Li *et al. Financial Innovation* (2022) 8:51
https://doi.org/10.1186/s40854-022-00356-3

Financial Innovation

RESEARCH Open Access

Detecting the lead–lag effect in stock markets: definition, patterns, and investment strategies

Yongli Li[1*], Tianchen Wang[1], Baiqing Sun[1*] and Chao Liu[1,2]

*Correspondence: liyongli@hit.edu.cn; baiqingsun@hit.edu.cn
[1] School of Economics and Management, Harbin Institute of Technology, Harbin 150001, People's Republic of China
Full list of author information is available at the end of the article

Abstract

Human activities widely exhibit a power-law distribution. Considering stock trading as a typical human activity in the financial domain, the first aim of this paper is to validate whether the well-known power-law distribution can be observed in this activity. Interestingly, this paper determines that the number of accumulated lead–lag days between stock pairs meets the power-law distribution in both the U.S. and Chinese stock markets based on 10 years of trading data. Based on this finding this paper adopts the power-law distribution to formally define the lead–lag effect, detect stock pairs with the lead–lag effect, and then design a pure lead–lag investment strategy as well as enhancement investment strategies by integrating the lead–lag strategy into classic alpha-factor strategies. Tests conducted on 20 different alpha-factor strategies demonstrate that both perform better than the selected benchmark strategy and that the lead–lag strategy provides useful signals that significantly improve the performance of basic alpha-factor strategies. Our results therefore indicate that the lead–lag effect may provide effective information for designing more profitable investment strategies.

Keywords: Power-law distribution, Lead–lag effect, Stock market, Complex network, Investment strategy

Introduction

The lead–lag phenomenon, a phenomenon in which a security leads the price movement of another with some time delay, has been empirically evidenced as widely existing in financial markets (Gong et al. 2016). Although the "lead–lag effect" concept has been adopted in many studies (Kobayashi and Takaguchi 2018), few have provided a formal definition of this concept, and its underlying meaning is not always consistent. Some studies have focused on how to generate greater stock returns by utilizing the "lead–lag phenomenon" (Stübinger 2019) but have often failed to mine its embedded features. To this end, this study aims to answer the following questions: (1) Are there several stable patterns in stock markets that are characterized by the lead–lag phenomenon? (2) How can we formally define the lead–lag effect to provide a solid foundation for detecting

such an effect? (3) Can detecting the lead–lag effect enable the design of more profitable investment strategies that are more likely to earn excess returns?

The definition of the "lead–lag effect" is not equivalent to that of the "lead–lag relationship." That is, if one stock's volatility today mimics another stock's volatility yesterday, the two stocks are said to have a "lead–lag relationship" over the two successive days in which the former is the follower, and the latter is the leader. In fact, it is quite common for one stock to follow another stock some days during a year. Thus, an occasional lead–lag relationship could be regarded as random, which would not be very meaningful. However, if the lead–lag days of one stock pair are long enough to differ significantly from a random event, an effect can be deemed to exist between the pair. Accordingly, the first motivation of our work is to define the lead–lag effect by providing a statistical testing model, the goal of which is to judge whether the days characterized by a lead–lag relationship (hereafter, "lead–lag days") are significantly long in statistics.

Once the definition of the lead–lag effect is scientifically determined, a method for detecting stock pairs characterized by the lead–lag effect can be proposed. However, two questions must first be addressed. These are: (1) how do external variables affect the detection results and (2) are the detection results sensitive to these influential external variables? The answers to these two questions will deepen our understanding of the proposed detection model. The patterns of external variables that influence the results will enable us to adopt the proposed model by selecting the appropriate variable values. The robustness of the proposed model is notable for its usage in investment practices in real-world stock markets because a model's robustness is desirable for designing investment strategies. Accordingly, the answers to these two questions will reveal the properties of the proposed model.

As a typical application, the detected lead–lag effect aims to be adopted in guiding investments in real-world stock markets. Apparently, detecting stock pairs with a significant lead–lag effect can benefit investors because the price movements of followers will mimic those of their leaders. Thus, this study will first examine the performance of the pure lead–lag strategy and then judge if it is satisfactory. If it is satisfactory, we will regard the detected lead–lag effect as an enhancement signal, and then add it to some classic investment strategies to propose enhancement investment strategies. Generally, when a basic strategy is enhanced by another strategy, we refer to it as a single-enhancement investment strategy. The alpha-factor strategy is selected as the basic strategy, and our proposed lead–lag strategy is adopted to enhance it. Accordingly, the third motivation is to design profitable investment strategies based on the detected lead–lag effect, and then test its performance in a pure investment strategy and the proposed enhancement strategies.

To sum up, the contributions of this study are as follows: (1) The features of the lead–lag phenomenon are explored in the context of both the U.S. and Chinese stock markets. As a result, the number of stock pairs characterized by the lead–lag relationship meets the well-known power-law distribution, which offers novel evidence that the power-law distribution exists widely in the real world (Clauset et al. 2009) and specifically in the financial domain (Gabaix et al. 2003). (2) A formal definition of the lead–lag effect is provided according to the principles of statistical testing, and a detection approach is proposed based on this definition. It is worth noting that most existing studies regard

the lead–lag relationship between stocks as a phenomenon (Scherbina and Schlusche 2020; Dao et al. 2018; Huth and Abergel 2014), whereas this study elevates this phenomenon into an effect. Accordingly, the lead–lag effect must be formally defined via statistical testing, which lays a foundation for future studies to compare and detect the lead–lag effect in various scenarios. The rationality and robustness of the proposed detection approach are carefully examined by determining how external variables influence the lead–lag effect. (3) A few profitable investment strategies are designed and validated based on the detected lead–lag effect, in parallel to previous design and validation studies such as Shen et al. (2017), Xiong et al. (2020), Flori and Regoli (2021), and Zhang et al. (2021). Here both the pure lead–lag strategy and the enhancement strategies report sound results regarding the functionality of the detected lead–lag effect.

The remainder of this paper is organized as follows. "Section Related work" reviews the related work to clearly delineate the aforementioned contributions; "Section Method for detecting the lead–lag effect" defines the lead–lag effect and proposes a detection methodology; "Section Main results and validation in real-world stock markets" explores the features of the lead–lag phenomenon based on a selected real-world dataset, applies the proposed detection method, and tests the method's robustness; "Section Investment strategies based on the detected lead–lag effect" designs investment strategies and validates their performance to reveal the functionality of the detected lead–lag effect; and "Section Conclusions and future work" summarizes the study and discusses potential future work.

Related work

Our work is directly related to two fields of existing studies: this includes the lead–lag phenomenon in stock markets and the focused alpha-factor strategy widely adopted in stock markets. Each field is reviewed individually in the following sections.

Lead–lag phenomenon in stock markets

The lead–lag phenomenon is a classic financial topic that has attracted the attention of numerous researchers (Conlon et al. 2018). First, one fundamental question has been widely examined in the literature: does the lead–lag phenomenon exist in the stock market? Generally, the lead–lag phenomenon can be observed in high-frequency data such as 5-min stock price movements. Both Jong and Nijman (1997) and Huth and Abergel (2014) deemed that the lead–lag relationship is an essential stylized fact at high frequencies. Fonseca and Zaatour (2017), Dao et al. (2018), Buccheri et al. (2019), Campajola et al. (2020), and many others do not mention the existence of the lead–lag phenomenon in high-frequency data, the influencing factors, or even its potential origins. However, when observed in high-frequency data, this phenomenon is often called the "lead–lag relationship" rather than the "lead–lag effect". In most cases, the lead–lag relationship is unstable in high-frequency data. Since, according to Tóth and Kertész (2006) and Curme et al. (2015), its appearance is likely to be occasional, our work aims to formulate a new approach to finding a stable lead–lag relationship over a long time period based on statistical testing and to rename such stable lead–lag relationship "the lead–lag effect" as an indication of its statistical significance.

Second, the existing literature explores how to take advantage of the lead–lag phenomenon in designing investment strategies for real-world stock markets. Typically, investment strategies that utilize the lead–lag phenomenon are often variations on the high-frequency trading strategy, which is in accord with the results. Stübinger (2019) developed an optimal causal path algorithm and designed statistical arbitrage strategies for high-frequency data based on the lead–lag phenomenon. However, designing an investment strategy based on high-frequency data still has drawbacks. According to Krauss (2017), high-frequency trading strategies are associated with greater commission fees and a higher transaction threshold for investors.

In contrast, the stable lead–lag effect discovered in low-frequency data facilitates the practices of small and medium investors because of its ample optional trading time and low technical threshold. Scherbina and Schlusche (2020) and Gupta and Chatterjee (2020) have pointed out that the lead–lag relationship enables out-of-sample forecasting and thus helps in the design of investment strategies. From this perspective, this study can also be seen as the development of our previously published work (Li et al., 2021), which focuses on identifying the factors that cause the lead–lag phenomenon. However, this study aims to develop new investment strategies by utilizing the lead–lag phenomenon, and thus the two have divergent research aims. In contrast to the successive lead–lag days analyzed in our previously published work, this study considers the number of cumulative lead–lag days that would benefit extending the application of the model in real-world stock markets.

According to the aforementioned literature and gaps in the existing research, we believe that it is meaningful and even necessary to study the lead–lag effect in low-frequency data for the following reasons: (1) the definition of the lead–lag effect is not unified or discussed in depth in the existing literature, and thus the underlying significance of the lead–lag phenomenon often differs despite their use of the same name; (2) traditional studies on detecting the lead–lag phenomenon are conducted using classical econometrics or empirical research methods, and thus the use of data-driven technical analysis to detect the lead–lag effect can supplement existing studies with a new perspective; and (3) building on the traditional methods of designing investment strategies by using the discovered lead–lag phenomenon, our study may identify effective signals, which will have a guiding significance for the development of investment strategy. Accordingly, this study contributes to the literature by providing a unified and solid definition of the lead–lag effect and by utilizing the lead–lag effect to design profitable trading strategies in real-world stock markets.

Alpha-factor strategy

Concerning our targeted enhancement investment strategies, the alpha-factor strategy that originated with the capital asset pricing model is chosen as the primary strategy due to its popularity and effectiveness in real-world investment practices (Sharpe 1964; Makarov and Plantin 2015). The alpha factor in the alpha-factor strategy reacts to one or some stock attributes; in other words, different alpha factors reflect different stock attributes. Thus, the alpha-factor strategy consists of numerous specific strategies using various alpha factors. Since alpha factors are used as buying and selling signals in the alpha-factor strategy, its choice is the core of the strategy. Generally, existing studies

focus mainly on the following two types of alpha factors: value alphas and transactional alphas.

Value alphas are derived from the fundamentals of one stock and describe its value attributes. Value alphas include but are not limited to value factors (Balatti et al. 2017; Eisdorfer et al. 2019), size factors (Liu et al. 2019), growth factors (Fama and French 1998), profitability factors (Hou et al., 2015; Fama and French 2015), and momentum factors (Fama and French 2012; Berggrun et al. 2020). Based on the mature factor model, value alphas provide not only a valuable tool for stock valuation, but also a reasonable explanation for the cross-section of stock returns (Harvey et al. 2016). Accordingly, when value alphas are adopted in a strategy, it indicates that the investor cares about the value investment's underlying factors (Fama and French 2016). In contrast to traditional value alphas, transactional alphas pay more attention to the patterns embedded in trading behaviors (Casgrain and Jaimungal 2019). Transactional alphas are obtained by means of technical analysis and derived from transaction data. With the current progression of computer science, millions of transactional alpha factors have been identified by automated algorithms. Despite the lack of a good explanation, the marginal revenue contributed by transactional alpha factors is relatively satisfactory (Kakushadze 2016); large financial institutions favor such transactional alphas. For example, the 101 alpha factors proposed by the World Quant and the 191 alpha factors from Guotai Junan Securities have been welcomed by many institutions and investors.

The alpha-factor strategy is always used for stock selection. The proposed lead–lag strategy in our work helps allocate the weight of each selected stock in an investment portfolio. Therefore, it is convenient to combine the two strategies when designing an enhancement strategy. Since the alpha-factor strategy includes numerous specific models with various alpha factors, we select it as the primary strategy to demonstrate representativeness. The lead–lag effect falls into the category of technology-driven analysis and therefore resonates with transactional alphas, which are determined by technical analysis.

For this reason, it would be more natural to combine the lead–lag trading strategy with transactional alphas. Accordingly, this study focuses mainly on transactional alpha-factor strategies, regarded as the basic strategies when designing enhancement investment strategies. Our work exploits the great potential of the existing alpha factors and provides a framework for enhancement strategies by integrating the lead–lag effect into the existing alpha-factor strategies.

Method for detecting the lead–lag effect

The daily lead–lag network

Let $r_{i,t}$ denote the yield rate of stock i on day t. Its mathematical expression is as follows:

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

where $p_{i,t}$ denotes the closing price of stock i on day t. Here, the adopted stock price is restoring the right price rather than the price of ex-rights. If a suspension occurs for stock i on day t, then both $r_{i,t}$ and $r_{i,t+1}$ are set to "NAN." Next, given a manufactured

threshold Δ ($0 \leq \Delta < 1$), the definition that stock j follows stock i on day t is defined as follows:

Definition 1 The conditions for forming a lead–lag link. If and only if the following condition holds:

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

then, stock i follows stock j on day t.

Definition 1 states that if the difference between the yield rate of stock i on day t and that of stock j on day t–1 is within the given threshold, Δ, stock i is judged to follow stock j on day t. Further, let $\mathbf{G}_t$ denote the lead–lag network on day t, and its element $g_{ij,t}$ reflects the status of stock i following stock j on day t. If stock i is judged as the follower of stock j on day t according to Definition 1, then $g_{ij,t}=1$; otherwise, $g_{ij,t}=0$. Our model allows one stock to follow itself, and thus it is possible that $g_{ii,t}=1$ holds. Then, given the closing prices of all concerned stocks during the sequential $T+1$ trading days, we can achieve T lead–lag networks according to Definition 1.

During the targeted period (e.g., the total number of $T+1$ trading days), the achieved T lead–lag networks can tell us how many days stock i follows stock j in total. Formally, let d_{ij} denote the number of accumulated days that stock i follows stock j during the targeted period, which can be calculated as follows:

$$d_{ij} = \sum_{t=1}^{T} g_{ij,t}. \tag{3}$$

$\mathbf{G}_t$ is an asymmetrical matrix in most cases, considering that d_{ij} is not often equal to d_{ji}.

Concerning the manufactured threshold Δ, a larger Δ will cause the achieved daily lead–lag networks to have more directed links than a smaller Δ, thus the threshold Δ affects network density. Because it is an artificial variable, we will explore how it affects the results and check whether our method is robust under different threshold values in Sect. 4.2.1. The mainstream literature defining the relationship between stock pairs, such as Huang et al. (2009), Kumar and Deo (2012), Peralta and Zareei (2016), Xia et al. (2018), Deev and Lyócsa (2020), and many others, has often adopted the correlation coefficient. Note that most existing literature related to this definition uses data from a defined period to calculate the so-called "correlation coefficient," whereas our study uses daily data to define each day's lead–lag relationship between stock pairs. Therefore, the novel idea of using the selected data (i.e., "daily usage to define the lead–lag relationship" or "usage together to calculate a correlation coefficient during the selected period") leads to one of the differences between our study and the existing literature.

Definition and detection of the lead–lag effect

As explained in the introduction, when d_{ij} (which is defined in Eq. (3)) is long enough to be significantly distinct from the amount achieved in a random event, we tend to believe that the lead–lag effect from stock j to stock i holds, where stock j is the leader and stock

i is the follower. However, the criterion for judging whether d_{ij} is sufficiently long or not should be determined before formally defining the lead–lag effect. Fortunately, statistical testing enables us to formulate the following criterion: the *null hypothesis* is set to "all the links in the daily lead–lag networks are randomly formed," the null hypothesis will allow us to obtain the distribution of the accumulated days of all stock pairs. Then, given the statistical significance level (e.g., 0.10, 0.05, 0.01, etc.), the criterion can be immediately achieved in the obtained distribution. To clarify, let $\hat{d}$ denote the criterion. The meaning of the term "lead–lag effect" is provided in Definition 2.

Definition 2 Lead–lag effect. Based on the calculated $\hat{d}$, for any pair of stocks (e.g., i and j), if the d_{ij} achieved from Eq. (3) satisfies $d_{ij} \geq \hat{d}$, the lead–lag effect from stock j to stock i is judged to hold. If $d_{ii} \geq \hat{d}$, the lead–lag effect from stock i to itself is judged to hold.

Note that the principles of statistical testing imply that it is almost impossible for a rare event to occur in one random trial. Given the null hypothesis and the statistical significance level, the criterion for judging whether an event is rare or not can be achieved. Then, if a rare event occurs in the analyzed real-world data, we can reject the null hypothesis under the given statistical significance level, or we can determine that the rare event has a statistically significant effect. In fact, few studies have formally defined the lead–lag effect. As mentioned in the Related Work section, the lead–lag phenomenon or relationship was more often examined in the existing literature rather than the lead–lag effect. We detected the lead–lag effect via formal statistical tests and null reference networks, but the existing literature adopted different approaches to detecting the lead–lag relationship, such as the Granger test (Scherbina and Schlusche 2020; Zeng and Atta Mills 2021) and the optimal causal path algorithm (Jiang et al. 2019). Accordingly, the approach adopted leads to different definitions, so our definition is new in this field.

Table 1 shows the detailed process of achieving criterion $\hat{d}$. Random networks are first generated to achieve the distribution of the accumulated days between all stock pairs under the null hypothesis. Then, criterion $\hat{d}$ can be obtained given the statistical significance level in Step (1). Here, we refer to the configuration model proposed by Newman

Table 1 The process of achieving the criterion $\hat{d}$

Input: the total number of stocks N, the daily lead–lag network $\mathbf{G}_t$ achieved from real data ($t = 1, 2, \ldots, T$), and the statistical significance level δ

Output: the criterion $\hat{d}$

Steps:

(1) Random directed network $\mathbf{RG}_t$ on day t is generated by retaining the node degree distribution of $\mathbf{G}_t$ via the configuration model. The code can be directly achieved from https://networkx.org/documentation/stable/modules/networkx/generators/degree_seq.html#directed_configuration_modmo. Note that the case of following oneself is also considered

(2) By repeating the steps (1) from $t = 1$ to T, a series of random networks $\mathbf{RG}_1, \mathbf{RG}_2, \ldots, \mathbf{RG}_T$ can be obtained. Then, the accumulated following days of each pair (denoted as rd_{ij}, $i, j = 1, 2, \ldots, N$) are achieved based on the above series of random networks via Eq. (3). As a result, one group of simulation is completed with the obtained set $\{rd_{ij}\}_{i,j=1}$

(3) Hundred groups of simulations can be conducted (e.g., 500 groups) as above and then all the accumulated following days of each pair obtained from each group are put together to get their distribution. Given the statistical significance level δ, the corresponding criterion $\hat{d}$ can be immediately achieved

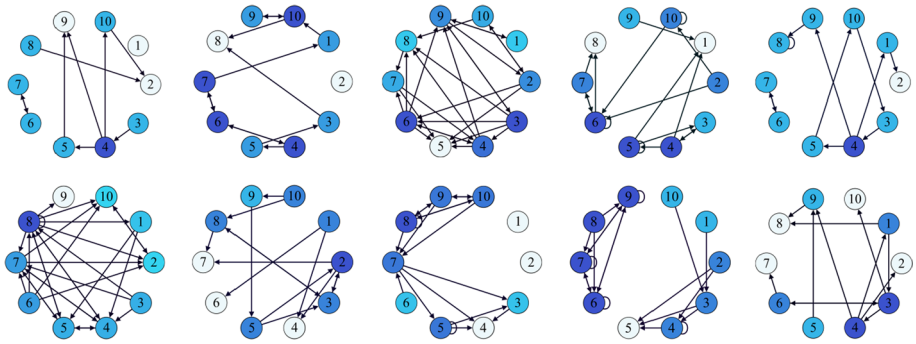


Fig. 1 Graphs of 10 daily lead–lag networks

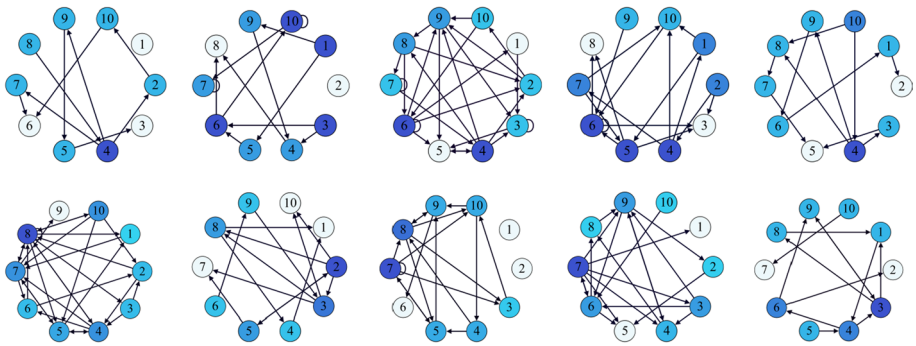


Fig. 2 Graphs of 10 sequential random networks

et al. (2001). The generated random networks retain the characteristics of the daily network as much as possible. Although the network indicator of the real-world lead–lag network changes each day, the adopted configuration model guarantees that each day's random network and the same day's real-world network share an almost identical node degree distribution, which is superior to the model that retains only the same edge number. Next, the statistical significance level δ is set to 0.001 since a lower significance level means a more rigorous criterion for determining the lead–lag effect. Once output $\hat{d}$ is achieved based on the process shown in Table 1, Definition 2 directly judges which stock pair features the lead–lag effect. Hereafter, the stock pairs detected with the lead–lag effect are called "*lead–lag stock pairs*."

Example

A simple example is presented to show how the proposed detection method works. This example analyzes the closing price of 10 stocks on 11 sequential trading days, and then obtains 10 daily lead–lag networks using Eq. (2). As displayed in Fig. 1, each node represents one stock, and the direction of the link points from the leader to the follower. The color of the node distinguishes between differences in the node out-degree (i.e., the number of followers): the more significant the out-degree, the darker the color.

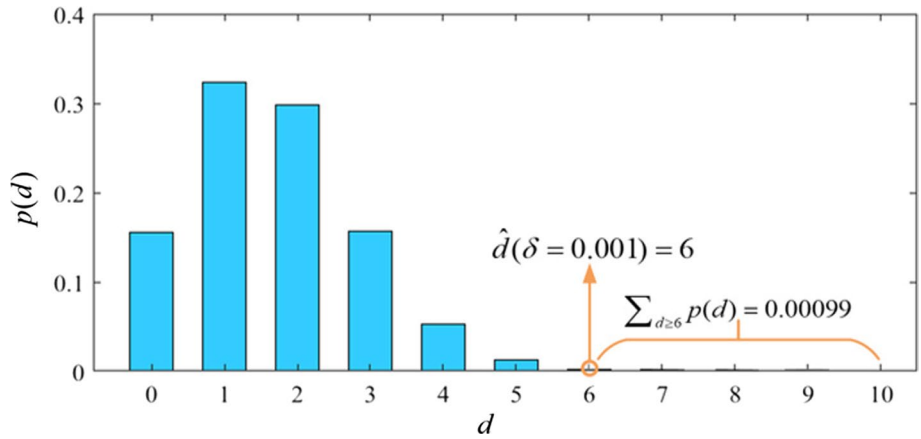


Fig. 3 Distribution of the accumulated following days via 500 groups of simulations

Following Steps (1) and (2) in Table 1, one group of simulations can achieve the following 10 sequential random networks, as shown in Fig. 2. Here, each day's random network retains the node degree distribution of the same day's real-world lead–lag network, which can be checked by comparing the counterpart in Figs. 1 and 2.

Then, by conducting 500 groups of simulations according to Step (3) in Table 1, the distribution of all the achieved accumulated days is achieved and displayed in Fig. 3. When the statistical significance level (δ) is set to 0.001, the criterion $\widehat{d}$ is equal to 6, based on the achieved distribution in Fig. 3. Accordingly, the detected leader–follower pairs are $3 \rightarrow 4$, $4 \rightarrow 5$, and $6 \rightarrow 7$.

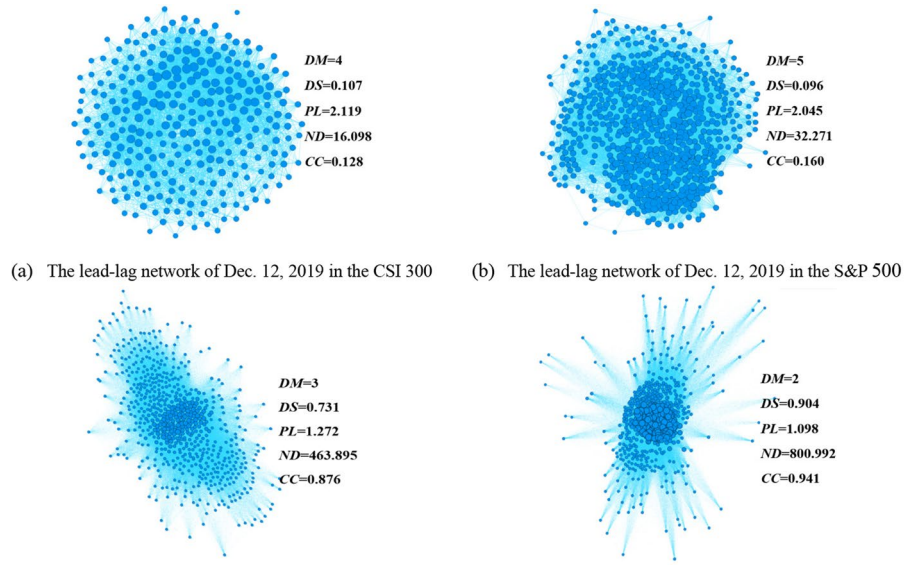
Main results and validation in real-world stock markets

To apply the previous simple example to real-world stock markets, this section selects the stock markets of mainland China and the U.S. as the targets of analysis. This section applies the proposed method to detect which stock pairs are characterized by the defined lead–lag effect and explores how the man-made variables embedded in the detection method affect the results. The following subsections introduce the process of data selection, report the main results in different stock markets, and discuss these findings.

Data preparation and main statistical results

Data preparation

Two stock sets are selected for the application and validation of the proposed method. One is the set of 300 stocks contained in the China Securities Index 300 (CSI 300), considering that these 300 stocks are the most liquid stocks in mainland China's A-share stock market; therefore, they are often used to reflect its overall performance. The other is the set of stocks included in Standard & Poor's 500 Index (S&P 500), which helps us understand the proposed method's performance in the U.S. stock market. Note that the stocks in the CSI 300 and the S&P 500 are not permanent, although adjustments to the stock set are quite infrequent.



(a) The lead-lag network of Dec. 12, 2019 in the CSI 300 (b) The lead-lag network of Dec. 12, 2019 in the S&P 500

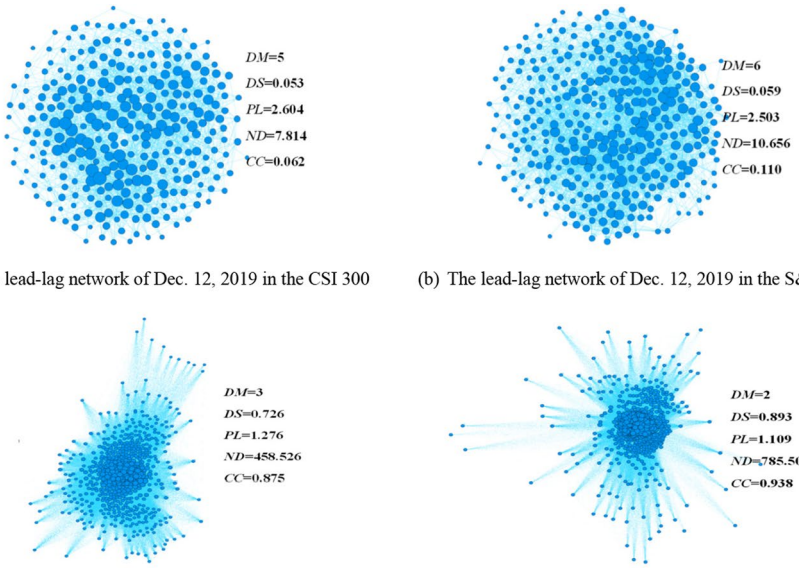
(c) The overall lead-lag network for 10 years in the CSI 300 (d) The overall lead-lag network for 10 years in the S&P 500

Fig. 4 The lead–lag networks achieved in the two selected stock sets ($\Delta = 0.20$)

The closing price of each stock in the two selected stock sets is collected on each trading day between January 1, 2010, and December 31, 2019 (i.e., over 10 years). The data were obtained from the Compustat database at https://wrds-www.wharton.upenn.edu/; each year has an average of 250 trading days. The stocks featured in each stock set changes over time because new stocks were added and others were removed during the chosen period. Almost every trading day witnessed stock suspensions due to some reason or rule, and thus the size of the daily lead–lag network fluctuates. As shown in Fig. 4, different stock sets feature different overall directed lead–lag networks in terms of their diameter (*DM*), density (*DS*), average path length, average node degree (*ND*), and clustering coefficient.

Recalling Eqs. (1–2) in Sect. 3.1, the daily lead–lag networks can be immediately achieved in each stock set based on the above-prepared data once the man-made threshold Δ is given. Here and hereafter, taking $\Delta = 0.20$ if no special statements are provided, the upper part of Fig. 4 displays the lead–lag networks achieved in each stock set on December 12, 2019. The overall lead–lag network can be obtained in each stock set by summing up each day's lead–lag network. The bottom part of Fig. 4 shows the overall lead–lag network of each stock set during the entire period; the link thickness is proportional to the number of cumulative days on which one stock followed the other in this directed link.

To display more results under different values of the man-made threshold, Δ, Fig. 5 shows the achieved lead–lag networks in the two selected stock sets when $\Delta = 0.10$. By comparing Figs. 4 and 5, we find that different values of Δ cause only slight changes in the overall lead–lag networks and their corresponding indicators in both markets, except that the average ND decreases with a decrease in Δ. However, the change in Δ has a significant impact on the daily networks of both markets because the daily network is not as robust as the overall network.



(a) The lead-lag network of Dec. 12, 2019 in the CSI 300 (b) The lead-lag network of Dec. 12, 2019 in the S&P 500

(c) The overall lead-lag network for 10 years in the CSI 300 (d) The overall lead-lag network for 10 years in the S&P 500

Fig. 5 The lead–lag networks achieved in the two selected stock sets ($\Delta = 0.10$)

Power-law distribution

Before formally describing the detailed analysis, we will first recall basic knowledge about the random network, the scale-free network, and the power-law distribution that is often seen in the fields of complexity science and network analysis. First, a random network indicates that the links in the network are randomly formed; in other words, the links are generated with a given probability (Barabási and Albert 1999). The random network is often used as a testable null hypothesis about network structure (Volz 2004). Its link distribution is thin-tailed, and our work follows this idea. In contrast to a random network, a scale-free network refers to one with a degree distribution that meets the power law, at least asymptotically (Barabási and Bonabeau 2003). Roughly speaking, the distribution discrepancy between a random network and a scale-free network often originates in human activity, that is, human activity causes the change from a thin-tailed to a heavy-tailed distribution (represented by the power-law distribution). In addition, human activity also makes the power-law distribution more prevalent and special in the field of complexity science, and even the power-law distribution is viewed as a signature of complexity by noting that such a distribution can reflect the underlying pattern of a complex process (Rickles 2011). Our study considers the function of human activity in the stock market and thus tests the power-law distribution as stated below.

Although the overall lead–lag network of each stock set is unique, we wonder whether some identical patterns exist for different stock sets. If they do, we can call the discovered identical pattern a feature, because different stock sets do not alter the features embedded in the lead–lag phenomenon. To answer this question, we will focus on the link thickness displayed in Fig. 4 and examine its distribution. The distribution of the concerned link thickness is equal to that of variable d_{ij} defined in Eq. (3) by carefully considering the meaning of link thickness. Figure 6 displays the distribution in each stock set using $\Delta = 0.20$. As displayed in Fig. 6, the points in the tail of each distribution are

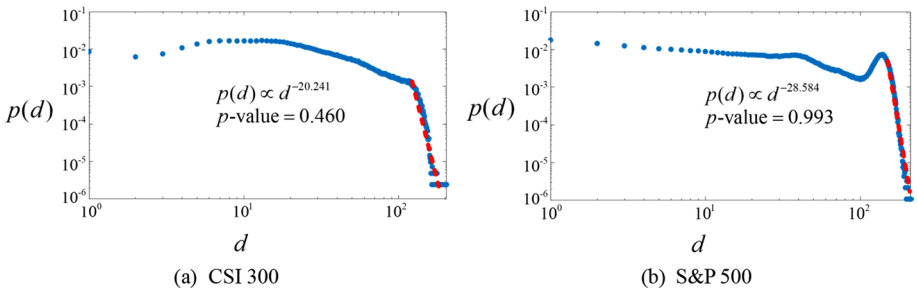


Fig. 6 Distribution of the link thickness in each stock set as well as the test results

almost in a line in the log–log coordinates (i.e., a feature of the power-law distribution), indicating that the tested distribution is quite likely to meet the power-law distribution.

Next, according to the mainstream testing methods used in the existing literature (Clauset et al. 2009; Malevergne et al. 2011; Toda 2012) to verify the power-law distribution, we apply three methods to obtain sound results: the Kolmogorov–Smirnov Test (K–S), the Kuiper Test (Kuiper 1960), and the Anderson–Darling test (A–D) (Scholz and Stephens 1987; Coronel-Brizio and Hernández-Montoya 2010). Recalling Eq. (2), the manufactured threshold, Δ, affects the achieved daily lead–lag networks as well as the overall lead–lag networks in both markets. Here, we test whether the power-law distribution holds under different values of Δ. Table 2 shows the results: none of the three methods rejects the null hypothesis that "the data meets the power-law distribution" at the statistical significance level of 0.05. Therefore, we believe that the power-law distribution can be regarded as a stable pattern underlying the lead–lag phenomenon.

In addition, we conduct additional tests to exclude the other possible distributions and provide additional evidence supporting the discovered power-law distribution. As both markets witnessed steep decays in the log–log coordinates shown in Fig. 6, two possible discrete and thin-tailed distributions such as the Poisson distribution or the binomial distribution are estimated and tested using the three testing methods. To make our results sound, we change the value of Δ to test the sensitivity of the results to this manufactured parameter. The results for the two tested distributions are shown in Tables 3 and 4. When the statistical significance level is 0.05, the two distributions are rejected in both markets in most cases, although several exceptions exist for the binomial distribution at $\Delta = 0.15$ under the Kuiper Test. In summary, these results provide more evidence that the verified distribution is likely to meet the power-law distribution.

Based on these findings, we now address why the discovered power-law distribution is important in our work. Our proposed detection approach is more meaningful when facing a power-law distribution than a thin-tailed distribution because very few stock pairs (the number is negligible) can be detected as having a lead–lag effect with a thin-tailed distribution, but the power-law distribution guarantees that a considerable number of stock pairs may be detected. As expected, more detected stock pairs implies more opportunities to utilize the information contained in the lead–lag effect

Table 2 The results of the power-law distribution test under different manufactured thresholds, Δ

Δ	CSI 300						S&P 500					
	$\hat{\lambda}$ of K–S	*p* value	$\hat{\lambda}$ of Kuiper	*p* value	$\hat{\lambda}$ of A–D	*p* value	$\hat{\lambda}$ of K–S	*p* value	$\hat{\lambda}$ of Kuiper	*p* value	$\hat{\lambda}$ of A–D	*p* value
0.10	− 10.946	0.088	− 13.524	0.999	− 12.684	0.097	− 19.124	0.466	− 11.476	0.999	− 11.476	0.127
0.15	− 15.844	0.392	− 16.582	0.999	− 18.298	0.427	− 29.637	0.998	− 20.201	0.999	− 25.722	0.364
0.20	− 20.241	0.460	− 16.577	0.999	− 20.241	0.458	− 28.584	0.993	− 23.252	0.999	− 27.778	0.752
0.25	− 24.245	0.999	− 16.510	0.999	− 20.301	0.686	− 30.000	0.994	− 26.959	0.999	− 30.000	0.719
0.30	− 27.579	0.927	− 18.807	0.999	− 24.206	0.506	− 21.927	0.962	− 29.009	0.999	− 23.228	0.654

The statistical significance level is set to 0.05. Here, $\hat{\lambda}$ is the estimated value of parameter λ from the probability function of the power-law distribution (i.e., $p(x = k) \propto k^{-\lambda}$)

Table 3 The results of the Poisson distribution test under different manufactured thresholds, Δ

Δ	CSI 300						S&P 500					
	$\hat{\alpha}$ of K–S	*p* value	$\hat{\alpha}$ of Kuiper	*p* value	$\hat{\alpha}$ of A–D	*p* value	$\hat{\alpha}$ of K–S	*p* value	$\hat{\alpha}$ of Kuiper	*p* value	$\hat{\alpha}$ of A–D	*p* value
0.10	75.236	0.000	78.195	0.000	71.451	0.000	38.094	0.000	36.464	0.000	36.389	0.000
0.15	102.343	0.000	107.535	0.038	99.510	0.000	115.679	0.000	122.484	0.014	114.548	0.000
0.20	134.573	0.000	128.299	0.034	135.362	0.000	160.827	0.000	170.017	0.000	169.185	0.000
0.25	167.018	0.000	175.649	0.000	176.570	0.000	200.461	0.000	200.094	0.000	207.783	0.000
0.30	204.753	0.000	208.141	0.000	216.652	0.000	239.250	0.000	248.230	0.000	242.992	0.000

The statistical significance level is set to 0.05. Here, $\hat{\alpha}$ is the estimated value of parameter α contained in the probability function of the Poisson distribution (i.e., $P(X = k) = \frac{\alpha^k}{k!} e^{-\alpha}$)

Table 4 The results of the binomial distribution test under different manufactured thresholds, Δ

Δ	CSI 300						S&P 500					
	$\hat{\beta}$ of K–S	*p* value	$\hat{\beta}$ of Kuiper	*p* value	$\hat{\beta}$ of A–D	*p* value	$\hat{\beta}$ of K–S	*p* value	$\hat{\beta}$ of Kuiper	*p* value	$\hat{\beta}$ of A–D	*p* value
0.10	0.409	0.000	0.420	0.003	0.392	0.000	0.343	0.000	0.347	0.000	0.352	0.000
0.15	0.536	0.000	0.553	0.180	0.510	0.000	0.710	0.001	0.695	0.178	0.714	0.000
0.20	0.660	0.000	0.641	0.049	0.660	0.000	0.759	0.000	0.782	0.046	0.752	0.000
0.25	0.723	0.000	0.739	0.047	0.765	0.000	0.777	0.000	0.773	0.013	0.782	0.000
0.30	0.819	0.000	0.835	0.006	0.803	0.000	0.792	0.000	0.793	0.007	0.766	0.000

The statistical significance level is set to 0.05. Here, $\hat{\beta}$ is the estimated value of parameter β contained in the probability function of the binomial distribution (i.e., $P(X = k) = \binom{n}{k} \beta^k (1-\beta)^{n-k}$)

Table 5 Robustness results in CSI 300

DD (*p*)	0.15	0.20	0.25	0.30	0.35
0.10	**0.112 (0.142)**	**0.191 (0.070)**	0.252 (0.000)	0.301 (0.000)	0.342 (0.000)
0.15		**0.081 (0.880)**	**0.144 (0.247)**	**0.196 (0.136)**	**0.240 (0.070)**
0.20			**0.064 (0.942)**	**0.117 (0.836)**	**0.162 (0.596)**
0.25				**0.054 (0.999)**	**0.100 (0.999)**
0.30					**0.047 (0.999)**

Table 6 Robustness results in S&P 500

DD (*p*)	0.15	0.20	0.25	0.30	0.35
0.10	**0.230 (0.999)**	**0.284 (0.974)**	**0.327 (0.774)**	**0.364 (0.390)**	**0.403 (0.348)**
0.15		**0.206 (0.999)**	**0.259 (0.969)**	**0.291 (0.606)**	**0.322 (0.556)**
0.20			**0.189 (0.999)**	**0.246 (0.957)**	**0.272 (0.929)**
0.25				**0.176 (0.999)**	**0.237 (0.998)**
0.30					**0.166 (0.999)**

to improve earnings, which lays a foundation for designing more profitable investment strategies.

Main results and validation

By recalling the proposed detection approach, two manufactured variables will affect the detection results: *the threshold* Δ and *the period* ζ. As we have explained, the threshold, Δ, influences the achieved daily lead–lag networks. The period, ζ, is also an influencing factor because the predictability is likely to differ when different periods are chosen. The following two subsections will explore how these variables affect the detection results. These findings can also partially answer questions related to the model's robustness and the predictability of the results.

Detection results as a function of Δ

Recalling Eq. (2), the manufactured threshold Δ will affect the link formation in a daily lead–lag network to further influence the distribution of the variable d_{ij}s (by recalling Eq. (3) or Fig. 6). This subsection focuses on how the manufactured threshold Δ affects the aforementioned distribution. If the distributions obtained under different values of Δ differ significantly, the output of our model is sensitive to Δ, or, in other words, is not

robust, and vice versa. To this end, $DD(\Delta_i, \Delta_j)$ is defined in Eq. (4) by following the K–S test (Massey 1951) to measure the difference in the distribution as follows:

$$DD(\Delta_i, \Delta_j) = \max_d \left| cdf(d; \Delta_i) - cdf(d; \Delta_j) \right|. \tag{4}$$

where $cdf(d, \Delta_i)$ and $cdf(d, \Delta_j)$ denote the cumulative distribution function under thresholds Δ_i and Δ_j, respectively. Because the measurement defined in Eq. (4) is a K–S statistic, the K–S test can be conducted to check whether the difference is significant. Considering different combinations of Δ_i and Δ_j, Tables 5 and 6 report the statistic $DD(\Delta_i, \Delta_j)$ of each combination and its corresponding p value using the K–S test.

The numbers in bold in Tables 5 and 6 indicate that the difference between the two distributions is not significant at the significance level of 0.05. In addition, when $|\Delta_i - \Delta_j| \leq 10\%$, none of the distribution differences under different combinations are significant, implying robustness, especially when the deviation of the two threshold values is not too large. Moreover, not surprisingly, $DD(\Delta_i, \Delta_j)$ increases with $|\Delta_i - \Delta_j|$ in all the combinations in the two stock markets, and, even if the deviation of the two threshold values is as great as 20%, the distribution under some combinations is also insignificant. Overall, the achieved distributions are robust considering that they are not quite sensitive to the parameter Δ.

Detection results as a function of ζ

Before discussing the function of ζ, we first focus on *the prediction task*: the detected leader during period ζ serves as a signal, and the price movements of the detected followers act as the predicted target. Specifically, if leader stock i and its follower stock j are one of these detected lead–lag stock pairs during the given period ζ (i.e., ζ months), the price movement of stock j on day t can be inferred from that of stock I on day t–1. Then, we compare the real price movement of stock j with the movement predicted by its leader i on each trading day in the targeted month; thus, the prediction accuracy of the month can be calculated. To simplify the problem, we use 1, − 1, and 0 to denote the three price movements without considering the degree. In addition, if one follower has multiple leaders, the movement direction of the follower is determined by the majority of the leaders. When half of the leaders move up and half move down, the movement of the follower is predicted to be 0. Finally, by averaging all followers' prediction accuracy,

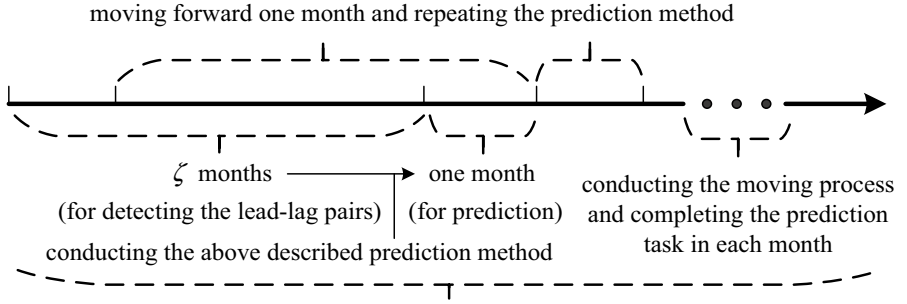


Fig. 7 The detailed process of the prediction task

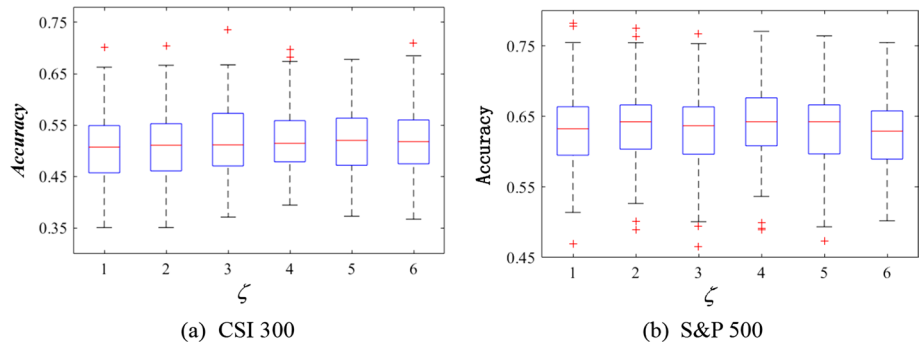


Fig. 8 Prediction accuracy under different values of ζ in each stock set

Table 7 Results of one-sample T tests by comparing the mean value to 0.50 in two stock sets

ζ	CSI 300			S&P 500		
	Mean value	Std error	T value	Mean value	Std error	T value
1	0.511	0.069	1.746*	0.628	0.054	26.095***
2	0.511	0.07	1.719*	0.635	0.051	29.176***
3	0.52	0.068	3.203***	0.628	0.054	25.784***
4	0.525	0.065	4.202***	0.639	0.054	28.277***
5	0.523	0.064	3.923***	0.634	0.055	26.448***
6	0.524	0.065	4.028***	0.623	0.053	25.627***

* $p<0.10$; ** $p<0.05$; *** $p<0.01$

we obtain the performance of the prediction task in the targeted month. The detailed process of the prediction task is displayed in Fig. 7.

Note that the detection results on the lead–lag stock pairs are dependent on the variable ζ. Thus, this subsection will explore the optimal value of ζ to achieve the best prediction performance. The performance is measured based on the overall accuracy shown in Fig. 7. On the one hand, the answer to this question will unveil the function of ζ on the detection results and even the accuracy of the simple prediction task, laying a foundation for designing profitable investment strategies; on the other hand, the answer will enable us to understand how much information is contained in the detected lead–lag stock pairs, although the prediction task is quite simple. If the mean overall prediction accuracy, as expected, is significantly greater than 50% (or say, a random guess), we tend to believe that the detected lead–lag stock pairs contain valuable information; a higher value means that they will be more helpful in designing profitable investment strategies in practice. Otherwise, we should consider how to better utilize and mine the information contained in the detection results.

Following the prediction task, Fig. 8 displays prediction accuracy under different values of ζ in each selected stock set. Here, the box plot under each value of ζ is achieved by 120 accuracy values, that is, the set of the overall accuracy obtained for each month (for prediction, as displayed in Fig. 7) over the 10 years between January 2010 and December 2019. As shown in Fig. 8, the medians of overall accuracy under different values of ζ are only a little higher than 0.50 for the CSI 300 and much higher than 0.50 for the

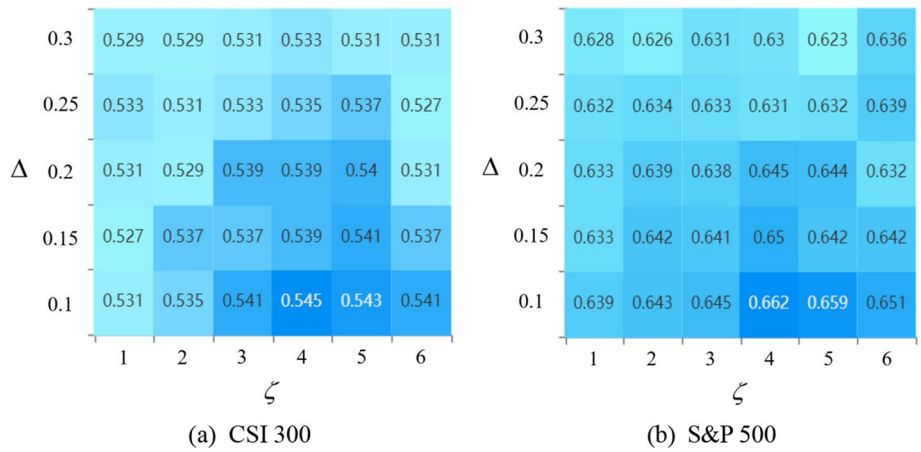


Fig. 9 Prediction accuracy under different combinations of two parameters in each stock set

S&P 500. Accordingly, the one-sample t-test is needed, especially for the CSI 300, to check whether the mean values of overall accuracy are significantly higher than 0.50 for every value of ζ, as stated in the previous paragraph. To this end, the results are listed in Table 7, showing that all the mean values are significantly higher than 0.50, at least under the significance level of 0.10, regardless of the stock set and the value of ζ.

Combining the results reported in Fig. 8 and Table 7, we find that the information contained in the detected lead–lag stock pairs helps design profitable investment strategies. In addition, the performance is robust to the manufactured variable ζ by noting that the discrepancy between the highest and lowest mean accuracy values is within 2% in both stock sets. Furthermore, the accuracies achieved in the S&P 500 are all higher than those in the CSI 300, implying that the detected lead–lag stock pairs will be much more beneficial in the S&P 500, which will be validated in "Section Investment strategies based on the detected lead–lag effect".

In addition, following the prediction task, different combinations of the two parameters (i.e., Δ and ζ) will yield varying accuracies. More importantly, the result in which combination has the best preformation will be useful for selecting parameters in designing investment strategies (see the next section). The two thermodynamic graphs displayed in Fig. 9 show the results for each stock set. According to Fig. 9, the prediction accuracy first increases and then decreases with an increase in ζ, in most cases, when Δ is fixed. The prediction accuracy increases with a decline in Δ, on average, but there are some exceptions when ζ is fixed. All the achieved accuracies are greater than 50%, demonstrating that the detected lead–lag stock pairs are helpful, even with the simple prediction task. Interestingly and more importantly, the best accuracy is achieved with the same parameter combination in the two stock sets; thus, the combination of $\Delta = 0.10$ and $\zeta = 4$ will lead to the most profitable lead–lag stock pairs, which will be adopted to design a more complicated investment strategy in the next section.

Investment strategies based on the detected lead–lag effect

The simple prediction task, described in Sect. 4.2.2, can be regarded as one of the most straightforward investment strategies because it only considers the direction of the predicted price movement without considering the trading details. In addition, the simple prediction task lays a foundation for designing more complicated investment strategies by revealing that the detected lead–lag stock pairs will yield more profitable information when $\Delta = 0.10$ and $\zeta = 4$. Based on the achieved parameter combination, this section extends the aforementioned simple prediction task into two types of practical investment strategies: the so-called "pure lead–lag strategy" and "enhancement strategies," which are determined by integrating the pure lead–lag strategy into well-known alpha-factor strategies. The following subsections will first present the two strategies designed in this study and then report their performance to guide investors' practices in real-world stock markets.

Pure lead–lag strategy

Our designed pure lead–lag strategy consists of three main steps based on the detected lead–lag stock pairs. The steps are listed in detail below.

Step 1: Calculate the strength of the influence of the leader on the follower.

A bipartite graph model is adopted to depict the leaders, followers, and their relationship, where the set of leaders and followers is denoted as N and M, respectively. For any $p \in N$ and $q \in M$, d_{pq} denotes the number of accumulated days that stock q follows stock p during the analyzed period by recalling Eq. (3). Then, let s_{pq} represent the influence strength of leader p on follower q with the following mathematical expression:

$$s_{pq} = \frac{d_{pq}}{\sum_{p \in N, q \in M} d_{pq}}. \tag{5}$$

According to Eq. (5), a greater number of accumulated days indicates a stronger influence. Once the detected bipartite graph is determined, the strength of the influence of all detected lead–lag stock pairs is also determined.

Step 2: Calculate each day's accumulated influence on the follower.

Let $w_{q,t}$ reflect the accumulated influence of the leader set on follower q on day t, which can be calculated as follows:

$$w_{q,t} = \sum_{p \in N} r_{p,t} s_{pq}, \tag{6}$$

where $r_{p,t}$ is the yield rate of stock p on day t according to Eq. (1). Then, similar to Eq. (5), normalizing the calculated $w_{q,t}$ achieves the following ratio variable $v_{q,t}$, which helps to determine the holding position of the follower stock q on day t. Equation (7) shows the specific expression for $v_{q,t}$ as follows:

$$v_{q,t} = \frac{w_{q,t}}{\sum_{q \in M} w_{q,t}}. \tag{7}$$

Step 3: Adjust the holding position of follower stock q based on $v_{q,t}$.

At the end of trading day t, the holding positions of all follower stocks can be adjusted based on the calculated $v_{q,t}$. Here, we assume that our adjustments can be instantly completed according to each follower's market price at the closing of the stock market that day. Let C_t denote total assets just before adjusting stock positions on trading day t; generally, C_t contains the holding stocks and currencies. Taking follower stock q as a representative example, the rule for adjusting stock positions is as follows: (1) when $v_{q,t} > 0$, the market value of stock q held in hand is adjusted to $v_{q,t}C_t$ through buying or selling, where the market value is measured at trading time; (2) when $v_{q,t} \leq 0$, the amount of stock q held in hand should be adjusted to 0.

Enhancement strategy

As mentioned in Sect. 2.2, the alpha-factor strategy selects stocks by calculating and ranking the value of the adopted alpha factor. Interestingly, the pure lead–lag strategy provides the selected stocks and the buy-and-sell signal. Naturally, the buy-and-sell signal can be adapted to the stock sets selected by both the alpha-factor and lead–lag strategies. As a result, the enhancement strategy can be designed by combining the buy-and-sell signal and stock selection, as explained previously. In addition, the trading framework of the commonly used alpha-factor strategy requires that the calculated value of the concerned alpha-factor should be updated each month. In practice, the value of the concerned alpha factor is calculated at the end of each month based on the technical data of that month, and then stock selection and trades are immediately made. Therefore, the trading day in the pure lead–lag strategy is identical to the alpha-factor strategy. The two strategies are coherent in terms of trading time when forming the enhancement strategy.

There are a total of four steps to conducting the enhancement strategy. The *first step* is to detect the lead–lag stock pairs based on the preceding work of this study and then determine the set of follower stocks, denoted as Q. The *second step* is to choose one alpha factor and then calculate the alpha value of each stock in set Q. Without any loss of generality, let α_q denote the calculated alpha value of stock q for any stock $q \in Q$. The *third step* is to achieve the variable value v_q (for any $q \in Q$, hereafter) by following the first two steps in the pure lead–lag strategy as stated in Sect. 5.1. The *last step* is to provide the trading rules based on α_q and v_q achieved in the foregoing steps and then use the rules to adjust the holding position of stock q.

Table 8 The process of calculating the alpha-01 value of each stock set Q in month TT

Input:	**For any stock *i* in one stock market, the two vectors OP_i and TV_i consist of daily opening price and daily trading volume of stock *i* on all the trading days of the month *TT*, respectively**
Output:	The value of alpha-01 of stock i, which is denoted as $\boldsymbol{\alpha}_i^{01}$
Process:	
(1)	For each stock i in the stock market, the correlation (denoted as $CORR_i$) between OP_i and TV_i is calculated
(2)	The achieved correlations of all the stocks are sorted in descending order, and then each stock i gets a rank number (denoted as RK_i) in the sorted vector
(3)	Then, $\boldsymbol{\alpha}_i^{01}$ can be calculated as $\boldsymbol{\alpha}_i^{01} = \frac{RK_i - 1}{N-1} - \frac{1}{2}$, (8) where N is the total number of stocks traded in the selected market

Here, we take our designed alpha-01 as an example to describe the aforementioned steps of the enhancement strategy to present them more clearly. We assume that the first step has been conducted and designated the follower stock set Q. Then, according to the second step, Table 8 describes the detailed process of calculating the alpha-01 value of each stock belonging to set Q.

Before conducting the steps of the proposed enhancement strategy, we first pay attention to the achieved α_i^{01}. On the one hand, if $CORR_i$ is low among all stocks, RK_i will be a large number and thus α_i^{01} will be high. In other words, if one stock's opening price is not consistent with its trading volume in the analyzed month, the calculated alpha-01 value of this stock will be high. On the other hand, Eq. (8) guarantees that the calculated alpha-01 value ranges between -0.5 and 0.5; half of the stocks traded on the stock market have a positive alpha-01 value. Note that different alpha factors have different calculation processes. Our study adopts 20 different alpha factors by following Kakushadze (2016) to ensure that our results are sound. Their detailed calculation processes are presented in Appendix 1.

Performance

This section aims to validate the performance of our proposed lead–lag strategy and test whether this strategy improves the performance of the pure alpha-factor strategies in the formed enhancement strategy. As in the work of Stübinger (2019), the trading cost is set to 0.25%, and the naïve buy-and-hold investment strategy (MKT) is chosen as the benchmark. In addition, we choose the upper 5% daily return rate, or the Sharpe ratio, of a series of random investment operations as another benchmark, where a random investment operation means buying one stock on a random day and selling it on a later random day. With the aim of obtaining sound results, we designed 20 different alpha-factor strategies (see Appendix 1) for validation and chose a testing period of 10 years (i.e., from January 2010 to December 2019).

Because the alpha-01 strategy is the example in Sect. 5.2, this section first focuses on the performance of the pure lead–lag strategy (PLL), the pure alpha-01 strategy (Pure-01), and the enhancement strategy of alpha-01 (Enhan-01). Table 9 reports their

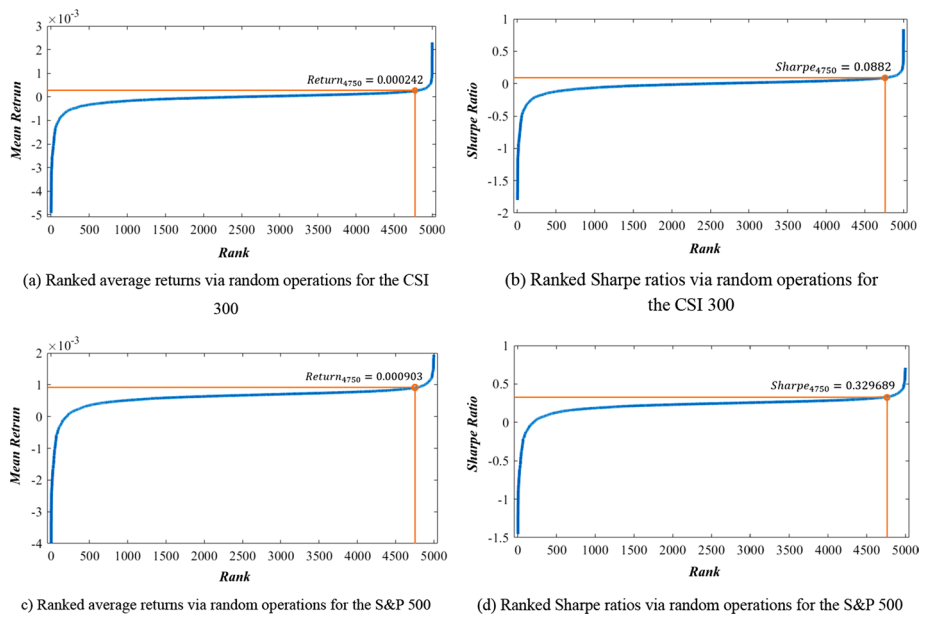
Table 9 The performance of PLL, Pure-01, Enhan-01, and MKT in each stock market

Indexes	CSI 300				S&P 500			
	PLL	Pure-01	Enhan-01	MKT	PLL	Pure-01	Enhan-01	MKT
Mean return	8.80E−05	−0.00011	0.00031	8.60E−05	0.00116	0.00101	0.00153	0.00074
Std error	0.00855	0.01542	0.00692	0.01021	0.00887	0.00897	0.00985	0.00930
Minimum	−0.06331	−0.09457	−0.16219	−0.06051	−0.06344	−0.06638	−0.05837	−0.06663
Quartile 1	−0.00335	−0.00689	−0.00197	−0.00531	−0.00330	−0.00339	−0.00368	−0.00327
Median	0.00017	0.00014	9.38E−05	0.00046	0.00085	0.00077	0.00094	0.00060
Quartile 3	0.00397	0.00759	0.00252	0.00582	0.00503	0.00515	0.00540	0.00505
Maximum	0.04431	0.06640	0.04199	0.05251	0.06771	0.05064	0.15993	0.04960
Skewness	−0.88929	−0.64429	−5.36223	−0.29767	−0.23049	−0.36344	1.30722	−0.40880
Kurtosis	11.05965	7.63471	132.602	5.76534	8.25482	7.43435	32.6182	7.45357
Sharpe ratio	0.14670	−0.17437	0.70335	0.14780	0.87648	0.82865	1.13450	0.80963
Max-draw	0.41724	0.56303	0.29421	0. 36,258	0.21234	0.19062	0.17300	0.19778

"Max-draw" is an abbreviation for "maximum drawdown."

performance in the two target stock markets, where "mean returns" are achieved by averaging the daily return rate. By comparing the mean returns of each strategy, we find that the enhancement strategy performs best in both markets. Therefore, this finding indicates that the proposed lead–lag strategy significantly improves the performance of the pure-01 factor strategy. Furthermore, PLL performs better than MKT in both markets in different degrees in terms of mean returns, meaning that PLL contains valuable information for investment. However, when the signals provided by the lead–lag strategy are added to the pure-01 factor strategy, the achieved enhancement strategy performs better, which implies that the value of the information contained in the lead–lag strategy is superior to that of the pure strategy. Considering the other indices listed in Table 9, the positively higher values of skewness and the Sharpe ratio in Enhan-01 displays a more desirable property for any potential investor in both markets (Cont 2010; Fievet and Sornette 2018). In addition, compared to Pure-01, the usage of the signal from the PLL significantly reduces the max-draw of Enhan-01, and thus, increases the advantage afforded by the enhancement strategy.

Next, we conduct random investment operations in each stock market during the selected 10-year period, record the average daily return rate (mean returns) and the Sharpe ratio of each operation, and then rank them in the figures. Figure 10 displays the results of 5,000 simulations, and the upper 5% of the ranked mean return value, or Sharpe ratio, is set as the threshold. When the corresponding value of one proposed strategy is above the threshold, we deem that the performance of the proposed strategy is significantly better than the benchmark at a significance level of 0.05. Recalling the performance results of Enhan-01 shown in Table 9, the mean values are 0.00036 and 0.01257 in both the U.S. and Chinese stock markets, respectively, which is higher than



(a) Ranked average returns via random operations for the CSI 300

(b) Ranked Sharpe ratios via random operations for the CSI 300

c) Ranked average returns via random operations for the S&P 500

(d) Ranked Sharpe ratios via random operations for the S&P 500

Fig. 10 The ranked mean returns and Sharpe ratios for each market via random investment operations

Table 10 The performance of Pure-02 to Pure-20 and Enhan-02 to Enhan-20 in the CSI 300

	Mean return	Std error	Min	Quartile 1	Median	Quartile 3	Max	Skewness	Kurtosis	Sharpe ratio	Max-draw
Pure-02	− 0.00010	0.00635	− 0.04265	− 0.00267	0.00006	0.00285	0.03063	− 1.01014	10.05995	− 0.31290	0.33828
Enhan-02	0.00025	0.01986	− 0.16023	− 0.00798	0.00037	0.00972	0.09742	− 0.56947	8.94921	0.17303	0.65359
Pure-03	− 0.00006	0.00658	− 0.04265	− 0.00280	0.00006	0.00308	0.03045	− 0.89533	9.10772	− 0.17437	0.30950
Enhan-03	0.00032	0.02062	− 0.14677	− 0.00803	0.00046	0.00962	0.12059	− 0.63243	9.37634	0.21623	0.62255
Pure-04	− 0.00007	0.00689	− 0.04366	− 0.00292	0.00012	0.00316	0.03191	− 0.86824	9.80510	− 0.19965	0.34254
Enhan-04	0.00023	0.02440	− 0.14135	− 0.00936	0.00028	0.01106	0.14547	− 0.25997	8.75513	0.15544	0.73340
Pure-05	− 0.00007	0.00689	− 0.04366	− 0.00292	0.00012	0.00316	0.03191	− 0.86824	9.80510	− 0.19965	0.34254
Enhan-05	0.00024	0.02434	− 0.14039	− 0.00972	0.00017	0.01109	0.15246	− 0.19857	8.27094	0.15963	0.72754
Pure-06	− 0.00007	0.00653	− 0.04611	− 0.00277	0.00005	0.00291	0.03106	− 0.98692	10.03960	− 0.22308	0.34649
Enhan-06	0.00022	0.02081	− 0.12979	− 0.00845	0.00045	0.00987	0.11703	− 0.36653	8.38719	0.16273	0.65000
Pure-07	− 0.00007	0.00669	− 0.04578	− 0.00284	0.00008	0.00301	0.03134	− 0.95963	10.08460	− 0.21202	0.34223
Enhan-07	0.00027	0.02056	− 0.13230	− 0.00828	0.00026	0.00964	0.10520	− 0.27751	8.64334	0.18308	0.60536
Pure-08	− 0.00004	0.00690	− 0.04521	− 0.00298	0.00014	0.00326	0.03281	− 0.88553	9.07309	− 0.12912	0.32702
Enhan-08	0.00022	0.02064	− 0.14743	− 0.00856	0.00030	0.01009	0.11220	− 0.57434	8.91242	0.13954	0.68266
Pure-09	− 0.00005	0.00684	− 0.04639	− 0.00281	0.00006	0.00306	0.03197	− 0.91544	9.83041	− 0.15928	0.35349
Enhan-09	0.00027	0.02034	− 0.13553	− 0.00835	0.00030	0.00986	0.10489	− 0.37692	8.96197	0.17850	0.60725
Pure-10	− 0.00005	0.00684	− 0.04637	− 0.00281	0.00007	0.00306	0.03210	− 0.91092	9.84131	− 0.15736	0.35235
Enhan-10	0.00023	0.02111	− 0.13966	− 0.00849	0.00024	0.01002	0.11000	− 0.35185	8.89457	0.15226	0.63761
Pure-11	− 0.00006	0.00684	− 0.04626	− 0.00281	0.00006	0.00305	0.03193	− 0.91458	9.81815	− 0.16033	0.35332
Enhan-11	0.00026	0.02091	− 0.13889	− 0.00844	0.00029	0.01002	0.10717	− 0.39298	9.00901	0.17365	0.62729
Pure-12	− 0.00006	0.00700	− 0.04808	− 0.00296	0.00013	0.00327	0.03244	− 0.98002	9.68904	− 0.16364	0.32964
Enhan-12	0.00023	0.02329	− 0.14660	− 0.00926	0.00035	0.01089	0.13428	− 0.38762	8.28422	0.14502	0.68820
Pure-13	− 0.00006	0.00688	− 0.04635	− 0.00293	0.00006	0.00318	0.03039	− 0.92702	9.66166	− 0.15429	0.34552
Enhan-13	0.00023	0.02030	− 0.15958	− 0.00821	0.00044	0.00992	0.10344	− 0.55992	8.89523	0.16151	0.65908
Pure-14	− 0.00001	0.00712	− 0.04879	− 0.00301	0.00014	0.00335	0.03215	− 0.91793	9.69364	− 0.02981	0.33667
Enhan-14	0.00037	0.01967	− 0.12254	− 0.00779	0.00022	0.00941	0.11605	− 0.35212	8.88372	0.23426	0.60974

Table 10 (continued)

	Mean return	Std error	Min	Quartile 1	Median	Quartile 3	Max	Skewness	Kurtosis	Sharpe ratio	Max-draw
Pure-15	− 0.00006	0.00672	− 0.04638	− 0.00286	0.00007	0.00294	0.03208	− 0.93275	10.03591	− 0.17884	0.34365
Enhan-15	0.00031	0.01940	− 0.12989	− 0.00801	0.00026	0.00921	0.10417	− 0.43907	8.76987	0.21523	0.57279
Pure-16	− 0.00004	0.00330	− 0.01903	− 0.00152	0.00005	0.00157	0.01280	− 0.75570	7.75709	− 0.19180	0.20181
Enhan-16	0.00034	0.02298	− 0.13176	− 0.00968	0.00029	0.01110	0.10692	− 0.30097	7.43824	0.20296	0.65323
Pure-17	− 0.00003	0.00425	− 0.02723	− 0.00172	0.00008	0.00188	0.01788	− 0.74147	7.64894	− 0.13885	0.21911
Enhan-17	0.00028	0.03022	− 0.18895	− 0.01235	0.00040	0.01405	0.19057	− 0.04576	8.26951	0.15887	0.81496
Pure-18	− 0.00005	0.00322	− 0.01854	− 0.00147	0.00002	0.00153	0.01262	− 0.77078	7.84633	− 0.27807	0.19892
Enhan-18	0.00025	0.02771	− 0.19643	− 0.01108	0.00050	0.01280	0.18315	− 0.14024	9.17196	0.14050	0.82105
Pure-19	− 0.00003	0.00735	− 0.05264	− 0.00303	0.00020	0.00344	0.04168	− 1.01673	11.68811	− 0.09121	0.32986
Enhan-19	0.00020	0.02401	− 0.13668	− 0.00977	0.00027	0.01125	0.13933	− 0.29755	8.15441	0.12926	0.71830
Pure-20	− 0.00008	0.00647	− 0.04293	− 0.00276	0.00008	0.00295	0.02841	− 0.94366	9.09652	− 0.24418	0.32205
Enhan-20	0.00027	0.02162	− 0.16371	− 0.00838	0.00033	0.01008	0.11411	− 0.51576	8.99948	0.19212	0.62835

Table 11 The performance of Pure-02 to Pure-20 and Enhan-02 to Enhan-20 in the S&P 500

	Mean return	Std error	Min	Quartile 1	Median	Quartile 3	Max	Skewness	Kurtosis	Sharpe ratio	Max-draw
Pure-02	0.00065	0.00822	−0.06276	−0.00294	0.00065	0.00436	0.14875	1.94494	48.89621	0.70804	0.20143
Enhan-02	0.00141	0.00982	−0.07164	−0.00383	0.00082	0.00549	0.14610	0.91193	25.80638	1.11447	0.21038
Pure-03	0.00079	0.01002	−0.06803	−0.00396	0.00075	0.00545	0.15073	0.96461	26.29571	0.74484	0.19733
Enhan-03	0.00099	0.00915	−0.06741	−0.00343	0.00078	0.00509	0.06991	−0.21806	8.44135	0.80842	0.19432
Pure-04	0.00120	0.00915	−0.06643	−0.00341	0.00080	0.00525	0.06604	−0.23041	7.92143	0.89153	0.20029
Enhan-04	0.00157	0.01020	−0.06963	−0.00378	0.00084	0.00574	0.12966	0.82719	18.82199	0.97031	0.22179
Pure-05	0.00072	0.00915	−0.07750	−0.00374	0.00090	0.00531	0.05338	−0.49598	8.75788	0.91645	0.22961
Enhan-05	0.00108	0.00893	−0.06946	−0.00335	0.00087	0.00509	0.06679	−0.23050	8.71492	0.77433	0.22811
Pure-06	0.00068	0.00958	−0.06913	−0.00376	0.00078	0.00532	0.07285	−0.20946	8.75767	0.70703	0.21381
Enhan-06	0.00095	0.00905	−0.06744	−0.00328	0.00078	0.00508	0.06658	−0.23617	8.50229	0.82611	0.20942
Pure-07	0.00090	0.00970	−0.06528	−0.00368	0.00081	0.00533	0.08442	−0.02074	8.84971	0.84558	0.20201
Enhan-07	0.00112	0.00888	−0.06636	−0.00333	0.00089	0.00502	0.06509	−0.22214	8.43601	0.88448	0.19332
Pure-08	0.00070	0.00998	−0.06530	−0.00371	0.00073	0.00539	0.15600	1.24370	30.69485	0.84949	0.20683
Enhan-08	0.00096	0.00899	−0.06576	−0.00341	0.00079	0.00519	0.04950	−0.36582	7.50026	1.06967	0.18900
Pure-09	0.00093	0.00874	−0.06590	−0.00334	0.00080	0.00506	0.04855	−0.38893	7.52609	0.81809	0.19038
Enhan-09	0.00103	0.00996	−0.06775	−0.00382	0.00083	0.00534	0.12956	0.82278	19.47913	0.91026	0.20666
Pure-10	0.00094	0.00874	−0.06600	−0.00333	0.00082	0.00504	0.04869	−0.38762	7.53035	0.82040	0.19076
Enhan-10	0.00106	0.00992	−0.06779	−0.00376	0.00087	0.00536	0.13221	0.87577	20.65819	0.93521	0.21022
Pure-11	0.00093	0.00874	−0.06590	−0.00334	0.00080	0.00506	0.04855	−0.38912	7.52605	0.81799	0.19038
Enhan-11	0.00108	0.00994	−0.06770	−0.00380	0.00090	0.00535	0.12949	0.82710	19.53265	0.91759	0.20667
Pure-12	0.00093	0.00908	−0.06667	−0.00347	0.00082	0.00522	0.07066	−0.20196	8.68972	0.89803	0.21751
Enhan-12	0.00104	0.00971	−0.07145	−0.00383	0.00077	0.00531	0.14379	0.88634	25.53199	1.04319	0.21414
Pure-13	0.00072	0.00979	−0.08557	−0.00391	0.00068	0.00540	0.06593	−0.49401	10.01950	0.79182	0.23344
Enhan-13	0.00092	0.00853	−0.06502	−0.00322	0.00073	0.00494	0.04622	−0.43357	7.57277	0.80047	0.20046
Pure-14	0.00078	0.00931	−0.07227	−0.00360	0.00084	0.00524	0.08287	−0.12111	10.06141	0.82279	0.22017
Enhan-14	0.00096	0.00844	−0.06419	−0.00323	0.00074	0.00497	0.04987	−0.35949	7.99071	0.96043	0.19602
Pure-15	0.00068	0.00928	−0.07368	−0.00366	0.00073	0.00523	0.06872	−0.28238	9.17828	0.84706	0.22266

Table 11 (continued)

	Mean return	Std error	Min	Quartile 1	Median	Quartile 3	Max	Skewness	Kurtosis	Sharpe ratio	Max-draw
Enhan-15	0.00097	0.00828	−0.06280	−0.00305	0.00081	0.00489	0.04454	−0.38062	7.76812	0.95942	0.19299
Pure-16	0.00086	0.00842	−0.06132	−0.00334	0.00074	0.00501	0.04532	−0.36673	6.95236	0.85057	0.18187
Enhan-16	0.00117	0.00991	−0.07078	−0.00386	0.00078	0.00555	0.13700	0.73753	20.63134	0.86131	0.21171
Pure-17	0.00115	0.00937	−0.06884	−0.00356	0.00083	0.00549	0.05036	−0.34672	7.43454	0.82053	0.21922
Enhan-17	0.00159	0.01026	−0.07222	−0.00386	0.00105	0.00593	0.14505	0.72666	21.73044	0.83085	0.24530
Pure-18	0.00109	0.00838	−0.05741	−0.00326	0.00078	0.00497	0.04304	−0.32879	6.56390	0.80609	0.18316
Enhan-18	0.00139	0.00974	−0.07954	−0.00390	0.00093	0.00557	0.09149	−0.08135	10.81406	0.87043	0.21145
Pure-19	0.00101	0.00980	−0.08135	−0.00369	0.00083	0.00535	0.11535	0.28924	16.40898	0.96011	0.25839
Enhan-19	0.00117	0.00927	−0.07072	−0.00344	0.00070	0.00525	0.06954	−0.19947	8.86839	0.89946	0.21460
Pure-20	0.00088	0.00835	−0.05780	−0.00329	0.00074	0.00486	0.04097	−0.38017	6.80294	0.79938	0.19112
Enhan-20	0.00101	0.00999	−0.07691	−0.00387	0.00078	0.00549	0.13801	0.98341	24.03536	0.85588	0.23047

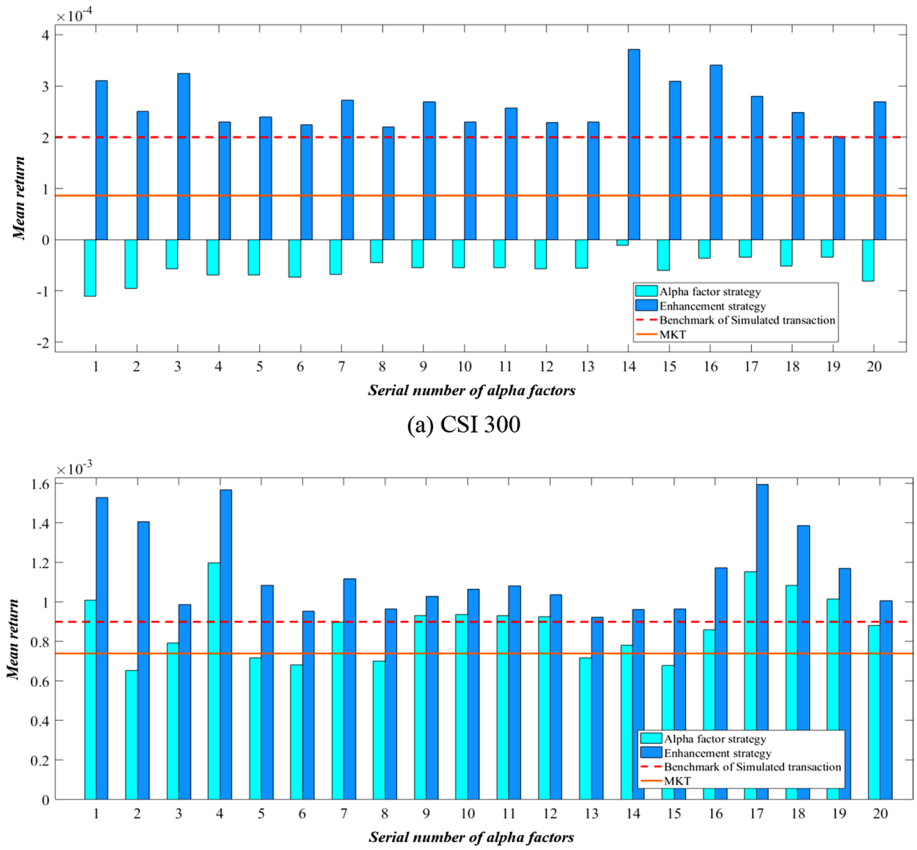


Fig. 11 Performance of each selected alpha strategy and its enhancement strategy in each market

the corresponding highlighted thresholds reported in Panels (a) and (c) of Fig. 10. Concerning the Sharpe ratio, a similar result holds.

Following this analytical process, the performance of the remaining 19 alpha-factor strategies and their corresponding enhancement strategies is also tested. All performance results are listed in Tables 10 and 11 for the two stock markets. To facilitate comparison, Fig. 11 displays the mean return of each selected alpha factor in each market as well as for the two types of benchmarks. Here, it is not difficult to find that the benchmark from random investment operations is higher than that from MKT. According to Fig. 11, we can obtain the following findings: (1) all of the enhancement strategies perform better than the two benchmarks in both markets, demonstrating the usefulness of the proposed strategies; and (2) almost all of the enhancement strategies perform better than the corresponding pure alpha-factor strategies, illustrating that the signal provided by the lead–lag strategy that we proposed does improve the performance of pure alpha-factor strategies in most cases.

Discussion

Section 4.2.2, Fig. 8, and Table 7 show that the overall prediction accuracy is significantly higher than 50% (i.e., better than a random guess) in each case, implying

that the stocks with the lead–lag effect provide useful information for prediction and strategy design. However, a degree of more than 50% is not exceptionally high, particularly in the CSI 300; thus, some stock pairs perform worse than a random guess in each prediction period. Inspired by this result, we provide a refined process in which the stock pairs (i.e., the detected lead–lag relationship with effect) with a prediction accuracy less than 50% are eliminated from the selected stock set. Accordingly, the refined process makes the set of stock pairs with the lead–lag effect more minor by deleting those with inferior prediction performance in the trained data. Then, putting the refined process into the enhancement strategy proposed above, the so-called refined strategy is proposed and the test of its performance is like what was done in Sect. 5.3.

All performance results are listed in Table 12 for each market, and Fig. 12 displays the mean returns of each enhancement strategy and its corresponding refined strategy in each market with the two selected benchmarks. According to Fig. 12, the refined strategies have different degrees of improvement in profit over the original enhancement strategies in the CSI 300, indicating that the refined process does provide more practical information for investing in the CSI 300. However, for the S&P 500, most refined strategies outperform the original enhancement strategies, but some perform worse than the original enhancement strategy, especially when the original strategy is already very profitable. This result implies that the refined process does not always work better than the original process, possibly because some helpful information may be dropped during the refinement process.

Furthermore, in the risk analysis, Table 12 shows that the refined strategy generally improves the Sharpe ratio and reduces the maximum draw-down rate in the CSI 300. In contrast, improvement is not evident in the S&P 500. These results indicate that the refinement process effectively discards risky lead–lag signals, but its performance depends on the application scenario. Overall, the refined strategy is more suitable for the CSI 300, but for the S&P 500, it may serve as an alternative to the original enhancement strategy.

Conclusions and future work

The power-law distribution is often observed in human activity, which explains its widespread existence in stock markets. Interestingly, this study finds that the number of accumulated lead–lag days between stock pairs fulfills the power-law distribution in both the U.S. and Chinese stock markets based on 10 years of data. Because the power-law distribution features a heavy tail, this study also formally defines the lead–lag effect via statistical testing and then proposes a new method for detecting stock pairs characterized by the previously defined lead–lag effect. Robustness and the functions of the parameters embedded in the detection method are tested and explored. As an application, a PLL investment strategy is first proposed based on stock pairs identified with the lead–lag effect. Although the proposed lead–lag strategy can beat a naïve buy-and-sell strategy, its leading edge is too limited to be satisfied. To this end, enhancement strategies are also designed by integrating the lead–lag strategy in the selected basic alpha-factor strategies. Then, a series of validations are conducted on 20 different alpha factors to guarantee sound results. The results demonstrate that the enhancement strategy significantly

Table 12 The performance of Refined-01 to Pefined-20 in both markets

	Mean return	Std error	Min	Quartile 1	Median	Quartile 3	Max	Skewness	Kurtosis	Sharpe ratio	Max-draw
CSI 300											
Refined-01	0.00043	0.01209	−0.13808	−0.00376	0.00023	0.00430	0.10412	−0.15767	19.11770	0.58097	0.34865
Refined-02	0.00030	0.01225	−0.20345	−0.00381	0.00023	0.00438	0.06775	−1.91137	39.55900	0.45806	0.43690
Refined-03	0.00034	0.01280	−0.23112	−0.00394	0.00029	0.00488	0.08230	−2.51640	53.00308	0.46194	0.43542
Refined-04	0.00027	0.01178	−0.13371	−0.00396	0.00026	0.00461	0.06859	−0.65453	16.31264	0.53790	0.41259
Refined-05	0.00028	0.01518	−0.36526	−0.00428	0.00027	0.00510	0.11025	−5.92353	147.26431	0.29302	0.59627
Refined-06	0.00024	0.01540	−0.38098	−0.00455	0.00028	0.00523	0.10242	−5.98321	161.46500	0.35753	0.54839
Refined-07	0.00029	0.01404	−0.25240	−0.00426	0.00030	0.00516	0.08402	−2.42830	51.45161	0.45391	0.48835
Refined-08	0.00023	0.01586	−0.39749	−0.00439	0.00025	0.00529	0.11566	−6.45872	170.19805	0.29923	0.57653
Refined-09	0.00027	0.01430	−0.27507	−0.00443	0.00031	0.00514	0.08186	−2.98920	64.45716	0.41838	0.51011
Refined-10	0.00032	0.01234	−0.21492	−0.00367	0.00022	0.00430	0.09186	−2.20615	47.95994	0.45276	0.45530
Refined-11	0.00034	0.01298	−0.24336	−0.00390	0.00026	0.00460	0.09726	−2.37009	60.10964	0.51303	0.43507
Refined-12	0.00040	0.01259	−0.23057	−0.00374	0.00024	0.00455	0.09099	−2.25038	54.56784	0.61923	0.40366
Refined-13	0.00030	0.01234	−0.22419	−0.00380	0.00025	0.00452	0.11710	−2.36556	56.52437	0.58254	0.43353
Refined-14	0.00044	0.01214	−0.14980	−0.00403	0.00025	0.00445	0.08214	−0.61242	19.31622	0.65979	0.39469
Refined-15	0.00040	0.01136	−0.21716	−0.00340	0.00024	0.00407	0.08660	−2.55120	63.68843	0.63325	0.41064
Refined-16	0.00039	0.01228	−0.24477	−0.00368	0.00024	0.00437	0.07994	−3.05135	73.26713	0.58568	0.43332
Refined-17	0.00033	0.01321	−0.25506	−0.00383	0.00025	0.00487	0.09458	−2.87147	66.50622	0.55404	0.49018
Refined-18	0.00027	0.01379	−0.33962	−0.00405	0.00028	0.00465	0.08289	−6.27706	158.92483	0.30812	0.53904
Refined-19	0.00029	0.01499	−0.31423	−0.00452	0.00027	0.00529	0.10389	−3.74322	88.56693	0.41267	0.49954
Refined-20	0.00032	0.01279	−0.26111	−0.00367	0.00025	0.00437	0.09343	−3.71691	81.73846	0.46131	0.48238
S&P 500											
Refined-01	0.00120	0.00927	−0.06280	−0.00343	0.00076	0.00527	0.05677	−0.25384	7.88097	0.93502	0.19632
Refined-02	0.00131	0.00910	−0.08505	−0.00320	0.00072	0.00494	0.05608	−0.44247	10.49970	0.96606	0.21925
Refined-03	0.00127	0.00900	−0.06692	−0.00323	0.00074	0.00512	0.05558	−0.31930	8.20261	0.93400	0.20136
Refined-04	0.00117	0.00937	−0.06606	−0.00330	0.00074	0.00526	0.05662	−0.29329	8.15174	0.84820	0.21106
Refined-05	0.00135	0.00846	−0.05400	−0.00307	0.00074	0.00489	0.05147	−0.22583	7.37853	1.04520	0.18261

Table 12 (continued)

	Mean return	Std error	Min	Quartile 1	Median	Quartile 3	Max	Skewness	Kurtosis	Sharpe ratio	Max-draw
Refined-06	0.00125	0.00918	− 0.06375	− 0.00337	0.00078	0.00526	0.05350	− 0.31981	7.32608	0.93895	0.19670
Refined-07	0.00125	0.00941	− 0.06717	− 0.00336	0.00076	0.00522	0.05445	− 0.31532	8.16291	0.94691	0.19897
Refined-08	0.00126	0.00914	− 0.05901	− 0.00326	0.00075	0.00531	0.05471	− 0.28155	7.60900	1.07543	0.19805
Refined-09	0.00117	0.00943	− 0.07142	− 0.00338	0.00079	0.00542	0.05784	− 0.41517	7.87385	0.91230	0.23837
Refined-10	0.00140	0.00880	− 0.07860	− 0.00305	0.00073	0.00501	0.05520	− 0.39093	9.52000	1.02659	0.20285
Refined-11	0.00122	0.00940	− 0.06740	− 0.00318	0.00076	0.00521	0.05717	− 0.21821	8.46796	0.88664	0.21298
Refined-12	0.00120	0.00968	− 0.08762	− 0.00341	0.00081	0.00539	0.05901	− 0.55531	9.55022	0.96708	0.24213
Refined-13	0.00130	0.00882	− 0.05322	− 0.00318	0.00077	0.00504	0.05297	− 0.21640	7.41976	1.06135	0.19842
Refined-14	0.00128	0.00868	− 0.06482	− 0.00312	0.00077	0.00501	0.05642	− 0.25954	7.99324	0.93207	0.18052
Refined-15	0.00127	0.00894	− 0.07473	− 0.00309	0.00074	0.00516	0.05317	− 0.35336	8.45123	1.00452	0.19280
Refined-16	0.00122	0.00958	− 0.06828	− 0.00349	0.00080	0.00540	0.05473	− 0.38037	7.72348	0.90882	0.24047
Refined-17	0.00124	0.00900	− 0.06820	− 0.00320	0.00077	0.00516	0.05031	− 0.35907	8.14505	1.04566	0.22087
Refined-18	0.00138	0.00890	− 0.08204	− 0.00320	0.00075	0.00506	0.05620	− 0.51347	9.56916	1.01977	0.20690
Refined-19	0.00121	0.00917	− 0.05942	− 0.00326	0.00075	0.00517	0.05396	− 0.29153	7.69752	0.94226	0.19388
Refined-20	0.00124	0.00903	− 0.06379	− 0.00318	0.00074	0.00504	0.04752	− 0.38454	7.65719	0.91660	0.20369

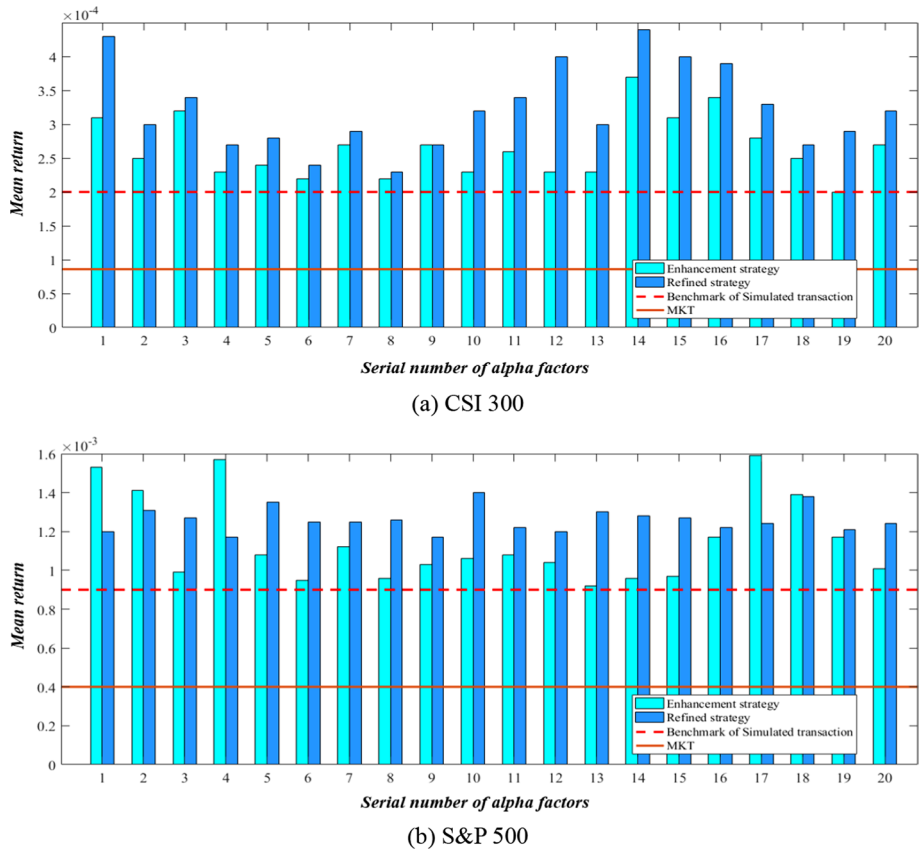


Fig. 12 Performance of each enhancement strategy and its refined strategies in each market

improves the performance of the basic alpha-factor strategies and the PLL strategy in most cases.

In theory, the discovered power-law distribution implies that the lead–lag phenomenon common in stock markets is attributable not only to random factors but is also influenced by human behaviors such as irrationality, herding, gaming behavior, and many others. Importantly, this finding provides new evidence in support of behavioral finance theory. The proposed detection method can be considered solid and credible because it originates from the principle of statistical testing, contributing to the existing methods for detecting the lead–lag phenomenon or effect. In practice, because the lead–lag effect is demonstrated in this study to provide effective information, it will benefit the designing of innovative and effective investment strategies that are especially suitable for low-frequency data due to its maneuverability. The idea of an enhancement strategy (i.e., a basic strategy supplemented by the lead–lag strategy) provides investors with a new framework for strategy design with potentially positive back-tested performance and practicality.

Our study does have some limitations. Although we selected two representative stock markets as the targets for this examination, the analysis and validation of additional

stock markets is required. In future work, we will study the characteristics of the lead–lag phenomenon in different emerging markets at different stages of economic development. Many previous studies have confirmed that there are more opportunities for profit in emerging markets than in mature markets; thus, if our proposed strategy will be effective in various emerging markets remains a question of interest. Although our proposed lead–lag strategy and enhancement strategies exhibit a significant improvement compared to the selected benchmarks, the type of basic investment strategy (i.e., the alpha strategy) is relatively singular in this work. In future work, other investment strategies can be selected as the basic strategy that will be enhanced by implementing the lead–lag strategy to design more competitive stock market investment strategies. As this is a preliminary examination into these two directions, more colorful findings and profitable investment strategies are welcome in the future.

Appendix 1: The designed 20 alpha factors and their expressions

We present the designed 20 alpha factors in details here by providing their formulaic expressions. Specifically, Table 13 shows symbolic descriptions of the variables related to the collected data, and Table 14 shows the operators and functions adopted in these formulaic expressions of alpha factors.

Accordingly, the designed 20 alpha formulas are expressed one by one as below.

- Alpha-01: $\text{Rank}(-1 * \text{Correlation}(Open, Volume, 20)) - 0.5$
- Alpha-02: $\text{Rank}(-1 * \text{Correlation}(\text{Rank}(\text{Delta}(Volume, 10)), \text{Rank}(\frac{Close-Open}{Open}), 20)) - 0.5$
- Alpha-03: $\text{Rank}(-1 * \text{Ts\_Rank}(Low, 20)) - 0.5$
- Alpha-04: $\text{Rank}(-1 * \text{Correlation}(Open, Volume, 20)) - 0.5$
- Alpha-05: $\text{Rank}\big(\text{Sign}(\text{Delta}(Volume, 1)) * (-1 * \text{Delta}(Close, 20))\big) - 0.5$
- Alpha-06: $0.5 - 1 * \text{Rank}(\text{Covariance}(\text{Rank}(Close), \text{Rank}(Volume), 20))$
- Alpha-07: $\text{Rank}((-1 * \text{Rank}(\text{Delta}(Returns, 10))) * \text{Correlation}(Open, Volume, 20)) - 0.5$
- Alpha-08: $0.5 - 1 * \text{Rank}(\text{Correlation}(\text{Rank}(High), \text{Rank}(Volume), 20))$
- Alpha-09: $0.5 - 1 * \text{Rank}(\text{covariance}(\text{Rank}(High), \text{Rank}(Volume), 20))$
- Alpha-10: $0.5 - 1 * \text{Rank}(\text{Correlation}(\text{Ts\_Rank}(Volume, 5), \text{Ts\_Rank}(High, 5), 15))$
- Alpha-11: $\text{Rank}(\text{Correlation}(Adv20, Low, 5) + ((High + Low)/2) - Close) - 0.5$
- Alpha-12: $\text{Rank}(\text{Correlation}(\text{Delay}((Open - Close), 1), Close, 20)) + \text{Rank}((Open - Close)) - 0.5$
- Alpha-13: $\text{Rank}(-1 * \text{Rank}(\text{Std}(High, 20)) * \text{Correlation}(High, Volume, 20)) - 0.5$
- Alpha-14: $\text{Rank}(-1 * \text{Correlation}(High, \text{Rank}(Volume), 20)) - 0.5$
- Alpha-15: $\text{Rank}(-1 * \text{Delta}((2 * Close - Low - High)/(Close - Low), 20)) - 0.5$
- Alpha-16: $\text{Rank}(\text{Correlation}((Low - Close) * (Open^5), (Low - High) * (Close^5), 20)) - 0.5$
- Alpha-17: $0.5 - \text{Rank}(\text{Correlation}(\text{Rank}(\frac{Close-\text{Min}(Low,12)}{\text{Max}(High,12)-\text{Min}(Low,12)}), \text{Rank}(Volume), 6))$
- Alpha-18: $\text{Rank}(\text{Correlation}\big(Close - Open, High - Low\big), 20) - 0.5$
- Alpha-19: $\text{Rank}(2 - \text{Rank}(\text{Std}(Returns, 7)/\text{Std}(Returns, 20)) - \text{Rank}(\text{Delta}(Close, 7))) - 0.5$
- Alpha-20: $\text{Rank}(\text{Ts\_Rank}(Volume/Adv20, 20) * \text{Ts\_Rank}(-1 * \text{Delta}(Close, 7), 7)) - 0.5$

Table 13 Symbolic descriptions of the data variables

Name	Descriptions
Open	Daily open price
Close	Daily close price
High	Daily high price
Low	Daily low price
Volume	Daily trading volume
Returns	Daily returns
Adv20	Average daily trading volume in past 20 days

Table 14 Operators and functions in the formulaic expressions

Operator and function	Description	Type
+, −, *, /, ^	Add, subtract, multiply, divide, power	
Correlation(x, y, n)	Correlation of the variables x and y for the past n days	Scalar
Covariance(x, y, n)	Covariance of the variables x and y for the past n days	Scalar
Delay(x, n)	x value of n days ago	Scalar
Delta(x, n)	x value of current day minus its value of n days ago	Scalar
Rank(x)	Rank value of the variable x of all the stocks and the achieved rank value is transformed into the range between 0.0 and 1.0. For example, Rank([20.2, 15.6, 10.0, 5.7, 50.2, 18.4]) is [0.8, 0.4, 0.2, 0.0, 1.0, 0.6]	Vector
Sign(x)	1 if $x>0$, −1 if $x<0$, and 0 if $x=0$	Scalar
Std(x, n)	Standard deviation of the variable x for the past n days	Scalar
Ts_Rank(x, n)	Rank the values of the variable x over the past d days and then all the rank values are transformed into the range between 0.0 and 1.0. Finally, the rank value of the variable x in current day is returned	Scalar
Max(x, n)	The maximum value of the variable x over the past d days	Scalar
Min(x, n)	The minimum value of the variable x over the past d days	Scalar

Appendix 2: List of ababreviations

See Table 15.

Table 15 Abbreviations and their full names

Abbreviation	Full name
CSI 300	China Securities 300 index
S&P 500	Standard & Poor 500 Index
K–S test	Kolmogorf–Smirnoff test
MKT	The naïve buy-and-hold investment strategy
PLL	The pure lead–lag strategy
Pure-01	The pure alpha strategy No. 01
Enhan-01	The enhancement strategy based on the alpha strategy No. 01
Max-draw	Maximum drawdown
DM	Diameter of a network
DS	Density of a network
PL	Average of path length
ND	Average of node degree
CC	Clustering coefficient

Author contributions
YL: Conceptualization, Methodology, Formal analysis, and Writing - Original Draft. TW: Methodology, Software, Formal analysis, and Writing - Original Draft. BS: Writing - Review & Editing, Supervision, Validation. CL: Visualization, Software, Validation, and Data Curation. All authors read and approved the final manuscript.

Funding
This work was supported by the National Natural Science Foundation of China (72171059, 71771041), the Fundamental Research Funds for the Central Universities (FRFCU5710000220) and the Natural Science Foundation of Heilongjiang Province, China (No. YQ2020G003).

Availability of data and material
Data and codes are available at https://github.com/liuchaos03/Power-law-distribution-Lead-lag-effect-and-Investment-strategies-in-Stock-Markets.

Competing interests
The authors declare that they have no competing interests.

Author details
[1]School of Economics and Management, Harbin Institute of Technology, Harbin 150001, People's Republic of China.
[2]School of Business Administration, Northeastern University, Shenyang 110169, People's Republic of China.

Received: 5 October 2021 Accepted: 6 April 2022
Published online: 20 May 2022

References

Balatti M, Brooks C, Kappou K (2017) Fundamental indexation revisited: new evidence on alpha. Int Rev Financ Anal 51:1–15
Barabási AL, Albert R (1999) Emergence of scaling in random networks. Science 286(5439):509–512
Barabási AL, Bonabeau E (2003) Scale-free networks. Sci Am 288(5):60–69
Berggrun L, Cardona E, Lizarzaburu E (2020) Profitability of momentum strategies in Latin America. Int Rev Financ Anal 70:101502
Buccheri G, Corsi F, Peluso S (2019) High-frequency lead–lag effects and cross-asset linkages: a multi-asset lagged adjustment model. J Bus Econ Stat. https://doi.org/10.1080/07350015.2019.1697699
Campajola C, Lillo F, Tantari D (2020) Unveiling the relation between herding and liquidity with trader lead–lag networks. Quant Finance 20(11):1765–1778
Casgrain P, Jaimungal S (2019) Trading algorithms with learning in latent alpha models. Math Financ 29(3):735–772
Clauset A, Shalizi CR, Newman MEJ (2009) Power-law distributions in empirical data. SIAM Rev 51(4):661–703
Conlon T, Cotter J, Gencay R (2018) Long-run wavelet-based correlation for financial time series. Eur J Oper Res 271(2):676–696
Cont R (2010) Empirical properties of asset returns: stylized facts and statistical issues. Quant Finance 1(2):223–236
Coronel-Brizio HF, Hernández-Montoya AR (2010) The Anderson-Darling test of fit for the power-law distribution from left-censored samples. Physica A Stat Mech Appl 389(17):3508–3515
Curme C, Tumminello M, Mantegna RN, Stanley HE, Kenett DY (2015) Emergence of statistically validated financial intraday lead–lag relationships. Quant Finance 15(8):1375–1386
Dao TM, Mcgroarty F, Urquhart A (2018) Ultra-high-frequency lead–lag relationship and information arrival. Quant Finance 18(5):725–735
Deev O, Lyócsa Š (2020) Connectedness of financial institutions in Europe: a network approach across quantiles. Phys A Stat Mech Appl 550:124035–124041
Eisdorfer A, Goyal A, Zhdanov A (2019) Equity misvaluation and default options. J Financ 74(2):845–898
Fama EF, French KR (2012) Size, value, and momentum in international stock returns. J Financ Econ 105(3):457–472
Fama EF, French KR (1998) Value versus growth: the international evidence. J Financ 53:1975–1999
Fama EF, French KR (2015) A five-factor asset pricing model. J Financ Econ 116(1):1–12
Fama EF, French KR (2016) Dissecting anomalies with a five-factor model. Rev Financ Stud 29(1):69–103
Fievet L, Sornette D (2018) Decision trees unearth return sign predictability in the S&P 500. Quant Finance 18(11):1797–1814
Flori A, Regoli D (2021) Revealing pairs-trading opportunities with long short-term memory networks. Eur J Oper Res. https://doi.org/10.1016/j.ejor.2021.03.009
Fonseca DJ, Zaatour R (2017) Correlation and lead–lag relationships in a hawkes microstructure model. J Futur Mark 37(3):260–285
Gabaix X, Gopikrishnan P, Plerou V, Stanley HE (2003) A theory of power-law distributions in financial market fluctuations. Nature 423(6937):267–270
Gong CC, Ji SD, Su LL, Li SP, Ren F (2016) The lead–lag relationship between stock index and stock index futures: a thermal optimal path method. Physica A 444:63–72
Gupta K, Chatterjee N (2020) Selecting stock pairs for pairs trading while incorporating lead–lag relationship. Phys A Stat Mech Appl 551:124103
Harvey CR, Liu Y, Zhu H (2016) … and the cross-section of expected returns. Rev Financ Stud 29(1):5–68
Hou K, Xue C, Zhang L (2015) Digesting anomalies: an investment approach. Rev Financ Stud 28(3):650–705
Huang WQ, Zhuang XT, Yao S (2009) A network analysis of the Chinese stock market. Physica A 388(14):2956–2964
Huth N, Abergel F (2014) High frequency lead/lag relationships—empirical facts. J Empir Financ 26:41–58
Jiang T, Bao S, Li L (2019) The linear and nonlinear lead–lag relationship among three SSE 50 Index markets: the index futures, 50ETF spot and options markets. Physica A Statis Mech Appl 525:878–893

Jong DF, Nijman T (1997) High frequency analysis of lead–lag relationships between financial markets. J Empir Financ 4(2–3):259–277
Kakushadze Z (2016) 101 formulaic alphas. Wilmott 2016(84):72–81
Kuiper NH (1960) Tests concerning random points on a circle. Proc Ser A 63(1):38–47
Kobayashi T, Takaguchi T (2018) Social dynamics of financial networks. EPJ Data Sci 7(1):15
Krauss C (2017) Statistical arbitrage pairs trading strategies: review and outlook. J Econ Surv 31(2):513–545
Kumar S, Deo N (2012) Correlation and network analysis of global financial indices. Phys Rev E 86(2):026101
Li Y, Liu C, Wang T, Sun B (2021) Dynamic patterns of daily lead–lag networks in stock markets. Quant Finance 21(12):2055–2068
Liu J, Stambaugh RF, Yuan Y (2019) Size and value in china. J Financ Econ 134(1):48–69
Makarov I, Plantin G (2015) Rewarding trading skills without inducing gambling. J Financ 70(3):925–962
Malevergne Y, Pisarenko V, Sornette D (2011) Testing the Pareto against the lognormal distributions with the uniformly most powerful unbiased test applied to the distribution ofcities. Phys Rev E 83(3):
Massey FJ Jr (1951) The Kolmogorov–Smirnov test for goodness of fit. J Am Stat Assoc 46(253):68–78
Newman ME, Strogatz SH, Watts DJ (2001) Random graphs with arbitrary degree distributions and their applications. Phys Rev E 64(2):026118
Peralta G, Zareei A (2016) A network approach to portfolio selection. J Empir Financ 38:157–180
Rickles D (2011) Econophysics and the complexity of financial markets. In: Hooker C (ed) Philosophy of complex systems. North-Holland, Amsterdam, pp 531–565
Scherbina A, Schlusche B (2020) Follow the leader: using the stock market to uncover information flows between firms. Rev Finance 24(1):189–225
Scholz FW, Stephens MA (1987) K-sample Anderson-Darling tests. J Am Stat Assoc 82(399):918–924
Sharpe WF (1964) Capital asset prices: a theory of market equilibrium under conditions of risk. J Financ 19(3):425–442
Shen D, Zhang Y, Xiong X, Zhang W (2017) Baidu index and predictability of Chinese stock returns. Financ Innov 3(1):1–8
Stübinger J (2019) Statistical arbitrage with optimal causal paths on high-frequency data of the S&P 500. Quant Finance 19(6):921–935
Toda AA (2012) The double power law in income distribution: explanations and evidence. J Econ Behav Org 84(1):364–381
Tóth B, Kertész J (2006) Increasing market efficiency: Evolution of cross-correlations of stock returns. Physica A 360(2):505–515
Volz E (2004) Random networks with tunable degree distribution and clustering. Phys Rev E 70(5):056115
Xia L, You D, Jiang X, Chen W (2018) Emergence and temporal structure of Lead-Lag correlations in collective stock dynamics. Phys A Statis Mech Appl 502:545–553
Xiong X, Cui Y, Yan X, Liu J, He S (2020) Cost-benefit analysis of trading strategies in the stock index futures market. Financ Innov 6(1):1–17
Zeng K, Atta Mills EFE (2021) Can economic links explain lead–lag relations across firms? Int J Finance Econ. https://doi.org/10.1002/ijfe.2480
Zhang W, Yan K, Shen D (2021) Can the Baidu Index predict realized volatility in the Chinese stock market? Financ Innov 7(1):1–31

Publisher's Note