


Inferring Microscopic Financial Information from the Long Memory in Market-Order Flow: A Quantitative Test of the Lillo-Mike-Farmer Model

Yuki Sato¹ and Kiyoshi Kanazawa^{1*}

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

 (Received 5 January 2023; revised 2 August 2023; accepted 7 September 2023; published 8 November 2023)

In financial markets, the market-order sign exhibits strong persistence, widely known as the long-range correlation (LRC) of order flow; specifically, the sign autocorrelation function (ACF) displays long memory with power-law exponent γ , such that $C(\tau) \propto \tau^{-\gamma}$ for large time-lag τ . One of the most promising microscopic hypotheses is the order-splitting behavior at the level of individual traders. Indeed, Lillo, Mike, and Farmer (LMF) introduced in 2005 a simple microscopic model of order-splitting behavior, which predicts that the macroscopic sign correlation is quantitatively associated with the microscopic distribution of metaorders. While this hypothesis has been a central issue of debate in econophysics, its direct quantitative validation has been missing because it requires large microscopic datasets with high resolution to observe the order-splitting behavior of all individual traders. Here we present the first quantitative validation of this LMF prediction by analyzing a large microscopic dataset in the Tokyo Stock Exchange market for more than nine years. On classifying all traders as either order-splitting traders or random traders as a statistical clustering, we directly measured the metaorder-length distributions $P(L) \propto L^{-\alpha-1}$ as the microscopic parameter of the LMF model and examined the theoretical prediction on the macroscopic order correlation $\gamma \approx \alpha - 1$. We discover that the LMF prediction agrees with the actual data even at the quantitative level. We also discuss the estimation of the total number of the order-splitting traders from the ACF prefactor, showing that microscopic financial information can be inferred from the LRC in the ACF. Our Letter provides the first solid support of the microscopic model and solves directly a long-standing problem in the field of econophysics and market microstructure.

DOI: [10.1103/PhysRevLett.131.197401](https://doi.org/10.1103/PhysRevLett.131.197401)

Introduction.—Can a statistical-physics approach help in understanding macroscopic phenomena in financial markets from their microscopic dynamics [1,2]? In posing this challenging thought, physicists have greatly benefitted from recent high-frequency data for econophysics modeling of market microstructure [3,4], even at the level of individual traders [5,6]. In this Letter, we provide the first quantitative evidence of a historic econophysics theory regarding the long-range correlation (LRC) in the market-order flow [7–9].

Let us briefly review the trading rules in recent financial markets, where traders have two options. The first option is the limit order, by which traders provide the market liquidity and show the potential prices at which they are willing to transact. The second option is the market order, by which traders immediately consume the liquidity and transact at the best prices (i.e., the highest bid or the lowest ask prices). This Letter tests an econophysics microscopic

model for the market-order flow, particularly on their statistical persistence.

The strong persistence of the market-order flow underscores an established empirical law in financial markets [3,8,9]: i.e., once a buy (sell) market order is observed, a buy (sell) market order is likely to be observed (see Fig. 1). This predictability regarding market orders is mathematically characterized by a power-law decay for the order-sign autocorrelation function (ACF):

$$C(\tau) := \langle \epsilon(t)\epsilon(t+\tau) \rangle \approx c_0\tau^{-\gamma}, \quad 0 < \gamma < 1, \quad (1)$$

for large time-lag $\tau \gg 1$. Here $\epsilon(t)$ is the market-order sign at time t defined by $\epsilon(t) = +1$ [$\epsilon(t) = -1$] for the buy (sell) market order, $\langle \dots \rangle$ represents the ensemble average, c_0 is the prefactor, and γ is the power-law exponent for the LRC. Because it is ubiquitously observed across broad markets, the LRC is believed essential to a market microstructure.

Then, what is the microscopic origin of the LRC as a macroscopic phenomenon? One promising response is the order-splitting hypothesis for individual traders' behaviors [7] [Fig. 1(a)]. This hypothesis claims the LRC appears because some traders split large metaorders into a long

Published by the American Physical Society under the terms of the Creative Commons Attribution 4.0 International license. Further distribution of this work must maintain attribution to the author(s) and the published article's title, journal citation, and DOI.

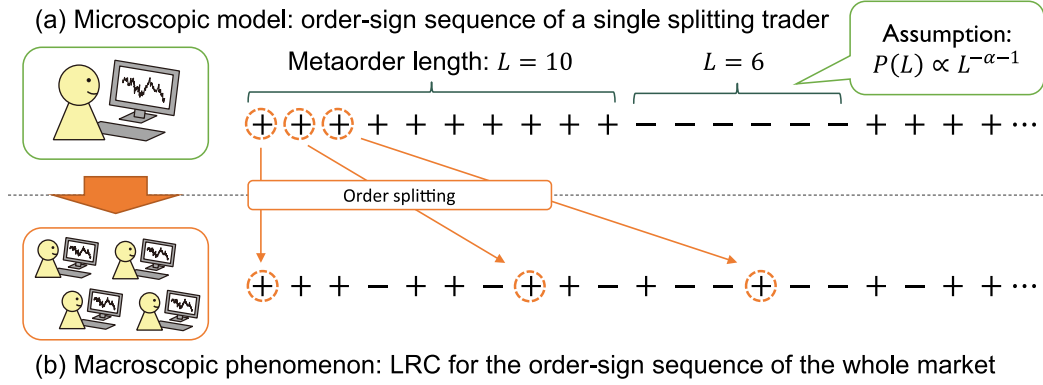


FIG. 1. Schematic of the LRC of the market-order flow and the order-splitting hypothesis (particularly, the LMF model); as a shorthand notation for +1 (−1), “+” (“−”) signifies a buy (sell). (a) As a microscopic model, we assume the presence of STs. Also, STs successively submit the child orders with the same sign for L times, where L is called the metaorder length and obeys power law statistics, $P(L) \propto L^{-\alpha-1}$. (b) Consequently, the LRC appears as a macroscopic phenomenon. The LMF theory predicts a quantitative relation $\gamma = \alpha - 1$, which we empirically establish in this Letter through data analysis.

series of small child orders. Because all the child orders share for a while the same sign, there is weak predictability of the future order sign, which is ultimately reflected in the power-law decay of the ACF as summarized in Eq. (1) [Fig. 1(b)]. Furthermore, Lillo, Mike, and Farmer (LMF) proposed a simple microscopic theory based on the order-splitting hypothesis. They assumed (i) the presence of splitting traders (STs), and (ii) the power-law probability density function (PDF) for the metaorder length L such that $P(L) \propto L^{-\alpha-1}$ with microscopic exponent $\alpha > 1$. By assuming random order submissions, the ACF macroscopically exhibits a power-law decay (1). Specifically, they showed

$$\gamma = \alpha - 1, \quad (2)$$

which in this Letter we refer to as the *quantitative LMF prediction*. The prediction (2) is beautiful and quantitatively powerful because it connects the macroscopic and microscopic parameters in alignment with the central spirit of statistical physics.

While the plausibility of this scenario was confirmed qualitatively in [10] (i.e., a decomposition of the ACF into an order-splitting component and the remainder), the detailed verification of the quantitative prediction (2) has been missing for 18 years. The original LMF paper [7] reported an initial attempt to test their prediction. However, they only confirmed a minimum consistency of their theory (i.e., the theoretical line passes through the center of the mass in the scatterplot; see Fig. 5 and Sec. III B in [11] for a brief review) when lacking suitably large datasets.

In this Letter, together with companion paper [11], we solve this long-standing econophysics problem precisely by analyzing a large comprehensive microscopic dataset of the Tokyo Stock Exchange (TSE). We accessed a special microscopic dataset, including trading-account identifiers

(IDs) on the TSE, enabling us to track effectively the behavior of trading accounts. Using our microscopic dataset, we first applied a strategy clustering of individual traders to test assumption (i). In regard to market orders, and after classifying all traders as STs or random traders (RTs), we confirmed the presence of STs in most of the TSE markets. We next studied the empirical metaorder-length PDF $P(L)$ to test assumption (ii), which we validated from our dataset. With the measured microscopic parameter α , we generated a scatterplot between α and γ to test the quantitative LMF prediction (2). Finally, we found the prediction (2) agreed with our dataset, providing quantitatively the first solid support for the LMF model as the minimal microscopic description of the order-splitting behavior. As the last discussion, we estimate the total number of the STs from the observed prefactor c_0 . Our findings imply that the long memory in the market-order ACF is useful in inferring microscopic financial information.

Data description.—Let us briefly describe our dataset provided by the Japan Exchange (JPX) Group, the platform manager of the TSE. The TSE being the biggest stock market in Japan, our dataset covers all the order flows in the TSE (market orders, limit orders, and cancellations), enabling us to track their complete life cycle for all the stocks for nine years (from the 4th January 2012 to the 30th December 2020). Furthermore, this dataset includes virtual server IDs (VSIDs), a unit of trading accounts on the TSE. The VSID is not technically equivalent to the membership ID, because any trader may have several VSIDs. However, we can effectively define trader IDs to track individual trader behavior with high resolution by appropriately aggregating VSIDs [12,13] (e.g., if a limit order is submitted by VSID 1 and is cancelled from VSID 2, both VSIDs are associated with the same trader); see also [11] for more technical details.

Our study focused on the sign sequences of market orders during double auctions from 09:00–11:30 and 12:30–15:00 Japan Standard Time. A yearly segmented

order-sign sequence was extracted for each stock to obtain one market data point. We only used data points with more than 0.5×10^6 transactions and removed transaction data from the opening and closing 10 min of auctions to suppress the intraday-seasonality effect.

Assumptions of the LMF model.—As summarized in Fig. 1, there are two key assumptions in the LMF model: (i) the presence of STs who have large latent demand (*metaorders*) and split them into small *child orders*, which are assumed to share the same sign for L successive times, and (ii) the *metaorder length* L obeys a power law $P(L) \propto L^{-\alpha-1}$ with $\alpha > 1$.

In previous literature, there was no solid direct evidence of assumption (i), although [10] shows indirect but promising evidence based on the ACF decomposition. Also, the plausibility of assumption (ii) was studied in [7] by analyzing the off-book data for the London stock exchange market as an “imperfect proxy.” However, with the absence of appropriate datasets at that time, the precise estimation of α became a technical problem for LMF verification. To verify assumptions (i) and (ii) directly, it is necessary to identify STs by strategy clustering at the level of individual traders and then study their metaorder-length PDF to measure α precisely.

Presence of STs.—We proceeded with strategy clustering to identify STs. We studied the order-sign sequence for each ST [Fig. 1(a)] to construct the metaorder length L by defining L as a length of successively equal signs. Concerning exceptional handling, if there was more than one business day between two successive orders, we assume they belong to different metaorders [14] to avoid overestimating metaorder length.

For a given metaorder-length sequence, we apply the binomial test for strategy clustering; the null hypothesis is that the order-sign sequence is purely random (obeying a symmetric Bernoulli process) and, thus, the trader belongs to the RT set. The trader is regarded as an ST if the null hypothesis is rejected with a significance level $\theta := 0.01$.

On the basis of this clustering scheme, we identified the ST set for each market data point. With summary statistics across all the markets during nine years, we evaluated the empirical PDF for the ST percentage in each market [Fig. 2(a)], and the contribution to market orders from the ST set [Fig. 2(b)]. We concluded that typically a quarter of all traders are STs, but they dominate the total market orders. Via this strategy clustering, we thus validated assumption (i) directly.

Metaorder-length PDF.—Having identified the set of STs, we measured the aggregated empirical PDF for the metaorder length of all STs. Most of the aggregated complementary-cumulative distribution functions (CCDFs) for the metaorder length of STs obey a power law $P_>(L) \propto L^{-\alpha}$ with the CCDF defined by $P_>(L) := \int_L^\infty P(L')dL'$. As a typical example, we plotted the metaorder-length CCDF for the Toyota Motor Corporation (with ticker number

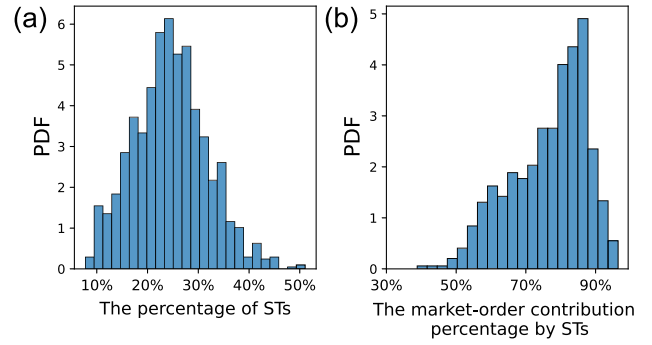


FIG. 2. Presence of the STs by our strategy clustering. (a) Empirical PDF for the percentage of STs in each market, showing direct evidence of the presence of STs. Typically, 25% of all traders were STs. (b) Empirical PDF for STs contribution to market orders in each market. Typically, 80% of all the market orders were issued by the STs, implying their overwhelming contribution to market orders.

7203) in 2020 [Fig. 3(a)]; it features a power-law asymptotic tail for large L . We then evaluated α using Clauset’s algorithm [15,16] to plot the empirical PDF of α [Fig. 3(b)] across all the stocks. Typically, the exponent α is distributed over values $1 < \alpha < 2$, in agreement with the standard assumption for the LMF model. We thus validated assumption (ii) for our dataset.

Power-law exponent in the ACF.—Having measured the microscopic power-law exponent α , we next measured the macroscopic power-law exponent γ of the ACF, which we did by fitting directly the sample order-sign ACF as follows (see Ref. [11] for details of the method): We first calculated the sample ACF from $C_{\text{sample}}(\tau) := \sum_{t=1}^{N_e-\tau} \epsilon(t)\epsilon(t+\tau) / (N_e - \tau)$ with time-lag τ and total number of market orders

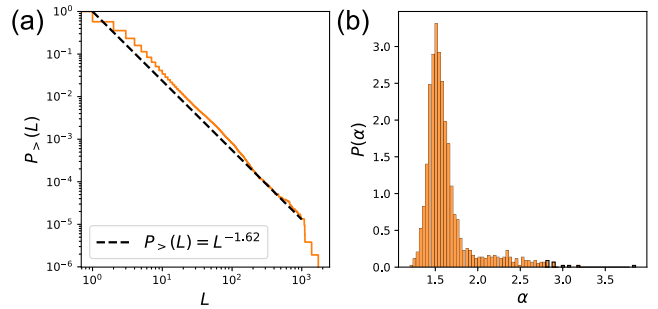


FIG. 3. The aggregated CCDFs for STs with our strategy clustering. (a) Empirical CCDF aggregated regarding metaorder length L among STs using data for the Toyota Motor Corporation in 2020 as a typical example. The CCDF obeys the power law $P_>(L) \propto L^{-\alpha}$ with $\alpha \approx 1.62$. Likewise, most empirical aggregated CCDF for STs obey similar power laws. (b) Empirical PDF of the power-law exponent α for all the markets. The power-law exponents were evaluated systematically using Clauset’s algorithm [15,16] across all the markets. The exponent α typically satisfies $1 < \alpha < 2$, consistent with the standard assumption for the LMF model.

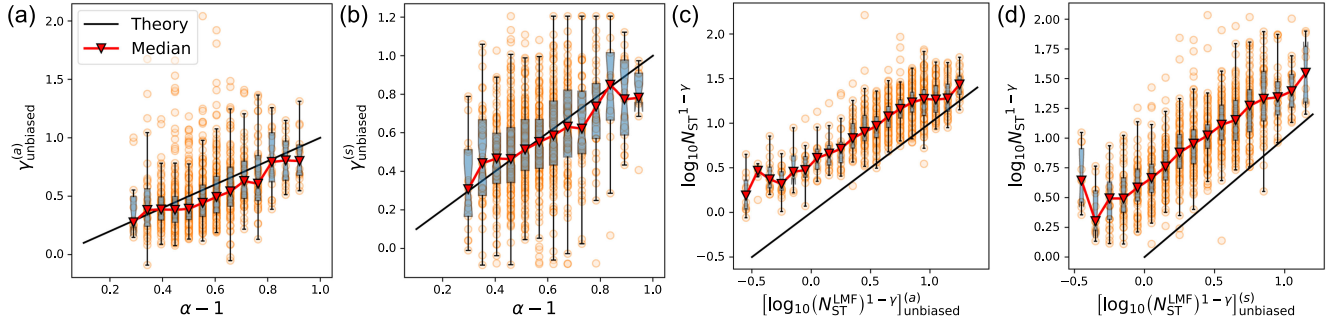


FIG. 4. (a),(b) Scattered box plots between α and γ with the median, the first and third quartiles for (a) the ACF and (b) the PSD methods, exhibiting excellent agreement with the LMF prediction (2) (black line). γ was evaluated using the approximate unbiased estimator γ_{unbiased} , based on the NLLS estimator and the LMF model. (c),(d) Scattered box plots between the LMF estimator $N_{\text{ST}}^{\text{LMF}}$ and the actual total number of the STs N_{ST} for (c) the ACF and (d) the PSD methods for the datapoints with $\alpha < 2$. The LMF estimator is highly correlated with the true value of N_{ST} as the classical theory predicts, but systematically underestimates N_{ST} , such that $N_{\text{ST}}^{\text{LMF}} \lesssim N_{\text{ST}}$. This observation is consistent with a generalized LMF theory [18] with heterogeneous intensities $\{\lambda^{(i)}\}_i$.

N_e . We fixed the fitting range $[\tau_{\text{th}}^-, \tau_{\text{th}}^+]$ automatically such that only the power-law decay is observed in the ACF for $[\tau_{\text{th}}^-, \tau_{\text{th}}^+]$. We applied logarithmic smoothing and a final fitting $C(\tau) = C_0 \tau^{-\gamma_{\text{NLLS}}}$ for the range $[\tau_{\text{th}}^-, \tau_{\text{th}}^+]$ employing the relative nonlinear least square (NLLS) estimation.

Although the NLLS estimator γ_{NLLS} gives numerical consistency for the LMF model, we noticed that the NLLS estimator γ_{NLLS} has a finite-sample-size bias. To remove this bias, we constructed heuristically an approximate unbiased estimator γ_{unbiased} based on the LMF model (see companion paper [11] for details). For this Letter, we used this unbiased estimator γ_{unbiased} for the final scatter plot.

As a robustness check, we also measured the power-law exponent γ via the power-spectral density (PSD) method (see Ref. [11]). The exponent measured by the ACF and PSD fittings are, respectively, denoted by $\gamma_{\text{unbiased}}^{(a)}$ and $\gamma_{\text{unbiased}}^{(s)}$. Both methods exhibit reasonable and consistent results, implying the statistical robustness of our results.

Scatter plot.—Having evaluated the microscopic and macroscopic power-law exponents α and γ using our huge TSE dataset, we are ready to draw the scatter plot between α and γ and test the LMF prediction (2). As the main result, we provide the scattered box plots [Figs. 4(a) and 4(b) for the ACF and PSD methods, respectively] between α and γ_{unbiased} with focus on the range $1 < \alpha < 2$ in accordance with the standard LMF assumption [17]. These figures exhibit excellent agreement with the theoretical line (2). From these figures, we conclude that with our microscopic dataset the LMF prediction (2) has quantitative validity.

Discussion on the prefactor.—While we extracted the microscopic information α from the ACF power-law exponent γ via Eq. (2), is it possible to extract other microscopic information from the prefactor c_0 ? The LMF theory predicts $c_0 \approx N_{\text{ST}}^{\alpha-2}/\alpha$ with the total number of the STs N_{ST} , implying that N_{ST} can be estimated by the LMF estimator

$$N_{\text{ST}}^{\text{LMF}}(c_0, \gamma) := \frac{1}{[(\gamma + 1)c_0]^{\frac{1}{1-\gamma}}}, \quad (3a)$$

where γ and c_0 can be observed from publicly available data.

Note that original LMF work made an assumption of the homogeneity of order-splitting intensities $\{\lambda^{(i)}\}_i$ among traders in [7], such that $\lambda^{(i)} = 1/N_{\text{ST}}$ for all i . While we noticed that this homogeneity assumption is unrealistic, we tested this prediction in our dataset by drawing the scattered box plots [Figs. 4(c) and 4(d) based on the ACF and PSD methods, respectively] between $\log_{10} N_{\text{ST}}^{1-\gamma}$ and the LMF estimator $\log_{10} (N_{\text{ST}}^{\text{LMF}})^{1-\gamma}$ with the finite-sample size bias removed (see Ref. [11]). We find that the LMF estimator $N_{\text{ST}}^{\text{LMF}}$ is highly correlated with the true value N_{ST} , implying that the ACF prefactor is a useful resource to infer N_{ST} . At the same time, the LMF estimator $N_{\text{ST}}^{\text{LMF}}$ systematically underestimates the true value N_{ST} , such that $N_{\text{ST}}^{\text{LMF}} \lesssim N_{\text{ST}}$.

Interestingly, our finding is consistent with a generalized LMF model with the heterogeneous intensity distribution $\{\lambda^{(i)}\}_i$. Indeed, Ref. [18] shows the ACF formula (3a) is nonrobust but sensitive to the heterogeneous intensity distribution, while the power-law-exponent formula (2) is robust. Furthermore, the LMF estimator $N_{\text{ST}}^{\text{LMF}}$ is shown to provide the lower bound of the true value of N_{ST} , such that

$$N_{\text{ST}}^{\text{LMF}} \lesssim N_{\text{ST}}, \quad (3b)$$

showing the consistency with Figs. 4(c) and 4(d). Thus, we have successfully confirmed the qualitative validity of the LMF picture even for the estimation of N_{ST} , while for better quantitative understanding it might require theoretical updates regarding the heterogeneity of trading strategies.

Conclusion.—Although the power-law memory character in the order-sign ACF has been a central issue in

econophysics, and with the absence of an appropriate huge microscopic dataset, no quantitative evidence had been provided for the corresponding microscopic model (the LMF model). In this Letter, we have provided the first solid evidence for the LMF model at the quantitative level (2) at least for the TSE market and, thus, solved this long-lasting problem.

Let us briefly discuss the implication of our findings. Our Letter shows that the microscopic parameters α and N_{ST} (usually unobservable because its direct estimation requires special microscopic datasets like ours) can be inferred via the LMF predictions (2) and (3b), where γ and c_0 are observable even for public data. This is reminiscent of Einstein's theory for physical Brownian motions: Avogadro's number N_A (unobservable) was indirectly estimated from the thermal fluctuations via the Einstein relation for the diffusion constant. The LMF theory can play a similar role in inferring microscopic financial parameters from financial fluctuations.

The microscopic parameter set (α, N_{ST}) quantifies how the latent demand is hidden in the long term. For markets with small α , the revealed liquidity on the limit-order book is insufficient for liquidity takers, and takers have no choice but to split their large metaorders into a longer series of child orders (see also [3] for a standard interpretation of the order-splitting behavior from the viewpoint of practitioners). In this sense, markets with smaller α and large N_{ST} might not be liquid enough because many large institutional investors are waiting for the liquidity to replenish during their order splitting. This characteristic of liquidity has not been captured in practice through conventional metrics such as market spread (the difference between the best bid and ask prices), market depth (the typical volume size at the best prices), and market impact (the average price movement after a market order). Thus, the parameter set (α, N_{ST}) is a new measure quantifying how the market is potentially illiquid due to the hidden demand by large institutional investors.

Remarkably, successful strategy clustering was the key to our data analysis at the individual trader level in revealing the market ecology from a microscopic viewpoint. This research direction aligns with the previous literature [19–21] proposing the need of market-ecology analyses. We believe that this direction of research holds promise, particularly for econophysics and sociophysics modeling [4] as it benefits from recent microscopic financial datasets.

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 Letter by the JPX Group, Inc. We declare no

financial conflict of interest. The JPX Group, Inc. provided the original data for this study without any financial support. We thank Richard Haase, Ph.D., from Edanz for editing a draft of this manuscript.

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

- [1] R. N. Mantegna and H. E. Stanley, *An Introduction to Econophysics* (Cambridge University Press, Cambridge, England, 2000).
- [2] F. Slanina, *Essentials of Econophysics Modelling* (Cambridge University Press, Cambridge, England, 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, England, 2018).
- [4] M. Jusup *et al.*, 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] F. Lillo, S. Mike, and J. D. Farmer, Theory for long memory in supply and demand, *Phys. Rev. E* **71**, 066122 (2005).
- [8] 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).
- [9] F. Lillo and J. D. Farmer, The long memory of the efficient market, *Stud. Nonlinear Dyn. Econom.* **8**, 1 (2004).
- [10] B. Tóth, I. Palit, F. Lillo, and J. D. Farmer, Why is equity order flow so persistent?, *J. Econ. Dyn. Control* **51**, 218 (2015).
- [11] Y. Sato and K. Kanazawa, companion paper, Quantitative statistical analysis of order-splitting behavior of individual trading accounts in the Japanese stock market over nine years, *Phys. Rev. Res.* **5**, 043131 (2023).
- [12] 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), https://www.waseda.jp/fcom/wbf/assets/uploads/2019/06/trading_cluster_June2019.pdf.
- [13] M. Hirano, K. Izumi, H. Matsushima, and H. Sakaji, Comparing actual and simulated HFT traders' behavior for agent design, *J. Artif. Soc. Soc. Simulat.* **23**, 6 (2020).
- [14] J. Donier and J. Bonart, A million metaorder analysis of market impact on the Bitcoin, *Mark. Microstruct. Liq.* **1**, 1550008 (2015).
- [15] A. Clauset, C. R. Shalizi, and M. E. Newman, Power-law distributions in empirical data, *SIAM Rev.* **54**, 661 (2009).
- [16] J. Alstott, E. Bullmore, and D. Plenz, powerlaw: A PYTHON package for analysis of heavy-tailed distributions, *PLoS One* **9**, e85777 (2014).

- [17] The scattered box plot is composed of both box plot and scatter plot. The data points are drawn on the center of the bins along the α axis.
- [18] Y. Sato and K. Kanazawa, Exact solution to a generalized Lillo-Mike-Farmer model with heterogeneous order-splitting strategies, [arXiv:2306.13378](https://arxiv.org/abs/2306.13378).
- [19] 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).
- [20] 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).
- [21] 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).