Quantitative statistical analysis of order-splitting behavior of individual trading accounts in the Japanese stock market over nine years

Yuki Sato and Kiyoshi Kanazawa*

Department of Physics, Graduate School of Science, Kyoto University, Kyoto 606-8502, Japan

(Received 5 January 2023; accepted 1 September 2023; published 8 November 2023)

Econophysics aims to understand the macroscopic behavior of financial markets from the underlying microscopic decision-making dynamics. In particular, the order splitting of large metaorders is one of the most important trading strategies in this literature: while traders have large potential metaorders, they split the large orders into small pieces (called child orders) to minimize market impact. This strategic behavior is believed to be important because it is a promising candidate for the microscopic origin of the long-range correlation (LRC) in the persistent order flow. Indeed, Lillo, Mike, and Farmer (LMF) [Phys. Rev. E **71**, 066122 (2005)] introduced a simple microscopic model of the order-splitting traders to predict the asymptotic behavior of the LRC from the microscopic dynamics, even quantitatively. The plausibility of this scenario has been investigated by Tóth *et al.* [J. Econ. Dyn. Control **51**, 218 (2015)] at a qualitative level. However, no solid support has been presented yet on the quantitative prediction by the LMF model in the lack of large microscopic data sets. In this paper, we have provided a quantitative statistical analysis of the order-splitting behavior at the level of each trading account. We analyze a large data set of the Tokyo stock exchange (TSE) market over nine years, including the account data of traders (called *virtual servers*). The virtual server is a unit of trading accounts in the TSE market, and we can effectively define the trader IDs by an appropriate preprocessing. We apply a strategy clustering to individual traders in terms of market orders to identify the order-splitting traders and the random traders. The length distribution of metaorders is empirically estimated for each stock every year. For most of the stocks, we find that the metaorder length distribution obeys power laws with exponent α, such that $P(L) \propto L^{-\alpha-1}$ with the metaorder length L, as theoretically assumed in the LMF model. By analyzing the sign correlation of order flow $C(\tau) \propto \tau^{-\gamma}$, we draw the scatterplot between α and γ, directly confirming the LMF prediction $\gamma \approx \alpha - 1$. Furthermore, we discuss how to estimate the total number of splitting traders only from public data via the autocorrelation function prefactor formula in the LMF model. Our work provides quantitative evidence of the LMF model, strongly supporting the order-splitting hypothesis as the origin of LRC.

DOI: 10.1103/PhysRevResearch.5.043131

I. INTRODUCTION

The ultimate goal of statistical physics is to reveal the macroscopic behaviors of physical systems from their microscopic dynamics, and physicists have broadly applied this concept to interdisciplinary topics, such as financial markets, beyond traditional physics [1–4]. Recently, econophysicists have greatly benefited from high-frequency financial data on the microscopic level of individual traders [5–8]. In this paper, we focus on the microscopic origin of the long-range correlation (LRC) of the order flow by providing a systematic statistical analysis of a large comprehensive data set on the level of individual trading accounts.

In recent financial markets, traders are required to submit limit orders or market orders for their trading activities. The limit order is an option to show the traders' potential will to buy or sell the stock by specifying their prices in advance. All the limit orders are collected as the limit-order book, which displays the current potential prices for transactions. Limit-order submissions are called *liquidity provision* in the economic context, and they are highly appreciated because they stabilize the market. On the other hand, if traders wish to transact immediately, they can submit market orders to buy or sell the stock at the best price (i.e., the highest bid or lowest ask price). Market-order submissions are called *liquidity consumption*, in contrast to limit-order submissions. In other words, financial markets are composed of the flows of limit orders and market orders, and the main target of this paper is the market-order flow, particularly regarding its persistence.

The market-order flow exhibits strong persistence in financial markets: the buy (sell) market orders tend to follow another buy (sell) market order for a long time. In other words, once you observe a buy (sell) order, you are more likely to observe buy (sell) orders in the future (e.g., a typical order-sign series is given by $\{\epsilon(t)\}_t = \{+1, +1, -1, +1, +1, +1, \dots\}$, where $\epsilon(t) = +1$ $[\epsilon(t) = -1]$ denotes a buy (sell) market order). More quantitatively, the power-law decay of the sign autocorrelation function (ACF) characterizes this

*kiyoshi@scphys.kyoto-u.ac.jp

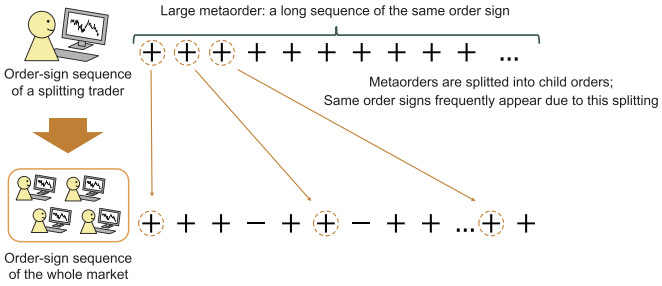

FIG. 1. Schematic of the order-splitting hypothesis. In this hypothesis, several traders hold large latent orders (called metaorders) and split them into small child orders. The plus (minus) sign "+" ("−") represents a buy (sell) market order. By definition, the child orders have the same order sign; thus, the order-sign sequence of the whole market exhibits a long memory due to this order splitting.

phenomenon [3,9–12], called the long-range correlation (LRC) in this paper:

$$C(\tau) := \lim_{t\to\infty} \langle \epsilon(t)\epsilon(t+\tau)\rangle \approx \frac{c_0}{\tau^\gamma}, \quad \gamma \in (0,1) \tag{1}$$

for a large time lag $\tau \gg 1$ with the characteristic power-law exponent γ and the prefactor c_0. Here $\langle A\rangle$ denotes the ensemble average of the stochastic variable A. This LRC is ubiquitously observed in various financial markets, such as stocks [9,10,13–15], foreign exchange (FX) [16], and cryptocurrency markets [17], and therefore it is believed to be an essential of the financial market microstructure.

What, then, is the microscopic origin of the LRC? One of the most promising hypotheses is the order-splitting behavior at the level of individual traders [3] (see Fig. 1 as a schematic). According to this hypothesis, several traders hold large latent orders (called *metaorders*). Typically, the size of a metaorder is much larger than the revealed liquidity on the order book, and therefore the traders have no choice but to split the metaorder into a long sequence of small market orders (called *child orders*) to minimize the transaction cost (called the *market impact*), naturally leading to the LRC in the sign ACF. From the standpoint of empirical analyses, this scenario has been supported qualitatively. While it is difficult to perform statistical analyses of comprehensive data, including all trader IDs, various fragmented data support the plausibility of the order-splitting hypothesis [3]. In addition, Ref. [18] provided crucial evidence on the qualitative importance of the order splitting based on a comprehensive data set: the authors of Ref. [18] decomposed the ACF $C(\tau)$ into the contribution by the same traders $C_{\rm same}(\tau)$ and that by other traders $C_{\rm other}(\tau)$. They finally showed that the former contribution is much larger than the latter one as $|C_{\rm other}(\tau)/C_{\rm same}(\tau)| \ll 1$ for large τ, suggesting the strong relevance of the order-splitting behaviors at least qualitatively

To organize this scenario more quantitatively and precisely, Lillo, Mike, and Farmer proposed a simple theoretical microscopic model (called the LMF model [19]) of the order-splitting behavior at the level of individual traders. They provided a clear explanation of the macroscopic LRC nature from the microscopic dynamics. Specifically, they assume that the length L of metaorders obeys the power-law distribution

$$P(L) \approx L^{-\alpha-1}, \quad \alpha > 1. \tag{2}$$

Under this assumption, they made a powerful quantitative prediction that the macroscopic behavior of the LRC should be directly related to the microscopic parameter of the model, such that

$$\gamma = \alpha - 1. \tag{3}$$

While the LMF model has been regarded for 18 years as a stylized microscopic model of order-splitting, its empirical foundation has not been fully verified, particularly for its quantitative prediction (3). While it is obviously appealing to provide its direct empirical verification, several severe difficulties have prohibited such empirical research: (i) Estimating the microscopic parameter α requires special comprehensive data sets, including all trader IDs. However, such data sets are scarce from the viewpoint of data availability. (ii) The quantitative confirmation of the prediction (3) is expected to require very large data sets. Indeed, because γ empirically distributes between 0 and 1, the estimation errors in the power-law exponents α and γ should be controlled roughly less than 0.1 even for drawing the scatterplot. This suggests that larger data sets are more necessary than the usual financial data analyses. (iii) Furthermore, the intrinsic long-memory character of the LRC essentially causes a slower convergence of its statistical estimator in estimating γ than usual (in our estimation, at least an order-sign sequence longer than 0.5 million transactions is necessary for obtaining even one data point of γ). Due to these three fundamental problems, the direct verification of the LMF model has been a crucial unsolved problem in econophysics.

In this paper, together with the companion paper [20], we present a quantitative verification of the LMF prediction (3) by analyzing a large high-frequency data set on the level of trading accounts. We have studied a large comprehensive order-book data set on the Tokyo Stock exchange (TSE) market, the biggest stock-exchange platform in Japan. This data set covers the nine-year period from 2012 to 2020 for all the stocks. Remarkably, this data set includes the *virtual server ID*, which is a unit of the trader accounts in the TSE platform. By appropriately analyzing the virtual server IDs, these data allow us to virtually track the trading behavior of all individual traders. Based on this data set, this paper addresses first the trading-strategy classification on the level of individual traders in terms of market orders. We classify all traders as either random traders (RTs) or splitting traders (STs) by the binomial test, directly confirming the presence of STs for most of the stocks. We next measure the metaorder-length (run-length) distribution among the STs for each stock. As assumed in the LMF model, we confirm that the metaorder length for the STs obeys power laws, such that $P(L) \propto L^{-\alpha-1}$ for large L with $\alpha > 1$. By measuring the power-law exponent γ in the LRC [$C(\tau) \propto \tau^{-\gamma}$], we provide the scatterplot between α and γ and then directly verify the LMF prediction (3) even at the quantitative level. Finally, we study the estimation of the total number of order-splitting traders from public data via the LMF theory regarding the prefactor c_0.

This paper is organized as follows. We describe our data set, the TSE market rule, and our mathematical notation in

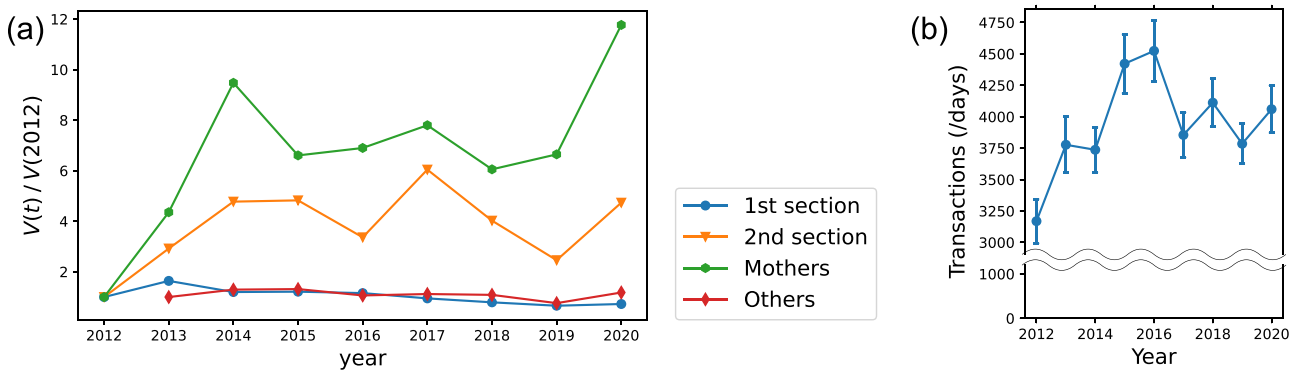

FIG. 2. Summary statistics on the TSE markets from 2012 to 2020. (a) Daily transaction volume on the arrowhead system $V(t)$ normalized by the volume in 2012 [i.e., $V(2012)$]. This figure is based on the public data provided by the TSE about the monthly/yearly transaction-volume statistics. These data show the increase in transaction volumes, particularly in the second section and Mothers markets. (b) Typical daily transaction numbers each year for the stock markets used for this study. As described later, we studied only stocks whose yearly transaction number is over 0.5 million. We calculated each stock's yearly average transaction number and then took its simple average across stocks for this plot.

Sec. II. In Sec. III, we provide a short review on the LMF model and its related literature. In Sec. IV, we apply our strategy-clustering algorithm to measure α for each stock. In Sec. V, we describe our statistical method to measure γ for each stock. The scatterplot between α and γ is provided in Sec. VI as the main result. We conclude in Sec. VII. A total of 10 Appendixes follow as a supplement to the main text.

II. DATA DESCRIPTION AND THE MARKET RULE

A. Data set

Here we describe our high-frequency data set on the TSE market in detail. Our data set was provided by the Japan Exchange (JPX) Group, Inc., which is the platform manager of the TSE market. This data set covers all the stocks in the TSE market during the nine-year period from January 4, 2012 to December 30, 2020. This comprehensive data set includes the order ID (i.e., the unique identifier to track the life cycle of any order), type (i.e., buy or sell), order type (i.e., limit order, cancellation order, and market order), price, and virtual server ID.

The TSE trading system is called the *arrowhead*. The arrowhead system was updated three times in our data set. For example, while the reaction speed of the arrowhead was 2 ms at the beginning of our data set, it was updated to be 1, 0.3, and 0.2 ms on July 17, 2012, September 24, 2015, and November 5, 2019, respectively. The high reaction speed of the arrowhead facilitates tradings, and the number of transactions increases in the TSE second section and Mothers [see Fig. 2(a)]. Because sufficient observations are crucial for precise measurement of the power-law exponent γ, it is expected that the measurement precision of γ will be better as time goes by, particularly after 2015 [see Fig. 2(b)].

B. Definition of trader IDs: Virtual server IDs and trading desks

One of the remarkable advantages of our data set is that it includes the virtual server ID. The virtual server is a unit of the trader accounts in the TSE. The virtual server ID is a consistent identifier of the TSE participants, but technically it is not completely equivalent to the membership ID. Indeed, there is an option for any trader to possess several virtual server IDs. For example, there is a limit on the number of submissions from a single virtual server during a fixed time interval. Some traders possess several virtual servers to avoid this submission limit for high-frequency tradings.

One of the technical solutions to this problem is to use the *trading desks* as an effective proxy of the membership ID, which was introduced by the work by Goshima, Tobe, and Uno [21]. The outline of their idea is to aggregate several virtual server IDs to allocate a unified ID (i.e., the trading desk) if we detect that the virtual servers are associated with the same membership.[1] For example, let us consider the case in which a trader possesses two virtual servers "V1" and "V2," submits a new limit order, and then cancels it finally. Typically, the virtual server IDs are identical between the submission and

[1]While the virtual server IDs are kept identical for most of the periods, they were shuffled when the arrowhead system was updated on September 24, 2015 and on November 4, 2019.

TABLE I. Schematic idea of the *trading desk* ID. Let us consider the case in which a trader issues submission and cancellation orders. Typically, the virtual server IDs between these orders are identical (see the order flow of the order ID "O1" with the same virtual server ID "V1"). At the same time, if the trader possesses two virtual servers "V1" and "V2," the trader can issue the cancellation order from a different virtual server "V2." For such a case, we infer that both virtual servers "V1" and "V2" are issued by the same trader and then allocate a single trading desk ID "T1." In this paper, the trading desk ID is regarded as representing an effective membership and is called the *trader ID* for short.

Order ID	Virtual server ID	Type	Trading desk ID
O1	V1	submission	T1
O1	V1	cancellation	T1
O2	**V1**	submission	T1
O2	**V2**	cancellation	T1

cancellation orders (see the order flow in Table I with the order ID "O1"). Sometimes, however, there are nontypical cases in which the virtual server IDs are not identical between the submission and cancellation orders (see the order flow in Table I with the order ID "O2"): e.g., when the trader submits a submission order with the order ID "O2" from the virtual server "V1" and subsequently submits a cancellation order with the order ID "O2" from the virtual server "V2," it is reasonable to infer that both virtual server IDs "V1" and "V2" are associated with the identical trader. The concept of the trading desk is to merge these two server IDs to allocate a single label as the effective trader ID (i.e., "T1" in Table I). For a detailed implementation, see Refs. [21,22]. In this paper, the trading desk is regarded as the effective membership ID and is called the *trader ID* for short.

It should be noted that in Japan, the TSE is not the sole stock market available. Various venues, including the proprietary trading system (PTS), exist where identical stocks can be traded. In addition, if a "final client," e.g., a mutual fund, can trade with different market members, their multiple IDs might be aggregated. In such cases, our trader ID might combine multiple metaorders from diverse clients, which would then be treated as a unified metaorder in our analyses.

C. Market rule

Here we describe the market rule in the TSE market. The TSE provides three types of trading periods: (i) the opening auctions (during 08:00–09:00 and 12:05–12:30), (ii) the continuous double auctions (during 09:00–11:30 and 12:30–15:00), and (iii) the closing auctions (at 11:30 and 15:00). Throughout this paper, time is based on Japan Standard Time (JST, UTC+9).

During the opening auctions, all the orders are collected but wait for their transaction until the fixed transaction time 9:00 or 12:30. During the continuous double auctions, all orders can be immediately executed under the time priority rule if the supply and demand match. In this paper, we focus on the continuous double auction periods.

In TSE, there are three types of orders: the limit order, the cancellation order, and the market order. Any limit order is composed of the price, the volume, and the type (i.e., bid or ask). When a trader is potentially willing to buy (sell) the specified volume of the stock at the specified price, the trader will submit a bid (ask) limit order. The limit order can be canceled if the trader is unwilling to buy (sell) the stock anymore.

D. Limit order book

While the background knowledge of the limit order book (LOB) is not essential in understanding our main findings, we briefly explain several important concepts related to the LOB, since they are useful in discussing the possible implications of our findings.

All the live limit orders are collected to form the LOB. A part of the LOB is publicly displayed and is used as an information source for decision-making by traders. The most important part of the LOB is the best bid (ask) price, defined by the highest bid (lowest ask) price in the LOB. Also, the market spread, defined by the difference between the best ask and bid prices, is an important measure of the effective transaction cost. We note that submitting limit orders is regarded as a beneficial contribution to market liquidity. Indeed, if there are plenty of bid and ask limit orders, anyone will be able to make a large volume of transactions with a small transaction cost. In this sense, traders keeping plenty of bid and ask limit orders are sometimes called *liquidity providers* or *market makers*.

On the other hand, the market order is the order to make a transaction at the available best prices. For example, if a trader submits the buy (sell) market order, the trader immediately buys (sells) the stock at the best ask (bid) prices. In contrast to the limit-order submissions, submitting market orders are regarded as liquidity consumption. Therefore, traders who submit market orders are sometimes called *liquidity consumers* or *takers*.

E. Mathematical notation

1. The fundamental quantities

Here we explain the mathematical notation for our analyses of one data point. In this paper, we focus on the following fundamental quantities [see Fig. 3(a) for a scheme]:

Ω_{TR}: the set of all trader IDs. The total size of the traders is finite, such that $|\Omega_{\mathrm{TR}}| = N_{\mathrm{TR}} < \infty$. Therefore, the trader IDs can be rewritten as $\Omega_{\mathrm{TR}} = \{i \mid i = 1, 2, \ldots, N_{\mathrm{TR}}\}$ without losing generality.

$\epsilon(t)$: the market-order sign at the discrete time $t \in \boldsymbol{N}$ in the whole market, with the set of natural integers $\boldsymbol{N} = \{1, 2, \ldots\}$. Here, the order sign $\epsilon(t) = +1$ [$\epsilon(t) = -1$] signifies the buy (sell) market order, and the time t is measured as a positive integer time (called *tick time*), incremented every transaction. The total number of market orders is denoted by $N_\epsilon := |\{\epsilon(t)\}_t|$, which is finite for real data analyses.

$\epsilon^{(i)}(t)$: the market-order sign issued by the trader $i \in \Omega_{\mathrm{TR}}$ at time t. If the trader i did not issue any order at time t, $\epsilon_t^{(i)}$ is set to be zero: $\epsilon^{(i)}(t) = 0$. By definition, an identity holds such that

$$\epsilon(t) = \sum_{i \in \Omega_{\mathrm{TR}}} \epsilon^{(i)}(t). \tag{4}$$

Here, the fundamental set $\boldsymbol{\Gamma} := (\Omega_{\mathrm{TR}}, \{\epsilon(t)\}_{t \in \boldsymbol{N}}, \{\epsilon^{(i)}(t)\}_{t \in \boldsymbol{N}, i \in \Omega_{\mathrm{TR}}})$ completely characterizes our analyses. We note that the volume information on any market order is not used in this paper.

2. Other important quantities

In addition, we can define the following quantities as derivatives of the fundamental quantities $\boldsymbol{\Gamma}$ [see Fig. 3(b) for a schematic]:

$C(\tau)$: the market ACF with the time lag $\tau \geqslant 0$, defined by $C(\tau) := \langle \epsilon(t)\epsilon(t+\tau) \rangle$, where $\langle A \rangle$ denotes the ensemble average of any stochastic quantity A.

$\{\epsilon_k^{(i)}\}_{k \in \boldsymbol{N}}$: the reduced order-sign sequences, by removing zeros from the original order-sign sequences $\{\epsilon^{(i)}(t)\}_{t \in \boldsymbol{N}}$. The total number of market orders for the trader i is denoted by $N_{\mathrm{MO}}^{(i)} := |\{\epsilon_k^{(i)}\}_k|$, which can be finite.

(a) $\Omega_{\rm TR} = \{1, 2\}$, $N_{\rm TR} := |\Omega_{\rm TR}| = 2$

Time	t	1	2	3	4	5	6	7	8	...
Market signs	$\epsilon(t)$	+	+	−	+	−	+	−	−	...
Signs of trader 1	$\epsilon^{(1)}(t)$	+	+	0	+	0	+	−	0	...
Signs of trader 2	$\epsilon^{(2)}(t)$	0	0	−	0	−	0	0	−	...

(b)

Reduced signs of trader 1	$\epsilon_k^{(1)}$	+	+	+	+	−	−	−	+	...	
Runs of trader 1	$L_k^{(1)}$	$L_1^{(1)} = 4$				$L_2^{(1)} = 3$			...		
Reduced signs of trader 2	$\epsilon_k^{(2)}$	−	−	−	+	+	+	−	−	...	
Runs of trader 2	$L_k^{(2)}$	$L_1^{(2)} = 3$			$L_2^{(2)} = 3$			$L_3^{(2)} = 2$		...	

FIG. 3. (a) Schematic example of the fundamental quantities $\mathbf{\Gamma} := (\Omega_{\rm TR}, \{\epsilon(t)\}_t, \{\epsilon^{(i)}(t)\}_{t,i})$ for the case $N_{\rm TR} = |\Omega_{\rm TR}| = 2$. Here, + (−) is an abbreviation of +1 (−1), representing a buy (sell) order. (b) Reduced-sign sequences $\{\epsilon_k^{(i)}\}_k$ are defined by removing zeros from the original order sequences $\{\epsilon^{(i)}(t)\}_t$ for $i \in \Omega_{\rm TR}$. Runs $\{L_k^{(i)}\}_k$ are also defined as the numbers of successively the same signs for the trader i.

$\{L_k^{(i)}\}_{k\in\mathbb{N}}$: the runs for the reduced-sign sequences $\{\epsilon_k^{(i)}\}_{k\in\mathbb{N}}$ for the trader $i \in \Omega_{\rm TR}$. For a given reduced-sign sequence $\{\epsilon_k^{(i)}\}_{k\in\mathbb{N}}$ of the trader $i \in \Omega_{\rm TR}$, we define the *runs* similarly to the Wald-Wolfowitz runs test [23]. In other words, for $\{\epsilon_k^{(i)}\}_k$, we count the numbers of adjacent equal elements (e.g., $L_1^{(1)} = 4$ and $L_2^{(1)} = 3$ for $\{++++---+\cdots\}$) to define the runs $\{L_k^{(i)}\}_k$ (see Fig. 3).

As will be explained in Sec. IV, we apply a strategy clustering in terms of market orders to define the following classes of traders:

$\Omega_{\rm RT}$: the set of random traders.

$\Omega_{\rm ST}$: the set of splitting traders.

By definition, we have $\Omega_{\rm TR} = \Omega_{\rm RT} \cup \Omega_{\rm ST}$.

3. Sample label for the integrated statistical analysis

More technically, the fundamental quantities $\mathbf{\Gamma}$ are defined for each data point. By introducing a sample label $s \in \mathcal{S}$ to identify each data point with the sample set $\mathcal{S}$, $\{\mathbf{\Gamma}_s\}_{s\in\mathcal{S}}$ is finally analyzed as the integrated statistical analysis to produce Figs. 11 and 12 (e.g., the scatterplot between α and γ in Sec. VI). One data point $\mathbf{\Gamma}_s$ corresponds to a yearly order-sign sequence for one stock market (i.e., the label s signifies the set of stock ticker code and year). The total number of the data points is denoted by $N_{\mathcal{S}} := |\mathcal{S}|$. However, if the expression clearly makes sense in the context, the sample label s is often omitted for brevity.

4. Filter on the sample markets

In this paper, we focus on the markets whose total transaction number is over 0.5 million, such that $N_\epsilon > 5 \times 10^5$. This filter is introduced to suppress the estimation errors in the power-law exponents α and γ.

5. Other mathematical notation

We next describe our notation for the probability theory. The probability density function (PDF) characterizes the probability that the stochastic variable x' resides in the range $[x, x + dx)$ as $P(x)dx$. The complementary cumulative distribution function (CCDF) is defined by $P_>(x) := \int_x^\infty dy P(y)$.

For a given series $\{x_k\}_k$, we can define the empirical PDF and CCDF as

$$P(x) := \frac{1}{|\{x_k\}_k|}\sum_k \delta(x - x_k),$$
$$P_>(x) := \int_x^\infty dy P(y) = \frac{N_>(x)}{|\{x_k\}_k|}, \tag{5}$$

where $N_>(x) := \int_x^\infty \sum_k \delta(y - x_k)dy$ is the total number of elements larger than x, and $\delta(x)$ is the Dirac delta function.

F. Data preprocessing

Here we explain our data preprocessing to extract data by removing the influence of intraday seasonality. Intraday seasonality is one of the stylized facts in financial markets [3], and the market activity typically exhibits high intensity around the opening and closing times of the auctions (called the U-shape profile). Indeed, we confirmed the U-shape profile in terms of the market-activity statistics (see Appendix A).

This intraday seasonality should be considered in interpreting the results of any data analysis because there are various factors unique to the opening and closing times of the auctions (such as the lifestyle of traders and the position management [24], for example). Such factors are not included in the LMF model; therefore, the data during such high-activity periods are not suitable for the data calibrations.

For these reasons, we used the market-order sign sequence during the continuous double auction periods with the 10-min sequences excluded around the opening and closing auctions. In other words, we used the data from 9:10 to 11:20 and from 12:40 to 14:50 as a daily order-sign sequence. The daily order-sign sequences are segmented on a yearly basis for each stock to obtain one data point $\mathbf{\Gamma}$.

III. LITERATURE REVIEW ON THE LMF MODEL

This section reviews the LMF model in terms of the model setup, quantitative prediction, and the current qualitative empirical evidence. This section aims to provide background knowledge on this econophysics topic for the general audience to clarify the novelty of our results. Since this review section is prepared independently of the other sections, readers interested only in our main results may skip this section.

A. Microscopic model: The original LMF model

Let us assume that the total number of traders $N_{\rm TR} > 0$ is a time-constant positive integer and the volume of any market order is always the minimum executable unit for simplicity. For any trader $i \in \Omega_{\rm TR}$, two microscopic variables are defined: $z^{(i)}(t) := (\epsilon^{(i)}(t), R^{(i)}(t))$, where $\epsilon^{(i)}$ is the order sign of the metaorder and $R^{(i)}$ is the remaining volume of the metaorder. The macroscopic variable of the market is given by the market-order sign $\epsilon(t)$. The LMF model is formulated as the Markovian stochastic process for the state variable $\boldsymbol{Z} := (\epsilon; z_1, \ldots, z_{N_{\rm TR}})$ on the discrete time $t \in \boldsymbol{N}$.

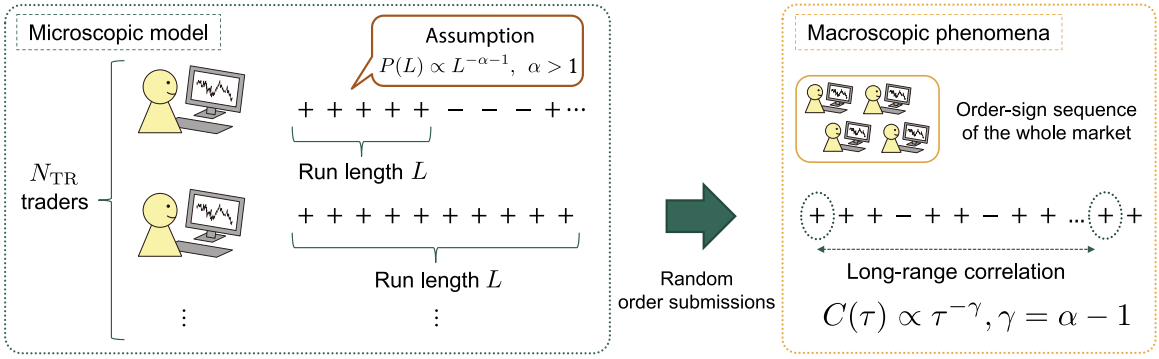

FIG. 4. Schematic of the Lillo-Mike-Farmer (LMF) model proposed in Ref. [19]. At the microscopic dynamics, the total number of traders is N_{TR} and all the traders are assumed to be splitting traders (STs). STs hold large metaorders and they randomly split them into small orders. Here we assume that the run length L obeys the power-law distribution $P(L) \propto L^{-\alpha-1}$ with $\alpha > 1$. At the macroscopic dynamics, the order-sign sequence of the whole market exhibits the long-range correlation $C(\tau) \propto \tau^{-\gamma}$. Furthermore, the LMF model theoretically predicts $\gamma = \alpha - 1$ as Eq. (7), connecting the macroscopic power-law exponent γ and the microscopic exponent α. The state variables and model parameters are summarized in Table II.

The concrete dynamics of this model is given by the following stochastic difference equations (SDEs; see Fig. 4): At the time $t+1$, a trader $i = \pi(t+1)$ is randomly selected with the uniform distribution, such that

$$P_{t+1}(\pi) = \frac{1}{N_{\mathrm{TR}}} \quad \text{for any } \pi \in \Omega_{\mathrm{TR}}. \tag{6a}$$

The $\pi(t+1)$th trader executes their metaorder with the order sign:

$$\epsilon(t+1) = \epsilon^{[\pi(t+1)]}(t). \tag{6b}$$

After the execution by the trader π, the remaining volume $R^{(\pi)}(t+1)$ decreases by 1 if $R^{(\pi)}(t) > 1$. If all the metaorder is executed [i.e., $R^{(\pi)}(t) = 1$], the metaorder and its sign are randomly reset for the trader π. In summary, the dynamics of $z^{(i)}$ is given as follows for all $i \in \Omega_{\mathrm{TR}}$:

$$R^{(i)}(t+1) = \begin{cases} R^{(i)}(t) & \text{if } i \neq \pi(t+1), \\ R^{(i)}(t) - 1 & \text{if } i = \pi(t+1) \text{ and } R^{(i)}(t) > 1, \\ L & \text{if } i = \pi(t+1) \text{ and } R^{(i)}(t) = 1;\ L \text{ obeys } P(L); \end{cases} \tag{6c}$$

$$\epsilon^{(i)}(t+1) = \begin{cases} \epsilon^{(i)}(t) & \text{if } i \neq \pi(t+1) \text{ or } R^{(i)}(t) > 1, \\ +1 & \text{with prob. } 1/2, \text{ if } i = \pi(t+1) \text{ and } R^{(i)}(t) = 1, \\ -1 & \text{with prob. } 1/2, \text{ if } i = \pi(t+1) \text{ and } R^{(i)}(t) = 1, \end{cases} \tag{6d}$$

with an independent and identically distributed (IID) random integer number $L > 0$ obeying the discrete PDF $P(L)$.

The set of the SDEs (6) completely characterizes the $(2N_{\mathrm{TR}}+1)$-dimensional Markovian dynamics with the state variable $\boldsymbol{Z}(t)$ on the discrete time $t \in \boldsymbol{N}$. In this sense, the SDEs (6) are the fundamental "equations of motion" for the LMF model at the microscopic level of the financial dynamics. We will consider the dynamics of this stochastic process until the final time $t = N_\epsilon := |\{\epsilon(t)\}_t|$ (i.e., the total number of transactions). See Table II for a summary of the state variables and the model parameters.

In this framework, all traders are assumed to simply split their metaorders without complicated strategies according to the order-splitting hypothesis, and the discrete PDF $P(L)$ can be interpreted as the distribution of the metaorder lengths (or the run lengths). For consistency with the realistic data analysis, it is customary to assume the power-law metaorder distribution:

$$P(L) \propto L^{-\alpha-1} \text{ for large } L$$

with a realistic value [3,25,26] around $\alpha \approx 1.5$ (see Appendix B for a detailed implementation in generating

TABLE II. The summary of the state variables and the model parameters for the LMF model (6).

State variable	Meaning	Model parameters	Meaning
$\epsilon(t)$	Order sign in the whole market	N_{TR}	Total number of the traders
$\{\epsilon^{(i)}(t)\}_i$	Order sign of the trader i	N_ϵ	Total number of the transactions
$\{R^{(i)}(t)\}_i$	Remaining metaorder length of trader i	$P(L) \propto L^{-\alpha-1}, \alpha > 1$	Metaorder length distribution

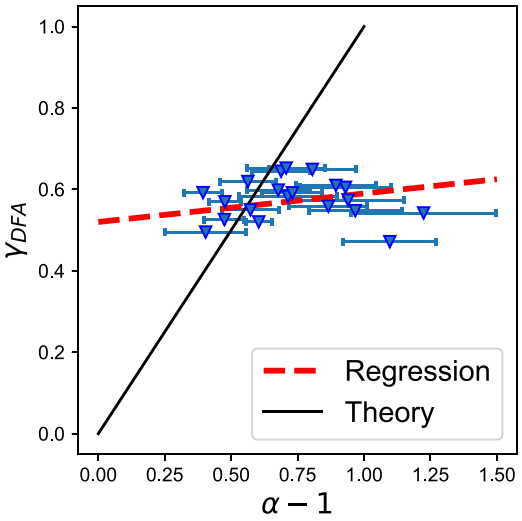

FIG. 5. Verification of the LMF prediction (3) in the previous work [19]. We extracted the data from figure 7 of Ref. [19] by measuring the coordinates of the data points using Adobe Illustrator. Here we additionally plot the regression line (led) with the slope 0.07 (statistically insignificant), far from the theoretical coefficient 1. We note that they measured γ based on the DFA (see the related discussion in Sec. V).

power-law random numbers). We can straightforwardly generalize this model to introduce heterogeneity of splitting strategies (see Ref. [27]).

B. Quantitative prediction: From micro to macro

Since the microscopic model is fixed as the high-dimensional Markovian stochastic process (6), the macroscopic character of this model can be deduced in principles. Such a statistical-mechanical program was provided by the original paper [19] by Lillo, Mike, and Farmer. Indeed, the ACF of the order-sign sequence $\{\epsilon(t)\}_t$ is asymptotically given by

$$C(\tau) := \lim_{t\to\infty} \langle \epsilon(t)\epsilon(t+\tau)\rangle \propto \tau^{-\gamma}, \quad \gamma := \alpha - 1 \tag{7}$$

for large τ. This formula implies that the macroscopic parameter γ is directly related to the microscopic parameter α. In this paper, the expression "the quantitative prediction of the LMF model" refers to this relationship (7). Note that this relation holds even for a generalized LMF model with heterogeneous strategies [27].

In the pioneering work in Ref. [19], a scatterplot was provided between α and γ by analyzing an off-book market data set as a proxy for hidden orders. We extracted the data in the figure in Ref. [19] and we plot it as Fig. 5 with the red regression line added. These data show two points:

(i) The theoretical line passes roughly through the center of the data points, suggesting the minimum qualitative consistency between data and theory.

(ii) At the same time, the theoretical line does not exhibit a good fit in explaining the "variations in the measured values." Indeed, the red regression line has the coefficient of the slope 0.07, which is far from the theoretical coefficient.[2] This may be due partly to their "improper proxy"[3] and the smallness of the sample size.

It should be noted that Refs. [25,26] showed that $\alpha \approx 1.5$ is empirically obtained on the basis of the aggregated distribution, suggesting the LMF prediction is consistent at least qualitatively. However, to establish the LMF quantitative prediction (7), it is necessary to solve the second problem by analyzing a large and proper data set.

C. Qualitative prediction and the corresponding empirical evidence

While the quantitative prediction (7) is interesting, it can be another option to examine a rather weaker prediction by the LMF model. According to Ref. [18], let us decompose the ACF $C(\tau)$:

$$C(\tau) := C_{\text{same}}(\tau) + C_{\text{other}}(\tau). \tag{8}$$

Here $C_{\text{same}}(\tau)$ is the contribution where the same trader issues orders at t and $t+\tau$, whereas $C_{\text{other}}(\tau)$ is the contribution where two distinct traders issue orders at t and $t+\tau$. If the order-splitting hypothesis is correct, the following relationship is expected to hold:

$$|C_{\text{same}}(\tau)| \gg |C_{\text{other}}(\tau)|, \quad \tau \gg 1. \tag{9}$$

In this paper, we call this relationship (9) "the qualitative prediction of the LMF model," in comparison to the qualitative prediction (7). Reference [18] addressed this problem and showed that the quantitative prediction (9) actually holds in their data set. This is the best empirical evidence supporting the order-splitting hypothesis, to the best of our knowledge.

D. Goal of this paper

The excellent evidence in [18] suggests the strong relevance of the order-splitting hypothesis as the microscopic origin of the LRC, at least on a qualitative level (9). At the same time, it is remarkable that the LMF model further provides the quantitative prediction (7), which is much stronger than the qualitative prediction (9). This relationship (7) is obviously appealing. However, there has been no systematic and solid evidence to support this prediction at the quantitative level.

The goal of this paper is to examine and establish the quantitative prediction (7) by analyzing our large data set on the TSE market. To prove the relationship (7), it is sufficient to draw a scatterplot between α and γ with a sufficiently large sample size. Therefore, we basically proceed with our data analysis in the following three steps (see also Appendix C for a summary of the technical problems to be solved): (i) measurement of the microscopic parameter α, (ii) measurement of the macroscopic parameter γ, and (iii) drawing the scatterplot between α and γ.

[2]They state, "As a stronger test, one might hope that variations in measured values of α might predict variations in measured values of γ. The model fails this test" in Ref. [19].

[3]They state, "Because we lack the proper data to test the model, we have used an imperfect proxy to test the model" in Ref. [19].

IV. MEASUREMENT OF THE THE METAORDER-LENGTH DISTRIBUTION

Here we classify the random and order-splitting strategies in terms of the market orders to finally measure the microscopic parameter α in the metaorder distribution $P(L) \propto L^{-\alpha-1}$ for large L.

A. Measurement of metaorder lengths for individual traders

We first define the metaorder series at the level of individual traders. Basically, we follow the rule described in Sec. II E 1 to extract the order-sign sequence $\{\epsilon_k^{(i)}\}_k$ for the ith trader and to construct the corresponding run sequences $\{L_k^{(i)}\}_k$. The run sequences are regarded the *metaorder-length sequences* in this paper.

For a practical reason, however, we introduce one exceptional rule: if the time interval between two successive orders is sufficiently longer, we assume that two orders belong to different metaorders according to Ref. [17]. This rule is introduced to avoid overestimation of unrelated orders. For example, let us consider the case in which a trader submits ten buy orders within a day, stops orders for one month, and then submits ten buy orders. It is not realistic to assume that the metaorder length is 20 because the one-month resting seems too long. We expect that this exceptional rule will reduce the risk of such overestimation. In this paper, we set this time threshold to be one business day.

B. Strategy clustering: Random versus order-splitting traders

We next identify the *order-splitting traders* (STs) at the level of individual traders. Our basic idea is to apply the binomial test in statistics to define the *random traders* (RTs) and then define STs as non-RTs. The details of our strategy-clustering methods and the corresponding results are described below.

1. Methods: The binomial test

Let us define the RTs by the binomial test as follows: If a trader i randomly issues market orders, it is expected that the sign sequence $\{L_k^{(i)}\}_k$ is generated according to the symmetric Bernoulli process. In other words, the sign sequence obeys the rule

$$P\big(\epsilon_k^{(i)} = +1 \mid \epsilon_{k-1}^{(i)}, \ldots, \epsilon_1^{(i)}\big) = \tfrac{1}{2} \tag{10}$$

for any $k \geqslant 1$.

On the basis of this picture, we set the following null hypothesis:

$$H_0: \text{ the sign sequence of the trader } i \text{ obeys the symmetric Bernoulli process.} \tag{11}$$

This hypothesis is examined by the one-sided binomial test with the significance level $\theta := 0.01$ as follows: Let us consider the reduced order-sign sequence $\{\epsilon_k^{(i)}\}_k$ of the ith trader. The total number of their market orders is given by $N_{\mathrm{MO}}^{(i)} := |\{\epsilon_k^{(i)}\}_k|$ and the corresponding run-length sequence is given by $\{L_k^{(i)}\}_k$. We focus here on the total number of runs defined by $N_{\mathrm{run}}^{(i)} := |\{L_k^{(i)}\}_k|$. If the null hypothesis H_0 is correct, the total number of runs $N_{\mathrm{run}}^{(i)}$ must obey the binomial distribution,

$$P(N_{\mathrm{run}}^{(i)}) = \frac{1}{2^{N_{\mathrm{MO}}^{(i)}-1}} \binom{N_{\mathrm{MO}}^{(i)} - 1}{N_{\mathrm{run}}^{(i)}}. \tag{12}$$

We thus apply the one-sided binomial test to testify the null hypothesis H_0. If this null hypothesis is rejected, we classify the trader i as an ST, such that $i \in \Omega_{\mathrm{ST}}$ with the set of the STs Ω_{ST}; otherwise, the trader i is classified as an RT (i.e., $i \in \Omega_{\mathrm{RT}}$ with the set of the RTs Ω_{RT}). The first-kind error (the false-positive rate) is controlled in our statistical test, and the clustering for the STs.

2. Results 1: The existence of the order-splitting traders

Let us show the overview of our clustering results. We applied the clustering algorithm in Sec. IV B 1 to all traders for all stock every year (i.e., one data point Γ) to obtain Ω_{ST} and Ω_{RT}. We can define the ratio of the STs $|\Omega_{\mathrm{ST}}|/|\Omega_{\mathrm{TR}}|$ for each data point. We first show the empirical distribution of the STs percentage as Fig. 6(a). The typical percentage of the STs is given by 25%, showing the direct evidence of the presence of the STs in our data set.

Interestingly, while the number of the STs is typically less than that of the RTs, the STs typically exhibits the dominant contribution to the market orders. To show this characteristic, let us define the market-order contribution percentage by the STs as the ratio of the number of market orders issued by the STs to the total number of market orders. Figure 6(b) shows the empirical distribution of the market-order contribution percentage by the STs, illustrating that the STs typically contribute 80% to the total market orders. In addition, the presence of STs shows a tendency to increase over the years [see Fig. 6(c)].

3. Results 2: Metaorder-length distributions

We then study the metaorder-length CCDF for the STs (see also Appendix D for the clustering results of RTs as a reference). Let us consider the joint run-length sequences for STs and the corresponding metaorder-length CCDF:

$$\{L_k^{\mathrm{ST}}\}_k := \bigcup_{i \in \Omega_{\mathrm{ST}}} \{L_k^{(i)}\}_k,$$

$$P_>(L^{\mathrm{ST}}) := \frac{N_>(L^{\mathrm{ST}})}{|\{L_k^{\mathrm{ST}}\}_k|}, \quad N_>(L^{\mathrm{ST}}) := \int_{L^{\mathrm{ST}}}^{\infty} dy \sum_k \delta\big(y - L_k^{\mathrm{ST}}\big). \tag{13}$$

The empirical metaorder distribution for STs is plotted in Fig. 7(a) for Toyota 2020, showing the power law $P_>(L) \approx L^{-\alpha}$. We confirm that this character is robustly observed even for other data points.

The empirical PDF $P(\alpha)$ of the power-law exponent α is shown in Fig. 7(b). Approximately 90% of the stocks have power-law exponent $\alpha < 2$ in our data set. This finding is consistent with the standard assumption that $\alpha < 2$ in the LMF model. The power-law exponent α is estimated by Clauset's algorithm [28,29] as one of the established statistical estimation methods. According to Ref. [28], the estimation errors in the power-law exponent are generally small, at least compared with the errors in another power-law exponent γ. We thus

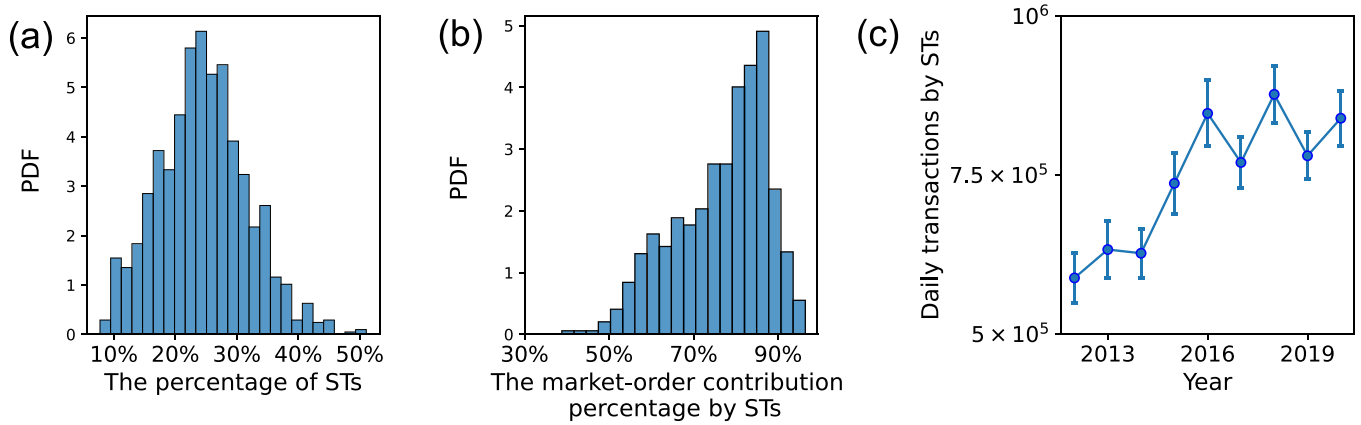

FIG. 6. Summary statistics of STs. (a) Empirical distribution of the percentage of STs in each market. Approximately 25% (10–50 %) of traders are classified as STs in each market. (b) Empirical distribution of the market-order contribution percentage by STs. 80% of the total market orders are typically submitted by STs. These two figures suggest that STs dominantly contribute to the market orders, while their number is relatively lower than that of RTs. (c) Yearly plot of average daily transaction numbers by STs, showing the growing presence of STs.

ignore the estimation errors of the power-law exponent α throughout this paper.

4. Parameter estimation for the LMF model

Here we describe our method to estimate the parameters for the LMF model.

N_{ST}: the total number of the active STs, trading at least 1000 times in the year, is estimated as the yearly average number of all the STs:

$$N_{\mathrm{ST}} = \frac{1}{D_{\mathrm{year}}} \sum_{i \in \Omega_{\mathrm{ST}}} D^{(i)}, \tag{14}$$

where D_{year} is the total number of business days in that year, and $D^{(i)}$ is the total number of active days by the ith ST. We used N_{ST} as a proxy for the parameter calibration of N_{TR} because the influence of inactive traders is expected to be negligible on the empirical autocorrelation function. In addition, it is numerically known that the asymptotic behavior of the autocorrelation function is robust regarding the total number of traders N_{TR} [3], and thus the technical details of the parameter calibration of N_{TR} are expected to be insensitive to the final results.

N_ϵ: we substitute the total number of market orders for the stock during the year into N_ϵ.

α: the power-law exponent of the metaorder length PDF $P_{\mathrm{ST}}(L) \propto L^{-\alpha-1}$ for large L. This power-law exponent is estimated by the Clauset algorithm as described in Sec. IV B 3.

This parameter estimation method was used for the numerical simulations in Sec. V.

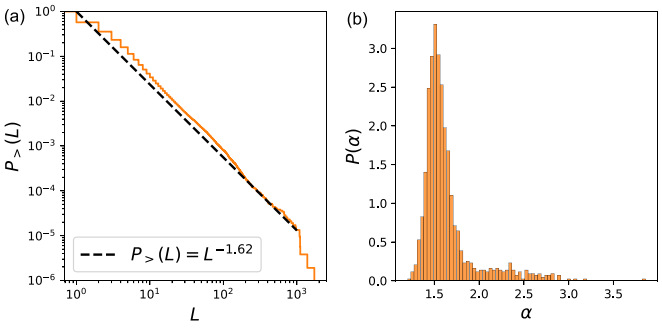

FIG. 7. Characters of metaorder-length distribution on STs in each markets. (a) The aggregated metaorder-length CCDF for all STs for Toyota Motor Corporation in 2020 as a typical example. The metaorder-length CCDF obeys the power law such that $P_>(L) \propto L^{-\alpha}$. (b) The empirical PDF of the power-law exponents α in our whole data set. The exponent α was measured by Clauset's algorithm [28,29] across all the markets. Typically, α distributes within $1 < \alpha < 2$, consistently with the standard assumption for the LMF model.

V. MEASUREMENT OF THE SIGN AUTOCORRELATION FUNCTION

Here we describe the measurement of the power-law exponent γ in the sign ACF $C(\tau) \propto \tau^{-\gamma}$ for large τ. Our method is composed of three steps: (i) application of the naive estimator γ_{NLLS} by the nonlinear least squares (NLLS) to the empirical ACF or power-spectral density (PSD), (ii) construction of an unbiased estimator based on the LMF model, and (iii) application of the unbiased estimator to obtain the final $\gamma_{\mathrm{unbiased}}$. Let us explain these steps one by one.

A. Estimations by the nonlinear least squares

We employed two NLLS estimation methods based on the ACF and PSD for our statistical analyses. Both methods show similar and consistent results, implying the robustness of our analyses. While there are many sophisticated estimation methods [such as the estimation based on the detrended fluctuation analysis (DFA) [10]], we employ this simple method because we find that the NLLS estimation has the consistency for an infinite sample size and has less bias for a finite sample size than other methods.

1. Estimation based on the sample ACF

Let us first describe the measurement method based on the sample ACF. The basic idea is to apply the power-law fitting

to the sample ACF $C_{\rm sample}(\tau)$, such that

$$C_{\rm sample}(\tau) := \frac{1}{N_\epsilon - \tau} \sum_{t=1}^{N_\epsilon - \tau} \epsilon(t)\epsilon(t+\tau) \propto \tau^{-\gamma^{(a)}_{\rm NLLS}}, \quad (15)$$

where the superscript (a) signifies the ACF estimator. A detailed implementation is described in Appendix E 1. The theoretical advantage of this method is that the sample ACF is expected to converge to the true ACF for the infinite sample size:

$$\lim_{N_\epsilon \to \infty} C_{\rm sample}(\tau) = \langle \epsilon(t)\epsilon(t+\tau) \rangle = C(\tau) \quad (16)$$

under the ergodicity assumption. In general, ergodicity is a weak assumption irrelevant to the underlying microscopic dynamics (in our case, the microscopic dynamics are assumed to be governed by the LMF model). Therefore, it is expected that the NLLS estimator has the consistency for general setups as shown in Sec. V B.

2. Estimation based on the sample PSD

Another method we employed is based on the PSD. The basic idea is to utilize the one-to-one correspondence between the PSD $S(\omega)$ and the ACF $C(\tau)$, guaranteed by the Wiener-Khinchin theorem,

$$S(\omega) = \int_{-\infty}^{\infty} C(\tau) e^{2\pi i \omega \tau} d\tau. \quad (17)$$

Considering the integral identity [30] for $\gamma \in (0, 1)$,

$$\int_{-\infty}^{\infty} e^{2\pi i \omega \tau} |\tau|^{-\gamma} d\tau = 2^\gamma \pi^{\gamma - 1} \Gamma(1-\gamma) \sin \frac{\pi \gamma}{2} |\omega|^{\gamma - 1}, \quad (18)$$

if the sample PSD obeys the power-law asymptotics for small ω,

$$S_{\rm sample}(\omega) \propto \omega^{-H}, \quad (19)$$

the Hurst exponent H is related to γ as $H = 1 - \gamma$. In other words, the NLLS estimator $\gamma^{(s)}_{\rm NLLS}$ is defined by

$$\gamma^{(s)}_{\rm NLLS} = 1 - H, \quad (20)$$

where H is determined by the NLLS method for the PSD. The superscript (s) signified the PSD estimator. As with the sample ACF, the sample PSD converges to the true PSD under the ergodicity assumption. Therefore, this estimation is expected to be robust regarding the consistency (see Sec. V B). A detailed implementation is described in Appendix E 2.

3. Comparison between the ACF and PSD methods

We discuss theoretical differences between the ACF and PSD methods for comparison. The NLLS fitting for the PSD sometimes provides negative values of $\gamma < 0 \Longleftrightarrow H > 1$ due to methodological artifacts. Negative γ implies monotonic increasing of the ACF for large τ, which does not make sense. We excluded such data points because of the obvious failure of the estimation.[4]

[4]The total number of data points was 16, which needs exceptional handling with $\gamma < 0$.

The PSD estimator is theoretically valid only for $H \in (0, 1)$, or equivalently $\alpha \in (1, 2)$. This fact means the estimation fails for $\alpha > 2$ in principle, even for infinite observations. Also, the estimation accuracy tends to be worse near the critical point $\alpha = 2$.

These disadvantages contrast with the ACF method, which is expected to work for any α in principle for infinite observations, and they provide only positive γ. However, the PSD method is broadly used to estimate the Hurst exponent and is a realistic option for the statistical estimation of γ. Indeed, as for the LMF simulations, we find that the overall bias due to the finite sample size was less in the PSD method than in the ACF method (see the value of β_1 in Sec. V B 2).

B. Consistency and biasedness of the NLLS estimator

In statistics, *consistency* and *unbiasedness* are two of the desirable characteristics of any statistical estimator. Here we numerically confirm these characteristics of the NLLS estimator based on the LMF model (see Figs. 8 and 9 for the ACF and PSD methods, respectively). While the NLLS estimator has consistency at least numerically, unfortunately it does not have unbiasedness. This problem is heuristically solved in Sec. V C by appropriate construction of an unbiased estimator.

1. Consistency for the infinite sample size

Any estimator T_n is called consistent if the estimated value converges to the true value θ for the infinite sample size $n \to \infty$: $\lim_{n\to\infty} T_n = \theta$. We have numerically confirmed the consistency of the NLLS estimator:

$$\lim_{N_\epsilon \to \infty} \gamma_{\rm NLLS} = \gamma. \quad (21)$$

To confirm this consistency (21), we have numerically generated the order-sign sequences by the LMF model with realistic parameters of our data set: we have measured the model parameter set $(N_{\rm ST}, \alpha)$ for all sample points according to Sec. IV B 3 except for N_ϵ. For N_ϵ, we employed $N_\epsilon = 10^8$ because realistic values $N_\epsilon \lesssim 10^7$ are not sufficient to confirm the consistency (21). Figures 8(a) and 9(a) illustrate our numerical simulation, showing that the NLLS estimator $\gamma_{\rm NLLS}$ numerically agrees with the theoretical formula $\gamma = \alpha - 1$. This numerical evidence supports the consistency of the NLLS estimator. Note that the consistency of the NLLS estimator is theoretically reasonable because the sample ACF (PSD) converges to the true ACF (PSD) for the infinite sample size under the assumption of ergodicity. Note that the PSD method works slightly worse near $\alpha \approx 2$ than the ACF method [see Fig. 9(a)] because $\alpha = 2$ is the critical point beyond which the PSD method fails to estimate α in principle.

2. Bias for the finite sample size

Any estimator T_n is called unbiased if the expectation of the estimator is equivalent to the true value θ for the finite sample size $n < \infty$: $\langle T_n \rangle = \theta$. Unfortunately, we have numerically confirmed that the NLLS estimator does not have unbiasedness:

$$\langle \gamma_{\rm NLLS} \rangle \neq \gamma \quad \text{for finite } N_\epsilon. \quad (22)$$

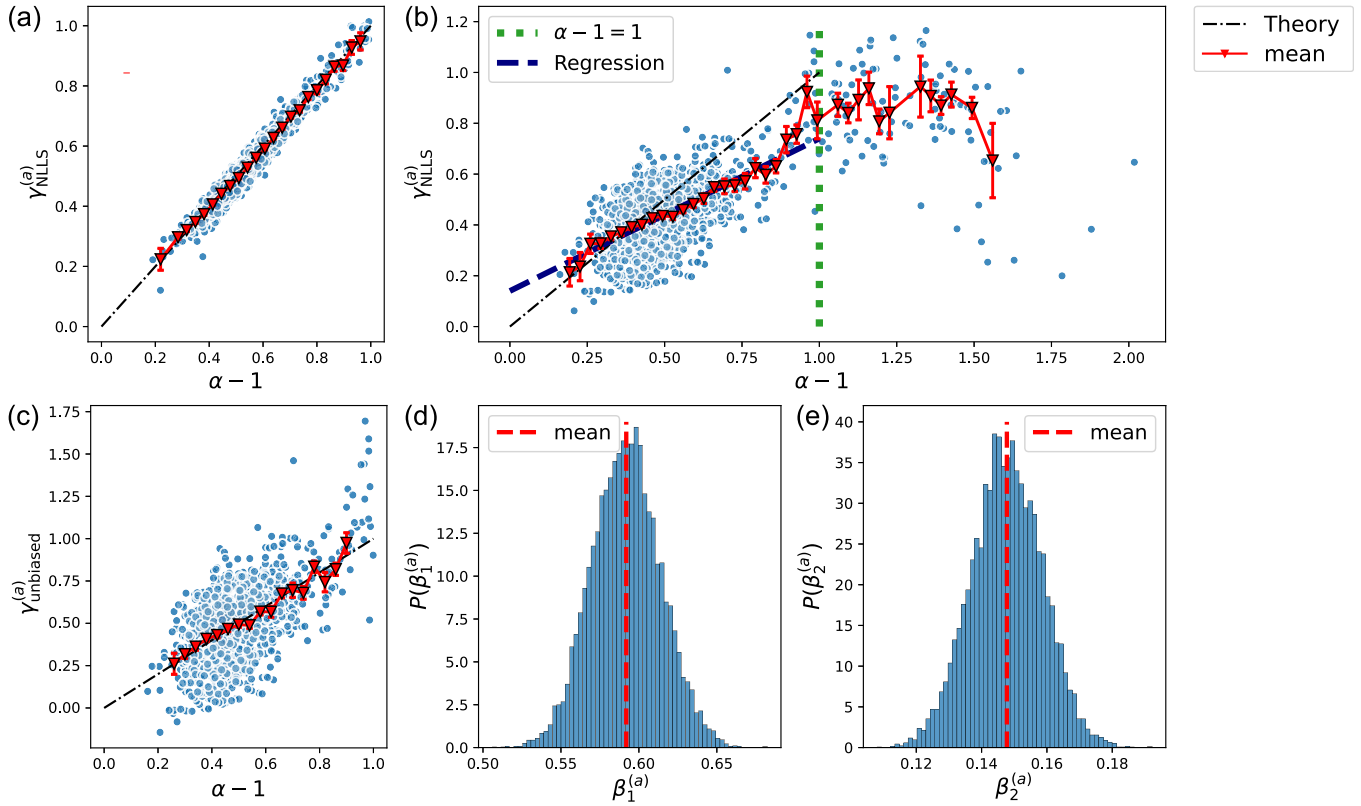

FIG. 8. Numerical simulations of the LMF model to test consistency and biasedness of the NLLS estimator $\gamma^{(a)}_{\rm NLLS}$ based on the ACF method. (a) Consistency of the ACF-NLLS estimator $\gamma^{(a)}_{\rm NLLS}$. The theoretical formula $\gamma^{(a)}_{\rm NLLS} = \alpha - 1$ holds for a sufficiently large sample size $N_\epsilon = 10^8$, supporting the consistency of the NLLS estimator $\gamma^{(a)}_{\rm NLLS}$. The parameter sets $(N_{\rm ST}, \alpha)$ are based on the measurement of our data set. (b) For realistic sample sizes $N_\epsilon \lesssim 10^7$, the ACF-NLLS estimator $\gamma^{(a)}_{\rm NLLS}$ exhibits biasedness due to the finite sample size. We have simulated the LMF model by assuming that the parameter sets $(N_{\rm ST}, N_\epsilon, \alpha)$ are identical to those measured in our data set. We find that the theoretical formula $\gamma^{(a)}_{\rm NLLS} = \alpha - 1$ does not hold due to the finite sample size. The deviation from the theoretical line is particularly serious for large $\alpha - 1 > 1$, showing the nonuniform convergence of the NLLS estimator. Based on this empirical finding, we focus on the range $\alpha - 1 \in (0, 1)$ and apply the linear regression (23) to obtain the navy line. (c) Check of the numerically constructed unbiased ACF estimator $\gamma^{(a)}_{\rm unbiased}$. The unbiased estimator is constructed by $\gamma^{(a)}_{\rm unbiased} := (\gamma^{(a)}_{\rm NLLS} - \beta^{(a)}_2)/\beta^{(a)}_1$. As expected, the theoretical formula $\gamma^{(a)}_{\rm unbiased} \approx \alpha - 1$ holds even for the realistic sample sizes. (d), (e) Parameter distribution of $\beta^{(a)}_1$ and $\beta^{(a)}_2$, the coefficients of the regression formula (23) in constructing the unbiased estimator $\gamma^{(a)}_{\rm unbiased}$. The means of $\beta^{(a)}_1$ and $\beta^{(a)}_2$ were given by 0.592 and 0.147, respectively.

To confirm this character, we have numerically generated the order-sign sequences by the LMF model with realistic parameters of our data set: we measured the model parameter set $(N_{\rm ST}, \alpha, N_\epsilon)$ for all sample points according to the method in Sec. IV B 3. Under the measured parameter sets, we numerically performed the Monte Carlo simulations of the LMF model. The scatterplots are generated 100 times as IID realizations, and take their ensemble average based on the bootstrap method to draw the final scatterplot.

Under realistic parameter sets, as shown in Figs. 8(b) and 9(b), we find the systematic deviation between the theoretical line $\gamma = \alpha - 1$ and the numerical data points. This suggests the NLLS estimator has finite-sample-size bias. In addition, we find that the convergence speed is very slow for finite N_ϵ and is not even uniform in terms of α. It is reasonable that the convergence is nonuniform in terms of α. Indeed, the decay speed of the ACF is so fast for larger α that the power-law part of the ACF cannot be observed for a wide range of τ. For this practical reason, we have restricted our analyses to the range $\alpha \in (1, 2)$, which agrees with the standard assumption of the LMF model.

By focusing on the range $\alpha \in (1, 2)$, let us apply the linear regression between $\gamma_{\rm NLLS}$ and α as shown in Fig. 8(b) according to the formula

$$\gamma_{\rm NLLS} = \beta_1(\alpha - 1) + \beta_2, \tag{23}$$

where β_1 and β_2 are regression coefficients. This relation is used to numerically construct an unbiased estimator [see Figs. 8(c) and 9(c) for the ACF and PSD methods, respectively] as shown in Sec. V C.

If the NLLS were an unbiased estimator, the relations $\langle\beta_1\rangle \approx 1$ and $\langle\beta_2\rangle \approx 0$ would hold. To test these relations, we repeated the numerical simulations of the LMF model and linear regressions (23) to obtain the empirical histograms of β_1 and β_2 [see Figs. 8(d) and 8(e) for the ACF method, and Figs. 9(d) and 9(e) for the PSD method]. We numerically find that $\langle\beta^{(a)}_1\rangle = 0.592$ and $\langle\beta^{(a)}_2\rangle = 0.147$ for the ACF method and $\langle\beta^{(s)}_1\rangle = 0.753$ and $\langle\beta^{(s)}_2\rangle = 0.083$ for the PSD method in our simulations. These values clearly show the biasedness of the NLLS estimator due to the finite sample size. To solve

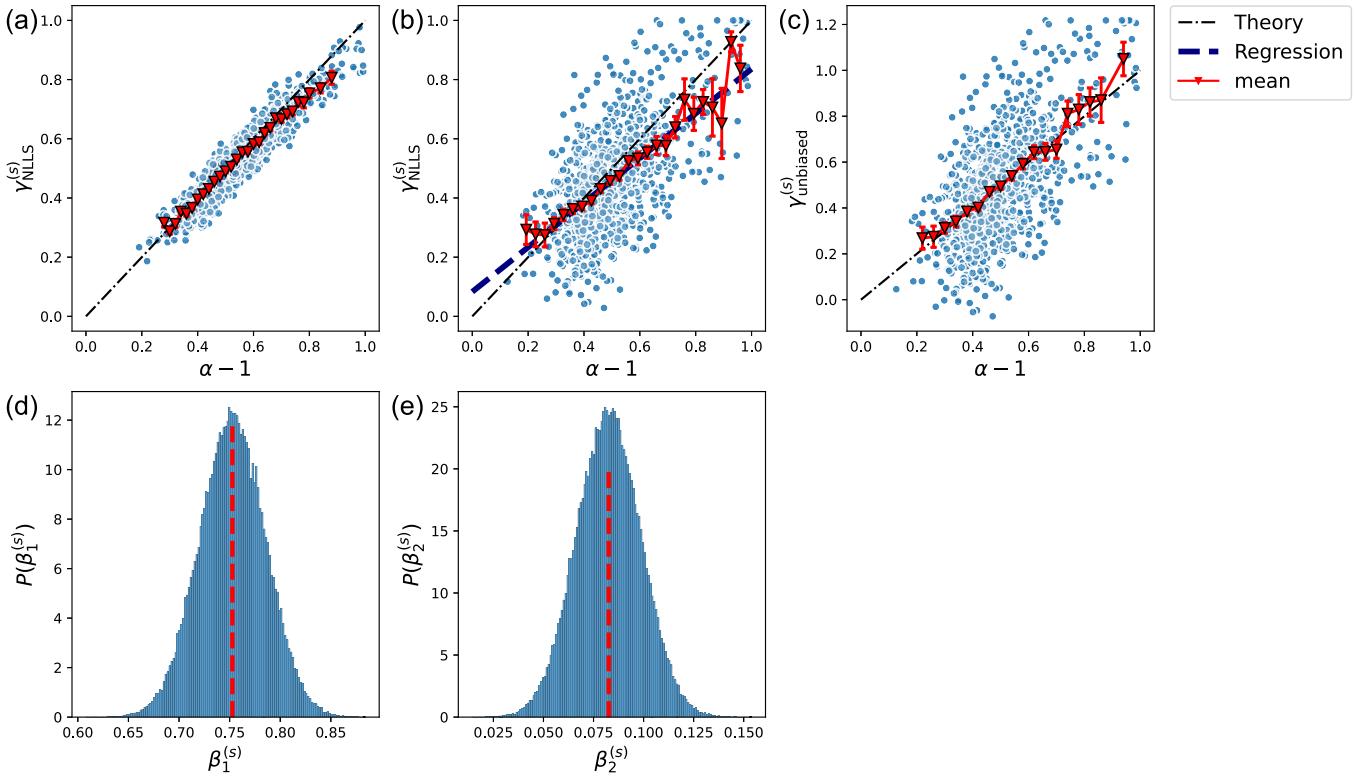

FIG. 9. Numerical tests for consistency and biasedness of the NLLS estimator $\gamma^{(s)}_{\rm NLLS}$ based on the PSD method. The test was based on the LMF model. (a) Consistency of $\gamma^{(s)}_{\rm NLLS}$ for a sufficiently large sample size $N_\epsilon = 10^8$. The parameter sets $(N_{\rm ST}, \alpha)$ are based on the measurement of our data set. (b) For realistic sample sizes $N_\epsilon \lesssim 10^7$, the PSD-NLLS estimator $\gamma^{(s)}_{\rm NLLS}$ exhibits biasedness due to the finite sample size. The LMF model was simulated by assuming that the parameter sets $(N_{\rm ST}, N_\epsilon, \alpha)$ are identical to those measured in our data set. We applied the linear regression (23) for the range $\alpha - 1 \in (0, 1)$ to obtain the navy line. (c) Confirmation of the numerically constructed unbiased PSD estimator $\gamma^{(s)}_{\rm unbiased}$. The unbiased estimator is constructed by $\gamma^{(s)}_{\rm unbiased} := (\gamma^{(s)}_{\rm NLLS} - \beta^{(s)}_2)/\beta^{(s)}_1$. As expected, $\gamma^{(s)}_{\rm unbiased} \approx \alpha - 1$ holds even for the realistic sample sizes. (d), (e) Parameter distribution of $\beta^{(s)}_1$ and $\beta^{(s)}_2$, the coefficients of the regression formula (23) in constructing the unbiased estimator $\gamma^{(s)}_{\rm unbiased}$. The means of $\beta^{(s)}_1$ and $\beta^{(s)}_2$ were given by 0.753 and 0.083, respectively.

this finite-sample-size bias problem, we will approximately construct a numerical unbiased estimator in Sec. V C.

3. *Other methods: The detrended fluctuation analysis*

There are several other methods to estimate the power-law exponent γ, and one of the famous methods is based on the Hurst exponent with the detrended fluctuation analysis (DFA) [31]. Indeed, some researchers claim that the DFA analysis provides much better results than the NLLS estimation [10,19]. In this paper, the estimated exponent by the DFA is called the DFA estimator and is denoted by $\gamma_{\rm DFA}$.

We do not use the DFA estimator because we numerically find a serious problem of the DFA estimator in terms of the finite-sample-size bias. We numerically generated the order-sign sequences by the LMF model and measured $\gamma_{\rm DFA}$ to obtain Fig. 10. The sample size is set to be $N_\epsilon = 10^8$ because the NLLS estimator showed consistency under this sample size.

The DFA estimator is numerically implemented by using the referred PYTHON package provided by Ref. [32]. Remarkably, the DFA estimator $\gamma_{\rm DFA}$ systematically deviates from the theoretical line (3). Since the theoretical line (3) is the exact solution for the LMF model, this deviation signifies the serious bias of the DFA estimator. Unfortunately, within

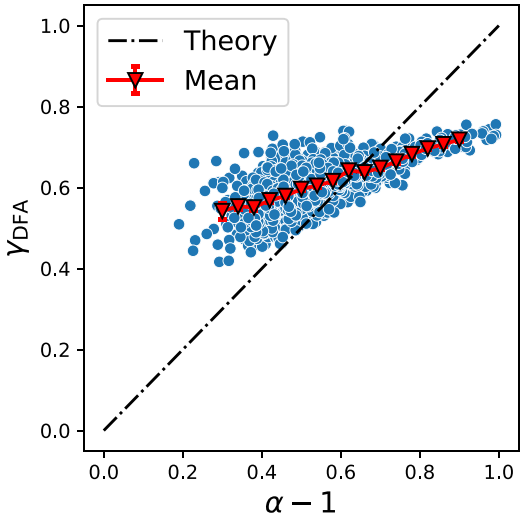

FIG. 10. Inconsistency of the DFA estimator $\gamma_{\rm DFA}$ even for the LMF model. We used the same parameters $(N_{\rm ST}, \alpha)$ as those measured in our data set and set $N_\epsilon = 10^8$. The theoretical relationship $\gamma_{\rm DFA} = \alpha - 1$ does not hold even for large sample size, rejecting the numerical consistency of the DFA estimator $\gamma_{\rm DFA}$ at least under a realistic computational resource.

our computational resource, we could not even confirm the consistency of the DFA estimator for a larger sample size.

We are not sure about its main reason currently, but one of the potential reasons might be related to the stronger statistical assumption required by the DFA estimator. While the consistency of the DFA estimator was recently proved for the fractional Brownian motion [33], it is nontrivial whether the consistency of the DFA estimator is still kept even for systems not obeying the fractional Brownian motion. In our case, there is no solid reason why the sign sequence generated by the LMF model can be regarded as the fractional Brownian motion. On the contrary, the NLLS estimator relies only on the ergodic assumption: the sample ACF converges to the true ACF for a large sample size. In this sense, the NLLS estimator requires weaker statistical assumptions than the DFA estimator. This might be a potential reason causing the difference between γ_{NLLS} and γ_{DFA}.

Here we do not claim inappropriateness of the DFA in a general context. However, since our aim is to verify the quantitative prediction (3) based on the LMF model, we should use a less biased estimator in terms of the scatterplot between γ and α at least for the LMF simulations. Thus, we do not use the DFA estimator in this paper.

C. Numerical construction of an approximate unbiased estimator

While the NLLS estimator numerically exhibits the consistency, it is biased for finite sample size with slow convergence speed. This problem should be solved before the direct verification of the LMF prediction (3). In this subsection, we approximately construct an unbiased estimator based on the LMF model.

Our idea for our unbiased estimator is based on our numerical observation of the scatterplot in Figs. 8(b) and 9(b) for the ACF and PSD methods, respectively. For the numerical simulations of our LMF model, we numerically find that γ_{NLLS} follows the linear regression relation (23) for the range $\alpha \in (1, 2)$ at least approximately. Since the relation (3) holds for the LMF model, the approximate relation $\gamma_{\mathrm{NLLS}} \approx \beta_1 \gamma + \beta_2$ should hold between the NLLS estimator γ_{NLLS} and the true γ. Therefore, we numerically construct an unbiased estimator $\gamma_{\mathrm{unbiased}}$ as

$$\gamma_{\mathrm{unbiased}} := \frac{\gamma_{\mathrm{NLLS}} - \beta_2}{\beta_1}, \tag{24}$$

which exhibits the approximate unbiasedness, at least for the LMF simulations [see Figs. 8(c) and 9(c) for the ACF and PSD methods, respectively], as

$$\langle \gamma_{\mathrm{unbiased}} \rangle \approx \gamma. \tag{25}$$

For the verification of the LMF prediction (3) in Sec. VI, we use the unbiased estimator $\gamma_{\mathrm{unbiased}}$.

VI. VERIFICATION OF THE LMF PREDICTION

Let us proceed with the main result of this paper, namely the direct verification of the LMF prediction (3). We then discuss the relationship to previous works, possible future implications, and some open questions.

A. Scatterplot about the power-law exponent

For the verification of the LMF prediction (3), we plot the scatterplot between α and $\gamma_{\mathrm{unbiased}}$ for $\alpha \in (1, 2)$ [see Figs. 11(a) and 11(d) for the ACF and PSD methods, respectively]. We set bins along the α axis and plotted the average $\gamma_{\mathrm{unbiased}}$ within each bin. The average line (red) agrees with the theoretical line (black) well, strongly supporting the validity of the LMF prediction even at the quantitative level. We also provide boxplots in Figs. 11(b) and 11(e) for the ACF and PSD methods, respectively, where the statistical quantities, such as the first, second, and third quartiles, are calculated within each bin along the α axis. Furthermore, we provide the empirical PDF of the errors $\eta := \alpha - 1 - \gamma_{\mathrm{unbiased}}$ as shown in Figs. 11(c) and 11(f) for the ACF and PSD methods, respectively. Since the average error is small, $\langle \eta^{(a)} \rangle = 0.03$ for the ACF method and $\langle \eta^{(s)} \rangle = 0.003$ for the PSD method, respectively, our statistical analysis is self-consistent.

For reference, the original scatterplot between α and γ_{NLLS} (i.e., the consistent but biased estimator) is provided in Appendix F. In addition, we provide the scatterplots aggregated for every three years between α and $\gamma_{\mathrm{unbiased}}$ as a robustness check (i.e., threefold cross-validation) in Appendix G.

We have shown that the power-law exponent γ in the ACF $C(\tau)$ is directly related to the microscopic power-law exponent α in the metaorder-length PDF $\rho(L)$. Since α is not observable from public data, our result implies that the LMF theory is useful for statistical estimation of microscopic parameters from public data.

B. Discussion 1: Estimation of the total number of traders

Since we show the feasibility of statistical estimation of α from the ACF power-law exponent γ, it is a natural idea to infer other microscopic quantities from the ACF prefactor c_0. In this subsection, we discuss the estimation of the total number of STs N_{ST} from the ACF prefactor c_0 based on the LMF theory.

1. Review of the original LMF theory on the prefactor c_0

The LMF theory predicts that the ACF prefactor should be given by

$$c_0^{\mathrm{LMF}} := \frac{1}{\alpha N_{\mathrm{ST}}^{2-\alpha}} \tag{26}$$

on the assumption that the intensity distribution $\{\lambda^{(i)}\}_{i \in \Omega_{\mathrm{ST}}}$ among the order-splitting traders is uniform, such that

$$\lambda^{(i)} = \frac{1}{N_{\mathrm{ST}}} \quad \text{for any } i \in \Omega_{\mathrm{ST}}. \tag{27}$$

This prediction is applicable to inferring the total number of the order-splitting traders N_{ST}, such that

$$N_{\mathrm{ST}} \approx N_{\mathrm{ST}}^{\mathrm{LMF}}(c_0, \gamma) := \left[\frac{1}{(\gamma + 1) c_0} \right]^{\frac{1}{1-\gamma}}, \tag{28}$$

where the right-hand side is composed of publicly available quantities from the sample ACF or PSD. Let us call $N_{\mathrm{ST}}^{\mathrm{LMF}}(c_0, \gamma)$ the LMF estimator for the total number of the STs. Since N_{ST} is not observable from public data, the prediction (28) is appealing from both academic and practical viewpoints.

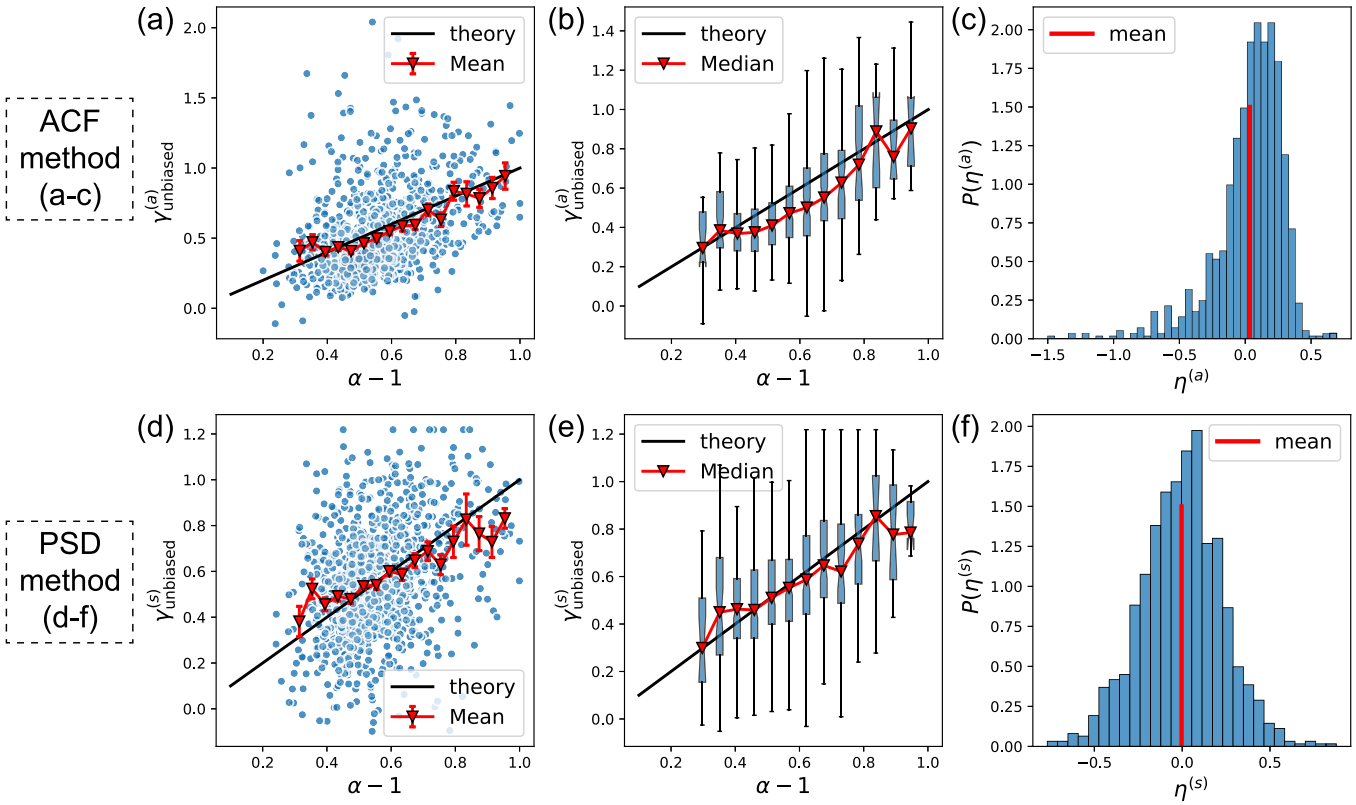

FIG. 11. Direct verification of the LMF prediction (3) based on the ACF method (a)–(c) and the PSD method (d)–(f). (a), (d) Scatterplot between α and $\gamma_{\rm unbiased}$. We set bins along the α axis, and the red line signifies the averages within the bins. (b), (e) Box plot between α and $\gamma_{\rm unbiased}$, where statistical quantities [e.g., the first, second (median), and third quartiles] are calculated within the bins along the α axis. (c), (f) Empirical PDF of the errors $\eta := \alpha - 1 - \gamma_{\rm unbiased}$. The average error is very small such that $\langle \eta^{(a)} \rangle = 0.03$ for the ACF method and $\langle \eta^{(s)} \rangle = 0.003$ for the PSD method, respectively.

Note that the LMF estimator has the singularity $\gamma = 1$ at which the estimation fails in principle. Therefore, it is more realistic to study

$$\left[N_{\rm ST}^{\rm LMF}(c_0, \gamma)\right]^{1-\gamma} = \frac{1}{(\gamma+1)c_0} \tag{29}$$

by removing the singularity at $\gamma = 1$.

2. *Review of a generalized LMF theory on the prefactor c_0*

During our data analysis, however, we noticed that the assumption (27) for homogeneous intensities is very unrealistic because time intervals between submissions are broadly distributed in our data set. Furthermore, in Ref. [27], the authors recently proposed a generalized LMF model by incorporating the inhomogeneous intensities and clarified the following points:

(i) Let us assume that all traders are order-splitting traders, but their intensity distribution is nonuniform, such that $\lambda^{(i)} \neq \frac{1}{N_{\rm ST}}$ for some $i \in \Omega_{\rm ST}$.

(ii) The ACF power-law exponent formula (3) robustly holds for any intensity distributions $\{\lambda^{(i)}\}_{i\in\Omega_{\rm ST}}$.

(iii) On the other hand, the ACF prefactor formula (26) is very sensitive to the system-specific details and does not hold anymore for general $\{\lambda^{(i)}\}_{i\in\Omega_{\rm ST}}$. Instead, the prefactor formula is replaced with

$$c_0^{\rm SK} := \frac{1}{\alpha} \sum_{i\in\Omega_{\rm ST}} (\lambda^{(i)})^{3-\alpha}. \tag{30}$$

(iv) Furthermore, the homogeneous LMF formula (26) systematically underestimates the actual prefactor, in the sense that

$$c_0^{\rm SK} \geqslant c_0^{\rm LMF}. \tag{31}$$

These heterogeneous LMF results imply that the LMF estimator $N_{\rm ST}^{\rm LMF}(c_0, \gamma)$ provides a lower bound of the true $N_{\rm ST}$:

$$N_{\rm ST}^{\rm LMF}(c_0, \gamma) \lesssim N_{\rm ST}. \tag{32}$$

3. *Scatterplot between $N_{\rm ST}^{\rm LMF}(c_0, \gamma)$ and $N_{\rm ST}$*

On the basis of the above theoretical predictions, we drew the scatterplot Fig. 12 between the true $\log_{10}(N_{\rm ST})^{1-\gamma}$ and the LMF estimator $\log_{10}(N_{\rm ST}^{\rm LMF})^{1-\gamma}$ after the finite-sample-size bias is removed (see Appendix H for details). Since we should stick to empirically available quantities, the estimators are based only on $\gamma_{\rm NLLS}$ and $c_{0,\rm NLLS}$ for both ACF and PSD methods. The ACF and PSD methods were employed Figs. 12(a) and 12(b) and Figs. 12(c) and 12(d), respectively, and they showed consistent results. We observed that the LMF estimator $N_{\rm ST}^{\rm LMF}$ is highly correlated with the true $N_{\rm ST}$. This

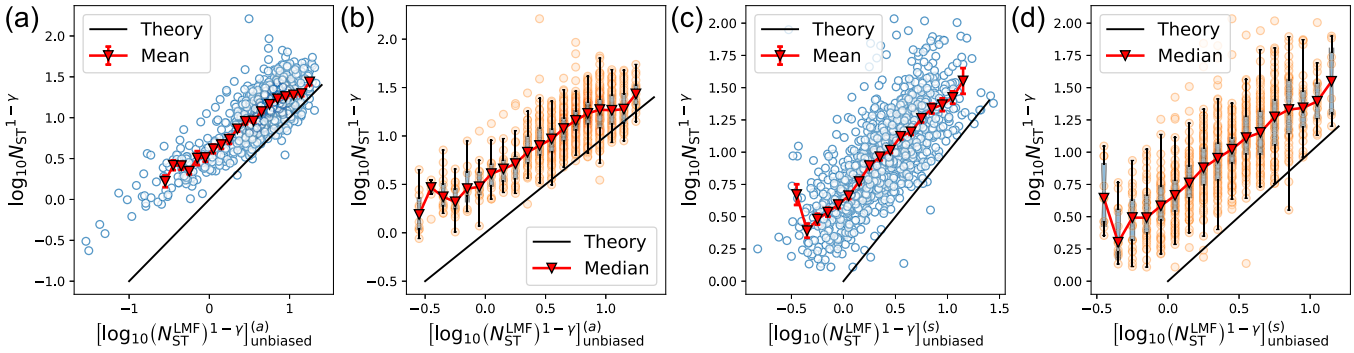

FIG. 12. Estimation of the total number of STs via the LMF theory for the ACF (a), (b) and PSD (c), (d) methods. (a), (c) Scatterplots between the unbiased LMF estimators $[\log_{10}(N_{\rm ST}^{\rm LMF})^{1-\gamma_{\rm NLLS}}]_{\rm unbiased}$ and the actual values of $\log_{10}(N_{\rm ST})^{1-\gamma_{\rm NLLS}}$. (b), (d) Corresponding boxplots. We find that the LMF estimator is strongly correlated with the actual $N_{\rm ST}$. However, the LMF estimator systematically underestimated the actual $N_{\rm ST}$, which is consistent with the theoretical inequality (32) for a generalized LMF model in [27].

fact implies that the ACF prefactor c_0 has potentially useful information on $N_{\rm ST}$ in principle. On the other hand, the LMF estimator $N_{\rm ST}^{\rm LMF}$ systematically underestimates the actual value of $N_{\rm ST}$, which is consistent with the theoretical expectation (32) for the heterogeneous LMF model. Therefore, the LMF estimator $N_{\rm ST}^{\rm LMF}$ should be interpreted as a lower bound of the total number of STs.

We thus conclude that the LMF theory qualitatively works even for the estimation of $N_{\rm ST}$ only from public data. However, its theoretical estimation is systematically biased due to traders' heterogeneity in order-splitting strategies. In this sense, the strategies' heterogeneity needs to be considered for a more quantitative estimation of $N_{\rm ST}$.

C. Discussion 2: Relation to previous results

In the original article [19], Lillo, Mike, and Farmer showed a scatterplot between α and γ by using the off-book market data as an imperfect proxy. While their figure does not statistically reject the LMF prediction (3) due to small sample size, it does not strongly support the validity of the strong LMF prediction (3) at the quantitative level. On the contrary, we have provided clear statistical evidence in Fig. 11 with enough sample size, which strongly supports the validity of the LMF prediction, even at the quantitative level. Furthermore, we have successfully demonstrated that the ACF prefactor has information on the total number of the order-splitting traders $N_{\rm ST}$ as shown in Fig. 12.

There are two more technical advantages of our statistical method than the previous one. The first advantage is that we directly estimated the metaorder-length (run-length) distribution at the level of individual traders, instead of the metaorder-volume distribution. This is in contrast to the statistical analysis in Ref. [19], which is based on the metaorder-volume distribution in the absence of an appropriate data set. When one uses the metaorder-volume distribution to estimate α, one has to assume that STs split their orders with constant volume for statistical analyses. However, this assumption is not realistic because volume specified by a market order is known to obey the power-law distributions empirically. On the other hand, we utilized a proper data set and directly measured the metaorder-length distribution. This analysis does not require the assumption of the constant volume splitting. In this sense, we believe that our estimated exponent α would be more reliable for the calibration to the LMF model.

The second advantage is that we used the NLLS-based unbiased estimator $\gamma_{\rm unbiased}$, instead of the DFA estimator $\gamma_{\rm DFA}$. In econophysics, several researchers use the Hurst-exponent analysis based on the DFA to measure γ, and the scatterplot of Ref. [19] (see Fig. 5) is also based on $\gamma_{\rm DFA}$. However, we find that $\gamma_{\rm DFA}$ has a serious problem in terms of the finite-sample-size bias at least in our numerical LMF simulations, and we have concluded that $\gamma_{\rm DFA}$ is an inappropriate estimator in validating the LMF model. Since $\gamma_{\rm DFA}$ is not consistent with the LMF prediction (3), we believe that our statistical estimation has a much greater advantage than the previous one.

D. Discussion 3: Implication for liquidity measurement

Our result strongly supports the validity of the LMF model even at the quantitative level (3). Since the LMF model is based on the order-splitting hypothesis, we believe that our result is relevant to quantitative measurement of the market liquidity from a different angle.

According to the order-splitting hypothesis, traders split their large metaorders in the lack of revealed liquidity: the volumes at the best prices are too small compared with the metaorder volume, and traders have no choice but to split their orders into pieces. In the LMF model, the order book for markets with smaller α and large $N_{\rm ST}$ is not thick enough for many institutional investors to immediately execute metaorders. Thus, the parameter set $(\alpha, N_{\rm ST})$ characterizes the illiquidity of markets regarding metaorder splittings. Particularly, $N_{\rm ST}$ characterizes how many institutional investors are waiting for the order books to replenish during their order splitting, which might have significant meaning for platform managers. Such an aspect of liquidity shortage is not measured by traditional liquidity measures, such as market spread and market impact. We believe that it might be interesting to develop some liquidity measures based on the order-splitting hypothesis as another direction.

E. Discussion 4: Open questions on statistical analyses

We approximately measured the power-law exponent of the sign ACF by the NLLS-based unbiased estimator $\gamma_{\rm unbiased}$.

While we believe that this estimator is practically reliable at least for our data set, we are not sure whether this is always the best option. Indeed, the approximate construction of the unbiased estimator is based on the numerical observation of the approximate linear regression relation (23) for $\alpha - 1 \in (0, 1)$, which is not theoretically proved yet. In addition, this construction of unbiased estimators may depend on the selection of underlying microscopic models (i.e., the LMF model in our case). Seeking the optimal unbiased estimator is an urgent topic as a technical issue in statistics.

In addition, we found a serious problem of the DFA estimator γ_{DFA} in terms of the finite-sample-size bias. Since we are not sure about its critical reason, this problem should be explored more deeply from the viewpoint of statistical analyses. In particular, we are interested in its robustness in terms of the consistency: i.e., does γ_{DFA} coincide with the true γ for the infinite sample size, even if the time series is generated by some microscopic model, instead of the fractional Brownian motion? In any case, the long-memory character of the LRC is a huge obstacle for statistical analyses, and thus development of statistical methods will be important.

VII. CONCLUDING REMARKS

While the LMF model has been a cornerstone to support the order-splitting hypothesis, its prediction (3) has not been verified at the quantitative level. In this paper, we have quantitatively established the validity of the LMF prediction (3) by analyzing a large data set of the TSE market over nine years. We first identified the RTs and STs by clustering analysis, and we measured the microscopic power-law exponent α in terms of the metaorder length for STs. We then developed a statistical method to measure the power-law exponent γ in the sign ACF. The scatterplot between α and γ is provided as the main result, strongly supporting the validity of the LMF prediction (3). Furthermore, we discuss a practical method to estimate the total number of order splitters from the ACF prefactor on the basis of the LMF theory.

Our study builds upon the stream of ecological analyses of financial markets, which is based on the trading-strategy clustering at the level of individual traders. In the literature, one of the pioneering studies on trading-strategy clustering was provided in Ref. [34] in 2012 by focusing on market-order submissions. As for limit-order submissions, strategy clustering was first provided by Refs. [5–8] for the EBS FX market in 2018 (i.e., regarding trend-following behavior), and it was also provided by Ref. [21] for the TSE market in 2019 (i.e., regarding market-making behavior). In this work, we classify traders into RTs and STs in terms of market-order submissions. It will be interesting to investigate the roles of RTs and STs in the ecology of the TSE market. We believe that this research direction would be promising in developing market microstructure for the future.

Finally, we remark on the availability of our dataset. The data supporting the findings of this study were provided by the JPX Group, Inc., and restrictions apply to the availability of these data, which were used under license for our projects. The authors are not allowed to distribute the data without the explicit permission of the JPX Group, Inc.

ACKNOWLEDGMENTS

Y.S. was supported by JST SPRING (Grant No. JPMJSP2110). K.K. was supported by JST PRESTO (Grant No. JPMJPR20M2), JSPS KAKENHI (Grants No. 21H01560 and No. 22H01141), and JSPS Core-to-Core Program (Grant No. JPJSCCA20200001). We greatly appreciate the data provision and careful review of this paper by the JPX Group, Inc.

Y.S. contributed the numerical and empirical analyses by program coding. K.K. designed the research plot and supervised the project. Y.S. and K.K. wrote the manuscript and agree with all the findings.

We declare no financial conflict of interest. The JPX Group, Inc. provided the original data for this study without any financial support.

APPENDIX A: INTRADAY SEASONALITY

In our data analysis, we excluded the data during the periods around the opening and closing auctions. This exceptional rule is applied to avoid the intraday-seasonality effect, which is a stylized fact in various financial markets [3]. In this Appendix, we show the statistical evidence of the intraday seasonality in TSE, called the U-shape profile of the temporal market-order activity, to justify our exceptional rule.

In this Appendix, we use the physical time t (minutes), representing the elapsed time from the starting time of the morning continuous double auction (9:00 JST) with the lunch break (11:30–12:30 JST) excluded. For example, $t = 0$ represents 9:00 JST, $t = 150$ represents 11:30 JST, $t = 151$ represents 12:31 JST, and $t = 300$ represents 15:00 JST.

Let us focus on Toyota Motor Corporation in 2020. The daily-total number of market orders is written as $N_{\mathrm{MO}}^{\mathrm{tot}}$ and the number of market orders during $[t, t + 1)$ is written as $N_{\mathrm{MO}}(t)$. The temporal market-order ratio is then defined by $r_{\mathrm{MO}}(t) := N_{\mathrm{MO}}(t)/N_{\mathrm{MO}}^{\mathrm{tot}}$. In Fig. 13, we plotted the yearly average of the temporal market-order ratio $r_{\mathrm{MO}}(t)$. This figure shows that the market-order submissions are active around the opening and closing times of the continuous double auctions (i.e., $t = 0$, 150, and 300), consistently with the empirical "U-shape profiles" in previous reports [3].

APPENDIX B: NUMERICAL IMPLEMENTATION OF THE RANDOM INTEGER NUMBER OBEYING A POWER LAW

Here we describe the numerical method to generate random integer numbers obeying a power-law relation

$$P(L) \propto L^{-\alpha-1} \quad \text{for large } L \tag{B1}$$

with an exponent $\alpha > 1$. Let us consider a continuous positive random number $x \in [1, \infty)$, which obeys the continuous Pareto distribution

$$P(x) = \alpha x^{-\alpha-1} \quad (x \in [1, \infty)). \tag{B2}$$

The Pareto random number x can be generated by

$$x := \frac{1}{(1-u)^{1/\alpha}} \tag{B3}$$

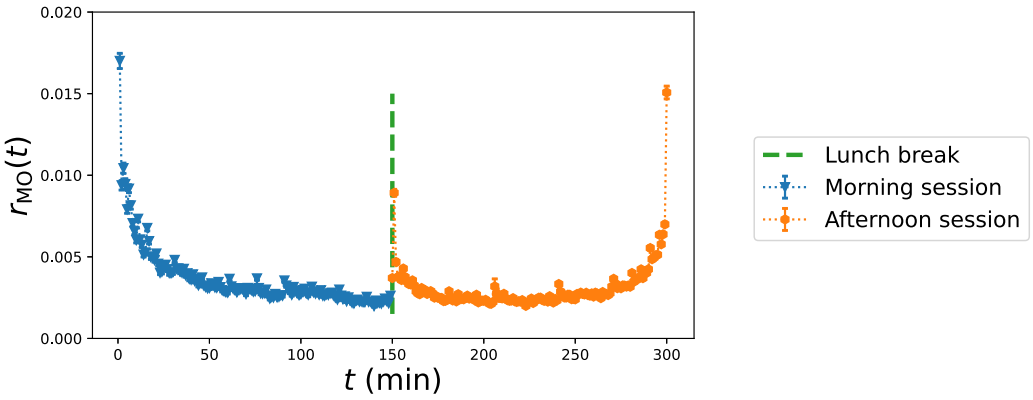

FIG. 13. Temporal market-order activity of Toyota 2020. The yearly average of the temporal market-order ratio r_{MO} is plotted to show the U-shape profile in its intensity: i.e., the transactions are active near the opening and closing times of the continuous double auction $t = 0$, 150, and 300. Note that $r_{MO}(t)$ is defined by $r_{MO}(t) := N_{MO}(t)/N_{MO}^{tot}$ with the daily-total number of market orders N_{MO}^{tot} and minutely total number of market orders $N_{MO}(t)$.

with a uniform random number $u \in [0, 1)$. Finally, the integer random number L is given by

$$L := \lfloor x \rfloor, \tag{B4}$$

where the floor function $\lfloor x \rfloor$ signifies the maximum integer not larger than x.

APPENDIX C: SUMMARY OF THE TECHNICAL PROBLEMS TO BE SOLVED

The validation steps for the LMF prediction are rather straightforward. Why has this relationship (7) not been verified yet? In our view, there are three technical problems in proving the quantitative prediction (7). Let us briefly summarize these technical problems one by one.

The first problem would be the scarcity of necessary high-quality data. To measure α, we are required to identify order-splitting traders (STs) at the level of individual traders by applying strategy clustering, and then measure the empirical PDF of the runs as $P(L)$. For example, let us write the set of STs as Ω_{ST}. The empirical PDF of the runs is obtained as

$$P(L) \propto \sum_{i \in \Omega_{ST}} \sum_{k} \delta\big(L - L_k^{(i)}\big). \tag{C1}$$

Since the STs Ω_{ST} and their run sequences $\{L_k^{(i)}\}_{k, i \in \Omega_{ST}}$ are the necessary inputs for the run PDF $P(L)$, we have to analyze the data sets enabling us to track orders at the level of trader accounts. However, such high-quality data are very scarce in terms of the data availability.

The second problem would be the necessary data size. While there are a few studies analyzing account-level data sets, the necessary data size would need to be huge in verifying Eq. (7). Indeed, the inputs of the scatterplot are the power-law exponents α and γ, and their accurate measurements are not easy: theoretically, they are expected to distribute typically within the range $\alpha \in (1, 2)$ and $\gamma \in (0, 1)$. Therefore, it would be necessary to control their estimation errors roughly less than 0.1. In particular, the accurate estimation of γ is very hard. Assuming that the data points of $C(\tau)$ are necessary to cover the range $\tau \in (10^1, 10^3)$, we have to suppress the noise in the ACF at a low level even at $\tau \sim 10^3$. Through our numerical simulations of the LMF model, we estimate that a long order-sign sequence $\{\epsilon(t)\}_t$ is necessary, such as $N_\epsilon > 5 \times 10^5$ at least, even to obtain one data point (α, γ) in the scatterplot.

The third problem is related to the fact that the LMF model belongs to the long-memory process (see Chap. 10 in Ref. [3]), implying that the convergence speed of any sample mean is slower than usual in terms of the sample size N_ϵ. Indeed, the long-memory character $C(\tau) \approx \tau^{-\gamma}$ with $\gamma \in (0, 1)$ suggests that $\epsilon(t)$ and $\epsilon(t + \tau)$ are not statistically independent with each other even for large $\tau > 0$. Thus, the estimation of γ will require a large data set from such a theoretical viewpoint.

To overcome such technical difficulties, in this paper we analyze a large TSE data set provided by the JPX Group, Inc. These data not only include the account-level information (i.e., the virtual-server IDs), but they also cover all the stocks over nine years. We then finally report the first verification of the quantitative LMF prediction (7) from the viewpoint of the big-data analysis.

APPENDIX D: METAORDER-LENGTH DISTRIBUTION FOR RTs

In Sec. IV B 1, we regard any trader as an RT when the binomial test was not rejected with the significance level $\theta = 0.01$. In the standard theory of statistical tests, it is often emphasized that passing tests does not necessarily mean the acceptance (proof) of the null hypothesis, and we should not draw hasty conclusions: the error of the first kind (false-positive rate) is controlled within the significance level θ for the rejection, but the error of the second kind (false-negative rate) is not controlled for the "acceptance" in the statistical tests. In this sense, while our clustering method is expected to be reasonable in identifying the set of STs within the significance level, the identification of the set of RTs might be incomplete; some small part of non-RTs, such as STs, might be included even in the RT cluster since we did not control the error of the second kind.

While we acknowledge this possible incompleteness of our clustering method for RTs, it would be helpful to check

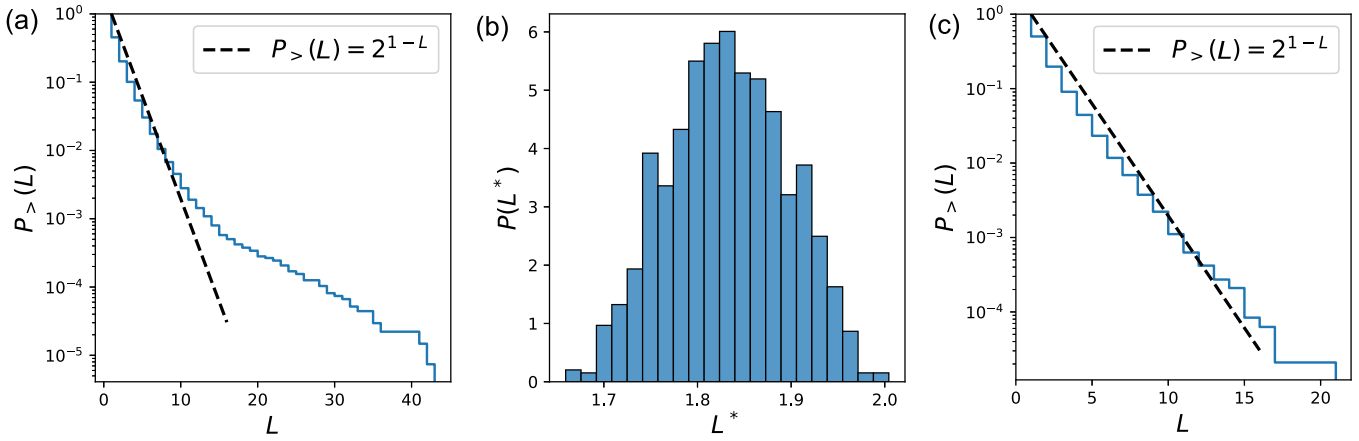

FIG. 14. Characters of metaorder-length distribution on RTs. (a) The metaorder-length (run-length) CCDF of all the RTs for Toyota Motor Corporation in 2020. The run-length CCDF obeys the exponential law at least for the body part $L \lesssim 10$, such that $P_>(L) \sim 2^{1-L}$, as theoretically expected. At the same time, we observed a theoretical discrepancy for the tail because the second-kind statistical error was not controlled in our binomial test; a small portion of STs might be included in the RT cluster. (b) The empirical PDF $P(L^*)$ of the decay length L^* in our data set, by assuming the exponential metaorder-length CCDF $P_>(L) \approx (L^*)^{1-L}$ for the RTs for each data point. The decay length L^* is measured by the maximum likelihood estimation (i.e., $L^* = \langle L^{\mathrm{RT}} \rangle$). The PDF has a sharp peak around $L^* \approx 1.8$, consistent with the theoretical prediction (D1). (c) The aggregated metaorder CCDF only for the active RTs (submitting more than 1000 orders annually), showing the exponential law without fat tails for Toyota 2020. This is an empirically successful symptom of reducing the second kind's error.

whether the set of RTs satisfies our theoretical expectation for reference. If the assumption of the null hypothesis (i.e., the symmetric Bernoulli process) is exactly correct, the CCDF of the run lengths for RTs should be given by the exponential distribution

$$P_>(L) = 2^{1-L} \tag{D1}$$

for any positive integer L. We check this character regarding the RT clusters.

We studied the metaorder-length distribution for the RTs: we consider the joint run-length sequences for RTs and the corresponding empirical metaorder-length CCDF:

$$\begin{aligned} \{L_k^{\mathrm{RT}}\}_k &:= \bigcup_{i \in \Omega_{\mathrm{RT}}} \{L_k^{(i)}\}_k, \\ P_>(L^{\mathrm{RT}}) &:= \frac{N_>(L^{\mathrm{RT}})}{|\{L_k^{\mathrm{RT}}\}_k|}, \\ N_>(L^{\mathrm{RT}}) &:= \int_{L^{\mathrm{RT}}}^{\infty} dy \sum_k \delta\big(y - L_k^{\mathrm{RT}}\big). \end{aligned} \tag{D2}$$

The empirical metaorder CCDF for the RTs is plotted in Fig. 14(a) for Toyota Motor Corporation in 2020. This plot shows that the metaorder CCDF exhibits the exponential law $P_>(L) \approx 2^{1-L}$ for the body part $L \lesssim 10$. In addition, the estimated decay length L^* shows a sharp peak around $L^* \approx 1.8$ based on the maximum-likelihood estimation $L^* = \langle L^{\mathrm{RT}} \rangle$ for the exponential law $P_>(L^{\mathrm{RT}}) \approx (L^*)^{1-L^{\mathrm{RT}}}$ [see Fig. 14(b)]. This result shows a minimum self-consistency of our clustering algorithm with $L^* \approx 2.0$. At the same time, we observe the discrepancy from the exponential law for the tail part $L \gtrsim 10$. This discrepancy is reasonable because we did not control the statistical error of the second kind, and a small portion of STs might be included in the RT cluster.

Improvement of our clustering algorithm is a future open issue regarding the RTs, and applying some filters would be desirable to control the second kind's error. As an initial attempt, we applied a simple filter by focusing only on active RTs submitting more than 1000 market orders a year (i.e., a few submissions everyday on average). We considered this filter a reasonable candidate, because a small portion of inactive RTs seemingly submitted large metaorders only a few times during the year while they behaved as RTs during most of the time. The aggregated metaorder CCDF only for the active RTs is plotted in Fig. 14(c) for Toyota 2020, where the discrepancy at the tail disappears. We checked all the stocks in 2012 and 2020 by eye and found similar observations.

APPENDIX E: DETAILED IMPLEMENTATION OF THE NONLINEAR LEAST SQUARES

In this Appendix, we describe the measurement of the power-law exponent γ and the prefactor c_0 in the sign ACF $C(\tau) \simeq c_0 \tau^{-\gamma}$ for large τ. Our methods are based on the nonlinear least squares (NLLS) for the ACF and PSD.

1. NLLS estimators based on the sample ACF

We first describe the NLLS estimation for the empirical ACFs. The basic idea is to systematically fix the fitting range $[\tau_{\mathrm{th}}^-, \tau_{\mathrm{th}}^+]$, and then apply the power-law fitting $C(\tau) \propto \tau^{-\gamma_{\mathrm{NLLS}}^{(a)}}$ to the sample ACF for the range $[\tau_{\mathrm{th}}^-, \tau_{\mathrm{th}}^+]$ (see Fig. 15 for the scheme). The detailed process is given as follows:

a. Step 1: The sample ACF

The sample ACF is defined by

$$C_{\mathrm{sample}}(\tau) := \frac{1}{N_\epsilon - \tau} \sum_{t=1}^{N_\epsilon - \tau} \epsilon(t)\epsilon(t+\tau) \tag{E1}$$

for positive $\tau > 0$ by assuming the symmetry $\langle \epsilon(t) \rangle = 0$ for the range $\tau \in [1, 10^4]$. This symmetric assumption is

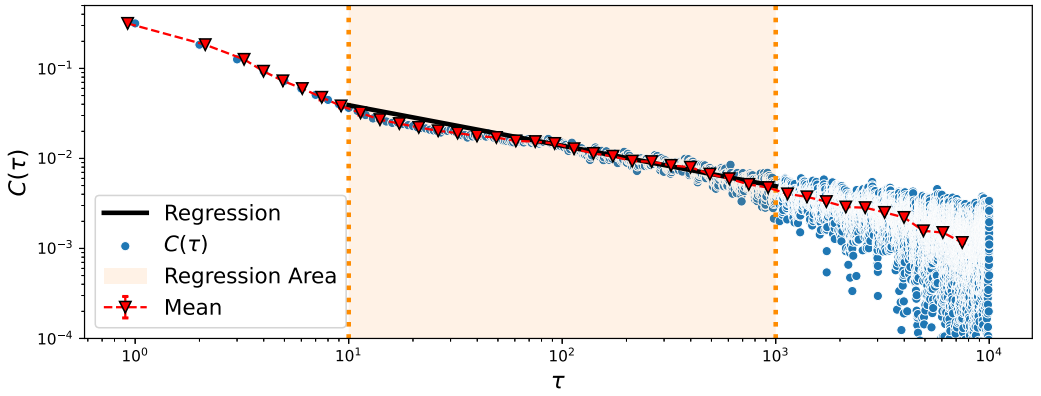

FIG. 15. Schematic figure of the ACF fitting based on the NLLS estimation for Toyota 2020. We used the orange area $\tau \in [\tau_{\rm th}^-, \tau_{\rm th}^+]$ for the fitting to obtain the power-law guideline with exponent $\gamma_{\rm NLLS}$ (solid black).

commonly used in other literature [18] and its validity is also checked in our data set. For nonpositive τ, we define $C(\tau) = 0$.

b. Step 2: The lower threshold

As reported in various data sets [3], the sample ACF initially exhibits a relatively rapid decay for small τ and the power-law decay follows for large τ in our data sets. To estimate the power-law exponent, it will be useful to estimate the lower bound $\tau_{\rm th}^-$ for the final fitting regime. This threshold is estimated as follows: let us first estimate the initial decay timescale $\tau_{\rm temp}$ by using the NLLS fitting of the sample ACF with tentative fitting function

$$C_{\rm model}(\tau) = C_{\rm temp}^{(0)} e^{-\tau/\tau_{\rm temp}} + C_{\rm temp}^{(1)} \tau^{-\gamma_{\rm temp}^{(1)}} \tag{E2}$$

with the temporary fitting parameters $C_{\rm temp}^{(0)}$, $C_{\rm temp}^{(1)}$, $\tau_{\rm temp}$, and $\gamma_{\rm temp}^{(1)}$ for the range $\tau \in [1, 10^3]$. These parameters are estimated by the relative least-squares error (RLS) method (see Appendix I).

Based of the tentative fitting formula (E2), we next fix the lower threshold $\tau_{\rm th}^-$ between the exponential and power-law decays as follows: since we would like to estimate the lower threshold $\tau_{\rm th}^-$ that satisfies $|C_{\rm temp}^{(0)} e^{-\tau_{\rm th}^-/\tau_{\rm temp}}| \ll |C_{\rm temp}^{(1)} (\tau_{\rm th}^-)^{-\gamma_{\rm temp}^{(1)}}|$, let us consider the area where the power-law part is dominant:

$$A_{\rm pow} := \left\{ \tau \;\middle|\; \left| \frac{C_{\rm temp}^{(0)} e^{-\tau/\tau_{\rm temp}}}{C_{\rm temp}^{(1)} \tau^{-\gamma_{\rm temp}^{(1)}}} \right| < \epsilon_{\rm th}, \;\; 10 \leqslant \tau \leqslant 10^2 \right\} \tag{E3}$$

with a small parameter $\epsilon_{\rm th} := 0.1$. The lower threshold $\tau_{\rm th}^-$ is defined by

$$\tau_{\rm th}^- := \begin{cases} \min_{A_{\rm pow}} \tau & \text{if } A_{\rm pow} \text{ is not empty,} \\ 10^2 & \text{if } A_{\rm pow} \text{ is empty.} \end{cases} \tag{E4}$$

c. Step 3: Logarithmic smoothing

The sample ACF exhibits fluctuations, particularly for large τ, due to the finite sample size. To remove such statistical fluctuations, we define the smoothed sample ACF:

$$C_{\rm smooth}(\tau) := \sum_{t=-\infty}^{\infty} w_\delta(\tau; t) C_{\rm sample}(t), \quad w_\delta(\tau; t) := \begin{cases} \dfrac{1}{\tau_{\rm smooth}^+(t) - \tau_{\rm smooth}^-(t)} & [t \in (\tau_{\rm smooth}^-(\tau), \tau_{\rm smooth}^+(\tau))], \\ 0 & [t \notin (\tau_{\rm smooth}^-(\tau), \tau_{\rm smooth}^+(\tau))]. \end{cases} \tag{E5a}$$

Since we are interested in the estimation of the power-law exponent, we use the logarithmic smoothing based on

$$\tau_{\rm smooth}^+(\tau) = \lfloor \tau \, 10^{+\delta/2} \rfloor, \quad \tau_{\rm smooth}^-(\tau) = \lfloor \tau \, 10^{-\delta/2} \rfloor \tag{E5b}$$

with the smoothing window size $\delta = 0.05$. This smoothing method is a discrete-time version of logarithmic smoothing for continuous time (see Appendix J).

d. Step 4: The upper threshold

It would be desirable to observe the power-law decay in the region of about two digits on the log-log ACF plot, such as by setting $\tau_{\rm th}^+ = 10^2 \tau_{\rm th}^-$. However, the sample ACF will be statistically insignificant for very large τ and such a naive setting of $\tau_{\rm th}^+$ might be inappropriate in general.

Indeed, even if the order-sign sequence were generated in a completely random manner (i.e., the white noise), the sample ACF could take nonzero values, such that $|C(\tau)| \simeq N_\epsilon^{-1/2}$, due to statistical errors. In this sense, if the absolute values of the sample ACF are smaller than $N_\epsilon^{-1/2}$, it is reasonable that the values of the sample ACF are regarded as statistically insignificant.

Based on this idea, we estimate an upper cutoff $\tau_{\rm stat}^+$ in terms of the statistical significance. The area where the

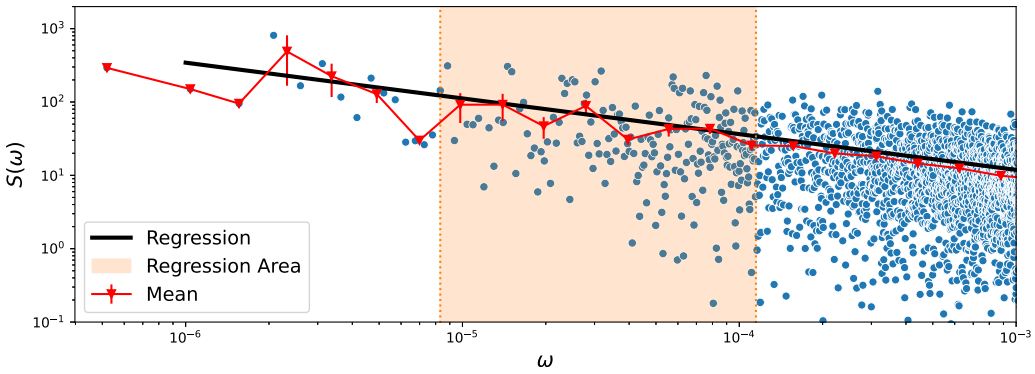

FIG. 16. Schematic figure of the PSD fitting based on the NLLS estimation for Toyota 2020. We used the orange area $\tau \in [\omega_{\rm th}^-, \omega_{\rm th}^+]$ for the fitting to obtain the power-law guideline with exponent $H_{\rm NLLS}$ (solid black).

smoothed ACF is statistically significant is estimated by

$$A_{\rm stat} := \left\{ \tau \;\middle|\; C_{\rm smooth}(\tau) < \frac{1}{\sqrt{N_\epsilon}}\ ,\ 10^3 \leqslant \tau \right\}. \tag{E6}$$

The upper cutoff $\tau_{\rm stat}^+$ for statistical significance is estimated as

$$\tau_{\rm stat}^+ := \begin{cases} \min_{A_{\rm stat}} \tau & \text{if } A_{\rm stat} \text{ is not empty,} \\ 10^4 & \text{if } A_{\rm stat} \text{ is empty.} \end{cases} \tag{E7}$$

Finally, the upper threshold for our final fitting $\tau_{\rm th}^+$ is given by

$$\tau_{\rm th}^+ := \min\{\tau_{\rm stat}^+, 10^2 \tau_{\rm th}^-\}. \tag{E8}$$

e. Step 5: The determination of $\gamma_{\rm NLLS}^{(a)}$

The power-law exponent $\gamma_{\rm NLLS}^{(a)}$ is finally estimated by the RLS fitting of the smoothed ACF $C_{\rm smooth}(\tau)$ for the range $[\tau_{\rm th}^-, \tau_{\rm th}^+]$ by the power-law fitting function

$$C(\tau) \propto \tau^{-\gamma_{\rm NLLS}^{(a)}} \tag{E9}$$

with fitting parameters $\gamma_{\rm NLLS}^{(a)}$.

f. Step 6: The determination of $c_{0,\rm NLLS}^{(a)}$

Finally, we determine $c_{0,\rm NLLS}^{(a)}$ by integration[5] of the smoothed ACF $C_{\rm smooth}(\tau)$ as

$$c_{0,\rm NLLS}^{(a)} := \frac{(1-\gamma_{\rm NLLS}^{(a)})}{(\tau_{\rm th}^+)^{1-\gamma_{\rm NLLS}^{(a)}} - (\tau_{\rm th}^-)^{1-\gamma_{\rm NLLS}^{(a)}}} \int_{\tau_{\rm th}^-}^{\tau_{\rm th}^+} C_{\rm smooth}(\tau')d\tau'. \tag{E10}$$

2. NLLS estimation based on the sample PSD

We next describe the measurements based on the sample PSD (see Fig. 16). According to [30], for $\gamma \in (0, 1)$ [or equivalently $\alpha \in (1, 2)$], the theoretical PSD of the LMF model is given by

$$S(\omega) \approx \int_0^\infty d\tau e^{2i\pi\omega\tau} \left(\frac{N_{\rm ST}^{\alpha-2}}{\alpha} |\tau|^{-\gamma} \right) = c^{(s)} \omega^{\gamma-1}, \tag{E11}$$

$$c^{(s)} = \frac{N_{\rm ST}^{\alpha-2}}{\alpha} 2^\gamma \pi^{\gamma-1} \Gamma(1-\gamma) \sin\left(\frac{\pi\gamma}{2}\right) \sim \omega^{-H}, \tag{E12}$$

$$H = 1 - \gamma = 2 - \alpha \quad \text{for small } \omega, \tag{E13}$$

where the Wiener-Khinchin theorem (17) is used. Similarly to the ACF method, we apply the power-law fitting $S(\omega) \propto \omega^{-H_{\rm NLLS}}$ to the sample PSD for the range $[\omega_{\rm th}^-, \omega_{\rm th}^+]$. The fitting range is automatically fixed as follows.

a. Step 1: The sample PSD

The sample PSD $S_{\rm sample}(\omega)$ was estimated by the periodogram method using *scipy* [35].

b. Step 2: Linear smoothing of the PSD

The sample PSD fluctuates due to the finite sample size. We apply normal smoothing of the empirical PSD:[6]

$$S_{\rm smooth}(\omega) := \sum_{\omega=-\infty}^{\infty} w_\delta(\omega) C_{\rm sample}(\omega),$$

$$w_\delta(\omega) := \begin{cases} \dfrac{1}{2\delta+1} & (\omega \in [\omega - \delta\Delta_\omega, \omega + \delta\Delta_\omega], \\ 0 & (\omega \notin [\omega - \delta\Delta_\omega, \omega + \delta\Delta_\omega], \end{cases}$$

$$\Delta_\omega = \frac{1}{N_\epsilon} \tag{E14a}$$

with the smoothing window size $\delta = 5$.

c. Step 3: The lower and upper thresholds

Let us determine the lower and upper thresholds $\omega_{\rm th}^-$ and $\omega_{\rm th}^+$. First, we describe the method to fix the lower threshold $\omega_{\rm th}^-$. The smoothed PSD $S_{\rm smooth}(\omega)$ fluctuates near the lowest frequency $\omega \approx \Delta_\omega$, and we discarded some of the low-frequency data points. We set $\omega_{\rm th}^- = 15\Delta_\omega$.

[5]The dimension of $c_{0,\rm NLLS}^{(a)}$ is given by [time$^{-\gamma}$], which is automatically consistent with the dimension analysis in this integration method. In addition, the integration of the ACF has a global smoothing effect, by which we expect that the estimation is more stable.

[6]We used normal smoothing instead of logarithmic smoothing because we are interested in the low-frequency regime of the PSD.

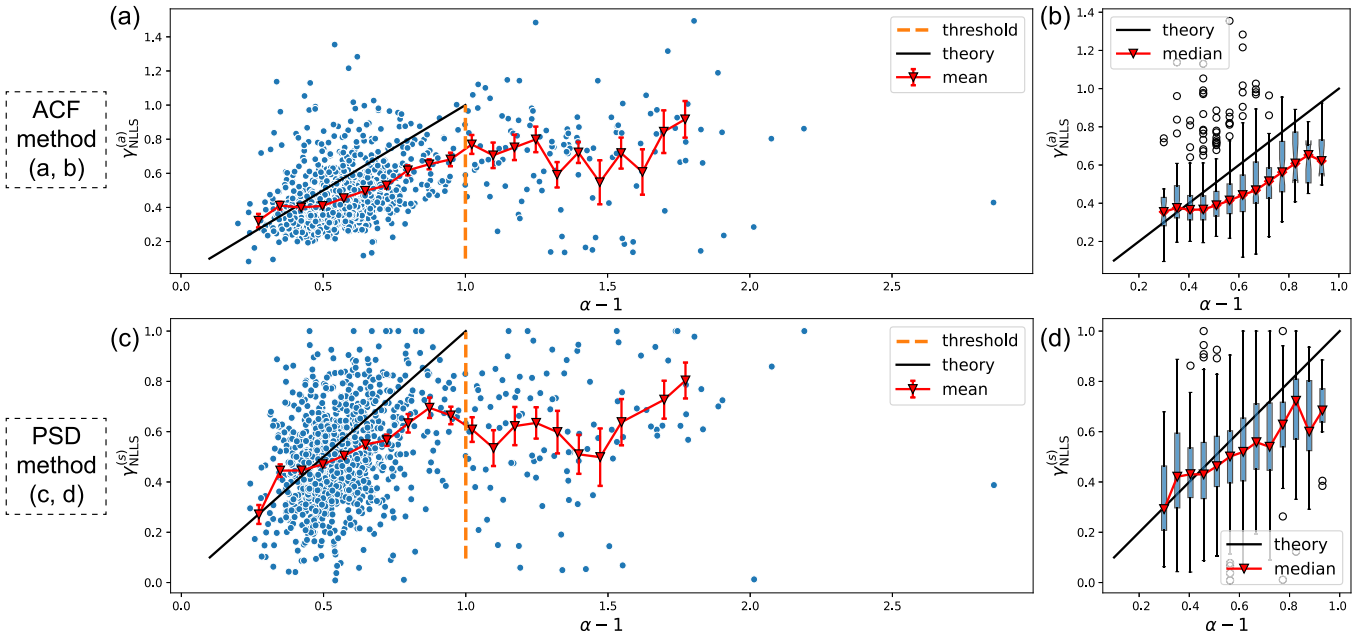

FIG. 17. Scatterplot between α and the NLLS estimator $\gamma_{\rm NLLS}$ for our data set based on the ACF method [upper panels, (a), (b)] and the PSD method [lower panels, (c), (d)]. (a), (c) The full scatterplot between α and $\gamma_{\rm NLLS}$. This figure illustrates that the NLLS estimator is actually biased, particularly for $\alpha > 2$. (b), (d) The boxplot between $\alpha - 1 \in (0, 1)$ and $\gamma_{\rm NLLS}$. This boxplot shows an approximate linear relation but shows the systematic deviation, as expected by the LMF simulations.

We next fix the upper threshold $\omega^+_{\rm th}$. According to the original LFM theory [10], the asymptotic relationship (E13) is valid up to $\tau \gg |\Omega_{\rm TR}|$ (or equivalently $2\pi\omega \ll |\Omega_{\rm TR}|^{-1}$). In our data set, the typical number of the splitting traders was 10^2 (see Sec. IV B 2). We therefore assume that $\omega^+_{\rm th}$ should be set smaller than 10^{-3}.

In addition, the PSD fluctuates for large frequency ω due to the finite sample size. Let us define the half-bandwidth $\omega_{\rm half}$ as the characteristic decay frequency of the PSD as

$$\omega_{\rm half} := \min_B \omega,$$
$$B := \{\omega \big| S_{\rm smooth}(\omega) < S_{\rm median}, \quad 10^2\Delta_\omega \leqslant \omega \leqslant 10^3\Delta_\omega\} \tag{E15}$$

if B is not empty, where $S_{\rm median}$ is the median of $\{S_{\rm smooth}(\omega)\}_{\omega\in[\omega^-_{\rm th}, 10^{-3}]}$. Considering the possibility that B might be empty in general, we set the upper threshold as

$$\omega^+_{\rm th} := \begin{cases} \omega_{\rm half} & \text{if } B \text{ is not empty,} \\ 10^3\Delta_\omega & \text{if } B \text{ is empty.} \end{cases} \tag{E16}$$

d. Step 4: The determination of $\gamma^{(s)}_{\rm NLLS}$

The Hurst exponent was estimated by the RLS fitting (see Appendix I) to the smoothed PSD $S_{\rm smooth}(\omega)$ for the range $[\omega^-_{\rm th}, \omega^+_{\rm th}]$, such that

$$S(\omega) \propto \omega^{-H_{\rm NLLS}} \tag{E17}$$

with fitting parameter $H_{\rm NLLS}$. The NLLS power-law exponent $\gamma^{(s)}_{\rm NLLS}$ is measured by the PSD method using the asymptotic relationship (E13):

$$\gamma^{(s)}_{\rm NLLS} := 1 - H_{\rm NLLS} \tag{E18}$$

if $H_{\rm NLLS} < 1$.

Due to methodological artifacts, we sometimes obtain $H_{\rm NLLS} > 1$, implying negative γ (i.e., a monotonically increasing ACF, which does not make sense). Since this is an obvious symptom of estimation failure, we excluded such data points as exceptional handling.

e. Step 5: The determination of $C^{(s)}_{\rm NLLS}$

Finally, the ACF prefactor $c^{(s)}_{0,\rm NLLS}$ is determined by the integration of the PSD, such that

$$c^{(s)}_{0,\rm NLLS} := \frac{2}{(2\pi)^{\gamma^{(s)}_{\rm NLLS}-1}\Gamma(1-\gamma^{(s)}_{\rm NLLS})\sin\left(\frac{\pi\gamma^{(s)}_{\rm NLLS}}{2}\right)} \times \frac{\gamma^{(s)}_{\rm NLLS}}{(\omega^+_{\rm th})^{\gamma^{(s)}_{\rm NLLS}} - (\omega^-_{\rm th})^{\gamma^{(s)}_{\rm NLLS}}} \int_{\omega^-_{\rm th}}^{\omega^+_{\rm th}} du S(u). \tag{E19}$$

APPENDIX F: THE SCATTERPLOT BASED ON THE NLLS ESTIMATOR

For our data analysis in the main text, we focused on the scatterplot between α and the naive estimator $\gamma_{\rm unbiased}$. This is because the NLLS estimator $\gamma_{\rm NLLS}$ has a statistical bias, and the unbiased estimator $\gamma_{\rm unbiased}$ is a better basis for our study. For reference, we show the scatterplot between α and $\gamma_{\rm NLLS}$ for our data set as Figs. 17(a) and 17(c) for the ACF and PSD methods, respectively. These figures illustrate that the naive estimator $\gamma_{\rm unbiased}$ is actually biased due to the finite sample size. For reference, we also show the boxplot in Figs. 17(b) and 17(d) for the ACF and PSD methods, respectively. As expected, the bias is much more serious for $\alpha > 2$.

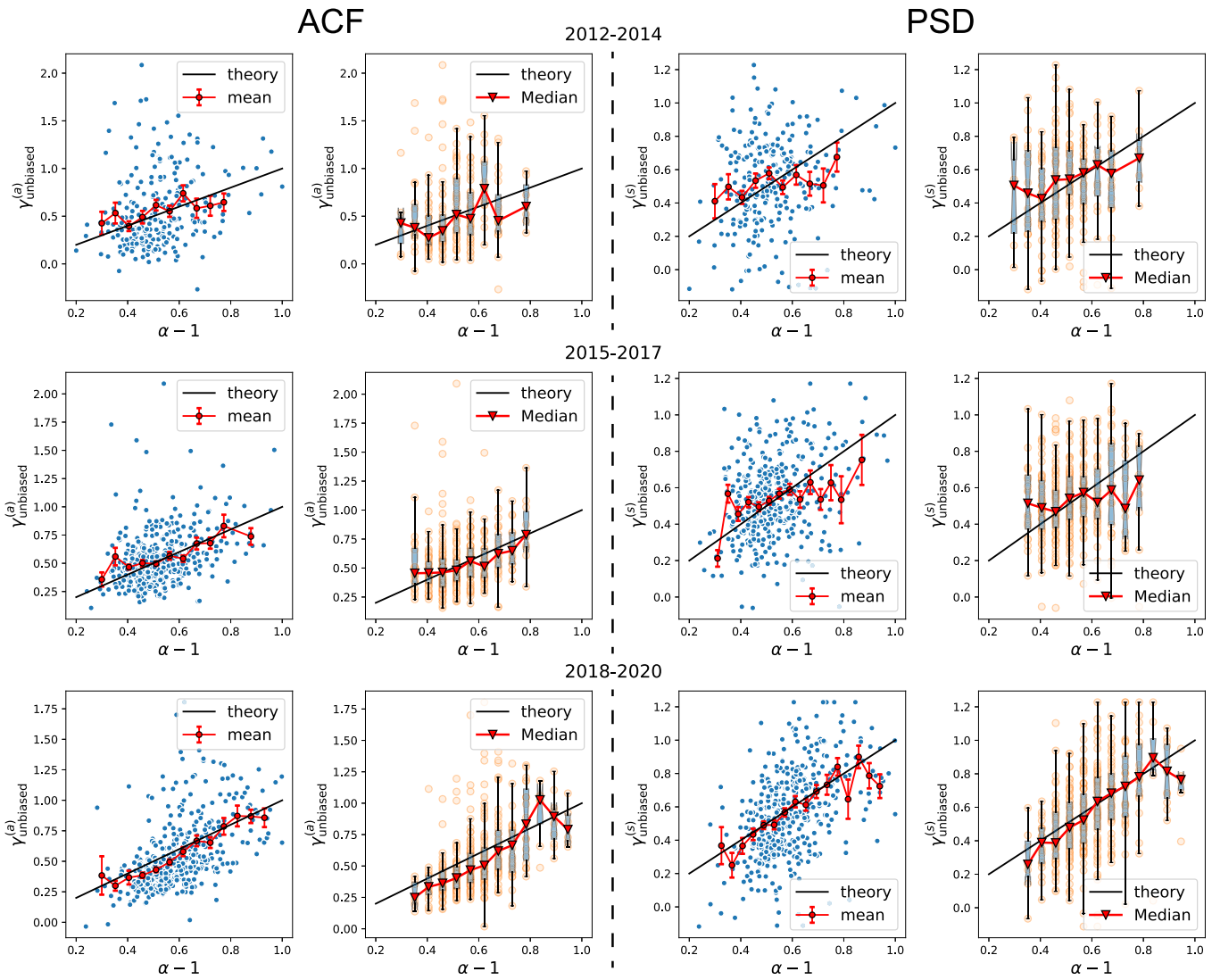

FIG. 18. Robustness check of our statistical analysis. Our nine-year data set was split into three data sets from 2012 to 2014, from 2015 to 2017, and from 2018 to 2020. The unbiased estimator $\gamma_{\rm unbiased}$ agrees with the theoretical line $\gamma = \alpha - 1$ for the three periods. The left (right) panels are based on the ACF (PSD) methods.

APPENDIX G: ROBUSTNESS CHECK OF OUR STATISTICAL ANALYSIS

In the main text, we tested the validity of the LMF prediction $\gamma = \alpha - 1$ by using the nine-year data. On the other hand, it would be more scientifically sound to check its statistical robustness. In this Appendix, we examined the temporal robustness of the LMF prediction.

For the robustness check, the nine-year data set was split into three data sets: (i) from 2012 to 2014, (ii) from 2015 to 2017, and (iii) from 2018 to 2020. We apply the same method as in Sec. V to these three independent data sets to test whether the LMF prediction holds for these three periods. The results are summarized in Fig. 18 [see the left (right) panels for the results based on the ACF (PSD) methods]. The LMF prediction $\gamma = \alpha - 1$ consistently holds for the three independent periods, suggesting the statistical robustness of our results. We have an impression that the goodness of fit improved for the most recent data set (2018–2020), which might be related to the increasing numbers of transactions, particularly by the STs.

APPENDIX H: THE LMF UNBIASED ESTIMATOR FOR THE TOTAL NUMBER OF SPLITTING TRADERS

This Appendix describes the detailed construction method of an LMF unbiased estimator for the total number of STs, $N_{\rm ST}$. Based on either the ACF or PSD method, let $\gamma_{\rm NLLS}$ and $c_{0,\rm NLLS}$ be the NLLS estimators for the ACF power-law exponent and the ACF prefactor as available quantities even from public data. The LMF theory predicts that the total number of traders $N_{\rm ST}$ is equal to the LMF estimator $N_{\rm ST}^{\rm LMF}(c_0, \gamma) := [(\gamma + 1)c_0]^{1/(\gamma-1)}$. We therefore study the relationship between the true value of $\log_{10}(N_{\rm ST})^{1-\gamma}$ and $\log_{10}(N_{\rm ST}^{\rm LMF})^{1-\gamma}$. The NLLS estimator is constructed as

$$\left[\log_{10}\left(N_{\rm ST}^{\rm LMF}\right)^{1-\gamma}\right]_{\rm NLLS} := \log_{10}\left[\frac{1}{(\gamma_{\rm NLLS}+1)c_{0,\rm NLLS}}\right]. \tag{H1}$$

1. Consistency for the infinite sample size

Let us check the consistency of the NLLS estimator by assuming the LMF model with realistic parameters regarding $(N_{\rm ST}, \alpha)$. We performed the numerical simulations of the LMF model for a sufficiently large sample size $N_\epsilon = 10^8$. We plotted the true values of $\log_{10}(N_{\rm ST})^{1-\gamma_{\rm NLLS}}$ for the vertical axis and those of $[\log_{10}(N_{\rm ST}^{\rm LMF})^{1-\gamma}]_{\rm NLLS}$ for the horizontal axis in Fig. 19(a) [Fig. 19(f)] for the ACF method (the PSD method). The figure shows the agreement between our numerical result and the theory, supporting consistency of the NLLS estimator $[\log_{10}(N_{\rm ST}^{\rm LMF})^{1-\gamma}]_{\rm NLLS}$.

2. Bias for the finite sample size

On the other hand, if we set realistic parameters regarding $(N_{\rm ST}, \alpha, N_\epsilon)$ with $N_\epsilon \lesssim 10^7$, the numerical result in Fig. 19(b) [Fig. 19(g)] shows systematic deviations from the theoretical line. Indeed, let us apply the regression

$$\log_{10}(N_{\rm ST})^{1-\gamma_{\rm NLLS}} = \beta_3\big[\log_{10}\left(N_{\rm ST}^{\rm LMF}\right)^{1-\gamma}\big]_{\rm NLLS} + \beta_4. \quad \text{(H2)}$$

We measured the values of (β_3, β_4) 100 times and took their ensemble average as $(\langle\beta_3\rangle, \langle\beta_4\rangle) = (0.847, 0.058)$ for the ACF method [$(\langle\beta_3\rangle, \langle\beta_4\rangle) = (1.117, -0.188)$ for the PSD method]. See also Figs. 19(d) and 19(e) [Figs. 19(i) and 19(j)] for the histogram of (β_3, β_4) regarding the ACF (PSD) method. This result implies that the NLLS estimator $[\log_{10}(N_{\rm ST}^{\rm LMF})^{1-\gamma}]_{\rm NLLS}$ is biased due to the finite sample size effect.

3. Construction of unbiased estimators

We next construct the unbiased estimators for $\log_{10}(N_{\rm ST})^{1-\gamma_{\rm NLLS}}$. The unbiased estimator $[\log_{10}(N_{\rm ST}^{\rm LMF})^{1-\gamma}]_{\rm unbiased}$ is constructed as

$$\Big[\log_{10}\left(N_{\rm ST}^{\rm LMF}\right)^{1-\gamma}\Big]_{\rm unbiased} := \beta_3\Big[\log_{10}\left(N_{\rm ST}^{\rm LMF}\right)^{1-\gamma}\Big]_{\rm NLLS} + \beta_4, \quad \text{(H3)}$$

which shows an approximate unbiasedness by definition,

$$\left\langle\Big[\log_{10}\left(N_{\rm ST}^{\rm LMF}\right)^{1-\gamma}\Big]_{\rm unbiased}\right\rangle \approx \log_{10}(N_{\rm ST})^{1-\gamma_{\rm NLLS}}. \quad \text{(H4)}$$

See Figs. 19(c) and 19(h) for the numerical check of the unbiasedness for the LMF simulations. Figure 12 is based on this approximate unbiased estimator $[\log_{10}(N_{\rm ST}^{\rm LMF})^{1-\gamma}]_{\rm unbiased}$.

APPENDIX I: NONLINEAR RELATIVE LEAST-SQUARES METHOD

In our fitting, we use the nonlinear relative least-squares (RLS) method, which is formulated as follows: let us consider the data points $\{(\tau_i, y_i)\}_{i=1,\ldots,N_{\rm dat}}$ and consider the fitting function $f(\tau|\boldsymbol{p})$ with the parameters $p := (p_1, \ldots, p_K)$. We fix the optimal parameter p^* as

$$\boldsymbol{p}^* = \arg\min_{\boldsymbol{p}} J_{\rm RLS}(\boldsymbol{p}), \quad J_{\rm RLS}(\boldsymbol{p}) := \sum_{i=1}^{N_{\rm dat}} \left(\frac{y_i - f(\tau_i|\boldsymbol{p})}{f(\tau_i|\boldsymbol{p})}\right)^2, \quad \text{(I1)}$$

where $J_{\rm RLS}(\boldsymbol{p})$ is the cost function for the RLS method. Note that the ordinary least-squares (OLS) method is formulated by

$$\boldsymbol{p}^* = \arg\min_{\boldsymbol{p}} J_{\rm OLS}(\boldsymbol{p}), \quad J_{\rm OLS}(\boldsymbol{p}) := \sum_{i=1}^{N_{\rm dat}} [y_i - f(\tau_i|\boldsymbol{p})]^2. \quad \text{(I2)}$$

Their difference comes from the cost functions between $J_{\rm RLS}(\boldsymbol{p})$ and $J_{\rm OLS}(\boldsymbol{p})$.

We did not employ the OLS method because the tail of the fitting function [i.e., $f(\tau|\boldsymbol{p})$ for large τ] contributes much less to the cost function than the head of the fitting function [i.e., $f(\tau|\boldsymbol{p})$ for small τ]. Since we are interested in the power-law exponent of the tail with $f(\tau|\boldsymbol{p}) \approx \tau^{-\gamma}$ for large τ, the contribution from the tail should not be underestimated. To clarify this point mathematically, let us rewrite the cost function of the OLS as

$$J_{\rm OLS}(\boldsymbol{p}) := \sum_{i=1}^{N_{\rm dat}} f^2(\tau_i|\boldsymbol{p})\left(\frac{y_i - f(\tau_i|\boldsymbol{p})}{f(\tau_i|\boldsymbol{p})}\right)^2 = J_{\rm weighted}(\boldsymbol{p}\,|\{f^2(\tau|\boldsymbol{p})\}_\tau), \quad \text{(I3)}$$

where we define the weighted cost function

$$J_{\rm weighted}(\boldsymbol{p}\,|\{w(\tau)\}_\tau) := \sum_{i=1}^{N_{\rm dat}} w(\tau_i)\left(\frac{y_i - f(\tau_i|\boldsymbol{p})}{f(\tau_i|\boldsymbol{p})}\right)^2. \quad \text{(I4)}$$

This representation highlights that the tail of the fitting function contributes much less to the cost function in the OLS method since $w(\tau) \ll w(0)$ for large τ. In contrast, the RLS has a better character than the OLS because the contributions to the cost function are theoretically expected to be the same between the head and the tail. We note that the cost function of the RLS method can be rewritten as

$$J_{\rm RLS}(\boldsymbol{p}) = J_{\rm weighted}(\boldsymbol{p}\,|\{1\}_\tau). \quad \text{(I5)}$$

APPENDIX J: SMOOTHING ON THE LOGARITHMIC TIME AXIS

Let us consider a smoothing method based on the logarithmic time. For simplicity, let us consider the AFC $C(\tau)$ for the continuous time $\tau \in (0, \infty)$. If the ACF asymptotically obeys the power-law decay $C(\tau) \approx C_0\tau^{-\gamma}$, its log-log plot should be linear as

$$\ln C(\tau) \approx \ln C_0 - \gamma \ln \tau. \quad \text{(J1)}$$

Therefore, it is customary to plot the log-log plot of the ACF to confirm the power-law decay. Based on this mathematical fact, we consider a smoothing of the ACF in the logarithmic time: by defining the logarithmic time $x := \ln\tau$, we introduce a smoothed ACF for a given τ as

$$C_{\rm smooth}(\tau) := \int_0^\infty \tilde{w}_\delta(x(\tau); x')C(\tau(x'))dx', \quad \tau(x) := e^x, \quad \tau(x') := e^{x'} \quad \text{(J2)}$$

with the weight function $\tilde{w}_\delta(x; x')$ and the smoothing window size $\delta > 0$. By assuming that $\tilde{w}_\delta(x; x')$ is uniform on the

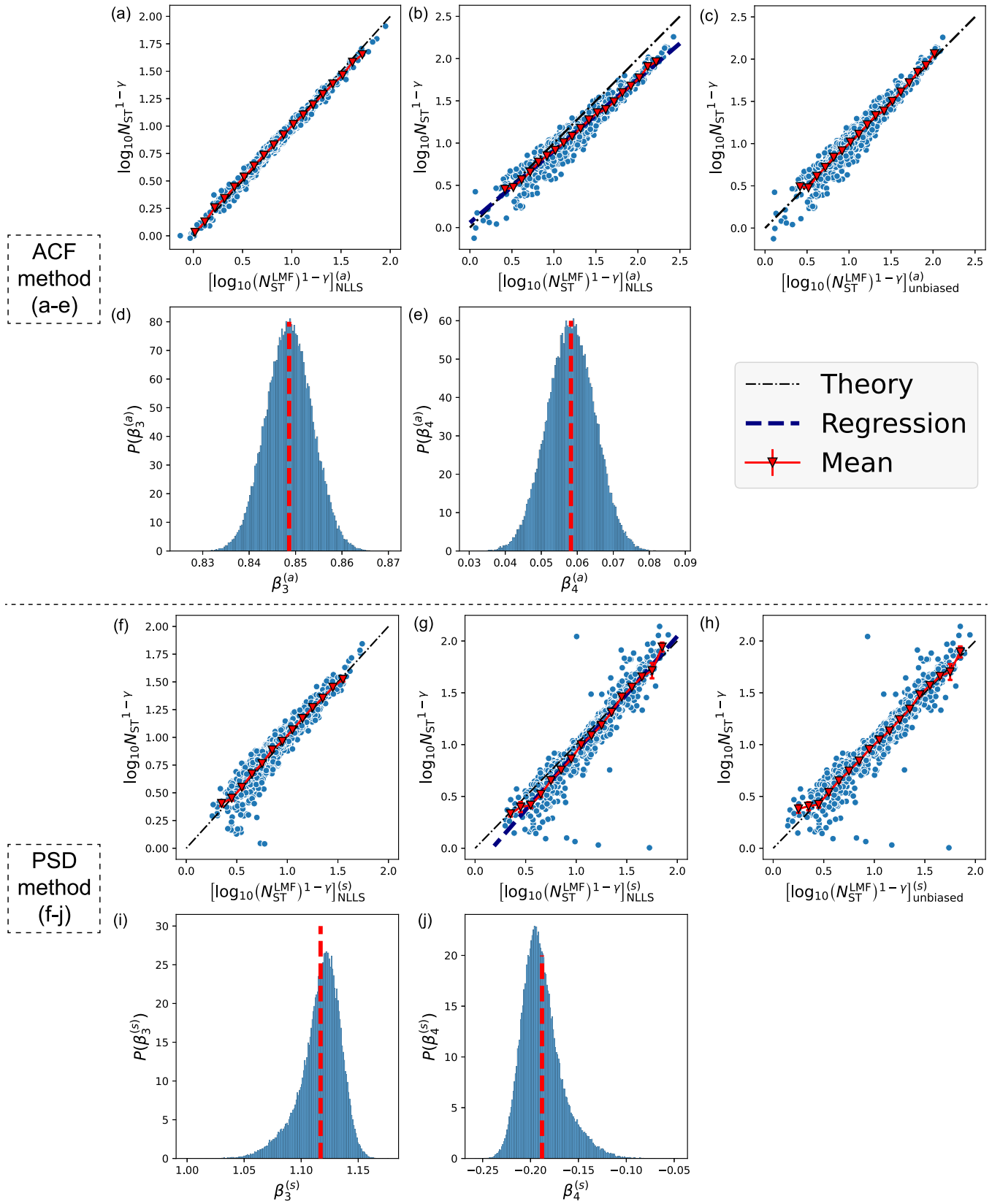

FIG. 19. Numerical study of the NLLS estimator $[\log_{10}(N_{\rm ST}^{\rm LMF})^{1-\gamma}]_{\rm NLLS}$ based on the ACF (a)–(e) and PSD (f)–(j) methods for the LMF simulations. The superscript $X^{(a)}$ ($X^{(s)}$) signifies that the estimator is based on the ACF (PSD) method. (a) Consistency check of the ACF-NLLS estimator for $N_\epsilon = 10^8$. The parameters $(N_{\rm ST}, \alpha)$ are the same as in our data set. (b) The ACF-NLLS estimator is biased for $N_\epsilon \lesssim 10^7$. The parameters $(N_{\rm ST}, \alpha, N_\epsilon)$ are the same as in our data set. (c) ACF-based unbiased estimator $[\log_{10}(N_{\rm ST}^{\rm LMF})^{1-\gamma}]_{\rm unbiased}$ approximately shows the unbiasedness as expected. (d), (e) Histogram of the regression coefficients (β_3, β_4) in Eq. (H2) for the ACF method. (f)–(j) Corresponding figures for the PSD method.

logarithmic time x as

$$\tilde{w}_\delta(x;x') = \begin{cases} \frac{1}{\delta} & (x' \in [x-\delta/2, x+\delta/2)), \\ 0 & (x' \notin [x-\delta/2, x+\delta/2)), \end{cases} \tag{J3}$$

we obtain the ACF formula of the logarithmic smoothing:

$$C_{\mathrm{smooth}}(\tau) = \int_{x(\tau)-\delta/2}^{x(\tau)+\delta/2} \frac{dx'}{\delta} C(\tau(x')) = \int_{\tau e^{-\delta/2}}^{\tau e^{+\delta/2}} \frac{d\tau'}{\delta\tau'} C(\tau') \tag{J4}$$

with the variable transformation $\tau' := e^{x'}$. This is equivalent to

$$C_{\mathrm{smooth}}(\tau) = \int_0^\infty w_\delta(\tau;\tau') C(\tau') d\tau',$$

$$w_\delta(\tau;\tau') := \begin{cases} \dfrac{1}{\delta\tau'} & (\tau' \in [\tau^-_{\mathrm{smooth}}(\tau), \tau^+_{\mathrm{smooth}}(\tau))), \\ 0 & (\tau' \notin [\tau^-_{\mathrm{smooth}}(\tau), \tau^+_{\mathrm{smooth}}(\tau))) \end{cases} \tag{J5a}$$

with

$$\tau^-_{\mathrm{smooth}}(\tau) := \tau e^{-\delta/2}, \quad \tau^+_{\mathrm{smooth}}(\tau) := \tau e^{+\delta/2}. \tag{J5b}$$

Thus, Eq. (E5) is the natural extension of the smoothed ACF formula for the discrete time τ.

[1] R. N. Mantegna and H. E. Stanley, *An Introduction to Econophysics* (Cambridge University Press, Cambridge, UK, 2000).

[2] F. Slanina, *Essentials of Econophysics Modelling* (Cambridge University Press, Cambridge, UK, 2014).

[3] J.-P. Bouchaud, J. Bonart, J. Donier, and M. Gould, *Trades, Quotes and Prices: Financial Markets Under the Microscope* (Cambridge University Press, Cambridge, UK, 2018).

[4] M. Jusup, P. Holme, K. Kanazawa, M. Takayasu, I. Romić, Z. Wang, S. Geĉek, T. Lipić, B. Podobnik, L. Wang, W. Luo, T. Klanjŝĉek, J. Fan, S. Boccaletti, and M. Perc, Social physics, Phys. Rep. **948**, 1 (2022).

[5] K. Kanazawa, T. Sueshige, H. Takayasu, and M. Takayasu, Derivation of the Boltzmann equation for financial Brownian motion: Direct observation of the collective motion of high-frequency traders, Phys. Rev. Lett. **120**, 138301 (2018).

[6] K. Kanazawa, T. Sueshige, H. Takayasu, and M. Takayasu, Kinetic theory for financial Brownian motion from microscopic dynamics, Phys. Rev. E **98**, 052317 (2018).

[7] T. Sueshige, K. Kanazawa, H. Takayasu, and M. Takayasu, Ecology of trading strategies in a forex market for limit and market orders, PLoS One **13**, e0208332 (2018).

[8] T. Sueshige, D. Sornette, H. Takayasu, and M. Takayasu, Classification of position management strategies at the order-book level and their influences on future market-price formation, PLoS One **14**, e0220645 (2019).

[9] J.-P. Bouchaud, Y. Gefen, M. Potters, and M. Wyart, Fluctuations and response in financial markets: the subtle nature of 'random' price changes, Quant. Financ. **4**, 176 (2003).

[10] F. Lillo and J. D. Farmer, The long memory of the efficient market, Stud. Nonlin. Dyn. Econom. **8**, 1 (2004).

[11] J.-P. Bouchaud, J. D. Farmer, and F. Lillo, How markets slowly digest changes in supply and demand, in *Handbook of Financial Markets: Dynamics and Evolution* (North-Holland, Amsterdam, 2009), pp. 57–160.

[12] J. D. Farmer and F. Lillo, On the origin of power-law tails in price fluctuations, Quant. Financ. **4**, 7 (2004).

[13] B. Biais, P. Hillion, and C. Spatt, An empirical analysis of the limit order book and the order flow in the Paris Bourse, J. Financ. **50**, 1655 (1995).

[14] J.-P. Bouchaud, J. Kockelkoren, and M. Potters, Random walks, liquidity molasses and critical response in financial markets, Quant. Financ. **6**, 115 (2006).

[15] Z. Eisler, J.-P. Bouchaud, and J. Kockelkoren, The price impact of order book events: market orders, limit orders and cancellations, Quant. Financ. **12**, 1395 (2012).

[16] M. D. Gould, M. A. Porter, and S. D. Howison, The long memory of order flow in the foreign exchange spot market, Mark. Microstructure Liq. **02**, 1650001 (2016).

[17] J. Donier and J. Bonart, A million metaorder analysis of market impact on the Bitcoin, Mark. Microstructure Liq. **01**, 1550008 (2015).

[18] B. Tóth, I. Palit, F. Lillo, and J. D. Farmer, Why is equity order flow so persistent? J. Econ. Dyn. Control **51**, 218 (2015).

[19] F. Lillo, S. Mike, and J. D. Farmer, Theory for long memory in supply and demand, Phys. Rev. E **71**, 066122 (2005).

[20] Y. Sato and K. Kanazawa, companion paper, Inferring microscopic financial information from the long memory in market-order flow: A quantitative test of the Lillo-Mike-Farmer model, Phys. Rev. Lett. **131**, 197401 (2023).

[21] K. Goshima, R. Tobe, and J. Uno, Trader Classification by Cluster Analysis: Interaction between HFTs and Other Traders, Waseda University Institute for Business and Finance Working Paper Series 19 (2019).

[22] M. Hirano, K. Izumi, H. Matsushima, and H. Sakaji, Comparing actual and simulated HFT traders' behavior for agent design, JASSS **23**, 6 (2020).

[23] A. Wald and J. Wolfowitz, On a test whether two samples are from the same population, Ann. Math. Stat. **11**, 147 (1940).

[24] S. M. Hussein, Event-based microscopic analysis of the FX market, Ph.D. thesis, University of Essex, 2013.

[25] G. Vaglica, F. Lillo, E. Moro, and R. N. Mantegna, Scaling laws of strategic behavior and size heterogeneity in agent dynamics, Phys. Rev. E **77**, 036110 (2008).

[26] N. Bershova and D. Rakhlin, The non-linear market impact of large trades: Evidence from buy-side order flow, Quant. Financ. **13**, 1759 (2013).

[27] Y. Sato and K. Kanazawa, Exact solution to a generalised Lillo-Mike-Farmer model with heterogeneous order-splitting strategies, arXiv:2306.13378 (2023).

[28] A. Clauset, C. R. Shalizi, and M. E. Newman, Power-law distributions in empirical data, SIAM Rev. **51**, 661 (2009).

[29] J. Alstott, E. Bullmore, and D. Plenz, powerlaw: A Python package for analysis of heavy-tailed distributions, PLoS One **9**, e85777 (2014).

[30] I. S. Gradshteyn and I. M. Ryzhik, *Table of Integrals, Series, and Products, Table of Integrals, Series, and Products* (Academic, 2014).

[31] C.-K. Peng, S. V. Buldyrev, S. Havlin, M. Simons, H. E. Stanley, and A. L. Goldberger, Mosaic organization of DNA nucleotides, Phys. Rev. E **49**, 1685 (1994).

[32] L. Rydin Gorjão, G. Hassan, J. Kurths, and D. Witthaut, MFDFA: Efficient multifractal detrended fluctuation analysis in python, Comput. Phys. Commun. **273**, 108254 (2022).

[33] O. Løvsletten, Consistency of detrended fluctuation analysis, Phys. Rev. E **96**, 012141 (2017).

[34] M. Tumminello, F. Lillo, J. Piilo, and R. N. Mantegna, Identification of clusters of investors from their real trading activity in a financial market, New J. Phys. **14**, 013041 (2012).

[35] P. Virtanen *et al.*, SciPy 1.0: fundamental algorithms for scientific computing in Python, Nat. Methods **17**, 261 (2020).