Tests of scaling and universality of the distributions of trade size and share volume: Evidence from three distinct markets

Vasiliki Plerou and H. Eugene Stanley
Center for Polymer Studies and Department of Physics, Boston University, Boston, Massachusetts 02215, USA

(Received 23 April 2007; published 12 October 2007)

Empirical evidence for scale-invariant distributions in financial data has attracted the research interest of physicists. While the power-law tails of the distribution of stock returns $P\{R>x\}\sim x^{-\zeta_R}$ are becoming increasingly well documented, less understood are the statistics of other closely related microstructural variables such as q_i, the number of shares exchanged in trade i (termed the *trade size*) and $Q_{\Delta t}(t)=\Sigma_{i=1}^{N}q_i$, the total number of shares exchanged as a result of the $N=N_{\Delta t}$ trades occurring in a time interval Δt (termed *share volume*). We analyze the statistical properties of trade size $q\equiv q_i$ and share volume $Q\equiv Q_{\Delta t}(t)$ by analyzing trade-by-trade data from three large databases representing three distinct markets: (i) 1000 major U.S. stocks for the 2-y period 1994–1995, (ii) 85 major U.K. stocks for the 2-y period 2001–2002, and (iii) 13 major Paris Bourse stocks for the 4.5-y period 1994–1999. We find that, for all three markets analyzed, the cumulative distribution of trade size displays a power-law tail $P(q>x)\sim x^{-\zeta_q}$ with exponent $\zeta_q<2$ within the Lévy stable domain. Our analysis of the exponent estimates of ζ_q suggests that the exponent value is universal in the following respects: (a) ζ_q is consistent across stocks within each of the three markets analyzed, and also across different markets, and (b) ζ_q does not display any systematic dependence on market capitalization or industry sector. We next analyze the distributions of share volume $Q_{\Delta t}$ over fixed time intervals and find that for all three markets $P\{Q>x\}\sim x^{-\zeta_Q}$ with exponent $\zeta_Q<2$ within the Lévy stable domain. To test the validity for $\Delta t=1$ day of the power-law distributions found from tick-by-tick data, we analyze a fourth large database containing daily U.S. data, and confirm a value for the exponent ζ_Q within the Lévy stable domain.

DOI: 10.1103/PhysRevE.76.046109 PACS number(s): 89.75.Da, 89.65.−s, 05.45.Tp

I. INTRODUCTION

Understanding economic phenomena using concepts and methods of physics has attracted the research interest of physicists and practitioners alike [1–3]. In particular, empirical evidence of scaling and long-range correlations in financial data is interesting because of the analogies it suggests with collective phenomena in complex physical systems. Moreover, it is possible that the statistical physics methods used to understand collective phenomena may provide new insights into understanding large economic fluctuations [4–6].

The statistical properties of the time series of financial variables are quite unique. Stock returns, for example, display tails that are much more pronounced than a simple Gaussian. Events such as the 1987 stock market crash—where the leading U.S. index, the Standard & Poors (S&P) 500 index, dropped by a magnitude of over 20 standard deviations—signify the non-Gaussian nature of these fluctuations. This market crash was accompanied by $\approx 6\times 10^8$ shares that changed hands on the New York Stock Exchange alone. Indeed, it is a common practitioner saying that it takes volume to move stock prices.

Prior analysis of intraday stock returns data for both indices and single stocks shows that the tails of the return distribution decay as power laws with exponents outside the Lévy stable domain [7–11]. Although the precise nature of the relationship between volume and price fluctuations is not known, the presence of fat tails in the distribution of returns suggests that the distribution of volume is also fat tailed. Understanding the empirically observed positive equal-time correlation between volume and volatility has been a subject of active research in the literature [12–21]. Here we focus on quantifying the statistical properties of share volume; in particular, we analyze the tail statistics of the distribution of volume, which is important in evaluating the validity of different models of market microstructure [22].

We define the trade size $q\equiv q_i$ as the number of shares exchanged in trade i, and the share volume as the total number of shares exchanged in a time interval Δt,

$$Q\equiv Q_{\Delta t}(t)\equiv\sum_{i=1}^{N}q_i. \tag{1}$$

Here $N=N_{\Delta t}(t)$ denotes the number of trades in Δt.

Previous analysis [20] for U.S. stocks reports power-law distributions

$$P\{q>x\}\sim x^{-\zeta_q} \tag{2}$$

for trade size q and

$$P\{Q>x\}\sim x^{-\zeta_Q} \tag{3}$$

for share volume Q with average tail exponents $\zeta_q=1.53\pm0.07$ and $\zeta_Q=1.7\pm0.1$, both belonging to the Lévy stable domain [0,2]. In their analysis of the NASDAQ order book data, the authors of Ref. [23] note that market order sizes display a power-law distribution with exponent 1.4±0.1, which is consistent with the estimate of ζ_q reported in Ref. [20]. Reference [24] analyzes stocks traded in the London Stock Exchange and reports values of ζ_q consistent with the ζ_q estimates for U.S. stocks (similar results can also be found in Ref. [22]). In their analysis of the dollar volume ("traded value") for U.S. stocks, Ref. [25] confirms a power-law tail for the distribution of Q, although the exponent estimates are reported to be outside the Lévy stable domain.

The goal of this work is to examine the validity of the power-law tails of the distributions of trade size q and share volume Q, and compare the tail behavior of these distributions between three distinct markets. We first focus on the more fundamental quantity, the trade size q, and quantify its tail statistics. Using two different statistical estimators—Hill's estimator [26] and the threshold-independent Meerschaert-Scheffler (MS) estimator [27]—we find exponent estimates for ζ_q within the Lévy stable domain for all three markets analyzed. Moreover our analysis shows that the exponent ζ_q is universal in the following respects: (a) ζ_q does not display any systematic dependence on stock-specific quantities such as the market capitalization or industry sector, and (b) ζ_q estimates are consistent across all three markets.

Next we analyze the statistics of the share volume $Q_{\Delta t}(t)$ over short time intervals Δt. Our analysis shows that, with increasing Δt, standard methods such as Hill's estimator can give biased estimates for ζ_Q with results that depend significantly on the estimation threshold which specifies the domain where the power law is expected to hold. Instead, we use a different estimator, the MS estimator [27], which, unlike Hill's estimator, does not rely on an estimation threshold. Using the threshold-independent MS estimator, we find that, for short time intervals Δt, the distribution $P\{Q>x\}$ is consistent with a power-law behavior with exponent ζ_Q in the Lévy stable domain for all three markets. Moreover, using data from the Center for Research in Security Prices (CRSP) database, we confirm the Lévy stable behavior for ζ_Q for $\Delta t=1$ day.

II. DATA ANALYZED

We analyze the following databases.

(a) *Trades and Quotes (TAQ) database (U.S. stocks)*. Tick by tick data for the 1000 largest U.S. stocks from the TAQ database [28] for the 2-y period 1994–1995 [10]. These 1000 stocks cover a wide range of market capitalization and industry sectors.

(b) *London Stock Exchange (LSE) database (U.K. stocks)*. Tick by tick data [29] for 85 stocks from the London Stock Exchange (Stock exchange Electronic Trading System (SETS) traded) for the 2-y period 2001–2002, which were part of the Financial Times Stock Exchange (FTSE) 100 index on January 2001 and survived through the period analyzed.

(c) Paris Bourse database. Tick by tick data for 13 stocks traded in the Paris Bourse which are part of the "Cotation Assistie en Continu" CAC 40 index and survived through the 4.5-y period from 3 January 1995 to 22 October 1999 (this database is analyzed in Ref. [21]).

(d) *CRSP database (daily data for USA stocks)*. To examine the behavior of the distribution $P(Q>x)$ over a larger Δt, we analyze daily volume data from the CRSP database [30] for 252 stocks for the 42-y period January 1963–December 2005.

III. SCALE-INVARIANT DISTRIBUTION OF TRADE SIZE *q*

To understand the behavior of the distribution $P(Q)$, it is important to first understand the behavior of a more fundamental quantity—the trade sizes q_i which form Q [see Eq. (1)]. In the following analysis we normalize the trade size $q\equiv q_i$ and the share volume $Q\equiv Q_{\Delta t}(t)$ by the total number of outstanding shares to account for share splits.

A. Database 1: TAQ database (U.S. stocks)

1. Prior analysis

Based on an analysis of the 1000 largest U.S. stocks, Ref. [20] reports that the distribution of trade size q follows a power law

$$P_{\mathrm{USA}}\{q>x\} \sim x^{-\zeta_q}, \tag{4}$$

with an average exponent estimate ζ_q within the Lévy stable domain. Using Hill's estimator [26], Ref. [20] reports a mean value of

$$\zeta_q^{\mathrm{USA}} = 1.53 \pm 0.07 \quad \text{(Hill estimator)}. \tag{5}$$

An examination of scaling behavior of the moments of q shows anomalous scaling behavior just as would be expected for a stable distribution; estimating the tail exponent based on the behavior of moments, Ref. [20] finds an average value

$$\zeta_q^{\mathrm{USA}} = 1.45 \pm 0.03 \ \text{(Moments)}. \tag{6}$$

2. Universality: Lack of dependence on market capitalization and industry sector

For the same set of 1000 U.S. stocks, we first analyze the dependence of the power-law exponent ζ_q on market capitalization and industry sector. Figures 1(a) and 1(c) show the exponent ζ_q plotted against the average market capitalization for each stock. A logarithmic regression $\zeta_q=\alpha \log S+\epsilon$ shows no significant dependence (see the caption of Fig. 1). Figures 1(b) and 1(d) show that the exponent ζ_q does not show any systematic dependence for any particular industry sector. That the exponent estimates ζ_q do not show systematic variations with either market capitalization or industry sector is consistent with the possibility that the distribution $P(q)$ displays a universal functional form for all stocks.

Although the functional form of the distribution of q [Eq. (4)] and the exponent values do not depend on market capitalization $\langle S\rangle$, the "width" is stock specific; i.e., average trade size $\langle q_i\rangle$ for each stock displays a striking power-law dependence on capitalization (Fig. 2),

$$\langle q\rangle \sim S^{-\beta}, \tag{7}$$

with an exponent

$$\beta = 0.67 \pm 0.02. \tag{8}$$

More precisely, the dependence of the distribution $P(q|S)$ on the market capitalization can be expressed as

$$P(q|S) \sim S^{\beta} f(q/S^{-\beta}), \tag{9}$$

where f does not depend on S. We note that Eq. (7) is analogous to the dependence of the average volatility on the market capitalization [10]. While the functional form of the individual stock return distributions and the power-law

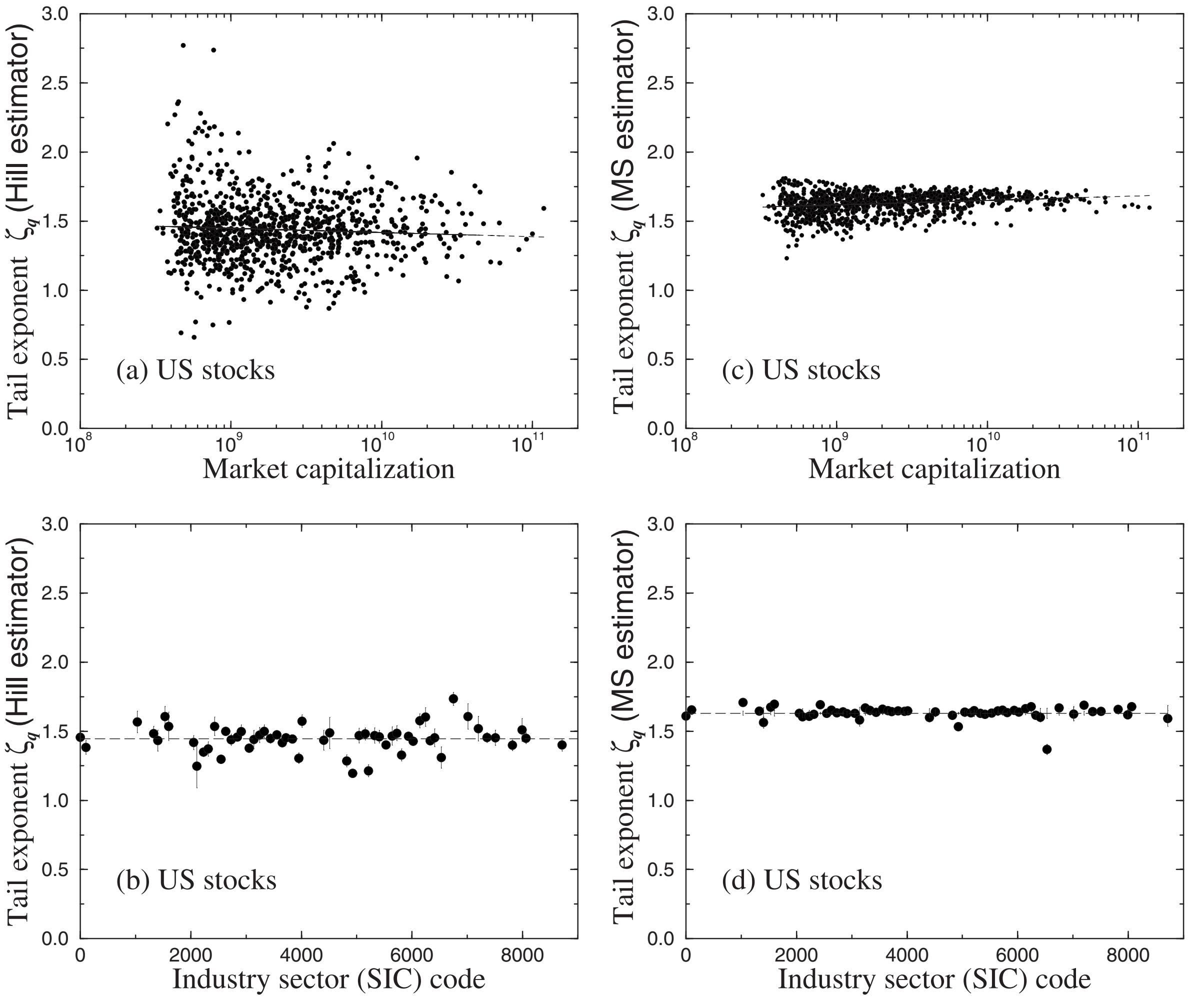

FIG. 1. (a) Hill estimate of exponents ζ_q for the 1000 U.S. stocks against average market capitalization for each stock. The line shows a logarithmic regression which shows no significant dependence (we obtain a slope -0.013 ± 0.006 with $R^2=0.004$). Note that the exponents in this plot are obtained using Hill's estimator using a tail threshold of five times the average value. Using estimation thresholds of up to 10, we obtain average values in the range 1.45–1.67 with no significant dependence on the capitalization. (b) Exponents ζ_q as a function of the Standard Industrial Classification (SIC) code shows no clear dependence on the industry sector. Here we have binned using the first 2 digits of the SIC code [37] which shows major industry sectors. The points at SIC code 0 show the 73 stocks in our sample of 1000 for which we did not have the corresponding SIC codes. (c) Same as (a) but using the threshold-independent MS estimator (see the Appendix) instead of Hill's estimator. The plot shows no dependence on market capitalization, consistent with part (a). A log-linear regression shows no statistically significant dependence; we find a slope of 0.014 ± 0.002 with negligible significance, $R^2=0.038$. (d) Same as (b) but using the MS estimator; there is no significant dependence on any particular industry sector. The dashed line shows the average value $\zeta_q=1.63$.

exponents that describe the tails do not display any clear dependency on the market capitalization, the average volatility (analogous to the average trade size) $\sigma(S)$ depends on the market capitalization S as a power law [10] with an exponent ≈ 0.20. A similar set of results for LSE data can be found in Ref. [31].

B. Database 2: LSE database (U.K. stocks)

In order to examine the robustness of the power-law exponent ζ_q for different markets, we next analyze the statistics of trade size for the U.K. data. We analyze the probability density function for each stock and find that it is consistent with a power-law decay. Scaling q for each stock by its first centered moment, we obtain a good data collapse. Using the normalized q for all stocks together, we find that the probability density function $P_{LSE}(q)$ is consistent with a power-law decay [Fig. 3(a)],

$$P_{\mathrm{LSE}}(q) \sim q^{-(\zeta_q+1)}, \tag{10}$$

with exponent ζ_q.

For the 85 stocks in our sample, we estimate the exponent ζ_q individually for each stock using Hill's estimator [Fig. 3(b)]. We obtain an average exponent estimate [32]

$$\zeta_q^{\mathrm{LSE}} = 1.57 \pm 0.02 \quad (\text{Hill estimator}), \tag{11}$$

consistent with our previous results [20,24,33].

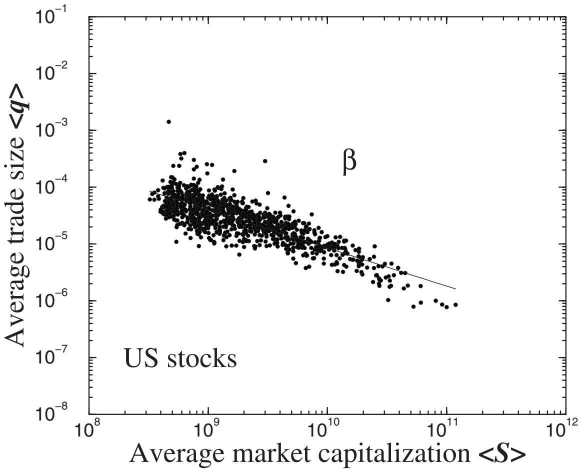

FIG. 2. Average trade size for U.S. stocks as a function of capitalization, showing a power-law dependence. To account for the effect of share splits, the trade sizes have been normalized by the total number of outstanding shares. The abscissa shows the average market capitalization in the 2 y period 1994–1995. The line shows a log-log regression, from which we obtain a slope $\beta=0.67\pm0.02$.

The Hill exponent estimates show some variability in the estimation threshold a, beyond which the power law $P(q>x)\sim x^{-\zeta_q}$ is expected to hold (see the Appendix). Figure 3(c) shows the average of the exponent estimate as a function of the estimation threshold. The mean value seems to increase and then plateau below $\zeta_q=2$, the upper bound for Lévy stability. The apparent increase of the exponent estimate with the estimation threshold is likely an artifact of Hill's estimation technique. While in the presence of sufficient number of data points Hill's estimator is a reliable method to estimate power-law exponents, the average of the exponent estimates across many stocks is biased when the estimation threshold is large and there are few points available for the tail estimation for each stock. This bias is enhanced in the presence of an outlier for any one stock, which biases the sample mean of the tail exponent to larger values.

A reliable estimate of the tail exponent is obtained if we use an estimator that does not depend on a threshold for obtaining the exponent estimate. One such estimator is the MS estimator [27] (see the Appendix). Figure 3(d) shows the estimate of ζ_q obtained from the MS estimator [27], from which we obtain

$$\zeta_q^{\mathrm{LSE}} = 1.58 \pm 0.01 \quad \text{(MS estimator)}, \tag{12}$$

which is consistent with Eq. (11), and within the Lévy stable domain. We shall further expand on the use of this estimator in a later section.

C. Database 3: Paris Bourse database

Next, we analyze the distribution of trade size for the Paris Bourse data. As in the analysis for the U.K. data, we find that the distribution $P(q)$ has the same functional form for each of the 15 stocks in our sample. Scaling q by the first centered moment, we find good data collapse. Under the assumption that all the underlying distributions are identical, we use the scaled data for all stocks to improve the tail statistics. Figure 4(a) shows that the probability distribution is consistent with a power law

$$P_{\mathrm{Bourse}}(q) \sim q^{-(\zeta_q+1)}, \tag{13}$$

with exponent ζ_q.

Using Hill's estimator, we obtain exponent estimates for ζ_q for each stock individually and find a mean value

$$\zeta_q^{\mathrm{Bourse}} = 1.53 \pm 0.04 \quad \text{(Hill estimator)}, \tag{14}$$

with only small dispersion among the stocks. Note that our estimate of ζ_q for the Paris Bourse data is consistent with our previous results for the New York Stock Exchange (NYSE) and for the LSE [Fig. 4(b)].

Figure 4(b) shows the exponent ζ_q as a function of the estimation threshold for Paris Bourse data. Here the exponent estimates plotted are the average values of the individual exponent estimates for each stock for a particular estimation threshold. Unlike for the U.K. data, we do not find any dependence with increasing threshold [Fig. 4(b)]. We find only statistical deviations around the mean value $\zeta_q=1.46$.

To compare the U.K. and Paris Bourse data to the U.S. data, we selected the 116 most actively traded stocks and analyzed the behavior of ζ_q with the estimation threshold. As in the U.K. data, we find an apparent tendency of the Hill exponent estimate to increase with the estimation threshold and plateau at a value <2 within the Lévy stable domain [Fig. 5(a)]. As we previously noted, the dependence of the Hill exponent estimate on the estimation threshold is an artifact of the technique for large estimation thresholds. Figure 5(b) shows that the threshold-independent MS estimator, when applied to the same set of 116 most actively traded stocks, provides an estimate $\zeta_q^{\mathrm{USA}}=1.65\pm0.01$, within the Lévy stable domain.

IV. SCALING OF SHARE VOLUME Q

A. Intraday time scales $\Delta t<1$ day

Previous work [20] analyzed the distribution of Q for U.S. stocks by splitting the 1000 stocks into different groups based on their liquidity (average time between trades) and reported a power-law distribution $P\{Q>x\}\sim x^{-\zeta_Q}$, where $\zeta_Q=1.7\pm0.1$, within the Lévy stable domain, on average for the 1000 largest stocks.

In a separate work [25], a similar analysis is performed on the dollar volume ("traded value") for U.S. stocks. While these authors confirm a power-law tail for the distribution of Q, they find a tail exponent that is outside the Lévy stable domain. Reference [25] concludes that the distribution $P(Q)$ is not Lévy stable since (a) larger time windows yield larger values of exponents, and (b) increasing the threshold used for estimating the Hill exponents gives rise to larger estimates.

Analyzing the tail behavior of Q over larger time windows and for large estimation thresholds is not as straightforward, and the reason for this can be seen from Eq. (1). As

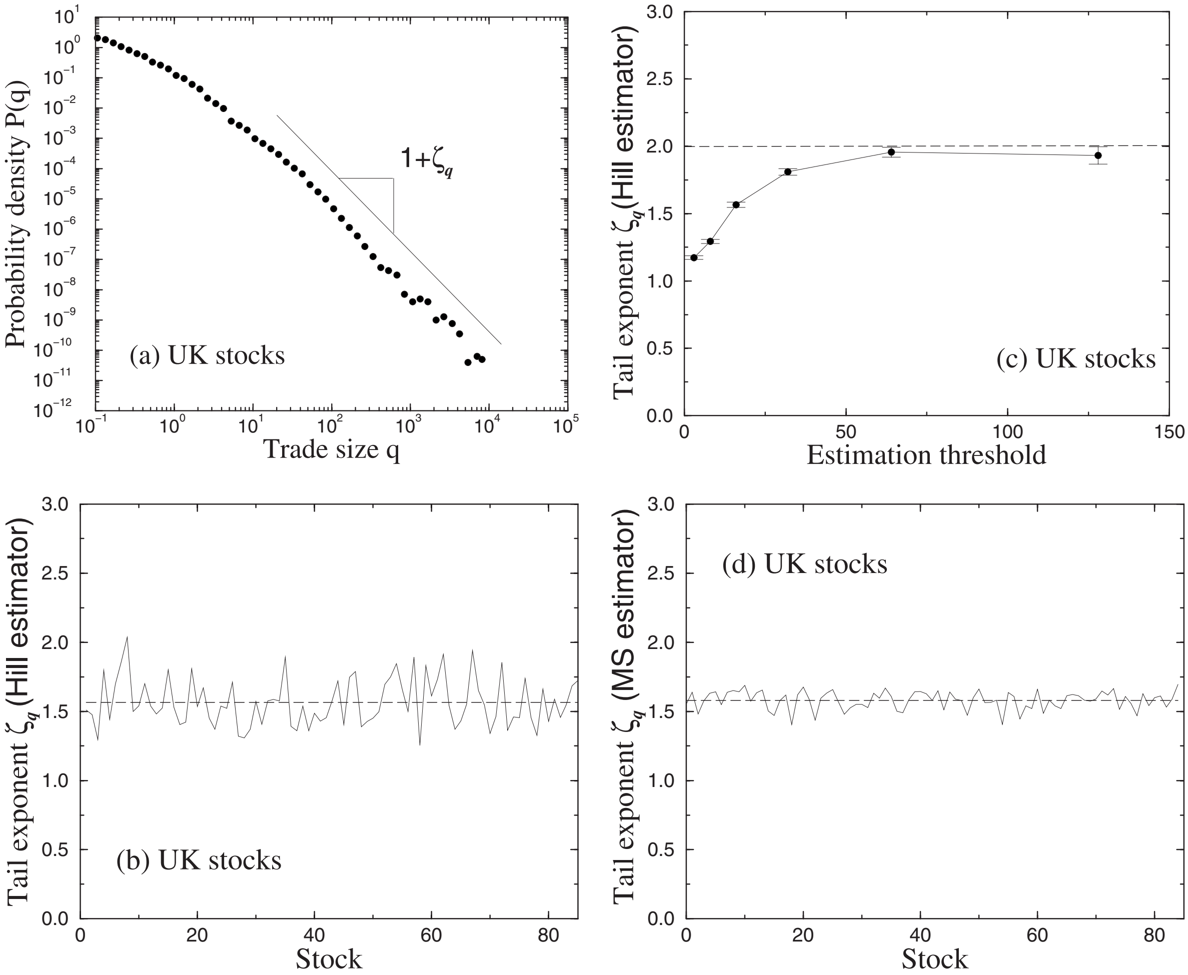

FIG. 3. (a) Probability density function of trade sizes for 85 SETS stocks that form part of the FTSE 100 index and survived through the 2 y period 2001–2002. The time series of q_i for each stock is normalized by its first centered moment. We find a power-law decay with an exponent $1+\zeta_q$. (b) We estimate the exponents ζ_q by applying Hill's estimator for each stock. For this plot, we have used a threshold of 16 times the average trade size for the estimation procedure. We obtain an average $\zeta_q=1.57\pm0.02$ [32]. (c) Average Hill exponent estimate ζ_q as a function of estimation threshold [in the same units as in the abscissa of (a)], showing an apparent increase of the exponent followed by a plateau within the Lévy stable domain. (d) Exponent estimate ζ_q obtained by using the MS estimator which, unlike Hill's estimator, does not depend on estimation threshold. We obtain a mean value $\zeta_q=1.58\pm0.01$.

Δt increases, the number of trades $N_{\Delta t}$ also increases, and consequently the mean of $Q_{\Delta t}$. Techniques such as Hill's estimator are reliable only for a pure power-law distribution. Unlike returns, which have almost zero mean, trade sizes under aggregation acquire increasingly larger mean values, so the effective region where the power law is expected to hold shrinks considerably. In addition, the number of data points decreases with aggregation, thereby providing hardly any resolution in the tail. Thus, for a distribution such as $P\{Q>x\}$, Hill's estimator gives biased results with increasing Δt, and with increasing estimation threshold. Therefore an increase in the Hill exponent estimate upon increasing Δt or upon increasing the estimation threshold may not be reflective of the actual behavior of the distribution.

To understand the behavior of Hill's estimator numerically, we consider the partial sum P_n defined as

$$P_n \equiv \sum_{i=1}^{n} x_i, \tag{15}$$

where $x_i>0$ is perfectly power-law distributed with $\zeta_x=1.5$. Since x_i is generated to have $\zeta_x=1.5$ in the Lévy stable domain, we expect the actual tail behavior to persist under aggregation. Figure 6(a) shows the Hill estimate of ζ_P as a function of estimation threshold for increasing n. For $n=1$, as expected, there is no dependence on threshold, and Hill's estimator yields $\zeta_P=1.5$. On increasing to $n=5$, we see that this curve starts to display a peak, i.e., an increase, which then decays and approaches the true asymptotic value of ζ_P $=1.5$. On further increasing n, we find that this peak is much more pronounced; moreover, since the number of data points shrinks with aggregation, the Hill estimate never reaches its true asymptotic value. This can be clearly seen for the case

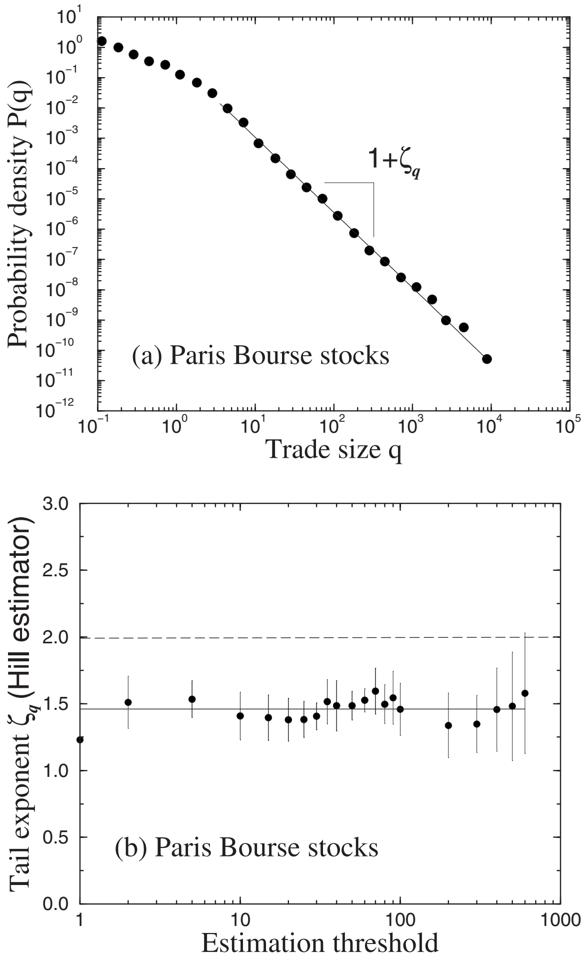

FIG. 4. (a) Probability density function of the trade size for 13 largest stocks listed in the Paris Bourse. Here we have normalized the time series of q_i for each stock by its first centered moment. Power-law fits yield $\zeta_q=1.49\pm0.03$. (b) Hill estimate of the exponent ζ_q averaged over the 15 Bourse stocks as a function of estimation threshold, showing no significant dependence, nor any indication of truncation. The line shows the mean value 1.46.

$n=100$, where the exponent estimate "stabilizes" even outside the Lévy stable domain. Thus, although we started with a genuine Lévy stable law, upon aggregation, Hill's estimator provides a biased estimate of the tail exponent that depends on the estimation threshold. Moreover, for a given estimation threshold, Fig. 6(a) shows that the Hill estimate of ζ_P increases monotonically with n, although ζ_P is within the stable Lévy domain.

We next analyze the behavior of the Hill estimate for the volume exponent ζ_Q when applied to the Paris Bourse data. We choose the Paris Bourse data for this analysis since this is the longest time series in our data, spanning more than four years. Figure 6(b) shows the mean of the Hill exponent estimates ζ_Q over all 13 stocks for the Paris Bourse data for $\Delta t=5$ min as a function of the estimation threshold. We first find an increase in the exponent estimate, which goes above the Lévy stable domain and then retreats back and stabilizes around the value of $\zeta_Q\approx1.5$, similar to the exponent estimate for ζ_q as expected for a stable distribution.

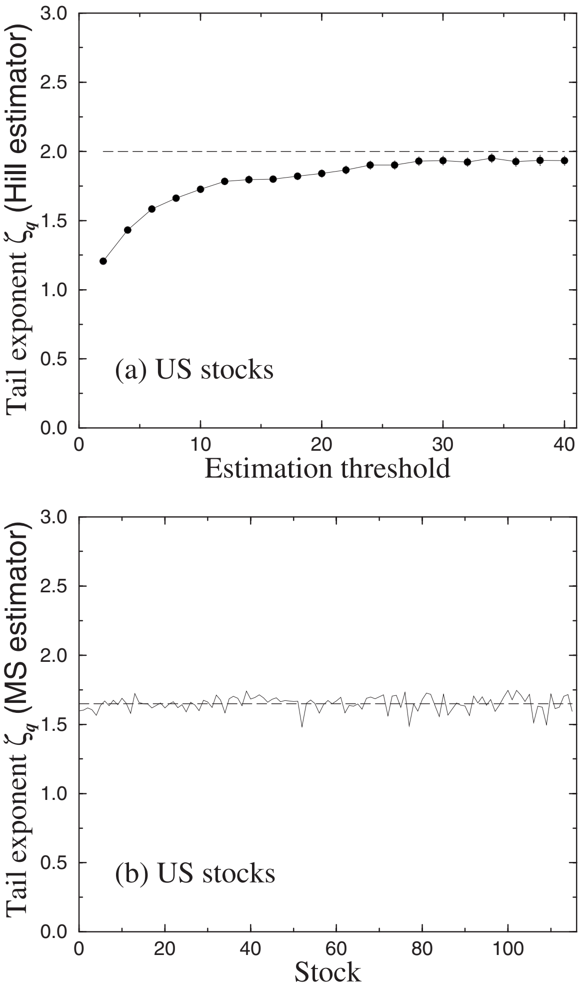

FIG. 5. (a) Hill estimate of the exponent ζ_q averaged over the 116 largest U.S. stocks as a function of the estimation threshold. The apparent increase followed by a plateau within the Lévy stable domain is similar to the behavior observed for the U.K. data in Fig. 3(c). (b) Using the MS estimator we obtain a mean $\zeta_q=1.65\pm0.01$.

Figure 6(c) (filled points) shows the same plot as Fig. 6(a) but for $\Delta t=30$ min. Here we see that the increase in exponent estimate is followed by a seeming plateau behavior around a value slightly larger than $\zeta_Q=2$, outside the Lévy stable domain. In order to compare with the surrogate data, instead of creating partial sums as in Eq. (15), which holds n constant, we construct a directly comparable case by defining

$$P_{\Delta t}\equiv\sum_{i=1}^{N_{\Delta t}}x_i, \tag{16}$$

where $N_{\Delta t}$ denotes the actual number of trades in Δt from the empirical data. By construction, the surrogate time series should also be Lévy stable since $N_{\Delta t}$ has tails that decay much more rapidly [19]. Figure 6(c) (empty symbols) shows the behavior of the Hill exponent estimate ζ_P, defined by $P\{P>x\}\sim x^{-\zeta_P}$, as a function of the estimation threshold. We find virtually indistinguishable behavior from the behavior of ζ_Q for the empirical data.

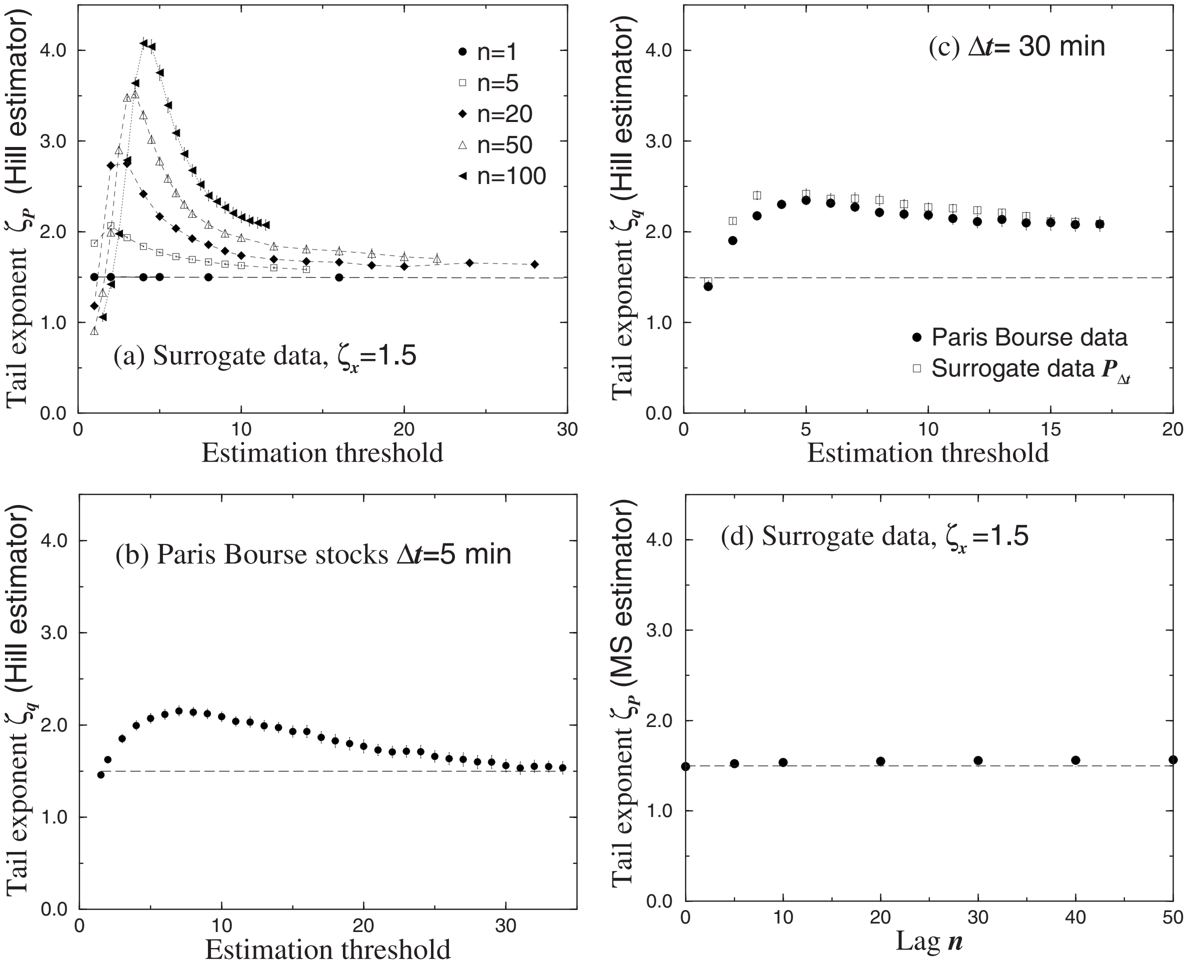

FIG. 6. (a) Hill exponent estimate as a function of estimation threshold for the surrogate data P_n defined in Eq. (15) for different n. We have considered 100 time series, each with 200 000 points at $n=1$. Each time series is generated to have a tail exponent of $\zeta_x=1.5$. For a given n and estimation threshold, each point denotes the average over the 100 time series of the exponent estimates of P_n. (b) Hill estimate of the tail exponent ζ_Q as a function of estimation threshold for the Paris Bourse data for $\Delta t=5$ min. We plot the average value of ζ_Q for the 13 Bourse stocks which are comparable in liquidity. (c) Hill exponent estimate for the Paris Bourse data for $\Delta t=30$ min (filled symbols) as a function of estimation threshold compared against the same technique applied to the surrogate data (empty symbols) constructed by Eq. (16). (d) MS exponent estimate for the surrogate data in Eq. (15) for different n. As in (a) we have considered 100 time series, each with 200 000 points generated to have a tail exponent of $\zeta_x=1.5$. Each point denotes the average over the 100 time series of the exponent estimates of P_n.

Thus, the behavior of the Hill exponent estimate for variables such as the aggregated Δt share volume $Q_{\Delta t}$ is unreliable, and increasingly so as Δt increases. Therefore it is likely that the exponent estimate at a level larger than the Lévy stable limit of $\zeta_Q=2$ [e.g., Fig. 6(c), or some of the results in Ref. [25], using methods that depend on estimation thresholds] is an artifact of the estimation procedure itself, as suggested by the surrogate data in the above numerical experiment.

To obtain a reliable estimate of the tail exponent, we use the MS estimator [27] introduced in the previous section, which does not have an estimation threshold, but instead uses all the data to estimate the tail exponent (see the Appendix). Figure 6(d) shows the behavior of the MS exponent estimate for the surrogate data P_n as a function of n. Clearly, the exponent estimate does not show dependence on n, as would be expected for a stable law. For subsequent analysis for Q, we rely on the MS estimator, since it avoids most of the problems that we encounter when using Hill's method for quantities such as Q.

Applying the MS estimator to the U.S. data over the time interval $\Delta t=15$ min, we obtain the average exponent estimate (Fig. 7)

$$\zeta_Q^{\mathrm{USA}} = 1.69 \pm 0.01 \quad \text{(MS estimator)}, \tag{17}$$

which is consistent with the results previously obtained in Ref. [20]. A similar analysis of ζ_Q for the 85 stocks in the U.K. data using the MS estimator yields a consistent value of

$$\zeta_Q^{\mathrm{LSE}} = 1.67 \pm 0.01 \quad \text{(MS estimator)}, \tag{18}$$

where $\Delta t=5$ min [Fig. 7(b)]. For comparison, we apply the MS estimator for the Paris Bourse data. Since we have 30-min data for 34 stocks (compared to the 13 for which we

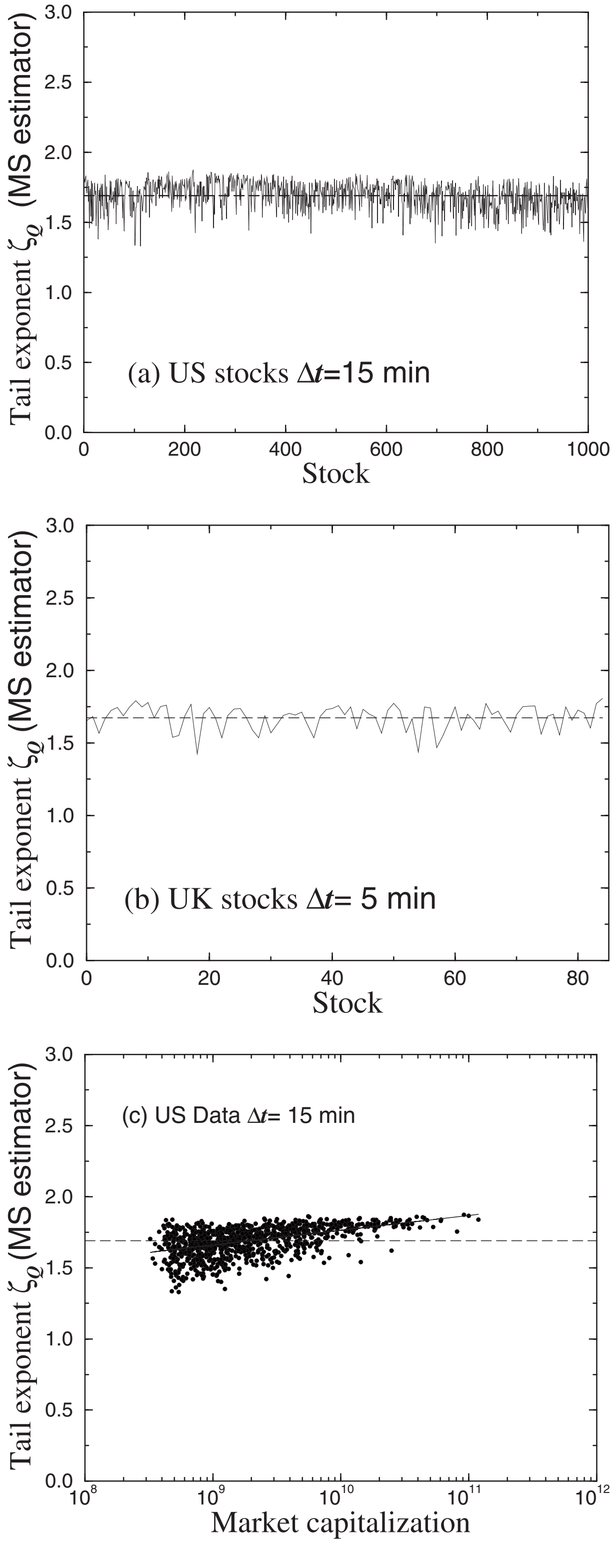

FIG. 7. (a) MS estimate of ζ_Q exponent for 1000 U.S. stocks over $\Delta t = 15$ min. Here we have normalized the data by the first centered moment. We obtain a mean value of $\zeta_Q = 1.69 \pm 0.01$. (b) Same as (a) for $\Delta t = 5$ min for the U.K. data. We find a mean value of $\zeta_Q = 1.67 \pm 0.01$. (c) MS exponent estimate of ζ_Q for the 1000 U.S. stocks at $\Delta t = 15$ min against the average market capitalization over the 2 y period. The solid line shows a logarithmic regression which gives a slope 0.045 ± 0.003, with $R^2 \approx 0.25$.

have tick data), we apply the MS estimator to all 34 stocks, and find

$$\zeta_Q^{\text{Bourse}} = 1.69 \pm 0.01 \quad \text{(MS estimator)}, \tag{19}$$

which is consistent within error bars with the estimates of ζ_Q for both the U.S. and the U.K. data.

Figure 7(c) shows the MS exponent estimate of ζ_Q for all 1000 U.S. stocks as a function of their average market capitalization over the 2 y period. A logarithmic regression shows a small dependence as evidenced by the slope 0.045 ± 0.003 [see the caption of Fig. 7(c)]. Although the statistical significance of this regression is small ($R^2 = 0.25$), it does not allow for a definitive conclusion regarding the universality of the exponent ζ_Q. Note that this is quite in contrast to the case of ζ_q, for which the dependence on market capitalization is clearly statistically insignificant [Figs. 1(a) and 1(c)]. In the small trend in Fig. 7(c) is genuine, this may be suggestive of an eventual truncation of the power-law at very large values.

B. Database 4: CRSP database, volume distribution for Δt=1 day

Given that the exponents ζ_q and ζ_Q are within the Lévy stable domain, one expects the distribution $P(Q)$ to be stable under aggregation. Conversely, an increase of exponent estimate at daily time scales would indicate the presence of a truncation of the distribution for large values [34].

To examine whether the stability is empirically observed, we analyze daily U.S. data which record the time series of volumes for 252 stocks from the CRSP database for the 43 y period 1962–2005. We have normalized the daily volumes by the number of outstanding shares to adjust for the effect of stock splits. Since the average volume increases considerably over this large time period, we separate each time series into blocks of 5 y windows and normalize each block by the average volume over that period.

We find that for each stock the distribution of the daily volumes displays a power-law decay,

$$P_{\text{USA daily}}\{Q > x\} \sim x^{-\zeta_Q}, \tag{20}$$

with exponent ζ_Q. To estimate for ζ_Q, we apply the MS estimator [27] to each stock individually. Figure 8 shows the exponents thus obtained. We find a mean value of

$$\zeta_Q^{\text{USA daily}} = 1.78 \pm 0.01 \quad \text{(MS estimator)}, \tag{21}$$

which is consistent with our estimate for the $\Delta t = 15$ min volume for the U.S. intraday data, suggesting a stable Lévy distribution of volume.

C. Time scaling of the distribution of volume

Since ζ_Q is within the Lévy stable domain, we expect the estimates of ζ_Q to remain stable with increasing Δt. We analyze the behavior of ζ_Q for U.S. data using the TAQ database for $\Delta t < 1$ day and the CRSP database for $\Delta t > 1$ day. Figure 9 shows that, for Δt varying from 15 min up to 8 days, ζ_Q consistently remains within the Lévy stable domain.

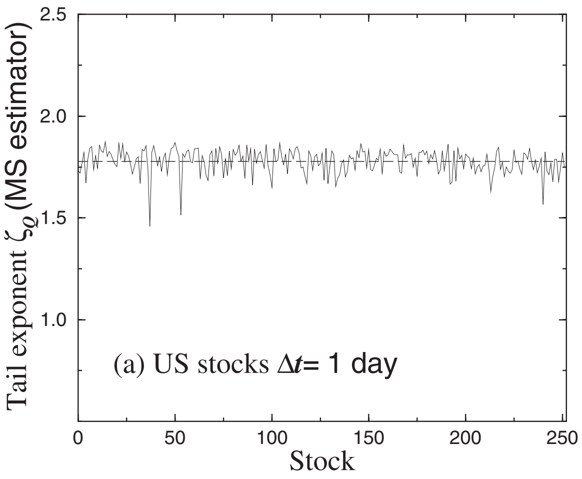

FIG. 8. MS estimate of the power-law exponent ζ_Q for 252 stocks from the daily U.S. data, showing a mean value of 1.78 and standard deviation of 0.06. We have normalized the data for each stock by first subdividing the 43 y period 1962–2005 into eight separate subperiods. Within each subperiod, we normalize Q by its average value within that period.

V. SUMMARY

We focused on quantifying the tail statistics of trade size and share volume. Analyzing tick data for three distinct markets, we find that the distribution of trade size q displays a power-law tail $P\{q>x\} \sim x^{-\zeta_q}$ with exponent ζ_q within the Lévy stable domain $[0,2]$. We next analyze the distributions of share volume $Q_{\Delta t}$ over short time intervals and find a power-law distribution $P\{Q>x\} \sim x^{-\zeta_Q}$ with exponent ζ_Q within the Lévy stable domain. We find consistent results for daily data from the CRSP database, confirming a Lévy stable behavior of the distribution of share volume for $\Delta t = 1$ day.

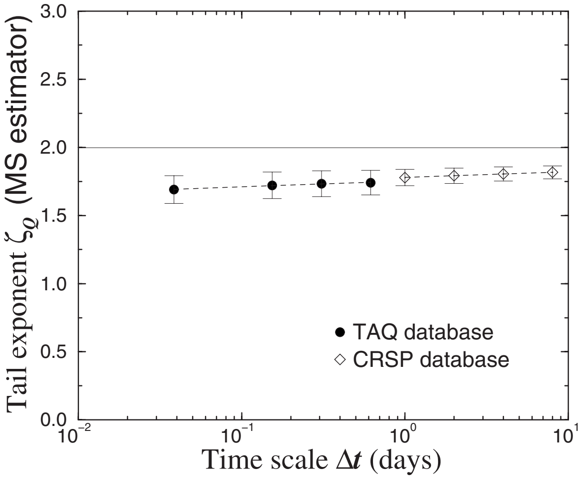

FIG. 9. MS estimate of the power-law exponent ζ_Q for time scales Δt ranging from 15 min up to 8 days, showing stable behavior within the Lévy stable domain. Here we have used the TAQ database for $\Delta t < 1$ day and the CRSP database for $\Delta t > 1$ day. The exponents seem mutually consistent within error bars for both the TAQ and the CRSP data. The dashed line shows a regression $y = A \log x + B$, and we obtain $A = 0.018 \pm 0.002$ for $\Delta t < 1$ day and $A = 0.019 \pm 0.001$ for $\Delta t > 1$ day.

Our analysis of the exponent estimates for ζ_q suggests that the exponent value is universal in the following respects: (a) ζ_q is consistent across stocks within each of the three markets, and also across different markets; (b) ζ_q does not display any systematic dependence on market capitalization or industry sector. These results are particularly interesting since universal behavior of tail exponents is suggestive of underlying mechanisms that are largely independent of microstructural details, analogous to scaling exponents that occur in strongly interacting physical systems [35].

ACKNOWLEDGMENTS

We are indebted to P. Gopikrishnan and X. Gabaix for many helpful discussions on this topic. This paper has also benefited from suggestions made by an anonymous referee, Sec. IV C in particular. We are grateful to Carl Hopman for the Paris Bourse data, and Thomas O'Brian of the London Stock Exchange for the LSE data. We thank the NSF and Morgan Stanley Microstructure Research Grant for support.

APPENDIX: METHODS FOR ESTIMATING POWER-LAW EXPONENTS

1. Hill's estimator

Hill's estimator is the conditional maximum likelihood estimator for a true power-law distribution based on the k largest order statistics [26]. Denote the data series as Y for which we expect a power-law behavior $P(Y>x) \sim x^{-\zeta}$. Define the inverse local slope of the cumulative distribution function $C(x) \equiv P(q>x)$,

$$\xi \equiv \zeta^{-1} = -[d \log C(x)/d \log x]^{-1}. \qquad \text{(A1)}$$

We obtain an estimator for ξ by sorting the Y by their size, $Y^{(1)} > Y^{(2)} > \cdots > Y^{(N)}$. The cumulative distribution can then be written as $P(Y^{(k)}) = k/N$, and we obtain for the local slope

$$\xi^{\text{Hill}} = \left((N-1) \sum_{i=1}^{N-1} \log Y^{(i)} \right) - \log Y^{(N)}, \qquad \text{(A2)}$$

where N is the number of tail events used. When applying Hill's estimator, k should be as large as possible but small enough that the chosen points are within the distributional tail where the power law holds. Instead of restricting ourselves to a fixed number of tail events, we specify an estimation threshold a beyond which we expect the power law to hold, i.e., $P(Y>x) \sim x^{-\zeta}$ for $x>a$. This procedure also allows us to compare the tail exponents of different stocks since we *normalize* the data series by the first centered moment. Thus we apply Eq. (A2) for all events such that $x>a$.

2. MS estimator

Hill's estimator can give misleading results when applied to stable data [36]; in particular, for the values $1.5 < \zeta < 2$, Hill's estimator can give estimates of ζ much larger than 2 although the true exponent is less than 2. In particular, one of the drawbacks of Hill's estimator is its dependence on the number k of order statistics (or equivalently the estimation threshold) that is used in the estimation.

Reference [27] developed a method for estimating the thickness of heavy tails based on the asymptotics of sums. This robust estimator depends only on the tail exponent ζ and not on the exact form of the distribution. The estimator works for dependent data as well, and performs well when ζ approaches 2. The central advantage of this estimator is that it uses all the data for estimating ζ and does not depend on any estimation threshold, unlike Hill's estimator. Using data denoted Y_i, with $i=1,\ldots N$, the MS estimator for $\zeta \equiv \xi^{-1}$ is

$$\xi^{\mathrm{MS}} = \frac{\gamma + \log_{+} \sum_{i=1}^{N} (Y_i - \langle Y \rangle)^2}{2(\gamma + \log N)}, \tag{A3}$$

where $\langle Y \rangle$ is the sample mean, $\log_{+} Y \equiv \max(\log Y, 0)$, and $\gamma = 0.5772$ is Euler's constant.

[1] *Practical Fruits of Econophysics: Proceedings of The Third Nikkei Econophysics Symposium*, edited by H. Takayasu (Springer, New York, 2006).

[2] *The Science of Disasters*, edited by A. Bunde *et al.* (Springer, Berlin, 2002).

[3] E. Derman, *My Life as a Quant: Reflections on Physics and Finance* (Wiley, New York, 2004).

[4] H. E. Stanley and V. Plerou, Quant. Finance **1**, 563 (2001).

[5] L. Laloux, P. Cizeau, J. -P. Bouchard and M. Potters, Phys. Rev. Lett. **83**, 1467 (1999).

[6] V. Plerou, P. Gopikrishnan, B. Rosenow, L. A. Nunes Amaral, and H. E. Stanley, Phys. Rev. Lett. **83**, 1471 (1999); V. Plerou, P. Gopikrishnan, B. Rosenow, Luis A. Nunes Amaral, T. Guhr, and H. E. Stanley, Phys. Rev. E **65**, 066126 (2002).

[7] M. Loretan and P. C. B. Phillips, J. Empirical Finance **1**, 211 (1994).

[8] T. Lux, Appl. Financ. Econ. **6**, 463 (1996).

[9] P. Gopikrishnan, V. Plerou, L. A. Nunes Amaral, M. Meyer, and H. E. Stanley, Phys. Rev. E **60**, 5305 (1999).

[10] P. Gopikrishnan, M. Meyer, L. A. N. Amaral and H. E. Stanley, Eur. Phys. J. B **3**, 139 (1998); V. Plerou, P. Gopikrishnan, L. A. Nunes Amaral, M. Meyer, and H. E. Stanley, Phys. Rev. E **60**, 6519 (1999).

[11] J. A. Skjeltorp, Physica A **283**, 486 (2001).

[12] P. K. Clark, Econometrica **41**, 135 (1973).

[13] B. B. Mandelbrot and H. Taylor, Oper. Res. **15**, 1057 (1967).

[14] T. W. Epps and M. L. Epps, Econometrica **44**, 305 (1976).

[15] G. Tauchen and M. Pitts, Econometrica **57**, 485 (1983).

[16] J. Karpoff, J. Financ. Quant. Anal. **22**, 109 (1987).

[17] J. Stock, J. Am. Stat. Assoc. **83**, 77 (1988).

[18] C. Jones *et al.*, Rev. Financ. Stud. **7**, 631 (1994); A. R. Gallant *et al.*, *ibid.* **5**, 199 (1992).

[19] V. Plerou, P. Gopikrishnan, L. A. Nunes Amaral, X. Gabaix, and H. E. Stanley, Phys. Rev. E **62**, R3023 (2000).

[20] P. Gopikrishnan, V. Plerou, X. Gabaix, and H. E. Stanley, Phys. Rev. E **62**, R4493 (2000).

[21] C. Hopman (unpublished).

[22] X. Gabaix, P. Gopikrishnan, V. Plerou and H. E. Stanley, Quart. J. Econom. **121**, 461 (2006).

[23] S. Maslov and M. Mills, Physica A **299**, 234 (2001).

[24] V. Plerou, P. Gopikrishnan, X. Gabaix and H. E. Stanley, Quant. Finance **4**, C11 (2004); J. D. Farmer and F. Lillo, *ibid.* **4**, C7 (2004).

[25] Z. Eisler and J. Kertesz, Eur. Phys. J. B **51**, 145 (2006); Phys. Rev. E **73**, 046109 (2006); and (unpublished).

[26] B. M. Hill, Ann. Stat. **3**, 1163 (1975).

[27] M. Meerschaert and H. Scheffler, J. Stat. Plan. Infer. **71**, 19 (1998); see also M. Herrick *et al.*, Water Resour. Res. **38**, 9 (2002).

[28] *The Trades and Quotes Database* (New York Stock Exchange, New York, 1994–1995) (CD-ROM).

[29] Tick data for stocks traded in the London Stock Exchange can be obtained at their website, http://www.londonstockexchange.com/en-gb/products/informationproducts/historic/tickbest.htm

[30] Data provided by the Center for Research in Security Prices can be obtained at http://www.crsp.com

[31] G. Zumbach, Quant. Finance **4**, 441 (2004).

[32] The error bars throughout the remainder of this paper correspond to one standard deviation and are obtained by scaling the standard deviation of the exponent estimates for all stocks using the square root of the number of stocks. Note that this procedure implicitly assumes uncorrelated observations and therefore underestimates the true magnitude of the error since the ζ_q estimates will in reality be correlated across different stocks.

[33] X. Gabaix, P. Gopikrishnan, V. Plerou and H. E. Stanley, Nature (London) **423**, 267 (2003).

[34] R. N. Mantegna and H. E. Stanley, Phys. Rev. Lett. **73**, 2946 (1994).

[35] H. E. Stanley, Rev. Mod. Phys. **71**, S358 (1999).

[36] J. McCulloch, J. Bus. Econ. Stat. **15**, 74 (1997).

[37] More information about SIC codes can be found at http://www.osha.gov/pls/imis/sic-manual.html