Special Proceedings: {ISBN #: 978-81-950383-0-5}
23rd Annual Conference, 24–28 February 2021; pp 185–199

Inferring the "Laws" of Finance from High-frequency Data

R K Singh[1,2,3] **and Sitabhra Sinha**[1,2]
[1] *The Institute of Mathematical Sciences, CIT Campus, Taramani, Chennai 600113, India*
[2] *Homi Bhaba National Institute, Anushaktinagar, Mumbai 400 094, India*
[3] *Department of Physics, Bar-Ilan University, Ramat-Gan 5290002, Israel*

Received: 16 August 2021; Revised: 31 August 2021; Accepted: 04 September 2021

Abstract

Financial markets represent a prototypical example of complex systems in which a large number of heterogeneous agents are involved in mutual interactions (namely, trading assets with each other) each having a specific goal (maximizing their own profit). However, despite the unpredictability inherent in the individual components, the system as a whole can exhibit universal features that are invariant across different markets, asset classes and period of observation. One of the most prominent examples of such "stylized facts" is the appearance of fat tails in distributions of price fluctuations, often referred to as the *inverse-cubic law*. However, as such features have mostly been reported in studies using low-resolution data, we look for them using high-frequency data of equities trading in the National Stock Exchange of India, one of the world's leading financial markets. We find that the distribution of trade sizes (the number of stock units involved in a single transaction) possess heavy tails, decaying as a power law with a characteristic exponent. Moreover, the distribution is in general stationary for the market as a whole, even though those of individual stocks may differ significantly from one period to the next. We also investigate the distribution of waiting times between successive trades and find it to be decaying slower than exponential in the case of individual stocks. We relate this to the frequent occurrences in succession of transactions involving large returns (price changes). The correlation between the intervals separating successive trades and the magnitude of price fluctuations that we observe implies that the distributions of the waiting times and that of step lengths in the walk executed by the price of a financial asset may not be completely independent. We also find that the cumulative volatility of price movements increases linearly with time within a trading day, but with deviation from linearity at the ends. This suggests that the non-Gaussian character of the return distribution (reflected in the inverse cubic law) arises from the significant volume of end-of-day trading.

Key words: Power laws; Quantitative finance; Price returns; Inverse cubic law; Heavy-tailed distributions; Trade-size distribution.

AMS Subject Classifications: 91B80, 91G15, 62P05
Journal of Economic Literature Classification: C40, G10

Corresponding Author: Sitabhra Sinha
Email: sitabhra@imsc.res.in

1. Introduction

Financial markets are one of the best known examples of complex systems which are characterized by a large number of interacting components and exhibiting nonlinear dynamics that is inherently unpredictable (Sinha *et al.* (2010)). Even though deterministic descriptions of the time evolution of individual components may not lead easily to an understanding of how an assembly of such components will behave, paradoxically the interactions between many constituents may make it possible to obtain well-defined statistical properties of the system as a whole, i.e., the market (Sinha *et al.* (2016)). Indeed, robust statistical features have been reported for the trading dynamics of different stock markets, notable among them being the so-called "inverse cubic law" describing the nature of the distributions of fluctuations in stock prices (and market indices) in developed (Lux (1996)), as well as, developing economies (Pan & Sinha (2007) and Pan & Sinha (2008)). In general, distributions having heavy tails have been reported for price fluctuations seen across many different asset classes, including currency exchange rates (Chakraborty *et al.* (2018) and Chakraborty *et al.* (2020)) and cryptocurrencies such as bitcoin (Dixit *et al.* (2015)).

Given the price p_t of some stock at time t, price fluctuations are characterized in terms of *logarithmic returns* or *log-returns*, *viz.*,

$$r_t = \ln p_t - \ln p_{t-\Delta t}, \tag{1}$$

where Δt is the time interval separating the two prices (Figure 1). The variation in the fluctuations of price of a stock is defined in terms of *volatility*, defined as the variance of log-returns, *viz.*,

$$\sigma_t^2 = \langle r_t^2 \rangle - \langle r_t \rangle^2. \tag{2}$$

If q_i denotes the number of stocks traded in the i-th transaction (also referred to as the *trade size*) then the volume of stocks $V_{\Delta t}$ traded over a time interval $[t, t + \Delta t]$ is defined as

$$V_{t,\Delta t} = \sum_{i=1}^{N_{t,\Delta t}} q_i, \tag{3}$$

where $N_{t,\Delta t}$ is the number of trades that have occurred during the interval $[t, t + \Delta t]$. The above relation implies that quantities characterizing a trading event, viz., trade sizes, trading volumes and number of trades are not independent of each other.

Most early studies of the empirical statistical properties of markets have used daily (or end of day) trade data which does not take into consideration the dynamical behavior of intra-day trading (Vijayraghavan & Sinha (2011)). In recent times, the availability of high-frequency (HF) data containing information about every transaction taking place in the market has made it possible to uncover the properties of financial markets at the highest possible resolution (Dacorogna *et al.* (2001)). In general, empirical analysis using such data can reveal features that are not possible to observe using data collected at temporal

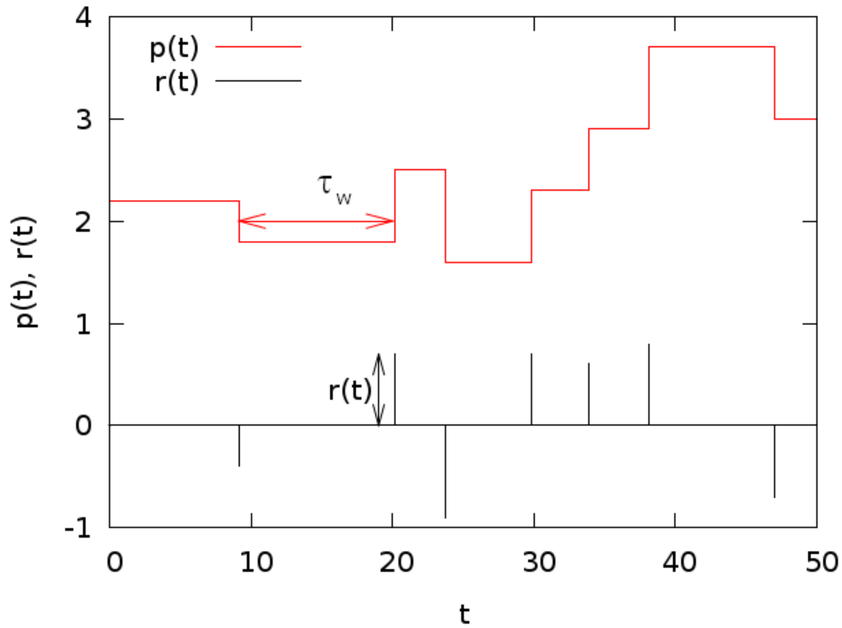

Figure 1: Schematic figure illustrating the time-evolution of price $p(t)$ and its fluctuation, measured as log-return $r(t)$, of an asset. Price changes every time a trade occurs, the time between two successive trades being denoted as τ_w.

resolutions of a single day. For example, the transactions in a stock market do not occur at regularly spaced intervals, implying that the events defining transactions of an asset could themselves be a stochastic process with a given distribution of *waiting times* τ_w, which are defined as the duration between two successive transactions. Motivated by this, we have used the HF equities trading data of the National Stock Exchange (NSE) of India (NSE (2020)) to uncover its principal statistical features. We focus on properties of the market as a whole, as well as, that of individual stocks, *e.g.*, the distribution of trade sizes, the distribution of waiting times between two successive trades, as well as the relation between price fluctuations and waiting times. Our study has implications for the current understanding of the dynamics of a developing market, in particular, the intra-day behavior of trades and price movements, as well as, in modeling such behavior using the tools of statistical physics. Our paper is organized as follows. In Section 2, we briefly describe the data used in our analysis. The results are described in Section 3, with successive subsections dealing with distributions of trade sizes, waiting times, logarithmic returns and their inter-relations. We conclude with an outline of the main findings in Section 4.

2. Description of the Data

In order to study the statistical properties of NSE, we use tick-by-tick HF data for the month of December in each year during the period 1999 to 2012 (the reason for focusing on a particular month is to enable comparison between the behavior seen in different years without the confounding factor of intra-annual seasonal variations). Note that the HF data comes with its own set of challenges, e.g., overwhelming data size, unevenly spaced time series, etc. This is visible from a small sample of the data set for December 2003 shown below:

```
20031201|MTNL|09:56:29|122.20|10
20031201|MTNL|09:56:29|122.25|50
20031201|MTNL|09:56:29|122.30|40
20031201|SATYAMCOMP|09:56:30|335.25|1000
20031201|SAIL|09:56:30|42.70|700
20031201|M&M|09:56:30|355.95|100
20031201|SATYAMCOMP|09:56:30|335.25|500
20031201|SATYAMCOMP|09:56:30|335.25|100
20031201|RAINCALCIN|09:56:30|25.40|500
20031201|VDOCONINTL|09:56:30|78.85|70
```

where the columns separated by "|" represent respectively the date, name of the company whose equity is being traded, the time of the transaction, price per stock of the equity traded, and the number of stocks that changed hands during the transaction (i.e., the trade size q).

It can be observed from the above sample that many transactions share the same time-stamp. This is because the temporal resolution of recording the transactions is 1 second, so that if two transactions occur within a duration of less than a second of each other, they will have the same time-stamp. However, the ordering of transactions is reported in the correct time-sorted order. For the purpose of the present analysis, we assume that the transactions sharing the same time-stamp take place at the same instant. As the market is open for the trading of common stocks only between 0950 hours and 1530 hours, the results reported here are obtained by exclusively considering trades that took place between 0950 hours and 1530 hours on a given day.

In order to study the properties of individual stocks we choose four representative stocks, *viz.*, $HDFCBANK$ (Finance sector), $INFOSYS$ (Infotech sector), $RELIANCE$ (Energy sector) and $SUNPHARMA$ (Pharmaceutical sector), as these belong to a few of the most important industrial sectors in NSE in terms of market capitalization. We show in Figure 2 the variation of price $p(t)$, returns $r(t)$ and trade size $q(t)$ as a function of time for $RELIANCE$ for the first 30 minutes of trading on December 1, 2005 as a typical example of the financial time series.

3. Results

3.1. Distribution of Trade Sizes

We report the cumulative probability distribution $P(Q \geq q)$ of the trade sizes q for the entire market (i.e., aggregating over all equities traded) for the month of December for years from 1999 to 2012 in Figure 3. We see that each of the distributions (corresponding to every year between 1999-2012) when shown on a log-log graph has a substantial linear portion, implying that the tails for the distribution of trade sizes calculated over the entire market follows a power law.

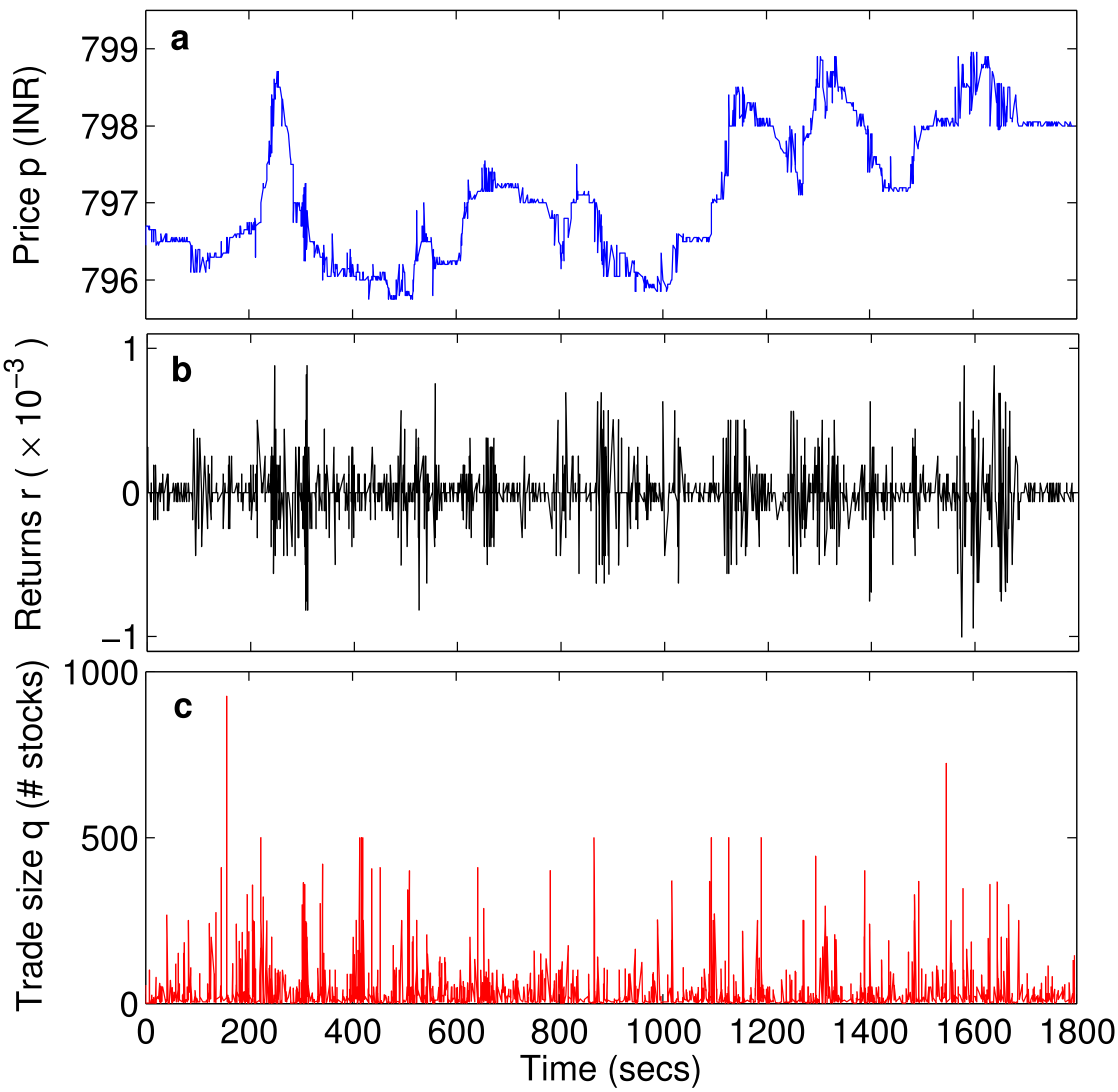

Figure 2: Representative time-series of $RELIANCE$ stock for 30 minutes beginning at 0950 hrs on December 1, 2005 showing (a) price $p(t)$ of the stock, (b) log-returns $r(t)$, and (c) trade size $q(t)$ as a function of time. Time is measured in seconds, with the origin (0) set at 0950 hrs.

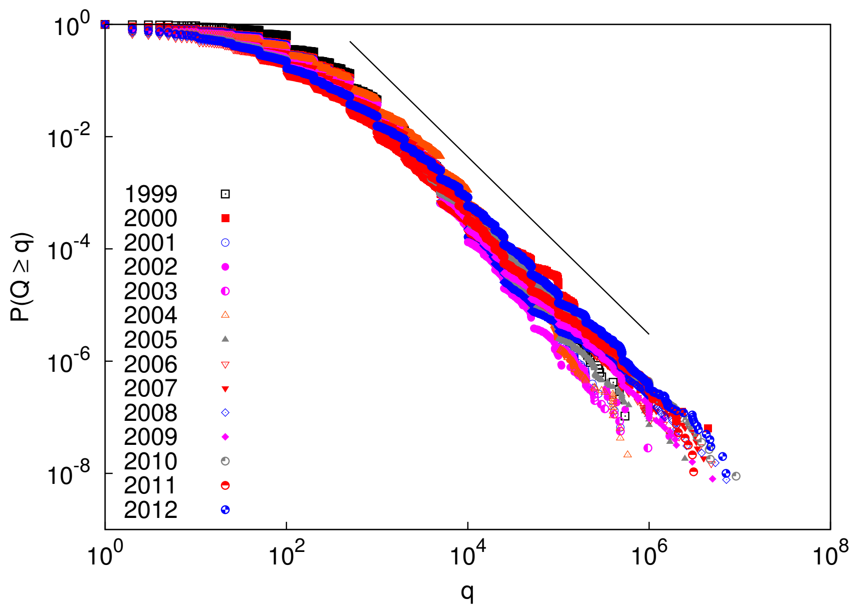

Figure 3: Cumulative probability distribution $P(Q \leq q)$ of the trade sizes q for NSE for the month of December for the years 1999 to 2012. Note the doubly logarithmic axis of the graphs, which means that a linearly decaying tail implies the existence of power-law decay of the distribution (the line represents a power-law fit with exponent 1.6).

Figure 4 shows the distribution of trade sizes q for equities of the four representative companies for the month of December in the years 2003, 2007 and 2010. It is evident from the figure that the tails of the distributions of each of the stocks also exhibit power law decay. Hence, the tails of the distribution of the trade sizes q are of the form:

$$P(Q \geq q) \sim q^{1-\alpha}, \tag{4}$$

where α is the exponent characterizing the power law distribution. The maximum likelihood estimates of the exponents (Clauset *et al.* (2009)) are shown in Figure 5. It is seen from Figure 5(a) that $\alpha \leq 3$ for almost all periods for the entire market (except 2003). This implies that for these distributions, moments other than the first do not exist. As this property holds true for almost the entire period under consideration, we can say that the distribution of trade sizes for the market is stationary, in the sense that it is Lévy stable with second and higher moments diverging. However, this is not true for the case of individual stocks, as seen from Figure 5 (b), where we see that the values taken by α for different stocks vary between 2 and 4. In addition, values taken by α for a particular stock exhibit widely different values depending on the period at which it is being observed, thus suggesting that the statistical behavior of trading dynamics for individual equities is non-stationary.

3.2. Distributions of waiting times and price fluctuations

An important quantity associated with a given stock is its price $p(t)$ at any given time t and which generally fluctuates over time. A generic time-series representing price variations of a given equity is shown in Figure 6. We see in panel (a) that the price $p(t)$ at time

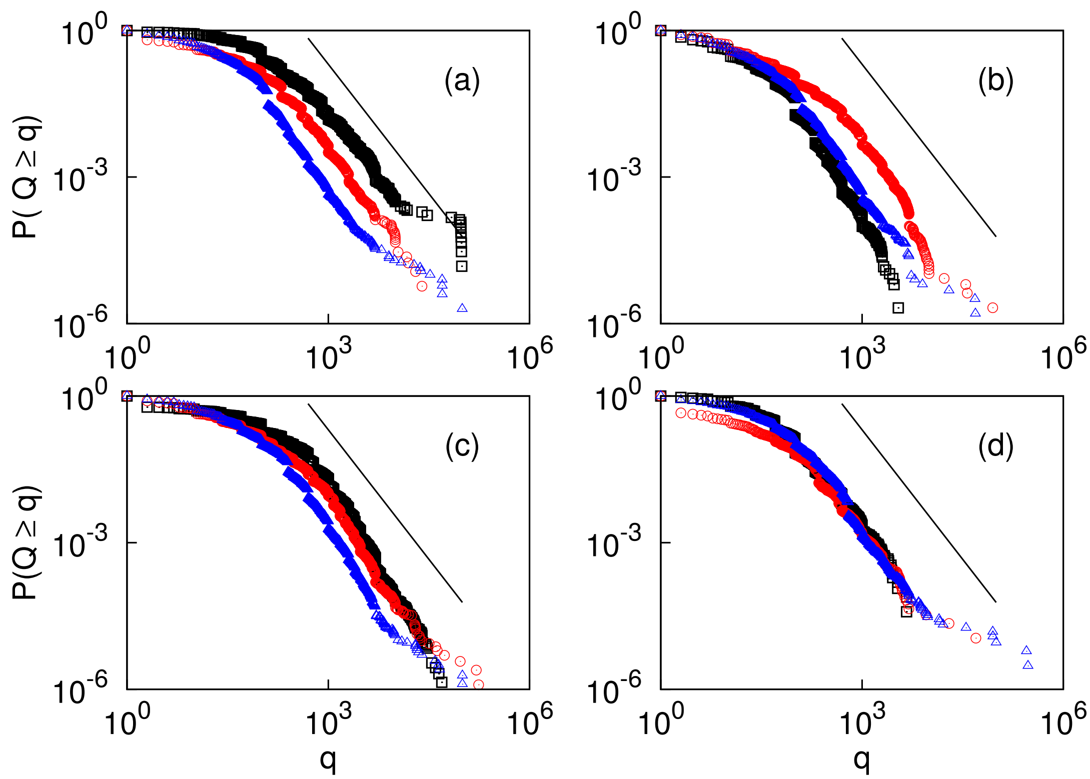

Figure 4: Cumulative probability distribution of trade sizes q for the equities of (a) $HDFCBANK$, (b) $INFOSYS$, (c) $RELIANCE$, and (d) $SUNPHARMA$. The distributions are shown for the month of December for the years 2003 (black squares), 2007 (red circles) and 2010 (blue triangles). As in the case of the entire market, the distributions of trade sizes for individual equities also exhibit a power law form.

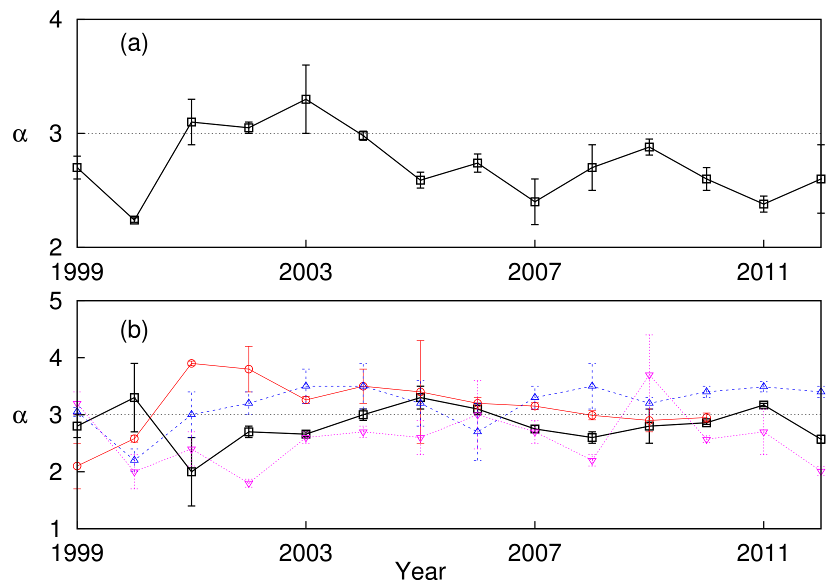

Figure 5: Maximum likelihood estimates of the exponents characterizing the power law nature of the tail for the distribution of trade sizes for (a) the entire market and (b) the equities of $HDFCBANK$ (black squares), $INFOSYS$ (red circles), $RELIANCE$ (blue triangles), and $SUNPHARMA$ (maroon inverted triangles), for the month of December for all years between 1999 to 2012. Note that the data for $INFOSYS$ was not available for the years 2011 and 2012. Bars represent the error in estimating the exponents, obtained using bootstrap technique. The horizontal broken line indicates $\alpha = 3$ which demarcates distributions with Levy-stable nature from those that will eventually converge to a Gaussian.

t and at time $t + \tau_w$ can be same or different, where τ_w is the waiting time between the two transactions. The tick-by-tick log-return associated with the successive price changes is measured as

$$r(t) = \ln p(t) - \ln p(t - \tau_w) \tag{5}$$

and is shown in Figure 6 (b). The random walk nature of price changes is easily seen from the figure, and the presence of irregular waiting times τ_w makes it an effectively continuous time process. Characterizing the distribution of waiting times is of fundamental importance in order to understand the price dynamics of a given asset. For that purpose, we show the distribution of waiting times for $RELIANCE$ for the month of December 2005 in Figure 7.

It can be seen that the distribution of waiting times $P(\tau > \tau_w)$ cannot be fit by a single exponential distribution having a characteristic period $\langle \tau_w \rangle$. This implies that the system has inherent long-range memory and that the occurrence of successive transactions are not independent events. This becomes clear upon fitting the empirical distribution of waiting times with a theoretical curve having the form of a sum of exponentials, *viz.*,

$$P(\tau \geq \tau_w) = \sum_i a_i \exp(-(\tau_w / b_i)). \tag{6}$$

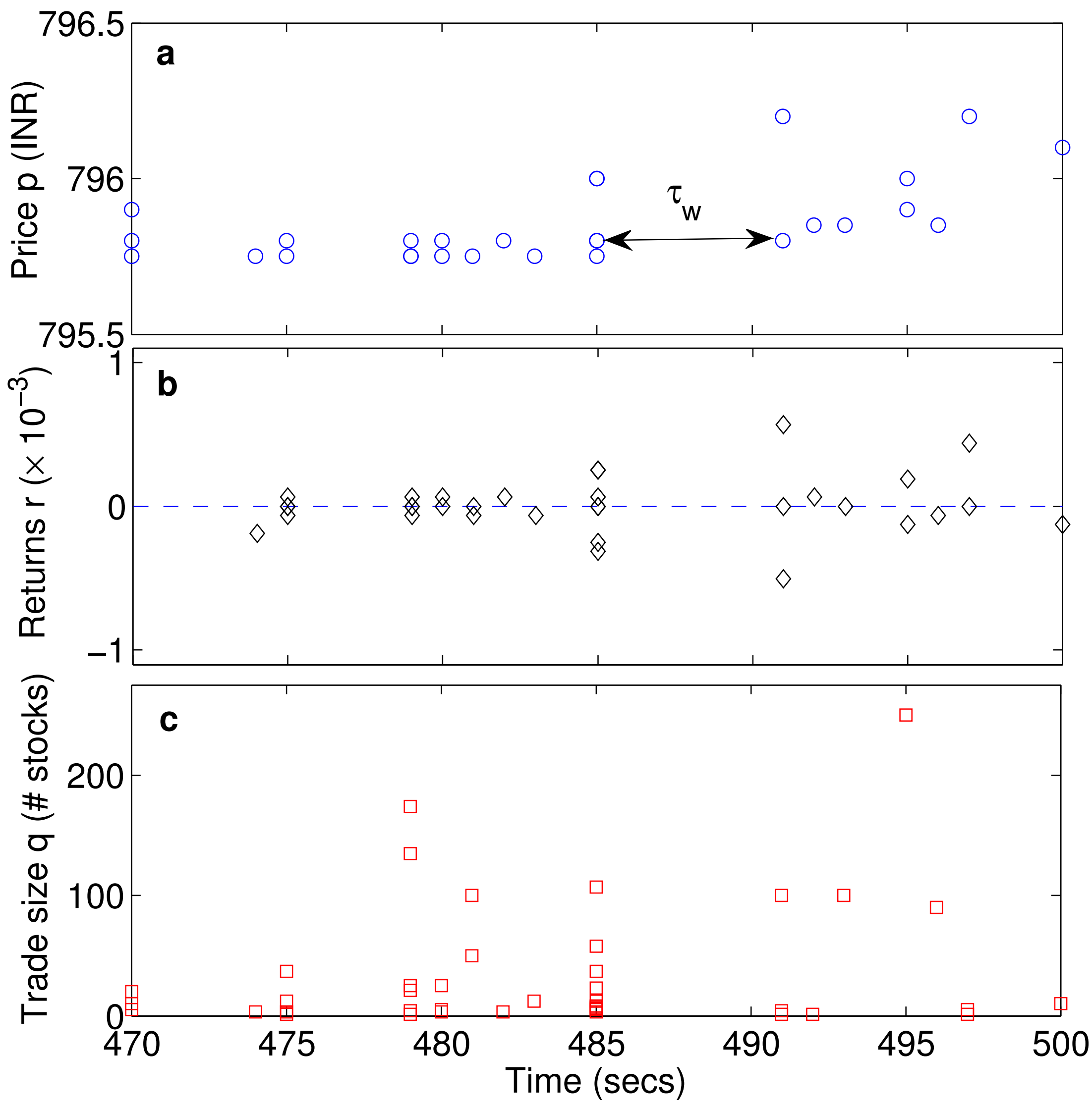

Figure 6: A magnified view of the time series shown in Figure 2 indicating that several transactions can share the same time stamp because the temporal resolution of the recording is limited to 1 second. The waiting time τ_w between two successive trades specifies an interval during which no transaction takes place.

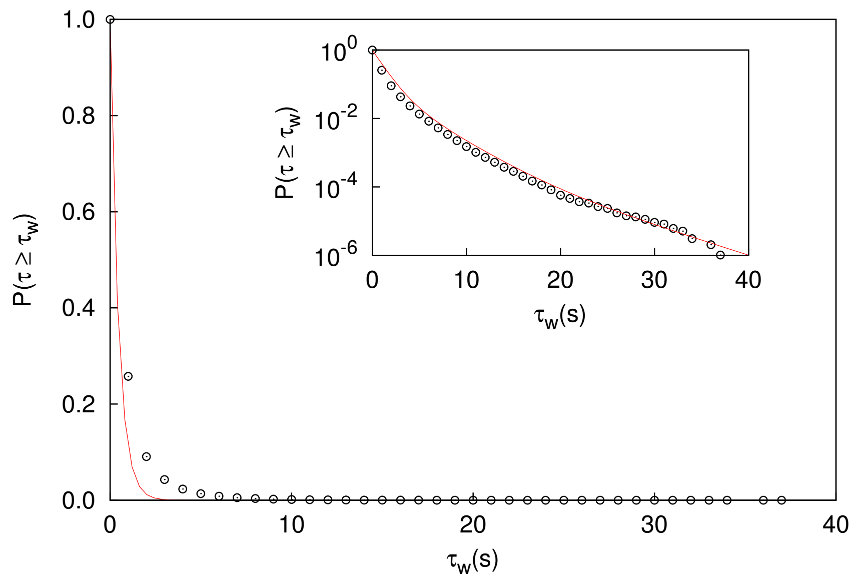

Figure 7: Cumulative probability distribution of waiting times for successive trades in $RELIANCE$ stock for the month of December in 2005. The empirically obtained distribution (circles) has been fitted with an exponential distribution (solid curve) having the same mean waiting time $\langle \tau_w \rangle \approx 0.453$. The inset shows a fit with a theoretical distribution that is a sum of three exponentially decaying curves $\sum_i a_i \exp(-(\tau_w/b_i))$ having different characteristic times τ_w/b_i $(i = 1, 2, 3)$.

Such theoretical distributions have been used earlier to describe systems having memory (Goychuk (2009)). The values of the coefficients used to fit the data shown in Figure 7 are $a_1 = 0.897, a_2 = 0.1, a_3 = 0.003; b_1 = 1, b_2 = 0.4, b_3 = 0.2$.

We also see from Figure 8 that the distribution of returns $Pr(X \geq x)$ has power law decaying tails with exponent close to 3. This is in accordance with the well-known inverse cubic law of price fluctuations reported for financial markets (Lux (1996), Gopikrishnan *et al.* (1998) and Pan & Sinha (2007)).

3.3. Relation between price fluctuations and waiting times

The relation between price fluctuations and waiting times can be discerned from Figure 9 which shows a scatter plot of log-returns r measured for successive transactions against the corresponding time-interval τ_w between them for a particular equity. It can be observed from the diagram that transactions resulting in large price changes generally occur closer to each other. To further quantify this observed behavior we look at the distribution of the log-returns conditioned on waiting times, i.e., $P(r|\tau_w)$ in Figure 10. It is evident from the distribution $P(r|\tau_w)$ that larger returns generally occur close to each other in time (Scalas (2006)), thus suggesting that the waiting-times and returns may not be independent of each other. This property is also observed for different stocks and for different years. This has implications towards the modeling of price dynamics by continuous time random walks, as market transactions may be better modeled by walks whose step lengths are not chosen

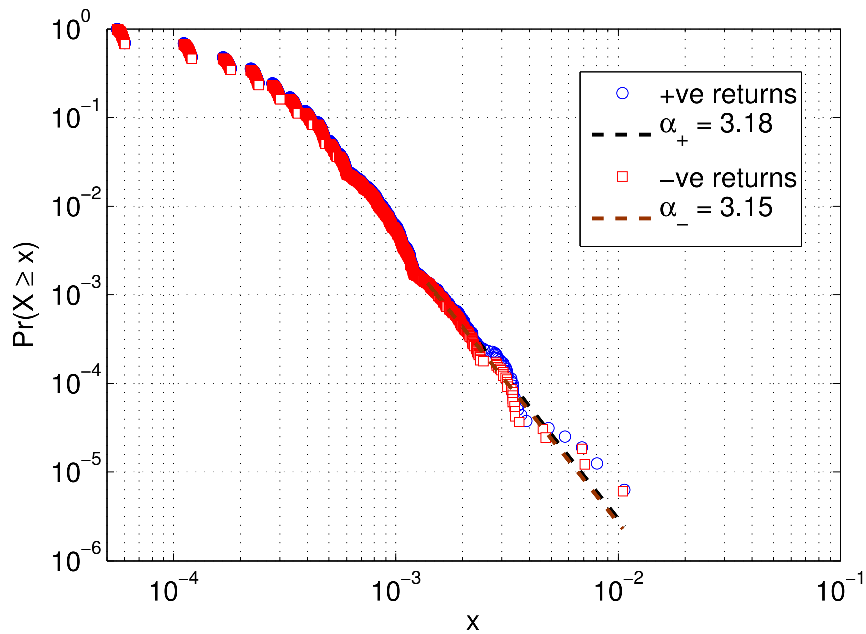

Figure 8: Distribution of log-returns of *RELIANCE* (each return being measured over an interval of 1 tick) for the month of December 2005.

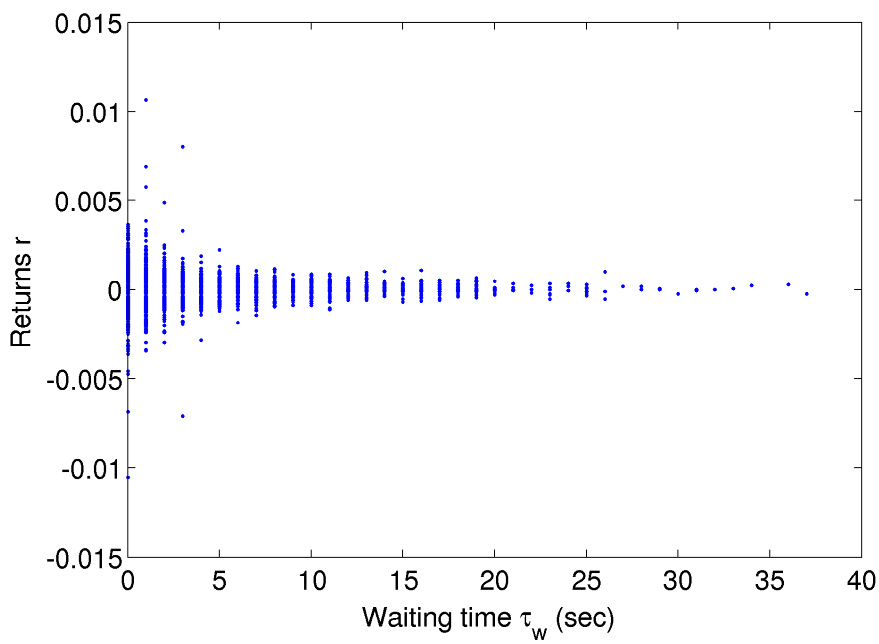

Figure 9: Scatter plot of waiting times τ_w between successive trades and the corresponding log-return r of *RELIANCE* shares for all transactions that took place in December 2005.

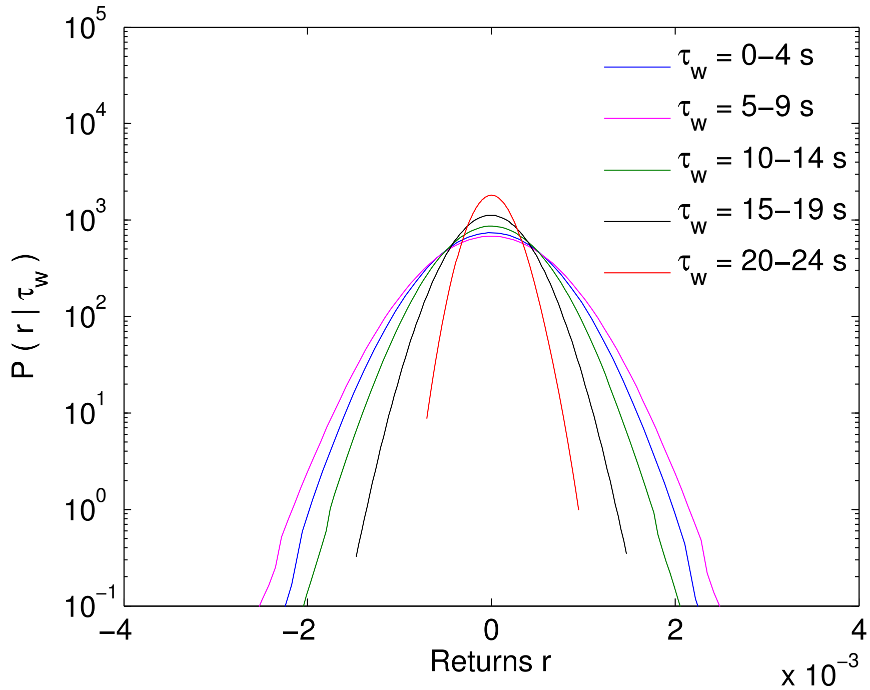

Figure 10: Conditional distribution $P(r|\tau_w)$ of log-returns $r(t)$ conditioned on the waiting times τ_w for successive trades in equities of $RELIANCE$ that were carried out in December 2005.

independently of the waiting time between successive steps, contrary to what is generally assumed (Merton (1976) and Masoliver *et al.* (2000)).

In order to characterize the dynamics of intra-day trading we define the variance of log-returns over an interval Δt as:

$$\sigma^2(t) = \frac{1}{N_{\Delta t}(t) - 1} \sum_i (r_i - \langle r_i \rangle)^2, \tag{7}$$

where $N_{\Delta t}(t)$ is the number of trades occurring in the interval of length Δt and $\langle r_i \rangle$ is the average return over the interval. We observe from Figure 11 that the variance over the intervals is nearly independent of the length of the interval Δt, and find that the scaled cumulative variance $\sigma_c^2(t)\Delta t$ is independent of Δt, as shown in Figure 12. In addition, we also find that the cumulative variance grows linearly with time, thus implying that for the major part of the day, price fluctuations of individual stocks are inherently Gaussian. We can see, however, a perceptible deviation from linearity at the beginning of the trading day, and more prominently towards its end. This suggests that the non-Gaussian nature of the return distribution, as evident from the heavy tails characterized by the inverse cubic law, is possibly an outcome of the significant volume of transactions that occur at either end of a trading day.

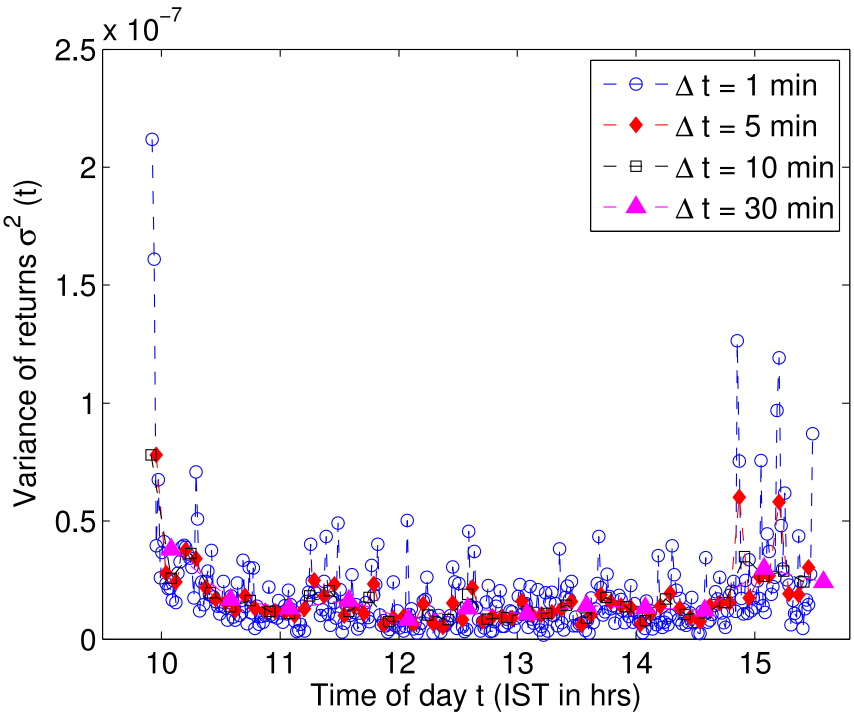

Figure 11: Intra-day volatility $\sigma^2(t)$ (measured by variance of the log-returns) during a period Δt shown as a function of time of day t for $RELIANCE$ on December 1, 2005, for different intervals Δt.

4. Conclusions

In this paper we investigate the statistical behavior of financial markets which constitute prototypical examples of complex systems having large number of components with unpredictable dynamics. Despite such unpredictability, statistically regular properties for the entire system can be observed as has indeed been reported for many different observables associated with market dynamics. Notable among such invariant features are the power laws describing the tails in the distributions of asset price fluctuations, as well as, the distributions of trade sizes. However, most early studies reporting such features have used low-resolution daily data, and thus do not take into account the information about intra-day trading. Motivated by this, in this paper we have reported our preliminary analysis of the high-frequency equities trading data obtained from the National Stock Exchange of India. Such data provides information about market movements at the highest possible resolution thereby revealing vital clues for understanding the underlying dynamics of this complex system.

Using data for the month of December for all years between 1999-2012 we show that gross statistical properties of the market as a whole are in general stationary, even though those of its constituents, i.e., equities of individual companies, may not be. In particular, we see that the distribution of trade sizes aggregated over all equities does not change its nature over time, with the exponent characterizing the power law nature of the tails taking values from the interval $(2, 3)$, i.e., it is Lévy stable with second and higher moments diverging.

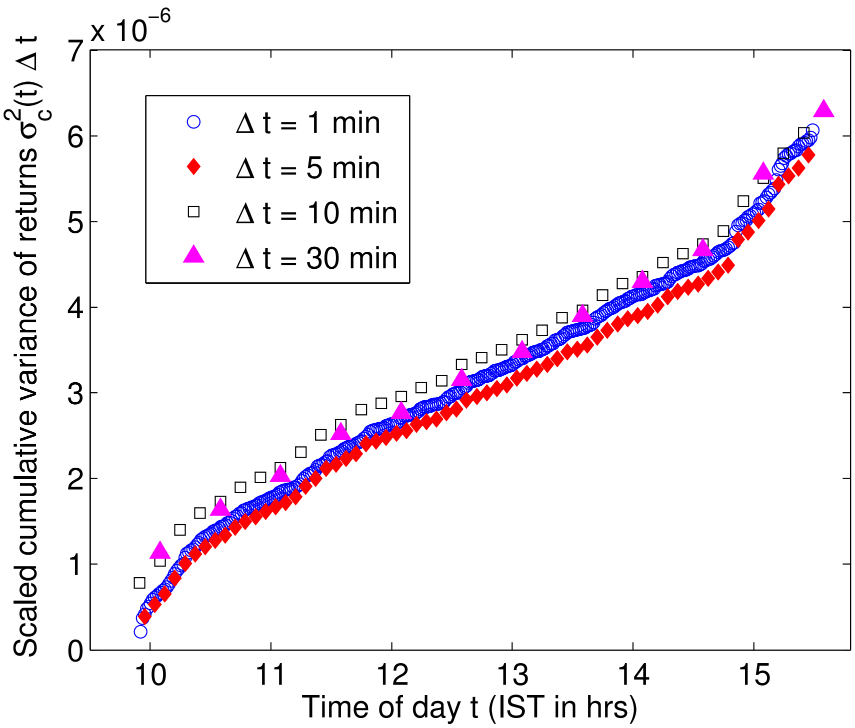

Figure 12: Scaled cumulative variance of the log-returns for intra-day trading $\sigma_c^2(t)\Delta t$ as a function of time of day t for $RELIANCE$ on December 1, 2005. It is observed that scaling by interval Δt over which the variance is calculated, the curves for cumulative variance $\sigma_c^2(t)$ for different choices of Δt overlap.

However, for individual stocks, this distribution can differ significantly between one period and another. We also find that waiting-times of transactions of individual stocks exhibit non-exponential character, and are related to log-returns in that transactions involving larger returns occur close to each other. This implies that the distribution of waiting times and step lengths of the random walk executed by a financial asset are not independent of each other, as has been assumed in many studies. We also find that cumulative volatility of returns increases linearly with time within a day. This implies that for a major part of the day, price fluctuations are Gaussian in nature. However, a significant deviation from linearity is seen towards the ends, suggesting that the genesis of the heavy-tailed nature of return distributions (reflected in the inverse cubic law) lies in the significant volume of trade that occurs at the beginning and towards the end of a trading day.

Acknowledgements

We thank Frederic Abergel, Trilochan Bagarti, Abhijit Chakraborty and Soumya Easwaran for helpful discussions. The work was supported in part by the Center of Excellence in Complex Systems and Data Science, The Institute of Mathematical Sciences, funded by the Department of Atomic Energy, Government of India.

References

Chakraborty, A., Easwaran, S. and Sinha, S. (2018). Deviations from universality in the fluctuation behavior of a heterogeneous complex system reveal intrinsic properties of components: The case of the international currency market. *Physica A*, **509**, 599–610.

Chakraborty, A., Easwaran, S. and Sinha, S. (2020). Uncovering hierarchical structure of international FOREX market by using similarity metric between fluctuation distributions of currencies. *Acta Physica Polonica A*, **138(1)**, 105–115.

Clauset, A., Shalizi, C. R. and Newman, M. E. J. (2009). Power-law distributions in empirical data. *SIAM Review*, **51(4)** 661–703.

Dacorogna, M. M., Gençay, R., Müller, U. A., Olsen, R. B. and Pictet, O. V. (2001). *An Introduction to High-frequency Finance.* Academic Press, San Diego.

Easwaran, S., Dixit, M. and Sinha, S. (2015). Bitcoin dynamics: The inverse square law of price fluctuations and other stylized facts. In *Econophysics and Data Driven Modelling of Market Dynamics*, Springer, Cham, 121–128.

Gopikrishnan, P., Meyer, M., Amaral, L. A. N. and Stanley, H. E. (1998). Inverse cubic law for the distribution of stock price variations. *European Physical Journal B*, **3(2)**, 139–140.

Goychuk, I. (2009) Viscoelastic subdiffusion: From anomalous to normal. *Physical Review E*, **80(4)**, 046125.

Lux, T. (1996). The stable Paretian hypothesis and the frequency of large returns: an examination of major German stocks. *Applied Financial Economics*, **6(6)**, 463–475.

Masoliver, J., Montero, M., Perelló, J. and Weiss, G. H. (2000). The continuous time random walk formalism in financial markets. *Journal of Economic Behavior & Organization*, **61(4)**, 577–598.

Merton, R. C. (1976). Option pricing when underlying stock returns are discontinuous. *Journal of Financial Economics*, **3(1-2)**, 125–144.

NSE (2020). *NSE India: About NSE Data & Analytics.* https://www1.nseindia.com/supra_global/content/dotex/about_dotex.htm.

Pan, R. K. and Sinha, S. (2007). Self-organization of price fluctuation distribution in evolving markets. *Europhysics Letters*, **77(5)**, 58004.

Pan, R. K. and Sinha, S. (2008). Inverse-cubic law of index fluctuation distribution in Indian markets. *Physica A*, **387(8)**, 2055–2065.

Scalas, E. (2006). The application of continuous-time random walks in finance and economics. *Physica A*, **362(2)**, 225–239.

Sinha, S., Chakrabarti, A. S. and Mitra, M. (2016). What is economics that physicists are mindful of it? *The European Physical Journal Special Topics*, **225(17-18)**, 3087–3089.

Sinha, S., Chatterjee, A., Chakraborti, A. and Chakrabarti, B. K. (2010). *Econophysics: An Introduction.* Wiley-VCH, Weinheim.

Vijayraghavan, V. S. and Sinha, S. (2011). Are the trading volume and the number of trades distributions universal ? In *Econophysics of Order-Driven Markets*, Springer, Milan, 17–30.