Quantitative Finance, 2021
https://doi.org/10.1080/14697688.2021.1916067

Dynamic patterns of daily lead-lag networks in stock markets

YONGLI LI *†, CHAO LIU†‡, TIANCHEN WANG† and BAIQING SUN†

†School of Economics and Management, Harbin Institute of Technology, Harbin 150001, People's Republic of China
‡School of Business Administration, Northeastern University, Shenyang 110169, People's Republic of China

(*Received 27 September 2020; accepted 5 April 2021; published online 17 June 2021*)

The lead-lag relationship between stocks is an interesting phenomenon, which has been empirically seen to widely exist in stock markets. This paper aims to discover the dynamic patterns of the daily lead-lag relationships between stock pairs, to detect the features of the discovered dynamic patterns, and to explore which factors significantly affect the emergence of the feature. To this end, a series of statistical analyses is conducted to find that the (longest) successive lead-lag days satisfy a power-law distribution in the two mainland stock markets in China, which answers the question regarding the dynamic pattern. Note that the heavy tail of the power-law distribution is the core of the discovered dynamic pattern. A formal and solid definition of the lead-lag effect is provided by statistical testing, and then the corresponding detection method is designed and applied to obtain the heavy tail. Finally, an empirical study of the detected stocks with lead-lag effect is further conducted via an exponential random graph model (ERGM). Our work adds new knowledge to the lead-lag phenomenon in the financial domain, provides a formal definition of the lead-lag effect and proposes a new detection method benefiting future studies on the lead-lag relationship in financial markets. It further contributes to the existing relevant literature by a deep understanding of which factors cause the emergence of the power-law distribution discovered.

Keywords: Stock market; Lead-lag network; Lead-lag effect; Complex network; Power-law distribution

JEL Classification: C59, C81, D85, G19

1. Introduction

The lead-lag effect, a phenomenon where one security leads the price movement of another security with some time delay, has been empirically noted to widely exist in financial markets generally (Tolikas 2018) and particularly in stock markets (Wang *et al.* 2017). Within a given stock market, if a stock is regarded as a node and a lead-lag relationship between a stock pair is regarded as a direct link, then a lead-lag network can be established by the stocks and their relationships. When a lead-lag relationship is detected on each trading day, the lead-lag network obtained is called a daily lead-lag network. There is no doubt that the lead-lag networks obtained will evolve day by day. However, do stable patterns emerge during daily evolutionary processes? Furthermore, can we find some factors that can explain the discovered patterns if they do indeed exist? Accordingly, this paper aims to explore the dynamic patterns of the daily lead-lag networks obtained in several targeted stock markets and then to examine which factors have a significant influence on the dynamic patterns discovered.

Once the above mentioned research questions are answered, we will gain new insights into macroscopic emergence in financial markets. In fact, this phenomenon of macroscopic emergence has been widely discovered in social systems. For example, the reply times of e-mail have been found to satisfy a power-law distribution when the e-mail reply times of numerous persons are statistically analyzed, although the reply times of one single person or several persons would present no stable pattern and even seems random (Barabasi and Albert 1999, Barabasi 2005). In other words, so-called macroscopic emergence means that one event is random or no regular pattern is seen from the individual level, but will present stable patterns if seen from the group level (Ryan 2007). This phenomenon inspires us to explore whether macroscopic emergence also exists in the financial domain. Then, we choose the daily lead-lag relationships between stock pairs as the analyzed target. It is worth noting that the daily lead-lag relationship between one single stock pair would be randomly formed and would present no

*Corresponding author. Email: liyongli@hit.edu.cn particular pattern during a period, but when these relationships between all stock pairs in a market are considered, some interesting patterns are likely to emerge from such a macro perspective.

Moreover, answering our research question is beneficial to risk management in stock markets since the answer provides knowledge of the driving factors that lead to the dynamic pattern of the lead-lag relationship mentioned above. The lead-lag relationships between stock pairs will be a channel for transforming the risk by considering the inner factors that lead to the lead-lag relationship. Once the driving factors in stock markets are found, the predictability of market risk will be improved, which is a critical step for the effective management of financial stability. Although the topic of the lead-lag relationship has been discussed in much existing literature such as the works of Stübinger (2018), Challet *et al.* (2018), and Basnarkov *et al.* (2020), only a few studies have focused on mining the driving factors that cause the lead-lag relationship. For example, a famous work by Kobayashi and Takaguchi (2018) found the stable dynamic pattern of the lead-lag relationship in interbank credit networks and explained the origin of the "social" dynamics pattern. Parallel to these studies, this paper adopts the dynamic network evolution model, i.e. the exponential random graph model (see Harris (2013) for a brief introduction or Lusher *et al.* (2013) for a more detailed study), to analyze the sequence of daily lead-lag networks by empirically examining the driving factors.

Although the research questions to be explored sound meaningful and exciting, it is not easy to provide convincing answers. First, the definition of the lead-lag effect has not yet been unified, and thus we should propose an explicit definition which is often ignored in the existing literature. Second, many factors may be potential influencing factors that explain the phenomenon of macroscopic emergence, but a large amount of data is needed to determine the significant factors. After addressing these difficulties, the main findings and our contributions can be summarized into three aspects.

First, one stock pair can form a lead-lag relationship on successive trading days (often many times) during the analyzed period. Then, considering all stock pairs' successive trading days as mentioned above, their distribution is found to follow a power-law distribution in two stock markets-the Shanghai Stock Exchange and the Shenzhen Stock Exchange. The discovered distribution can be regarded as a stable pattern of the dynamic lead-lag relationships among all the stock pairs in the two stock markets, which fulfills the examples of macroscopic emergence in the financial domain from the perspective of network science.

Second, the heavy tail is the key part of one power-law distribution, which hints that the stock pairs in the scope of the heavy tail will be critical to analyzing the formed power-law distribution in particular or the system's macroscopic emergence in general. Note that if all the lead-lag relationships are randomly formed at the individual level, the distribution of the successive lead-lag days will be an exponential distribution at the system level rather than the discovered power-law distribution. In order to distinguish the key part of the power-law distribution from the exponential distribution, a new approach for detecting the lead-lag effect, which explicitly specifies the definition of the lead-lag effect according to the principle of statistical test, is proposed in this paper. This contribution of providing a solid definition lays a foundation for detecting the lead-lag effect in the relevant research fields.

Third, if some factors are found to significantly affect the formation of the heavy tail of the observed power-law distribution via empirical analysis, they are influencing factors that will contribute to risk management in stock markets. Note that the heavy tail refers to the special parts that are worthy of our attention because the number of stock pairs within the heavy tail is not too small to be ignored in a power law distribution. Although several risk management tools for stock markets, such as those by Acemoglu *et al.* (2015), Berndsen *et al.* (2016) and Li *et al.* (2018), have been proposed, the factors we examine will become new indicators that can predict price fluctuations, risk transmission and even market stability through the lead-lag effect.

In order to logically and clearly present our work and contributions, the remaining parts are organized as follows: Section 2 reviews the related work on the lead-lag relationship studied in the financial domain. Section 3 reports the dynamic patterns of daily lead-lag networks by selecting data and performing statistical analysis. Section 4 defines the lead-lag effect from the stable dynamic patterns revealed and further provides the approach for detecting stock pairs with the defined lead-lag effect. Section 5 empirically examines which factors significantly explain the formation of the lead-lag effect via the exponential random graph model (ERGM, for short). Section 6 concludes and discusses future work.

2. Related work

The lead-lag network studied in this paper is essentially one kind of stock network, and thus, we first briefly review the existing research on stock networks. Generally, the existing literature in this direction can be classified according to the network types. As shown in Figure 1, two common network types are often seen: synchronous networks and asynchronous networks. The so-called synchronous stock network is an efficient tool for summarizing and visualizing the correlations between stocks or stock markets by utilizing the synchronous data of the analyzed targets (Tse *et al.* 2010), such as the closing prices of different stocks on the same day (Boginski *et al.* 2006, Kinnunen 2017). In contrast to a synchronous stock network, an asynchronous stock network, such as the lead-lag network (Basnarkov *et al.* 2020) and the asynchronous

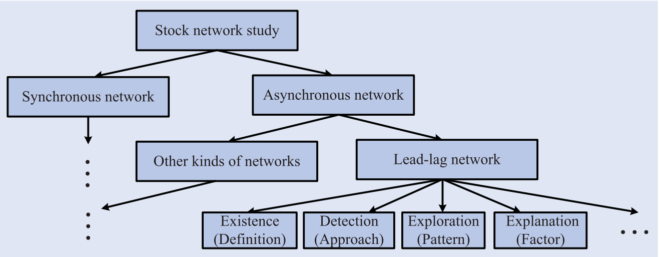

Figure 1. Structure of the related work.

trading network (Liu *et al.* 2010), focuses on the stocks' asynchronous data (Cai *et al.* 2017).

Further focusing on our studied lead-lag relationship in stock markets, one early study is Tóth and Kertész (2006), which analyzed how the lead–lag relationships between daily returns of stocks evolved. In addition to the abovementioned analysis of daily data, a series of studies have been conducted on high-frequency data, such as Jong and Nijman (1997), Huth and Abergel (2011, 2014), Pomponio and Abergel (2013), Buccheri *et al.* (2019) and many others not mentioned. Since the lead-lag in financial markets is a classical topic, there have been numerous relevant researches on this topic. Although we cannot review all of them in this paper, we will select the most relevant ones with the lead-lag networks from them. In order to review these selected aspects logically, four aspects (i.e. existence, detection, exploration and explanation), as shown in Figure 1, are assessed and organized as follows.

Existence. The lead-lag effect has been empirically evidenced to widely exist in stock markets. For example, Gong *et al.* (2016) studied the lead-lag relationship between the China Securities Index 300 (CSI 300), Hang Seng Index (HSI), Standard and Poor's 500 (S&P 500) Index and their associated futures to reveal the dynamic patterns of their relationships over time. In addition, Hayashi and Koike (2018) proposed wavelet-based methods to conduct high-frequency lead-lag analysis between stocks in the selected stock markets. Similar existing studies can also be found in Fonseca and Zaatour (2017), Dao *et al.* (2018), Scherbina and Schlusche (2018), O'Neill and Rajaguru (2019), Yao and Li (2020), and many others not mentioned. These existing studies have laid a solid foundation for our work by revealing that the lead-lag effect is quite likely to exist in our targeted stock markets and is therefore worth exploring and discovering. However, there is no precise definition of the lead-lag effect, although many literatures have focused on this phenomenon. At least, the various definitions of the lead-lag effect will cause difficulties in comparing the findings of the existing studies. Thus, this paper potentially contributes by providing a precise definition of the lead-lag effect based on a generally accepted principle or criterion.

Detection. Apart from the abovementioned empirical findings, many approaches have been proposed to detect the lead-lag effect. The existing methods include but are not limited to the following: Fiedor (2014) created an information-theoretic approach to detect the lead-lag effect in financial markets; Curme *et al.* (2015) proposed a numerical method to statistically validate links in correlation-based networks to detect the lead-lag relationship; the DTW (dynamic time warping) algorithm used in the asynchronous time series analysis became a common approach for constructing the lead-lag network (Ito and Sakemoto 2020); the Granger causality test also made a significant contribution to the discovery of the lead-lag relationship (Výrost *et al.* 2015, Basnarkov *et al.* 2020); and recently, O'Neill and Rajaguru (2019) designed a new response surface analysis of critical values for the lead-lag ratio based on high-frequency and non-synchronous financial data. Different from these excellent and somewhat complicated detection approaches, a solid statistical test model is designed in this paper to detect stock pairs with a lead-lag effect. More importantly, the rationality, robustness and even predictability of the proposed model are deeply studied. From this viewpoint, our work has contributed a new approach to the collection of approaches for detecting the lead-lag effect.

Exploration. Earlier studies often explored the functions of the lead-lag relationship on stock return correlations and market performance. For example, Tóth and Kertész (2007, 2009) explored the functions of the lead-lag relationship in forming the Epps effect, and a similar study can also be found in Huth and Abergel (2011). However, with the development of dynamic network analysis technology, it is a new trend to study the dynamic pattern of the lead-lag correlations in stock markets. For example, Xia *et al.* (2018) examined the emergence and temporal structure of lead-lag correlations in collective stock dynamics, Curme *et al.* (2019) answered how lead-lag correlations affect the intraday pattern of collective stock dynamics, and Campajola *et al.* (2020) unveiled the relation between herding and liquidity with trader lead-lag networks. Following this trend, this paper explores the dynamic pattern of lead-lag networks in the targeted Chinese stock markets, which will contribute new findings to this stream of literature.

Explanation. In addition to exploring dynamic patterns, this paper further aims to reveal the causes of collective dynamic patterns by mining the significant explanatory factors. One direction of the related work is to unveil the formation mechanism of the dynamic patterns discovered by simulation analysis (Kobayashi and Takaguchi 2018), and the other direction is to conduct empirical analysis to statistically test the significance of the potential influencing factors and trading mode (Pomponio and Abergel 2013, Huth and Abergel 2014). Our work uses empirical analysis and adopts the exponential random graph model (ERGM) to examine the influencing factors in the context of dynamic daily lead-lag networks. Note that ERGM is a powerful tool for modeling dynamic networks, especially in the stock markets, and only a few papers have adopted this model in the financial domain (Deev and Lyocsa 2020). Accordingly, this paper, on one hand, will mine and examine the influencing factors that drive the formation of the dynamic patterns discovered, compare them with existing literature and discuss the new findings; on the other hand, our work also adds a new application of the ERGM in stock markets, contributing to both relevant fields.

3. Basic model and statistical results

3.1. Daily lead-lag networks

A lead-lag network is one type of directed network, where a node represents a stock and a link represents the lead-lag relationship. We choose the daily stock return rate as the indicator to define the daily lead-lag relationship. Taking stock i and j as examples, if the return rate of stock j on day t is quite close to that of stock i on day t-1, the stock pair is deemed to possess a lead-lag relationship on the two successive days. Here, stock i is the leader and stock j is the follower, and a directed link is formed from stock i to stock j in the established lead-lag network on day t.

Let $v_{i,t}$ denote the return rate of stock i on day t, and it can be calculated by

$$v_{i,t} = \frac{p_{i,t} - p_{i,t-1}}{p_{i,t-1}}, \tag{1}$$

where $p_{i,t}$ denotes the closing price of stock i on day t. In this paper, the adopted closing price is the restoration of the rights price rather than the ex-rights price in order to eliminate the effects of ex-dividends and ex-rights. Then, given a man-made threshold variable $\varepsilon > 0$ reflecting the degree of the abovementioned closeness, the condition for forming a directed link from stock i to stock j in the lead-lag network on day t is

$$\begin{cases} (1-\varepsilon)v_{i,t-1} \leq v_{j,t} \leq (1+\varepsilon)v_{i,t-1}, & \text{when } v_{i,t-1} \geq 0; \\ (1+\varepsilon)v_{i,t-1} \leq v_{j,t} \leq (1-\varepsilon)v_{i,t-1}, & \text{when } v_{i,t-1} < 0. \end{cases} \tag{2}$$

Furthermore, let $\boldsymbol{G}_t$ denote the daily lead-lag network of day t. If stock i and stock j meet the conditions in Formula (2), it holds that $g_{ij,t} = 1$; otherwise, $g_{ij,t} = 0$. Here, $g_{ij,t}$ is the element of $\boldsymbol{G}_t$ and represents the lead-lag relationship from leader stock i to follower stock j. Accordingly, the obtained daily lead-lag network $\boldsymbol{G}_t$ can be expressed as a 0–1 matrix since all its elements are either 0 or 1. In particular, our work allows one stock to follow itself.

Note that it takes two days' stock closing prices to determine one daily lead-lag network, and thus T successive days' trading information leads to T-1 daily lead-lag networks. Then, all the formed daily lead-lag networks are combined, according to their time-stamp order, to constitute a time series of networks. The series of lead-lag networks are easy to form owing to the very little information needed so as to facilitate the analysis and its extensions. Besides, the time series of networks lay the foundation for analyzing their dynamic patterns, which is promising for providing some insightful unknown findings.

3.2. *Data set and basic statistical results*

3.2.1. Data preparation. Two famous stock markets are selected in this paper for analysis and comparison: the Shanghai Stock Exchange (SSE) and the Shenzhen Stock Exchange (SZSE) in China. The analyzed time period is from the beginning of 2015 to the end of 2019; and on each trading day in the five years, the closing prices of all the stocks traded in the two stock markets are collected from http://cndata1.csmar.com/. Once the data is obtained, the daily lead-lag networks in each market can be immediately obtained based on Equation (1) and Formula (2). Here, there are a total of 1219 trading days from the SSE or the SZSE during the targeted time period.

During data preparation process, there are two things worth noting: (1) Many new stocks listed in two markets and many delisted during the five years, and thus, the total number of stocks traded is not constant. As a result, the total number of nodes in the daily lead-lag network could be different. (2) On almost every trading day, there are some stocks that are suspended due to some reasons or rules, and we use "Null" to mark the closing prices of these suspended stocks. Accordingly, the suspended stocks are isolated nodes in the lead-lag network on the day of their suspension.

3.2.2. Basic statistical results. Based on the prepared data and the given ε, these daily lead-lag networks can be achieved according to Equation (1) and Formula (2). In order to illustrate the characteristics of the obtained daily lead-lag networks, we explore the answers to the following two questions: (a) the relationship between the daily lead-lag network's size N and number of edges M, and (b) the ratios of three types of nodes in the daily lead-lag network: the pure leader, the pure lagger, and the intermediary node that acts both the leader and the follower. The main results are displayed in Figures 2 and 3, where ε is set as 20%. Since ε is a man-made variable by recalling Formula (2), its robustness test will be conducted in the later part of this section.

The size of the lead-lag network increases on average in both stock markets over time, as displayed in Figure 2. More importantly, five years of data lead to numerous daily combinations (N, M) that allow us to fit a function between N and M. As a result, superlinearity is revealed in both stock markets but with different parameters: in the SSE and $M \propto N^{1.475}$ in the SZSE. These findings suggest that the average degree of a daily network (i.e. M/N) increases with orders $N^{0.733}$ and $N^{0.475}$ in the two s$M \propto N^{1.733}$tock markets, respectively. Accordingly, the difference between the two markets shows that as network size increases, the growth rate of the network density of the SSE is faster than that of the SZSE.

There are three types of nodes in a daily lead-lag network: the pure leader, the pure follower and the intermediary. The former two are easy to understand by their names, and the

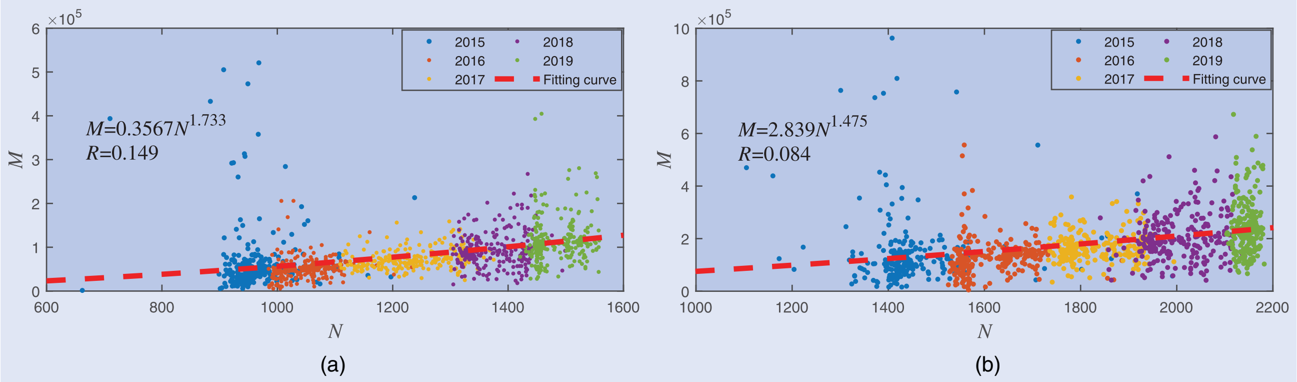

Figure 2. Relationship between network size N and edge number M in two stock markets.

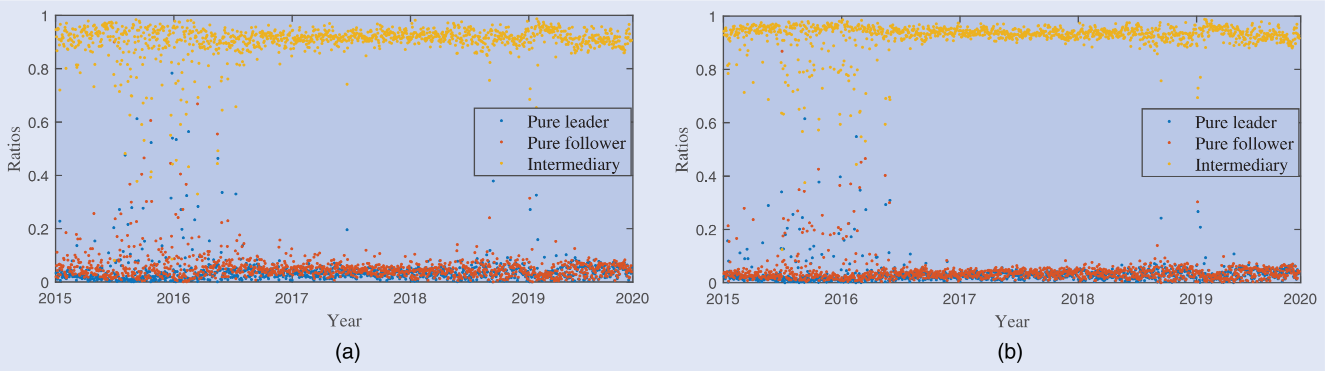

Figure 3. Ratios of pure leader, pure follower and intermediary in two stock markets.

latter acts as both a leader and a follower. The existence of three types of nodes rather than two types implies that the triangle would be a common structure in the daily lead-lag network, which suggests that we choose the network structure variable in our empirical analysis. Furthermore, the ratios of the three types of nodes are almost stable during the five years as shown in Figure 3, although fluctuations sometimes appear. Besides, the ratio of intermediaries in any stock market is much higher than the ratios of the other two types, but some differences exist between the two markets: the ratio of pure leaders in the SSE is higher than that in the SZSE, and the fluctuation in the SSE is slightly more dramatic than that in the SZSE.

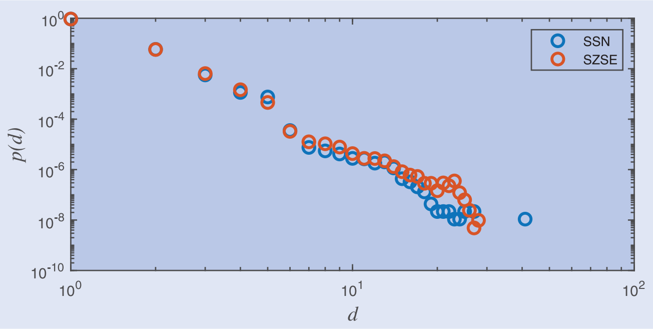

Figure 4. Distribution of edge duration (the variable d) aggregated over all pairs.

3.3. Dynamics patterns of daily lead-lag networks

Among all the potential variables reflecting the dynamics patterns, we choose the variable of *edge duration* in order to highlight the critical finding. The so-called *edge duration*, denoted as d, is defined as the number of successive trading days on which a leader stock and a follower stock form an edge in the successive daily lead-lag network. Note that the value of d of the same stock pair is quite likely not to be unique during the analyzed period because a stock pair can form the successive lead-lag relationship many times. Next, the following subsection will present the distribution of the *edge duration* and test the possible distribution assumption, and then the last subsection will discuss how the different values of ε affect the results as robustness analysis.

3.3.1. Distributions of the edge duration (variable d) in the two stock markets. Aggregating all pairs in the obtained daily lead-lag networks, the distributions of d are obtained and displayed in Figure 4, where blue circles and red circles indicate the results for the SSE and the SZSE, respectively. The two symbols show the same pattern in the graph: they both present almost a line curve in the adopted double-logarithm (i.e. log–log) coordinates, which implies that the variable d is likely to satisfy a power-law distribution. Accordingly, regression analysis is conducted to find that $p(d) \propto d^{-1.213}$ for the SSE and $p(d) \propto d^{-1.238}$ for the SZSE, showing that they are quite close in the two stock markets. Note that the exponents of the two discovered power-law distributions are both less than 2, and their generic properties are quite different from those with an exponent greater than 2 (Seyed-Allaei *et al.* 2006). In detail, the expected value of a power-law distribution with an exponent greater than 2 is finite and independent of the number of elements considered, whereas the opposite is true if the exponent is less than 2. Accordingly, the mean edge durations of the two discovered power-law distributions will increase with the network size N (or the total number of stock pairs N^2), which implies that the generic properties of the discovered power-law distributions are dependent on N or N^2.

More precisely, both the Kolmogorov–Smirnov test (K-S test, for short; Clauset *et al.* 2009) and the Kuiper test (Kuiper 1960) are conducted to test the discovered power-law distribution with the null hypothesis “the tested distribution satisfies a power-law distribution”. Then, the obtained p values are 0.442 for the SSE and 0.936 for the SZSE via the K-S test and meanwhile 0.969 for the SSE and 0.780 for the SZSE via the Kuiper test, which cannot reject the null hypothesis. To ensure the correctness of the finding, we further conduct the K-S test and the Kuiper test to check whether the assumption of an exponential distribution can be accepted, which is often done to compare the results with the power-law distribution. Then, the obtained p values of the two stock markets are 6.1E-05 and 8.0E-08 via the K-S test and 4.8E-06 and 5.3E-07 via the Kuiper test, and thus, the assumed exponential distribution is rejected. Therefore, the edge duration (or the variable d) satisfies a power-law distribution in both stock markets, suggesting that the number of pairs with quite long successive lead-lag days is not too small to be ignored. This finding inspires us to focus on the heavy tail of the power-law distribution obtained, i.e. such pairs that have a "sufficiently" long edge duration.

3.3.2. Robustness analysis. Recalling Equation (2), the man-made threshold ε will affect the edge formation in a daily lead-lag network and further influence the distribution of the edge duration. Here, the robustness analysis focuses on how the man-made threshold ε affects the distribution of the edge duration: if the distributions obtained under different values of ε differ greatly, we will conclude that the output of our model is sensitive to the man-made variable ε, or that it is not robust, and vice versa. To this end, $KD(\varepsilon_i, \varepsilon_j)$ is defined as Equation (3). By following the Kuiper statistic test (Kuiper 1960), the difference between two distributions obtained under ε_i and ε_j is measured.

$$KD(\varepsilon_i, \varepsilon_j) = \max_d(cdf(d; \varepsilon_i) - cdf(d; \varepsilon_j)) + \max_d(cdf(d; \varepsilon_j) - cdf(d; \varepsilon_i)), \tag{3}$$

where $cdf(d; \varepsilon_i)$ and $cdf(d; \varepsilon_j)$ denote the cumulative distribution functions of the edge duration d under the given thresholds ε_i and ε_j, respectively. Besides, since the measurement defined in Equation (3) is a Kuiper statistic, the Kuiper test can further be conducted to check whether the difference between two distributions is significant. Considering different combinations of ε_i and ε_j, Tables 1 and 2 report the statistic $KD(\varepsilon_i, \varepsilon_j)$ of each combination and its corresponding p value via the Kuiper test.

The results marked with "bold type" in Tables 1 and 2 mean that the two distributions cannot be deemed to be different under the 0.05 significance level. As a result, when $|\varepsilon_i - \varepsilon_j| \leq 10\%$, the difference of the distributions obtained under two threshold values is not significant, which implies that our model's output is robust if the deviation of two threshold values is not too large. Besides, unsurprisingly, $DD(\varepsilon_i, \varepsilon_j)$ increases as $|\varepsilon_i - \varepsilon_j|$ increases in all the combinations of ε_i and ε_j in the two stock markets; and even if the deviation of two threshold values is as large as 20%, the difference of two distributions under some combinations is also not significant. In summary, our model's outputs can be deemed to be robust to some extent by considering that they are not very sensitive to the man-made threshold ε. Since the subsequent work is mainly dependent on the distribution of the edge duration, the robustness test in this subsection is fundamental for the following section.

4. Detecting the lead-lag effect from the observed power-law distribution

4.1. Definition of lead-lag effect

As the discovered power-law distributions imply, a fair number of stock pairs have long successive lead-lag days (i.e. a large d) such that these pairs should not be ignored. Accordingly, our definition of the lead-lag effect concerns the "length" of successive lead-lag days; in other words, the lead-lag effect means that the successive lead-lag days are long enough so that their length significantly differs from the lengths caused by random events. Once the lead-lag effect can be detected, the core of our studied dynamic pattern can be mastered, which not only facilitates the origins of the dynamic pattern, but also benefits the prediction of stock fluctuation and risk transmission. Although the direction sounds exciting and promising, we should first answer *one important question*: how to formally define the lead-lag effect since the above words "long enough" are vague.

Before formally answering the question, let us recall the following fact: the successive lead-lag relationship between the same stock pair often occurs more than once, and therefore one stock pair is likely to have more than one value of d. We use $d_{ij,\omega}$ to express the successive lead-lag days from leader i to follower j that occurred at the ωth time. In order to highlight the longest period of lead-lag days of one stock pair and eliminate the confusion caused by its multiple successive lead-lag days, we use d_{ij} to denote the maximal value of $d_{ij,\omega}$ among all the occurrences via the following mathematical expression:

$$d_{ij} = \max_{\omega}\{d_{ij,\omega}\}. \tag{4}$$

Here, d_{ij} could be zero if no lead-lag edge exists from leader i to follower j in all daily lead-lag networks. Note that the graph

Table 1. Robustness results in SSE.

KD (p)	15%	20%	25%	30%	35%
10%	0.148 (0.973)	0.258 (0.255)	0.326 (0.039)	0.366 (0.009)	0.390 (0.004)
15%		0.110 (0.999)	0.178 (0.851)	0.218 (0.539)	0.304 (0.077)
20%			0.070 (0.999)	0.155 (0.957)	0.247 (0.319)
25%				0.084 (0.999)	0.177 (0.856)
30%					0.093 (0.999)

Table 2. Robustness results in SZSE.

KD (p)	15%	20%	25%	30%	35%
10%	0.154 (0.948)	0.259 (0.222)	0.318 (0.042)	0.349 (0.014)	0.383 (0.004)
15%		0.1045 (0.999)	0.1632 (0.914)	0.2370 (0.362)	0.3374 (0.022)
20%			0.0819 (0.999)	0.1762 (0.841)	0.2765 (0.145)
25%				0.0942 (0.999)	0.1946 (0.704)
30%					0.1004 (0.999)

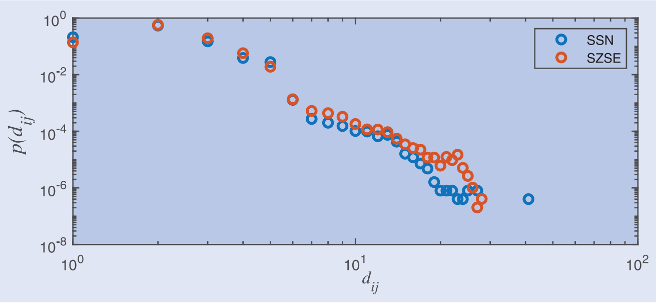

Figure 5. Distribution of the longest successive lead-lag days (i.e. d_{ij}) between all pairs.

displayed in Figure 4 is the distribution of $d_{ij,\omega}$ aggregating all pairs and all occurrences. Different from Figure 4, Figure 5 displays the two distributions of d_{ij} in two stock markets. Similar to the above statistical analysis process, the variables d_{ij}s in the two markets both satisfy a power-law distribution,† which inherits the property illustrated in Figure 4. Hereafter, when we mention the distribution of successive lead-lag days, the distribution of d_{ij} is as defined in Equation (4).

Returning to our question, its core is to find an objective criterion for judging the concept of "long enough". Without serious considerations, it seems reasonable to set a statistical significance level (e.g. 0.01) and obtain the statistically significant node pairs from the distributions shown in Figure 5. However, the above approach does not correctly understand the meanings of the "lead-lag effect" because whether one effect holds or not requires statistically testing whether one event can be deemed to be a rare event in a random trial. In other words, the lead-lag effect should be defined by comparing it with a random event: if the successive lead-lag days of one pair can be judged as a rare event based on a random trial, the corresponding pair is defined to hold the lead-lag effect.

In order to make the abovementioned comparison fair, the generated daily random network should keep the same degree sequence with its same day's real lead-lag network. Hereafter, the real lead-lag network is named the "referential network" by considering that its link density and degree distribution should be inherited from its same day's random network. Then, the distribution of the successive lead-lag days can be determined from the daily random networks, similar to what we have done based on the daily real lead-lag networks. Given a statistical significance level, it is not difficult to judge whether one successive lead-lag day is long enough to be a rare event based on the achieved distribution. Accordingly, the definition of the lead-lag effect is formally provided in Definition 1.

Definition 1 Lead-lag effect *Given a significance level Δ of the statistical test, the criterion d_Δ can be determined from the obtained distribution of the successive lead-lag days. For any stock pair (i, j), the lead-lag effect from i to j holds if and only if $d_{ij} \geq d_\Delta$, and the detected pairs are called the lead-lag pairs.* □

4.2. Detection method and an example

4.2.1. Detection method. According to Definition 1, the main task of the detection method is to generate the daily random network preserving the full degree sequence of its referential network. Fortunately, the configuration model (Newman *et al.* 2001) adapts for solving this task and its codes can be directly achieved from https://networkx.org/documentation/stable/_modules/networkx/generators/degree_seq.html#directed_configuration_model. Note that the adopted configuration model can only obtain approximate solutions in our context due to deleting the duplicate edges; thus, a larger network size implies a more accurate solution by considering that the likelihood of duplicate edges appearing decreases as the network size increases. In other words, the random network achieved via the configuration model can only have an approximately identical link density and degree distribution with its referential network; however, the network sizes of the two analyzed stock markets are large enough to guarantee that the difference is within limits of acceptability.

Next, one group of simulations will generate each day's random network that constitutes a time series set denoted as $\{G_1^R, G_2^R, \cdots, G_t^R, \cdots, G_T^R\}$, where T denotes the total number of trading days and equals 1219 in this paper (recalling Section 3.2.1) and the superscript "R" highlights the random generation distinguished with the real ones. Then, the distribution of d_{ij} can be achieved from $\{G_1^R, G_2^R, \cdots, G_t^R, \cdots, G_T^R\}$, similar to what has done in Figure 5. In order to make the results sound, hundreds of groups of simulations are conducted to obtain hundreds of the abovementioned distributions. Next, one comprehensive distribution of d_{ij} can be obtained by averaging all the achieved distributions. In the last step, given a significance level Δ, the threshold value d_Δ can be achieved from the comprehensive distribution as the judgment criterion. Until now, all the variables in Definition 1 have been determined, and the pairs with the lead-lag effect can be detected.

4.2.2. An explanatory example. An example is provided here in order to explain and reveal the above proposed detection method. This example only analyzes 10 stocks as well as their closing prices on 6 successive trading days for concision. Based on the input, five daily lead-lag networks are obtained as displayed in Figure 6. In this graph, each node represents one stock, the node size reflects the node degree and the directed link points from the leader to the follower. Besides, the arc represents the loop showing that the case of following oneself is allowed.

According to the proposed detection method, the abovementioned configuration model is adopted to generate the daily random networks by taking the daily networks displayed in Figure 6 as referential networks. Figure 7 displays the result via one group of simulation, where each day's random network preserves the same degree sequence of its corresponding referential network.

Then, 500 groups of simulations will generate 500 groups of random lead-lag networks that immediately lead to the distribution of the successive lead-lag days. When the statistical significance level Δ is set as 0.01, the criterion d_Δ equals 4 based on the obtained distribution in Figure 8. Here,

† The result of the statistical tests for satisfying the power-law distribution in both markets is 0.093 and 0.576 (p value), respectively.

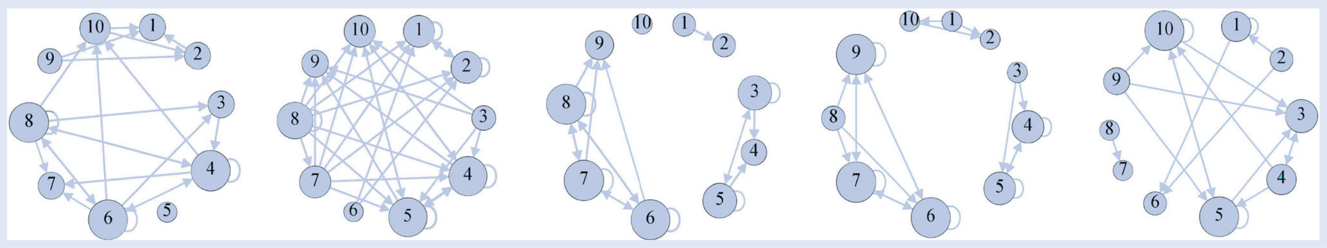

Figure 6. Graphs of five successive lead-lag networks given in this example.

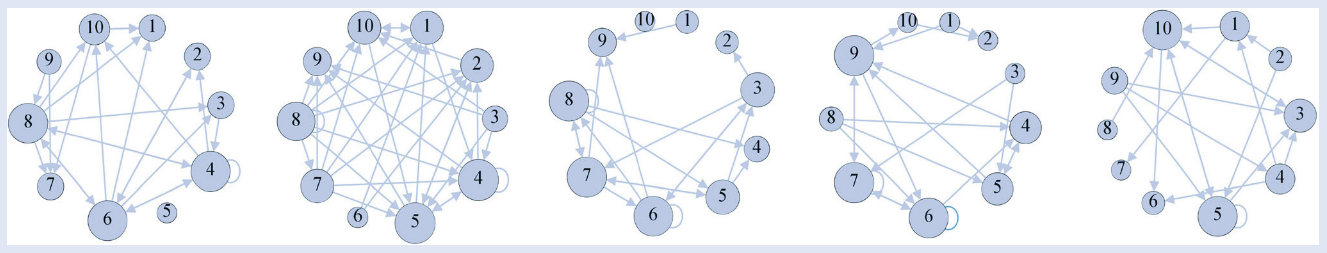

Figure 7. Graghs of five successive random lead-lag networks via one group of simulation.

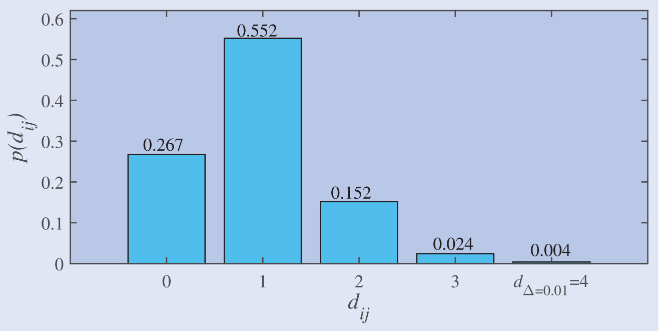

Figure 8. Distribution of the successive lead-lag days via 500 groups of simulations.

$d_{\Delta=0.01} = 4$ in our example means that, if and only if a directed link between the same pair appears in at least four successive real lead-lag networks shown in Figure 6, the pairs connected by the link can be judged to possess a lead-lag effect. As a result, the detected leader-follower pairs are $3 \to 4$, $5 \to 5$ and $8 \to 7$.

Table 3. Network-based statistical indexes in SSE.

Δ	0.10	0.01	0.001
Node number	1574	1565	1389
Edge number	545574	171870	75080
Density	0.220	0.070	0.039
Diameter	1.692	2.258	3.123
Clustering coefficient	0.526	0.373	0.455
Ratio of triad	0.206	0.047	0.020
Ratio of reciprocity	0.526	0.652	0.483

Table 4. Network-based statistical indexes in SZSE.

Δ	0.10	0.01	0.001
Node number	2210	2209	2121
Edge number	1340975	391655	109948
Density	0.275	0.080	0.024
Diameter	1.603	2.039	2.868
Clustering coefficient	0.580	0.402	0.384
Ratio of triad	0.298	0.049	0.007
Ratio of reciprocity	0.550	0.551	0.344

4.3. Results based on the detection method

Based on the data sets described above and ε in Equation (2) being set as 20%, the distributions of the real d_{ij}s and simulated d_{ij}s in the SSE (Shanghai Stock Exchange) are obtained via the above proposed detection method and combined in Figure 9, where different significance levels Δs cause different criterions d_Δs.

It is not hard to understand that a smaller significance level will result in a larger criterion. To illustrate this point more clearly, Figure 9 displays the directed networks formed by these detected lead-lag pairs under different significance levels. Essentially, the detected lead-lag pairs can be understood as the origin of the heavy tail contained in the observed power-law distribution displayed in Figure 5. Note that the networks formed by these detected lead-lag pairs are different from the daily real lead-lag networks or the daily random lead-lag networks. In order to distinguish them, we call the network formed by the detected lead-lag pairs by the name "heavy-tailed network". Intuitively, by comparing the three displayed heavy-tailed networks, we find that their total numbers of nodes are quite similar although they are formed under different significance levels, whereas the number of edges decreases when Δ is smaller. Table 3 illustrates the above findings and reports the other statistical indexes of the three heavy-tailed networks obtained in the SSE. Similar results can also be found in the SZSE (Shenzhen Stock Exchange), which are shown in Figure 10 and Table 4.

In addition to the above findings, we also find that the distribution of d_{ij} from real lead-lag networks is quite different from that from the simulated random lead-lag networks, which implies that the dynamics of lead-lag networks cannot be formed purely from the random mechanism. The distributions of d_{ij} from real lead-lag networks both satisfy

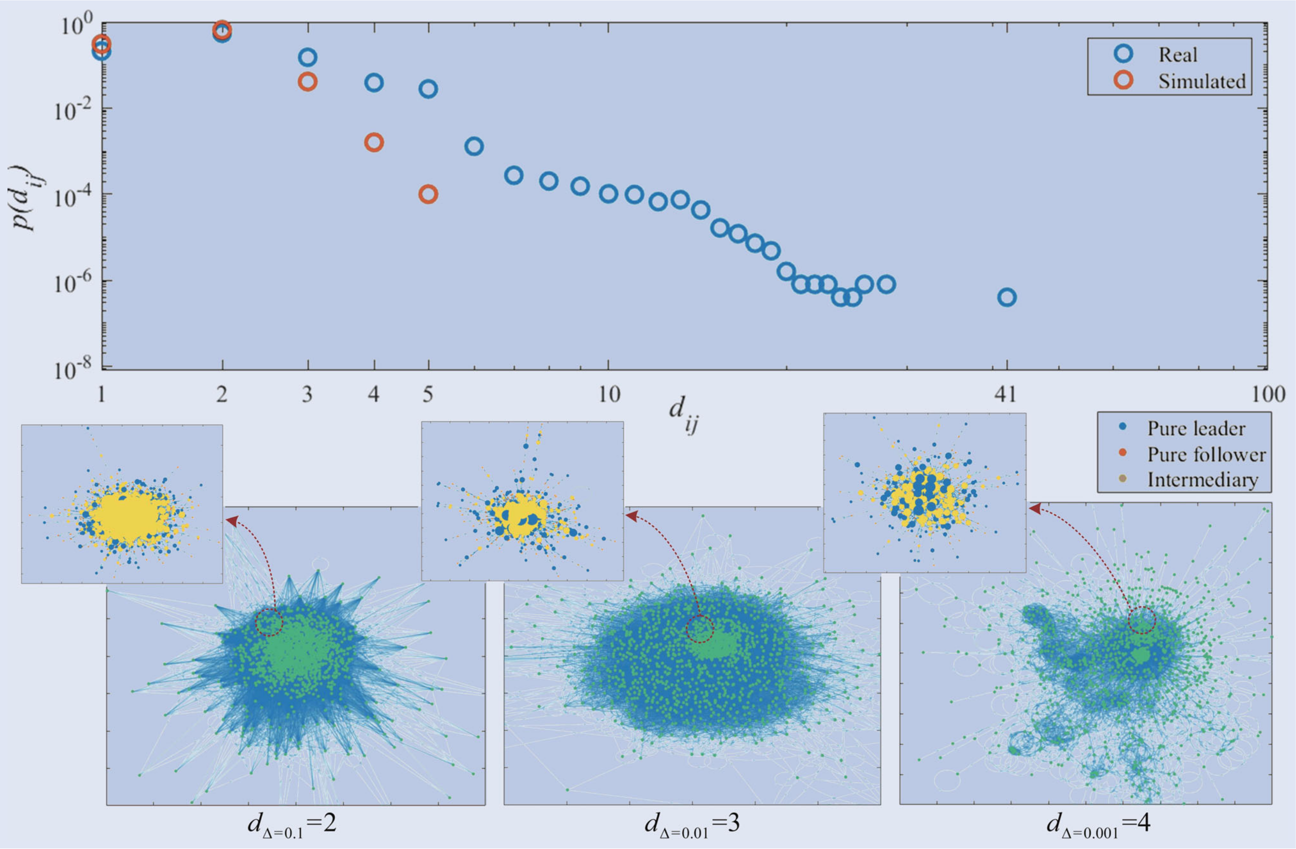

Figure 9. The detection results in SSE.

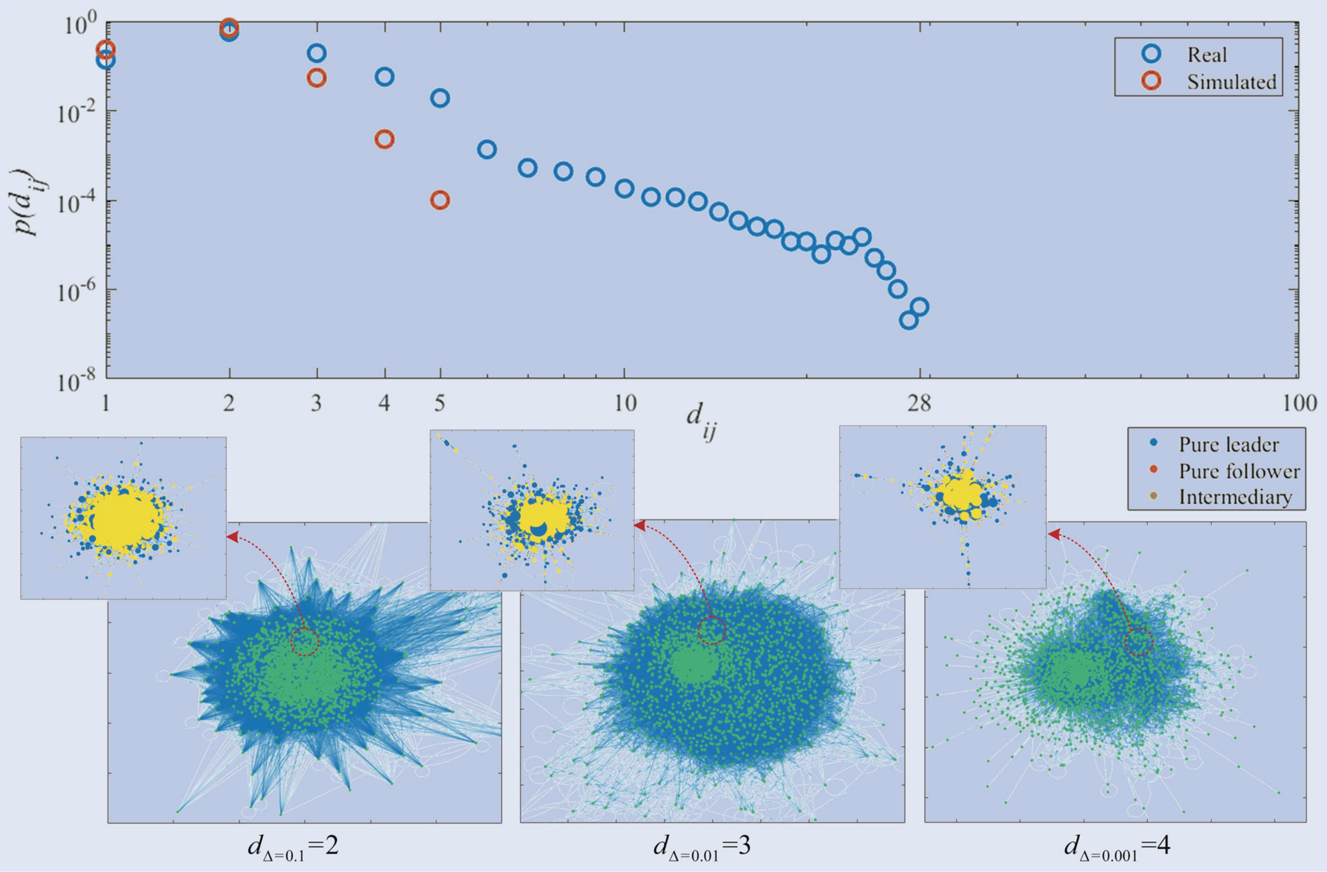

Figure 10. The detection results in SZSE.

a power-law distribution in the two targeted stock markets whereas those from the simulated random lead-lag networks both satisfy the exponential distribution.† Our work focuses on their distinctions and detects the above-defined heavy-tailed networks to highlight the features of the observed social dynamics of real lead-lag networks. Recalling Figures 9 and 10 again, even if the significance level Δ is as small as 0.001, there are also a large number of stock pairs meeting the lead-lag effect and being detected in both stock markets. Besides, the displayed heavy-tailed networks and their local enlarged drawings show that intermediary nodes have larger degrees than the other two types of nodes and become the cores of the formed networks, noting that the node size is proportional to its degree in our graphs. This finding essentially reveals that the transmission structure widely exists in the formed heavy-tailed networks and that many transmission paths share the same middle nodes. This result is helpful not only to design an empirical model to analyze the influencing factors that affect the formation of the heavy-tailed networks obtained, but also to understand the price fluctuation and risk transmission patterns in the targeted stock markets.

As the adopted significance level Δ decreases, some network-based indexes change very little, but others change substantially. In both stock markets, the node number, the clustering coefficient and the ratio of reciprocity are not sensitive to the change in Δ whereas the edge number, the network density and the ratio of the triad sharply decrease and the network diameter significantly rises as Δ decreases. In other words, there was a significant change in network global indexes (such as the edge number, the ratio of triad, the network density and diameter) and a tiny change in network local indexes (such as the clustering coefficient). Besides, it is an interesting finding that the total node number and the ratio of reciprocity are almost unchanged under different significance levels. All the above findings hint at the structural pattern of the obtained heavy-tailed networks to some extent, which benefits the design of the empirical analysis.

5. Empirical analysis: exploring the driving factors

As we have explained, the heavy-tailed network, consisting of lead-lag pairs and their detected lead-lag links, act as the core of the observed dynamics pattern of lead-lag networks. This section will empirically demonstrate which driving factors can significantly affect the formation of heavy-tailed networks and then reveal the origin of the observed power-law distribution.

5.1. *Empirical model*

When assessing stocks or stock markets, we can usually think of the following variables: earnings per share (EPS), turnover rate (TR), market value (MV), region (REG) and industry (IND).‡ In this section, whether these variables affect the formation of heavy-tailed networks will be examined. In fact, some studies have explored the influence of one or several variables mentioned above on the formation of stock networks in stock network research. For example, Chordia and Swaminathan (2000) found that the trading volume is a significant determinant of the lead-lag patterns observed in stock returns, where the so-called trading volume is an index similar to the turnover rate (TR) reflecting the degree of trading activity. In another example, Kinnunen (2017) demonstrated that firm size is a significant factor influencing the formation of a lead-lag relationship. Besides, Scherbina and Schlusche (2018) indicated that industry and size are not good predicators to identify leader-follower pairs. However, we find that the conclusions of existing studies vary with the different research objects and data; in addition, different from the definition of stock networks in previous studies, this paper focuses on the analysis of the heavy-tailed part of a power-law distribution rather than the whole network. It is worth noting that a power-law distribution is the core feature of the social dynamics of the lead-lag network studied here. Due to the particularity of the network analyzed in this article, the object of our study is different from those mentioned above; however, it has a certain degree of inheritance regarding selecting the explanatory variables of interest.

Furthermore, in view of the purpose of this paper, i.e. to conduct dynamic modeling analysis on the formation of network edges, the exponential random graph model (ERGM) is chosen to explain the formation and disconnection of network edges. The model covers not only the attitudes of individual stocks as the exogenous variables, but also the network structures as the endogenous explanatory variables. Considering the network structure effects in the model is the characteristic that distinguishes the ERGM from the widely adopted logistic regression model. Accordingly, the ERGM pays more attention to the overall perspective of dynamic network formation rather than the micro perspective of edge changes. Analyzing network formation with network structure variables as control variables helps to identify the role of exogenous explanatory variables more accurately. Specifically, similar to many literatures on the applications of the ERGM (Windzio 2018, Krichene *et al.* 2019), the following commonly used network structure variables are selected: reciprocity (REC), 2-path (PA), transitional triads (TT) and cyclic triads (CT). The diagrams of the four network structure variables are shown in Figure 11.

Based on the above explanation of exogenous explanatory variables and network structure variables, the specific ERGM used for empirical analysis in this work is established as follows:

$$p_\theta(\boldsymbol{G}) = \frac{1}{\kappa(\theta)} exp(\theta_1 L(\boldsymbol{G}) + \theta_2 REC(\boldsymbol{G}) + \theta_3 PA(\boldsymbol{G}) + \theta_4 TT(\boldsymbol{G}) + \theta_5 CC(\boldsymbol{G}) + \theta_6 \sum_{i<j} IS(IND_i - IND_j)$$

† Given the null hypothesis that the variables follow an exponential distribution, the results of the statistical tests are 0.991 and 0.989 (p value), respectively, thus not rejecting the null hypothesis in both markets.

‡ Since there are many standards for the division of stocks by industry, the standard chosen in this paper comes from the China Securities Regulatory Commission (CSRC).

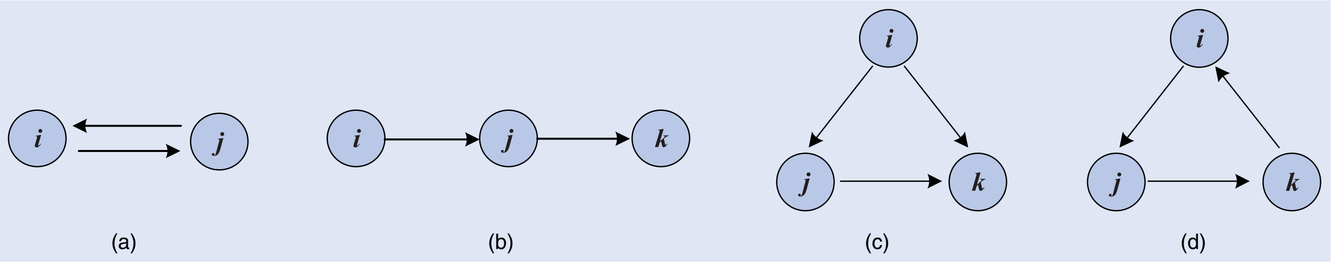

Figure 11. The selected variables of network structures.

$$
\begin{aligned}
&+\theta_7 \sum_{i<j} IS(REG_i - REG_j) + \theta_8 \sum_{i<j} |EPS_i - EPS_j| \\
&+\theta_9 \sum_i \sum_j g_{ij} EPS_j + \theta_{10} \sum_j \sum_i g_{ij} EPS_i \\
&+\theta_{11} \sum_{i<j} |TR_i - TR_j| + \theta_{12} \sum_i \sum_j g_{ij} TR_j \\
&+\theta_{13} \sum_j \sum_i g_{ij} TR_i + \theta_{14} \sum_{i<j} |MV_i - MV_j| \\
&+\theta_{15} \sum_i \sum_j g_{ij} MV_j + \theta_{16} \sum_j \sum_i g_{ij} MV_i). \qquad (5)
\end{aligned}
$$

Here, $L(\boldsymbol{G})$ represents the total number of links contained in G and is similar to the constant item in the classical linear regression; and $REC(\boldsymbol{G})$, $PA(\boldsymbol{G})$, $TT(\boldsymbol{G})$ and $CC(\boldsymbol{G})$ denote the four network structure variables displayed in Figure 11. Therefore, the first row of Equation (5) models the network structure variables. Next, $IS(x)$ is a sign function that is defined as $IS(x = 0) = 1$ and $IS(x \neq 0) = 0$, and thus, the second row of Equation (5) tests whether the link between a pair of stocks is more likely to form when the pair shares the same industry or the same region. Then, the last three rows of Equation (5) model the effects of the three concerned explanatory variables, i.e. earnings per share (EPS), turnover rate (TR) and market value (MV), respectively. Taking the third row as an example, the three items in this row, one by one, aim to demonstrate how the difference in the EPSs between stock pairs affects the formation of links and how the followers' and the leaders' EPSs affect the formations of the directed links in the observed network G. Meanwhile, the remaining two rows can also be understood by following this example. Note that $\kappa(\theta)$ is adopted for normalization and is not important for our analysis. Fortunately, the package "statnet" in the R software can analyze and fit the established ERGM, and then all the results reported in the following section are obtained via the R software.

5.2. Results and discussions

The data obtained for the empirical analysis come from http://cndata1.csmar.com/ for the trading data and https://www.wind.com.cn/ for the individual stock attributes. Recalling that the lead-lag networks are analyzed year by year in Section 3.2.1, the empirical study remains consistent and conducts year-by-year analysis. Since some of the individual stock variables mentioned above are not collected on an annual basis, some calculations and explanations are necessary: annual earnings per share (EPS) can be collected directly, annual turnover rate (TR) is the average of the daily turnover rates for all trading days of the targeted year, and the annual market value (MV) is calculated similar to the annual TR. Besides, the industry and the region to which one stock belongs do not frequently change. If there are some changes, they will be updated in the corresponding year.

By fitting the model shown in Equation (5), the empirical results of the two stock exchange markets from 2015 to 2019 are obtained and listed in Tables 5 and 6, respectively. Comparing Tables 5 and 6, the results of the network structure effects in the two markets and in different years are relatively consistent. Specifically, the edge is similar to the constant term that appears in the classical linear regression, reflecting the level of the link density. Through a year-on-year comparison, it can be found that the coefficient of edge in the two markets is larger in 2015, indicating that the link density of the targeted heavy-tail in this year is higher than those in subsequent years.

The regression results on the network structure effects are listed as below: (1) The coefficients of reciprocity are all significantly positive in different years and in different markets, indicating that the bidirectional links formed between stock pairs are significantly greater than those in random networks. This finding means that the underlying trading behavior of investors makes the leader and the follower contained in one stock pair transpose during one year's time. (2) The coefficient of 2-path is either significantly negative or equals zero, which means that the number of 2-path is either smaller than that of a random network or not different from it. (3) The coefficient of transitive triad is significantly positive, implying that the transitive cluster formed by investor behaviors is one of the driving factors for the formation of the lead-lag effect. (4) Unlike the transitive triad, the coefficient of the cyclic triad is significantly negative, meaning that the cyclic cluster is not the driving factor that leads to the link formation with the defined lead-lag effect. In essence, the network structure variables reflect the investors' potential behaviors difficult to observe by partly uncovering the investors' potential attention process and the information transmission process implied in the network structures.

Furthermore, Tables 5 and 6 also contain the individual stock attitudes. Different from network structure variables, the coefficients of these variables are not completely consistent in different markets and in different years. However, it is worth mentioning that the coefficient of turnover rate (TR) shows a consistent result. Specifically, the coefficients of the variable that reflect the heterogeneity of TR (i.e. $|TR_i - TR_j|$)

Table 5. Empirical results in the market of SSE from 2015 to 2019.

	Estimates (Standard errors)				
Parameters	2015	2016	2017	2018	2019
Edge	– 4.43 (0.02)***	– 6.92 (0.07)***	– 6.75 (0.04)***	– 6.69 (0.05)***	– 6.35 (0.04)***
Network structural effects					
Reciprocity	1.40 (0.04)***	4.69 (0.13)***	4.21 (0.11)***	3.25 (0.15)***	2.88 (0.14)***
2-path	– 0.03 (0.01)***	– 0.05 (0.01)***	– 0.04 (0.00)***	0.00 (0.01)	0.02 (0.003)***
Transitive triads	0.01 (0.00)***	0.72 (0.01)***	0.49 (0.01)***	1.21 (0.03)***	0.66 (0.01)***
Cyclic triplets	– 0.02 (0.00)***	– 1.06 (0.02)***	– 0.61 (0.02)***	– 1.17 (0.06)***	– 0.89 (0.02)***
Stock-relations effects					
Homophily – REG	0.05(0.05)	0.19 (0.09)**	0.13 (0.08)*	– 0.07 (0.08)	– 0.01 (0.07)
Homophily – IND	– 0.14 (0.04)***	0.07 (0.07)	0.03 (0.06)	0.06 (0.05)	– 0.02 (0.04)
Heterophily – EPS	– 0.14 (0.02)***	– 0.03 (0.04)	– 1.00 (0.92)	0.40 (1.06)	– 0.02 (0.01)**
Receiver – EPS	– 0.23 (0.02)***	0.01 (0.04)	1.85 (0.84)**	– 1.49 (1.04)	– 0.10 (0.02)***
Sender – EPS	0.20 (0.01)***	0.04 (0.04)	1.12 (0.88)	– 0.63 (1.01)	– 0.07 (0.02)***
Heterophily – TR	– 0.50 (0.05)***	– 2.50 (0.20)***	– 2.86 (0.18)***	– 2.39 (0.17)***	– 1.38 (0.16)***
Receiver – TR	0.91 (0.06)***	2.18 (0.18)***	1.59 (0.14)***	1.30 (0.15)***	0.78 (0.15)***
Sender – TR	1.11 (0.03)***	2.55 (0.18)***	2.19 (0.16)***	2.17 (0.16)***	0.92 (0.13)***
Heterophily – MV	– 4.05 (0.14)***	10.60 (3.38)	3.30 (3.04)	– 2.13 (1.01)**	– 0.62 (1.31)
Receiver – MV	3.20 (0.15)***	– 11.79 (3.41)	– 2.79 (2.98)	2.10 (0.97)**	0.82 (1.27)
Sender – MV	2.01 (0.12)***	– 11.71 (3.46)	– 3.05 (2.98)	2.01 (0.97)**	1.03 (1.28)

Note. *, ** and *** denotes significant in the level of 0.1, 0.05 and 0.01, respectively.

Table 6. Empirical results in the market of SZSE from 2015 to 2019.

	Estimates (Standard errors)				
Parameters	2015	2016	2017	2018	2019
Edge	– 5.24 (0.02)***	– 7.12 (0.06)***	– 6.80 (0.03)***	– 6.66 (0.04)***	– 6.36 (0.03)***
Network structural effects					
Reciprocity	1.75 (0.42)***	4.21 (0.11)***	4.32 (0.08)***	2.68 (0.11)***	3.13 (0.09)***
2-path	– 0.02 (0.00)***	– 0.05 (0.00)***	– 0.03 (0.00)***	0.00 (0.00)	0.00 (0.00)
Transitive triads	0.01 (0.00)***	0.45 (0.01)***	0.43 (0.00)***	1.06 (0.02)***	0.64 (0.01)***
Cyclic triplets	– 0.01 (0.00)***	– 0.46 (0.02)***	– 0.57 (0.01)***	– 0.96 (0.07)***	– 0.83 (0.02)***
Stock-relations effects					
Homophily – REG	0.02 (0.05)	– 0.03 (0.06)	– 0.04 (0.05)	– 0.04 (0.04)	– 0.03 (0.04)
Homophily – IND	– 0.01 (0.00)***	– 0.01 (0.05)	0.01 (0.04)	0.01 (0.03)	– 0.05 (0.03)
Heterophily – EPS	– 0.07 (0.02)***	– 0.08 (0.05)*	– 0.08 (0.03)***	0.01 (0.01)	– 0.02 (0.01)**
Receiver – EPS	– 0.02 (0.01)**	– 0.10 (0.05)**	0.08 (0.03)***	– 0.01 (0.02)	– 0.01 (0.01)
Sender – EPS	0.09 (0.04)**	0.03 (0.05)	0.05(0.03)*	– 0.01 (0.12)	0.00 (0.01)
Heterophily – TR	– 0.43 (0.03)***	– 1.62 (0.17)***	– 2.57 (0.14)***	– 1.86 (0.12)***	– 1.59 (0.11)***
Receiver – TR	1.32 (0.11)***	2.04 (0.19)***	1.71 (0.12)***	0.60 (0.12)***	1.19 (0.12)***
Sender – TR	0.81 (0.02)***	2.56 (0.17)***	2.45 (0.13)***	2.20 (0.10)***	1.78 (0.11)***
Heterophily – MV	– 1.30 (0.35)***	– 1.22 (0.61)**	– 1.09 (0.67)*	0.53 (0.63)	– 0.02 (0.93)
Receiver – MV	1.45 (0.53)***	1.27 (0.58)**	1.30 (0.61)***	– 0.88 (0.64)	– 1.24 (0.96)
Sender – MV	0.87 (0.26)***	0.88 (0.56)*	0.99 (0.67)	0.18 (0.62)	0.74 (0.88)

are significantly negative, demonstrating that a greater difference in the turnover rate between two stocks will lead to a lower likelihood of forming a lead-lag link between them. Furthermore, both coefficients are significantly positive from both the sender's perspective and the receiver's perspective, and thus, a stock with a high TR is more likely to establish a lead-lag link. By further comparing their coefficients (i.e. sender-TR and receiver-TR), the likelihood of acting as a leader is higher than that of acting as a follower. On the other hand, industries and regions are not significant in most years in the two stock markets, and thus, industries and regions do not significantly affect the formation of lead-lag links in most years. Furthermore, earnings per share (EPS) and market value (MV) do not follow the same pattern in different markets and in different years, adding new findings to the related literature.

6. Conclusions and future work

This paper focuses on studying the successive lead-lag phenomenon of stock price movements and the driving factors causing this observed phenomenon. First, the distribution of the successive following days among all stock pairs is found and validated to meet a power-law distribution in the two selected stock markets-the Shanghai Stock Exchange and the Shenzhen Stock Exchange. Note that the heavy tail is at the heart of a power-law distribution and it acts as the key to analyzing and identifying the social dynamics of the lead-lag networks. Second, a rigorous definition is proposed to identify the lead-lag effect according to the principle of statistical test, and it is further adopted to detect the heavy tail of the observed power-law distributions. Third, a series of empirical analyses are conducted based on the ERGM and they help explain the origins of the observed power-law distribution. In the empirical analysis, the above detected heavy-tailed network is regarded as the dependent variable, and the network structure effects and the individual stock attributes are regarded as the independent variables. Furthermore, the empirical analysis indicates that among the four network structure variables (i.e. reciprocity (REC), 2-path (PA), transitive triads (TT) and cyclic triads (CT)), REC, TT and CT significantly and consistently affect the formation of heavy-tailed networks in the two markets throughout the selected years, reflecting some inherent behavioral characteristics of investors. Besides, in terms of individual stock attributes, the influence of the turnover rate is consistent and significant in the two markets throughout the selected years, the industry and the region have no significant influence on the formation of the lead-lag effect, and the influences of the other concerned individual stock attributes are inconsistent in different markets and in different years.

We recommend that future work validates the model and empirical findings in more stock markets. Moreover, the observed power-law distribution can also be useful for designing investment strategies in stock markets, because the price fluctuation of the leader can effectively predict that of the follower in the next day. Finally, as a preliminary study, our empirical study only validates some of the potential stock attributes and several network structure effects, so in-depth and rigorous empirical studies are still needed.

7. Open Scholarship

This article has earned the Center for Open Science badges for Open Data and Open Materials through Open Practices Disclosure. The data and materials are openly accessible at https://github.com/liuchaos03/ssz-szse-stock-data and https://github.com/liuchaos03/Lead-lag-Networks-in-Stock-Markets.

8. Availability of data and material

Data available at https://github.com/liuchaos03/ssz-szse-stock-data, and Codes available at https://github.com/liuchaos03/Lead-lag-Networks-in-Stock-Markets.

Acknowledgements

We greatly thank the helpful discussions from Dr. Minheng Ni, Dr. Yiru Jiao, Dr. Ningyuan Fan and Ms. Xi Chen.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

This work was supported by National Natural Science Foundation of China under [grant number 71771041] and Natural Science Foundation of Heilongjiang Province, China [grant number YQ2020G003].]

ORCID

Yongli Li http://orcid.org/0000-0002-1979-9057

References

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

Barabasi, A., The origin of bursts and heavy tails in human dynamics. *Nature*, 2005, **435**(7039), 207–211.

Barabasi, A. and Albert, R., Emergence of scaling in random networks. *Science*, 1999, **286**(5439), 509–512.

Basnarkov, L., Stojkoski, V., Utkovski, Z. and Kocarev, L., Lead–lag relationships in foreign exchange markets. *Physica A*, 2020, **539**, 122986.

Berndsen, R. J., Leon, C. and Renneboog, L., Financial stability in networks of financial institutions and market infrastructures. *J. Financ. Stabil.*, 2016, **35**, 120–135.

Boginski, V., Butenko, S. and Pardalos, P. M., Mining market data: A network approach. *Comput. Oper. Res.*, 2006, **33**(11), 3171–3184.

Buccheri, G., Corsi, F. and Peluso, S., High-frequency lead-lag effects and cross-asset linkages: A multi-asset lagged adjustment model. *J. Bus. Econ. Stat.*, 2019. doi:10.1080/07350015.2019.1697699.

Cai, C. X., Mobarek, A. and Zhang, Q., International stock market leadership and its determinants. *J. Financ. Stabil.*, 2017, **33**, 150–162.

Campajola, C., Lillo, F. and Tantari, D., Unveiling the relation between herding and liquidity with trader lead-lag networks. *Quant. Finance*, 2020, **20**(11), 1765–1778.

Challet, D., Chicheportiche, R., Lallouache, M. and Kassibrakis, S., Statistically validated lead-lag networks and inventory prediction in the foreign exchange market. *Adv. Complex Syst.*, 2018, **21**(08), 1850019.

Chordia, T. and Swaminathan, B., Trading volume and cross-autocorrelations in stock returns. *J. Financ.*, 2000, **55**(2), 913–935.

Clauset, A., Shalizi, C. R. and Newman, M. E. J., Power-law distributions in empirical data. *SIAM Rev.*, 2009, **51**(4), 661–703.

Curme, C., Tumminello, M., Mantegna, R. N., Stanley, H. E. and Kenett, D. Y., Emergence of statistically validated financial intraday lead-lag relationships. *Quant. Finance*, 2015, **15**(8), 1375–1386.

Curme, C., Tumminello, M., Mantegna, R. N., Stanley, H. E. and Kenett, D. Y. How Lead-Lag correlations affect the intraday pattern of collective stock dynamics. Available at SSRN 2648490, 2019.

Fonseca, D. J. and Zaatour, R., Correlation and lead–lag relationships in a hawkes microstructure model. *J. Futures Markets*, 2017, **37**(3), 260–285.

Dao, T. M., Mcgroarty, F. and Urquhart, A., Ultra-high-frequency lead–lag relationship and information arrival. *Quant. Finance*, 2018, **18**(5), 725–735.

Deev, O. and Lyocsa, S., Connectedness of financial institutions in Europe: A network approach across quantiles. *Physica A*, 2020. doi:10.1016/j.physa.2019.124035.

Fiedor, P., Information-theoretic approach to lead-lag effect on financial markets. *Eur. Phys. J. B*, 2014, **87**(8), 168.

Gong, C. C., Ji, S. D., Su, L. L., Li, S. P. and Ren, F., The lead–lag relationship between stock index and stock index futures: A thermal optimal path method. *Physica A*, 2016, **444**, 63–72.

Harris, J. K., *An Introduction to Exponential Random Graph Modeling (Vol. 173)*, 2013 (Sage Publications: London).

Hayashi, T. and Koike, Y., Wavelet-based methods for high-frequency lead-lag analysis. *SIAM J. Financ. Math.*, 2018, **9**(4), 1208–1248.

Huth, N. and Abergel, F., High frequency correlation modelling. In *Econophysics of Order-Driven Markets*, edited by F. Abergel, B.K. Chakrabarti, A. Chakraborti and M. Mitra, pp. 189–202, 2011 (Springer: Milano).

Huth, N. and Abergel, F., High frequency lead/lag relationships—empirical facts. *J. Empir. Financ.*, 2014, **26**, 41–58.

Ito, K. and Sakemoto, R., Direct estimation of lead–lag relationships using multinomial dynamic time warping. *Asia Pac. Financ. Mkts*, 2020, **27**(3), 325–342.

Jong, D. F. and Nijman, T., High frequency analysis of lead-lag relationships between financial markets. *J. Empir. Financ.*, 1997, **4**(2-3), 259–277.

Kinnunen, J., Dynamic cross-autocorrelation in stock returns. *J. Empir. Financ.*, 2017, **40**, 162–173.

Kobayashi, T. and Takaguchi, T., Social dynamics of financial networks. *EPJ Data Sci.*, 2018, **7**(1), 15.

Krichene, H., Fujiwara, Y., Chakraborty, A., Arata, Y., Inoue, H. and Terai, M., The emergence of properties of the Japanese production network: How do listed firms choose their partners? *Soc. Networks*, 2019, **59**(10), 1–9.

Kuiper, H. N., Tests concerning random points on a circle. *Indagat. Math.*, 1960, **63**, 38–47.

Li, Y., Liu, G. and Pin, P., Network-based risk measurements for interbank systems. *Plos One*, 2018, **13**(7), 200–209.

Liu, Z., Hu, H., Liu, Y., Ross, K. W., Wang, Y. and Mobius, M. P2P trading in social networks: The value of staying connected. In *2010 Proceedings IEEE INFOCOM* (pp. 1-9). IEEE, 2010, March.

Lusher, D., Koskinen, J. and Robins, G. (Eds.), *Exponential Random Graph Models for Social Networks: Theory, Methods, and Applications*, 2013 (Cambridge University Press: Cambridge).

Newman, M. E., Strogatz, S. H. and Watts, D. J., Random graphs with arbitrary degree distributions and their applications. *Phys. Rev. E*, 2001, **64**(2), 026118.

O'Neill, M. and Rajaguru, G., A response surface analysis of critical values for the lead-lag ratio with application to high frequency and non-synchronous financial data. *Account. Financ.*, 2019, **60**(4), 3979–3990.

Pomponio, F. and Abergel, F., Multiple-limit trades: Empirical facts and application to lead–lag measures. *Quant. Finance*, 2013, **13**(5), 783–793.

Ryan, A., Emergence is coupled to scope, not level. *Complexity*, 2007, **13**(2), 67–77.

Scherbina, A. and Schlusche, B., Follow the leader: Using the stock market to uncover information flows between firms. *Rev. Financ.*, 2018, **24**(1), 189–225.

Seyed-Allaei, H., Bianconi, G. and Marsili, M., Scale-free networks with an exponent less than two. *Phys. Rev. E*, 2006, **73**(4), 046113.

Stübinger, J., Statistical arbitrage with optimal causal paths on high-frequency data of the S&P 500. *Quant. Finance*, 2018, **19**, 921–935.

Tolikas, K., The lead-lag relation between the stock and the bond markets. *Eur. J. Financ.*, 2018, **24**(10), 849–866.

Tóth, B. and Kertész, J., Increasing market efficiency: Evolution of cross-correlations of stock returns. *Physica A*, 2006, **360**(2), 505–515.

Tóth, B. and Kertész, J., On the origin of the Epps effect. *Physica A*, 2007, **383**(1), 54–58.

Tóth, B. and Kertész, J., The Epps effect revisited. *Quant. Finance*, 2009, **9**(7), 793–802.

Tse, C. K., Liu, J. and Lau, F. C., A network perspective of the stock market. *J. Empir. Financ.*, 2010, **17**(4), 659–667.

Výrost, T., Lyocsa, S. and Baumohl, E., Granger causality stock market networks: Temporal proximity and preferential attachment. *Physica A*, 2015, **427**, 262–276.

Wang, D., Tu, J., Chang, X. and Li, S., The lead–lag relationship between the spot and futures markets in China. *Quant. Finance*, 2017, **17**(9), 1447–1456.

Windzio, M., The network of global migration 1990–2013 using ergms to test theories of migration between countries. *Soc. Networks*, 2018, **53**(5), 20–29.

Xia, L., You, D., Jiang, X. and Chen, W., Emergence and temporal structure of lead–Lag correlations in collective stock dynamics. *Physica A*, 2018, **502**, 545–553.

Yao, C. Z. and Li, H. Y., Time-varying lead–lag structure between investor sentiment and stock market. *N. Am. J. Econ. Financ.*, 2020, **52**, 101148.