Scaling of the distribution of price fluctuations of individual companies

Vasiliki Plerou,[1,2] Parameswaran Gopikrishnan,[1] Luís A. Nunes Amaral,[1] Martin Meyer,[1] and H. Eugene Stanley[1]
[1]*Center for Polymer Studies and Department of Physics, Boston University, Boston, Massachusetts 02215*
[2]*Department of Physics, Boston College, Chestnut Hill, Massachusetts 02167*
(Received 14 July 1999)

We present a phenomenological study of stock price fluctuations of individual companies. We systematically analyze two different databases covering securities from the three major U.S. stock markets: (a) the New York Stock Exchange, (b) the American Stock Exchange, and (c) the National Association of Securities Dealers Automated Quotation stock market. Specifically, we consider (i) the trades and quotes database, for which we analyze 40 million records for 1000 U.S. companies for the 2-yr period 1994–95; and (ii) the Center for Research and Security Prices database, for which we analyze 35 million daily records for approximately 16 000 companies in the 35-yr period 1962–96. We study the probability distribution of returns over varying time scales Δt, where Δt varies by a factor of $\approx 10^5$, from 5 min up to ≈ 4 yr. For time scales from 5 min up to approximately 16 days, we find that the tails of the distributions can be well described by a power-law decay, characterized by an exponent $2.5 < \propto < 4$, well outside the stable Lévy regime $0 < \alpha < 2$. For time scales $\Delta t \gg (\Delta t)_\times \approx 16$ days, we observe results consistent with a slow convergence to Gaussian behavior. We also analyze the role of cross correlations between the returns of different companies and relate these correlations to the distribution of returns for market indices. [S1063-651X(99)11412-0]

PACS number(s): 05.40.Fb, 05.45.Tp

I. INTRODUCTION

The study of financial markets poses many challenging questions. For example, how can one understand a strongly fluctuating system that is constantly driven by external information? And how can one account for the role of the feedback between the markets and the outside world, or of the complex interactions between traders and assets? An advantage for the researcher trying to answer these questions is the availability of huge amounts of data for analysis. Indeed, the activities at financial markets result in several observables, such as the values of different market indices, the prices of the different stocks, trading volumes, etc.

Some of the most widely studied market observables are the values of market indices. Previous empirical studies [1–12] show that the distribution of fluctuations—measured by the returns—of market indices has slow decaying tails, and that the distributions apparently retain the same functional form for a range of time scales [1,2,6,7]. Fluctuations in market indices reflect the average behavior of the price fluctuations of the companies comprising them. For example, the S&P 500 is defined as the sum of the market capitalizations (stock price multiplied by the number of outstanding shares) of 500 companies representative of the U.S. economy.

Here we focus on a more ''microscopic'' quantity: individual companies. We analyze the tick-by-tick data [13] for the 1000 publicly-traded U.S. companies with the largest market capitalizations, and systematically study the statistical properties of their stock price fluctuations. A preliminary study [14] reported that the distribution of the 5-min returns for 1000 individual companies and the S&P 500 index decays as a power-law with an exponent $\alpha \approx 3$—well outside the stable Lévy regime ($\alpha < 2$). Earlier independent studies on individual stock returns on longer time scales yield similar results [15]. These findings raise the following questions.

First, how does the nature of the distribution of individual stock returns change with increasing time scale Δt? In other words, does the distribution retain its power-law functional form for longer time scales, or does it converge to a Gaussian, as found for market indices [7,16]? If the distribution indeed converges to Gaussian behavior, how fast does this convergence occur? For the S&P 500 index, for example, one finds the distribution of returns to be consistent with a *nonstable* power-law functional form ($\alpha \approx 3$) for approximately 4 day, after which an onset of convergence to Gaussian behavior is found [16].

Second, why is it that the distribution of returns for individual companies and for the S&P 500 index have the same asymptotic form? This finding is unexpected, since the returns of the S&P 500 are the weighted sums of the returns of 500 companies. Hence, we would expect, the S&P 500 returns to be distributed approximately as a Gaussian, unless there were significant dependencies between the returns of different companies which prevent the central limit theorem from applying.

To answer the first question, we extend previous work [14] on the distribution of returns for 5-min returns by performing an empirical analysis of individual company returns for time scales up to 46 month. Our analysis uses two distinct databases detailed below. We find that the cumulative distribution of individual-company returns is consistent with a power-law asymptotic behavior with an exponent $\alpha \approx 3$, which is outside the stable Lévy regime. We also find that these distributions appear to retain the same functional form for time scales up to approximately 16 day. For longer time scales, we observe results consistent with a slow convergence to Gaussian behavior.

To answer the second question, we randomize each of the 500 time series of returns for the constituent 500 stocks of the S&P 500 index. A surrogate ''index return,'' thus constructed from the randomized time series, shows fast conver-

gence to Gaussian behavior. Further, we find that the functional form of the distribution of returns remains unchanged for different system sizes (measured by the market capitalization), while the standard deviation decays as a power law of the market capitalization.

The organization of this paper is as follows. Section II describes the databases studied and the data analyzed. Sections III, IV, and V present results for the distribution of returns for individual companies for a wide range of time scales. Section VI discusses the role of cross-correlations between companies and possible reasons why market indices have statistical properties very similar to those of individual companies. Section VII contains some concluding remarks.

II. DATA ANALYZED

We analyze two different databases covering securities from the three major U.S. stock markets, namely, (i) the New York Stock Exchange (NYSE), (ii) the American Stock Exchange (AMEX), and (iii) the National Association of Securities Dealers Automated Quotation (Nasdaq) stock market. The NYSE is the oldest stock exchange, tracing its origin to the Buttonwood Agreement of 1792 [17]. The NYSE is an agency auction market, that is, trading at the NYSE takes place by open bids and offers by Exchange members, acting as agents for institutions or individual investors. Buy and sell orders are brought to the trading floor, and prices are determined by the interplay of supply and demand. As of the end of November 1998, the NYSE listed over 3100 companies. These companies have over 2×10^{11} shares, worth approximately 10^{13} in U.S. dollars, available for trading on the Exchange.

In contrast to the NYSE, Nasdaq uses computers and telecommunication networks which create an electronic trading system wherein the market participants meet over the computer rather than face to face. Nasdaq's share volume reached 1.6×10^{11} shares in 1997, and dollar volume reached 4.4×10^{12} in U.S. dollars. As of December 1998, the Nasdaq Stock Market listed over 5400 U.S. and non-U.S. companies [18]. Nasdaq and AMEX merged in October 1998, after the end of the period studied in this work.

The first database we consider is the trades and quotes (TAQ) database [19], for which we analyze the 2-yr period January 1994 to December 1995. The TAQ database, which has been published by the NYSE since 1993, covers *all* trades at the three major U.S. stock markets. This huge database is available in the form of CD-ROM's. The rate of publication was one CD-ROM per month for the period studied, but recently has increased to 2–3 CD-ROM's per month. The total number of transactions for the largest 1000 stocks is of the order of 10^9 in the 2-yr period studied.

The second database we analyze is the Center for Research and Security Prices (CRSP) database [20]. The CRSP stock files cover common stocks listed on the NYSE beginning in 1925, on the AMEX beginning in 1962, and on the Nasdaq Stock Market beginning in 1972. The files provide complete historical descriptive information and market data including comprehensive distribution information; high, low, and closing prices; trading volumes; shares outstanding; and total returns [21].

The CRSP stock files provide monthly data for the NYSE beginning in December 1925 and daily data beginning in July 1962. For the AMEX, both monthly and daily data begin in July 1962. For the Nasdaq stock market, both monthly and daily data begin in July 1972.

We also analyze the S&P 500 index, which comprises 500 companies chosen for market size, liquidity, and industry group representation in the U.S. In our study, we analyze data with a recording frequency of less than 1 min that cover the 13 yr from January 1984 to December 1996. The total number of data points in this 13-yr period exceeds 4.5×10^6.

III. DISTRIBUTION OF RETURNS FOR $\Delta T<1$ DAY

The basic quantity studied for individual companies—$i=1,2,\ldots,1000$—is the market capitalization $S_i(t)$, defined as the share price multiplied by the number of outstanding shares. The time t runs over the working hours of the stock exchange—removing nights, weekends, and holidays [22]. For each company, we analyze the return

$$G_i \equiv G_i(t,\Delta t) \equiv \ln S_i(t+\Delta t) - \ln S_i(t). \tag{1}$$

For small changes in $S_i(t)$, the return $G_i(t,\Delta t)$ is approximately the forward relative change

$$G_i(t,\Delta t) \approx \frac{S_i(t+\Delta t)-S_i(t)}{S_i(t)}. \tag{2}$$

For time scales shorter than 1 day, we analyze the data from the TAQ database. We consider the largest 1000 companies [23], in decreasing order of values of their market capitalization on the first trading day, 3 January 1994. We sample the price of these 1000 companies at 5-min intervals [24]. In order to obtain time series for market capitalization, we multiply the stock price of each company by the number of outstanding shares for that company at each sampling time. We thereby generate a time series, sampled at 5-min intervals, for the market capitalizations of each of the largest 1000 companies. Each of the 1000 time series has approximately 40 000 data points—corresponding to the number of 5-min intervals in the 2-yr period—or about 40 million data points in total. For each time series of market capitalizations, we compute the 5-min returns using Eq. (1). We filter the data to remove spurious events, such as occur due to the inevitable recording errors [25].

A. Distribution of returns for $\Delta t=5$ min

Figure 1(a) shows the cumulative distributions of returns G_i for $\Delta t=5$ min—the probability of a return larger than or equal to a threshold—for ten individual companies randomly selected from the 1000 companies that we analyze. For each company i, the asymptotic behavior of the functional form of the cumulative distribution is ''visually'' consistent with a power law,

$$P(G_i>x) \sim \frac{1}{x^{\alpha_i}}, \tag{3}$$

where α_i is the exponent characterizing the power-law decay. In Fig. 1(b) we show the histogram for α_i, obtained

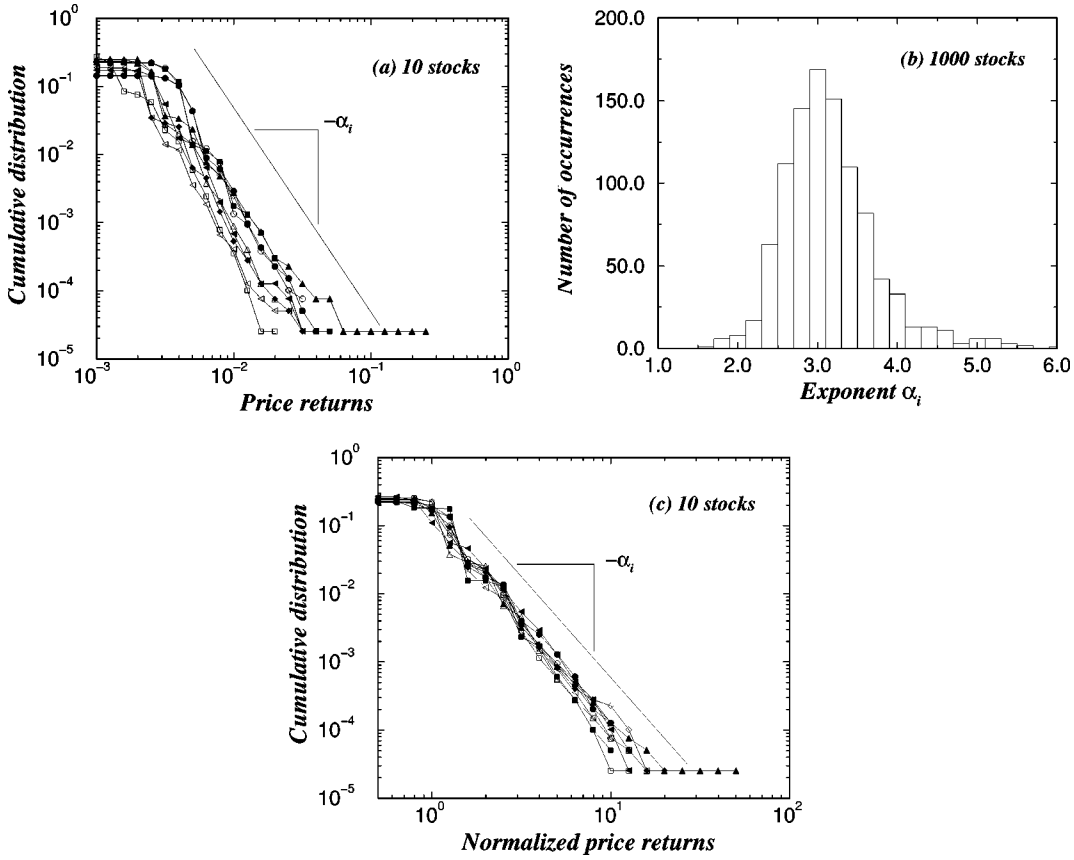

FIG. 1. (a) Cumulative distributions $P(g>x)$ for the positive tails of ten randomly selected companies. Note that they are all consistent with a power-law asymptotic behavior. (b) The histogram of the power-law exponents obtained by power-law regression fits to the individual cumulative distribution functions, where the fit is for all x larger than two standard deviations. Note that this histogram is not normalized—the y axis indicates the number of occurrences of the exponent. (c) Cumulative distributions of the ten randomly chosen companies in (a) scaled by the standard deviation calculated from the entire 2-yr period.

from power-law regression fits to the positive tails of the individual cumulative distributions of all 1000 companies studied. The histogram has the most probable value $\alpha_{MP} \approx 3$.

Next we compute the time-averaged volatility $v_i \equiv v_i(\Delta t)$ of company i as the standard deviation of the returns over the 2-yr period

$$v_i^2 \equiv \langle G_i^2 \rangle_T - \langle G_i \rangle_T^2, \tag{4}$$

where $\langle \cdots \rangle_T$ denotes a time average over the 40 000 data points of each time series, for the 2-yr period studied. Figure 1(a) suggests that the widths of the individual distributions differ for different companies; indeed, companies with small values of market capitalization are likely to fluctuate more. In order to compare the returns of different companies with different volatilities, we define the normalized return $g_i \equiv g_i(t,\Delta t)$ as

$$g_i \equiv \frac{G_i - \langle G_i \rangle_T}{v_i}. \tag{5}$$

Figure 1(c) shows the ten cumulative distributions of the normalized returns g_i for the same ten companies as in Fig 1(a). The distributions for all 1000 normalized returns g_i have similar functional forms to these ten. Hence, to obtain better statistics, we compute a *single* distribution of all the normalized returns. The cumulative distribution $P(g>x)$ shows a power-law decay [Fig. 2(a)]

$$P(g>x) \sim \frac{1}{x^{\alpha}}. \tag{6}$$

Regression fits in the region $2 \leqslant g \leqslant 80$ yield

$$\alpha = \begin{cases} 3.10 \pm 0.03 & \text{(positive tail)} \\ 2.84 \pm 0.12 & \text{(negative tail).} \end{cases} \tag{7}$$

These estimates [26] of the exponent α are well outside the stable Lévy range, which requires $0<\alpha<2$.

In order to obtain an alternative estimate for α, we use the methods of Refs. [12,14–16,27]. We first calculate the inverse of the local logarithmic slope of $P(g)$, $\zeta^{-1}(g) \equiv d \ln P(g)/d \ln g$, where g is rank ordered. We then estimate the asymptotic slope α by extrapolating ζ as a function of $1/g \to 0$. Figure 3 shows the results for the negative and positive tails, for the 5-min returns for individual companies, each using all returns larger than five standard deviations. Extrapolation of the linear regression lines yields

$$\alpha = \begin{cases} 2.84 \pm 0.12 & \text{(positive tail)} \\ 2.73 \pm 0.13 & \text{(negative tail).} \end{cases} \tag{8}$$

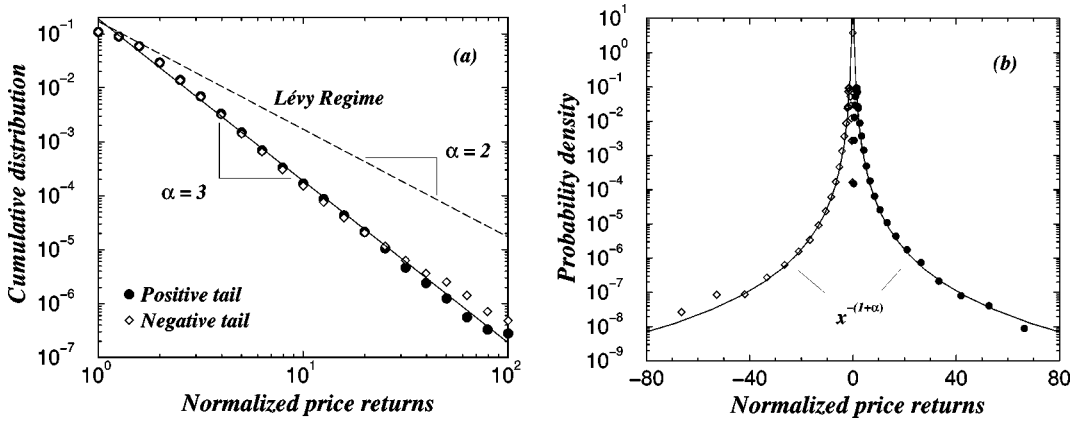

FIG. 2. (a) Cumulative distributions of the positive and negative tails of the normalized returns of the 1000 largest companies in the TAQ database for the 2-yr period 1994–1995. The solid line is a power-law regression fit in the region $2 \leqslant x \leqslant 80$. (b) Probability density function of the normalized returns. The values in the center of the distribution arise from the discreteness in stock prices, which are set in units of fractions of U.S. dollars, usually 1/8, 1/16, or 1/32. The solid curve is a power-law fit in the region $2 \leqslant x \leqslant 80$. We find $\alpha = 3.10 \pm 0.03$ for the positive tail, and $\alpha = 2.84 \pm 0.12$ for the negative tail.

B. Scaling of the distribution of returns for $\Delta t \leqslant 1$ day

The next logical step would be to extend the previous procedure to time scales longer than 5 min. However, this approach leads to unreliable results, the reason being that the estimate of the time averaged volatility—used to define the normalized returns of Eq. (5)—has estimation errors that increase with Δt. For the distribution of 5-min returns, the previous procedure relies on 40 000 data points per company for the estimation of the time averaged volatility. For 500-min returns the number of data points available is reduced to 400 per company which leads to a much larger error in the estimate of $v_i(\Delta t)$.

To circumvent the difficulty arising from the large uncertainty in $v_i(\Delta t)$, we use an alternative procedure for estimating the volatility [28–31] which relies on two observations. The first is that volatility decreases with market capitalization (Fig. 4). The second is that companies with similar market capitalization typically have similar volatilities. Based on these observations, we make the hypothesis that the market capitalization is an influential factor in determining the volatility:

$$v_i = v_i(S, \Delta t). \tag{9}$$

Hence we group the returns of all the companies into ''bins'' according to the market capitalization of each company at the beginning of the interval for which the return is computed. We then compute the conditional probability of the Δt returns for each of the bins of market capitalization. We define $G_S \equiv G_S(t, \Delta t)$ as the Δt returns of the subset of all companies with market capitalization S, and we then calculate the cumulative conditional probability $P(G_S \geqslant x | S)$. Figure 5(a) shows $P(G_S \geqslant x | S)$ for 30-min returns for four different bins of S. The functional form for each of the four distributions is consistent with a power-law.

We define a normalized return

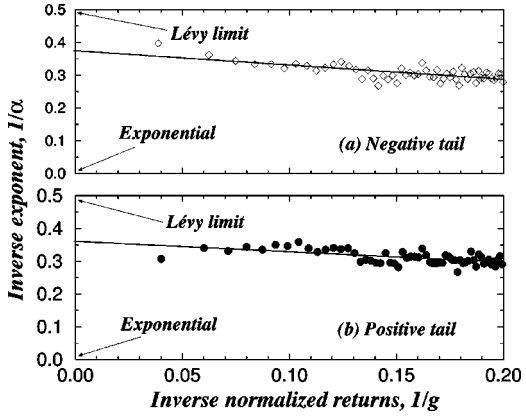

FIG. 3. The inverse local slope of $P(g)$, $\zeta^{-1}(g) \equiv -[d \ln P(g)/d \ln g]$ as a function of the inverse normalized returns $1/g$ for (a) the negative tail and (b) the positive tail [16,27]. Each data point shown is an average over 1000 events and the lines are linear regression fits to the data. The linear regression fit over the range $0 \leqslant g \leqslant 0.2$ yields the values of the inverse asymptotic slopes, $1/\alpha$; we find $\alpha = 2.84 \pm 0.12$ for the positive tail and $\alpha = 2.73 \pm 0.13$ for the negative tail. Note that the average over all events used would be identical to the estimator for the asymptotic slope proposed by Hill [27].

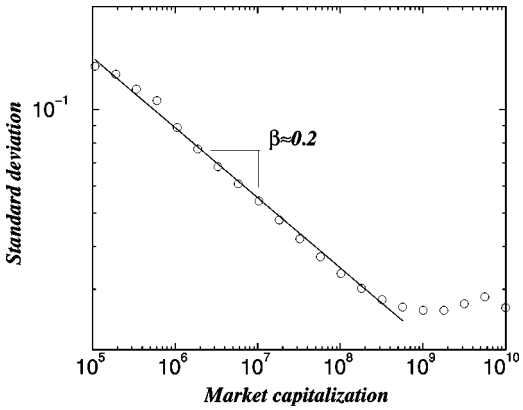

FIG. 4. Log-log plot of the standard deviation of the distribution of returns as a function of market capitalization for $\Delta t = 1$ day. Our preliminary data suggest a power-law dependence with exponent $\beta \approx 0.2$. This value is not unlike what was observed for the firm sales ($\beta \approx 1/6$) [28], GDP of countries ($\beta \approx 1/6$) [29], and research budgets ($\beta \approx 1/4$) [30]. For large values of market capitalization, this power law is followed by a ''flat'' region.

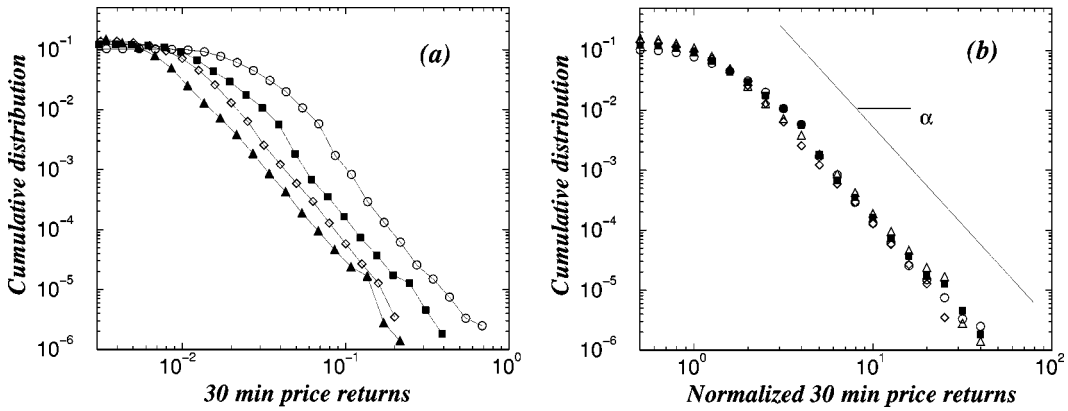

FIG. 5. (a) Cumulative distribution of the conditional probability $P(g>x|S)$ of the 30-min returns, for companies with market capitalization S, from the TAQ database. We define uniformly spaced bins on a logarithmic scale. We show the distribution of returns for the four bins, $10^{9.8}<S\leqslant 10^{10.2}$, $10^{10.2}<S\leqslant 10^{10.4}$, $10^{10.4}<S\leqslant 10^{10.6}$, and $10^{10.6}<S\leqslant 10^{10.8}$. (b) Cumulative conditional distributions of returns normalized by the average volatility $v_S(\Delta t)$ of each bin. Note that we find the same functional form for the different values of S.

$$g_S \equiv g_S(t,\Delta t) \equiv \frac{G_S(\Delta t) - \langle G_S(\Delta t)\rangle_S}{v_S(\Delta t)}, \tag{10}$$

where $\langle \cdots \rangle_S$ denotes an average over all returns of all companies with market capitalization S. The average volatility $v_S \equiv v_S(\Delta t)$ is defined through the relation

$$v_S^2 \equiv \langle G_S^2\rangle_S - \langle G_S\rangle_S^2. \tag{11}$$

In Fig. 5(b) we show the cumulative conditional probability of the normalized 30-min returns $P(g_S \geqslant x|S)$ for the same four bins shown in Fig. 5(a). Visually, it seems clear that these distributions have power-law functional forms with similar values of α. Hence, to obtain better statistics,

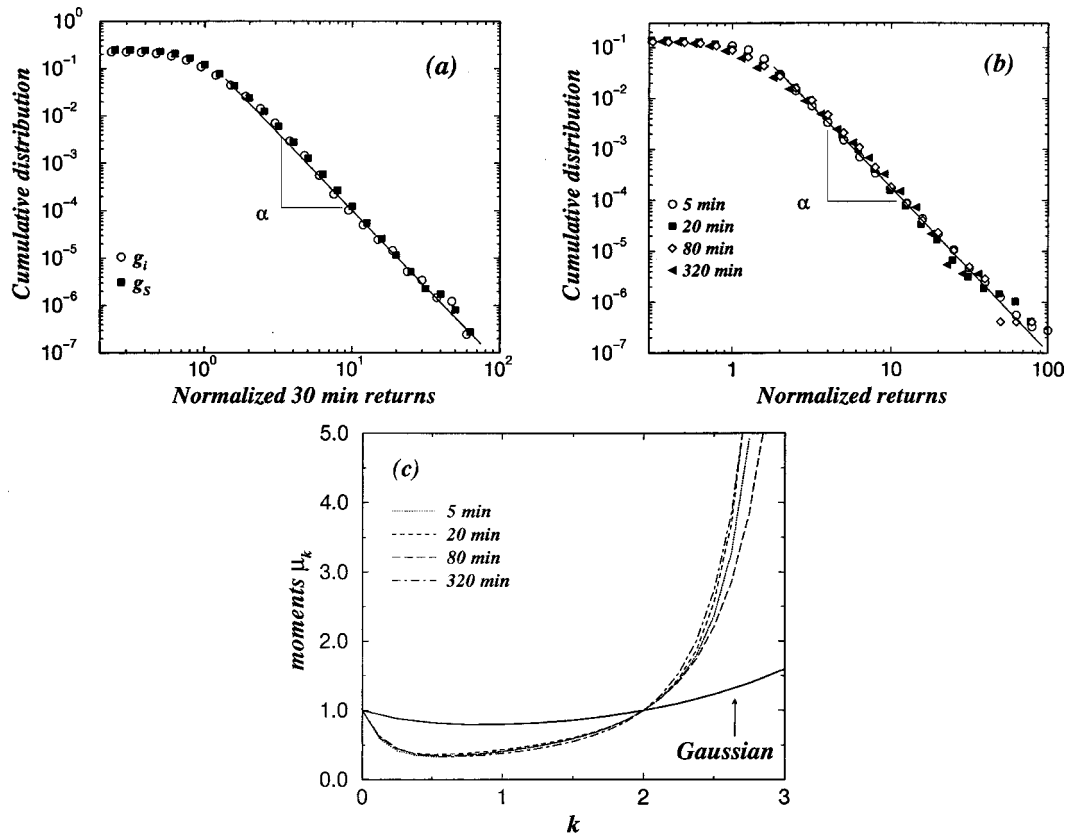

FIG. 6. (a) Cumulative distribution of normalized returns for $\Delta t = 30$ min. The filled squares show the distribution for returns normalized by the time-averaged volatility for each company, as defined in Eq. (5). The circles show the distribution for returns normalized by the average volatility for each size bin [Eq. (10)], showing the consistency of these two methods. (b) The distribution of returns for different time scales $\Delta t \leqslant 1$ day. The exponents from the power-law regression fits are summarized in Table I. (c) Fractional moments from $0 \leqslant k < 3$ for the normalized returns for the same scales as in (b). Note that the moments are not converging to Gaussian behavior, for example, at large k the moments for $\Delta t = 80$ min is to the right of $\Delta t = 320$ min. The thick full line shows the Gaussian moments.

TABLE I. The values of the exponent α for different time scales Δt obtained by (a) a power-law regression fit to the cumulative distribution and (b) the Hill estimator. The non-daggered values are computed using the TAQ database, which contains tick data, while the daggered values are computed using the CRSP database, which contains records with $\Delta t = 1$ day and $\Delta t = 1$ month sampling. Note that we use the conversion 1 day = 390 min and 1 month = 22 day.

Δt (min)	Power-law fit		Hill estimator	
	Positive	Negative	Positive	Negative
5	3.10 ± 0.03	2.84 ± 0.12	2.84 ± 0.12	2.73 ± 0.13
10	3.32 ± 0.08	2.89 ± 0.13	3.14 ± 0.10	2.68 ± 0.14
20	3.25 ± 0.08	2.75 ± 0.10	3.32 ± 0.18	2.41 ± 0.10
40	3.28 ± 0.08	2.61 ± 0.10	3.39 ± 0.16	2.62 ± 0.11
80	3.50 ± 0.13	2.49 ± 0.11	3.65 ± 0.26	2.53 ± 0.14
160	3.47 ± 0.08	2.42 ± 0.09	2.9 ± 0.4	2.53 ± 0.17
320	3.60 ± 0.10	2.54 ± 0.10	3.32 ± 0.08	3.19 ± 0.05
390†	2.96 ± 0.09	2.70 ± 0.10	3.05 ± 0.13	2.95 ± 0.15
780†	3.09 ± 0.03	2.62 ± 0.04	3.11 ± 0.09	2.90 ± 0.12
1560†	3.18 ± 0.05	2.75 ± 0.09	3.20 ± 0.08	2.90 ± 0.10
3120†	3.31 ± 0.08	2.71 ± 0.03	3.25 ± 0.06	2.94 ± 0.09
6240†	3.43 ± 0.04	2.74 ± 0.12	3.35 ± 0.04	2.93 ± 0.07
12 480†	3.73 ± 0.04	2.63 ± 0.06	3.54 ± 0.05	2.93 ± 0.08
24 960†	3.98 ± 0.09	2.78 ± 0.07	3.89 ± 0.09	3.00 ± 0.10
49 920†	4.24 ± 0.09	2.84 ± 0.07	4.52 ± 0.22	3.10 ± 0.18
99 840†	5.06 ± 0.07	3.01 ± 0.07	4.5 ± 0.6	2.92 ± 0.19
199 680†	5.24 ± 0.12	3.32 ± 0.06	5.6 ± 1.0	3.14 ± 0.13
399 360†	6.43 ± 0.29	3.48 ± 0.07	5.11 ± 0.03	3.45 ± 0.02

we consider the normalized returns for all values of S and compute a *single* cumulative distribution.

Figure 6(a) shows the distribution of normalized 30-min returns. We test if our alternative procedure of normalizing the returns by the time averaged volatility for each bin of market capitalization S is consistent with the previous procedure of normalizing by the time averaged volatility for each company through Eq. (5). To this end, in Fig. 6(a), we also show the distribution of normalized 30-min returns using the normalization of Eq. (5). The distribution of returns obtained by both procedures are consistent with a power-law decay of the same form as Eq. (6). Power-law regression fits to the positive tail yield estimates of $\alpha = 3.21 \pm 0.08$ for the former method and $\alpha = 3.23 \pm 0.05$ for the latter, confirming the consistency of the two procedures. The values of the exponent for 30-min time scales, $\alpha = 3.21 \pm 0.08$ (positive tail) and $\alpha = 3.01 \pm 0.12$ (negative tail), are also consistent with the estimates [Eq. (7)] for 5-min normalized returns.

Next we compute the distribution of returns for longer time scales Δt. Figure 6(b) shows the cumulative distribution of the normalized returns for time scales from 5 min up to 1 day. We observe good ''data collapse'' with consistent values of α which suggests that the distribution of returns appears to retain its functional form for larger Δt, beyond a lower bound which is approximately the same for different Δt. Note that we do not find any indication of the distribution converging to Gaussian behavior for small returns. For example, consider the sum of independent, nonstable, power-law distributed random variables. Then the power-law behavior is pushed further out into the tails with an increasing number of variables summed, with Gaussian behavior for low values. However, our results are inconsistent with this possibility, as we do not observe any indication of such convergence to Gaussian behavior. The estimates of the exponent α from power-law regression fits to the cumulative distribution and from the Hill estimator are listed in Table I. Note also that the scaling of the distribution of returns for individual companies is consistent with previous results for the distribution of the S&P 500 index returns [7,16].

C. Scaling of the moments for $\Delta t < 1$ day

In Sec. III B we reported that the distribution of returns retains the same functional form for 5 min $< \Delta t <$ 1 day. We can further test this scaling behavior by analyzing the moments of the distribution of normalized returns g,

$$\mu_k \equiv \langle |g|^k \rangle, \tag{12}$$

where $\langle \cdots \rangle$ denotes an average over all the normalized returns for all the bins. Since $\alpha \approx 3$, we expect μ_k to diverge for $k \geq 3$, and hence we compute μ_k for $k < 3$.

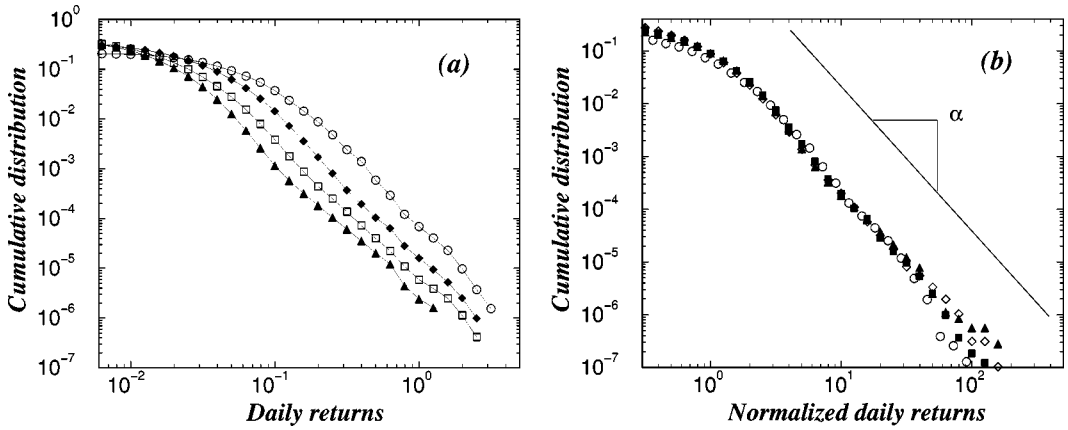

FIG. 7. (a) Cumulative distribution of the conditional probability $P(g > x|S)$ of the returns for companies with starting values of market capitalization S for $\Delta t = 1$ day from the CRSP database. We define uniformly spaced bins on a logarithmic scale and show the distribution of returns for the bins, $10^5 < S \leq 10^6$, $10^6 < S \leq 10^7$, $10^7 < S \leq 10^8$, and $10^8 < S \leq 10^9$. (b) Cumulative conditional distributions of returns normalized by the average volatility $v_S(\Delta t)$ of each bin.

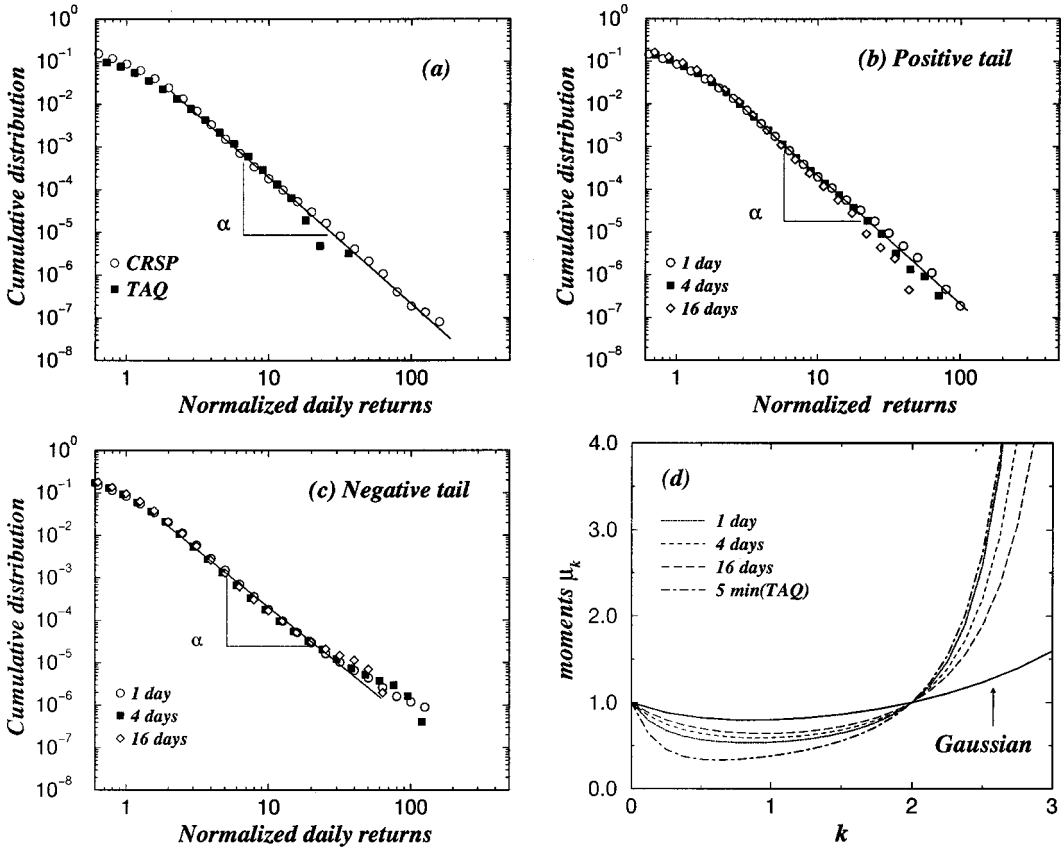

FIG. 8. (a) Cumulative distribution of normalized daily returns computed from the CRSP database contrasted with the same distribution from the TAQ database, normalized by the average volatility. Regression fits yield estimates $\alpha=2.96\pm0.09$ (positive tail), and $\alpha=2.70\pm0.10$ (negative tail) for the CRSP data, and $\alpha=3.27\pm0.19$ (positive tail) and $\alpha=2.98\pm0.21$ (negative tail) for the TAQ data. The regression fits were performed for the region $2\leqslant g\leqslant 80$. (b) Positive and (c) negative tails of the cumulative distribution of normalized returns for $\Delta t=1$,4, and 16 day. Estimates of the exponents are listed in Table I. (d) The fractional moments $\mu_k\equiv\langle|g|^k\rangle$ for the normalized returns for the same time scales. The thick full line shows the Gaussian moments.

Figure 6(c) shows the moments of the normalized returns g for different time scales from 5 min up to 1 day. The moments do not vary significantly for the above time scales, thus confirming the scaling behavior of the distribution observed in Fig 6(b).

IV. DISTRIBUTION OF RETURNS FOR 1 DAY $\leqslant\Delta t\leqslant$16 DAY

For time scales of 1 day or longer, we analyze data from the CRSP database. We analyze approximately 3.5×10^7 daily records for about 16 000 companies for the 35-yr period 1962–96. We expect the market capitalization of a company to change dramatically in such a long period of time. Further, we expect small companies to be more volatile than large companies. Hence large changes that might occur in the market capitalization of a company will lead to large changes in its average volatility. To control these changes in market capitalization, we adopt the method that was used in the previous subsection for $\Delta t>5$ min.

Thus, we compute the cumulative conditional probability $P(G_S\geqslant x|S)$ that the return $G_S\equiv G_S(t,\Delta t)$ is greater than x, for a given bin of average market capitalization S. We first divide the entire range of S into bins of uniform length in logarithmic scale. We then compute a separate probability distribution for the returns G_S which belong to a bin of average market capitalization S.

Figure 7(a) shows the cumulative distribution of daily returns $P(G_S>x|S)$ for different values of S. Since the widths of these distributions are different for different S, we analyze the normalized returns g_S, which were defined in Eq. (10).

Figure 7(b) shows the cumulative distribution $P(g_S>x)$ of the normalized daily returns g_S. These distributions appear to have similar functional forms for different values of S. In order to improve statistics, we compute a *single* cumulative distribution $P(g_S>x)$ of the normalized returns for all S. We observe a power-law behavior of the same form as Eq. (6). Regression fits yield estimates for the exponent, $\alpha=2.96\pm0.09$ for the positive tail and $\alpha=2.70\pm0.10$ for the negative tail.

Figure 8(a) compares the cumulative distributions of the normalized 1-day returns obtained from the CRSP and TAQ databases. The estimates of the power-law exponents obtained from regression fits are in good agreement for these two databases.

Figures 8(b) and 8(c) show the distributions of normalized returns for $\Delta t=1$, 4, and 16 day. The estimates of the exponent α increase slightly in value for the positive tail, while for the negative tail the estimates of α are approximately

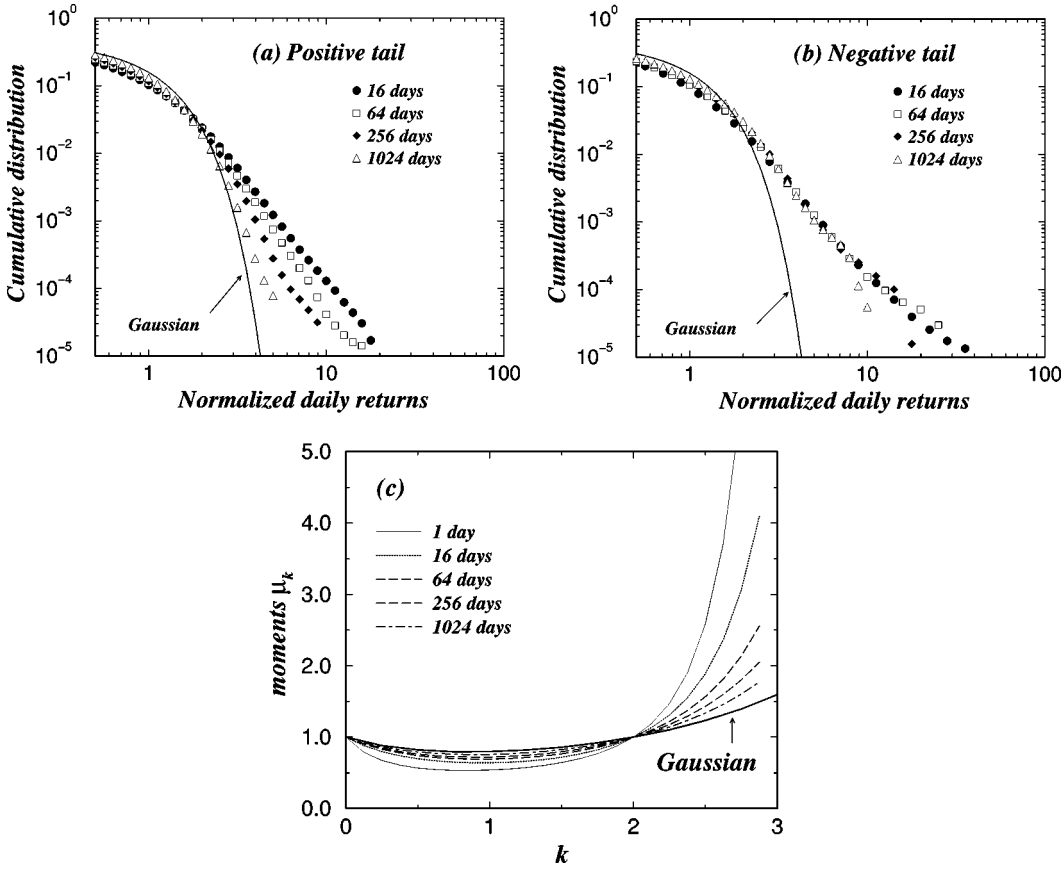

FIG. 9. (a) Positive and (b) negative tails of the cumulative distribution of the normalized returns for $\Delta t = 16,\ 64,\ 256$, and 1024 day. The positive tail shows clear indication of convergence to Gaussian behavior, whereas for the negative tail the power-law behavior still seems to hold, although the statistics at the tail are limited for the longer time scales. Estimates of the exponents are listed in Table I. (c) The fractional moments μ_k, $0 \leqslant k < 3$, of the normalized returns for $\Delta t = 16,\ 64,\ 256$, and 1024 day show clear indication of convergence to Gaussian behavior with increasing Δt.

constant. The increase in α for the positive tail is also reflected in the moments [Fig. 8(d)].

V. DISTRIBUTION OF RETURNS FOR $\Delta t \geqslant 16$ DAY

The scaling behavior of the distributions of returns appears to break down for $\Delta t \geqslant 16$ day, and we observe indications of slow convergence to Gaussian behavior. In Figs. 9(a) and 9(b) we show the cumulative distributions of the normalized returns for $\Delta t \geqslant 16$ day. For the positive tail, we find indications of convergence to a Gaussian, while the negative tail appears not to converge. The convergence to Gaussian behavior is also apparent from the behavior of the moments for these time scales [Fig. 9(c)].

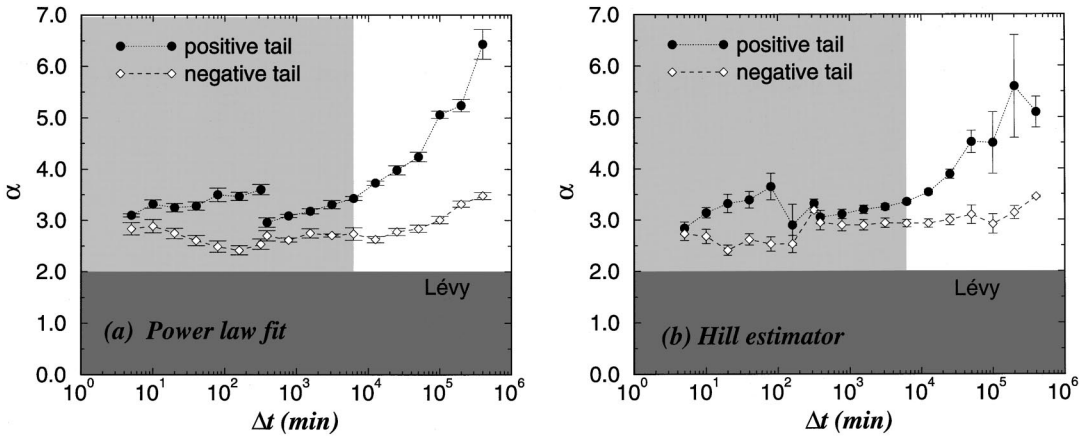

FIG. 10. The values of the exponent α characterizing the asymptotic power-law behavior of the distribution of returns as a function of the time scale Δt obtained using (a) a power-law fit, and (b) the Hill estimator. The values of α for $\Delta t < 1$ day are calculated from the TAQ database, while for $\Delta t \geqslant 1$ day they are calculated from the CRSP database. The unshaded region, corresponding to time scales larger than $(\Delta t)_\times \approx 16$ day (6240 min), indicates the range of time scales where we find results consistent with slow convergence to Gaussian behavior.

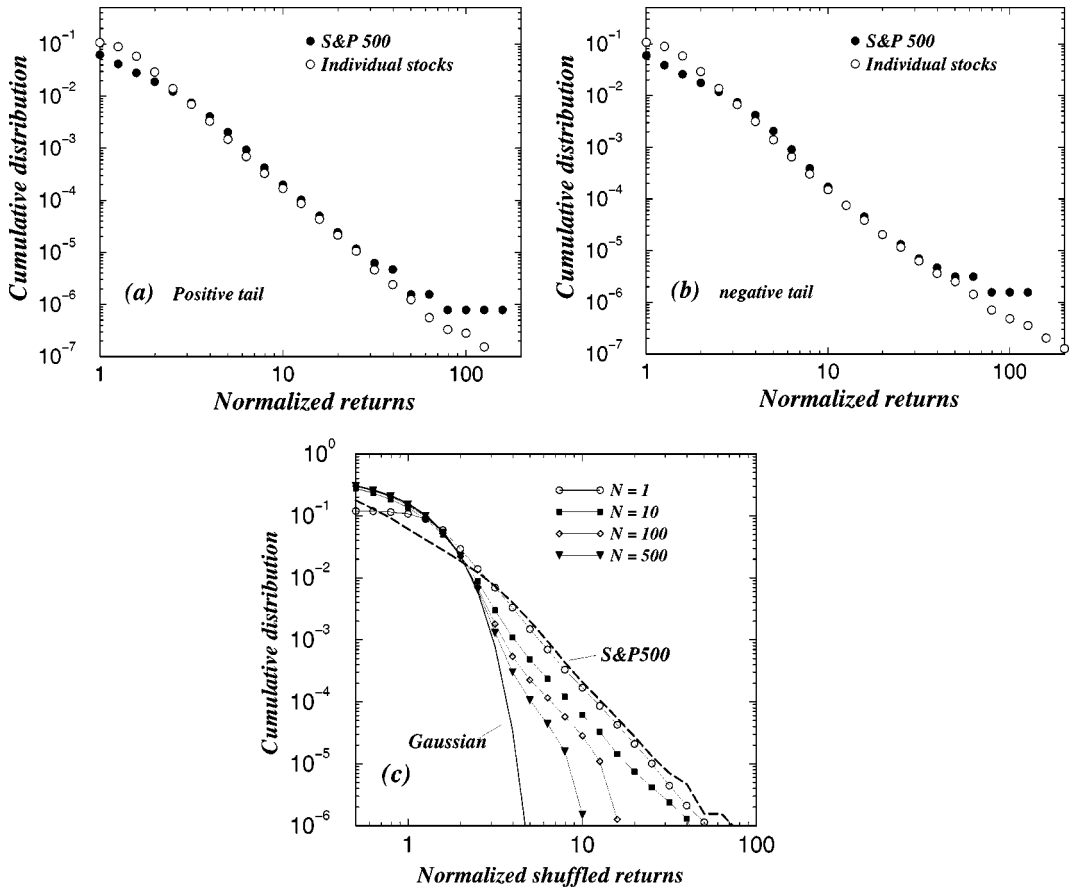

FIG. 11. (a) Positive and (b) negative tails of the cumulative distribution for the normalized returns for the individual companies and the S&P 500 index. Both the distributions show the same functional form, in spite of being a nonstable law. (c) Cumulative distribution for the shuffled returns $\tilde{g}^{(N)}(t,\Delta t)$ for $N=1,\ 10,\ 100,$ and 500. The dotted curve is the cumulative distribution for the S&P 500. With increasing N the curves progressively approach a Gaussian, implying that without the cross-dependencies between companies, the cumulative distribution for the S&P 500 would be almost Gaussian.

To summarize our results for the distribution of individual company returns, we find that (i) the distribution of normalized returns for individual companies is consistent with a power-law behavior characterized by an exponent $\alpha\approx 3$; (ii) the distributions of returns retain the same functional form for a wide range of time scales Δt, varying over three orders of magnitude, 5 min $\leqslant\Delta t\leqslant$6240 min = 16 day; and (iii) for $\Delta t>16$ day; the distribution of returns appears to slowly converge to a Gaussian [Fig. 10].

VI. CROSS-CORRELATIONS

In this section we address the second question that we posed initially. That is, why is it that the distribution of returns for individual companies and for the S&P 500 index have the same asymptotic form? In the previous sections, we presented evidence that the distribution of returns scales for a wide range of time intervals. In a previous study [16], we demonstrated that this scaling behavior is possibly due to time dependencies, in particular, volatility correlations. Next, we will show that as the time correlations lead to the time scaling of the distributions of returns, so do cross-correlations among different companies lead to a functional form of the distribution of returns of indices similar to that for single companies.

A direct way of analyzing the cross-correlations is by computing the cross-correlation matrix [32–34]. Here we take a different approach, by analyzing the distribution of returns as a function of market capitalization.

First we compare the distributions of the S&P 500 index and that of individual companies. Figures 11(a) and 11(b) show the cumulative distribution $P(g\geqslant x)$ for individual companies and for the S&P 500 index. The distributions show the same power-law behavior for $3\leqslant g\leqslant 80$. This is surprising, because the distribution of index returns $G_{SP500}(t,\Delta t)$ does not show convergence to Gaussian behavior—even though the 500 distributions of individual returns $G_i(t,\Delta t)$ are not stable. Consider the family of index returns defined as the partial sum [35]

$$G_{(N)}(t,\Delta t)\equiv\sum_{i=1}^{N} w_i\, G_i(t,\Delta t), \tag{13}$$

where the weights $w_i\equiv S_i/\Sigma_{j=1}^{N}S_j$ have weak time dependencies [36]. From the central limit theorem for random variables with finite variance, we expect that the probability distribution of $G_{(N)}$ would change systematically with N and approach a Gaussian for large N, provided there are no significant dependencies among the returns G_i for different i. Instead, we find that the distribution of $G_{(N)}$ has the same asymptotic behavior as that for individual companies.

In order to show that the scaling behavior may be due to cross-correlations between companies, we first destroy any existing dependencies among the returns of different companies by randomizing each of the 1000 time series $G_i(t)$. By adding up the shuffled series, we construct a shuffled index return $G_{(N)}^{sh}(t)$ out of statistically independent companies with the same distribution of returns. Figure 11(c) shows the cumulative distribution of the shuffled index returns $G_{(N)}^{sh}(t,\Delta t)$ for increasing N and $\Delta t = 5$ min. The distribution changes with N, and approaches a Gaussian shape for large N, which indicates that the scaling in Fig. 11(a) is caused by nontrivial dependencies between different companies.

VII. DISCUSSION

We have presented a systematic analysis, on two different databases, of the distribution of returns for individual companies for time scales Δt ranging from 5 min up to ≈ 4 yr. We find that the distribution of returns is consistent with a power-law asymptotic behavior, characterized by an exponent $\alpha \approx 3$—well outside the stable Lévy regime $0 < \alpha < 2$—for time scales up to approximately 16 day. For longer time scales, the scaling behavior appears to break down and we observe ''slow'' convergence to Gaussian behavior.

We also find that the distribution of returns of individual companies and the S&P 500 index have the same asymptotic behavior. This scaling behavior does not hold when the cross-correlations between companies are destroyed, suggesting the existence of correlations between companies —as occurs in strongly interacting physical systems where power-law correlations at the critical point result in scale-invariant properties [37]. Recent studies of the cross-correlation matrix using methods of random matrix theory [32–34] also show the existence of correlations that are present through a wide range of time scales from 30 min [34] up to 1 day [32,33]. These studies [32–34] show that the largest eigenvalue of the cross-correlation matrix corresponds to correlations that pervade the entire market, and a few other large eigenvalues correspond to clusters of companies that are correlated amongst each other.

ACKNOWLEDGMENTS

We thank J.-P. Bouchaud, M. Barthélemy, S. V. Buldyrev, P. Cizeau, X. Gabaix, I. Grosse, S. Havlin, K. Illinski, C. King, C.-K. Peng, B. Rosenow, D. Sornette, D. Stauffer, S. Solomon, J. Voit, and especially R. N. Mantegna for stimulating discussions and helpful suggestions. We thank X. Gabaix, C. King, J. Stein, and especially T. Lim for help with obtaining the data. We are also very grateful to L. Giannitrapani of the SCV at Boston University for her generous help in allocating the necessary computer resources, and to R. Tomposki for his help throughout this work. M.M. thanks DFG, and L.A.N.A. thanks FCT/Portugal for financial support. The Center for Polymer Studies is supported by the NSF.

APPENDIX: DEPENDENCE OF VOLATILITY ON SIZE

We find that the average volatility for each bin, $v_S(\Delta t)$ shows an interesting dependence on the market capitalization. In Fig. 4, we plot the standard deviation as a function of size on a log-log scale for $\Delta t = 1$ day. We find a power-law dependence of the standard deviation of the returns on the market capitalization, with exponent $\beta \approx 0.2$ very similar to the values reported for the annual sales of firms [28–31], the Gross Domestic Product of countries [29] and the university research budgets [30]. For larger time scales the exponent gradually decreases, approaching the value $\beta \approx 0.09$ for $\Delta t = 1000$ day.

[1] J. P. Bouchaud and M. Potters, *Theorie des Risques Financiéres*, (Alea-Saclay, Eyrolles, 1997).
[2] R. N. Mantegna and H. E. Stanley, *An Introduction to Econophysics: Correlations and Complexity in Finance* (Cambridge University Press, Cambridge, 1999).
[3] *Econophysics: An Emerging Science*, edited by I. Kondor and J. Kertész (Kluwer, Dordrecht, 1999).
[4] *Proceedings of the International Workshop on Econophysics and Statistical Finance*, edited by R. N. Mantegna, special issue of Physica A **269**, 1 (1999).
[5] *Application of Physics in Financial Analysis*, edited by J.-P. Bouchaud, P. Alström, and K. B. Lauritsen, special issue of Int. J. Theor. Appl. Finance (to be published).
[6] B. B. Mandelbrot, J. Business **36**, 294 (1963).
[7] R. N. Mantegna and H. E. Stanley, Nature (London) **376**, 46 (1995).
[8] S. Ghashghaie, W. Breymann, J. Peinke, P. Talkner, and Y. Dodge, Nature (London) **381**, 767 (1996); see also R. N. Mantegna and H. E. Stanley, *ibid.* **383**, 587 (1996); Physica A **239**, 255 (1997).
[9] A. Arneodo, J.-F. Muzy, and D. Sornette, Eur. Phys. J. B **2**, 277 (1998).
[10] N. Vandewalle and M. Ausloos, Int. J. Mod. Phys. C **9**, 711 (1998); Eur. Phys. J. B **4**, 257 (1998).
[11] E. Egenter, T. Lux, and D. Stauffer, Physica A **268**, 250 (1999); D. Chowdhury and D. Stauffer, Eur. Phys. J. B **8**, 477 (1999); I. Chang and D. Stauffer, Physica A **264**, 1 (1999); D. Stauffer and T. J. P. Penna, *ibid.* **256**, 284 (1998).
[12] A. Pagan, J. Empirical Finance **3**, 15 (1996).
[13] By ''tick-by-tick data'' we refer to data for *every* transaction.
[14] P. Gopikrishnan, M. Meyer, L. A. N. Amaral, and H. E. Stanley, Eur. Phys. J. B **3**, 139 (1998).
[15] T. Lux, Appl. Financial Economics **6**, 463 (1996); M. Loretan and P. C. B. Phillips, J. Empirical Finance **1**, 211 (1994); J. Kahler (unpublished).
[16] P. Gopikrishnan, V. Plerou, L. A. N. Amaral, M. Meyer, and H. E. Stanley, Phys. Rev. E **60**, 5305 (1999).
[17] Details can be found at http://www.nyse.com.
[18] Details can be found at http://www.nasdaq.com.
[19] Details can be found at http://www.nyse.com/public/search/07ix.htm.
[20] Details can be found at http://www.crsp.com.
[21] The CRSP links all former and current company identifiers to a unique permanent CRSP identifier allowing uninterrupted time-series analysis.

[22] The New York Stock Exchange is open from Monday through Friday 9:30 a.m. to 4:00 p.m. The time runs over working hours only. Nights, weekends, and holidays are removed. Overnight returns are not removed. Removing overnight returns yields a slightly larger value for $\propto$.

[23] Only the companies that existed throughout the 2-yr period 1994–95 were considered.

[24] The trading frequency increases, on average, with market capitalization. For the largest companies there are several trades that occur within each 5-min interval. On the other hand, for the smallest companies we consider, the typical time between trades is of the order of 15 min.

[25] The analyzed data are affected by several types of recording errors. The most common error is missing digits, which appear as a large spike in the time series of returns. These are much larger than usual fluctuations, and can be removed by choosing an appropriate threshold. We tested a range of thresholds and find no effect on the results. Additionally we checked *individually* that the removed events correspond to missing digits in entering the data. There are also stock splits and takeovers which always occur overnight. To account for these, we take to be zero *all* the returns that happen overnight that are merely due to change in the number of outstanding shares.

[26] The errors on the exponent estimates are the errors given by the regression fits to the cumulative distribution. A more realistic error bar is the width of the histogram of exponents from Fig. 1(b).

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

[28] M. H. R. Stanley, L. A. N. Amaral, S. V. Buldyrev, S. Havlin, H. Leschhorn, P. Maass, M. A. Salinger, and H. E. Stanley, Nature (London) **379**, 804 (1996).

[29] Y. Lee, L. A. N. Amaral, D. Canning, M. Meyer, and H. E. Stanley, Phys. Rev. Lett. **81**, 3275 (1998).

[30] V. Plerou, L. A. N. Amaral, M. Meyer, P. Gopikrishnan, and H. E. Stanley, Nature (London) **400**, 433 (1999).

[31] L. A. N. Amaral, S. V. Buldyrev, S. Havlin, M. A. Salinger, and H. E. Stanley, Phys. Rev. Lett. **80**, 1385 (1998).

[32] S. Galluccio, J.-P. Bouchaud, and M. Potters, Physica A **259**, 449 (1998).

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

[34] V. Plerou, P. Gopikrishnan, B. Rosenow, L. A. N. Amaral, and H. E. Stanley, Phys. Rev. Lett. **83**, 1471 (1999).

[35] $S_{N=100}(t)$ and $S_{N=500}(t)$ are not exactly identical to the S&P100 and S&P 500 indices, because the latter sums the market values of the companies representing major industries at time t, which are not necessarily the largest, while the former sums over a *fixed* set of companies which are the largest in market values on January 3, 1994. However, the difference between the two is negligible for the period studied.

[36] If the weighted sum $G_{SP500}(t,\Delta t)$, in the presence of the weights $w_i \propto S_i$, were to be dominated by just a few of the companies—the ones with the largest S_i—then the collapse would be trivial. To show that this is not so, we compute the cumulative distribution for the returns of $X(t,\Delta t) \equiv \Sigma_{i=1}^{N} G_i(t,\Delta t)$, which we find to coincide with the cumulative distribution for $G_{SP500}(t,\Delta t)$.

[37] H. E. Stanley, *Introduction to Phase Transitions and Critical Phenomena* (Oxford University Press, Oxford, 1971).