Chapter 1

Introduction to Complex and Econophysics Systems: A navigation map

T. Aste[1,2,3] and T. Di Matteo[1,2]

[1] *Department of Applied Mathematics, School of Physical Sciences, The Australian National University, Canberra, ACT 0200, AUSTRALIA;*
[2] *Department of Mathematics, King's College, The Strand, London, WC2R 2LS, UK;*
[3] *School of Physical Sciences, University of Kent, Canterbury, Kent, CT2 7NH, UK.*

This Chapter is an introduction to the basic concepts used in complex systems studies. Our aim is to illustrate some fundamental ideas and provide a *navigation map* through some of the cutting edge topics in this emerging science. In particular, we will focus our attention to *econophysics* which mainly concerns the application of tools from statistical physics to the study of financial systems.

Contents

COMPLEX PHYSICAL, BIOPHYSICAL AND ECONOPHYSICAL SYSTEMS - Proceedings of the 22nd Canberra International Physics Summer School

http://www.worldscibooks.com/physics/7340.html

1.1. Introduction

We all experience complexity in everyday life, where simple answers are hard to find and the consequences of our actions are difficult to predict. Modern science recognizes that problems involving the collective behavior of many interacting elements are often "complex." These systems typically display collective, organized behaviors that cannot be predicted from traditional atomistic studies of their components in isolation. As a result, phenomena in complex systems are often counterintuitive. Their explanations require the use of new tools and new paradigms which range from network theory to multi-scale analysis.

There is no uniquely agreed definition of complex system, but if we look up the word "complex" in a dictionary we might find *"a whole made up of complicated or interrelated parts"* (from Merriam-Webster Online Dictionary). Indeed, complex systems are in general made up of several parts, which are interrelated and often complex themselves. Financial systems provide great examples of complex systems, being systems with a very large number of agents that interact in complicated ways, the agents themselves being complex individuals who act following rules and feelings, applying both knowledge and guesswork.

To properly introduce complexity from a common, well established,

COMPLEX PHYSICAL, BIOPHYSICAL AND ECONOPHYSICAL SYSTEMS - Proceedings of the 22nd Canberra International Physics Summer School

ground it is probably better to start from a time when the universe was perceived to be "harmonious" and it was believed that the laws of nature should tend to produce order. In the introduction of the *Principia* (1687) Newton writes: *"I wish we could derive the rest of the phenomena of nature by the same reasoning from mechanical principles for I am induced by many reasons to suspect that they may all depend on certain forces."* Laplace (1749-1827) was indeed persuaded that *"An infinitely intelligent mathematician would be able to predict the future by observing the present state of the universe and using the laws of motion."* This idea has been central in the foundation of modern science and we can still find a very similar idea expressed by Einstein and Infield in 1938:[1] *"(...) all these things contributed to the belief that it is possible to describe all natural phenomena in terms of simple forces between unalterable objects."* Within this framework, the final aim of a scientific endeavor is to *"(...) reduce the apparent complexity of natural phenomena to some simple fundamental ideas and relations."*[1]

However, together with the evolution of our scientific knowledge it has become increasingly clear that sometimes the reduction of the system into smaller and simpler parts with known relations might not lead to any valuable knowledge about the overall system's properties. Probably, in this respect, the most revealing examples are living biological systems. There are some simple organisms of which we know essentially everything from the molecular level, through the cellular organization, up to the animal behavior. However, we still miss one fundamental emerging point: the animal is *alive* and, although we can understand each single part of it, we cannot describe the property of being alive from the simple assemblage of the inanimate parts. Even Newton, well after the glorious years of the Principia, had to admit that *"I can calculate the motions of the heavenly bodies, but not the madness of people."* This was in 1720, when after making a very good profit from the stocks of the South Sea Company, he reinvested right at the top of what is now known the "South Seas Bubble," which crashed and made him loose 20,000 pounds.

The study of complex systems is a very challenging endeavor which requires a paradigmatic shift. Indeed, in these systems the global behavior is an "emergent property" which is not simply related to the local properties of its components. Energy and information are constantly imported and exported across system boundaries. History cannot be ignored, even a small change in the current circumstances can lead to large deviations in the future. In complex systems the "equations of motion" themselves

COMPLEX PHYSICAL, BIOPHYSICAL AND ECONOPHYSICAL SYSTEMS - Proceedings of the 22nd Canberra International Physics Summer School

can evolve and change: in response to external (or internal) changes, the system can reorganize itself without breaking and it can adapt to the new environment. Complex systems have multiple (meta) (stable) states where small perturbations lead to recovery and larger ones can lead to radical changes of properties. In these systems, dynamics does not average simply. Complex systems are multi-scale and hierarchical. Complex systems are characterized by, and often dominated by, unexpected, unpredictable, adaptive, emerging behaviors, which can span over several orders of magnitude. Dynamics can propagate through scales, both upwards, when the system is hierarchically organizing, and downwards, when large fluctuations may melt down such hierarchy. The presence of large, scale-free, power-law fluctuations makes often impossible from a finite set of measures to calculate the parameters characterizing these statistical fluctuations. Complex systems are disordered in the sense that there is no compact and concise way to encode the whole information contained in the system. Even the measure of complexity is a complex task per se. We can say that the complexity of a system scales by its number of elements, its number of interactions, by the complexity of the element, by the complexity of the interaction. This is a "recursive" measure which reflects the multi-scale nature of these systems.

We are all aware that accurate predictions of real world complex phenomena are very challenging. Forecasting tomorrow's weather or the house market trends for the next five years, seems to be more a form of art than an exact science. Indeed, as Neils Bohr emphatically warned, *"Prediction is very difficult, especially about the future."* As the physicist Giorgio Parisi pointed out, the meaning of the word *prediction* has evolved with the evolution of science and it has assumed a new meaning in the context of complex systems.[2] At the times of Newton or Laplace, when *classical mechanics* was founded, prediction meant the accurate forecast of the position and velocity of any object for the infinite future, given a precise knowledge of the present position and velocity. The first change in the meaning of prediction happened already in the 19$^{\text{th}}$ century with the beginning of *statistical mechanics.* In that context, prediction becomes a probabilistic concept where the theory no longer aims to describe the behavior of each single molecule, but the statistical properties of the ensemble of the molecules. With the advent of *quantum mechanics* the probabilistic description of natural phenomena became the predictive framework for atomic and subatomic phenomena, and the uncertainty principle introduced the idea that some variables might not be measurable simultaneously with arbitrary precision. More recently, the word prediction assumed another new significance in the context of the theory of nonlinear dynamics. In deterministic chaos, despite the fact that the system behavior is fully determined by the initial conditions, the high sensitivity of trajectories over large time intervals to infinitesimal changes of the initial conditions makes such a deterministic prediction meaningless. In these systems, prediction concerns the identification of classes of long-time behaviors associated with given regions of the space of initial conditions, and statistical statements about these behaviors.

The science of complex systems has introduced another paradigmatic change to the concept of prediction and consequently it has changed the meaning of physical investigation. For these systems the main question is the dependence of the emerging behaviors at system level on the set of rules and constraints imposed at local level. Hierarchy and emergence imply that we must describe the system at different levels of abstraction. In these systems, models and theories often apply only to a given abstraction level and different theories might apply to different levels. This is a different and somehow lesser power of predictability compared with traditional physical theories. On the other hand, this opens the way to apply physical investigation to new kinds of systems: the modeling of brain functions, the study of financial markets, or the study of the influence of a given gene on some biological functions, are among the topics currently investigated by physicists.[3,4]

A few years ago, Stephen Hawking declared that *"the next century will be the century of complexity."* Indeed, science is changing in response to the new problems arising from the study of complex systems. The scientific community now faces new expectations and challenges, the nature of problems has forced modern scientists to move beyond the conventional reductionist approaches. Complex systems studies have opened a new scientific frontier for the description of the social, biological, physical, and engineered systems on which human society has come to depend.

1.2. An Example of Complex Systems: Financial Markets

Financial markets are open systems in which many subunits interact nonlinearly in the presence of feedback. Financial systems are archetypal of complexity. In markets, several different individuals, groups, humans and machines, generically called "agents," operate at different frequencies with different strategies. The agents interact both individually and collectively at different levels within an intricate network of complicated relations. The emerging dynamics continuously evolves through bubbles and crashes with unpredictable trends. Although intrinsically "complex," financial systems are very well suited for statistical studies. Indeed, the governing rules are rather stable and the time evolution of the system is continuously monitored, providing therefore a very large amount of data for scientists to investigate.

Since the mid '90s a growing number of physicists have undertaken the challenge of understanding financial and economic systems. A new research area related to complex systems research has emerged and it has been named "Econophysics." This is a relatively recent discipline, but it has already a rich history, with a variety of approaches, and even controversial trends.[5–11] Econophysics is an interdisciplinary field which applies the methods of statistical physics, nonlinear dynamics, and network theory to macro-micro/economic modeling, to financial market analysis and social problems.

There are several open questions that econophysicists are actively investigating. Some of the main topics concern:

- development of theoretical models able to encompass empirical evidence;
- statistical characterization of the stochastic process describing price changes of a financial asset;
- search for scaling and universality in economic systems;
- implementation of models for wealth and income distributions;
- use of network theory and statistical physics tools to describe collective fluctuations in financial assets prices;
- development of statistical mechanics approaches in socio-economic systems;
- exploration of novel theoretical approaches for interacting agents;
- investigation of new tools to evaluate risk and understand complex system behaviors under partial information and uncertainty.

In this Chapter we will focus on a few examples of financial system studies from an Econophysics perspective.

1.3. Probabilities and Improbabilities

Let us start our "navigation" from some fundamental concepts and theorems from probability theory that are of relevance to the study of complex systems. In particular we focus our attention on large fluctuations and the probability of their occurrence, that, as we shall see shortly, is an important characterizing aspect of complex systems phenomena.

COMPLEX PHYSICAL, BIOPHYSICAL AND ECONOPHYSICAL SYSTEMS - Proceedings of the 22nd Canberra International Physics Summer School

http://www.worldscibooks.com/physics/7340.html

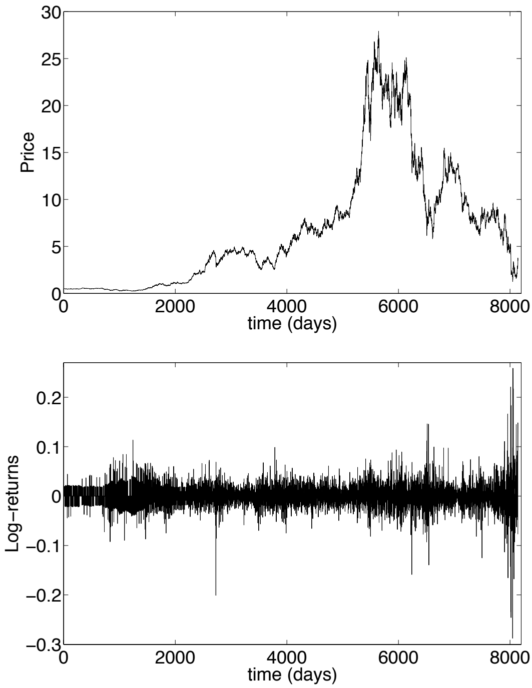

Fig. 1.1. (top) Daily closing adjusted prices for the stock Ford Motor in the New York Stock Exchange market. The time-period ranges from 3 Jan 1977 to 7 April 2009 (there are 8142 points in total). (bottom) Log-returns over the same time-period. The data are from http://finance.yahoo.com.

1.3.1. *Log-returns*

In order to let the reader focus on a practical example, let us start with the study of the statistical properties of daily closing prices of an equity over a given period of time (let us, for instance, analyze the prices of Ford Motor Company in the New York Stock Exchange as shown in Fig. 1.1). From a statistical point of view we want to characterize the daily variations of these prices, and to this end it is convenient to look at the so-called log-returns, defined as:[5,12]

$$r(t, \tau) = \log(\text{price}(t + \tau)) - \log(\text{price}(t)) \quad ; \tag{1.1}$$

where, in this case, we take $\tau = 1$ day. We can calculate from the data plotted in Fig. 1.1 that these log-returns fluctuate around zero with a sample

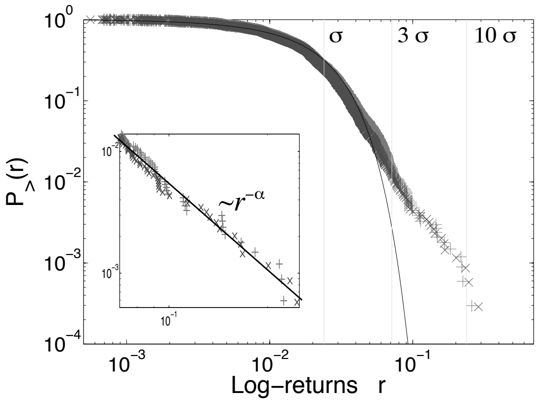

Fig. 1.2. Complementary distribution of the log-returns for the Ford Motor stock prices reported in Fig. 1.1. The "+" is the distribution of the positive returns $P(R \geq r)$ and the "×" the distribution of the negative ones $P(R \leq -r)$. The line is the comparison with a normal distribution with the same mean and variance. The inset is the tail region and the linear behavior in log-log scale highlights that there is a characteristic power-law decreasing trend $P_>(r) \sim r^{-\alpha}$. The best fit reveals an exponent $\alpha \sim 2.4$. The vertical lines correspond to deviation from the mean of respectively one, three and ten standard deviations (σ).

mean[a] $\mu = \langle r(t,\tau) \rangle \sim 2.4 \times 10^{-4}$ which is very small if compared with the estimated standard deviation $\sigma \sim 0.0237$. On the other hand, the distribution has a rather large fourth central moment[13] $\mu_4 = \left\langle (r(t,\tau) - \mu)^4 \right\rangle \sim 16.9$.

1.3.2. *Leptokurtic distributions*

The fact that higher moments have increasingly large relative values is a good indication that the distribution might deviate from a normal distribution. Such a deviation is often measured by using the excess kurtosis:

$$\gamma_2 = \frac{\mu_4}{\sigma^4} - 3 \ ; \tag{1.2}$$

In the case of Ford Motor we obtain $\gamma_2 \sim 5 \times 10^7$, which is a very sizable deviation. In fact, let us stress that the excess kurtosis of a *normal* distribution is zero. Distributions with large kurtosis are called *leptokurtic*. They are characterized by larger-than-normal probability of very small

[a] The "sample mean" $\langle x \rangle$ of a given set of values $\{x_1, ...x_n\}$ is calculated as the sum of all entries divided by their number $\langle x \rangle = (1/n)\sum_{i=1}^{n} x_i$. More generally, the sample mean of any function $f(x)$ is $\langle f(x) \rangle = (1/n)\sum_{i=1}^{n} f(x_i)$.

fluctuations, but also by larger-than-normal probabilities of very large fluctuations.

1.3.3. *Distribution tails*

The deviation of the fourth central moment μ_4 from that expected for a normal distribution is a very good indication that these systems should have special non-normal statistical properties. However, we must stress that it is very important in complex systems studies to look at the whole distribution and not only at the moments. In particular, for financial analysis and risk assessment the most important part of a distribution, which must be studied with attention, is is that describing large fluctuations, the so-called "tail of the distribution."

An idea of the deviation from the normal distribution in the tail region is given in Fig. 1.2 where we plot the complementary cumulative distribution, which is defined as:

$$P(R \geq r) = P_{>}(r) = \int_r^\infty p(s)ds \tag{1.3}$$

with $p(s)$ the probability density function. One can see from Fig. 1.2 that both the positive and negative tails deviate from a normal probability function with the same average and standard deviation. We see that the probability to find a deviation from the mean of the order of 1 standard deviation is about 30% (once every few days) and it is comparable for both the measured distribution and for the normal one. A sizable deviation is instead observed if we look at the probability to observe fluctuations larger or equal than 3 standard deviations. The normal distribution predicts 0.3% which is essentially once every year or two, in average. On the other hand, the observed frequency is in average once every 4 months. The deviation between the normal statistics and the observed one becomes huge if we move further away from the average. For instance, fluctuations larger than 10 standard deviations are observed several times during the investigated period 03/01/77 - 07/04/09 ($\sim$ 8000 days). Conversely, the normal statistics predicts an extremely small probability for such event (probability $\sim 10^{-23}$). In practice, it predicts that it would be very unlikely to observe such a fluctuation even if one waits for a time longer than the age of the universe.

1.4. Central Limit Theorem(s)

In the previous section we have compared the observed probability distributions with the normal one. Indeed, normal distributions are commonly observed in a very wide range of natural and artificial phenomena throughout statistics, natural science, medicine and social science. One of the reasons for this wide occurrence of normal distributions is that in several phenomena the observed quantities are sums of other (hidden) variables that contribute with different weights to the observed value. The Central Limit Theorem guarantees us that, *under some conditions*, the aggregate distribution in this kind of additive process tends towards the normal distribution. Therefore, a deviation from normal distribution is a good indication that we are dealing with a particular class of phenomena.

Let us first discuss the case where the distribution converges towards the normal one and then let us understand the origin of the observed deviations.

1.4.1. *Tendency towards normal distribution*

The Central Limit Theorem (CLT) states that if the various contributing variables are distributed independently and if they all follow the same identical distribution [i.e. they are *independent and identically-distributed* (i.i.d.) random variables] with *finite variance*, then the sum of a large number of such variables will tend to be normally distributed.[13] Which is, given n i.i.d. variables $\{x_i\}$ with mean[b] $E(x_i) = \mu$ and finite variance $E((x_i - \mu)^2) = \sigma^2$, the probability distribution of the sum

$$y = \sum_{i=1}^{n} x_i \tag{1.4}$$

is approximated by the probability density function:

$$p_n(y) \sim \Phi(y) = \frac{1}{\sqrt{2\pi n\sigma^2}} \exp\left(-\frac{(y - n\mu)}{2n\sigma^2}\right) , \tag{1.5}$$

for large n. The Berry–Esseen theorem guarantees that, if the third absolute moment $\mu_3 = E(|x_i - \mu|^3)$ is finite, then such a convergence is such that the difference goes as $1/\sqrt{n}$. More precisely: $|p_n(y - E(y)) - \Phi(y - E(y))| \leq 3c\mu_3/(\sigma^3\sqrt{n})$ with c a constant smaller than 1 and larger than 0.409.

The conditions dictated by the Central Limit Theorem to obtain normal distributions are quite broad and normal distributions are indeed

[b]The symbol $E(X)$ is the expectation value (or mean, or first moment) of the random variable X.

widespread. However, they are not commonly observed in complex systems, where strong deviations from the normal behavior are routinely found especially for large fluctuations.

1.4.2. *Violation of central limit theorem*

The central limit theorem (CLT) applies to a sum of random variables and it relies on three assumptions. These conditions are often violated in complex systems. Let us discuss each one of these conditions schematically.

- The CLT applies to a *sum* of variables. However, there are several processes which are not purely additive. One extreme case is a purely multiplicative process where the observable is a product of random variables. Incidentally, this extreme case is also a particular one because the product can be transformed into a sum by applying the logarithm and the CLT can be applied on the distribution of the log of the variable resulting in a log-normal distribution. However, in general, the process can be a mix of multiplicative and additive terms. Moreover, several different variables can contribute in an interrelated way through a network of "interactions."
- A condition is that the variables should be *independent*. On the other hand, often the variables are correlated and therefore not independent. Sometimes these correlations are associated with cascading events (one event triggers the other, which causes another, etc.) that can produce "avalanches" characterized by very large fluctuations in sizes with distributions having power law behaviors.
- A second condition requires the variables to be *identically distributed*. Again, often this is not the case and a broad range of distributions can sometimes shape the resulting aggregate distribution in an almost arbitrary way. However, we will see in the following Sections that the statistical behavior simplifies if one limits the study to extreme fluctuations only.
- The last condition requires *finite variance*. On the other hand, it is now widely recognized that in many complex systems the probability of large fluctuations often decreases slower than an exponential, usually with a power law trend $p(x) \sim x^{-\alpha-1}$, and the variance becomes undefined when $\alpha \leq 2$. To this class of distributions with non-defined variance an extension of the CLT applies.

1.4.3. *Extension of central limit theorem*

The Central Limit Theorem can be extended to a more general class of additive processes by dropping the condition of finite variance. Given n i.i.d. variables the sum $y = \sum_{i=1}^{n} x_i$ tends to a stable probability density function $f(y)$ which has characteristic function[14,15]

$$\hat{f}(q) = \exp\left(iq\mu - |cq|^\alpha \left(1 - i\beta \frac{q}{|q|}\Phi(\alpha, q)\right)\right) \ , \tag{1.6}$$

for $0 \leq \alpha \leq 2$ and with $\Phi(\alpha, q) = \tan(\pi\alpha/2)$ when $\alpha \neq 1$ or $\Phi(\alpha, q) = 2/\pi \log(|q|)$ when $\alpha = 1$. Let us recall that the characteristic function is defined by the transformation: $\hat{f}(q) = \frac{1}{2\pi}\int_{-\infty}^{+\infty} e^{-iqy} f(y)dy$ and, vice versa, $f(y) = \int_{-\infty}^{+\infty} e^{iqy}\hat{f}(q)dq$. The parameter $c > 0$ is a scale factor which is a measure of the width of the distribution. The parameter $-1 \leq \beta \leq 1$ is called skewness and is associated with asymmetry. In the symmetric case, when $\beta = 0$, the distribution becomes a stretched exponential function. In the case $\alpha = 2$, Eq. (1.6) gives the normal distribution. When $\alpha > 2$ the variance is finite and the central limit theorem applies predicting therefore the convergence towards the normal distribution. In general, the distribution defined by the characteristic function in Eq. (1.6) has no compact analytic form for $f(y)$ in the direct space. However, it is rather simple to show that, in the asymptotic limit of large fluctuations, the probability density function decreases as a power law, $f(y) \sim y^{-\alpha-1}$, where the exponent α is the same exponent as the one from the tails of the distributions of the variables x_i.

1.4.4. *Stable distributions*

The normal distribution and the distribution in Eq. (1.6) are "stable distributions." As a general property, stable distributions must satisfy the following condition: a distribution is stable if and only if, for any $n > 1$, the distribution of $y = x_1 + x_2 + ... + x_n$ is equal to the distribution of $n^{1/\alpha}x + d$ with $d \in \mathbb{R}$.[14] This implies

$$p_n(y) = \frac{1}{n^{1/\alpha}} p\left(\frac{y-d}{n^{1/\alpha}}\right) \ , \tag{1.7}$$

where $p_n(y)$ is the aggregate distribution of the sum of the n i.i.d. variables and $p(x)$ is the distribution of each of the variables x_i, with $i = 1, ...n$. The distribution is called strictly stable if $d = 0$. It is rather simple to prove that the distribution in Eq. (1.6) satisfies the scaling in Eq. (1.7) and it is indeed a stable distribution.

COMPLEX PHYSICAL, BIOPHYSICAL AND ECONOPHYSICAL SYSTEMS - Proceedings of the 22nd Canberra International Physics Summer School

1.5. Looking for the Tails

A key question that we all might wish to answer is: *what is the maximum loss that we might possibly experience from our financial investments?* When dealing with risk we must be aware that in the market extremely large fluctuations can happen with finite probability. We often underestimate risk because extreme fluctuations are rather unusual in natural phenomena described by normal statistics. Large fluctuations described by non-normal statistics are instead rather common in financial systems, and in complex systems in general, representing one of the distinctive features of these systems. Indeed, a crucial key to many risk management problems is the understanding of the occurrence of extreme losses. It is for instance important in the evaluation of insurance losses from natural disasters such as hurricanes or earthquakes. Extreme losses happen rarely but they can be catastrophically deadly. However, the very fact that they are rare means that there is little statistics to rely on, which makes very difficult to predict the probability of their occurrence with precision. It is therefore very important to implement methods which can help to precisely estimate the behavior of the probability distribution in the region of large and rare variations, the "tail" of the distribution.

1.5.1. *Extreme fluctuations*

Let us consider a sequence of events $x_1, .., x_n$ characterized by a probability distribution function $p(x)$. We are interested in estimating the probability of the *maximum value* of such events $\max\{x_1, .., x_n\}$ for a given number n of occurrences. (For instance the largest size of the loss in the most catastrophic hurricane over a series of n hurricanes.)

A priori the probability of the events x_i can follow any kind of distribution. However, we are asking for the probability of the *maximum* and, in this case, we have an important general result, which is valid for asymptotic distributions of extreme order statistics.

The *Fisher–Tippet–Gnedenko, extreme value theorem*[6,13] states that the maximum of a sample of independent and identically distributed random variables, after proper renormalization, converges in distribution to one of three possible distributions, the *Gumbel* distribution, the *Fréchet* distribution, or the *Weibull* distribution.

1.5.2. *Extreme value distribution*

These distributions are particular cases of the generalized extreme value distribution (GEV), whose complementary cumulative distribution is:

$$G(x) = \exp\left(-\left(1+\frac{1}{\alpha}\left(\frac{x-\mu}{\sigma}\right)\right)^{-\alpha}\right) \tag{1.8}$$

for $1+(1/\alpha)(x-\mu)/\sigma > 0$. This is the general limit distribution of properly normalized maxima of a sequence of i.i.d. random variables. The subfamilies defined by $\alpha > 0$ and $\alpha < 0$ correspond, respectively, to the *Fréchet* and *Weibull* distributions whereas the *Gumbel* distribution is associated with the limit $\alpha \to \infty$.

1.5.3. *Fat-tailed distributions*

For the study of price fluctuations in financial markets, and specifically for risk analysis, we are interested in "fat-tailed" distributions where the complementary cumulative distribution $P_>(x)$ tends to $1-G(x)$ (Fréchet) in the tail region of large x. It is easy to show that for large x this distribution behaves as a power law with

$$1 - G(x) \sim ax^{-\alpha} . \tag{1.9}$$

Therefore, the tail of the distribution is associated with one parameter only: the exponent α which fully characterizes the kind of extreme value statistics.

From a general perspective, as far as extreme event statistics is concerned, we can classify probability distributions in three broad categories with respect to the value of the tail index α.

1) Thin-tailed distributions, for which all moments are finite and whose cumulative distributions decrease at least exponentially fast in the tails, they have $\alpha \to \infty$.
2) Fat-tailed distributions whose cumulative distribution function declines as a power law in the tails. For these distributions only the first k moments with $k < \alpha$ are bounded, and in particular:
   - for $\alpha > 2$ the standard deviation is finite and the distribution of a sum of these variables will converge towards the normal form in Eq. (1.5) (Central Limit Theorem[13]);
   - for $0 < \alpha \leq 2$ the standard deviation is not defined and the distribution of a sum of these variables will converge towards the Levy

COMPLEX PHYSICAL, BIOPHYSICAL AND ECONOPHYSICAL SYSTEMS - Proceedings of the 22nd Canberra International Physics Summer School

stable distribution in Eq. (1.6) (extension of the Central Limit Theorem[13]).

3) Bounded distributions, which have no tails. They can be associated with $\alpha < 0$.

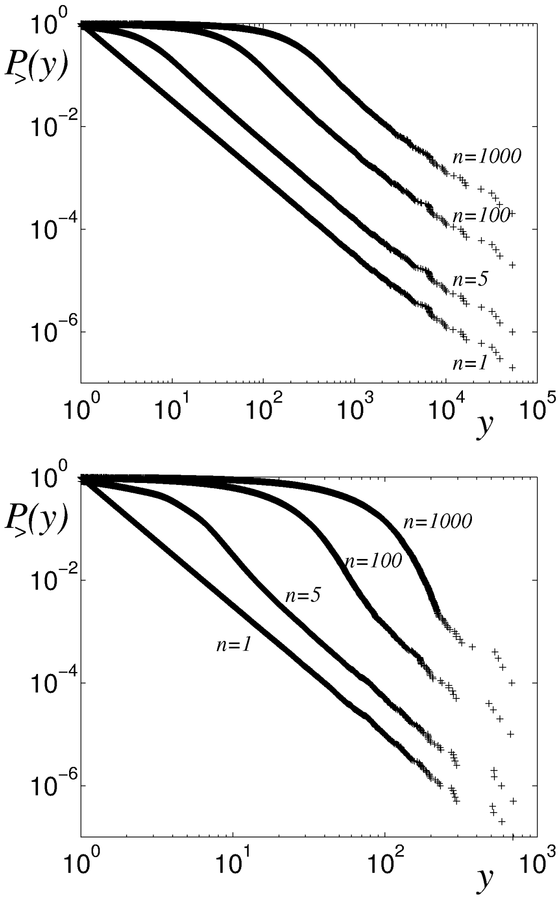

Fig. 1.3. Complementary cumulative distribution of the aggregate statistics resulting from a sum of n i.i.d. power law distributed variables. Specifically, we have, $y = \sum_{i=1}^{n} x_i$ with x_i independent random variables with probability distribution $p(x) = ax^{-\alpha-1}$. The top figure refers to the case $\alpha = 1.5$ whereas the bottom to the case $\alpha = 2.5$. Different aggregation sizes ($n = 1, 5, 100, 1000$) are shown.

1.5.4. *Sum of power-law-tailed distributions*

A very important consequence of the extreme value theorem is that the tails of a fat-tailed distribution (for i.i.d. processes) are invariant under addition even if the distribution as a whole is varying with aggregation. For instance, if we observe that daily returns are well fitted with a Student-t distribution,[13] then the Central Limit Theorem tells us that the monthly returns should be well approximated by a normal distribution and *not* a Student-t distribution. Yet the tails of the monthly returns are like the tails of the daily returns with the same exponent α. However, we must be aware that estimation of the tail exponent is not an easy task and a precise measurement of α requires a large sample size. This is why the use of data recorded every few seconds, or even tick by tick data (high frequency data), is highly recommended in this kind of analysis.[6]

Evidences of heavy tails in financial assets return distributions is plentiful ever since the seminal work of Mandelbrot on cotton prices.[16] However, the debate is still highly active and controversial, in particular on whether the second moment of the distribution of returns converges. Which requires to establish whether the exponent α is larger than 2 (σ defined) or smaller than 2 (σ not defined). It is clear that this question is central to many models in finance that specifically rely on the finiteness of the variance of returns. Indeed, as discussed in Section 1.4, there is a fundamental difference between additive i.i.d. processes with finite or infinite variance.

Let us here investigate further these differences with a simple example. We take the sum of n i.i.d. random variables x_i characterized by the following power-law probability density function:

$$p(x) = \frac{\alpha x_{\min}^{\alpha}}{x^{\alpha+1}} \ , \tag{1.10}$$

with $x \geq x_{\min}$ for some arbitrary $x_{\min} > 0$. Let us study the two cases $\alpha < 2$ and $\alpha > 2$. Figure 1.3 (top) shows that in the first case, for $\alpha = 1.5$, the distribution of the sum of the variables rests persistently a power law in most of the tail region (the complementary cumulative distribution decreases linearly in log-log scale) with the same exponent α of the individual random variables in Eq. (1.10) (which is the case $n = 1$ in Fig. 1.3). Indeed, in this case the distribution tends to the stable distribution [Eq. (1.6)] which behaves as a power law in the tail region $p_n(y) \to f(y) \sim y^{-\alpha-1}$.

We can see from Fig. 1.3 (bottom) that the second case, for $\alpha = 2.5$, is remarkably different. The shape of the distribution changes rapidly with the gathering of variables displaying a steeper decrease with x than for a

COMPLEX PHYSICAL, BIOPHYSICAL AND ECONOPHYSICAL SYSTEMS - Proceedings of the 22nd Canberra International Physics Summer School

power law distribution. Indeed, in this case, the Central Limit Theorem predicts a convergence of the aggregate distribution towards the normal one. However, in the tail region, below the Berry–Esseen convergence threshold, the extreme value theorem predicts a Fréchet distribution for the extreme variations and therefore we observe persistence of the power law trend in this extreme region. This is indeed evident from Fig. 1.3.

1.6. Capturing the tail

Have we already seen the worst or are we going to experience even larger losses? The answer of this question is essential for any good management of risk. We now have the instruments to answer this question. Indeed, we can apply the extreme value theory *outside* our sample to consider extreme events that have not yet been observed. To this purpose it is essential to be able to properly measure the tail index α.

1.6.1. *Power law in the tails*

We can see from the inset in Fig. 1.2, that the tails of the distributions of both the positive and negative fluctuations for the daily log-returns of the Ford Motor prices are decreasing linearly in log-log scale. This is an indication of a power law kind of behavior [i.e. $P_>(r) \sim ar^{-\alpha}$]. The main issue is how to quantify precisely the tail exponent α. There are several established methods to estimate the exponent α.[6,12] Let us here mention that a good practical rule is first to look qualitatively for the signatures of a linear trend in the log-log plot of $P_>(r)$ and afterwards, check the goodness of the quantitatively estimated α by comparing the data in the plot with the straight line from the power law function $ar^{-\alpha}$.

1.6.2. *Rank frequency plot*

Let us first point out that the plot of $P_>(r)$ in Fig. 1.2 is a so-called "rank-frequency" plot. This is a very convenient, and simple, method to analyze the tail region of the distribution without any loss of information which would instead derive from gathering together data points with an artificial binning. In order to make this plot one first sorts the n observed values in ascending order, and then plot them against the vector $[1, (n-1)/n, (n-2)/n, ..., 1/n]$. Indeed, for a given set of observations $\{x_1, x_2, ..., x_n\}$, we have that $\text{Rank}(x_i)/n = 1 - P_>(x_i)$.

A best fit of the exponent for the data in Fig. 1.2 reveals a value $\alpha \sim 2.4$. Values of exponents between 2 and 4 are very typical for these kinds of systems. These distributions typically have finite second moment but diverging fourth moment, and this is the reason why they reveal very high excess kurtosis.

1.7. Random Walks

So far we have discussed some relevant statistical features typically associated with complex system phenomena and in particular with stock price fluctuations. In this section we introduce a technique to model such fluctuations.

There are many factors that contribute to the "formation" of the price of a given equity and to its variation during time. This is, per se, the subject of several books and huge efforts have been dedicated to better understand this issue. From a very general and simple perspective we can say with some confidence that the price of a given equity is changing every time it is traded. At each successful transaction the price is fixed for that given time. A future transition will be processed at a marginally different price depending on the market expectations (rational or irrational) regarding that specific asset. This information is in part reflected in the bid and ask and their volumes on the order book and in part it lies in the mind and in the hearts of the human traders, and in the algorithms of the automatic traders as well.

1.7.1. *Price fluctuations as random walks*

An asset price model that has been widely used assumes that the logarithm of the asset price $x(t) = \log[\text{price}(t)]$ at a given time t results with some probability $p(\eta)$ at some value η above or below the logarithm of the price at the previous trading time. Formally we can write:

$$x(t + \Delta) = x(t) + \eta(t) \;\; ; \tag{1.11}$$

where $\Delta > 0$ is the time-step. Equation 1.11 defines a *random walk* which is a particular case of a stochastic process. Sometime the random variable η is called "noise."

Random walk kinds of processes have been widely used in modeling complex systems. The term "random walk" was first used by Karl Pearson in 1905. He proposed a simple model for mosquito infestation in a forest:

at each time step, a mosquito moves a fixed length at a randomly chosen angle. Pearson wanted to know the mosquitos distribution after many steps. The paper (a letter to Nature[17]) was answered by Lord Rayleigh, who had already tackled the problem for sound waves in heterogeneous materials. As a matter of fact, the theory of random walks was developed a few years before (1900) in the PhD thesis of a young economist: Louis Bachelier. He proposed the random walk as the fundamental model for financial time series. Bachelier was also the first to draw a connection between discrete random walks and the continuous diffusion equation. Curiously, in the same year as the paper of Pearson (1905), Albert Einstein published his paper on Brownian motion, which he modeled as a random walk, driven by collisions with gas molecules. Smoluchowski in 1906 also published very similar ideas.

Note that Eq. (1.11) assumes discrete time and uses equally spaced time-steps Δ. In reality, the market time is indeed not continuous since transactions are registered at discrete times. However, these transaction times are not equally spaced, having periods with high activity and others with a relatively small number of transactions. Furthermore, the price variations at two consecutive times might be related. For instance, in periods of large volatility (large price fluctuations) the size of $|\eta(t)|$ is likely to be consistently larger than average for extended periods of time (a phenomenon called volatility clustering). Generally speaking, Eq. (1.11) must be considered as a basic model, which has however the advantage of being relatively easy to treat both analytically and numerically. The model can then be extended to consider continuous time and/or non-uniform spacing between time-steps and/or time correlations.

One more specific question about the random walk model in Eq. (1.11) concerns the size of the discrete time step Δ. In the market a stock can be traded several times in a second, however there can be intervals of several seconds where the stock is not traded. This “granularity” of the trading time is difficult to handle; as a general rule we must consider Δ of the order of a few seconds. The exact value is not particularly relevant in the present context, but the order of magnitude is very important, as we shall see hereafter.

1.7.2. *Log-return as sum of random variables*

Given that $x(t)$ in Eq. (1.11) is the log-price, the log-returns are $r(t, \tau) = x(t + \tau) - x(t)$ [Eq. (1.1)] and they can be written as

$$r(t, \tau) = \sum_{s=0}^{\tau/\Delta - 1} \eta(s\Delta + t) \ . \tag{1.12}$$

They are therefore sums of $n = \tau/\Delta$ random variables and, if the $\eta(t)$ are i.i.d., the Central Limit Theorem must apply to $r(t, \tau)$. We have seen in Section 1.4 that we have two broad cases: (1) the probability distribution function of $\eta(t)$ has finite variance and therefore the distribution of $r(t, \tau)$ should approximate a normal distribution for large τ; (2) the variance is not defined and therefore the distribution of $r(t, \tau)$ should approximate a Levy Stable distribution for large τ. If we have fat-tailed distributions, as the ones described in Sections 1.5 and 1.6, then the parameter that distinguishes between these two classes is the tail index α. The case $\alpha \geq 2$ leads to normal distributions, whereas $\alpha < 2$ yields to Levy Stable distributions.

We have seen in our example with the Ford Motor data (Section 1.6.2), that in this specific example the tail index is best fitted with $\alpha \sim 2.4$ which is therefore larger than 2. In this case, the Central Limit Theorem tells us that a sum of n of these variables (where $\tau = n\Delta$) will converge towards a normal form and the Berry–Esseen theorem guarantees that this convergence is in $1/\sqrt{n}$. This implies that if we look for deviations from the normal statistics, we should explore the tail region where, roughly speaking, $P_>(x) < 1/\sqrt{n}$. Since in Fig. 1.2 we are reporting the statistics of daily returns, we have $\tau = 1$ day which corresponds to about 6 market hours and therefore ~ 22000 seconds, we expect to still observe power law behaviors in the tail region where $P_>(x) < 1/\sqrt{22000} \sim 10^{-2}$, which is indeed where the distribution starts to differ substantially from the normal statistics as one can clearly see in Fig. 1.2.

1.7.3. *High frequency data*

It is clear that a better estimate of the tail exponent can be obtained by reducing the interval τ, and this requires the use of infra-day data. Nowadays, there is a great availability of "high" frequency financial data, up to the whole order book, where every single bid, ask and transaction price is registered together with the volumes (amount of capital traded). However, to work with infra-day data poses some new technical challenges.[6]

For instance, the opening prices are affected by the events occurred during the closure and by the night electronic trading. The closure prices are also affected by the same reasons, in expectations. There are periods in the day that are very active and others that are instead pretty gloomy. For instance, large activities and sudden price variations are likely at times when other markets open or close. It is beyond the purposes of this Chapter to give any account of the so-called "seasonality" effects in the infra-day data. However, it is important that readers bear in mind that infra-day data should be handled with care. A highly recommended "minimal-trick" is to eliminate from the analysis data the first 20 min after opening and the last 20 min before closure.[6]

1.8. Scaling

We have seen that the statistical analysis of price fluctuations can be performed over different time scales. In general, the overall statistical properties of the log-returns are changing with the time interval τ. Indeed, there are different factors governing the variation at short- or long- time scales. However, it is also clear that the two must be related. Such a relation between the different probability distributions of the fluctuations at different time intervals is called *scaling* of the distribution.

The presence of large fluctuations and in particular power law noise with exponent $\alpha < 2$ affects dramatically the overall dynamics of the process and it is reflected in the *scaling* properties of the aggregate statistics. Let us first note that the log-returns $r(t, \tau)$ from the random walk process in Eq. (1.11) can be written as a sum of $n = \tau/\Delta$ noise terms [as explicitly shown in Eq. (1.12)]. Therefore, the changes of the statistical properties of $r(t, \tau)$ with τ (the so-called scaling of the distribution) correspond to the changes of the aggregate statistics of the sum of $n = \tau/\Delta$ i.i.d. variables. We know already from the previous discussion in Sections 1.4 and 1.6 that there is a difference in the aggregate statistics of random variables with finite or undefined variance. Such a difference is reflected in the diffusion dynamics of the random walker.

1.8.1. *Super-diffusive processes*

In additive stochastic processes [such as Eq. (1.12)] with fat-tailed noise (i.e. $p(\eta) \sim |\eta|^{-\alpha-1}$, with $0 < \alpha < 2$), all the motion is dominated by the large fluctuations and it results in a super-diffusive behavior where the mean square displacement increases faster than τ. Let us here give an heuristic simple derivation of this fact which shows clearly the origin of anomalous diffusion behavior in presence of power law noise.

Given a power law probability $p(\eta) \sim |\eta|^{-\alpha-1}$, the probability of a jump of size L which is larger or equal than a given $L_{\max}$ is given by the complementary cumulative distribution:

$$P(L \geq L_{\max}) = P_{>}(L_{\max}) = \int_{L_{\max}}^{\infty} p(\eta) d\eta \sim \frac{1}{L_{\max}^{\alpha}} \quad . \tag{1.13}$$

We can infer an idea of the time-dependence of $L_{\max}$ by noticing that if we monitor the process for an interval of time $\tau = n\Delta$ we will have a *finite* probability to observe a jump of size $L_{\max}$ if $nP(L \geq L_{\max}) \sim 1$ and therefore, from Eq. (1.13), $n/L_{\max}^{\alpha} \sim 1$, yielding

$$L_{\max} \sim \left(\frac{\tau}{\Delta}\right)^{1/\alpha} \quad . \tag{1.14}$$

We can now use the same argument to calculate the mean square displacement after $n = \tau/\Delta$ time steps: $E(r^2) - E(r)^2 = E(r^2) = nE(\eta^2)$ [having $E(r) = 0$]. When $0 < \alpha < 2$, we have $E(\eta^2) = \int_{L_{\min}}^{L_{\max}} \eta^2 p(\eta) d\eta \sim L_{\max}^{2-\alpha}$ and therefore $E(r^2) \sim nL_{\max}^{2-\alpha} = L_{\max}^2$. This indicates that the whole average movement in the process is the size of the largest jump. In other words, the evolution is entirely dominated by the largest jumps. By using Eq. (1.14) we have

$$E(r^2) \sim \left(\frac{\tau}{\Delta}\right)^{2/\alpha} \quad . \tag{1.15}$$

We see that for $0 < \alpha < 2$ the mean square displacement increases faster than τ and the system is “super-diffusive.” For $\alpha \geq 2$ the arguments above do not hold any longer and the mean square displacement grows linearly with τ as for any diffusive process.

1.8.2. *Sub-diffusive processes*

Let us here also mention that an opposite kind of scaling is observed when the mean square displacement increases slower than τ. This case is referred to as “sub-diffusive” behavior and it can be obtained from additive kinds of models when the time-step intervals between subsequent variations are unequally distributed following a power-law kind of distribution. It is also the result of time-correlated processes.

1.8.3. *Uni-scaling*

The random walk is a very simple and useful model to introduce and study stochastic processes. However, it must be stressed that most of the real stochastic processes are *correlated*, and therefore they are not random walks.[18] We have seen that random walk processes are associated with scaling laws that describe the way the distribution changes when the variables are aggregated. For instance for a stable process the probability distribution of the log-returns should scale with τ accordingly with:

$$p_\tau(r) = \left(\frac{\Delta}{\tau}\right)^{1/\alpha} p\left(\left(\frac{\Delta}{\tau}\right)^{1/\alpha} r\right) \quad . \tag{1.16}$$

This is a direct consequence of Eq. (1.7) and the fact that $r(t,\tau)$ is the sum of τ/Δ random variables [Eq. (1.12)]. Accordingly, the q-moments scale as $E(|r(t,\tau)|^q) = (\tau/\Delta)^{q/\alpha} E(|r(t,1)|^q)$. This is one particular form of scaling, which applies to stable distributions and is analogous to that discussed in the previous section.

More generally, in analogy with the previous scaling law, one can define a stochastic process where the probability distribution of $\{x(ct)\}$ is equal to the probability of $\{c^H x(t)\}$. Such a process is called self-affine.[19] In self-affine processes, with stationary increments, the q moments must scale as

$$E(|r(t,\tau)|^q) = E(|x(t+\tau) - x(t)|^q) = c(q)\left(\frac{\tau}{\Delta}\right)^{qH} \quad . \tag{1.17}$$

The parameter H is called the self-affine index or scaling exponent (or Hurst exponent—see chapter by Henry *et al.* in this volume). It is related to the fractal dimension by $D_f = D+1-H$, where D is the process dimensionality ($D = 1$ in our case). It is clear from Eq. (1.16) that in the case of stable distributions $E(|r(t,\tau)|^q) = E(|x(t+\tau) - x(t)|^q) = (\tau/\Delta)^{q/\alpha} E(|r(t,1)|^q)$ and we can identify $H = 1/\alpha$.

A process satisfying the scaling in Eq. (1.17) is called *uniscaling*.[19]

1.8.4. *Multi-scaling*

The kinds of processes encountered in finance, and in complex systems in general, often scale in an even more complicated way indicating that the kind of observed scaling is not simply a fractal. In contrast to the more conventional fractals, in these processes we need more than one fractal dimension depending on the aspect we are considering. In order to properly model real world processes we must use *multiscaling* processes where, for stationary increments, the q moments scale as[19–21]

$$E(|x(t+\tau) - x(t)|^q) = c(q) \left(\frac{\tau}{\Delta}\right)^{qH(q)} . \tag{1.18}$$

with $H(q)$ a function of q (for $q > -1$). The function $qH(q) - 1$ is called the scaling function of the multi-scaling process and it is in general concave.

Most of the processes in financial systems are multiscaling. This implies that the moments of the distribution scale differently according to the time horizons (i.e. the distribution of the returns changes its shape with τ) revealing that the system properties at a given scale might not be preserved at another scale.

1.9. Complex Networked Systems

Let us now continue our "navigation" by introducing a new fundamental factor which concerns the collective dynamics of the system. As we already mentioned in the introduction, studies of complex systems have ranged from the human genome to financial markets. Despite this breadth of systems—from food chains to power grids, or voting patterns to avalanches—a unifying and characterizing aspect has emerged: all these systems are comprised of many interacting elements. In recent years, it has become increasingly clear that in complex system studies it is of paramount importance to analyze the dynamics of all the elements highlighting their emerging collective properties.

The study of collective dynamics requires the simultaneous investigation of a large number of different variables. So far, in this Chapter, we have focused on the complex behavior of a *single* element (i.e. the price of a given stock), however it is clear that any information from such an individual property is meaningless if not properly compared with the information from the rest of the elements constituting the system (i.e. all the other stocks in the market). Typically, in these systems each element is not evolving in isolation and therefore the collective dynamics is reflected in the individual behavior as much as the individual changes affect the global variations. The understanding of the properties of such a network of interactions and co-variations is one of the key elements to understand complex systems. Indeed, one of the most significant breakthroughs in complex systems studies has been the discovery that all these systems share similar structures in the network of interactions between their constitutive elements.[22–24]

1.9.1. *Scale-free networks*

It results that in a large number of complex systems the probability distribution of the number of contacts per vertex (called *degree distribution* $p(k)$) is "fat tailed" with $p(k) \sim k^{-\alpha-1}$, with exponents α typically ranging between 1 and 2.[25] Such networks are widespread including internet, world wide web, protein networks, citation networks, world trade network etc.

It is important to stress that these distributions are different from what it would result by randomly connecting pair of vertices in a *random graph* which will instead yield a Poissonian distribution with exponentially-fast-decreasing tails. The power law degree distribution implies that there is no "typical" scale and all scales of connectivity are represented. These "scale-free" networks are characterized by a large number of poorly connected vertices but also by a few very highly connected "hubs."[25]

1.9.2. *Small and ultra-small worlds*

The properties of such networks are also different from the properties of regular or random networks. For instance, the average distance ($\langle d \rangle$) between two vertices scales with the total number of vertices in the network (V) as $\langle d \rangle \sim \log \log V$ ($1 < \alpha < 2$). This means that the network is very compact and only very few steps are necessary to pass from an individual to another. This property is called "ultra small world" in contrast with the "small world" property where $\langle d \rangle \sim \log V$, which holds for $\alpha \geq 2$ or for random networks ($\alpha \to \infty$). In contrast, regular lattices in D dimensions are "large worlds" with $\langle d \rangle \sim V^{1/D}$. Without entering into any detail, it should be quite clear to anyone just from an intuitive perspective that the difference between a large world and an ultra small world is huge, and can have very dramatic implications. For instance, the dynamical properties, such as the rate of spreading of a disease through the network, are strongly affected by the network structure. A world pandemic on a "large world" contact network will require a very long chain made of hundreds of thousands of contacts to infect all the individuals from a single source. By contrast, on an "ultra small world" it would take a chain of only a few steps to infect all the individuals in the world. Alarmingly, it has been shown[26] that the airport-network system has a small world structure.

1.9.3. *Extracting the network*

The extraction of the network of interrelations associated with a given complex system can be a challenging task. Indeed, except for a few cases where the network has an unambiguously given property, there are a large number of other cases where the network is not as clearly defined. The network of relations between elements of the systems can have *weights*, can be *asymmetric*, there might be *missed* unknown links, there might be asynchronous interactions, feedback and fluctuations. In this section we investigate a specific case which is rather general and widespread in the study of these systems.

We consider a system with a large number of elements where, a priori, every element can be differently affected by any other and we aim to infer the network of most relevant links by studying the mutual dependencies between elements from the analysis of the collective dynamics. In other words, we search for variables that behave similarly and we want to link them with edges in the network. Conversely, we do not want to directly connect the variables that behave independently. To this purpose we must first look at methods to define and quantify dependency among variables.

1.9.4. *Dependency*

Generally speaking, the mutual dependence between two variables x and y should be measurable from the "difference" between the probability to observe them simultaneously and the probability to observe them separately. Let us call $P(X \leq x, Y \leq y)$ the joint cumulative distribution to observe both the values of the variables X and Y to be less than or equal to two given values x and y. We must compare this joint probability with the marginal cumulative probabilities $P_X(X \leq x)$ and $P_Y(Y \leq y)$ of observing the variables independently from each other. A theorem[12] guarantees us that two random variables X and Y are *independent* if and only if:

$$P(X \leq x, Y \leq y) = P_X(X \leq x) P_Y(Y \leq y) \ . \tag{1.19}$$

This identity reflects the intuitive fact that when the variables are independent the occurrence of one event must make it neither more nor less probable that the other occurs. Therefore, given two variables X and Y, the difference, or the distance between the joint cumulative probability $P(X \leq x, Y \leq y)$ and the product of the two marginal cumulative probabilities $P_X(X \leq x)$ and $P_Y(Y \leq y)$, should be a measure of dependence between the two variables. Indeed, when $P(X \leq x, Y \leq y) > P_X(X \leq$

$x)P_Y(Y \leq y)$ we have the so called *positive quadrant dependency*, which expresses the fact that "when X is large, Y is also likely to be large." Conversely, $P(X \leq x, Y \leq y) < P_X(X \leq x)P_Y(Y \leq y)$ we have the so called *negative quadrant dependency*, which expresses the fact that "when X is large, Y is likely to be small." One simple quantification of such a measure of dependency is the *covariance*:

$$\text{Cov}(X,Y) = \iint [P(X \leq x, Y \leq y) - P_X(X \leq x)P_Y(Y \leq y)]\, dxdy \,. \qquad (1.20)$$

A positive covariance indicates that the two variables are likely to behave similarly whereas a negative covariance indicates that the two variables tend to have opposite trends. However, it should be stressed that this is a measure of linear dependency and there are nonlinear cases where dependent variables have zero covariance [e.g. $y = x^2 - 1$, with $E(x) = 0$ and $E(x^2) = 1$].

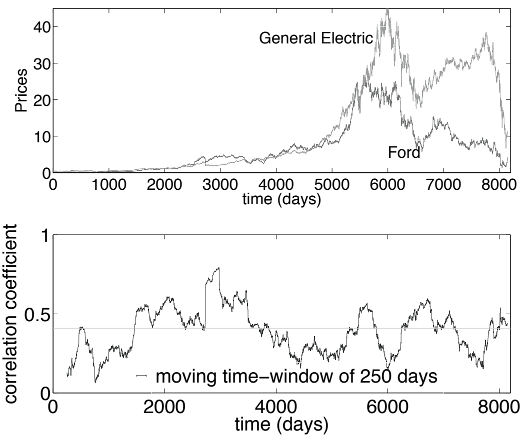

Fig. 1.4. (Top) Daily closing adjusted prices for the two stocks Ford Motor and General Electric in the New York Stock Exchange. The time-period ranges from Jan 3 1977 to April 7 2009. (Bottom) Correlation coefficient calculated over a moving window of 250 days ($\sim$ 1 year). The horizontal line is the correlation coefficient calculated over the whole period.

1.9.5. *Correlation coefficient*

A measure of dependency directly proportional to the covariance is the Pearson product-moment correlation coefficient $\rho_{i,j}$. Given two random variables x_i and x_j with expectation values μ_i and μ_j, and standard deviations σ_i and σ_j, the correlation coefficient is defined as

$$\rho_{i,j} = \frac{\mathrm{Cov}(x_j, x_j)}{\sigma_i \sigma_j} = \frac{E\left((x_i - \mu_i)(x_j - \mu_j)\right)}{\sigma_i \sigma_j} \quad . \tag{1.21}$$

Analogously to the covariance, positive values for $\rho_{i,j}$ indicate that the two variables are likely to behave similarly, whereas negative $\rho_{i,j}$ indicate that the two variables tend to have opposite trends. The correlation coefficient has however the advantage of being bounded between $[-1, 1]$, with the two limits corresponding to perfectly anti-correlated and perfectly correlated variables. For example, one can verify that the two variables $x_j = a + bx_i$ have $\rho_{i,j} = b/|b|$ giving $\rho_{i,j} = +1$ when $b > 0$ and $\rho_{i,j} = -1$ when $b < 0$.

1.9.6. *Significance*

In practice, the correlation coefficient is estimated over a finite set of data points: the time series $x_i(t)$ with $t = t_0 + s\Delta$ with $s = 1, 2, ...T$. The Pearson estimator $\rho_{i,j}$ is calculated from Eq. (1.21) by substituting the expectation values $E(...)$ with the sample averages $\langle(...)\rangle$ and by using the sample means and standard deviations. Clearly, the smaller the observation time T, the larger will be the inaccuracy on the estimated coefficient.

The use of the correlation coefficient to measure dependence between variables is very common and widespread, and it turns out to be a very efficient measure in a large number of domains. However, this measure can be very problematic and it might sometimes lead to serious faults. We have already mentioned that, in nonlinear cases, completely dependent variables can have zero covariance and consequently zero correlation coefficient. Other problems might arise with non-normally distributed variables. Indeed, we already noticed that the standard deviation is not defined for random variables with fat-tailed power law distributions and tail exponent smaller than or equal to 2. This implies that for these variables the correlation coefficient is not defined as well. Moreover, when the tail index belongs to the interval $\alpha \in (2, 4]$, the correlation coefficient exists but its Pearson estimator is highly unreliable because its distribution is fat tailed, with undefined second moments, and therefore it can have unbounded large variations.

Moreover, in complex systems studies we are often observing systems that are not stationary and the interrelations between the elements are themselves changing during the observation time. As a general rule, one must assume that these changes happen on a longer time scale than the one within which the correlations are measured. Therefore, T should not be too small, in order to improve the statistics, but it should not be too long either, in order to avoid being influenced by the long-term changes.

A practical example is given in Fig. 1.4 (top) where the historical data for Ford Motor (same as in Fig. 1.1) are plotted together with the data for General Electric. One can see that there are similarities and differences. The cross correlation coefficient for the log-returns over the entire period is $\rho \sim 0.4$. On the other hand, Fig. 1.4 (bottom) shows that the values over sub-periods calculated on a moving window of 1 year (~ 250 days) fluctuate around this value, showing significant variations depending on the market evolution.

1.9.7. *Building the network*

In practice, we often have more than two variables and the dependency problem is in general a high dimensional challenge. However, the extension from two to n variables is straightforward, with the exception that the names change and the joint distribution takes the name of *multivariate distribution* when $n > 2$ (bivariate for $n = 2$). As far as we are interested in the dependencies between couples of variables x_i and x_j we can apply straightforwardly Eq. (1.21) to each couple of variables, obtaining an $n \times n$ correlation matrix which is symmetric and has all ones on the diagonal. We have therefore $n(n-1)/2$ distinct entries.

Let us here concentrate on one precise example: the system of companies quoted on the same equity market as Ford Motor (the NYSE) and, for simplicity, let us take only a set of 100 among the most capitalized ones. Even with such a reduction, the system of correlations between the various stock prices has $4,950$ entries, most of which are redundant. In terms of the network of interactions we are now looking at the *complete graph* where every node is connected by one *weighted* edge to every other node in the network. We must simplify such a system extracting only a subset of the most relevant interrelations: the "backbone" network.

COMPLEX PHYSICAL, BIOPHYSICAL AND ECONOPHYSICAL SYSTEMS - Proceedings of the 22nd Canberra International Physics Summer School

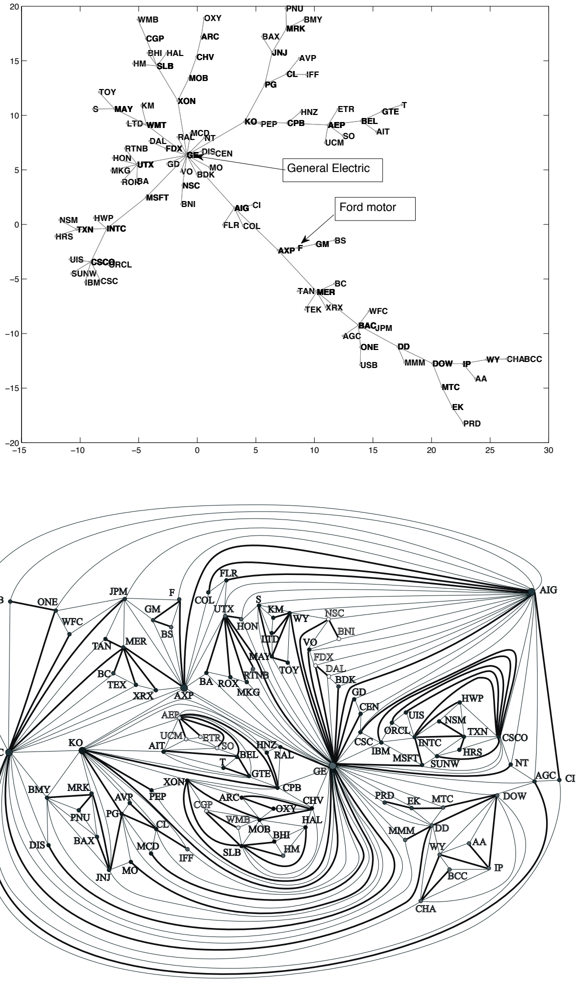

Fig. 1.5. (top) Minimum Spanning Tree network built from the cross-correlation matrix for 100 stocks in the US equity market. (bottom) Planar Maximally Filtered Graph built for the same cross-correlation data.

1.9.8. *Disentangling the network: minimum spanning tree*

We want to build a network whose topological structure represents the correlation among the different elements. All the important relations must be represented, but the network should be as "simple" as possible. The simplest connected graph is a spanning tree (a graph with no cycles that

COMPLEX PHYSICAL, BIOPHYSICAL AND ECONOPHYSICAL SYSTEMS - Proceedings of the 22nd Canberra International Physics Summer School

connects all vertices). It is therefore natural to choose as representative network a spanning tree which retains the maximum possible correlations. Such a network is called the Minimum Spanning Tree (MST).

There are several algorithms to build a MST, the two most common being Prim's algorithm[27] and Kruskal's algorithm[28] both from the '50s, but there are also older ones. Remarkably, there are also very recently discovered ones such as the one proposed by Chazelle[29] in the year 2000, which is, so far, the algorithmically most efficient, running in almost linear time with the number of edges.

The general approach for the construction of the MST is to connect the most correlated pairs while constraining the network to be a tree. Let us here describe a very simple algorithm (similar to Kruskal's) which is very intuitive and will help to clarify the concept.

(1) Make an ordered list of pairs i, j, ranking them by decreasing correlation $\rho_{i,j}$ (the largest first and the smallest last).
(2) Take the first element in the list and add the edge to the graph.
(3) Take the next element and add the edge if the resulting graph is still a forest or a tree, otherwise discard it.
(4) Iterate the process from step 3 until all pairs have been exhausted.

The resulting MST has $n - 1$ edges and it is the spanning tree that maximizes the sum of the correlations over the connected edges.

The resulting network for the case of the 100 stocks quoted in the NYSE studied during the period 3 January '95 to 31 December '98 is shown in Fig. 1.5.[30,31] We can see that in the MST the stock Ford Motor (F) is linked to the stock General Motor (GM). They form a separate branch together with Bethlehem Steel (BS), and the branch is attached to the main "trunk" through the financial services provider American Express (AXP). This structure of links that we have here extracted with the MST is economically very meaningful because we know that cars need steel to be built and consumers need credit from financial companies to buy the cars. What is remarkable is that these links have been extracted from the cross-correlation matrix without any a priori information on the system.

It is clear that the same method can potentially be applied to a very broad class of systems, specifically in all cases where a correlation (or even, more simply, a similarity measure) between a large number of interacting elements can be assigned.

1.9.9. *Disentangling the network: planar maximally filtered graph*

Although we have just shown that the MST method is extremely powerful, there are some aspects that might be unsatisfactory. In particular the condition that the extracted network should be a tree is a strong constraint. Ideally, one would like to be able to maintain the same powerful filtering properties of the MST but also allowing the presence of cycles and extra links in a controlled manner.

A recently proposed solution consists in building graphs embedded on surfaces with given genus.[32] (Roughly speaking the genus of a surface is the number of holes in the surface: $g = 0$ corresponds to the embedding on a topological sphere; $g = 1$ on a torus; $g = 2$ on a double torus; etc.) The algorithm to build such a network is identical to the one for the MST discussed previously except that at step 3 the condition to accept the link now requires that the resulting graph must be embeddable on a surface of genus g. The resulting graph has $3n-6+6g$ edges and it is a triangulation of the surface. It has been proved that the MST is always a subgraph of such a graph.[31]

It is known that for large enough genus any network can be embedded on a surface. From a general perspective, the larger the genus, the larger is the complexity of the embedded triangulation. The simplest network is the one associated with $g = 0$ which is a triangulation of a topological sphere. Such planar graphs are called Planar Maximally Filtered Graphs (PMFG).[31] PMFG have the algorithmic advantage that planarity tests are relatively simple to perform.

The PMFG network for the case of the 100 stocks studied previously is reported in Fig. 1.5.[31] We can observe that in this network Ford Motor (F) acquires a direct link with Bethlehem Steel (BS), it acquires a new link with the bank JPMorgan Chase (JPM) and it also acquires a link with the very influential insurance services American International Group (AIG). We note that F, AXP and BS form a 3-clique (a triangle), which becomes a 4-clique (a tetrahedron)[c] by adding GM. As one can see from Fig. 1.5, the PMFG is a network richer in links and with a more complex structure than the MST, of which it preserves and expands some hierarchical properties.

[c]An r-clique is a complete graph with r vertices where each vertex is connected with all the others.

1.10. Conclusion

In conclusion, in this Chapter we have introduced a few fundamental notions useful for the study of complex systems, with the aim of providing a sort of referential "navigation map." We have presented and discussed a large number of concepts, theorems and topics. However, the most important aspects that we have treated under different perspectives can be summarized in the following two points:

1) the ubiquitous presence in these systems of fat-tail probability distributions and their effects on the statistics of extreme events and on the scaling properties of additive processes;
2) the importance in these systems of the collective cross-correlated dynamics and the need of novel investigation techniques which combine statistical methods with network theory.

Let us here conclude by summarizing all the different aspects discussed in this Chapter by including them within a single, compact formula that is a sort of *constitutive equation for complex systems*:

$$x_i(t_{s+1}) = x_i(t_s) + \eta_i(t_s) + \sum_{u=0}^{s} \sum_{k=1}^{n} J_{i,k}(t_s, t_u) x_k(t_u) \quad , \tag{1.22}$$

with $J_{i,k}$ an exchange matrix associated with a weighted, directed network of interactions between the variables. This equation describes a process where n variables $x_1(t), ..., x_n(t)$ start at t_0 with values $x_i(t_0) = x_i^0$ and evolve in time through the time points $t_1, t_2, ...$, which are not necessarily equally spaced. The term $\eta_i(t_s)$ is an additive noise equivalent to the one described in Eq. (1.11). On the other hand, the last term describes the interaction between variables and can be used to introduce multiplicative noise, feedback and autocorrelation effects. All the characterizing features of complex systems that we have been discussing in this Chapter can be modeled and accounted by means of Eq. (1.22). This equation is compact but not simple and cannot be solved in general. On the other hand, it is rather straightforward to implement numerically and can be applied to model a very broad range of systems. Indeed, equations of the form of Eq. (1.22) have been proposed in the literature to model very different kinds of complex systems, from the spread of epidemics to productivity distribution. However, to our knowledge, the great generality and the wide applicability of such an equation has not so far been pointed out.

There are of course several other important topics, methods and techniques very relevant in the study of complex systems but which we have been unable to squeeze inside this Chapter. In this respect, the rest of this book provides a great example of the variety, diversity and breadth of this fast-expanding field.

Acknowledgments

We wish to acknowledge the strong support from the ARC network RN0460006. This work was partially supported by the ARC Discovery Projects DP0344004 (2003), DP0558183 (2005) and COST MP0801 project.

References

1. A. Einstein and L. Infield, *The Evolution of Physics: The Growth of Ideas from Early Concepts to Relativity and Quanta*. (Simon and Schuster, New York, 1938).
2. G. Parisi, Complex systems: a physicist's viewpoint, *Physica A*. **263**, 557–564, (1999).
3. N. Boccara, *Modeling Complex Systems*. (Springer-Verlag, Berlin, 2004).
4. fURL `www.santafe.edu/research/topics-physics-complex-systems.php`.
5. R. N. Mantegna and H. E. Stanley, *An Introduction to Econophysics: Correlations and Complexity in Finance*. (Cambridge University Press, Cambridge, UK, 2000).
6. M. M. Dacorogna, R. Gençay, U. A. Müller, R. B. Olsen, and O. V. Pictet, *An Introduction to High Frequency Finance*. (Academic Press, San Diego CA, 2001).
7. T. Di Matteo, T. Aste and S. T. Hyde. Exchanges in complex networks: income and wealth distributions. In *The Physics of Complex Systems (New advances and Perspectives)*, Proceedings of the International School of Physics "Enrico Fermi." Eds. F. Mallamace and H. E. Stanley, 435–442 (IOS Press, Amsterdam 2004).
8. P. Ball, Culture crash, *Nature*. **441**, 686–688, (2006).
9. M. Gallegati, S. Keen, T. Lux, and P. Ormerod, Worrying trends in econophysics, *Physica A*. **370**, 1–6, (2006).
10. T. Di Matteo and T. Aste, "No Worries": Trends in Econophysics, *Eur. Phys. J. B*. **55**, 121–122, (2007).
11. T. Di Matteo and T. Aste, Econophysics Colloquium, *Physica A*. **370**, xi–xiv., (2006).
12. E. Platen and D. Heath, *A Benchmark Approach to Quantitative Finance*. (Springer-Verlag Berlin, 2006).
13. W. Feller, *An Introduction to Probability Theory and Its Applications*. (Wiley, New York, 1968).

14. J. P. Nolan. URL `academic2.american.edu/~jpnolan/stable/stable.html`.
15. J.-P. Bouchaud and M. Potters, *Theory of Financial Risk and Derivative Pricing: From Statistical Physics to Risk Management (2nd ed.)*. (Cambridge University Press, Cambridge, UK, 2003).
16. T. V. of Certain Speculative Prices, B. B. Mandelbrot, *Journal of Business*. **36**, 394–419, (1963).
17. K. Pearson, The problem of the random walk, *Nature*. **72**, 294–342, (1905).
18. A. W. Lo and A. C. MacKinlay, *A Non-Random Walk Down Wall Street*. (Princeton University Press, 2001).
19. T. Di Matteo, Multi-scaling in finance, *Quantitative Finance*. **7**, 21–36, (2007).
20. T. Di Matteo, T. Aste, and M. M. Dacorogna, Scaling behaviors in differently developed markets, *Physica A*. **324**, 183–188., (2003).
21. T. Di Matteo, T. Aste, and M. M. Dacorogna, Long term memories of developed and emerging markets: using the scaling analysis to characterize their stage of development, *Journal of Banking & Finance*. **29/4**, 827–851, (2005).
22. M. E. J. Newman, The structure and function of complex networks, *SIAM REVIEW*. **45**, 167–256, (2003).
23. G. Caldarelli, *Scale-Free Networks Complex Webs in Nature and Technology*. (Oxford Univesity Press, 2007).
24. L. Amaral and J. Ottino, Complex networks, *Eur. Phys. J. B*. **38**, (2004).
25. R. Albert and A.-L. Barabási, Statistical mechanics of complex networks, *Rev. Mod. Phys.* **74**, 47–97, (2002).
26. R. Guimerà, S. Mossa, A. Turtschi, and L. A. N. Amaral, The worldwide air transportation network: Anomalous centrality, community structure, and cities' global roles, *Proceedings of the National Academy of Sciences of the United States of America*. **102**(22), 7794–7799, (2005). doi: 10.1073/pnas.0407994102. URL `http://www.pnas.org/content/102/22/7794.abstract`.
27. R. C. Prim, Shortest connection networks and some generalizations, *Bell System Technical Journal*. **36**, 1389–1401, (1957).
28. J. B. Kruskal, On the shortest spanning subtree of a graph and the traveling salesman problem, *Proceedings of the American Mathematical Society*. **7**, 48–50, (1956).
29. B. Chazelle, A minimum spanning tree algorithm with inverse-ackermann type complexity, *Journal of the ACM (JACM)*. **47**, 1028–1047, (2000).
30. R. N. Mantegna, Hierarchical structure in financial markets, *Eur. Phys. J. B*. **11**, 193–197, (1999).
31. M. Tumminello, T. Aste, T. Di Matteo, and R. N. Mantegna, A tool for filtering information in complex systems, *PNAS*. **102**, 10421–10426, (2005).
32. T. Aste, T. Di Matteo, and S. T. Hyde, Complex networks on hyperbolic surfaces, *Physica A*. **346**, 20–26, (2005).