due to the presence of $(x)$, Eq. (6) defines the ‘mixed’ Bose-Fermi distribution where establishing a different ‘cost’ than the reinforcement of an already existing link. This feature is due to the existence of both binary and weighted constraints, making ECM suitable for modelling real networks.

Next, this research applies the maximum-likelihood approach to implement the ECM. Consider a particular real weighted network $W^*$ with only degrees $k^*_i \equiv k_i(W^*)$ and strengths $s^*_i \equiv s_i(W^*)$. The log-likelihood of the ECM defined by Eqs. (5) and (6) is

$$\mathcal{L}(\bar{x}, \bar{y}) \equiv \ln P(W^*|\bar{x}, \bar{y}) = \sum_{i<j} \ln q_{ij}(w_{ij}|\bar{x}, \bar{y})$$

$$= \sum_{i=1}^{N} (k^*_i \ln x_i + s^*_i \ln y_i) + \sum_{i<j} \ln \left( \frac{1 - y_{ij}}{1 - y_{ij} + x_i x_j y_{ij}} \right).$$

Now, look for the specific parameter values $\bar{x}^*$, $\bar{y}^*$ that maximize $\mathcal{L}(\bar{x}, \bar{y})$. Similar to the solving process of other random network models [27], the real solutions of $\bar{x}^*$, $\bar{y}^*$ can be obtained from the following 2 N coupled equations:

$$k_i = \sum_{i<j} \frac{x_i x_j y_{ij}}{1 - y_{ij} + x_i x_j y_{ij}} = k^*_i \forall i$$

$$s_i = \sum_{i<j} \frac{x_i x_j y_{ij}}{(1 - y_{ij})(1 - y_{ij} + x_i x_j y_{ij})} = s^*_i \forall i$$

Therefore, the likelihood-maximizing values $\bar{x}^*$, $\bar{y}^*$ are precisely those ensuring that the expected degree and strength sequences are consistent with the observed sequences $\bar{k}$, $\bar{s}$. $\bar{x}$, $\bar{y}$ are exactly the values we want to find.

It is worth noting that the values $\bar{k}^*$ and $\bar{s}^*$ contain all the information needed to rebuild the network. Therefore, the maximum-likelihood approach equivalently translates the time-consuming and bias-prone problem of generating multiple reconstructed networks into the much simpler problem of maximizing the function $\mathcal{L}(\bar{x}, \bar{y})$ of 2N variables. To find $\bar{x}^*$ and $\bar{y}^*$, the assignment method implemented by MATLAB is applied to solve Eqs. (8)–(9). Note that finding $\bar{x}^*$ and $\bar{y}^*$ only requires the information of degrees and strengths and not that of the entire network $W^*$. This is consistent with the fact that $\bar{k}^*$ and $\bar{s}^*$ are sufficient statistics.

2.3.4 Empirical tests based on real networks

Next, the superiority of our enhanced model over the naive model will be determined based on the reproduction of different types of real network topological properties. This study considers two classic social networks from [36], including research group social network (RGSN) and fraternity social network (FSN), and six food networks collected in [37], including Chesapeake Bay food network (CBFN), Crystal River food network (CRFN), Michigan Lake food network (MiLFN), Maspalomas Lagoon food network (MaLFN), Mondego Estuary food network (MEFN) and Everglades Marshes food network (EMFN). In Fig. 3, the real and reconstructed values of the topological properties (pure and weighted) of all networks in the sample are compared.

The purely topological property we choose is the simplest non-local one, i.e. the average nearest neighbour degree (ANND) measured by the correlation between the degrees of adjacent nodes, defined as

$$\text{ANND}_i(W) \equiv \frac{\sum_{j \neq i} a_{ij} k_j}{k_i} = \frac{\sum_{j \neq i} \sum_{k \neq j} w_{ij}^0 w_{jk}^0}{\sum_{j \neq i} w_{ij}^0}$$

where $W$ is the full weighted matrix, and $w_{ij}$ is the weight of the link between node $i$ and node $j$. $a_{ij}$ is a Boolean variable with the value of 1 if $w_{ij} > 0$ and 0 if $w_{ij} = 0$ to judge whether there is a connection between nodes $i$ and $j$ (compactly, we can write $a_{ij} \equiv w_{ij}^0$ with the convention $0^0 \equiv 0$). The corresponding weighted
The property is the average nearest neighbour strength (ANNS) defined as

\[ \text{ANNS}(W) = \frac{\sum_{j \neq i} a_{ij} s_j}{k_i} = \frac{\sum_{j \neq i} \sum_{k \neq j} w_{ij}^0 w_{jk}}{\sum_{j \neq i} w_{ij}^0} \]

where \( s_i = \sum w_{ij} \).

For all nodes and all networks, the measured values of the two properties for the real network and the corresponding reconstructed network predicated by the WCM (ECM) are illustrated in the left (right) two panels of Fig. 3. Each point in the graph is a node. The goal of a good reconstruction approach is to place all the points along the identity line. In most cases, the reconstructed values of the WCM for all nodes in a given network are along the horizontal line, i.e. they are almost equal, completely independent of the true value. In contrast, our enhanced approach achieves a significant improvement over the naive approach.

Most points are near the identity, meaning that our approach can successfully reconstruct. Figure 3 presents the comparison of real (x-axis) and reconstructed (y-axis) values of topological properties for all eight sample networks, including RGSN, FSN, CBFN, CRFN, MiLFN, MaLFN, MEFN and EMFN. Top left: comparison of ANND between real network and WCM. Top right: comparison of ANND between real network and ECM. Bottom left: comparison of ANNS between real network and WCM. Top right: comparison of ANNS between real network and ECM.

The above findings indicate that the reconstructing networks from local node-specific information is intrinsically problematic, possibly because of the involvement of higher-order mechanisms in the formation of real networks. In fact, WCM is often used as a null model in important high-order attribute
detection to filter local heterogeneity of nodes, thus interpreting difference between real data and WCM as important features of non-local models. However, our enhancement method has now proved to reach a highly satisfactory level, at least for the networks considered here. Therefore, it can be expected that if the ECM is viewed as an improved null model to detect higher-order patterns such as communities or motifs, the result will be quite different from those obtained routinely using WCM prediction in the modular definition. Besides representing an improved reconstruction method, the ECM has the potential to become a non-trivial tool as a null model of networks with local constraints.

2.3.5 Information-theoretic tests

The superiority of our enhanced reconstruction method has been analysed on basis of its increased accuracy, with respect to the naive approach. A rigorous goodness-of-fit approach will be used in this section to confirm these results, comparing whether ECM preserves better efficiency from WCM. Precisely, the degrees as extra parameters introduced into the ECM will be tested for non-redundancy.

To start with, the likelihood of the ordinary WCM need to be compared with that of ECM. Note that the WCM can be regarded as a special case of the ECM by setting $x_i = 1 \forall i$. The log-likelihood of the WCM is therefore the reduced function $L(\bar{1}, \bar{y})$ of $N$ variables, and is maximized by a new vector $\bar{y}^* \neq \bar{y}$ which is also the solution of Eq. (9) with $x = \bar{1}$. In the WCM, Eq. (9) no longer works.

If the maximized likelihoods of WCM and ECM are simply compared, the conclusion $L(\bar{x}, \bar{y}) \geq L(\bar{1}, \bar{y}^*)$ can be easily obtained since ECM always improves the fit of the real network, given that it includes the WCM as a special case. However, statistical and information-theoretic criteria can be used to evaluate whether the increase in accuracy of a model with more parameters is the result of overfitting. The most popular and simplest criterion is Akaike’s information criterion (AIC). This test strictly achieves the optimal trade-off between accuracy and parsimony, which is achieved by discounting the number of free parameters from the maximized likelihood. AIC for our two competing null models is defined as

$$AIC_{ECM} = -2L(\bar{x}, \bar{y}) + 4N$$

$$AIC_{WCM} = -2L(\bar{1}, \bar{y}^*) + 4N$$

The optimal model is the one with minimum AIC. Even though the difference between AIC values is small, the two models are still comparable. A converted form of quantitative criterion is given by the so-called AIC weights, which in our case read

$$w_{AIC}^{ECM} = \frac{e^{-AIC_{ECM}/2}}{e^{-AIC_{ECM}/2} + e^{-AIC_{WCM}/2}}$$

$$w_{AIC}^{WCM} = 1 - w_{AIC}^{ECM}$$

where $w_{AIC}^{ECM}$ is the probability of supporting ECM as the best model, and $w_{AIC}^{WCM}$ is the one of supporting WCM as the best model.

The AIC weights of the two reconstruction methods are shown in Table 1 for all networks. With the exception of a social network, our enhanced approach is always superior to a naive approach and achieves the best unit probability in the two models. A closer look at this social network with the opposite result reveals that it is almost fully connected. This explains why the degree sequence is redundant for this network. The local constraints are reduced to strength sequences, so the “naive” WCM is preferable. However, most real-world networks are often not fully connected, and the strengths and the degrees must be separately specified for better reconstruction. Therefore, it can be expected that degree sequences are irreducible to strength sequences for most real-world networks. In such case, the inclusion of degree sequences in our enhanced method is non-redundant, which explains why our method can maintain higher information efficiency.

In addition, corrected Akaike’s Information Criterion (AICc) is also used to correct small samples. Based on the value of AICc, the corresponding weights are calculated in analogy with Eqs. (12) and (13). For all the networks in our sample, the AICc weights are the same as the AIC ones.
3 Results and analysis

3.1 Summary statistics

Panel A of Table 2 shows the summary statistics of shareholdings by different types of shareholders. FIC shareholders are the most frequent, accounting for approximately 62%, while SOE shareholders are the least frequent. QFII shareholders hold the largest number of shares on average. SOE shareholders hold the highest average market capitalization of their shares. Panel B shows the summary statistics of strong links. The top five strong links all include FIC shareholders, of which FIC-FIC links account for the highest proportion at 36.92%. Panel C shows the summary statistics of weak links. The top four weak links all include FIC shareholders. Similar to Panel B, the FIC-FIC links have the highest proportion. Panel D shows descriptive statistics of network topology characteristics. The average number of nodes in the SCN is 1237.30, and the average number of edges is 8234.98.

3.2 Motif identification

3.2.1 Consider the link weights

The motif structures help to understand local information transmission among shareholders. Considering the link weights, the important motifs of the SCN are explored in this section. Table 3 provides the results. Although with a low proportion, the four types of closed motifs were found to be important motifs with statistical significance in the network. In contrast, although the proportion of chain motifs is very high, they are not important motifs in the SCN. These results indicate that the closed motifs play an essential role in information transmission among shareholders.

3.2.2 Consider the node attribute

Next, further considering the node attributes, the components of the important motifs (closed motifs) in the SCN are explored. Table 4 lists the top five closed motif types with the highest frequency of occurrence in all quarters. The results show that the motifs with three FIC nodes have the highest proportion. Other types of motifs contain at least two FIC nodes. In brief, FIC shareholders are important components of closed motifs. Additionally, the proportion of closed motifs in 2013–2017 is lower than that of the other two periods, which may be related to the stock market crash in 2016. A large number of shareholders sold their shares, causing the motif to disappear.

3.3 Motif evolution

Next, the motif evolution of the SCN is investigated to reveal the behavioural laws of shareholders in different motif structures. The motif evolution is divided into three steps: First, the shareholder intersection of the previous and subsequent periods is calculated. Second, the local structure of motif 1–7 in the SCN formed by shareholder intersection in the previous period is excavated. Finally, the corresponding evolved motif structure is mined in the later period. Two time scales, short cycle (quarterly) and long cycle (annual), are selected to explore the evolution of motifs.

3.3.1 Motif evolution in the short cycle

The motif in the later period corresponds to three evolution states: invariant motif, vanishing motif, and transforming motif. Specifically, the invariant motif refers to a motif that maintains the same structure in the two periods. The vanishing motif refers to a motif with at least one isolated node due to edge fracture. The transforming motif refers to a motif change between two periods due to edge fracture or reconnection.

Figure 4 shows the results of the motif evolution on the quarterly time scale. The first column corresponds
to the motifs with at least one isolated node, representing the vanishing motif. The diagonal line represents invariant motifs that have the same structure in the front and later periods. The other positions in Fig. 4 are the transforming motifs.
Figure 4 shows the motif evolution in the short cycle. The vertical axis is the motif structure in the previous period, and the horizontal axis is the motif structure in the later period. The circle (and the shade of the colour) represents the evolution probability, which is the average of all quarters.

As shown in the first column, all motif structures have a higher tendency to disappear. With the change in the link weights from weak to strong and the structure from chain to closed (from Motif 1 to Motif 7), the trend of motif disappearance gradually decreases. Second, as shown from the diagonal, all motif structures have a relatively high propensity to remain unchanged, especially Motif 4. Third, in the transforming motif, the transforming probability from a motif with high information transmission speed (lower triangle) to a motif with low information transmission speed (upper triangle) is higher. Nevertheless, the possibility of the reverse is relatively lower. In short, edges tend to break rather than reconnect. Such evolution results are in line with the principle of entropy increase in physics. The system always tends to transform from an ordered state to a disordered state [38].

Table 3 Motif identification

|       | Freq(%) | Z-score | Freq(%) | Z-score | Freq(%) | Z-score | Freq(%) | Z-score |
|-------|---------|---------|---------|---------|---------|---------|---------|---------|
| 2007  | 62.69   | -134.55 | 25.29   | -29.69  | 2.29    | -28.94  | 4.34*** | 138.79  |
| 2008  | 66.73   | -187.35 | 23.95   | -31.80  | 1.92    | -31.37  | 3.62*** | 172.71  |
| 2009  | 61.55   | -165.76 | 27.13   | -15.32  | 2.87    | -27.55  | 3.87*** | 144.28  |
| 2010  | 63.58   | -257.27 | 26.57   | -12.31  | 3.10    | -25.63  | 3.44*** | 192.98  |
| 2011  | 63.73   | -294.81 | 26.25   | -6.41   | 3.34    | -26.30  | 3.86*** | 214.66  |
| 2012  | 62.92   | -249.12 | 26.66   | -5.71   | 3.33    | -35.49  | 4.07*** | 212.72  |
| 2013  | 67.90   | -266.91 | 23.04   | -10.76  | 2.16    | -53.25  | 4.37*** | 289.66  |
| 2014  | 70.53   | -378.90 | 19.99   | 3.62    | 1.84    | -55.66  | 5.47*** | 432.68  |
| 2015  | 66.75   | -219.27 | 25.12   | -27.23  | 2.38    | -30.64  | 3.80*** | 115.20  |
| 2016  | 65.89   | -196.07 | 28.08   | -3.59   | 2.47    | -14.55  | 1.82*** | 177.74  |
| 2017  | 69.15   | -314.15 | 25.53   | -5.73   | 1.88    | -21.65  | 1.95*** | 183.43  |
| All   | 65.58   | -242.20 | 25.24   | -13.18  | 2.51    | -31.91  | 3.69*** | 206.80  |

Table 3 shows the significance of motif occurrences. Freq represents the frequency of the motif, and the Z-score represents the Z score of the motif. The proportion of motifs in each year is the average of the four quarters, and the last row is the average of all quarters. *** indicates the significance level of 1%.
Next, considering the node attributes, the composition of each motif evolution state are calculated. The results are shown in Fig. 5. To ensure the accuracy of the results, motif types whose total number is less than 100 are deleted. Among the invariant motifs, the proportion of yellow and green nodes is relatively high, indicating that the holdings of GC and II are stable and tend to remain unchanged during motif evolution. Especially for Motif 4, the probability that the motif with all three nodes as II remains unchanged is 0.98. Furthermore, among the vanishing motifs, the motifs containing red nodes account for a relatively high proportion, indicating that FIC has active market transactions and is likely to disappear during motif evolution. Finally, among the transforming motifs, the holdings of FIC and QFII are unstable, so motifs containing these two types of nodes are easily transformed into other varieties in the evolution process.

Figure 5 shows motif evolution in the short cycle considering node attributes. 1st, 2nd, and 3rd represent the evolution probability of the top 3 motifs. The number in parentheses indicates the probability of motif evolution. Taking 0.81 in the upper left corner of the figure as an example, it means that in the evolutionary motif where Motif 1 remains invariant, the probability of the motif with all three nodes being GC is the highest, which is 0.81. In other words, there are 100 motifs 1 with all three nodes as GC, and 81 motifs 1 remain unchanged in the next quarter. The node colour represents the shareholder category. The red node represents FIC, the yellow node represents GC, the green node represents II, the black node represents SOC, and the blue node represents QFII.

The above analysis shows that motifs containing GC and II are stable during the quarterly evolution, and motifs containing FIC tend to transform or even disappear. In contrast, motifs containing QFII tend to transform into other types.

**3.3.2 Motif evolution in the long cycle**

Next, the motif evolution is discussed on the annual scale. The calculation steps are the same as those in the quarterly period. The results are shown in Fig. 6.
Figure 6 shows the motif evolution in the long cycle. The vertical axis is the motif structure in the previous period, and the horizontal axis is the motif structure in the later period. The circle (and the shade of the colour) represents the evolution probability, which is the average of all quarters.

As shown in the first column of Fig. 6, all the motif structures have a high tendency to disappear. Compared with the motif evolution results in a short period, the probability of the vanishing motif in a long period is better. These results show that investors’ shareholding behaviour is more unstable in a long period. Furthermore, as shown in the diagonal, all the motif structures have a relatively higher tendency to remain unchanged. Nevertheless, the probability is reduced compared with the situation in a short period. Third, the transforming probability of high information transmission speed motifs (lower triangle) to low information transmission speed motifs (upper triangle) is higher. These results are the same as the conclusion of the short-period motif evolution. The system always tends to evolve to a state of increasing entropy.

Next, considering the node attributes, the compositions of each motif evolution state are counted. The results are shown in Fig. 7. The conclusion of the motif evolution in the long period is the same as that in the short period. However, the probability of motif evolution remaining unchanged in the long period is less than that in the short period. In contrast, the probability of the motif vanishing and transforming in the long period is greater than that in the short period.

Figure 7 shows motif evolution in the long cycle considering node attributes. 1st, 2nd, and 3rd represent the evolution probability of the top 3 motifs. The number in parentheses indicates the probability of
motif evolution. The node colour represents the shareholder category. The red node represents FIC, the yellow node represents GC, the green node represents II, the black node represents SOC, and the blue node represents QFII.

3.4 Risk warning of important motifs in the financial crisis

The stock market is essentially an information market, and information guides stock prices [39]. As important investors in listed companies, shareholders often play the role of information carriers [32]. Major financial risk events are often caused by unpredictable random events that trigger minority shareholders to sell stocks. This behaviour spread throughout the network and ultimately led to the collapse of the entire stock market. In 2015, the Chinese stock market suffered the worst crash in history. The Shanghai Composite index fell as much as 2000 points in a two-month period from June to July. The huge short-term volatility of the stock market caused heavy losses for investors. This section will explore the evolution of important motifs during the stock market crash.

Regardless of the node attributes, only the link weights are considered. Each panel in Fig. 8 demonstrates the evolution law of the different important motifs. To establish the proposed scheme, the motif ratios of the real network and the null network (ECM) are shown in the same figure. Intuitively, the proportion of Motif 4 in the real network compared to that in the null model shows an abnormal decline from 2014 before the stock market crash to 2016 after the stock market crash. In contrast, Motif 5, Motif 6 and Motif 7 show no significant decreasing trend.

Furthermore, the node attributes of Motif 4 are distinguished to explore the evolution of specific motif categories. The top 10 Motif 4 categories per year considering node attributes are calculated and shown in Fig. 9. The Motif 4 category composed of GC, II and SOC maintained a high proportion until 2014. However, in the years before and after the stock market crash (2014–2016), the motif experienced an abnormal decline. In contrast, the Motif 4 category, which consists only of GC and II nodes, was unusually high around the stock market crash. These results suggest that SOCs did not play a “ballaststone” role in stabilizing the market during the financial crisis. In contrast, GC and II became “scapegoat” during the financial crisis and suffered huge losses.

Figure 8 displays the evolution of important motifs. The horizontal axis is the time, and the vertical axis is the motif ratio. The green line represents the motif.

| Motif | Invariant motif | Vanishing motif | Transforming motif |
|-------|----------------|----------------|------------------|
| 1st   |                |                |                  |
| 2nd   |                |                |                  |
| 3rd   |                |                |                  |
|       | (0.58)         | (0.53)         | (0.52)           |
|       | (0.56)         | (0.53)         | (0.53)           |
|       | (0.55)         | (0.43)         | (0.50)           |
|       | (0.96)         | (0.91)         | (0.84)           |
|       | (0.64)         | (0.58)         | (0.57)           |
|       | (0.60)         | (0.50)         | (0.31)           |
|       | (0.50)         | (0.35)         | (0.35)           |

Fig. 7 Motif evolution in the long cycle considering node attributes
ratio in the real network, and the red line represents the motif ratio in the null model (ECM). Top left: Motif 4 ratio. Top right: Motif 5 ratio. Bottom left: Motif 6 ratio. Bottom right: Motif 7 ratio.

Figure 9 shows the top 10 Motif 4 categories per year considering node attributes. The horizontal axis is the time, and the vertical axis is the top ten Motif 4 categories considering node attributes. The red node represents FIC, the yellow node represents GC, the green node represents II, the black node represents SOC, and the blue node represents QFII.

4 Conclusion

Based on shareholding data from 2007 to 2017, this study mines the motif structures of the SCN from a micro-perspective. Specifically, considering the link weights, the closed motifs are confirmed to be important motifs with statistical significance in the network. Further considering the node attributes, the motifs with all three nodes being FIC have the highest proportion.

Furthermore, the evolution of different types of motifs is investigated on two time scales: the quarterly and the annual cycle. All motif structures are found to have a higher tendency to vanish. The longer the period, the greater the possibility of vanishment. Furthermore, considering the node attributes, motifs containing FIC nodes have a higher tendency to disappear, while motifs containing GC nodes tend to keep the structure unchanged.

Once again, important motifs are being explored to determine if abnormal quantitative mutations exist before and after the financial crisis. Motif 4, which contains SOC, GC, and II nodes, was exceptionally
low during the financial crisis. In contrast, the Motif 4 category, which consists only of GC and II nodes, was unusually high around the financial crisis. These results show that SOCs did not play a role in stabilizing the market during the financial crisis.

In addition, at the theoretical level, this research introduces an analytical and unbiased maximum-entropy technique, which uses the knowledge of both strengths and degrees to reconstruct the unbiased ensemble of weighted networks. In applying our enhanced method to several real networks of different nature, the results show that it leads to a significantly improved reconstruction. Moreover, rigorous information-theoretic criteria are introduced to confirm that the joint specification of the strengths and degrees cannot be reduced to that of the strengths alone.

In short, this research explores the microstructure and evolution mechanism of information interaction among shareholders from a micro-perspective. It is of great significance to understand how information is transmitted in the network.

Acknowledgements This article is funded by the National Natural Science Foundation of China (U1811462, 71771170, 71790594), Youth Foundation for Humanities and Social Sciences Research of the Ministry of Education (20YJC790062), Open Project of Jiangsu Key Laboratory of Financial Engineering (NSK2021-18), Jiangsu Planning Office of Philosophy and Social Science (20GLC007) and the Applied Economics of Nanjing Audit University of the Priority Academic Program Development Phase III of Jiangsu Higher Education Institutions (Office of Jiangsu Provincial People’s Government, No. [2018]87).

Funding This article is funded by the National Natural Science Foundation of China (U1811462, 71771170, 71790594), Youth Foundation for Humanities and Social Sciences Research of the Ministry of Education (20YJC790062), Open Project of Jiangsu Key Laboratory of Financial Engineering (NSK2021-18), Jiangsu Planning Office of Philosophy and Social Science (20GLC007) and the Applied Economics of Nanjing Audit University of the Priority Academic Program. Development Phase III of Jiangsu Higher Education Institutions (Office of Jiangsu Provincial People’s Government, No. [2018]87).

Declarations

Conflict of interest The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be constructed as a potential conflict of interest. The authors have no affiliation with any organization with a direct or indirect financial interest in the subject matter discussed in the manuscript.

Data availability The authors confirm that data will be made available on reasonable request.
Appendix

See Table 5.

Table 5  List of symbol

| Symbol | Definition |
|--------|------------|
| H      | Shareholder |
| w      | Link weight |
| W      | Matrix      |
| M      | Motif       |
| n      | The number  |
| N      | The total number |
| f      | Function    |
| Z      | Z-score     |
| N_{real} | The number of motif in the real network |
| N_{rand} | The number of motif in the random network |
| \sigma_{rand} | The standard deviation of the number of motif in an ensemble of random network |
| s      | Strength of node |
| k      | Degree of node |
| P      | Probability  |
| q      | Probability density |
| S      | Shannon’s entropy |
| \tilde{x} | N-dimensional Lagrange multiplier |
| L      | The log-likelihood estimation |

References

1. Hong, H., Kubik, J.D., Stein, J.C.: Thy neighbor’s portfolio: word-of-mouth effects in the holdings and trades of money managers. J. Finance 60(6), 2801–2824 (2005)
2. Shue, K.: Executive networks and firm policies: evidence from the random assignment of MBA peers. Rev. Financ. Stud. 26(6), 1401–1442 (2013)
3. Fracassi, C.: Corporate finance policies and social networks. Manag. Sci. 63(8), 2420–2438 (2017)
4. Li, H., An, H., Huang, J., Huang, X., et al.: The evolutionary stability of shareholders’ co-holding behavior for China’s listed energy companies based on associated maximal connected sub-graphs of derivative holding-based networks. Appl. Energy 162, 1601–1607 (2016)
5. Crane, A.D., Koch, A., Michenaud, S.: Institutional investor cliques and governance. J. Financ. Econ. 133(1), 175–197 (2019)
6. Pareek, A.: Information networks: implications for mutual fund trading behavior and stock returns. Working paper (2012)
7. Milo, R., Shen-Orr, S., Itzkovitz, S., Kashtan, N., et al.: Network motifs: simple building blocks of complex networks. Science 298(5594), 824–827 (2002)
8. Guan, Q., An, H., Liu, N., An, F., et al.: Information connections among multiple investors: evolutionary local patterns revealed by motifs. Sci. Rep. 7, 14034 (2017)
9. Battiston, S., Farmer, J.D., Flache, A., Garlaschelli, D., et al.: Complexity theory and financial regulation. Science 351(6275), 818–819 (2016)
10. Lee, T.I., Rinaldi, N.J., Robert, F., Odom, D.T., et al.: Transcriptional regulatory networks in Saccharomyces cerevisiae. Science 298(5594), 799–804 (2002)
11. Harrigan, N., Achananuparp, P., Lim, E.P.: Influentials, novelty, and social contagion: the viral power of average friends, close communities, and old news. Soc. Netw. 34(4), 470–480 (2012)
12. Kovanen, L., Kaski, K., Kertesz, J., Saramaki, J.: Temporal motifs reveal homophily, gender-specific patterns, and group talk in call sequences. Proc. Natl. Acad. Sci. 110(45), 18070–18075 (2013)
13. Jiang, Z.Q., Xie, W.J., Xiong, X., Zhang, W., et al.: Trading networks, abnormal motifs and stock manipulation. Quant. Finance Lett. 1(1), 1–8 (2013)
14. Squartini, T., Lelyveld, I., Garlaschelli, D.: Early-warning signals of topological collapse in interbank networks. Sci. Rep. 3, 3357 (2013)
15. Bargigli, L., Di Iasio, G., Infante, L., Lillo, F., et al.: The multiplex structure of interbank networks. Quant. Finance 15(4), 673–691 (2015)
16. Takes, F.W., Kosters, W.A., Witte, B., et al.: Multiplex network motifs as building blocks of corporate networks. Appl. Netw. Sci. 3(1), 39 (2018)
17. Liu, S., Huang, S., Chi, Y., Feng, S., et al.: Three-level network analysis of the North American natural gas price: a multiscale perspective. Int. Rev. Financ. Anal. 67, 101420 (2020)
18. Bargigli, L., Gallegati, M.: Random digraphs with given expected degree sequences: a model for economic networks. J. Econ. Behav. Organ. 78, 396–411 (2011)
19. Veld, D.I., Lelyveld, I.: Finding the core: network structure in interbank markets. J. Bank. Finance 49, 27–40 (2014)
20. Nobi, A., Lee, S., Kim, D.H., Lee, J.W.: Correlation and network topologies in global and local stock indices. Phys. Lett. A 378, 2482–2489 (2014)
21. Minoiu, C., Kang, C., Subrahmanian, V.S., Berea, A.: Does financial connectedness predict crises? Quant. Finance 15(4), 607–624 (2015)
22. Samitas, A., Kampouris, E., Kenourgios, D.: Machine learning as an early warning system to predict financial crisis. Int. Rev. Financ. Anal. 71, 101507 (2020)
23. Li, S., Wang, C.: Network structure, portfolio diversification and systemic risk. J. Manag. Sci. Eng. 6(2), 235–245 (2021)
24. Saracco, F., Clemente, R.D., Gabrielli, A., Squartini, T.: Detecting early signs of the 2007–2008 crisis in the world trade. Sci. Rep. 6, 30286 (2016)
25. Park, J., Newman, M.E.: Statistical mechanics of networks. Phys. Rev. E 70, 066117 (2004)
26. Garlaschelli, D.: The weighted random graph model. New J. Phys. 11, 073005 (2009)
27. Squartini, T., Garlaschelli, D.: Analytical maximum-likelihood method to detect patterns in real networks. New J. Phys. 13, 083001 (2011)
28. Squartini, T., Fagiolo, G., Garlaschelli, D.: Randomizing world trade. II. A weighted network analysis. Phys. Rev. E 84, 046118 (2011)
29. Garlaschelli, D., Loffredo, M.I.: Generalized Bose-Fermi statistics and structural correlations in weighted networks. Phys. Rev. Lett. 102, 038701 (2009)
30. Roberts, E.S., Coolen, A.C.C.: Unbiased degree-preserving randomization of directed binary networks. Phys. Rev. E 85, 046103 (2012)
31. Mastrandrea, R., Squartini, T., Fagiolo, G., Garlaschelli, D.: Enhanced reconstruction of weighted networks from strengths and degrees. New J. Phys. 16, 043022 (2014)
32. Li, J., Zhang, Y., Feng, X., An, Y.: Which kind of investor causes comovement? J. Int. Finan. Mark. Inst. Money 61, 1–15 (2019)
33. Newman, M.E.: Scientific collaboration networks. I. Network construction and fundamental results. Phys. Rev. E 64(1), 016131 (2001)
34. Khan, N.S., Kumam, P., Thounthong, P.: Computational approach to dynamic systems through similarity measure and homotopy analysis method for renewable energy. Curr. Comput.-Aided Drug Des. 10, 1086 (2020)
35. Khan, N.S., Kumam, P., Thounthong, P.: Magnetic field promoted irreversible process of water based nanocomposites with heat and mass transfer flow. Sci. Rep. 11, 1692 (2021)
36. Killworth, P.D., Bernard, H.R.: Informant accuracy in social network data. Hum. Organ. 35(3), 269–286 (1976)
37. Pajek datasets http://vlado.fmf.uni-lj.si/pub/networks/data/bio/foodweb/foodweb.htm
38. Li, X., Sun, M., Boersma, K.: Policy spillover and regional linkage characteristics of the real estate market in China’s urban agglomerations. J. Manag. Sci. Eng. 4(3), 189–210 (2019)
39. Dow, J., Gorton, G.: Stock market efficiency and economic efficiency: Is there a connection? J. Finance 52(3), 1087–1129 (1997)

Publisher’s Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.