

**TECHNISCHE
UNIVERSITÄT
DRESDEN**

Modeling and predicting time series of social activity with fat-tailed distributions

Dissertation
zur Erlangung des wissenschaftlichen Grades
Doctor rerum naturalium

vorgelegt von

Lic. José María Miotto
geboren am 25.10.1986 in Mailand, Italien.

erstellt am
Max-Planck-Institut für Physik komplexer Systeme
Dresden
2016

Eingereicht am 19 Mai 2016.

Verteidigt am 17 August 2016.

Gutachter:

Prof. Dr. Holger Kantz

Prof. Dr. Roland Ketzmerick

Prof. Dr. Joachim Peinke

Abstract

Fat-tailed distributions, characterized by the relation $\mathbb{P}(x) \propto x^{-\alpha-1}$, are an emergent statistical signature of many complex systems, and in particular of social activities. These fat-tailed distributions are the outcome of dynamical processes that, contrary to the shape of the distributions, is in most cases are unknown. Knowledge of these processes' properties sheds light on how the events in these fat tails, i.e. extreme events, appear and if it is possible to anticipate them. In this Thesis, we study how to model the dynamics that lead to fat-tailed distributions and the possibility of an accurate prediction in this context. To approach these problems, we focus on the study of attention to items (such as videos, forum posts or papers) in the Internet, since human interactions through the online media leave digital traces that can be analyzed quantitatively. We collected four sets of time series of online activity that show fat tails and we characterize them. Of the many features that items in the datasets have, we need to know which ones are the most relevant to describe the dynamics, in order to include them in a model; we select the features that show high predictability, i.e. the capacity of realizing an accurate prediction based on that information. To quantify predictability we propose to measure the quality of the optimal forecasting method for extreme events, and we construct this measure. Applying these methods to data, we find that more extreme events (i.e. higher value of activity) are systematically more predictable, indicating that the possibility of discriminate successful items is enhanced. The simplest model that describes the dynamics of activity is to relate linearly the increment of activity with the last value of activity recorded. This starting point is known as proportional effect, a celebrated and widely used class of growth models in complex systems, which leads to a distribution of activity that is fat-tailed. On the one hand, we show that this process can be described and generalized in the framework of Stochastic Differential Equations (SDE) with Normal noise; moreover, we formalize the methods to estimate the parameters of such SDE. On the other hand, we show that the fluctuations of activity resulting from these models are not compatible with the data. We propose a model with proportional effect and Lévy-distributed noise, that proves to be superior describing the fluctuations around the average of the data and predicting the possibility of an item to become an extreme event. However, it is possible to model the dynamics using more than just the last value of activity; we generalize the growth models used previously, and perform an analysis that indicates that the most relevant variable for a model is the last increment in activity. We propose a new model using only this variable and the fat-tailed noise, and we find that, in our data, this model is superior to the previous models, including the one we proposed. These results indicate that, even if present, the relevance of proportional effect as a generative mechanism for fat-tailed distributions is greatly reduced, since the dynamical equations of our models contain this feature in the noise. The implications of this new interpretation of growth models to the quantification of predictability are discussed along with applications to other complex systems.

Abbreviations and most used variables

$X \sim Y$ Random variable X is Y -distributed.

\mathbb{P} Probability.

f_X Probability density function of the variable X .

\mathbb{E} Expectation.

\mathbb{V} Variance.

α Index of fat-tailed distributions ($\mathbb{P}(X > x) \propto x^{-\alpha}$).

A Alarm.

E Event.

Π Predictability.

X_t Activity aggregated up to time t .

dX_t Increments in activity ($X_{t+1} - X_t$).

W_t Wiener process.

L_t Lévy-stable process.

AR Autoregressive model.

k -NN k Nearest Neighbors method.

Distributions:

GP Generalized Pareto.

S Lévy-stable.

N Normal.

LN Lognormal.

CEV Constant Elasticity of Variance.

Contents

1. Introduction	1
2. Fat-tailed distributions in social activity: evidence and models	5
2.1. Distributions with fat tails	5
2.1.1. Definitions	5
2.1.2. Fat-tailed distributions in nature	6
2.1.3. Fat-tailed distributions in social activity	7
2.1.4. Social activity in Online Media	8
2.1.5. Measuring attention	9
2.2. General arguments for the prevalence of fat tails	10
2.2.1. Fat-tailed distribution as attractors of stochastic processes	11
2.2.2. Mechanistic Models for fat-tailed distributions	13
2.3. Estimation of the Proportional effect	15
2.4. Forecasting social activity	18
3. Datasets used and relevant properties	21
3.1. Characterization of data	21
3.2. Fit of fat-tailed distributions	23
3.3. YouTube	24
3.4. Stack Overflow	25
3.5. Usenet	26
3.6. PLOS ONE	27
4. Extreme events predictability	29
4.1. Introduction	29
4.2. Robust estimation and Extreme Events	30
4.3. Predictability of Events	31
4.4. Proof that strategy LD (Bayes classifier) is dominant	34
4.5. Computation of Predictability for the optimal strategy	35
4.6. Application to Data	35

4.7. Discussion	38
4.7.1. Dependence of Predictability with respect to extreme events	38
4.7.2. Probabilistic Forecast	39
4.7.3. Conclusions	40
5. Stochastic dynamics of growth processes	43
5.1. Introduction	43
5.2. Allocation model of Proportional Effect	45
5.2.1. Solution of the Allocation model	45
5.3. Stochastic Differential Equations framework	49
5.3.1. Solution of the SDE with proportional effect: Lognormal and CEV distributions	51
5.3.2. A SDE model with Lévy fluctuations	52
5.4. Estimation of the SDE parameters	53
5.4.1. Averaging estimation	53
5.4.2. Estimation by Maximum Likelihood	55
5.5. Model selection	58
5.6. Forecasts of popularity	60
5.6.1. Optimal deterministic forecast	61
5.6.2. Predictability of big hits	62
5.6.3. Correlations in the S model	63
5.6.4. Dependence of correlation with respect to lag and time	65
5.7. Discussion and Conclusions	66
6. Generalized growth models	69
6.1. Introduction	69
6.2. Forecasting popularity with longer histories	69
6.2.1. Autoregressive models	70
6.2.2. k -Nearest Neighbors algorithm	72
6.2.3. Results	74
6.2.4. Summary	77
6.3. A new dynamical model: the Daily model	78
6.3.1. Parametric proposal of the Daily model	79
6.3.2. Estimation by Maximum Likelihood	80
6.4. Forecast of popularity	82
6.4.1. Prediction of big hits	83
6.5. Correlations of the models' fluctuations	84
6.6. Discussion and Conclusions	86

7. Conclusion	89
7.1. Summary of the results	89
7.2. Discussion and outlook	91
7.3. Open issues and directions for future work	93
Appendices	97
A. Computation of the Likelihood	97
B. Scaling of fluctuations is incompatible with Wiener noise	99
C. Packages released	102
C.1. datagram	102
C.2. pyLevy	102
List of figures	104
Bibliography	117

1. Introduction

The present Thesis is concerned with the challenges to dynamical models and prediction posed by fat-tailed distributions. In particular, we focus on the implications for these two problems that this statistical feature has in time series of social activities, where it is typical. This introductory chapter motivates the importance of understanding the causes and effects of fat tails in physical and social systems.

Fat-tailed distributions are a prominent pattern in the statistics of complex systems. There is a variety of distributions that are fat-tailed, but in general they are approximated by $\mathbb{P}(X = x) \propto x^{-\alpha-1}$ for large x . From a purely theoretical point of view, such distributions will have no defined n -th moment, $\langle X^n \rangle$, for any $n > \alpha$, which is very relevant if we consider that moments are usually used to characterize distributions. From the point of view of statistical physics, fat tails are a characteristic feature of a system at a critical state, e.g. the distribution of clusters' size in a second order phase transition like in percolation [SA94], or when the system is far from equilibrium, e.g. distribution of turbulent wind gusts [BRWP03, KHRV04] or earthquakes' magnitude [GR56]. In the study of social activity, these distributions are even more common, and have been used to describe a wide range of phenomena, from the distribution of population in cities to the distribution of citations that papers accrue [Gab99, Red98]. This ubiquity in both social and natural systems generated in the last two decades an ever increasing interest in topics related to heavy tails, which range from establishing statistically the presence or absence of such tails in empirical data [GMY04, CSN09] to studying their consequences in each of the fields where it was found, and in particular in the field of complex networks [BA99, ALPH01].

The importance of understanding fat-tailed distributions in social activity can be hardly overestimated. As an example, consider the *income distribution*, the frequency of people with a given amount of income; this distribution was first studied by Vilfredo Pareto in 1896, who observed that it was fat-tailed [Par96]. This implies that few people concentrate a large amount of wealth, i.e. there is a strong (likely undesirable) inequality, with a vast impact in the organization of economy and society as a whole. Pareto himself called the group of people in the far end of the tail the *elite*, and considered that society was ruled always by a particular elite, claiming that changes in history were fundamentally caused by changes in the dominant cliques of society (theory of the circulation of elites). However, the starting point of Pareto (his *null model*) was that the distribution of income should have been approximately random, and by random he meant *Normally distributed*. Although Pareto could not conceive a form of randomness other than the Gaussian

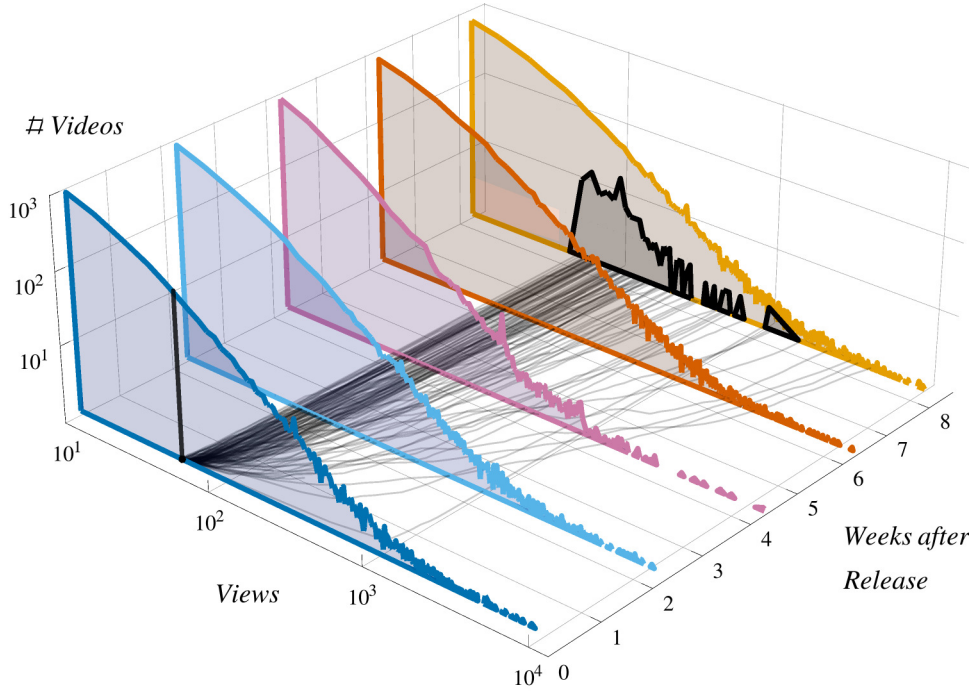


Figure 1.1.: Dynamics of views in YouTube. Colored histograms: distributions of videos' views at fixed times after publication (0.3 million videos from our database). Gray lines at the bottom: trajectories of 120 videos which had the same early success (50 views 2 days after publication). Black histogram: distribution of views of the 120 selected videos 2 months after publication.

error distribution, the question of the nature of the fat tail remains, namely, up to which point the distribution of income (or any other) is the result of a random process? How does this distribution changes? Is it possible for an individual to move within its ranks? Is not difficult to see that the answers to these questions have an implication on the interpretation of how society works.

Fat tails are then associated with *extreme values*, rare items or events that display a disproportionate importance in the system [AJK06]. Since these items are rare, two main issues must be addressed:

- how do these extreme values appear?
- is it possible to anticipate them?

These two questions define the themes of this Thesis, namely the *models* and *prediction* in social activity with fat tails. In order to address them, we have to go beyond the statistical analysis and understand the dynamical process that leads to these distributions.

The most celebrated dynamical models of ensembles that lead to fat-tailed distributions are the growth processes with proportional effect, i.e. models where the variable X of an item grows with time, such that the increment dX is proportional to X . These models are almost as old as

the observation of fat tails itself, with Gibrat proposing it for the income distribution studied by Pareto in 1931 [Gib30]. The argument of proportional effect has been used as a general underlying principle to explain the appearance of fat tails in the large variety of situations studied; due to its flexibility, many times it has been modified accordingly to fit the particularities of each observation. However, as we will see, these models generally show a *predictability* (the capacity of predicting accurately the variable of interest) much higher than in real data; this is because these models are usually focused in explaining the existence of the fat-tailed distributions. It is important to remark that predictability is not absolute (a yes/no question) but can be quantified, although it is required to do it in a way that is robust also for systems exhibiting fat-tailed distributions. We are also interested in prediction: in that case the dynamical features of the system have to be modeled as well, in order to reproduce the randomness, not only in the overall distribution, but also in the movement among its ranks.

Online social media, such as forums, websites' views or online videos, are a natural starting point for studying aggregate behavior of social activity: each interaction in these media is tracked, generating an unprecedented amount of data that can be analyzed. The variable of interest in this case will be a proxy for *attention*, the time that a user invests in viewing a given item. Thus the objects of study consist of an ensemble of time series of users' activity. The dynamics of the activity are, at first sight, non-deterministic, as seen in Fig. 1.1, where a schematic representation of the views of YouTube videos (one of the four databases used in this Thesis) is shown. Videos that earlier in their lifetime have the same amount of views rapidly take radically different, seemingly random, paths. Therefore, a probabilistic modeling approach seems better suited to account the fluctuations from the average behavior; one framework that fit these needs is the one of Stochastic Differential Equations (SDE), a formal way of dealing with *Langevin equations*. It is possible to model activity as a first order SDE, or using higher order equations. In first order SDEs, the activity at a given time is only dependent on the immediately previous value; proportional effect models would enter in this class. While higher order equations are more complex, and would in principle allow for a more detailed description of the dynamics, the improvement in the description's accuracy should be estimated, in order to quantify the net advantage of this approach. Beyond the specific results of our analysis that clarify the dynamics of this *economy of attention*, we aim to establish a framework to study the dynamics of systems (also biological and physical) where fat-tailed distributions appear.

This Thesis is organized as follows. Fat-tailed distributions are formally introduced in Chapter 2, along with a review of the mechanisms for their appearance, and their implications for forecasting social activity. In Chapter 3 the data used in the Thesis is introduced; four databases of online social media were collected, and a statistical characterization of them is presented. In Chapter 4 a measure of predictability with respect to features that is robust against heavy tails is proposed, based on the possibility of classifying extreme events correctly, independently from the forecasting method chosen. Applications to data are presented, where it is established which are the more

important features to be modeled. Proportional effect is studied in Chapter 5 in the YouTube database. Although the data present proportional effect (the views in one day are proportional to the aggregated views since publication), it is shown, by taking into account the fluctuations of the data, that it is not the main cause of fat-tailed distributions. Accordingly, a model that reproduces the features of data, namely the presence of Lévy-stable fluctuations, is presented in the context of SDE. The parameter estimation of this model from data is discussed in detail, and a rigorous statistical comparison with respect to previous competing models is performed. The problem of modeling activity with generalized models, also beyond first order SDEs, is approached in Chapter 6, where other time-series models are discussed and a new model based on daily activity is introduced. A comparison with the model proposed in Chapter 5 is presented; the results indicate that evidence for proportional effect can be observed in systems that are better described by models with Lévy fluctuations, even without an explicit dependence on the total activity. A summary of the main conclusions, an outlook, and a discussion on some open questions is presented in Chapter 7.

This Thesis is the result of studies performed at the Max Planck Institute for the Physics of Complex Systems between 2011 and 2016 under the supervision of Dr. Habil. Eduardo G. Altmann.

2. Fat-tailed distributions in social activity: evidence and models

Fat-tailed distributions are ubiquitous in social activities, where they are used to describe very diverse variables, such as cities' population, personal income and wealth, and, in particular, online activity patterns. The prevalence of these distributions triggered plenty of research devoted to understand why they are so common, and aiming to unveil, if any, the universal mechanism underlying their emergence. This problem is one of the oldest in social complex system research and is considered a fundamental one as well, due to the variety of systems where fat tails appear, from physics, biology and social sciences, and the wide implications that these particular mechanisms have.

The problem of the prevalence of fat tails in social activity is intimately related with the problem of how this activity dynamically evolve, because many different mechanisms can originate such distributions. In this chapter these ideas will be briefly reviewed, and one in particular, the class of growth models with proportional effect, will be explored more in detail. Moreover, the implications for the problem of forecasting will be mentioned, pointing out the necessity of methods that are robust with respect to fat-tailed distributed variables.

2.1. Distributions with fat tails

2.1.1. Definitions

Let X be a continuous random variable, with a probability density function f_X . We say that X is *fat-tailed distributed* if f_X is such that some of the moments of X diverge, i.e. if exists a value α such that the expectation value $\mathbb{E}X^\beta$, defined as

$$\mathbb{E}X^\beta \equiv \int x^\beta f_X(x) dx, \quad (2.1)$$

diverges for all β such that $\beta > \alpha > 0$ but is finite for all $\beta < \alpha$. We make a subtle distinction with *heavy-tailed distributions*, which are distributions that decay slower than exponentially, i.e.

$$\lim_{x \rightarrow \infty} e^{\lambda x} \mathbb{P}(X > x) = \infty \quad (2.2)$$

for all $\lambda > 0$, which is equivalent to state that their Moment Generating Functions is infinite for any positive argument. (A notable example of a heavy-tailed, but not fat-tailed distribution is the Lognormal.) Note that there is no universal consensus in these two definitions, so fat and heavy tails in this Thesis should be intended in the word sense above described.

If X is a discrete random variable, then the probability mass function is used, which we note simply by $\mathbb{P}(X = x)$; the above definitions hold, replacing the integral with the sum in the expectation value calculation,

$$\mathbb{E}X^\beta \equiv \sum_x x^\beta \mathbb{P}(X = x) \quad . \quad (2.3)$$

In both cases (continuous and discrete) the cumulative density function will be denoted as $F_X(x)$ or simply as $\mathbb{P}(X < x)$. A related term often used is the *power-law distribution*, which refers to any distribution that decays as a power of X :

$$\mathbb{P}(X > x) \propto x^{-\alpha} \quad , \quad (2.4)$$

with $\alpha > 0$ (throughout the Thesis, α will always denote this exponent). This notion falls into the category of fat-tailed distribution, which is the notion that we discuss in this Thesis.

2.1.2. Fat-tailed distributions in nature

Fat-tailed distributions are present in a variety of situations. In statistical mechanics they are typically considered as signatures of systems in their critical state [Sor06], in particular in a second order phase transition. Is the case, for example, of clusters' distribution in the simplest models of percolation [SA94] and spin systems as the Ising model [CN86]; the lack of a typical scale in a cluster of sites is directly related to an infinitely large correlation length, which can be regarded as the hallmark of a critical point [Bax07]. Clusters in percolation and Ising models exhibit fractal properties as well, i.e. a non-trivial scaling of the surface of the cluster with its size, a notion that relates with surface growth in interfaces [BS95].

While Ising model and percolation have a fundamental importance in order to understand strictly physical problems such as magnetism, conductivity, and porous media, they have been used as paradigmatic models for a much larger class of phenomena of other natural sciences in geophysics and ecology. In these systems, fat tails are observed but the fundamental mechanisms are not well established, because they are a result of a complex interplay of many elements (see Fig. 2.1 for examples of these distributions). This is the case of the Gutenberg-Richter law [GR56], which states that the amount of earthquakes is inversely proportional to their magnitude (the released energy) up to an exponent to be estimated from data (a power-law); it has been proposed [SRS93] that the earthquakes' magnitudes are related with the geometry of faults' patterns, which is given by percolation theory.

There are also fat-tailed distributions in the natural sciences that are not related with these statistical physics models. A well-known example belongs to ecology, where a power-law shape has been found for the frequency distribution for sizes of genera of flowering plants; this topic was studied by Udny Yule in 1922 [WY22], and constitutes one of the earliest examples of the observation of fat tails in nature. Notice that this is a case where the objects of study (flowers) are already a complex system on its own.

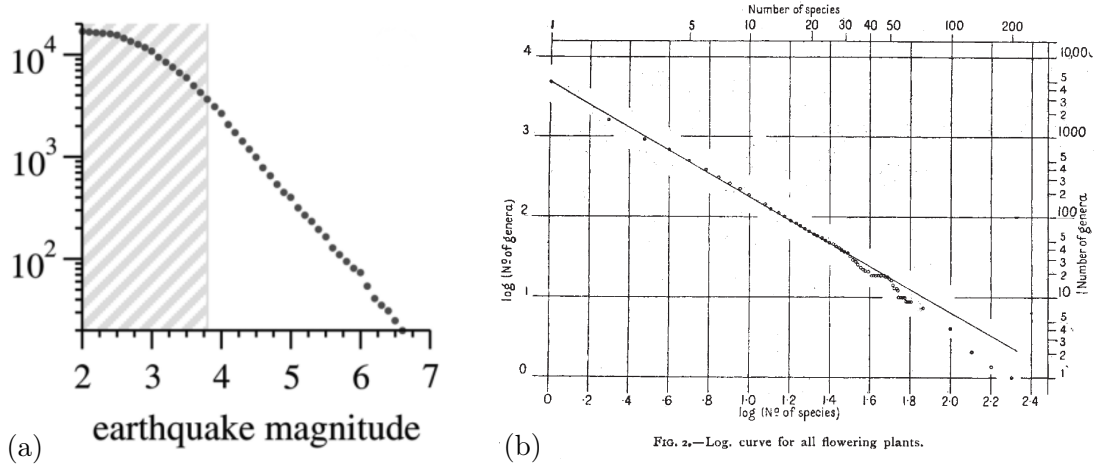


Figure 2.1.: Examples of fat-tailed distributions in natural sciences. (a) Frequency of earthquakes' magnitudes, from [New05], logarithmic scale for the frequency, magnitudes are in Richter scale (proportional to the logarithm of the released energy). (b) Frequency of genera's number of species, from [WY22], logarithmic scale in both axes.

2.1.3. Fat-tailed distributions in social activity

In an even higher level of complexity, we have the so-called *social systems*, where the individual elements of the distribution are the result of the interactions among humans, which are not just complex in the biological sense, but they are also conscious of their own interactions (see Fig. 2.2 for examples of these distributions). Two examples of social systems with fat-tailed distributions are notable, since their analysis were seminal in their respective fields. Vilfredo Pareto found that the income distribution [Par96] was fat-tailed in 1896; his study was focused on data of England, Prussia, Saxony, and Italy, already pointing at some universal feature of income distribution that would be confirmed for many other times and places (for modern analysis see Ref. [CG05]). George Zipf, instead, found the same property for the frequency distribution of words in 1936 [Zip36], which, up to certain limitations, is considered valid for a large range of datasets from different years and languages [GA13].

It is difficult to overestimate the importance of the observation of fat-tailed distributions in these disciplines. For instance, heavy tails in the income distribution and the lack of a characteristic scale

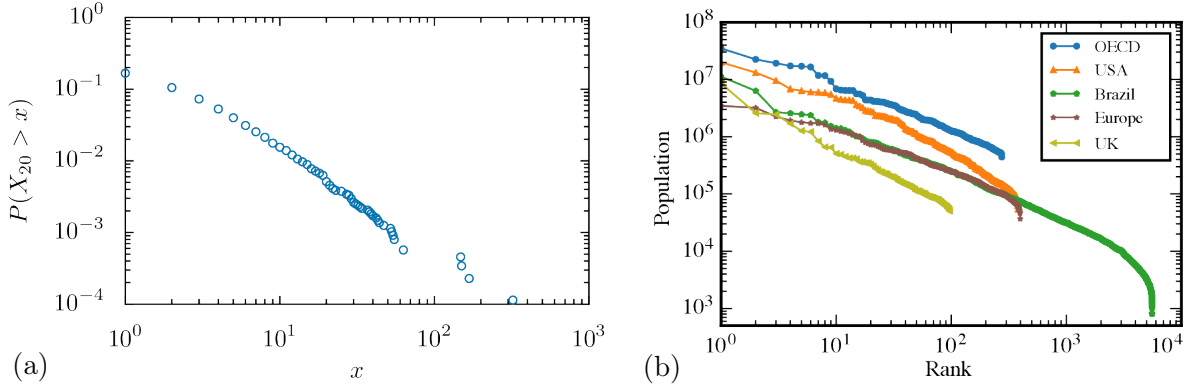


Figure 2.2.: Examples of fat-tailed distributions in social sciences. (a) Distribution of citations up to the year 2010 on papers published in the APS in 1990 (20 years after the publication; see Section 2.3 for details about the data). (b) Rank-frequency distribution of cities' population for five datasets of cities, from [LMGA16]. The Rank-frequency distribution is built by ordering the cities according to their population, where the rank 1 is given to the most populated city. This representation is equivalent to the cumulative density function [AH02] (in this plot the axis are inverted).

of income imply high inequality, where the index α of the distribution is a proxy for concentration of wealth [DY00].

2.1.4. Social activity in Online Media

A particular type of social systems are the *Online Media*. The development of Internet had a wide impact on the society as a whole. Intended as a form of communication, Internet is based on the digitization of its content, which has as a consequence not only the storage of such content, but also the storage of the information about the people who communicate. The appearance, and increase over time, of communication data lead to scientific research that analyzes social systems from this data, since it mediates a great deal of our activity. In fact, many social scientists understand data in Internet as a proxy for features in society that would be otherwise hard to observe [FC10, AG05, SDW06]. This fact does not only explain the rise of the study on complex networks in the last twenty years [BA99, WS98], but also the central role occupied by machine learning methods. These approaches, respectively fulfill the need for models of interpersonal links on the one hand, and for tools suited to cope with complex, high-dimensional data on the other.

Mass media are characterized by their easy access, so that they can be consumed by the so-called masses. Examples of it include the Internet, newspapers, radio, and television. It has been proposed [Sim71, DB13] that when a product is practically free, the *attention* (of the consumers) becomes the scarce commodity of the market, because people have only a limited amount of it, a *capacity* [Kah73]. This is the concept that regulates the economy of mass media, an *economy of attention*, where companies profit mostly from advertisement and where the time a person

dedicates to the product is the valuable asset.

In the Internet, in particular, typical mass broadcasting is intertwined with social media, where the publication (in a wide sense) of content is left to the public. This innovation is due to the relatively cheap, as well as technically undemanding, possibilities of publication, usually through platforms that enable it. A familiar example for the reader, although not directly related with Internet, may be the one of scientific publishing. This modality, where scientists instead of a passive audience are both writers and readers, is now common to other types of communication, such as general prose writing, music and videos.

The massive increase of producers and content poses naturally the problem of how the attention of the consumers will be distributed in this new situation. Not surprisingly, in view of the examples described above, data analysis shows that attention is fat-tailed distributed (see for example Refs. [LAH07, RFF⁺10, Pri76] and Table 2.1). This is, again, a problem with a clear commercial interest, but also of scientific interest, because the scale of the reach of online media is the highest (worldwide), and the impact of the dominant items can be determinant for the overall culture and/or politics.

It remains to explain how these fat tails appear, how they evolve, and why do they appear, i.e. to model the distribution of attention (*activity*). The second problem to address is what happens with the individual objects (*items*), if it is possible to predict the attention that they gather, what are the factors that determine this prediction, and how they are related with the overall distribution. Modeling and predicting social activity are the focus of this Thesis.

2.1.5. Measuring attention

Attention, defined as the fraction of time that people spend in a particular item from the media, is not usually measured in a direct way. There are, however, different proxies for attention, mostly traces of online activity by individuals. Table 2.1 shows some examples of proxies for attention where fat tails have been reported. In Chapter 3 we describe in detail the datasets that are used in this Thesis. In some of these reported contexts, attention can be understood as well as *popularity* or *success* [WSB13, WH07a]. This flexibility in the terms actually points out a double nature of the origin of attention. Take the example of scientific papers. Scientific papers (and scientists) often are evaluated by the number of citations they accrue. By being numerical, citations have the advantage of being objective measures of the impact of a paper in its field, if the intention of measuring the intrinsic quality of the researcher with a proxy. Many criticisms to this interpretation can be made: the field of a particular paper may be more popular than other fields (thus being popular among fields), some papers become cited as references because other papers made it already, some citations are “generic” (usually used as examples that can be replaced) thus leaving space for choosing the citation with non-scientific considerations, etc. This criticisms reveal that citation dynamics, and attention dynamics in general, is governed by an

interplay of quality and popularity mechanisms [WSB13].

System	Item	Attention measure/activity (X)	Refs.
Online Videos	video	views, likes	[CS08]
Discussion Groups	threads	posts, answers	[APM11]
Publications	papers	citations, views	[Pri76, SSPA10, WSB13, PPP ⁺ 13]
Twitter	tweet	retweets	[WFVM12]
WWW	web page	views	[RFF ⁺ 10]
Online Petitions	petition	signers	[YHM13]

Table 2.1.: Examples of reported fat-tailed distributions of proxies for attention.

Another aspect of the dynamics of citations is the distinction between *endogenous* and *exogenous* disturbances [CS08, SDGA04, GGMA14]. Exogenous bursts are triggered by a sudden interest in a particular topic, or item, while the endogenous bursts are a result of a dynamical process that has as main variable the past activity itself. This distinction has the advantage that if data about a related observable is available (an example of an exogenous factor may be the mentions of a certain research paper in the mass media), it is possible to recover, in principle, the evolution of the attention given just by the sharing of people among each other, something usually modeled as an epidemic-like process. In general, in online media there is a coexistence among many forms of communication, as broadcasting, direct sharing, public sharing, and navigating, and additionally, there is a feedback among these forms. To distinguish these factors is a difficult task, because it usually requires detailed information of user activities [WFVM12, CSF10].

Proxies, or measures of attention will be noted with the variable X , where $X_t^{(i)}$ is the amount of activity accrued up to time t by the i -item. We will use $dX_t^{(i)}$ as the difference in the consecutive values of $X_t^{(i)}$,

$$dX_t^{(i)} \equiv X_t^{(i)} - X_{t-1}^{(i)} . \quad (2.5)$$

2.2. General arguments for the prevalence of fat tails

Reviews on the subject of fat-tailed distributions in data [CSN09] and their origin [Mit04, Per14] focus mostly on observations from social systems; in fact, it is often said that fat tails are ubiquitous in social systems. The reason for the ubiquity of fat-tailed distribution in social activity can be approached in two ways. The first is the mathematical approach, where fat-tailed distributions are the attractor of Central Limit-like theorems, detailed in Section 2.2.1. A second less mathematical approach is based on models, i.e. when fat-tailed distributions are the result of mechanistic models, which is discussed in Section 2.2.2.

2.2.1. Fat-tailed distribution as attractors of stochastic processes

The observation of distributions with fat tails can be explained by probabilistic arguments. The first of them, involves Extreme Value Theory, which deals with the probability of having very high values in an unknown stochastic process. Two fundamental theorems in this theory describe this probability in different ways. The so-called First Extreme Value Theorem (Fisher-Tippett-Gnedenko [DHF07]) states that the probability distribution of the maximum in a collection of values can be of only three types, among the Frechét, Gumbel, or the inverse Weibull distributions. These three distributions define basins of attraction in the space of all possible distributions.

A *Second Extreme Value Theorem* (Pickands-Balkema-de Haan [BDH74, PI75]) states that the probability distribution of values above a given threshold is the Generalized Pareto distribution. More formally, consider a random variable X with cumulative density function $F(x)$ that is non trivial in all the range of the real numbers, i.e. $F(x) \neq 1, \forall x \in \mathbb{R}$; for a sufficiently large threshold x_p , the distribution function for $X > x_p$ can be approximated by

$$1 - F(x) \approx (1 - F(x_p)) \left(1 - H_\gamma \left(\frac{x - x_p}{\sigma(x_p)} \right) \right), \quad (2.6)$$

where $\sigma(x_p)$ is a constant that depends on the threshold x_p , and H_γ is the Generalized Pareto distribution, defined as

$$H_\gamma(x) = 1 - (1 + x\gamma)^{-1/\gamma}, \quad (2.7)$$

where for $\gamma = 0$ the right hand side is interpreted as $1 - e^{-x}$. For example, if X is normally distributed, i.e. $X \sim N(0, 1)$, then $\gamma = 0$; we are interested in the case where the variables do not follow normal statistics, i.e. $\gamma \neq 0$, so we will replace γ by $\alpha = 1/\gamma$ to ease the notation.

Since $1 - F(x)$ is the probability of X being higher than x , the conditional probability of X being higher than x given that X is higher than the threshold can be written as

$$\mathbb{P}(X > x \mid X > x_p) \approx \left(1 + \frac{x - x_p}{\sigma(x_p)\alpha} \right)^{-\alpha}. \quad (2.8)$$

With $\alpha > 0$, Eq. (2.8) becomes a power-law distribution and, as such, a fat-tailed one. This theorem is valid for any random variable X , implying that if we have a set of empirical data obtained from any stochastic process, a threshold can be set such that we observe the Generalized Pareto distribution, which, if $\alpha > 0$, is a fat-tailed distribution.

A second probabilistic argument involves the sum of variables instead of their maxima. A distribution is called *Stable* (or *Lévy-stable*, or α -stable) if the sum of i.i.d. (independent, identically distributed) Stable random variables is also Stable-distributed. Notably, there is no analytical form for the probability density function (and the cumulative) of a Stable random variable X

$(X \sim S(\alpha, \beta, \mu, \sigma))$, which has instead a characteristic function, $\phi_X(k) \equiv \mathbb{E}[\exp(ikX)]$:

$$\log \phi_X(k) = \begin{cases} i\mu k - s^\alpha |k|^\alpha \left[1 - i\beta \tan\left(\frac{\pi\alpha}{2}\right) \text{sign}(k)\right] & \alpha \neq 1 \\ i\mu k - s|k| \left[1 + i\beta \frac{2}{\pi} \text{sign}(k) \log(|k|)\right] & \alpha = 1 \end{cases}, \quad (2.9)$$

where μ is its expected value $\mathbb{E}X$, σ is a positive scale parameter, β regulates the asymmetry ($\beta \in [-1, 1]$) and α is the index of the tail, which is heavy if $\alpha \in [0, 2)$; if $\alpha = 2$ this distribution coincides with the Normal distribution, where β has no effect and σ is the usual standard deviation. (These parameters' choice correspond to the so-called parametrization 1 of Ref. [Nol12].) The previous expression can be deduced from the property of stability, which is easily tractable by means of the characteristic function (equivalent to the Fourier transform). Although the probability function of the stable distribution has no analytical form, it is possible to approximate its tail ($x \rightarrow \infty$), by the formula

$$\mathbb{P}(X > x) \approx \sin\left(\frac{\pi\alpha}{2}\right) \frac{\Gamma(\alpha)}{\pi} (1 + \beta) \left(\frac{x}{\sigma}\right)^{-\alpha}. \quad (2.10)$$

It is visible from this approximation that the Stable distribution is fat-tailed. The Central Limit Theorem in its usual form gives the asymptotic distribution of the sum of random variables when the variance ($\mathbb{V}(X) = \mathbb{E}X^2 - (\mathbb{E}X)^2$) is bounded. A *Generalized Central Limit Theorem* [Zol86, Nol12], which instead makes no requirements on the variance, states that the sum converges to a Stable distribution (also known as Lévy-stable or α -stable). If we have a sequence $\{x_i\}_i$ of i.i.d. realizations of X , such that $\mathbb{E}X < \infty$, the distribution of the sum of N terms converges to a Stable distribution (convergence in distribution), when N tends to infinity:

$$\sum_i^N x_i \xrightarrow{d} S(\alpha, \beta, N\mu, N^{1/\alpha}\sigma) \quad (2.11)$$

When $\mathbb{V}(X) < \infty$, the resulting distribution has $\alpha = 2$, which is the Normal distribution.

It is worth to notice that these two theorems (Second Extreme Value Th. and Generalized Central Limit Th.) use the notion of *stability*, i.e. applying a function between two i.i.d. random variables returns the same distribution. In the case of extreme values, that function is the maximum, which is used in the First Extreme Value Theorem, and it leads to the Generalized Pareto distribution; in the case of the Central Limit Theorem, the function is the sum that preserves the Stable distribution, given that they belong to the same class (same α). The stability of these (similar) distributions implies that they are attractors in the probability-distribution space of the repeated application of the functions with respect to which they are stable.

The generality of these arguments account for the emergence of fat-tailed distributions, although it gives no explanation on how these distributions appear, i.e. why the extreme events are in the basin of attraction that correspond to $\alpha > 0$ or why the variance is not finite.

2.2.2. Mechanistic Models for fat-tailed distributions

Models that lead to fat-tailed distributions are almost as old as their observation, and although time has brought rigorousness and variety derived from different applications to specific problems, the underlying ideas are essentially the same and can be reduced to three classes: growth processes, multiplicative random walks, and cost optimization.

The first two classes rely on proportional effect, and are being detailed below, while the cost optimization approach states that this type of distributions exists because they are minimizing a certain cost function, as originally stated by Mandelbrot [Man53]. In more recent times, the argument has been generalized to designed systems in the Highly Optimized Tolerance framework [CD99], where heavy-tailed distributed configurations are also very robust under perturbations, and other models [PDD13, MMR13], although different, maintain the same spirit. A recent family of models added to the list of generators of fat-tailed distributions is the Self-Organized Criticality [BP95] (SOC). The inspiration for these models clearly comes from Physics: physical systems in a critical state show fat-tailed distributions, implying that these statistical patterns are indication of criticality. However, to reach a critical state in physics it is needed to tune a given order parameter to its critical value, which is a fine tuning of the system. The idea behind SOC is that in certain systems the critical state can be an attractor of the dynamical evolution of the system.

If we are interested in modeling not only the distributions of activity, but also the dynamics of how this activity evolves for each item, the ideas above will not play a big role because they lack a mechanistic dynamical model. A class of stochastic processes is often applied in describing such mechanisms by which activity grows, the *growth processes* [Per14]. Udney Yule proposed first a model for the distribution of species among genera of plants [Yul25], in the context of an evolutionary process, where new species are naturally created, the growth element of the process. The number of species per genus would increase exponentially with time, since genera with more species are more likely to produce new mutations, introducing a dynamic element. With these simple assumptions it is possible to obtain a power-law distribution of species dynamically, from an initial condition where each genus have only one species. The essential ingredient here is the proportionality between the probability of having new species and the number of already existing species in a genus, later dubbed as *proportional effect*¹ by Gibrat [Gib30], who used, perhaps unknowingly, the same idea to explain the fat tail of the income distribution. Simon [Sim55] generalized this mechanism and proposed it as a model to explain the many fat-tailed distributions observed in social systems, while at the same time formalized the mathematics of the problem. Models with proportional effect were since then widely used for different applications with many

¹the word *proportionate* is also used; little difference between the two terms exists, but while *proportional* is mostly used quantitatively or in a relative sense (*the hand is proportional to the body*), *proportionate* can be used in absolute way (with the sense of *appropriate*) or with concepts not so clearly measurable (*that human body is proportionate; proportionate response*) [OED16]; in this Thesis the word *proportional* is preferred.

variations that adapt it to each particular situation [Per14]. An example worth to be mentioned is that known as *preferential attachment* [BA99, New01, JNB03], due to its leading role in the study of Complex Networks.

A second class of processes was originally proposed by Champernowne, who was also interested in the income distribution [Cha53]. He arrived to the same conclusions as Gibrat through a master equation approach, where the basic postulate is that people move from one (logarithmic) class of income to another as a random walk. This argument was formalized by Kesten [Kes73] in general terms and by Gabaix [Gab99] for cities' population distribution, and in modern term it can be formulated as follows: consider a stochastic process where the variable $X_t^{(i)}$ is the fraction of total income of the person i at time t . A growth rate of income is assigned to each person; this growth rate should account for the different mechanisms in the economy, and we assume that the total effect of these mechanisms can be described by a Normally distributed random variable. Moreover, a reflecting barrier of income x_0 is set, such that if the person is below the barrier, it can receive only a positive growth rate. Since the growth rate is an increase relative to $X_t^{(i)}$, and it is random, it is said that $X_t^{(i)}$ is subject to a multiplicative noise and a reflecting barrier at some low x_0 . The average growth rate when the income is above x_0 should be (slightly) negative, in order to compensate the positive growth guaranteed to the items below x_0 . The process can be written for continuous time as a *Stochastic Differential Equation* (SDE),

$$dX_t^{(i)} = \begin{cases} aX_t^{(i)}dt + bX_t^{(i)}dW_t^{(i)} & X_t^{(i)} > x_0 \\ \max\left(aX_t^{(i)}dt + bX_t^{(i)}dW_t^{(i)}, 0\right) & X_t^{(i)} \leq x_0 \end{cases}, \quad (2.12)$$

where $dW_t^{(i)}$ is the infinitesimal increment of the Wiener process associated to the item i (informally, it can be thought as an infinitesimal step of a random walk), and $a < 0$, $b > 0$. (This particular SDE is known as Geometric Brownian Motion [KPS12].) The distribution converges (for large t) to a power-law [Gab99]

$$\mathbb{P}(X > x) = \left(\frac{x}{x_0}\right)^{-\alpha}, \quad (2.13)$$

where $\alpha = 1/(1 - x_0/\mathbb{E}X)$. The barrier is important, since it makes the difference between the typical Lognormal distribution output of a simple multiplicative noise process and the power-law distribution obtained here. Paradoxically, in the limit where the barrier $x_0 \rightarrow 0$, the exponent $\alpha \rightarrow 1$, although if it is exactly zero, the distribution is Lognormal.

It has been early noted by Simon himself [Sim55], that the two previous classes of models (growth processes and the Geometric Brownian Motion with barrier) share the same property, namely the linear proportionality of the average increment dX_t with respect of the current value of the variable of interest, X_t . However, both approaches have advantages and disadvantages. The

growth processes are better suited to define sophisticate models, since it is possible to introduce complex mechanisms that modify the amount of X assigned to a particular item, and how this item is chosen, as well as the possibility to control the number of items in the system. The SDE approach, instead, even if less flexible, allows to use the mathematical machinery developed for this kind of equations [KPS12]. Additionally, mechanistic models with proportional effect are usually defined discretely (both X and the holders of X (the items) are discretized), while SDE are continuous both in time as in space. It is possible, however, to use SDEs as an approximate formalism to the more general class of growth processes; a particular case where this approximation can be done will be shown in Section 5.2 and Section 5.3.

As stated above, the common feature between the two classes of models previously described is that the expectation of the increment of an item's activity is proportional to the total activity

$$\mathbb{E}dX_t^{(i)} = aX_t^{(i)}, \quad (2.14)$$

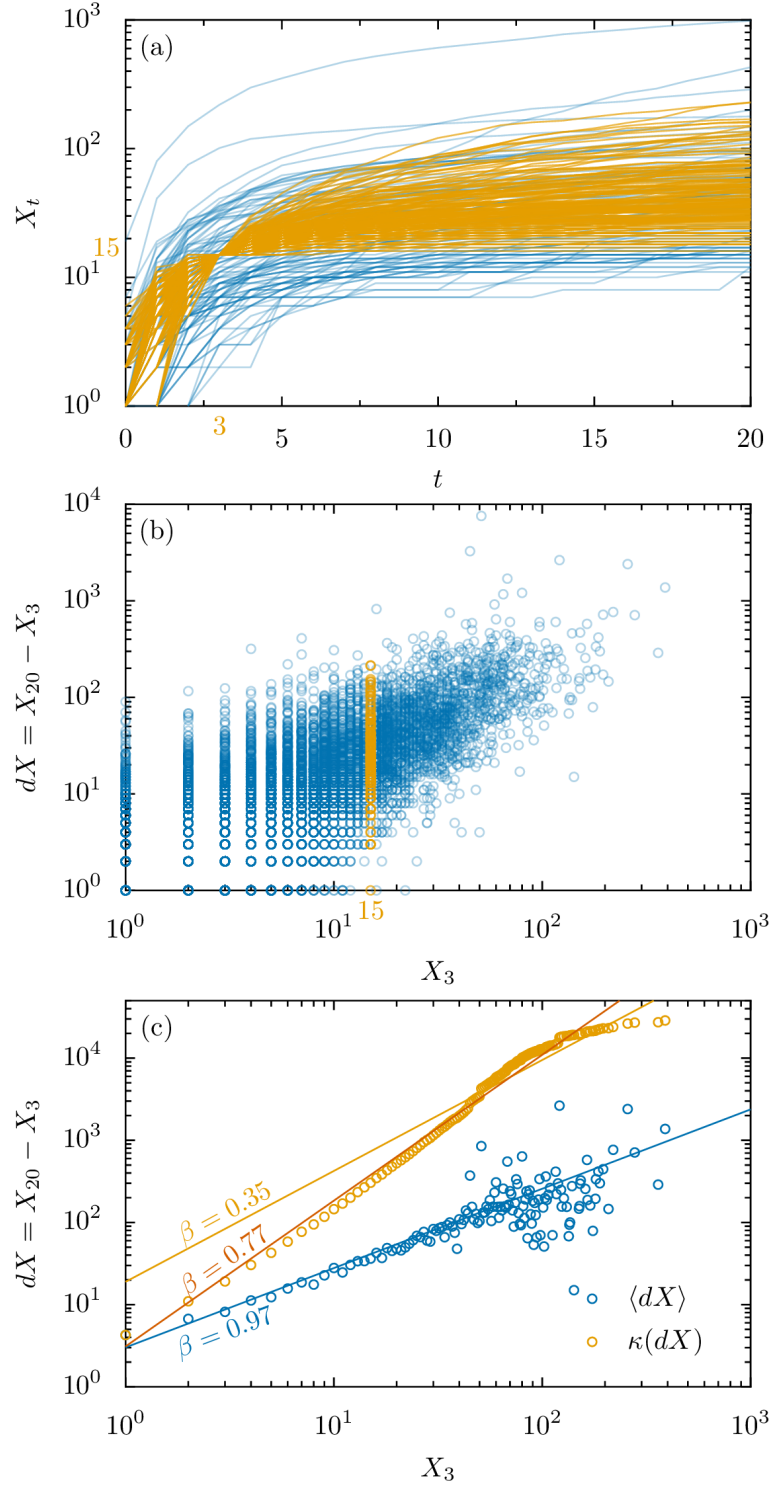
for some proportionality constant a , which for us defines the class of processes with proportional effect. We focus now on how to model both the dynamics and the fat-tailed distributions of attention in empirical data.

2.3. Estimation of the Proportional effect

We can continue the example of citations to illustrate how the proportional effect is established in empirical data. From the database of Web of Science (provided via the Max Planck Digital Library) it is possible to extract the total amount of citations a paper i gathered since its publication up to t years, $X_t^{(i)}$; these time series are shown in Fig. 2.3(a). By taking two time points, in this case 3 and 20 years, we can try to establish a relation between X_3 and the increment up to 20 years, $dX = X_{20} - X_3$, variables that we expect to be proportional. This relation is illustrated in Fig. 2.3(b) and suggests the presence of the proportional effect $dX \sim X_3$.

Two methods are typically used to estimate proportional effect from empirical data [Per14, GS13]: the averaging method and the cumulating method.

- The averaging method consists in averaging the values of dX for each of a set of predefined bins [New01], $\langle dX \mid X_t \in \text{bin} \rangle$. In Fig. 2.3(c) these bins were selected to have a width 1, thus effectively computing an average of dX for each different X_t , with the consequence of having noisy data in the end of the curve. A different selection of bins can be made, but it has to be noted that any selection is arbitrary, and as such will have its own problems. Two issues are: to which X_t the average dX has to be assigned and the disparity on the amount of items that fall in each bin, which makes some points statistically more reliable than others. A partial solution to this issue is detailed in Section 5.4.1.



- The method of cumulation [JNB03, EL03] consist just in summing the averages of dX

correspondent to all the items with $X_t < x$, defining a cumulative function

$$\kappa(x) = \sum_{i: X_t \leq x} \langle dX \mid X_t \rangle \quad (2.15)$$

shown in Fig. 2.3(c). This cumulative function will have the functional form of the integral of the relation between X_t and dX , so in the case of proportional effect, we expect a quadratic function. Since more items are summed in each point of κ , the noise is greatly reduced by this method.

Both methods operate on the data creating a new set of points with a lower noise than the original; in this sense, cumulation is equivalent to use the cumulative probability distribution instead of the probability density. The result of any of these processes is typically fitted by a non-linear function of form $y(x) \propto x^\beta$, and it has been found that $\beta \approx 1$ in a variety of situations.

However, the implementation of the estimation procedures mentioned above yields results that are not always consistent, as illustrated in the citations example. The points that result from averaging and cumulating are fitted with a non-linear regression, shown in Fig. 2.3(c) with the best fit for the function $y(x) = a \propto x^\beta$ by the weighted least-squares method, where the weight is given by the amount of points at each value of X_t . The results of fitting both datasets (cumulated and averaged data) is clearly not the same; while the fit of the averaged data results in $\beta = 0.974 \pm 0.006$, the fit of the cumulated data yield $\beta = 0.35 \pm 0.02$ because it is dominated by the points of high X_t . A commonly used procedure is avoiding the non-linear fit altogether, applying the logarithm to linearize the data [JNB03]; this type of regression on the cumulated data results in $\beta = 0.772 \pm 0.001$.

The inconsistency among methods can be understood by noticing that these fitting procedures are generally inadequate to account for fluctuations, except in very few cases [MSSVK08]. In these methods, the weights in the fit fulfill the role of assigning a fluctuation to the data points, making more certain averaged values of more data. The aim of cumulating data is to reduce the error, but that is achieved by summing the increments dX , which correlates the points to be fitted; at the same time, it is not clear what the weight should be (the amount of the cumulated data up to X , or the amount of data that is being added exactly for X ?); a fit with constant weights results in a value of β even further from $\beta = 1$. In Section 5.4.2 these issues will be addressed by the introduction of the Maximum Likelihood method for this particular problem, where an explicit model for the data fluctuations is defined.

Fluctuations away from proportional effect can be analyzed in two general frameworks, where we consider X_t to be the result of a dynamical process. The first option is to consider the fluctuations as a random noise to be added to an average behavior common to all the items; this approach is considered and implemented in Chapter 5, where the limitations of SDE models are discussed and it is shown that the choice of the form of the fluctuations is crucial to the correct interpretation

of the dynamics of the system, and the formation of fat tails as well.

The second possibility is to consider the fluctuations as the aggregated effect of mixing items that are not in the same state, but just the same X_t . The assumption here is that the state of an item is given by more than one value of the activity, e.g. X_t and X_{t-1} , generalizing the proportional effect described until now. The implications of this possibility are discussed in Chapter 6, where models in which the increment depends on many previous times are implemented.

The interpretation of these fluctuations is not important only because of the accuracy of the model, but also because these models would forecast in a radically different way the unobserved future.

2.4. Forecasting social activity

Forecasting social activity is, on the one hand, a way of testing the accuracy of models, and on the other hand, a goal on itself, where most of the commercial interest is focused [YK12]. A *forecasting method* is any algorithm that estimates the value of a variable (e.g. X_t) for a time still not observed; forecasting methods will use a set of variables, that we call in this context *features*, to make predictions.

In the context of time series prediction, forecasts can be deterministic (e.g. the paper i will have 100 citations 10 years after its release, $\hat{X}_{10}^{(i)} = 100$) or probabilistic (e.g. the probability of paper i to have 100 citations 10 years after the release is p , $\hat{\mathbb{P}}(X_{10}^{(i)} = n) = p$). The deterministic forecast can be thought as an extreme case of a probabilistic forecast (the probability of paper i to have 100 citations 10 years after the release is 1), so it carries potentially more information. Both types of forecasts will use information available before the predicting time t . Let's note with $H_t^{(i)}$, the set of values $X_{t-1}^{(i)}, X_{t-2}^{(i)}, \dots, X_0^{(i)}$, the *history* of the i item and with $M^{(i)}$ all the features of the item that do not depend explicitly with time (e.g. observables that depend on the content or properties of the item such as the publication date). The most general forecasting method is a function f that uses as input $H_t^{(i)}$ and $M^{(i)}$, it has some parameters θ , and returns an estimate for the activity at a desired future time, $X_t^{(i)}$, or a probability distribution of its value (deterministic or probabilistic forecast, respectively). As an intermediate step, the parameters θ of the method must be established; the general name for this process is *learning*, where θ are adjusted such that the accuracy of the forecast is maximized.

The choice of the *accuracy measure* of the prediction is a fundamental point in the forecasting problem, since the measure of accuracy chosen will determine the optimal forecasting method. For a deterministic forecast at time t , one natural and widely used measure is the mean squared error,

$$\epsilon^2 = \frac{1}{N} \sum_{i=1}^N \left(\hat{X}_t^{(i)} - X_t^{(i)} \right)^2 \quad (2.16)$$

which is the euclidean distance between the set of predictors $\hat{X}_t^{(i)}$, and the real values, $X_t^{(i)}$, for each item.

In general, the parameters θ are used to weight the importance of each of the available information, i.e. the components of $H_t^{(i)}$ and $M^{(i)}$. Having a high number of parameters makes the forecasting method more flexible to adequate the prediction to the real value; however, this flexibility may fit the model to data that is actually a fluctuation from the standard behavior. For this reason the evaluation of a forecasting method is done by dividing the set in two subsets, the *training* and the *target* sets. The training set is the set of data used to choose the parameters of the method, maximizing the accuracy, and the target set is the set of data used to measure the real accuracy of the method (out-of-sample error). Usually increasing the complexity of the method (most of the times this is equivalent to increase the number of parameters) at first improves the accuracy measured both on the target and on the training set, but then decreases the accuracy on the target while still increasing the accuracy in the training set, a behavior named *overfitting*. In order to avoid overfitting, it is important to find the most important features that determine the accuracy of the forecast, in order to use only them in the forecasting method. Notice that when the forecasting method is based in a model (some methods are not), solving this problem is equivalent to finding the minimal set of relevant features that explain the observations. It is possible in certain cases to approximate the out-of-sample error by a *regularization* technique [FHT01], penalizing more complex models (see Section 5.4.2).

The problem of finding the most important features is usually addressed in the context of a particular forecasting method [MJG90], determined by the chosen accuracy measure. However, the question can be framed regardless of it, i.e. what are the features that contain information about the activity dynamics? Here we get to the idea of *Predictability* [KAH⁺06], the capacity of a variable to be predicted given a certain feature. The study of the predictability is a main theme of this Thesis, and has an importance in itself, since it is part of our daily experience that seemingly ordinary items (videos, news, publications, etc.) unexpectedly gain an enormous amount of attention, propelling the idea that systems with fat-tailed distributions are unpredictable [Tal07, COS⁺13, Sor09]. Therefore, there is a need for a measure of the predictability that is independent from the forecasting method; this topic is addressed in Chapter 4.

It remains the problem of estimating the accuracy in presence of fat tails. When the accuracy measure is the mean squared error, there is a trivial forecasting method that consists in predicting the same value for each item, regardless of the other available information. If such method is used, then the predictor that maximizes the accuracy can be found by deriving Eq. (2.16) with respect to \hat{X}_t and equating to 0, the result being exactly the mean of the training set,

$$\hat{X}_t = \frac{1}{N} \sum_{i=1}^N X_t^{(i)} . \quad (2.17)$$

(Notice that the forecasting method is not using the information that is trying to predict directly, but it uses indirectly if the parameters θ are fitted using the forecast accuracy; here for example, the method has an output equal to a constant value, that when fitted results in this equation.) The mean squared error computed with this forecasting method is equal to the variance of the training set, and can be used as a baseline to compare other forecasting methods [Brö09]. However, if X_t is distributed with a fat-tail, the variance is not defined if the exponent $\alpha < 2$, a common situation in social data. In finite datasets, this has the consequence that the variance of the training set can have large fluctuations since the empirical variance is a convergent estimator of a quantity that is not defined (in particular, the variance will depend on the size of the training set, N – see Section 5.4.1), another challenge to cope with if we desire to measure predictability.

Evaluation of accuracy in fat-tailed data can be approached in two ways. One of them is to explicitly measure the impact of the sample size. This can be done by using a *resampling technique* such as bootstrap aggregating [Bre96], i.e. creating sets of pairs training/target sets from the original dataset and considering the statistics of the accuracy of each pair. By these methods, for each size $N' < N$, a probability distribution of the accuracy is obtained; this approach is taken in Section 6.4.

The other approach is to consider instead forecasts that are robust with respect to the type of distribution that the variables have. If we assume that the target and the training sets have the same distribution (which is reasonable if the process is stationary or if both sets are an unbiased partition of the original dataset), then the simplest robust variables are the quantiles, i.e. the values of the type $\mathbb{P}(X_t > x)$, which, for values of x chosen accordingly, is the definition of extreme events given by the Second Extreme Value theorem. In summary, the prediction of extreme events, which in the context of attention are the *big hits*, is not valuable only because of the interest that these events have in itself, as mentioned before, but also because they constitute a robust test of accuracy for forecasting methods and, more generally, for models. This idea is a recurring theme throughout this Thesis.

3. Datasets used and relevant properties

In this chapter we describe the four datasets used throughout this Thesis. The datasets were chosen because they fit into the phenomenon that we want to study: each of them has proxies for attention, which are fat-tailed distributed, and can be tracked along time. As a result, from each original source, a dataset consisting of a large set of time series of online activity was produced. Two of these sources were collected by the author, while the other two are used with permission. A characterization of the datasets in terms of extreme value distributions is realized, and a partition by features is described. The datasets consist of

- views received by 16.2 million videos in YouTube.com between Jan. 2012 and Apr. 2013;
- posts written in 0.8 million threads in 9 different Usenet discussion groups between 1994 and 2008;
- votes to (all) 4.6 million questions published in Stack-Overflow between Jul. 2008 and Mar. 2013;
- views of (all) 72246 papers published in the journal PLOS ONE from Dec. 2006 to Aug. 2013.

All these datasets are openly provided in Ref. [MA14b].

3.1. Characterization of data

Each dataset is composed by a collection of time series of activity (views, posts, etc.), each corresponding to one of the items (videos, threads, etc.) in the original source. As explained in Section 2.4, the items have particular features, that characterize them: the *content*, the *metadata* (together denoted by $M^{(i)}$), and, of course, their past activity (denoted as a whole by $H_t^{(i)}$).

The content is, naturally, the item itself, what the user is actually interested in. The content is not always accessible, and in some cases, even if accessible, is not useful to make a prediction. An example of lack of accessibility is the previous example of papers (see Chapter 2): individually they are available to read, but a large collection of them to perform statistical analysis is more difficult to get and to automatically read. An example of low utility is the one of online videos: they can be collected in bulk, but its content is highly difficult to interpret by machines, because it is a process

that involves speech and image recognition, very complex fields on their own. If the content is available in machine-readable format (imagine a dataset of full-text of papers), this content has to be processed and converted into some numeric or categorical variables usable to forecast; in this process is where concepts like keyword spotting [LFL98] or sentiment analysis [DD10, CST⁺11] play an important role, since their goal is to create measures that represent features of the content that most humans do understand and most machines do not.

Metadata, instead, is simpler by definition. The metadata of an item is a set of machine-readable information about it that serves to categorize, i.e. to be able to find the item fast, by filtering of the relevant metadata values. The title of some item (or the abstract of a paper) is usually the most complex metadata that can be found, but typical examples include the length of a text, or a video, the publisher's name, his/her location and any other information available about him/her, the size of a file, the date of publication, the journal of a paper, etc. In this sense, if there is an algorithm that translates features of the content into alphanumerical values, then this values would become metadata, so there is metadata given at the moment of publication, and the metadata computed from the content. Here we are not analyzing the content, so we consider only the original metadata (given by the dataset without further analysis).

A partition in groups by a single feature of the metadata is performed to each dataset, for the purpose of analyzing later its impact on the predictability (see Chapter 3). In order to characterize in general terms the data, a fixed time after publication t_* was chosen for each database, so in this chapter X will be used to note X_{t_*} . The tails of the distribution $\mathbb{P}(X)$ of activity X received by the items at a time t_* after publication is characterized without loss of generality using Extreme Value Theory, as explained in Section 2.2.1. For large thresholds x_p the probability $P(X > x | X > x_p)$ follows a Generalized Pareto distribution [Col01]:

$$P(X > x | X > x_p) \approx \left(1 + \frac{x - x_p}{\sigma\alpha}\right)^{-\alpha} \quad (3.1)$$

for $x > x_p$. In our analysis it is essential to consider the discretization of the observations (specially for small values), since it has a direct impact on the estimation of the parameters [CSN09]. Therefore a discretized version of the Generalized Pareto Distribution is used, which has a probability mass function

$$\mathbb{P}(X = x) = \frac{(\sigma\alpha)^{-(\alpha+1)}}{\zeta(\alpha+1, \sigma\alpha)} \left(1 + \frac{x - x_p}{\sigma\alpha}\right)^{-\alpha-1}, \quad (3.2)$$

where ζ is the Hurwitz Zeta function.

3.2. Fit of fat-tailed distributions

Our goal is to characterize the tails of the distributions of the collected datasets, by fitting the Generalized Pareto distribution to the data at a given time t_* ; particular relevance has the parameter α , that defines how fast the distribution decays (the higher the α the faster is the decay of $P(X = x)$ with respect to x). The parameters of each distribution $P(X = x)$ are estimated by Maximum Likelihood Estimation [Gri93] (MLE). MLE consists in maximizing the probability of the data given the model parameters α and σ

$$\mathcal{L} = \prod_{i=1}^N \mathbb{P}(X = X^{(i)} \mid \alpha, \sigma; x_p) \quad . \quad (3.3)$$

This maximization is computationally done by minimizing the related quantity, $\ell = -\ln \mathcal{L}$, therefore equal to

$$\ell = - \sum_{i=1}^N \ln \mathbb{P}(X = X^{(i)} \mid \alpha, \sigma; x_p) \quad . \quad (3.4)$$

The value x_p is not a parameter, but is selected through a modification of the criterion proposed in Ref. [CSN09], which consists in selecting the lowest possible value of x_p for which the fit is statistically significant; here, x_p is selected as the lowest threshold that guarantees statistical significance for at least 80% of the groups analyzed for each database. Significance is estimated by the computation of the p -value, the probability of measuring in data sampled from the model a given statistic as extreme as in the data; the chosen statistic is the Kolmogorov-Smirnov [CSN09] (KS) distance between two distributions with cumulative density function F and G ,

$$KS = \max_x (|F_X(x) - G_X(x)|) \quad . \quad (3.5)$$

The result of the fitting procedure is reported in Table 3.1, where a time after publication t of interest is selected.

Additionally to the fit of the overall distribution, a fit of each partition of the database is realized. These fits yield $\alpha \in [0.50, 4.36]$ and are statistically significant already for relatively small x_p 's, with a p -value > 0.05 in 52 out of 59 fits. These results confirm the presence of fat tails, and indicate that our databases are representative of social media more generally (while scientific publications are usually not classified as social media items, from the point of view of their online views, they are subject to a similar attention-gathering process).

Database	α	σ	x_p	KS	p	N_{fit}	N	t_*
YouTube	0.69	55.3	20	0.003	0.00	7621233	13284409	20 days
Usenet	1.56	5.15	1	0.028	0.96	344608	871931	20 days
Stack-Overflow	2.03	7.7	15	0.027	0.38	12193	4397194	1 year
PLOS ONE	1.41	1083.22	2300	0.009	0.14	5948	72244	2 years

Table 3.1.: Summary of all the databases. Additionally to the parameters of the Generalized Pareto distribution, are reported the selected threshold x_p , the Kolmogorov-Smirnov distance between the data and the fit (KS), the p -value, the amount of items fitted by the distribution, N_{fit} (i.e. the ones that exceed x_p), the total amount of items in the database N , and the time t_* after publication that is selected to fit the distributions.

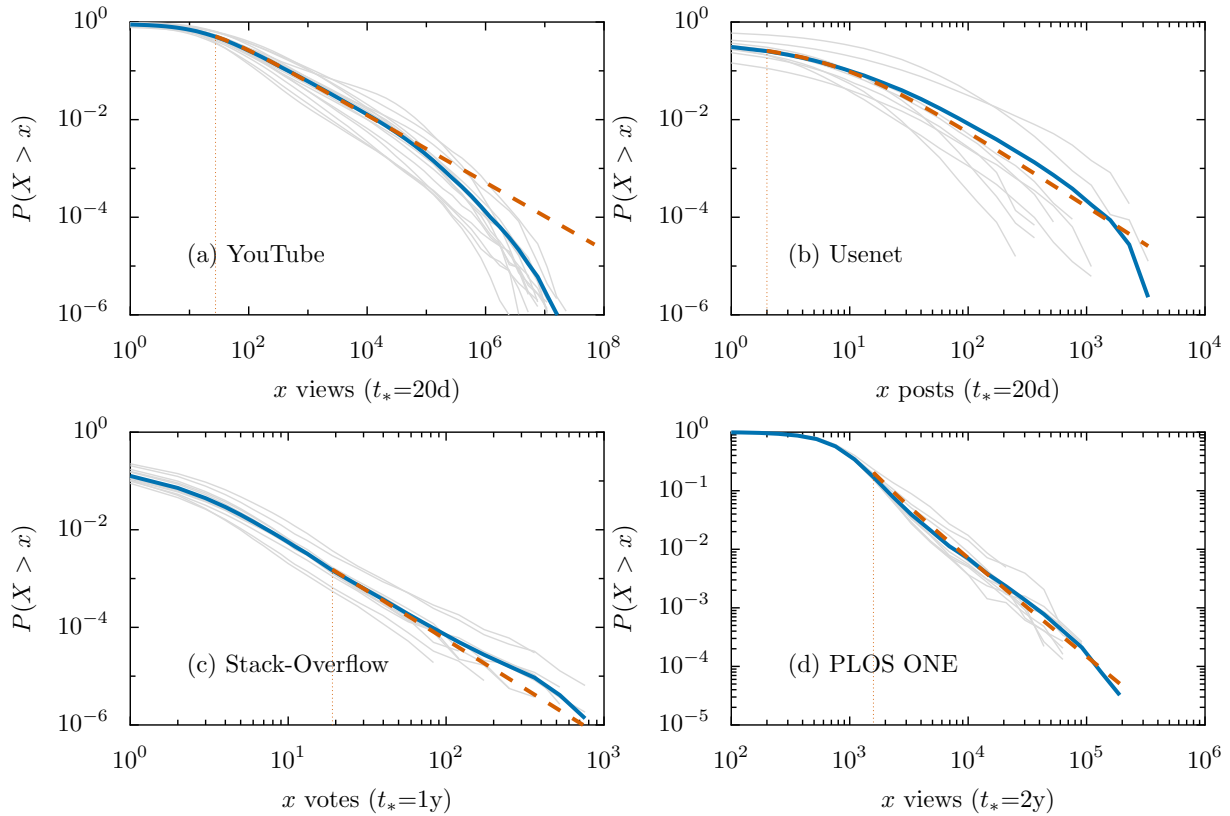


Figure 3.1.: Cumulative distribution functions for each dataset. Blue solid line: distribution for the combined data; red dashed line: fit of the generalized Pareto distribution for the combined data; dotted vertical red line: location of x_p , the threshold of the fitted distribution; gray solid lines: distribution for each of the metadata features selected.

3.3. YouTube

YouTube (www.youtube.com) is the main platform for sharing videos online, being the third more viewed website (at the time of the writing of this Thesis, according to Alexa rank). Some of

the videos published in it reached more than 1 billion people, and the data of views is recorded daily. The data was obtained through the YouTube API, a service that the parent company, Google, provides, using two particular functions: one that allows to retrieve a list of the last videos published with their metadata, updated approximately every four hours, and another one that retrieves the time series of views, shares and ratings.

The metadata available for each video includes the publication date, publisher, and *category*, one of fifteen predefined thematic groups selected by the users. The categories will be used to partition the dataset. The threshold used in the reported results is $x_p = 20$. The fit of the overall distribution here is not statistically significant: visually (see Fig. 3.1(a)), this can be observed in the deviation of the fit with respect to the data (around $x \sim 10^5$); nevertheless, the fit is significant for most of the individual groups (Table 3.2), whose deviations are tolerated by the lower amount of data available. The exact description of the distributions is not the goal here, but is worth to notice that the distribution of the data does not decay abruptly in the tail; instead, it seems that there is a second fat-tailed regime.

3.4. Stack Overflow

Stack Overflow (www.stackoverflow.com) is the most used platform of question and answers among programmers; these questions and answers compose the dataset. Additionally, the questions

Group	α	σ	KS	p	N_{fit}	$\mathbb{P}(g)$
Autos & Vehicles	0.75	62.9	0.005	0.78	711022	7.4%
Comedy	0.84	55.2	0.005	0.85	571719	8.3%
Education	0.61	45.3	0.009	0.07	502299	7.0%
Entertainment	0.88	46.2	0.007	0.54	494318	7.7%
Film & Animation	0.64	63.5	0.006	0.81	375084	5.1%
Gaming	0.58	74.8	0.009	0.12	575361	6.3%
Howto & Style	0.69	48.8	0.006	0.81	445804	6.4%
Music	0.70	59.3	0.006	0.84	530635	6.4%
News	0.50	51.7	0.017	0.02	232925	2.9%
Nonprofits & Activism	0.92	40.4	0.004	1.00	428223	7.3%
People & Blogs	0.97	33.8	0.007	0.57	476004	7.6%
Pets & Animals	0.71	101.1	0.017	0.00	752916	8.2%
Science & Technology	0.65	76.8	0.012	0.01	655954	7.1%
Sports	0.86	53.5	0.005	0.94	479078	6.6%
Travel & Events	0.70	43.8	0.005	0.99	389891	5.6%

Table 3.2.: Summary of the YouTube database. Along the estimated parameters of Eq. (3.2), we report the corresponding KS distance, the p -value of the fit, the amount of items fitted N_{fit} and $\mathbb{P}(g)$, the proportion of items of a given category with respect to the total.

and the answers are tagged to improve search and classification and can be voted, according to their quality and insight. The whole website is available at <http://archive.org/details/stackexchange>, and from their databases it is possible to extract time series of votes for each question.

A grouping of questions based on *tags* was used, and since each question has many tags, a classification procedure was performed. In particular, programming languages were used as the grouping factor, since the majority of tags are related to them. The 10 more common tags of this type were selected (see list in the table below), and the items with a tag if that tag was a substring of a longer, different, tag were merged into them. Similarly, tags that could be associated to a single programming language were also merged. The remaining tags were grouped in a group labeled **rest**, which included also all cases in the intersection of two or more programming language groups; this **rest** group is therefore much larger than others. The complete grouping of the tags can be seen in our publication, Ref. [MA14b] in the file "Stack-Overflow Lemmas". The threshold used for the fits is $x_p = 15$, and all these fits are significant (see Table 3.3), which can be corroborated by visual inspection in Fig. 3.1(c).

3.5. Usenet

Usenet is a worldwide distributed discussion system established in 1980, and is a widely used precursor to Internet forums, where users would debate on the most variate topics. Discussions are *threaded*, which means that there are original posts that can be replied by the other users, creating a sequence of posts on a particular topic.

Group	α	σ	KS	p	N_{fit}	$\mathbb{P}(g)$
.net	2.06	7.4	0.039	0.74	1802	12.2%
c	1.54	8.0	0.030	1.00	348	1.5%
c++	1.82	7.1	0.036	0.94	1125	3.5%
css	2.35	7.9	0.049	1.00	163	1.4%
html	1.86	6.5	0.050	1.00	114	1.0%
java	1.87	7.8	0.045	0.81	909	7.6%
javascript	1.83	7.9	0.031	0.99	747	8.6%
php	2.10	7.9	0.040	0.99	293	5.6%
python	2.11	6.5	0.055	0.91	516	4.1%
rest	2.14	8.0	0.029	0.76	5969	49.7%
sql	3.14	8.9	0.054	0.97	207	4.8%

Table 3.3.: Summary of the Stack Overflow database. Along the estimated parameters of Eq. (3.2), we report the corresponding KS distance, the p -value of the fit, the amount of items fitted N_{fit} and $\mathbb{P}(g)$, the proportion of items with a given language tag with respect to the total.

Topics are divided in *Discussion Groups*, which are a natural partition of the dataset, and is the partition used here. The dataset is a collection of threads used originally in Ref. [APM11], and retrieved through Google Groups (for example, for the discussion group on the operative system Linux, `comp.os.linux`, visit <https://groups.google.com/forum/#!forum/comp.os.linux>). The threshold used is $x_p = 1$, and while most of the fits are significant (see Table 3.4), the amount of items in the groups is heterogeneous, specially for the amount of posts that have 0 replies (see Fig. 3.1(b)).

3.6. PLOS ONE

The PLOS ONE dataset is a collection (72246) of publicly available scientific publications of the journal PLOS ONE, between Dec. 2006 and Aug. 2013. The data was retrieved through the API provided by PLOS (<http://api.plos.org>), but is identical to the one published in Ref. [FL13].

The amount of authors in a paper was chosen as the grouping factor (the group labeled 12 contains all the papers with 12 or more authors). The threshold used for the fits is $x_p = 1400$; the data and the fits for each category are quite similar (see Fig. 3.1(d) and Table 3.5), indicating that the number of authors is not a strong indicator of differences in the amount of views.

Group	α	σ	KS	p	N_{fit}	$\mathbb{P}(g)$
alt.rap	4.36	4.6	0.046	0.25	23112	7.3%
comp.os.linux.misc	3.60	3.4	0.061	0.00	69099	18.6%
rec.arts.poems	1.67	4.0	0.028	1.00	42199	23.5%
rec.arts.sf.written	1.15	8.2	0.020	0.23	39040	9.1%
rec.music.classical	2.03	5.6	0.030	0.60	36151	10.0%
rec.music.country.western	2.02	4.9	0.032	0.84	29131	8.6%
rec.music.hip-hop	3.36	5.8	0.033	0.83	48792	12.1%
sci.physics.fusion	2.08	4.5	0.033	0.89	4622	1.7%
talk.origins	1.63	12.7	0.017	0.10	52454	9.1%

Table 3.4.: Summary of the Usenet database. Along the estimated parameters of Eq. (3.2), we report the corresponding KS distance, the p -value of the fit, the amount of items fitted N_{fit} and $\mathbb{P}(g)$, the proportion of items of a given discussion group with respect to the total.

Group	α	σ	KS	p	N_{fit}	$\mathbb{P}(g)$
1	1.64	1262.7	0.043	0.46	187	1.1%
2	1.27	897.9	0.015	0.82	941	6.7%
3	1.61	904.5	0.016	0.44	1393	10.5%
4	1.47	686.5	0.009	0.97	1498	12.2%
5	1.52	723.4	0.019	0.10	1565	12.6%
6	1.59	672.3	0.014	0.66	1313	11.4%
7	1.50	682.6	0.018	0.49	1142	10.3%
8	1.65	643.5	0.023	0.13	931	8.6%
9	1.85	666.9	0.034	0.03	794	6.8%
10	1.60	642.7	0.033	0.07	587	5.3%
11	1.74	693.8	0.040	0.04	474	3.9%
12+	1.88	767.1	0.020	0.08	1665	10.6%

Table 3.5.: Summary of the PLOS database. Along the estimated parameters of Eq. (3.2), we report the corresponding KS distance, the p -value of the fit, the amount of items fitted N_{fit} and $\mathbb{P}(g)$, the proportion of items with a given amount of authors with respect to the total.

4. Extreme events predictability

4.1. Introduction

An important question is how to quantify the extent into which prediction of individual items is possible, i.e. their *predictability* [KAH⁺06]. Of particular interest –in social and natural systems– is the predictability of extreme events [AJK06, HAHK07, HK08, Sor02, GYH⁺11, BB11], the small number of items in the tail of the distribution that gather a substantial portion of the public attention.

As motivated in Section 2.4, measuring predictability is difficult because it is usually impossible to disentangle how multiple factors affect the quality of predictions. For instance, predictions of the future activity of individual items rely on (i) information on properties of the item (e.g., metadata or previous activity) and (ii) a prediction strategy that converts the information into predictions. The quality of the predictions reflect the interplay between these two factors and the dynamics of attention in the system. In particular, the choice of the prediction strategy is crucial. Instead, predictability is a property of the system and is by definition independent of the prediction strategy (it is the upper bound for the quality of any prediction based on the same information on the items). A proper measure of the predictability should provide direct access to the properties of the system, enabling a quantification of the importance of different information on the items in terms of their predictive power.

In this chapter we introduce a method to quantify the predictability of extreme events and apply it to data from social media, motivated in Section 4.2. This is done by formulating a simple prediction problem which allows for the computation of the optimal prediction strategy. In Section 4.3 we consider the problem of providing a binary (yes/no) prediction to whether or not an item will be an extreme event (attention passes a given threshold). Predictability is then quantified in Section 4.4 and Section 4.5 as the quality of the optimal strategy. In Section 4.6 we apply this method to the four different systems introduced in Chapter 3: views of YouTube videos, comments in threads of Usenet discussion groups, votes to Stack-Overflow questions, and number of views of papers published in the journal PLOS ONE. The most striking empirical finding is that in all cases the predictability increases for more extreme events (increasing threshold). We show that this observation is a direct consequence of differences in (the tails of) the distributions of attention conditioned by the known property about the items, as summarized in Section 4.7.

4.2. Robust estimation and Extreme Events

An important property of a statistic (a measure realized on a sample of data) is *robustness*, the capacity of the statistic to have a good performance even if outliers are present in the data. This quality is usually quantified through the concept of *breakdown point*, the smallest proportion of observations that can result in the statistic to take an arbitrary value (statistically incompatible with the true distribution). The outliers here are these problematic observations, uncommon points in the data that stands out against the rest.

As an example, if we consider the estimation of the (well-defined) mean of a variable X , μ_X , we can see that the breakdown point of the sample mean is $1/N$, i.e. a single point of the data is enough to change arbitrarily the value of the estimate of the mean. The mean is, from this point of view, the worst possible statistic that represent the location of an unknown distribution; in contrast, the median is the best possible statistic, since it has a breakdown point of $1/2$, the maximum achievable. The median is just the point of X that divides the total distribution in two parts with equal mass, a particular *quantile* of the distribution.

While the mean is the optimal statistic of the location of the normal distribution, it can be substantially sub-optimal for distributions close to the normal [VR13]. Filtering the outliers out of the data is not a solution to this problem. In the first place, a method to identify outliers from an unknown distribution has to be defined, but this is a rather subjective task, since a data point is an outlier always with respect to some underlying distribution, that must be assumed. In the second place, in some distributions, big, rare values cannot be considered outliers at all. In particular, fat-tailed distributions can be approximated by power-law scalings $\mathbb{P}(x) \approx x^{-\alpha-1}$, thus there is no characteristic scale after which an outlier can be defined. The mean does not even exists for $\alpha < 1$, and for $\alpha < 2$, the standard deviation does not exists, which means that the Central Limit Theorem cannot be used to estimate the distribution of the mean. In systems with fat-tailed distributions, such as social systems, if predictions are issued based on estimators of the moments there is a high chance of having unreliable results, therefore, there is a need for a different approach to prediction of social activity.

A solution comes from the previous example, using the median to replace the mean as an estimator of location. More generally, the quantiles' measurements are essentially robust: the quantile corresponding to q , x_q , is defined as the value such that

$$\mathbb{P}(X < x_q) = q, \tag{4.1}$$

with a the breakdown point of q if $q < 1/2$ and $1 - q$ if $q > 1/2$. From Eq. (4.1), we see that the quantile is the threshold that a value has to surpass to be considered extreme, when the probability of having such extreme values is $1 - q$. We therefore consider the problem of *event prediction* instead of the value prediction.

We say an extreme event E happens at time t if the cumulative activity until time t , $X_t^{(i)}$, is bigger than a fixed threshold x_* , $X_t > x_*$. The variable to be predicted for each item is then binary: E or \bar{E} (not E). As the observable is binary, we issue binary predictions for each item (E will occur or not), which is equivalent to a classification problem and different from a probabilistic prediction (E will occur with a given probability). Fat tails do not affect the robustness of the method because all items for which $X_t > x_*$ count the same (each of them as one event), regardless of their size x . Indeed, the tails of $\mathbb{P}(X_t > x_*)$ determine simply how the probability of an event $\mathbb{P}(E)$ depends on the threshold x_* .

4.3. Predictability of Events

In this section we introduce a method to quantify predictability based on the binary prediction of extreme events. This is done by arguing that, despite the apparent freedom to choose among different prediction strategies, it is possible to compute a single optimal strategy for this problem. We then show how the quality of prediction can be quantified and argue that the quality of the optimal strategy is a proper quantification of predictability.

Predictions are based on information on items which generally lead to a partition of the items in groups $g \in \{1, \dots, G\}$ that have the same feature [SB94]. As a simple example of our general approach, consider the problem of predicting at publication time $t = 0$ the YouTube videos that at $t_* = 20$ days will have more than $x_* = 1000$ views (about $\mathbb{P}(E) \approx 6\%$ of all videos succeed). As items' information, we use the category of a video so that, e.g., videos belonging to the category *music* correspond to one group g and videos belonging to *sport* correspond to a different group g' . Since the membership to a group g is the only thing that characterizes an item, predictive strategies can only be based on the probability of having E for that group, $\mathbb{P}(E | g)$.

In principle, one can think about different strategies on how to issue binary predictions on the items of a group g . They can be based on the likelihood (L) $\mathbb{P}(E | g)$ or on the posterior (P) probability $\mathbb{P}(g | E)$ [HAHK07], and they can issue predictions stochastically (S), with rates proportional to the computed probabilities, or deterministically (D), only for the groups with largest $\mathbb{P}(g | E)$ or $\mathbb{P}(E | g)$. These simple considerations lead to four (out of many) alternative strategies to predict events (raise alarms) for items in group g

- (LS)** stochastically based on the likelihood, i.e., with probability $\min\{1, \beta \mathbb{P}(E | g)\}$, with $\beta \geq 0$;
- (LD)** deterministically based on the likelihood, i.e., always if $\mathbb{P}(E | g) > p_*$, with $0 \leq p_* \leq 1$;
- (PS)** stochastically based on the posterior, i.e., with probability $\min\{1, \beta' \mathbb{P}(g | E)\}$, with $\beta' \geq 0$;
- (PD)** deterministically based on the posterior, i.e., always if $\mathbb{P}(g | E) > p'_*$, with $0 \leq p'_* \leq 1$.

In the limit of large number of predictions (items), the fraction of events that strategy (LS) predicts for each group g matches the probability of events $\mathbb{P}(E | g)$ and therefore strategy (LS) is

reliable [Brö09] and can be considered a natural extension of a probabilistic predictor. Predictions of strategies (LD), (PS) and (PD) do not follow $\mathbb{P}(E | g)$ and therefore they are not reliable.

Being now the observable to predict binary, a way of assessing the quality of prediction must be defined. Comparing predictions and observations gives four possible results, given by the combination of the prediction (positive or negative) and its success (true or false). If A denotes the prediction of an event (an alarm), the hit rate (or True Positive Rate) and the false alarm rate (or False Positive Rate) are defined as

$$\begin{aligned} \text{hit rate} &\equiv \frac{\text{number of true positives}}{\text{number of positives}} = \mathbb{P}(A | E), \\ \text{false alarm rate} &\equiv \frac{\text{number of false positives}}{\text{number of negatives}} = \mathbb{P}(A | \bar{E}). \end{aligned} \tag{4.2}$$

These are analogous to measures like Accuracy and Specificity or Precision and Recall [BYRN99]. Prediction strategies typically have a specificity parameter (e.g., controlling the rate of false positives). Varying this parameter, a prediction curve that goes from $(0,0)$ to $(1,1)$ is built in the hit \times false-alarm space, the *ROC curve* (see Fig. 4.1(a)). The amount of desired false alarms of the strategy (β, p_*, β' , and p'_* in the examples above) are the specificity parameters by definition.

The overall quality is usually measured by the area below this curve, known as Area Under the Curve (AUC) [HM82]. For convenience, we use the area between the curve and the diagonal (hits=false-alarms), $\Pi = 2\text{AUC} - 1$ (equivalent to the Gini coefficient). In this way, $\Pi_S \in (-1,1)$ represents the improvement of strategy S against a random prediction. In absence of information $\Pi_S = 0$ and perfect predictions lead to $\Pi = 1$. In the YouTube example considered above, we obtain $\Pi_{PS} < \Pi_{LS} < \Pi_{PD} < \Pi_{LD}$ (17%, 18%, 29%, 32%), indicating that strategy (LD) is the best one.

We now argue that, as it will be shown in detail in Section 4.4, that Strategy (LD) is optimal (or *dominant* [PFK98]), i.e., for any false alarm rate it leads to a larger hit rate than any other strategy based on the same set of $\mathbb{P}(E | g)$. The strategies listed lead to piecewise linear functions (see Fig. 4.1(b)), and the True Positive Rate can be maximized by using the Simplex algorithm, a very popular linear programming method. In fact, strategy (LD) is the only ordering of the groups that enforces convexity in the hit \times false-alarms rates space, a property that, as the Simplex algorithm shows, maximizes the True Positive Rate if the False Positive Rate is constrained (see Section 4.4 for a formal derivation).

The ranking of the groups by $\mathbb{P}(E | g)$ implies a ranking of the items, an implicit assumption in the measure of the performance of classification rules [HM82, HT01]. The existence of an optimal strategy implies that the freedom in choosing the prediction strategy argued above is not genuine and that we can ignore the alternative strategies. In our context, it implies that the performance of the optimal strategy measures a property of the system (or problem), and not simply the efficiency of a particular strategy. Therefore, we use the quality of prediction of

the optimal strategy ($\Pi \equiv \Pi_{LD}$) to quantify the predictability (i.e., the potential prediction) of the system for the given problem and information. By geometrical arguments we obtain from Fig. 4.1(b) (see Section 4.5)

$$\Pi = \sum_g \sum_{h < g} \frac{\mathbb{P}(g)\mathbb{P}(h) (\mathbb{P}(E | h) - \mathbb{P}(E | g))}{\mathbb{P}(E)(1 - \mathbb{P}(E))}, \quad (4.3)$$

where $\mathbb{P}(g)$ is the probability of group g and g is ordered by decreasing $\mathbb{P}(E | g)$, i.e., $h < g \Rightarrow \mathbb{P}(E | h) > \mathbb{P}(E | g)$.

The value of Π can be interpreted as the probability of a correct classification of a pair of E and \bar{E} items [HM82, HT01]. In practice, the optimality of this strategy is dependent on the estimation of the ordering of the groups according to $\mathbb{P}(E | g)$. Wrong ordering may occur due to finite sampling on the training dataset or lack of stationarity in the data. In fact, any permutation of indexes in Eq. (4.3) reduces Π .

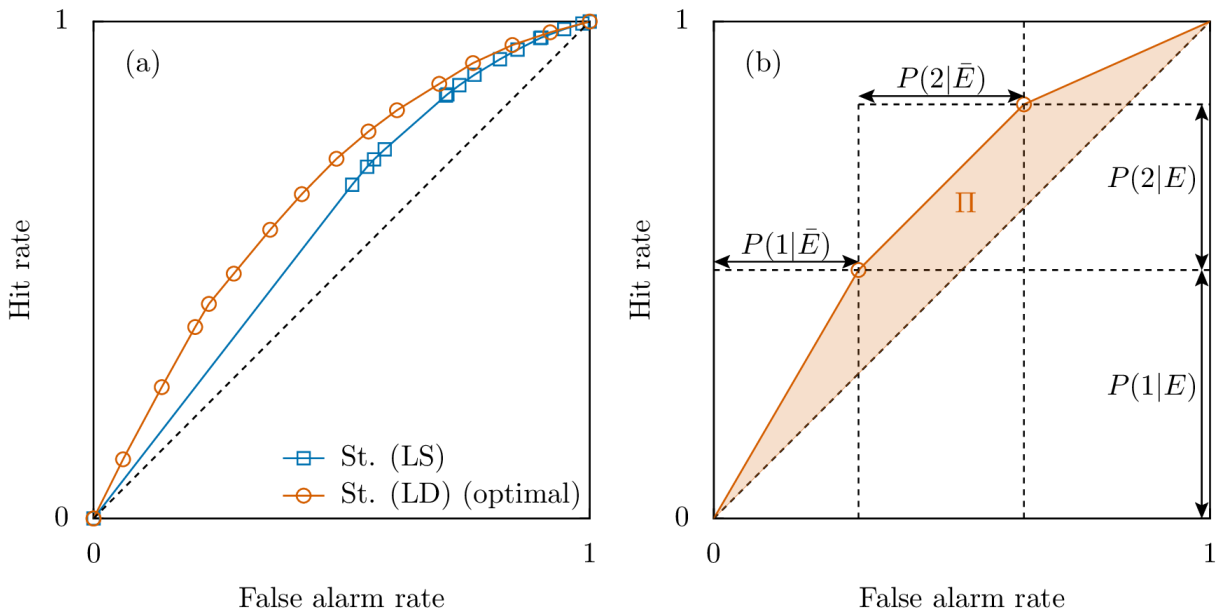


Figure 4.1.: Quantifying the quality of event-prediction strategies requires measuring both the hit and false alarm rates. (a) Performance of Strategy (LS) and Strategy (LD) for the problem of predicting views of YouTube videos 20 days after publication based on their categories. The symbols indicate where the rate of issued predictions for a given group equals 1 (the straight lines between the symbols are obtained by issuing predictions randomly with a growing rate). (b) Illustration of the prediction curve (red line) for an optimal strategy with three groups $g = 1, 2, 3$ with $\mathbb{P}(1) = \mathbb{P}(2) = \mathbb{P}(3) = 1/3$ and $\mathbb{P}(E | 1) = 0.3, \mathbb{P}(E | 2) = 0.2, \mathbb{P}(E | 3) = 0.1$.

4.4. Proof that strategy LD (Bayes classifier) is dominant

A strategy is dominant when for any given false alarm rate, the hit rate is maximized. Following the definition in Eq. (4.2), we write the x and y coordinates of the hit \times false-alarm plot as

$$\begin{aligned} \text{hit rate} &\equiv \mathbb{P}(A | E) = \sum_{g=1}^G \mathbb{P}(A | g) \mathbb{P}(g | E) = \sum_{g=1}^G \pi_g y_g \equiv y, \\ \text{false-alarm rate} &\equiv \mathbb{P}(A | \bar{E}) = \sum_{g=1}^G \mathbb{P}(A | g) \mathbb{P}(g | \bar{E}) = \sum_{g=1}^G \pi_g x_g \equiv x, \end{aligned} \quad (4.4)$$

where for notational convenience $y_g \equiv \mathbb{P}(g | E)$, $x_g \equiv \mathbb{P}(g | \bar{E})$, and $\pi_g \equiv \mathbb{P}(A | g)$. Since predictions are issued based only on the information about the groups, strategies (both deterministic and stochastic) are defined uniquely by π_g , while x_g and y_g are estimated from data. The computation of the dominant strategy corresponds to finding the set $\{\pi_g\}_{g < G}$ that maximize y with the constraint $\sum_{g=1}^G \pi_g x_g = x$. This problem can be solved exactly by applying the Simplex method [Dan98]. Define h such that $\sum_{g < h} x_g < x < \sum_{g \leq h} x_g$; we write Eq. (4.4) as:

$$\begin{aligned} y - \sum_{g < h} y_g &= - \sum_{g < h} (1 - \pi_g) y_g + \sum_{g > h} \pi_g y_g + \pi_h y_h, \\ x - \sum_{g < h} x_g &= - \sum_{g < h} (1 - \pi_g) x_g + \sum_{g > h} \pi_g x_g + \pi_h x_h. \end{aligned}$$

Isolating π_h in the lower equation and introducing it in the top one we obtain

$$y = \sum_{g < h} y_g + x \frac{y_h}{x_h} - \sum_{g < h} (1 - \pi_g) x_g \left(\frac{y_g}{x_g} - \frac{y_h}{x_h} \right) + \sum_{g > h} \pi_g x_g \left(\frac{y_g}{x_g} - \frac{y_h}{x_h} \right). \quad (4.5)$$

Notice that y_g/x_g is the contribution of the group g to the slope of the prediction curve in the hit \times false-alarm space. If the G groups are ordered by decreasing $\mathbb{P}(E | g)$, then y_g/x_g also decreases with g . Therefore $(y_g/x_g - y_h/x_h) > 0$ for $g < h$ and $(y_g/x_g - y_h/x_h) < 0$ for $g > h$ and Eq. (4.5) is maximized by choosing π_g such that the two last terms vanish. This is achieved choosing

$$\pi_g = \begin{cases} 1 & g < h, \\ \frac{x - \sum_{g < h} x_g}{x_h} & g = h, \\ 0 & g > h, \end{cases} \quad (4.6)$$

which corresponds to issuing positive predictions only to the h groups with largest $\mathbb{P}(E | g)$ and is equivalent to the strategy (LD). Positive events are predicted for the group h in Eq. (4.6) as much as needed to reach the required false positive rate x .

4.5. Computation of Predictability for the optimal strategy

As illustrated in Fig. 4.1(b), the partition performed by the optimal strategy defines G different intervals in the hit and false alarm axis (the points for which $\mathbb{P}(E | g) = P_*$, $g \in \{1 \dots G\}$) and therefore G^2 rectangles in the hit \times false-alarm space. The (g, h) rectangle has height $\mathbb{P}(h)\mathbb{P}(E | h) / \mathbb{P}(E)$, and therefore it is equal to $\mathbb{P}(h | E)$, and a width $\mathbb{P}(g | \bar{E})$. Its area is then $A_{g,h} = \mathbb{P}(h | E) \mathbb{P}(g | \bar{E})$. The curve of strategy (LD) is the union of the diagonals of the $g = h$ rectangles (which are obtained by increasing p_*). Π is two times the sum of the rectangles and triangles under this curve minus half of all the area:

$$\begin{aligned}
 \Pi &= 2 \left[\sum_g \sum_{h < g} A_{g,h} + \frac{1}{2} \sum_g A_{g,g} - \frac{1}{2} \sum_g \sum_h A_{g,h} \right] \\
 &= \sum_g \sum_{h < g} A_{g,h} - \sum_g \sum_{h > g} A_{g,h} \\
 &= \sum_g \sum_{h < g} (A_{g,h} - A_{h,g}) \\
 &= \sum_g \sum_{h < g} \mathbb{P}(h | E) \mathbb{P}(g | \bar{E}) - \mathbb{P}(h | \bar{E}) \mathbb{P}(g | E) \\
 &= \frac{\sum_g \sum_{h < g} \mathbb{P}(g) \mathbb{P}(h) (\mathbb{P}(E | h) - \mathbb{P}(E | g))}{\mathbb{P}(E)(1 - \mathbb{P}(E))},
 \end{aligned}$$

where we used $\sum_g \sum_h A_{g,h} = 1$.

4.6. Application to Data

Here we apply our methodology to the four social-media data described above. We consider the problem of predicting at time $t_1 \geq 0$ whether the attention $X_{t_*}^{(i)}$ of an item at time $t_* > t_1$ passes a threshold x_* . In practice, the calculation of Π from the data is done counting the number of items: (i) in each group g ($\mathbb{P}(g) = \{\# \text{ items in } g\} / \{\# \text{ items}\}$); (ii) that lead to an event ($\mathbb{P}(E) = \{\# \text{ items that cross the threshold } x_* \text{ at } t_*\} / \{\# \text{ items}\}$); and (iii) that lead to an event given that they are in group g ($\mathbb{P}(E | g) = \{\# \text{ items in } g \text{ that cross the threshold } x_* \text{ at } t_*\} / \{\# \text{ items in } g\}$). Finally, the groups are numbered as $g = 1, 2, \dots, G$ by decreasing $\mathbb{P}(E | g)$ and the sum over all groups is computed as indicated in Eq. (4.3). In Ref. [MA14b] a python script, which performs this calculation in the data, is provided.

We report the values of Π obtained from Eq. (4.3) considering two different information on the items:

- 1) the attention at prediction time X_{t_1} ;
- 2) information available at publication time $t = 0$ (metadata).

In case 1), a group g corresponds to items with the same X_{t_1} . These groups are naturally ordered in terms of $\mathbb{P}(E | g)$ by the value of X_{t_1} and therefore the optimal strategy is equivalent to issue positive prediction to the items with X_{t_1} above a certain threshold. In case 2), the groups correspond to items having the same meta-data (e.g., belonging to the same category). In this case, we order the groups according to the empirically observed $\mathbb{P}(E | g)$ (as discussed above). Before performing a systematic exploration of parameters, we illustrate our approach in two examples.

YouTube

Consider the case of predicting whether YouTube videos at $t_* = 20$ days will have more than $x_* = 1,000$ views. For case 1), we use the views achieved by the items after $t_1 = 3$ days and obtain a predictability of $\Pi = 90\%$. For case 2), we obtain that using the day of the week to group the items leads to $\Pi = 3\%$ against $\Pi = 31\%$ obtained using the categories of the videos. This observation, which is robust against variations of x_* and t_* , shows that the category but not the day of the week is a relevant information in determining the occurrence of extreme events in YouTube. Previous views, however, are much more informative than categories.

PLOS ONE

Consider the problem of identifying in advance the papers published in the online journal PLOS ONE that received at least 7500 views 2 years after publication, i.e. $t_* = 2$ years and $x_* = 7500$ (only $\mathbb{P}(E) = 1\%$ achieve this threshold). For case 1), knowing the number of views at $t_1 = 2$ months after publication leads to a predictability of $\Pi = 93\%$. For case 2), a predictability $\Pi = 19\%$ is achieved alone by knowing the number of authors of the paper –surprisingly, the chance of achieving a large number of views decays monotonously with number of author (g increases with number of authors).

The examples above show that formula Eq. (4.3) allows for a quantification of the importance of different factors (e.g., number of authors, early views to the paper) to the occurrence of extreme events, beyond correlation and regression methods (see also Ref. [PPP⁺13]). Besides the quantification of the predictability of specific problems, by systematically varying t_1 , t_* , and x_* we can quantify how the predictability changes with time and with event magnitude. The significant finding in this systematic analysis is that in all tested databases and grouping strategies the predictability increases with x_* , i.e., extreme events become increasingly more predictable, as shown in Fig. 4.2.

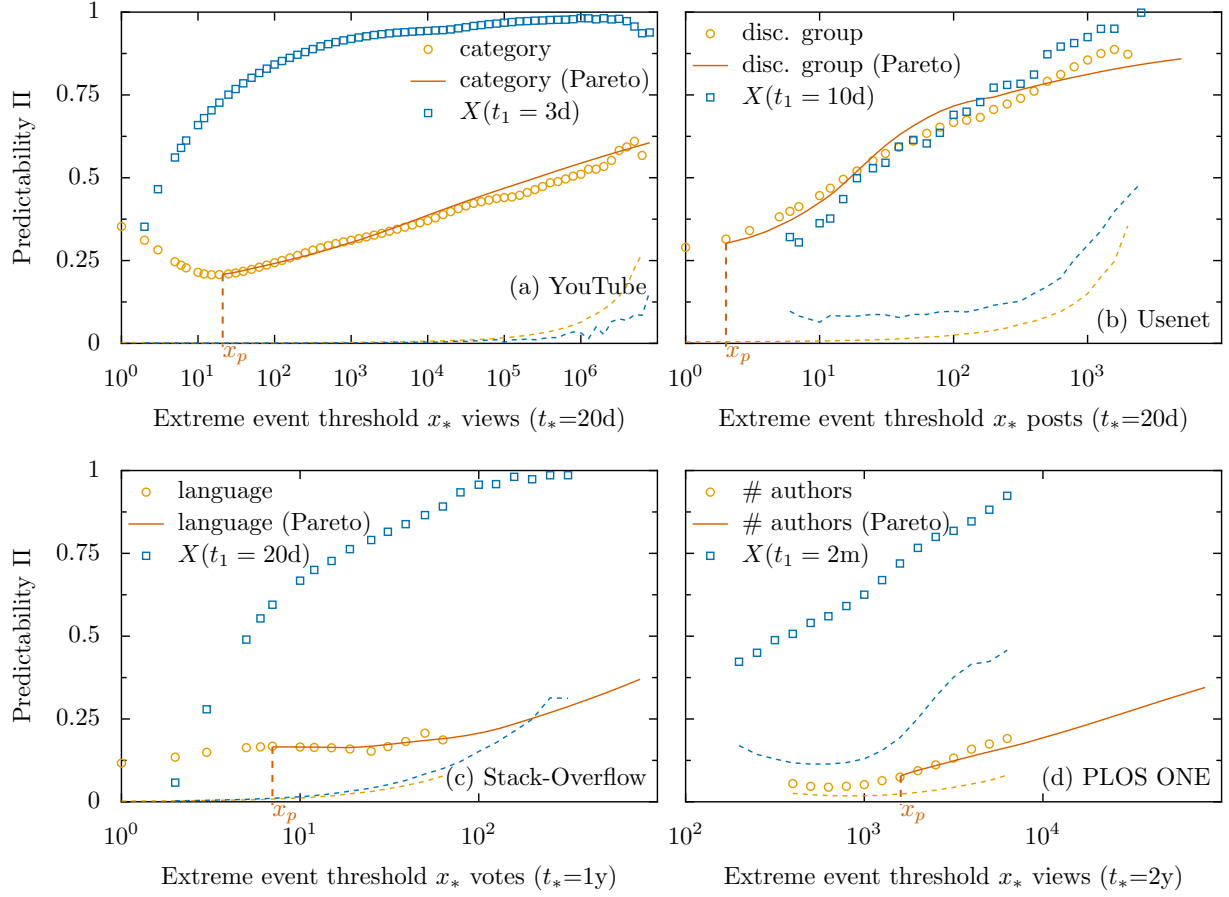


Figure 4.2.: Predictability increases for extreme events. If the attention an item receives at time t_* is above a threshold, $X_{t_*} > x_*$, an event E is triggered. The plots show how the predictability Π changes with x_* using two different information to combine the items in groups $\{g\}$. Black circles: Π at time $t = 0$ using metadata of the items to group them. The red lines are computed using as probabilities $\mathbb{P}(E | g)$ the Extreme Value distribution fits for each group at a threshold value x_p , see Eq. (2.7) and Section 3.1. Blue squares: Π at time $t_1 < t_*$ using X_{t_1} , i.e., the attention the item obtained at day t_1 . The dashed lines are the values of the 95% percentile of the distribution generated by measuring Π in an ensemble of databases obtained shuffling the attribution of groups (g) to items (the colors match the symbols and symbols are shown only where Π is at least twice this value). Results for the four databases are shown (see Chapter 3 for details): (a) YouTube (X : views of a video; metadata: video category); (b) Usenet discussion groups (X : posts in a thread; metadata: discussion group of the thread); (c) Stack-Overflow (X : votes to a question; metadata: programming language of the question); (d) PLOS ONE (X : online views of a paper; metadata: number of authors of the paper).

4.7. Discussion

4.7.1. Dependence of Predictability with respect to extreme events

We now explain why predictability increases for extreme events (increasing x_*). This increase is not due to the reduction of the number of events $\mathbb{P}(E)$. Consider the case in which E is defined in the interval $[x_f - \Delta x, x_f + \Delta x)$, instead of $[x_*, \infty)$. Assuming $\mathbb{P}(X)$ to be smooth in X , for $\Delta x \rightarrow 0$ at fixed x_f we have that $\mathbb{P}(E) \rightarrow \mathbb{P}(x_f)\Delta x$ and $\mathbb{P}(E | g) \rightarrow \mathbb{P}(x_f | g)\Delta x$ ($\mathbb{P}(g)$ remains unaffected), and Eq. (4.3) yields

$$\Pi = \frac{\sum_g \sum_{h>g} \mathbb{P}(g)\mathbb{P}(h) (\mathbb{P}(x_f | h) - \mathbb{P}(x_f | g))}{\mathbb{P}(E_f)[1 - \Delta x \mathbb{P}(x_f)]}, \quad (4.7)$$

which decreases with $\Delta x \rightarrow 0$. This shows that the increased predictability with x_* is not a trivial consequence of the reduction of $\mathbb{P}(E)$ ($\Delta x \rightarrow 0$), but instead is a consequence of the change in $\mathbb{P}(E | g)$ for extreme events E .

Systematic differences in the tails of $\mathbb{P}(X | g)$ lead to an increased predictability of extreme events. Consider the case of two groups with cumulative distributions $\mathbb{P}(E | g)$ that decay as a power law as in Eq. (2.7) with exponents α and $\alpha' = \alpha + \epsilon$, with $\mathbb{P}(1) = \mathbb{P}(2)$. From Eq. (4.3), Π for large x_* ($1 - \mathbb{P}(E) \approx 1$) can be estimated as

$$\Pi = \frac{1}{4} \frac{\mathbb{P}(E | 1) - \mathbb{P}(E | 2)}{\mathbb{P}(E | 1) + \mathbb{P}(E | 2)} = \frac{1}{4} \frac{x_*^{-\alpha} - x_*^{-(\alpha+\epsilon)}}{x_*^{-\alpha} + x_*^{-(\alpha+\epsilon)}} \approx \frac{1}{8} \log(x_*)\epsilon, \quad (4.8)$$

where the approximation corresponds to the first order Taylor expansion around $\epsilon = 0$. The calculation above can be directly applied to the results we obtained issuing predictions based on metadata. The logarithmic dependency in Eq. (4.8) is consistent with the roughly linear behavior observed in Fig. 4.2(a,b). A more accurate estimation is obtained using the power-law fits of Eq. (2.7) for each group g , which were performed in Chapter 3, and introducing the $\mathbb{P}(E | g)$ obtained from these fits in Eq. (4.3). The red line in Fig. 4.2 shows that this estimation agrees with the observations for values $x_* \gtrsim x_p$, the threshold used in the fit. Deviations observed for $x_* \gg x_p$ (e.g., for PLOS ONE data in panel (d)) reflect the deviations of $\mathbb{P}(E | g)$ from the Pareto distribution obtained for small thresholds $x_p \ll x_*$. This allows for an estimation of the predictability for large thresholds x_* even in small datasets (when the sampling of E is low).

A similar behavior is expected when prediction is performed based on the attention obtained at short times t_1 . Eq. (4.7) applies in this case too and therefore the increase in predictability is also due to change in $\mathbb{P}(E | g)$ with x_* for different g (and not, e.g., due to the decrease of $\mathbb{P}(E)$). For increasingly large x_* the items with significant probability of passing the threshold concentrate on the large x_{t_1} and increase the predictability of the system. We have verified that this happens already for simple multiplicative stochastic processes, such as the geometric Brownian motion

(see Fig. 4.3). This provides further support for the generality of this finding. The dynamics of attention in specific systems affect the shape of predictability growth with respect to the threshold.

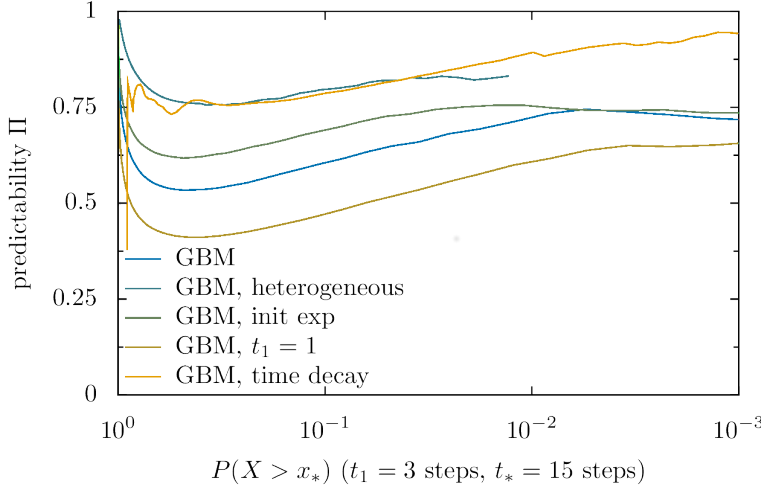


Figure 4.3: Predictability of simple stochastic processes. An ensemble of random walkers evolve through the dynamics $X_{t+1}^{(i)} = X_t^{(i)}(1 + \varepsilon)$, where $\varepsilon \sim \mathcal{N}(\mu_i, \sqrt{\mu_i})$, a Geometric Brownian Motion with Gaussian steps. The predictability of extreme events Π was computed for $t_1 = 3$ steps and $t_* = 15$ steps. GBM: $\mu_i = 2 \forall i$ and $X_0 \sim \mathcal{U}(0, 1)$; GBM heterogeneous: $\mu_i \sim \mathcal{N}(2, 0.7)$ and $X_0 \sim \mathcal{U}(0, 1)$, fixed in time; GBM, init exp: $\mu_i = 2 \forall i$ and $X_0 \sim \mathcal{E}(1/6)$; GBM, $t_1 = 1$ the same as GBM for $t_1 = 1$; GBM, time decay: model proposed in Ref. [WSB13], similar to GBM heterogeneous but with a rate that decays in time ($X_{t+1}^{(i)} = X_t^{(i)}(1 + \varepsilon f_i(t))$ with $\mu_i \sim \mathcal{N}(1, 0.5)$; $f_i(t)$ is a Lognormal surviving probability with parameters $\mu_i^t \sim \mathcal{LN}(6.5, 0.5)$ and $\sigma \sim \mathcal{LN}(1, 0.2)$).

Altogether, we conclude that the difference in (the tails of) the distribution of attention of different groups g is responsible for the increase in predictability for extreme events: for large x_* , any informative property on the items increases the relative difference among the $\mathbb{P}(E | g)$. This corresponds to an increase of the information contained in the grouping which leads to an increase in Π .

4.7.2. Probabilistic Forecast

In this chapter we used a deterministic forecast, i.e. the forecast is stating if the event E is going to happen or not. As discussed in Section 2.4, a different type of forecast is the probabilistic forecast, where the forecast consists in expressing the probability for something to happen, in our case, an extreme event.

The most used performance measure for probabilistic forecasts is the Brier Score [Bri50], which is defined as

$$BS = \frac{1}{N} \sum_{i=1}^N (\pi_i - o_i)^2, \quad (4.9)$$

where π_i is the forecasted probability of an event for the i -item, and o_i is an indicator variable of

the event, which value is 1 if the item i crosses the threshold and 0 otherwise. The Brier score can be conveniently decomposed [Bri50, Brö09] in three terms,

$$BS = \bar{o}(1 - \bar{o}) - \frac{1}{N} \sum_{g=1}^G N_g(\bar{o}_g - \bar{o})^2 + \frac{1}{N} \sum_{g=1}^G N_g(\pi_g - \bar{o}_g)^2 \quad , \quad (4.10)$$

where the items are partitioned in groups with the same π_g (equivalent to partition the groups arbitrarily and assign the same probability π_g to each item of the same group, $\pi_i = \pi_g \forall i \in g$). The three terms have an interpretation. The first, the *uncertainty*, is the baseline score, achievable by just predicting for all items the probability of an event to happen, \bar{o} . The second, the *resolution*, is the deviation of \bar{o}_g from respect to its average value \bar{o} . The third term is the *reliability*, and is the deviation of the forecasting method with respect to the rate for the group g .

The elements of Eq. (4.10) can be readily replaced by the concepts used in this chapter: the average probability of an event to occur is $\bar{o} = \mathbb{P}(E)$, and it can be conditioned by the group, $\bar{o}_g = \mathbb{P}(E | g)$. We get

$$BS = \mathbb{P}(E)(1 - \mathbb{P}(E)) - \sum_{g=1}^G \mathbb{P}(g)(\mathbb{P}(E | g) - \mathbb{P}(E))^2 + \sum_{g=1}^G \mathbb{P}(g)(\pi_g - \mathbb{P}(E | g))^2 \quad . \quad (4.11)$$

It becomes clear that the best strategy here is the one that makes the reliability term to vanish, i.e. $\pi_g = \mathbb{P}(E | g)$, which is related to the previously defined strategy (LS) (stochastic based on the likelihood $\mathbb{P}(E | g)$), although not equal, because here the forecast is a probability, while in the LS strategy the forecast is binary, and attributed stochastically among the items of the same group. The baseline score is the uncertainty $\mathbb{P}(E)(1 - \mathbb{P}(E))$, so the maximum increase in score is equal to the resolution term, which quantifies the difference between the average behavior and the conditional one, as the measure Π proposed does. In the context of binary deterministic prediction, there is no intrinsic baseline like the uncertainty, since the True Positive Rate would match the False Positive Rate if we would predict items with a fixed probability, resulting in a null predictability.

4.7.3. Conclusions

In summary, we propose a method, Eq. (4.3), to measure the predictability of extreme events for any given available information on the items. We applied this measure to four different social media databases and quantified how predictable the attention devoted to different items is and how informative are different properties of the items. We quantified the predictability due to metadata available at publication date and due to the early success of the items and found that usually the latter quickly becomes more relevant than the former. Our results can also be applied for combinations of different information on the items (e.g., a group g can be composed by videos

in the category *music* with a fixed x_{t_1}). In practice, the number of groups G should be much smaller than the observations in the training dataset to ensure an accurate estimation of $\mathbb{P}(E \mid g)$. The most striking finding is that extreme events are better predictable than non-extreme events, a result previously observed in physical systems [HK08] and in time-series models [HAHK07, BB11]. For social media, this finding means that for the large attention catchers the surprise is reduced and the possibilities to discriminate success enhanced.

These results are particularly important in view of the widespread observation of fat-tailed distributions of attention, which imply that extreme events carry a significant portion of the total public attention. Similar distributions appear in financial markets, in which case our methodology can quantify the increase in predictability due to the availability of specific information (e.g., in Ref. [PMS13] Internet activities were used as information to issue predictions). For the numerous models of collective behavior leading to fat tails [Pri76, SSPA10, WFVM12, ORT10, RFF⁺10, PPP⁺13, WSB13], the predictability we estimate is a bound to the quality of binary event predictions. Furthermore, the identifications of the factors leading to an improved predictability indicate which properties should be included in the models and which ones can be safely ignored (feature selection). For instance, the relevant factors identified in our analysis should affect the growth rate of items in rich-get-richer models [RFF⁺10, Per14] or the transmission rates between agents in information-spreading models [CFL09]. The use of Π to identify relevant factors goes beyond simple correlation tests and can be considered as a measure of causality in the sense of Granger [Gra80].

Predictability in systems showing fat tails has been a matter of intense debate. While simple models of self-organized criticality suggest that prediction of individual events is impossible [BP95], the existence of predictable mechanisms for the very extreme events has been advocated in different systems [Sor02]. In practice, predictability is not an yes/no question [SDW06, KAH⁺06] and the main contribution of this paper is to provide a robust quantification of the predictability of extreme events in systems showing fat-tailed distributions.

5. Stochastic dynamics of growth processes

5.1. Introduction

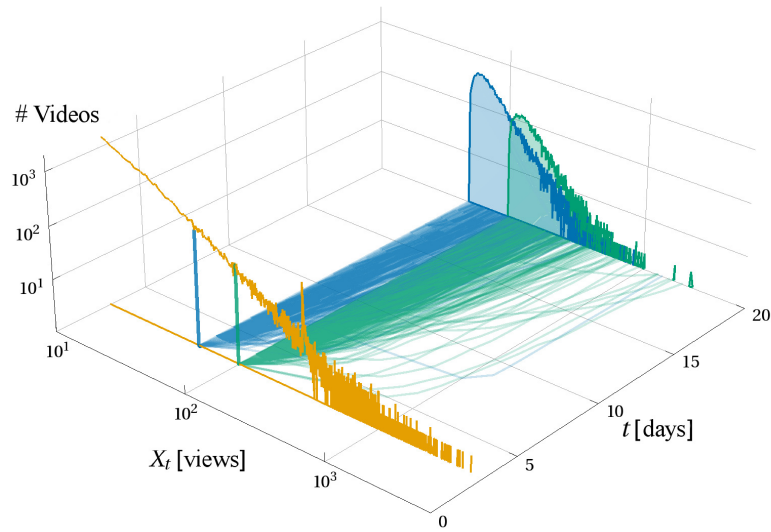


Figure 5.1.: Mixing of views dynamics The sets of videos selected according to their earlier success will rapidly mix, since they evolve into broad distributions. However, these final distributions are similar and depend on the initial success. Orange histogram: distribution of views 3 days after publication (0.3 million videos from our database). Blue and green lines: trajectories of the videos which had the same early success 3 days after publication (50 and 100 views respectively). Blue and green histograms: distributions of views 20 days after publication of the respective groups of items selected.

In this chapter we use our largest dataset (YouTube) to investigate the dynamics of activity with proportional effect, the idea on which most growth model rely (see Chapter 2). We consider models that are fully determined by the distributions $\mathbb{P}(dX_t | X_t)$ (i.e. with the *Markov property*), being X_t and dX_t the activity at time t and its increment respectively. Such distributions will define the overall evolution in time, as seen in Fig. 5.1, where the activity of videos with 50 and 100 views three days after publication is shown.

As described in Chapter 2, most growth models rely on the idea of proportional effect. We considered part of this class of models any process such that the expectation of the increment in

the activity is proportional to the total activity, Eq. (2.14)),

$$\mathbb{E}(dX_t | X_t) = aX_t, \quad (5.1)$$

where the increment $dX_t^{(i)}$ is defined as

$$dX_t^{(i)} = X_{t+1}^{(i)} - X_t^{(i)} \quad (5.2)$$

and the units of t are the minimal time step possible given by the data.

From the point of view of data analysis, this equation implies that the sample average of the increments over all the items with the same X_t has to be proportional with respect to X_t

$$\langle dX_t | X_t \rangle = aX_t. \quad (5.3)$$

This is the statistical foundation of the averaging method described in Section 2.3.

In Fig. 5.2 an example of this type of measure is shown, where the amount of views that videos receive in the third day after its release is related with the amount of views it received up to that point. However, if the aim is to analyze the full distribution $\mathbb{P}(dX_t | X_t)$, it is

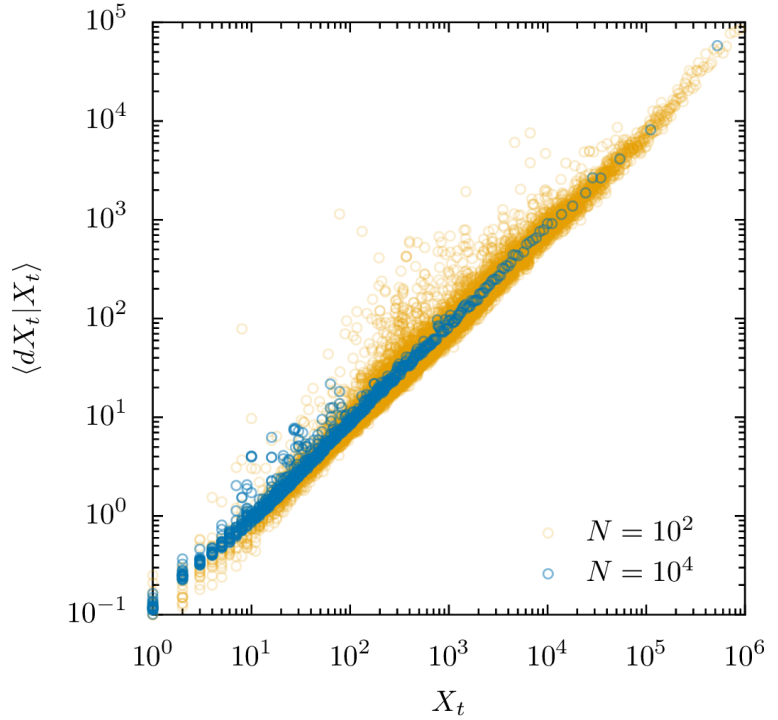


Figure 5.2: Mean of dX_t conditioned on X_t ; $t = 3$ days, $dt = 1$ day. Each measurement is made on bins of N items, as explained in Section 2.3. The mean follows roughly a linear shape, which is constant with respect to the sampling size. Note that, since the overall distribution of X_t is heavy-tailed, the points with low X_t are close, while for high X_t data is very sparse.

needed to use a process that models how the activity is accumulated. Two frameworks for such processes are presented: in Section 5.2 we discuss the model of allocation, and in Section 5.3 the Stochastic Differential Equations (SDE), which is a more general class of models, and in

particular we introduce a model with a fat-tail distribution of the increments, in order to model the fluctuations. In Section 5.4 we introduce the Maximum Likelihood estimation for SDE parameters, and in Section 5.5 we confirm that indeed the model with fat-tailed fluctuations is better at describing the data. In Section 5.6 we use the fitted models to issue predictions on the future items' activity, and we show that fat-tailed increment distributions are necessary in order to estimate correctly the amount of big hits, i.e. the extreme events of activity. The chapter is concluded with a discussion in Section 5.7.

5.2. Allocation model of Proportional Effect

In this section we discuss a model that results process of allocation where *tokens* (a view, a citation, a person) are assigned to items (a video, a paper, a city), with a probability proportional to their amount of tokens (this is the proportional element). This model is the core idea of the growth process described in Section 2.2.2, and we analyze it in detail since it relies on a microscopic picture that can be tuned in order to fit different phenomena. Formally, for a set of N_t items where $X_t^{(i)}$ is the amount of tokens of the i -th item at time t ,

$$\mathbb{P}(i) = \frac{X_t^{(i)}}{\sum_{i=1}^{N_t} X_t^{(i)}} . \quad (5.4)$$

Note that the amount of items N_t will increase with time, since it is a growth processes. The condition Eq. (5.1) is retrieved directly from Eq. (5.3) if it is assumed that the probabilities $\mathbb{P}(i)$ do not change when $X_t^{(i)}$ increases (like a multinomial process). Beyond this approximation, we derive the exact distribution $\mathbb{P}(dX_t | X_t)$ for this process in Section 5.2.1, from which the average and the fluctuations will be estimated.

5.2.1. Solution of the Allocation model

We want to estimate the distribution $\mathbb{P}(dX_t | X_t)$ of the model of allocation with proportional effect, in order to compare it with the data. Fortunately, this problem is part of the class of problems denominated *Polya's urn* (in our case the urns are the items, the balls are the tokens).

The setting of Polya's urn is exactly the same as ours: consider a box with colored balls belonging to M different colors, and there are exactly X_t balls of each color. (The balls represent the views and the colors the videos.) The process is defined by picking randomly a ball from the box, and returning k balls of the same color to the box. This action is repeated until N balls are added (or taken away). Note that $k = 0$ is sampling colors with replacement, $k = -1$ (removing the chosen ball) is sampling without replacement, and $k > 0$ is the usual setting of proportional effect, where the colors represent the items, and the balls represent the tokens. In fact, the probability of choosing a color (item) is proportional to the amount of balls (tokens), and at each step k balls

(tokens) are assigned to each color (item). We are interested in the final distribution of dX_t , the amount of new tokens of each item after N iterations of the process, which is equivalent to the distribution $\mathbb{P}\left(X_{t+1}^{(i)} \mid X_t^{(i)}\right)$.

In order to compute this probability, we will consider an operational time s , the time step of the process, the random variable $I_{i,s}$, which takes the value 1 if the item i is chosen at the time step s and 0 otherwise, and the value $n_{i,s}$, the amount of tokens that each item has at the i -item up to the operational time s . The initial value is $n_{i,0} = X_t^{(i)}$ and $dX_t^{(i)}$ will be equal to $n_{i,T} - n_{i,0}$.

Consider first the probability of an item i to be chosen. When the process starts ($s = 0$) all the items have the same probability,

$$\mathbb{P}(i, 0) = \mathbb{E}(I_{i,0}) = \frac{X_t^{(i)}}{\sum_i X_t^{(i)}} = \frac{1}{M}.$$

After the first round, we can compute this probability again:

$$\mathbb{E}(I_{i,1}) = \frac{n_{i,0} + k}{Mn_{i,0} + k} \mathbb{P}(I_{i,0}) + \frac{n_{i,0}}{Mn_{i,0} + k} (1 - \mathbb{P}(I_{i,0})) = \frac{1}{M} = \mathbb{E}(I_{i,0}).$$

This is a general property of Polya process, i.e. valid under any distribution of $n_{i,0}$, and implies that the probability of choosing an item at any time is constant, if considered without taking into account the history of the process (it can be understood as that the random variable $I_{i,s}$ possess the *martingale property*)[KPS12]. A corollary of this result is that the expectation of dX_t at any time is proportional to its initial value, up to a constant that represents the increment in the amount of tokens, as seen below.

We now show that the number of tokens of a given item in the end of the process is distributed with a Beta Binomial distribution, also called Polya distribution, which is a well known result in Probability Theory [EP23]. The proof is as follows. Consider $k = 1$ from now on; the amount of tokens added to the item i , up to the time step N is $n_{i,N}$,

$$n_{i,N} = \sum_{s \leq N} I_{i,s}.$$

We compute now the probability of the sequence $\{I_{i,s}\}_{s \leq N}$ to occur. We simply state it as a sequence of conditional probabilities:

$$\mathbb{P}(\{I_{i,s}\}_{s \leq N}) = \mathbb{P}(I_{i,1} = i_1) \mathbb{P}(I_{i,2} = i_2 \mid I_{i,1} = i_1) \cdots \mathbb{P}(I_{i,N} = i_N \mid I_{i,1} = i_1, \cdots, I_{i,N-1} = i_{N-1}).$$

Each conditional probability is known, and at first sight it depends on the whole history of chosen items. Actually, it only depends on the previous amount of the tokens assigned, because the

conditional probability is given by Eq. (5.4), which only depends on it. In fact,

$$\mathbb{P}(I_{i,t} = i_t \mid I_{i,1} = i_1, \dots, I_{i,t-1} = i_{t-1}) = \frac{n_{i,t-1}}{n_{i,0} + (t-1)} .$$

Notice that since the sequence $n_{i,t}$ increases by at most 1 at each time step; if the probability of the tokens having a particular value is computed, $n_{i,N} - n_{i,0} = x$, necessarily $n_{i,t}$ has to take values from $n_{i,0}$ to $n_{i,0} + t$, independently from the order. When N tokens are added, then x times the item i is chosen and $N - x$ is not. Not choosing the i -item can be considered as a second (merged) item, and the previous argument of the independence of the order can be used again. The values of the factors in the other item go from $(M-1)n_{i,0}$ to $(M-1)n_{i,0} - x + N$. Now we can see that if the last equation are replaced into the previous one, we will have in the numerator a product of all the the numbers between $n_{i,0}$ and $n_{i,0} + x$ and a product of all the numbers between $(M-1)n_{i,0}$ and $(M-1)n_{i,0} - x + N$. In the numerator, we have the product of all the numbers between M_0 and $Mn_{i,0} + N$; all together,

$$\mathbb{P}(\{I_t\}_{t \leq N}) = \frac{\prod_{j=n_{i,0}}^{n_{i,0}+x} j \prod_{j=(M-1)n_{i,0}}^{(M-1)n_{i,0}+N-x} j}{\prod_{j=Mn_{i,0}}^{Mn_{i,0}+N} j} = \frac{B(x + n_{i,0}, N - x - n_{i,0})}{B(n_{i,0}, N - n_{i,0})} \quad (5.5)$$

where B is the Beta function, defined as

$$B(x, y) = \frac{\Gamma(x)\Gamma(y)}{\Gamma(x+y)} = \frac{(x-1)!(y-1)!}{(x+y-1)!} \quad (5.6)$$

where the last form is valid only when $x, y \in \mathbb{N}$, by the equivalence of the Gamma function with the factorial in the natural numbers. Notice that the probability $\mathbb{P}(\{I_t\}_{t \leq N})$ is independent from the order of the variables I_t . This property allow us to write the probability of $n_{i,N}$ as the probability of having a sequence of I_t that sums x in N time steps, which is the probability of each individual sequence times the possible combinations,

$$\mathbb{P}(n_{i,N} - n_{i,0} = x) = \binom{N}{x} \frac{B(x + n_{i,0}, N - x - n_{i,0})}{B(n_{i,0}, N - n_{i,0})} . \quad (5.7)$$

This distribution is called Beta Binomial, referring to the functions present in the last equation. It can be readily computed the expectation and the variance of the increment $n_{i,N} - n_{i,0}$. The expectation is, as expected from the independence of $\mathbb{P}(i)$ with respect of time,

$$\mathbb{E}(n_{i,N} - n_{i,0}) = \mathbb{E}\left(dX_t^{(i)} \mid X_t^{(i)}\right) = \frac{N}{M} . \quad (5.8)$$

The variance is instead,

$$\mathbb{V}(n_{i,N} - n_{i,0}) = \mathbb{V}(dX_t^{(i)} | X_t^{(i)}) = \frac{(M-1)N(N + MX_t^{(i)})}{M^2(1 + MX_t^{(i)})} . \quad (5.9)$$

Interestingly, the number of tokens assigned to the items is not independent from $X_t^{(i)}$; in the context of proportional effect, the condition $\mathbb{E}(dX_t^{(i)} | X_t^{(i)}) = N/M = aX_t^{(i)}$ is set, which corresponds to equate the expectation of the distribution to Eq. (5.1). A new variance in this condition and its limit for $M \gg 1$ and $M \gg a$ (a is a number usually smaller than 1) can be computed:

$$\mathbb{V}(dX_t^{(i)} | X_t^{(i)}) = \frac{(M-1)a(1+a)(X_t^{(i)})^2}{(1 + MX_t^{(i)})} \rightarrow a(1+a)X_t^{(i)} . \quad (5.10)$$

This result is important, because it indicates that the standard deviation that we expect for $dX_t^{(i)}$ from a Polya-like process scales as the square root of $X_t^{(i)}$. This is the same scaling that we expect from a sum of variables with finite variance by Central Limit Theorem. In this case, the random variables being summed are the indicator functions on the items $I_{i,s}$, which are correlated, but their probability distribution is the same along the whole process.

The Beta Binomial distribution is discrete, but the limit to a continuous function, suitable for large values of h , can be easily taken by replacing the binomial factor by the equivalent Beta function:

$$\mathbb{P}(dX_t^{(i)} = x | X_t^{(i)}) = \frac{B(x + X_t^{(i)}, (M-1)X_t^{(i)} - x)}{(MX_t^{(i)} + 1)B(MX_t^{(i)} - x + 1, x + 1)B(X_t^{(i)}, (M-1)X_t^{(i)})} . \quad (5.11)$$

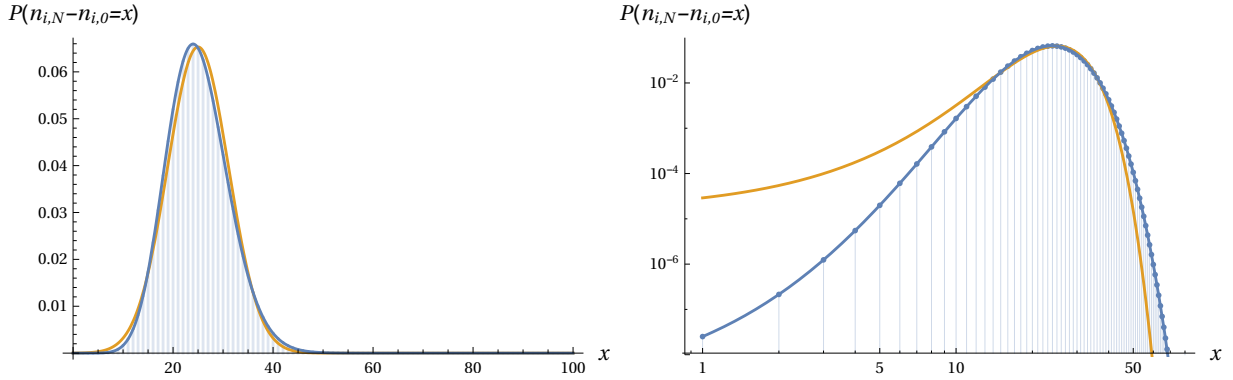


Figure 5.3.: Beta Binomial Distribution. Blue: probability mass function for parameters $X_t = 50$, $M = 200$ and $N = 5000$, and continuous probability density function with the same parameters. Orange: normal probability density function with same mean and variance as the Beta Binomial shown. Left panel: linear scale; Right panel: logarithmic scale.

The Fig. 5.3 shows an example of this function, the continuous version and a Normal distribution with the same mean and variance.

5.3. Stochastic Differential Equations framework

While simple generative processes where tokens are allocated, as preferential attachment, allow us to go from microscopic dynamics to macroscopic distributions, their simplicity will restrict considerably the range of possible observations that they can explain. In fact, the allocation process results in a particular type of fluctuations, which, as it will be seen, is unrealistic. Here we consider Stochastic Differential Equations (SDE) as a way to extend the type of dynamics that we can describe.

About SDEs

SDEs are the mathematical formalization of Langevin equations. A first order Langevin equation can be written as

$$\dot{x} = f(x, t) + g(x, t)\xi(t) \quad , \quad (5.12)$$

where x is the variable of interest, f, g are known function and ξ is a random variable such that

$$\langle \xi(t) \rangle = 0 \quad (5.13)$$

$$\langle \xi(t)\xi(t') \rangle = q\delta(t - t') \quad (5.14)$$

$$\cdot \quad (5.15)$$

This type of formulation is troublesome because it can generate *spurious drifts*. For instance, in the case where $f = 0$ and $g = ax$, the quantity $\frac{d}{dt}\langle x \rangle_{t=0} = a^2x$, is not zero as expected from $f = 0$. This apparent paradox arises from the integration of the stochastic term, a process that can be defined in many ways, or *interpretations* [KPS12]; the naive way of computing the derivative of the average position leads to this spurious drift, which corresponds to the *Stratonovich* interpretation. Here, we use always the so called *Ito* interpretation, where a given function F is integrated against a stochastic term ϕ according to

$$\int_0^\tau F(\phi(t'), t') d\phi(t') \equiv \lim_{\Delta \rightarrow 0} \sum_{i=0}^{N-1} F(\phi(\tau_i), \tau_i) (\phi(\tau_{i+1}) - \phi(\tau_i)) \quad (5.16)$$

where $\Delta = \max(t_{i+1} - t_i)$ and $0 = \tau_0 < \dots < \tau_N = \tau$. With this definition [Øks03], the derivative of the average is only given by the deterministic function f , meaning that the stochastic term of the equation is independent from previous realizations of itself.

The most common stochastic process to be integrated is the *Wiener process* W_t , a continuous process that has Normally distributed increments $W_{t+u} - W_u \sim N(0, \sqrt{u})$ which are independent. The Wiener process can be seen as the limit of a rescaled random walk, result of the Functional Central Limit Theorem [Don52], which explains its ubiquity.

When integrating a function f of the process X_t against Wiener noise, we have then equations of the type

$$Y_t = \int_0^t f(X_s, s) dW_s \quad . \quad (5.17)$$

The term dW_t is such that when integrated is equal to the Wiener noise difference, i.e.

$$\int_0^t dW_s = W_{t_1} - W_{t_0} \quad . \quad (5.18)$$

If X_t itself is the result of one of these integrals, we will have a (stochastic) differential equation, that can be written as well in differential form; for example, if $Y_t = X_t$, we would write Eq. (5.17) as

$$dX_t = f(X_t, t) dW_t \quad . \quad (5.19)$$

Proportional effect in SDEs

We will use first order SDEs to model the dynamics of videos' views; the most general form of (first order) SDEs with Wiener noise is

$$dX_t = \mu(t, X_t) dt + \sigma(t, X_t) dW_t \quad , \quad (5.20)$$

where $\mu(t, X_t)$ is the average growth, and $\sigma(t, X_t)$ scales the fluctuations. Since we want proportional effect to hold, we enforce Eq. (5.1) by using

$$\mu(X_t, t) = \mu_t X_t \quad . \quad (5.21)$$

We keep here a dependence of the proportionality constant with respect to time, allowing for effects such as attention decay [WH07b]. With respect to the fluctuations size dependence with respect to X_t , we use a dependence given by

$$\sigma(X_t, t) = \sigma_t X_t^\beta \quad , \quad (5.22)$$

which has been reported extensively in a variety of contexts, known as *Taylor's Law* (see Ref. [EBK08] for a review on this subject). This scaling of the fluctuations is usually dependent on time or on the size of the ensemble, while here, it is dynamically dependent on the variable itself. It is

worth to notice that the choice of σ as a power-law still gives a lot of flexibility. For example, $\beta = 1$ corresponds to the Geometric Brownian Motion, and $\beta = 0$ to the Ornstein-Uhlenbeck process [Ros14, Doo42]. Instead, $\beta = 1/2$ corresponds to a linear scaling in the variance of the fluctuations, a typical result of Central Limit Theorem; notice that the exact process described in Section 5.2.1 falls as well into this value of β . In summary, the SDE with proportional effect that we investigate here is

$$dX_t = \mu_t X_t dt + \sigma_t X_t^\beta dW_t \quad . \quad (5.23)$$

5.3.1. Solution of the SDE with proportional effect: Lognormal and CEV distributions

Here we show the result of integrating Eq. (5.23), also called *Constant Elasticity of Variance process* [CL10] (CEV), in a time unit $\Delta t = 1$ (e.g. 1 day). The complete derivation can be found in Ref. [CL10], and it involves changing variables a few times using Ito's lemma, yielding an SDE of the form

$$dX_t = a dt + b \sqrt{X_t} dW_t \quad , \quad (5.24)$$

where a and b depend on μ_t , σ_t and β . The result of this equation, which is known as the Cox-Ingersoll-Ross process [GJY03], is then transformed back to correspond to the original SDE. We evaluate the resulting distribution in a time $t + 1$ (exactly one time unit after t). When $\beta < 1$, this distribution has the form

$$\mathbb{P}(X_{t+1} = x | X_t = x_0) = 2(1 - \beta) k^{\frac{1}{2(1-\beta)}} \left(x z^{1-4\beta} \right)^{\frac{1}{4(1-\beta)}} e^{-x-z} I_{\left| \frac{1}{2(1-\beta)} \right|} (2\sqrt{xz}) \quad , \quad (5.25)$$

with

$$\begin{aligned} k &= \frac{\mu}{\sigma^2(1 - \beta)(e^{2\mu(1-\beta)} - 1)} \\ x &= k(x_0 e^\mu)^{2(1-\beta)} \\ z &= kx^{2(1-\beta)} \end{aligned}$$

where I is the modified Bessel function of the first kind. The expression simplifies using the substitution $p = 2(1 - \beta)$:

$$\mathbb{P}(X_{t+1} = x | X_t = x_0) = p k^{\frac{1}{p}} \left(x z^{2p-3} \right)^{\frac{1}{2p}} e^{-x-z} I_{\left| \frac{1}{p} \right|} (2\sqrt{xz}) \quad , \quad (5.26)$$

with

$$\begin{aligned} k &= \frac{2\mu}{\sigma^2 p(e^{\mu p} - 1)} \\ x &= kx_0^p e^{\mu p} \\ z &= kx^p . \end{aligned}$$

When $\beta > 1$, the probability function is just the same as above but multiplied by -1 . Note that the β parameter is the exponent of the power law tail that the distribution has asymptotically. To get dX_t here X_t is subtracted to obtain a distribution of dX_t . In the analysis of the YouTube data, the distributions are also truncated, discretized and normalized, in order to be compatible with the data.

When $\beta \rightarrow 2$, the solution to Eq. (5.23) is the Lognormal distribution, given by

$$\mathbb{P}(X_{t+1} = x | X_t = x_0) = \frac{1}{\sqrt{2\pi\sigma}X_{t+1}} \exp\left(-\frac{(\ln dX_t - \mu - \ln x_0)^2}{2\sigma^2}\right) \quad (5.27)$$

For numerical reasons these two models will be kept separated, because even if the fit of the CEV distribution can in principle result in a value of $\beta = 1$, that value of the fit has measure zero. We call the model that results in Lognormal distributions, Lognormal model (LN), and the one with $\beta \neq 1$ ($0 < \beta < 2$), the CEV model.

5.3.2. A SDE model with Lévy fluctuations

In this section we propose a model similar to Eq. (5.23) but with Stable (or Lévy-Stable) noise, instead of Wiener noise, which is a noise with a heavy-tailed distribution. (Lévy-Stable noise should not be confused by the noise of a Lévy process, given by the Lévy-Khintchine representation.) The motivation for introducing this model comes from empirical observations, and will become clear in the next section.

Stable random variables were introduced in Section 2.2.1, and are a result of summing variables without finite variance by means of a general Central Limit Theorem, which is why Stable distributions are preferred to any other fat-tailed distributions. The parameters of this distribution are: α , the index of the power-law tail; β , a parameter of asymmetry; μ , a location parameter equal to the average of the distribution; σ , a scale parameter; then, when a random variable X is Stable, we will denote it as $X \sim S(\alpha, \beta, \mu, \sigma)$.

The observation of fat-tailed noise in the fluctuations is due to Mandelbrot [Man63], who noticed it in financial assets, but only in the 1990s there was an intent to integrate it into a dynamical model [MS95]; nevertheless, it was ultimately shown that, even if at some time scale the return rate of financial assets is fat-tailed distributed, it is not a Lévy distribution [CPB97]. Formally, it is needed to replace the Wiener noise term dW_t by the term dL_t [JW93]. We can understand this

replacement as a generalization, since Normal distributions are the limiting case of Stable ones, for $\alpha = 2$. dL_t is such that, if integrated, leads to a Lévy-stable distribution, given by, in analogy with Eq. (5.18),

$$\int_{t_0}^{t_1} dL_t = L_{t_1} - L_{t_0} \quad , \quad (5.28)$$

where $L_{t_1} - L_{t_0} \sim S(\alpha, \beta, 0, (t_1 - t_0)^{1/\alpha})$, analogously to the distribution of the Wiener increment. SDEs with Lévy-Stable noise have been considered in . Having a noise term defined, we can construct a linear SDE,

$$dX_t = (\mu_t X_t + c_t)dt + (a_t X_t + b_t)dL_t \quad , \quad (5.29)$$

where the asymmetry parameter of dL_t is set to $\beta = 1$, based on empirical observations. Here we added two extra parameters, b_t and d_t ; this inclusion makes the model more general, although it will be seen that c_t is not very relevant. We call the model with the four parameters S4, with $c_t = 0$, S3, and with only μ_t and a_t , S2.

5.4. Estimation of the SDE parameters

We focus now on the estimation from data of the parameters of the SDEs equations

$$dX_t = \mu_t X_t dt + \sigma_t X_t^\beta dW_t \quad (5.23)$$

and Eq. (5.29), corresponding to SDE with Wiener noise and Lévy-stable noise respectively.

5.4.1. Averaging estimation

The parameters of the SDE can be estimated using the averaging method considered in Section 2.3 to investigate the proportional effect in the data. This method would first consider the set of points $(X_t^{(i)}, X_t^{(i)})$, then compute the average, $\langle X_t | X_t \in \text{bin} \rangle$ for a given set of bins, and finally fit the averaged points with the function $dX_t = \mu_t X_t$, through the method of least squares. In the context of time series analysis this method was considered extensively (see Ref. [FP97], and [FPST11] for a review of the main methods and applications).

The set of bins can be chosen arbitrarily; a trivial choice is to consider a bin for each X_t , but that leave us with a great fluctuation in the resulting points, since the distribution of the data with respect to X_t is very inhomogeneous. Here a binning was chosen such that the amount of items in each bin is the same, N ; avoiding the aforementioned problem. The averaged points are

given then by

$$(x, y) = \left(\frac{1}{N} \sum_{i=1}^N X_t^{(i)}, \frac{1}{N} \sum_{i=1}^N dX_t^{(i)} \right), \quad (5.30)$$

which are shown in Fig. 5.2 for two bin sizes, $N = 10^2, 10^4$, where the value of the average in the bigger window is converging, i.e. visually the noise over a straight line is decreasing. The straight line behavior over 7 decades confirm the linearity of the proportional effect.

In the same way that the expectation $\mathbb{E}(dX_t | X_t)$ is estimated, it is possible to estimate the fluctuation around it, as computed before, i.e. the standard deviation is computed in the defined window. From the relationship between the standard deviation and X_t , the form of the $\sigma(X_t, t)$ function can be estimated, in particular the parameter β if we assume X_t^β is the functional form. It is important to notice that in this type of estimation, if the standard deviation is an estimator for $\sigma(X_t, t)$, it means that we are implicitly assuming that $dX_t | X_t$ is normally distributed, with mean aX_t and variance $\sigma(X_t, t)$. This is a valid assumption only if we consider that an infinitesimal time passed from t to $t+1$, i.e. that the equation Eq. (5.23) is not integrated in time. If we do not assume this condition, it is needed to integrate equation Eq. (5.23), as explained in Section 5.3.1.

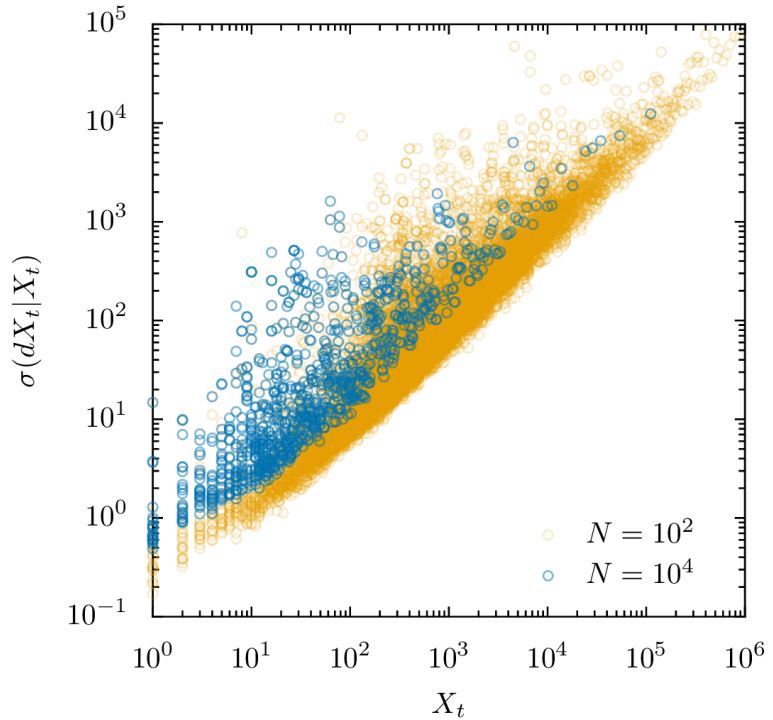


Figure 5.4: Standard deviation of dX_t conditioned on X_t at $t = 3$ days. Two different bin sizes, $N = 10^2, 10^4$ were chosen, and it is visible that aggregating the bins does not make the points less noisy.

In Fig. 5.4, the standard deviations of $dX_t | X_t$ for the bin sizes $N = 10^2, 10^4$ are shown; while visually a scaling given by $\beta = 1$ can be seen, the values are not converging, but are overall increasing with N . We do not expect this behavior for distributions that result from SDEs with Wiener noise, i.e. Eq. (5.23); a possible explanation is that the particular binning used is mixing

distributions with different X_t , that could lead to this observation. This argument is ruled out in the Appendix B, in favor of explaining this effect by the presence of fat tails in the fluctuations.

5.4.2. Estimation by Maximum Likelihood

A different way of estimating the parameters of the SDEs that considers the possibility of using other distributions to model the fluctuations, is by fitting each of the probability distributions $\mathbb{P}(dX_t | X_t)$. This is done by maximizing the likelihood of the data for each different bin (all the items in the bin are considered to have the same X_t),

$$\mathcal{L} = \mathbb{P}(\text{data} | \text{model}) = \prod_{i=1}^N \mathbb{P}(dX_t^{(i)} | X_t, \text{model}) \quad , \quad (5.31)$$

where here we refer to *model* as the choice of the distribution and the set of parameters that define it (i.e. Lognormal, CEV or Stable distribution). distributions [CSN09]. Most of the times, instead of the likelihood, minus the logarithm of the likelihood is minimized, because the probabilities tend to be quite small and it can cause troubles numerically,

$$\ell = -\ln \mathcal{L} = -\sum_{i=1}^N \log \mathbb{P}(dX_t^{(i)} | X_t, \text{model}) \quad . \quad (5.32)$$

This can be re-written as a sum over all the values that takes dX_t ,

$$\ell = -\sum_{dX_t} N(dX_t, X_t) \log \mathbb{P}(dX_t | X_t, \text{model}) \quad , \quad (5.33)$$

where $N(dX_t, X_t)$ is the amount of items in the data such that $dX_t^{(i)} = dX_t$ and $X_t^{(i)} = X_t$. This formulation is equivalent but is more convenient when applied numerically, since the data is already grouped in the observable $N(dX_t, X_t)$.

Once the parameters of the distributions for each bin are computed, it would be necessary to fit the resulting parameters with respect to X_t in order to obtain the parameters of the SDE model. This procedure is not rigorous and can be avoided by an improvement in the Maximum Likelihood Estimation.

However, we can look at a particular histogram of dX_t conditioned on X_t in Fig. 5.5; here the data corresponds to $t = 3$ and $X_t \in [513, 527]$ views, along with the best fits for the Lognormal, CEV and Stable distributions; Stable distributions, at first sight, provide a superior description of the tail of this distribution. In fact, the difference in the log-likelihood ℓ with respect to the CEV distribution is about 70, which means that the Stable distribution is e^{70} times more likely, i.e. it has a very high statistical support.

Before proceeding with the fit of the Eq. (5.29), we will analyze the quantiles of the distributions

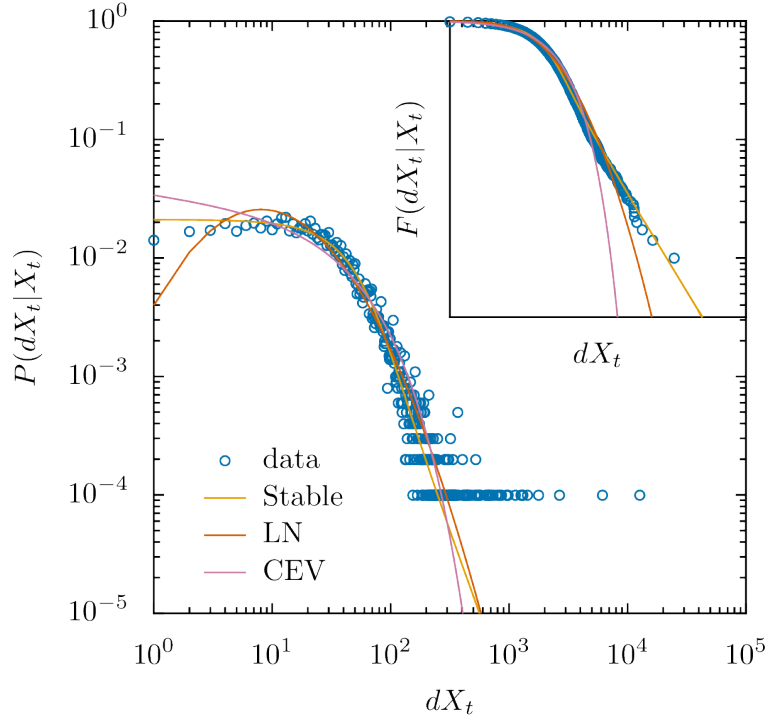


Figure 5.5: Data and fits of dX_t for a bin of size $N = 10^4$. S: Stable, LN: Lognormal, CEV: Constant Elasticity of Variance. In the main panel, the PDF; in the inset, the complementary CDF; $t = 3$ days, $dt = 1$ day. The size $N = 10^4$ for the bin $X_t \in [513, 527]$ is used because this size is suggested to be the minimal amount of data necessary to distinguish Lognormal from power law distributions [CSN09].

$\mathbb{P}(dX_t | X_t)$, in order to confirm that the functional form of Eq. (5.29) is consistent with the data. In Fig. 5.6, the quantiles 5%, 25%, 50%, 75% and 95% are shown. The quantiles are linearly

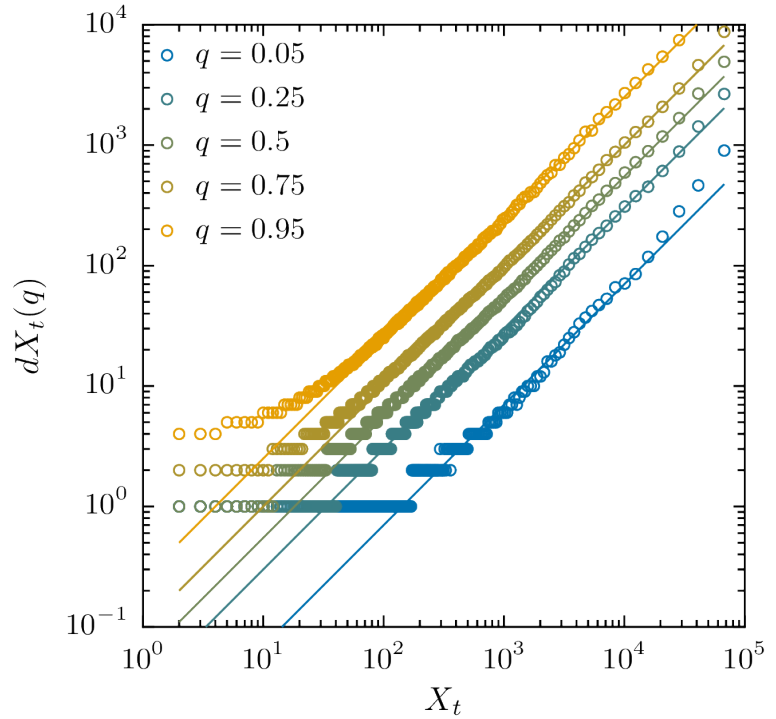


Figure 5.6: Quantiles of the distributions of dX_3 conditioned on X_3 . The quantiles are computed on groups of items of size $N \geq 10^4$. The lines are linear functions and are shown as references.

related with respect to the condition X_t , as indicated by the straight lines in Fig. 5.6, except for

the items with low amount of views. From the quantiles it is possible to estimate the Lévy-stable distribution parameters for each bin [FR71, McC86], however the aforementioned problems related with the binning persist.

The Maximum Likelihood Estimation can be improved, by fitting the parameters of model directly from data, without the need of binning. Here we fit actually X_{t+1} instead of dX_t ; this is necessary because the integration of the SDE results in distributions of X_{t+1} that cannot be readily translated in formulas for dX_t . From the parameters of the SDE, the parameters of the distributions for each X_t can be deduced.

This approach is pursued by Ref. [KFNP05], where it is proposed to iteratively minimize the Kullback-Leibler divergence between the empirical distributions and the distributions that result from solving the Fokker-Planck equation associated to the SDE for a finite time. In our case, these distributions are known (Lognormal and CEV), thus avoiding altogether the problems related with the estimation at finite time [RK01, FRSP02, HF11] and the numerical integration of the Fokker-Planck equation when the divergence is minimized. (The log-likelihood ℓ is equal to the Kullback-Leibler divergence between the empirical distributions and the distributions of the model.) If the SDE is not integrated in time, i.e. the fluctuations are Normal or Stable, then the distributions are just the distributions of the standard noise but translated by $\mu(X_t, t)$ and scaled by $\sigma(X_t, t)$. If instead the SDE is integrated in time, the fluctuations will be Lognormal or CEV, and the parameters of each distribution of X_{t+1} conditioned on X_t can be readily obtained from their definitions (see Equations 5.26 and 5.27).

Again a log-likelihood will be minimized, but we consider all the data points together; the total probability of the data with respect to the model is then

$$\ell = - \sum_{i=1}^{N_{tot}} \log \mathbb{P} \left(dX_t^{(i)} \mid X_t^{(i)} \right) \quad . \quad (5.34)$$

An equivalent form, more suitable to computation, is

$$\ell = - \sum_{X_t} \sum_{dX_t} N(dX_t, X_t) \log \mathbb{P}(dX_t \mid X_t) \quad . \quad (5.35)$$

In this way, all the X_t are considered, increasing to the maximum the size of the dataset to be fitted, thus eliminating the need of binning the data. Some numerical problems of this process are detailed in Appendix A.

We performed estimation of the parameters for the five models (LN, CEV, S3, S3 and S4), and for different times, $t \in (1, 30)$. As shown in the next section, the result of the model S3 ($b_t = 0$ in Eq. (5.29)) is the best. Agreement with respect to the data can be seen in Fig. 5.7, where a small set of the histograms $\mathbb{P}(dX_t \mid X_t)$ are being shown in the panels (a) and (c), as well as the collapse of the rescaled distributions in the panels (b) and (d). This collapse is due to the closure

of the Stable distribution under translation and scaling. (Both PDFs and CDFs are shown in the figure, since visual agreement can be present in one representation and not in the other.)

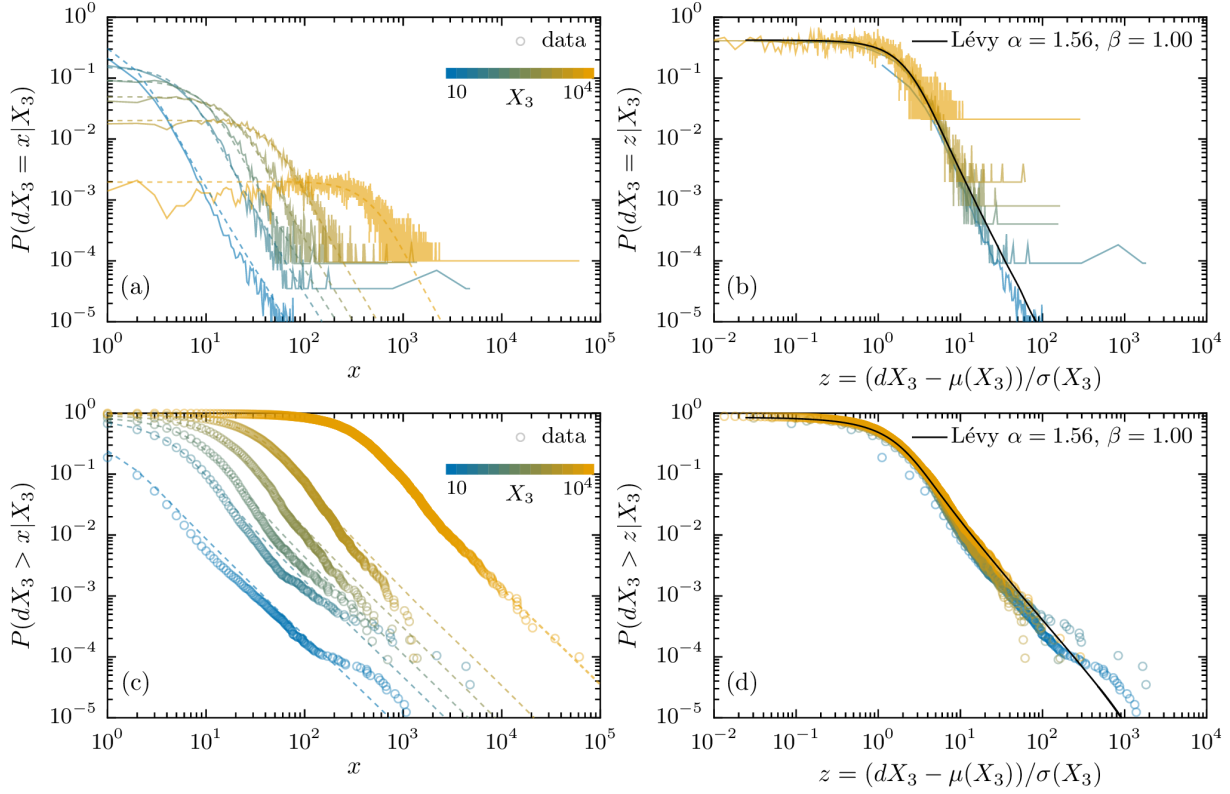


Figure 5.7.: Agreement between data and the S3 Lévy-Stable model. Five subsets of points are selected, where each subset is respectively the set of items with $X_{t=3} \in \{10, 55, 100, 196, 496\}$. (a) PDFs of $dX_{t=3}$ given X_t (solid lines), and Lévy PDFs with parameters that correspond to the same X_t (dashed lines), based on the global fit of the S model. (b) Rescaling of the PDFs according to the fitted parameters and Lévy CDF with $\mu = 0$, $\sigma = 1$. (c) CDFs of dX_t given X_t (points), and Lévy CDFs with parameters that correspond to the same X_t (dashed lines), based on the global fit of the S model. (d) Rescaling of the CDFs according to the fitted parameters and Lévy CDF with $\mu = 0$, $\sigma = 1$.

5.5. Model selection

In this section we compare quantitatively the agreement of the models proposed with respect to data. The models compared are summarized in Table 5.1. It has to be noted that the models that we compare have a different amount of parameters. We use the Bayes Information Criterion (BIC), a conservative way of penalizing additional parameters. The BIC is defined as

$$BIC = -2\ell + k \ln N \quad , \quad (5.36)$$

Model name	$\mathbb{P}(dX_t X_t)$ functional form	Parameters
LN	Lognormal	μ_t, σ_t
CEV	CEV	μ_t, σ_t, β_t
S2	Lévy-stable	α_t, μ_t, a_t
S3	Lévy-stable	$\alpha_t, \mu_t, a_t, b_t$
S4	Lévy-stable	$\alpha_t, \mu_t, a_t, b_t, c_t$

Table 5.1.: Summary of the models considered.

where \mathcal{L} is the likelihood of the data with respect to the model, k is the amount of parameters of the model, and N is the amount of data points; the BIC is the result of the Laplace approximation on the Bayes Factor [FHT01].

A similar approach for non-linear fit was implemented in Ref. [LMGA16], where non-linear scaling was studied in the context of scaling in cities, i.e. if a given observable measured on cities scales with respect to the population linearly or not. Integrating the fluctuations as part of a model naturally leads to consider the full likelihood function as a measure of quality, compensated by some factor that accounts for overfitting, like the $k \ln N$ in the BIC. Even if the function to estimate is relatively simple, the estimation process is particularly difficult to interpret: most of the items (or cities) have low activity (or low population), while few items concentrate most of the activity; if fluctuations decrease with size (as some models propose), then the few data points in the tail of the population distribution can be very important to determine the value of the fitted parameters. In our databases, the high volume of data allows us to have an educated guess on the shape of the fluctuations, given by the quantiles (see Fig. 5.6).

When we compare the models the BIC difference between any S model and the LN or CEV is, for different values of t , always bigger than 10^5 , indicating extremely strong support for our model with Stable fluctuations. In Fig. 5.8 the BIC difference among S3 and the other models is shown: S4 and S3 have a similar likelihood, while S2, LN and CEV are very unlikely in comparison. The addition of the parameter c_t is then not adding much with respect of S3, so we will select the model S3 among all. Altogether, these analysis support our proposal of stochastic differential equation with Lévy noise, Eq. (5.23), to describe the dynamics of popularity in YouTube.

We can, moreover, analyze the particular values of the parameters of the S3 model, shown in Fig. 5.9. The parameters show a strong dependence in t in the first week. In particular, μ_t decays in this period (reflecting a decay in the gain of views) and $\alpha_t \approx 1.75$ for $t > 5$. This decay is also observed in other media [WH07b, LBK09] where also there is a constant inflow of new items.

Figure 5.8: Quality of the models. The difference in Bayes Information Criterion (ΔBIC) is shown with respect to the model S3. Values as high as 10^5 indicate a very high support for the S3 model.

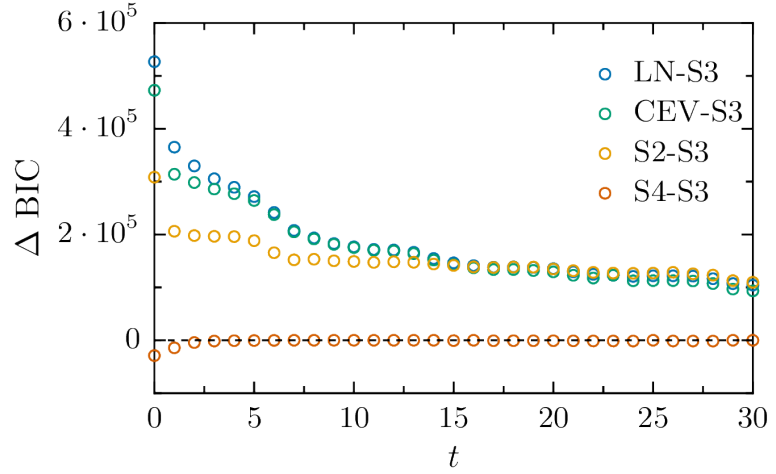
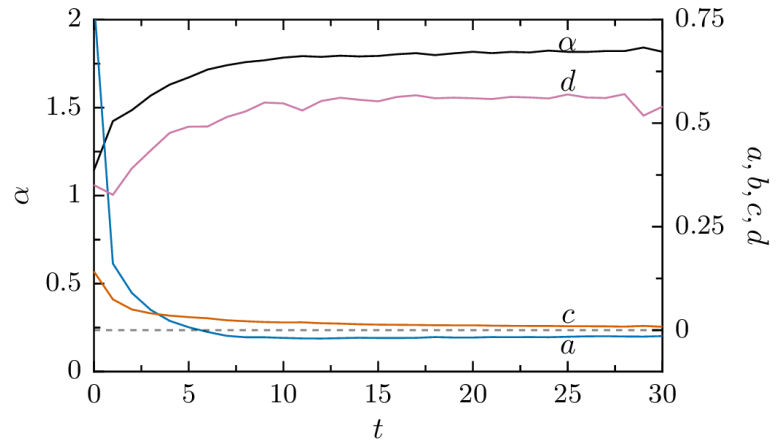


Figure 5.9: Evolution of the parameters of the model, $0 \leq t \leq 20$. After the first week, there is a tendency to stabilization.



5.6. Forecasts of popularity

We now focus on the possibility of forecasting popularity with the models proposed, and in particular with the S3 model (model S from now on).

Having a model for the probability $P(dX_t | X_t)$ enables us to predict the evolution of the videos' views, using the SDE equations. In order to do that, it is needed to know how many views an item had at a given time t , X_t , which is the initial condition; the probability of having x views in the next day is the probability $\mathbb{P}(dX_t = x - X_t | X_t)$, given by the SDE equations with the appropriate parameters.

To compute the probability of having x views at a time $t_f = t + \tau$ the probabilities of all the possible paths have to be integrated. Specifically, for $\tau > 1$ we will have the Chapman-Kolmogorov equation,

$$\mathbb{P}(X_{t+\tau} = x | X_t = x_0) = \sum_{x_1=0}^{\infty} \mathbb{P}(X_{t+\tau} = x | X_{t+\tau-1} = x_1) \mathbb{P}(X_{t+\tau-1} = x_1 | X_t = x_0). \quad (5.37)$$

This recursive equation allow us to numerically compute the probability distribution of views at any time, but requires, of course, the parameters for all the times between t and $t + \tau$, which were previously computed and shown in Fig. 5.9. This procedure is used in the next section to analyze the possibility of a video becoming a big hit.

5.6.1. Optimal deterministic forecast

If we want to issue a deterministic forecast of the views, instead of a probability, it is needed to define a cost with respect to which we issue an optimal predictor $\hat{X}_{t+\tau}^{(i)}$. The cost function usually used is the mean squared error (see Section 2.4, which has as optimal predictor the mean value of $\hat{X}_{t+\tau}^{(i)}$). The mean value can be estimated by computing the expectation value of $X_{t+\tau}$, an easy calculation, since the average increment of views scales linearly with the previous amount of views, Eq. (5.1). In practical terms, the optimal prediction for $X_{t+\tau}$ is given by $\mathbb{E}(X_{t+\tau} | X_t = x_0)$. Using Eq. (5.37), we evaluate the expression

$$\mathbb{E}(X_{t+\tau} | X_t = x_0) = \sum_{x_1=0}^{\infty} \mathbb{E}(X_{t+\tau} | X_{t+\tau-1} = x_1) \mathbb{P}(X_{t+\tau-1} = x_1 | X_t = x_0)$$

and now we use the fact that $\mathbb{E}(X_{t+\tau} | X_{t+\tau-1} = x_1) = a_{t+\tau-1}x_1$,

$$\mathbb{E}(X_{t+\tau} | X_t = x_0) = a_{t+\tau-1} \sum_{x_1=0}^{\infty} x_1 \mathbb{P}(X_{t+\tau-1} = x_1 | X_t = x_0) .$$

Here, it is assumed that $\langle dX_t \rangle = a_t X_t$, which is a valid assumption only for large X_t (because the truncation and the discretization of the Lévy distribution would shift the average from the 0; this effect is stronger for small values of X_t).

The sum left is just the expectation value of $X_{t+\tau-1}$,

$$\mathbb{E}(X_{t+\tau} | X_t = x_0) = a_{t+\tau-1} \mathbb{E}(X_{t+\tau-1} | X_t = x_0) .$$

In this way, it is possible to go back τ iterations, and as result the optimal predictor for a video such that $X_t = x_0$ is

$$\hat{X}_{t+\tau}^{(i)} = \mathbb{E}(X_{t+\tau} | X_t = x_0) = \prod_{i=0}^{\tau-1} a_{t+i} x_0 . \quad (5.38)$$

It has to be noted that this predictor is optimal, but its properties have to be considered carefully. If the mean is estimated directly from the data available (e.g. by taking the videos with same X_t and computing the average of $X_{t+\tau}$) it may be the case that this value is not close to the expected value previously computed. This happens because the average of Stable-

distributed random variables is Stable-distributed as well, with the same exponent α and with the scale decaying with $N^{1/\alpha-1}$, as result of the Central Limit Theorem mentioned in Section 5.3.2. In conclusion, means estimated directly from data may be unreliable estimators of the expected value, specially when the set of items considered is small (large X_t); for this reason we consider in the next section quantile predictions.

5.6.2. Predictability of big hits

We now focus on the estimation of the probability of an item becoming a big hit after a given time (larger than one time unit). We define as a big hit at time t the top $q\%$ videos with highest X_t ($X_t > x_t^q$), where we note with x_t^q the value of X_t for which only a fraction q of items is more popular. This is a definition of a big hit as an extreme event; notice however, that the threshold of activity is not given a priori as in Chapter 4, but it instead evolves with time, since naturally the total activity of any item increases.

We are particularly interested in estimating the probability $P(X_t > x_t^q | X_{t_0} = x_0)$ of videos that are not big hits at time $t_0 < t$ (i.e., $x_0 < x_{t_0}^q$) becoming big hits at time t . This probability quantifies how unpredictable the system is. For instance, in a deterministic (proportional growth) model, the rank of the videos does not change and therefore such probability is zero (perfect predictability).

As an example, the videos that had 100 views one day after publication are selected, $X_1 = 100$, which belong to a rank of $q \approx 15\%$. We are interested in the probability of these videos having $X_t \gg 100$ at $t > 1$. To obtain the expectations of the models, we computed $P(X_t | X_1 = 100)$ iteratively from $P(dX_s | X_s)$ for $s = 1, \dots, t$, using $X_1 = 100$ and the method described in the previous section. We also compute the same distribution for the LN and CEV models, in order to show the difference that these models imply in the long run. An example of this simulation is given in Fig. 5.10 where we show the expected distribution of views of the selected videos after 5 days ($t = 6$). We can see that there is a good agreement between our model and the real distribution, while the LN and CEV models fail to describe the empirical distribution after only around 300 views. Moreover, the Lévy model predicts a substantially higher probability for large X_t than the alternative models, as expected. The agreement of our model with the data also points out that the assumption of independent uncorrelated increments is valid, at least in the short run.

In order to investigate the temporal dependence, we focus on the probability of the videos improving their rank and being by day t in the top $q = 5\%$, using the previously computed probabilities from the models and the thresholds x_t^q estimated from data. The results are summarized in Fig. 5.11 and show that the Lévy-stable model succeeds in estimating this probability in the short-term, while for the long-term the data shows an even higher probability (mixing of ranks). The other models assign a video a substantially lower possibility of becoming a big hit, an effect of their highly predictable dynamics. The fact that our model provides a good account for short-time

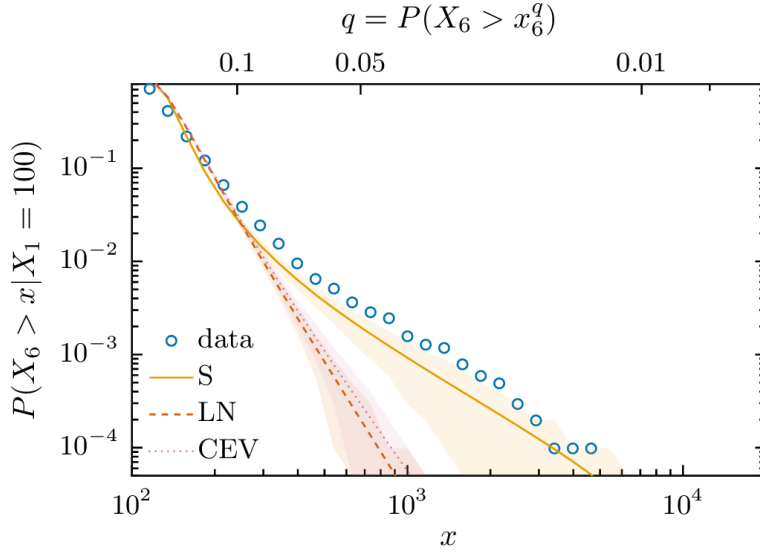


Figure 5.10: Expected distribution of views after 5 days for videos such that $X_1 = 100$.

intervals but not in the long run suggests the existence of correlations in the attribution of views that span multiple days and that are not accounted by our assumption of an independent noise.

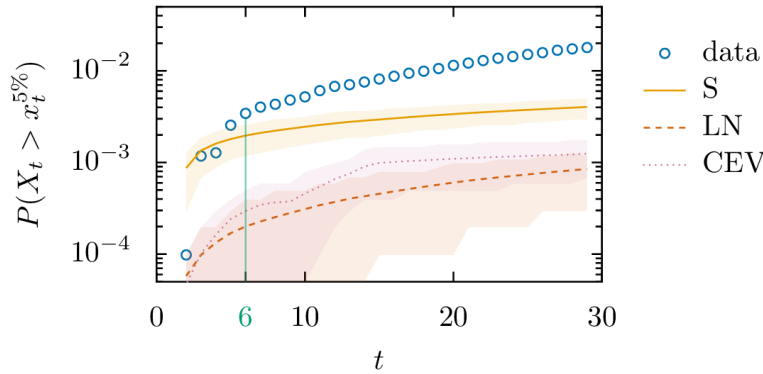


Figure 5.11: Expected amount of videos in the top 5% among the ones such that $X_1 = 100$.

5.6.3. Correlations in the S model

In this section we analyze the time correlations of the S model. We assumed until now that the noise in the SDEs is distributed identically and independently. If the first condition of *identity* is fulfilled, we say that the model agrees with the data in a descriptive way, as seen. The second condition, *independence*, is much more restrictive; measuring the correlations and testing if they are consistent with the model is a way to assess this condition.

Each item has its own time-series of stochastic elements $dL_t^{(i)}$; they can be estimated by extracting them from a stochastic differential equation whose drift and diffusion term depend on the

history $H_t^{(i)}$ through functions μ and σ (fitted for all items):

$$dL_t^{(i)} = \frac{dX_t^{(i)} - \mu(X_t^{(i)}, t)}{\sigma(X_t^{(i)}, t)} \quad (5.39)$$

The increments are Lévy distributed, as described previously, and now we want to get closer as we can to test for independence, which in the most cases is to test for correlations. The usual indicator used in time-series analysis is to estimate the autocorrelation function $\rho(t, \tau)$, defined as

$$\rho(t, \tau) = \frac{\mathbb{E} \left(dL_t^{(i)} dL_{t+\tau}^{(i)} \right)}{\mathbb{V} \left(dL_t^{(i)} \right)} \quad (5.40)$$

In this definition we used that $\mathbb{E} \left(dL_t^{(i)} \right) = 0$. The expectation value is taken on the sample space of the time series realizations; usually (e.g. time-series of temperature anomalies, or climate records applications) it is assumed that the time-series is stationary and ergodic, such that this expectation value can be estimated by an average in time. In our context this assumption cannot be made, so we will use as estimator an average in the ensemble of items, which we consider indistinguishable:

$$\hat{\rho}(t, \tau) = \frac{\sum_{i=1}^n dL_t^{(i)} dL_{t+\tau}^{(i)}}{\sum_{i=1}^n (dL_t^{(i)})^2} \quad (5.41)$$

However, there is a problem in using the autocorrelation function for time-series with heavy-tailed increments, namely that we are trying to estimate the variance (in the denominator) of a variable that has no defined variance. If an extreme event occurs at time t , that value will bias the estimation of all the correlation measures that involve that particular day; in other words, it is needed an estimation of correlation that is robust under the presence of fat tails.

A solution to this problem is given by the Spearman rank correlation [Ken48, Spe04]. As the name indicates, it consists in a correlation between two sets of ranks; this ranks are generated by ordering the values of $dL_t^{(i)}$ and $dL_{t+\tau}^{(i)}$, where average rank values are given to tied items, and the definition of correlation is as before, but instead of using the values dL , their ranks are used, $R_t^{(i)}$ (the position of the i -th item in an ordered list of items according to X_t):

$$\rho_S(t, \tau) = \frac{\mathbb{E} \left(R_t^{(i)} R_{t+\tau}^{(i)} \right)}{\mathbb{V} \left(R_t^{(i)} \right)} \quad (5.42)$$

Since we are evaluating the correlation among sets of ranks, the value of the particular dL is not important: on the one hand, the Spearman rank correlation allow us to measure correlation

when outliers are present, and in particular when values are heavy-tailed distributed; on the other hand, it is also more general than the usual correlation, because there is no dependence on the functional form of the relationship between the two sets. The correlation given by Eq. (5.40) measures the dispersion from a linear relationship between two sets of points, while the Spearman rank correlation measures the dispersion from a general monotonic relationship, that can be non linear. The estimator of the Spearman rank correlation is given by the formula [Spe04]

$$\hat{\rho}_S(t, \tau) = 1 - \frac{6 \sum_{i=1}^N \left(R_t^{(i)} - R_{t+\tau}^{(i)} \right)^2}{N(N^2 - 1)} . \quad (5.43)$$

If we assume the model as valid, the correlation between the sets of dL terms should be zero; we will use this assumption as a *null hypothesis*. In the case of the Spearman rank correlation is quite easy to compute the p -value of this hypothesis. It is known in fact [Ken48] that the statistic

$$t = \rho \sqrt{\frac{N-2}{1-\rho^2}} \quad (5.44)$$

is distributed as a Student t-distribution with $N-2$ degrees of freedom. In summary, the Spearman rank correlation provides a complete framework to study the presence of correlations between sets of heavy-tailed distributed values.

5.6.4. Dependence of correlation with respect to lag and time

In Fig. 5.12 show the dependence of the correlation with respect to the chosen lag τ is shown. In the S model the correlation between a fixed day t and the day $t + \tau$ is decreasing when time advances, but only when $t = 0$ it reaches zero, while for $t > 0$, the correlation never decays completely. This means that knowing if an item receives an increment in activity above the average at a particular time implies that is more likely that this item will receive an increment in activity above the average in the future, except for the information of the very first day, which decays in the first two weeks to zero. It has to be noted that activity at time $t = 0$ may not correspond to a full day (24 h), so this has to be considered as a special case.

The lag can be fixed, to observe the dependence with respect to the initial day of the comparison t . The results are shown in Fig. 5.13: the noise in the S model is increasingly correlated as time advances. In essence, the forecast attempted using the SDE model with Stable-distributed fluctuations is affected by long-term correlations, that are statistically significant. This type of analysis of correlations is usually overlooked in the context of proportional effect models, but we can see that is important, since the random process that the model has is strongly constrained. A new model that we propose in Chapter 6 will solve this problem.

Figure 5.12: Correlation of the estimated Lévy noise with respect to the lag τ . Different times of the first observation were used, $t = 1, 2, 5$. Where the correlation is not significantly different than zero ($p\text{-value} > 0.05$), the symbols are filled, while where the null hypothesis is rejected are not filled. The correlation in the S model decays to zero only when $t = 0$; this is the correlation between the activity of the first day and the activity of the day τ .

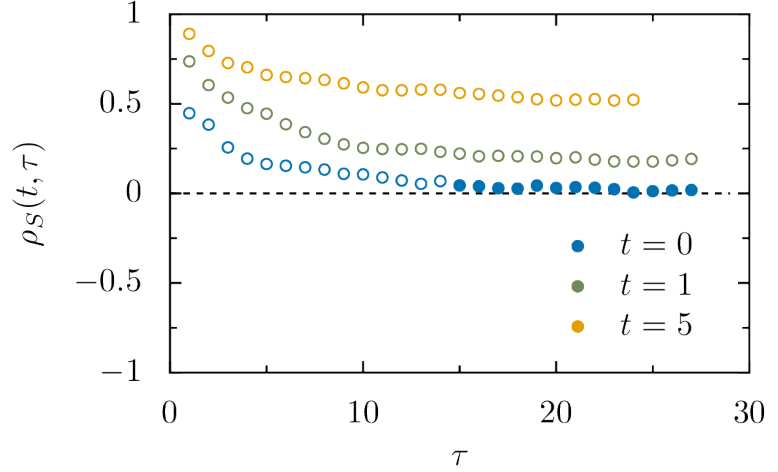
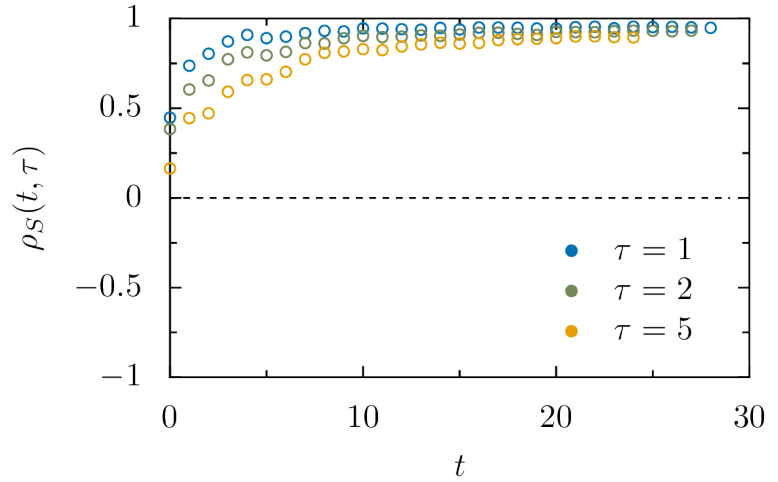


Figure 5.13: Correlation of the estimated Lévy noise with respect to time of the first observation t . Different lags were used, $\tau = 1, 2, 5$. Where the correlation is not significantly different than zero ($p\text{-value} > 0.05$), the symbols are filled, while where the null hypothesis is rejected are not filled.



5.7. Discussion and Conclusions

In this chapter, we established the presence of both proportional effect and Lévy fluctuations in the stochastic dynamics of the YouTube videos' views. Moreover, we showed that a model with both elements is better than previous proposals in terms of descriptive power and represents better the predictability of the dynamics.

These findings have important consequences for the mathematical modeling of complex systems. First, it shows that, even if proportional growth is present, it cannot be attributed as the responsible for the origin of the heavy tails because this is a feature already present in the fluctuations. Second, the use of Gaussian-based stochastic equations, such as Eq. (5.23) or traditional Fokker-Planck equations, overestimate the predictability of videos, by neglecting the mobility of popularity. We showed that better results are obtained in YouTube using a stochastic equation with Lévy noise, Eq. (5.29), an approach that has been previously proposed in Physics [JW93]. With respect to applications to empirical data, while the observation of fat-tailed distributed dis-

placements is a well-studied field of research (see Refs. [SZF95] and [ZDK15] for an overview on *Lévy flights* and *Lévy walks* respectively), the modeling of time series with this type of SDE is rare (see Ref. [Dit99] for an example in time series analysis of climate change). This formalism, and possibly also kinetic equations of the fractional type [MBK99, BS02], can be successfully applied to describe the dynamics of social media as well.

New insights are brought on the attention economy of the Internet as well. The fact that the multiple factors affecting the popularity of videos can be effectively modeled by a Lévy-stable distribution shows that the decision of different individuals are correlated to each other and lead itself to strong fluctuations. The Lévy-stable distribution is invariant under convolution, i.e. if X_1, X_2 are stable, also $X_1 + X_2$ is stable, and therefore it is a natural attractor for the combination of multiple (fat-tailed) processes, as the ones that creates the bursty activity patterns that characterize online social media as well. The presented analysis of fluctuations are enabled by the large availability of data in YouTube videos and we expect similar results to hold also in more general systems in which items compete for the attention of users.

6. Generalized growth models

6.1. Introduction

A forecast method to predict the activity, X_t , can use a variety of input variables. The analysis of Chapter 4 tells us that the variables that do not depend with time, M , are not very informative in comparison with the previous activity of the item. In the last chapter, instead, models based on a simple dependence with respect to the past were discussed, i.e. the future activity X_t comes from a distribution that depends only on the total activity one time unit before, X_{t-1} . However, more values of the past activity can be used; in fact, even the full history of the item, H_t can be used as input in a forecasting method. Moreover, the correlation found in the noise in Section 5.6.3 suggests that this generalization is needed.

In Section 6.2 we present two very general time series forecasting methods, the Autoregressive Model and the k -Nearest Neighbors Algorithm; the results of these methods indicate that the best forecast of X_t is achievable knowing only X_{t-1} and X_{t-2} , and in particular, that the forecast depends directly on the quantity $dX_{t-1} = X_{t-1} - X_{t-2}$. This result is be used to construct a new SDE-based model in Section 6.3, the D model. The estimation and the agreement of the D model is analyzed in Section 6.3.2, its usage for long-term forecast in Section 6.4 and the time correlations of its fluctuations in Section 6.5; in all this three points the D model proves to be superior to the best model of Chapter 5, which here is called the S model. The results have as implication that proportional effect can be measured and even modeled if an underlying, different, model is more likely, as discussed in Section 6.6.

6.2. Forecasting popularity with longer histories

Until now we focused on models where the the future activity X_{t+1} is dependent only on the previous activity X_t . This dependence is of the form

$$\mathbb{P}(dX_t \mid o) = f(dX_t, o) \quad , \quad (6.1)$$

where o is the observable used (X_t for S) and f is a probability mass function of dX_t (or probability density function if X is continuous) where o is taken as a parameter. The abundance of time series of the datasets considered allows the study of each individual distribution $\mathbb{P}(dX_t|o = X_t)$.

In the YouTube data, for example, the typical number of different values of X_t , K_X is around $K_X \sim 55000$, where the total number of items is $N_{tot} \sim 10^7$, leaving an average number of items per X_t of 200. Of these groups, most of them have only one item, and some of them will have many more items, since the distribution of X_t is heavy-tailed, which is the original motivation for merging the data in groups of similar X_t , described in Section 5.4.1, with the problems associated to it. By the modifications on the likelihood calculation introduced in Section 5.4.2, each step of the maximization requires to compute around 7000 Lévy-stable probability mass functions (instead of K_X).

If a second observable is included, say $q = X_{t-1}$, since the probability of having the same amount of activity on two consecutive days is virtually null, a very large number of distinct pairs (o, q) will exist ($K_{(o,q)} \sim 4.5 \cdot 10^5$ – roughly a tenfold increase), leaving the average number of items per key at 20. If an analogous method to the one described in Section 5.4.2 is used in this context, it would be needed to compute roughly 25000 Lévy-stable probability mass function in each maximization step, which is a much larger computational requirement.

In summary, if the goal is to model the fluctuations on a given model, increasing the amount of input information would make highly impractical the estimation procedure. Thus, in this section a different approach is pursued, where the fluctuations around the expected value are not modeled, but instead are indicators of the quality of the model, as in many time-series studies. We propose two simple models, which capture different features of the items' activity history, and we compare their predictive power, the criterion that ultimately selects the best model. The goal is to evaluate qualitatively which type of features o are more significant and if there is general information that these models can point out, such as if using larger histories (i.e. more information) results in better predictions.

The two forecasts, namely the Autoregressive and the k -nearest neighbors, are tailored to the particular case of an ensemble of time-series of items' activity. In the following paragraphs we motivate each of them, we describe how they are implemented, and then we summarize and discuss the results.

6.2.1. Autoregressive models

Autoregressive models are one of the classes of models most used in linear time-series analysis [BJRL15], and applied as well for YouTube data [PAG13]. The models of the class $AR(\tau)$ are the ones where the current value of an observable X is a linear function of its τ last values plus a noise

$$X_t = \sum_{s=1}^{\tau} a_s X_{t-s} + \sigma \xi_t \quad . \quad (6.2)$$

This definition can be seen as a discretization of an SDE, essentially equal to the S model for $\tau = 1$, i.e. X_t depends only on X_{t-1} , but with a different noise term. The usual context where this model is used is climate or finance research, where a particular observable (temperature anomaly in a geographic location or return rate of an equity) is registered for a long time, and this observable is assumed to be stationary. In the context of time series of items' activity, there are many, short, time dependent time-series. A consequence of this is that the parameters a_s will depend with time, changing the model to

$$X_t = \sum_{s=1}^{\tau} a_{s,t} X_{t-s} + \sigma \xi_t \quad . \quad (6.3)$$

It means that now there are τ parameters $a_{s,t}$ for each time step t modeled, which implies that the estimation of the parameters cannot be done in the usual way (i.e., through the Yule-Walker equations [Yul27, Wal31]), but it has to be considered as a common linear regression. The second consequence is that the statistical ensemble is not given by observations distributed in time, but at the same time over all the time series. In practice, for each time step t , the parameters a_{s,t_s} will be estimated as the coefficients of a linear regression between the set of $\tau \cdot N$ previous observations, namely the matrix $\mathbf{X}_{(t,\tau)}$ and the N observations at time t , the vector \mathbf{X}_t . The least squares estimator of the parameters (equivalent to the Maximum likelihood estimator if the fluctuations ξ_t are assumed i.i.d. normally distributed) is

$$\hat{\mathbf{a}}_t = \left(\mathbf{X}_{(t,\tau)}^T \mathbf{X}_{(t,\tau)} \right)^{-1} \mathbf{X}_{(t,\tau)}^T \mathbf{X}_t \quad . \quad (6.4)$$

This is the estimator that minimizes the squared distance between the prediction $\hat{\mathbf{X}}_t$ and the observations \hat{X}_t , where the predictor for a particular time series i is

$$\hat{X}_t^{(i)} = \sum_{s=1}^{\tau} \hat{a}_{s,t} X_{t-s}^{(i)} \quad . \quad (6.5)$$

The parameters represent then the average behavior of the item's activity time series. In the context of real-time forecasts, there are then two possible situations in which the predictor $\hat{X}_t^{(i)}$ is built, depending if the values $X_{t-s}^{(i)}$ used are the real ones (only possible if the forecast of the next day is desired) or result of a forecast themselves, which is necessary if a forecast more than one day ahead is desired, a very likely situation.

A fundamental assumption for this model is that the fluctuations $\sigma \xi_t$ must be i.i.d. normally distributed and independent from $X_s \forall s < t$. (A variation of this model would be to have heteroskedastic, and independent fluctuations, but it requires to have an *ansatz* for the size of the fluctuations.) However, if the forecast accuracy of the S model is measured in terms of the square distance between its predictor and the observation, the error made would be very large, and in-

creasing with the sample size, because of the presence of heavy-tails, as discussed in Chapter 5. This empirical observation is not directly modeled by the AR model. If the addition of more information, i.e. increasing τ , removes these properties, a model without heavy tails would have been reached, where the heavy tails observed previously would be explained as the result of a mixture of initial conditions and parameters of the items.

6.2.2. k -Nearest Neighbors algorithm

The $\text{AR}(\tau)$ process is the simplest way of introducing the history of a video $H_t^{(i)}$ into a model. Its simplicity is as well its limitation, since such a linear model cannot possibly reproduce many of the observed features of social activities. We introduce here a second class of methods which can address this issue by being solely based on the available data, without entering into the estimation of non-linear models, or even parameter estimation itself. The so called k -Nearest Neighbors (k -NN) algorithm [FHT01] relies on the intuitive idea that similar items should behave in similar fashion. (An essentially equal technique is known in the time-series domain as *Lorenz's method of analogues* [Lor69]; an important difference to the method of analogues is that for k -NN the amount of neighbors is fixed, no matter how far away they are from the item.) It is a very common and simple algorithm used in machine learning, and is completely data-based, in the sense that its prediction relies almost solely in data and not in any parametric model. As such, its main limitation is its lack of a clear physical interpretation; it has as well some other practical limitation, which will be discussed below.

For a given that a notion of distance between items in the features space $d(i, j)$, the method consists in choosing the k items that are closer to the i -th item, and predict for the i -th item the average of them,

$$\hat{X}_t^{(i)} = \frac{1}{k} \sum_{j=1}^k X_t^{(j)} = \langle X_t \rangle_k, \quad (6.6)$$

where the notation $\langle \cdot \rangle_k$ indicates the average over the k neighbors of the i -th time series. Since the space of features considered is the history of the videos H_t , a natural choice for distance is to think the features as an L^p space, so that the distance between two videos will be

$$d(i, j) = \left(\sum_{s=1}^{\tau} |X_{t-s}^{(i)} - X_{t-s}^{(j)}|^p \right)^{1/p}, \quad (6.7)$$

where $p = 2$ is the most common choice. A more general version of this algorithm introduces weights into these definitions. The motivation for doing this is to introduce some natural assumptions, like to give more importance to some videos more than others (e.g. closer to i), or to some days more than others (e.g. closer to t). In this sense, the above definitions will change accordingly

to

$$\hat{X}_t^{(i)} = \frac{1}{k} \sum_{j=1}^k w_{i,j} X_t^{(j)} \quad (6.8)$$

and

$$d(i, j) = \left(\sum_{s=1}^{\tau} \omega_{t,s} \left| X_{t-s}^{(i)} - X_{t-s}^{(j)} \right|^p \right)^{1/p} . \quad (6.9)$$

For now, these options will be ignored because a particular choice of the weights cannot be justified a priori.

While *k*-NN is a reasonable algorithm, its limitations surface when we are confronted with finite size datasets, since the algorithm assures convergence only asymptotically with *N*. In order to clarify this point, let's consider a dynamics only dependent on the last state ($\tau = 1$):

$$X_t = f(X_{t-1}) + \sigma \xi_t \quad (6.10)$$

For a given distance $d(i, j)$, the predictor of *i*-th time series is

$$\hat{X}_t^{(i)} = \langle X_t \rangle_k = \langle f(X_{t-1}) \rangle_k \quad (6.11)$$

The predictor is optimal only if its expectation matches the one of the model, i.e.,

$$\mathbb{E} X_t^{(i)} = \mathbb{E} \hat{X}_t^{(i)} = \langle f(X_{t-1}^{(j)}) \rangle_k \quad (6.12)$$

This is a necessary condition for the algorithm in order to give meaningful results. Although it seems trivial, it is not, since the average over the *k* neighbors is actually dependent on the sampling of X_{t-1} . If X_{t-1} is distributed according to $\rho(x)$, in a small interval Δ there will be approximately $N\rho(x)\Delta$ items. In order to have *k* neighbors in a space Δ the following condition must stand

$$\frac{k}{N} \sim \rho(x)\Delta \quad (6.13)$$

If $\rho(x)$ is small, then Δ must be big, decreasing the chances for Eq. (6.12) to hold. In the case of $\rho(x)$ being a heavy-tailed distribution, which is the case of interest, a large range of *x* where the data is sparse will always exist. More formally, Δ is given by

$$\frac{k}{N} = \int_{X_{t-1}-\Delta/2}^{X_{t-1}+\Delta/2} \rho(x) dx = \int \mathbb{I}_x(\Delta) \rho(x) dx \quad , \quad (6.14)$$

where $\mathbb{I}_x(\Delta)$ is a shorthand notation for the indicator function of *x* in the interval $X_t \pm \Delta/2$; the

expectation of the predictor is

$$\mathbb{E}\hat{X}_t = \int \mathbb{I}_x(\Delta)\rho(x)f(x)dx \Big/ \int \mathbb{I}_x(\Delta)\rho(x)dx = \frac{N}{k} \int \mathbb{I}_x(\Delta)\rho(x)f(x)dx \quad . \quad (6.15)$$

require that the real expected value to be equal to the expected value of the average over the neighbors, by integrating over the density ρ : This means that in order for k -NN to be a consistent predictor, the k neighbors have to be in a space Δ small enough such that the density ρ is approximately flat. Additionally, Δ is naturally given by the selection of the k neighbors: These two last equations regulate the k -NN algorithm sensitivity.

6.2.3. Results

Verification of the forecast

The forecasts produced by the AR and k -NN algorithms are single values, so their quality can be assessed just by measuring the distance between the forecast and the target,

$$\epsilon^2 = \frac{1}{N} \sum_{i=1}^N \left(\hat{X}_t^{(i)} - X_t^{(i)} \right)^2 \quad . \quad (6.16)$$

Note that this value is related to the log-likelihood of the AR model, because according to the model, $X_t^{(i)}$ should be distributed normally around $\hat{X}_t^{(i)}$. In addition to this absolute error, we consider also the relative error,

$$\delta^2 = \frac{1}{N} \sum_{i=1}^N \left(\frac{\hat{X}_t^{(i)} - X_t^{(i)}}{X_t^{(i)}} \right)^2 \quad . \quad (6.17)$$

This second type of error will indicate if the absolute error is driven by some items with very high activity.

The protocol to measure the errors in the forecast is the standard for this type of analysis, where two non-intersecting sets are used. The first, the *training set*, is used to fit the models, and the second, the *target set*, is used to verify the forecasts; in this section, N is the total number of items, so each set has $N/2$ items.

The analysis is repeated for different sizes of sets N , where the items have been randomly chosen among the ones with an aggregated activity bigger than 500 at the third day ($X_3 \geq 500$), in order to ensure that they have a non-trivial activity. This set has a size of roughly 107000 items, so the maximum size of training and target sets is set to 50000 items. This analysis was realized with the Python library `scikit-learn`, which significantly reduces the computational effort required.

Results for AR

In Fig. 6.1 the results for the AR forecasts are shown. The forecast is made for the day $t = 10$, using the information of the last τ days. The increase of the size of the sets N is visible in the shrinking of the range of the fluctuations of the errors, which do not decay in average. With respect to the dependence with τ , the error is generally higher for $\tau = 1$ than for $\tau = 2$, and increasing τ does not improve significantly the quality. This can be, more formally, assessed by the Akaike Information Criterion (AIC) [Aka74], which takes into account, in addition to the quality of the model, its complexity; the AIC is defined as

$$\text{AIC} = -2\ell + 2k, \quad (6.18)$$

where k is the number of parameters of the model. While similar to the BIC, the AIC is based on information theory arguments, and is particularly suited to the problem of complexity in the AR model [Aka74, BJRL15]. The correction due to the amount of parameters is $\tau/2$, marginally small with respect to the total value of the log-likelihood. Therefore, the model with $\tau = 2$ is the best model of the AR class for this data.

The choice $\tau = 2$ gives the best description, and since there are only two parameters in the model, it is possible to analyze in a straightforward manner how they are distributed in a bootstrap sample; the results are shown in Fig. 6.2. In this example, where $t = 10$, the average AR(2) model is

$$X_{10} = 1.9603X_9 - 0.9609X_8 + \xi_t = 0.9994X_9 + 0.9609dX_8 + \xi_t, \quad (6.19)$$

which can be approximated as

$$dX_9 \sim 0.9609dX_8 + \xi_t. \quad (6.20)$$

In Fig. 6.2, the sum of the parameters $a_{1,t}$, that multiplies X_{t-1} , and $a_{2,t}$, that multiplies X_{t-2} , is insignificantly different than 1. This result comes from bootstrapping the process 300 times, where the sets are sampled with $N = 10^5$; in the figure are shown the averages of the parameters and the 90% confidence intervals, which overlap with 1 for all the t considered. The confidence intervals are very narrow, which means that $a_{1,t} + a_{2,t}$ is close to one for each bootstrap (i.e. for each random selection of training and target set).

This property, $a_{1,t} + a_{2,t} = 1$, is remarkable, since the AR model coefficients are computed by a simple linear regression through least-squares; if dX_{t-1} is the variable on which dX_t depends, we expect to have fat tails the fluctuations, which is a feature that in principle would invalidate this estimation method. If a different type of fluctuation is desired, then a more detailed analysis is needed, likely through Maximum Likelihood Estimation, that requires a full model of the

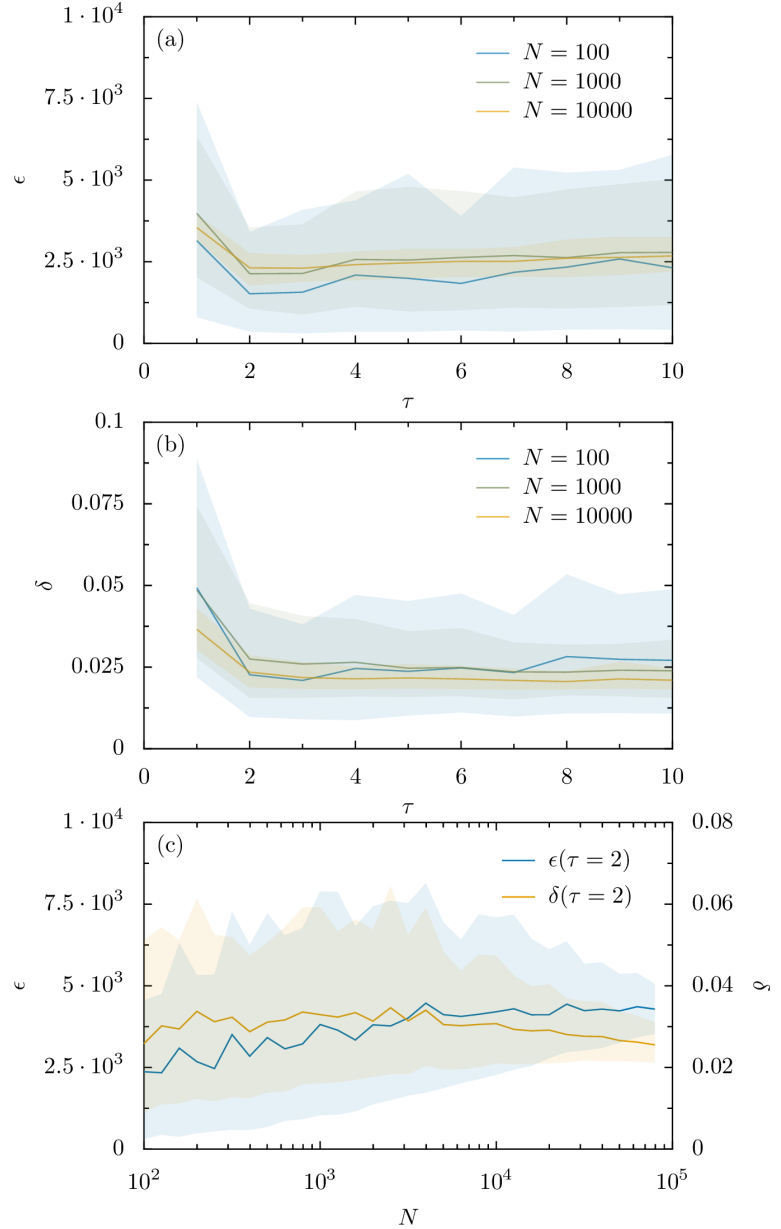


Figure 6.1: Quality of the AR forecasting method. Solid lines: average error; shaded area: 90% percentiles of the error. The time of the target is $t = 10$, and information of the previous τ days is used. (a) Absolute error ϵ as function of τ . (b) Relative error δ as function of τ . (c) Absolute and relative errors as function of N for a fixed $\tau = 2$.

fluctuations, which is increasingly complex as the number of features analyzed grows, as discussed before.

The presence of fat tails has still to be established, but for $\tau = 2$, there is an increase on the absolute forecast error with the increase of the dataset size N , as shown in Fig. 6.1(c) (different τ do not change this behavior – see Fig. 6.1(a)). This is an observation that encourages us to not discard fat-tailed fluctuations, because the absolute error is an estimator of the standard deviation of the fluctuations of the data with respect to the model, which has a dependence on N when fluctuations are fat-tailed, as explained in Section 5.4.1.

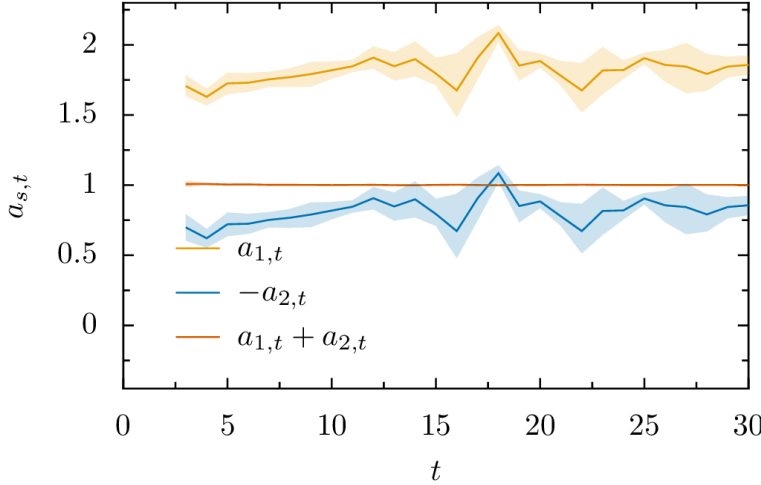


Figure 6.2: Bootstrap of the AR(2) model parameters. The sum of the parameters is very close to 1, an observation that supports the D model. Notice that the confidence intervals of the sum are very narrow, indicating that the sum is close to 1 *for each subset* of the bootstrap, and not only in average. The parameter a_2 is multiplied by -1 for graphical convenience. Solid lines: average (300 times, $N = 10^5$); shaded areas: 90% confidence interval.

Results for k -NN

Although the k Nearest Neighbors is an algorithm that can, in principle, overcome non-linearities and large fluctuations, Fig. 6.3 shows that the quality of the forecast is much worse than for the AR model; in fact, the average error is an order of magnitude bigger. Nevertheless, this forecast scales in a different way with respect to N . While in the AR model τ indicates the complexity of the model (and has an impact on the AIC), in the k -NN method the complexity is given by k [FHT01], while a change in τ modifies just the feature space.

The k -NN method performs marginally worse for high τ , indicating that the new features added, although not increasing the complexity, are not relevant for a better prediction. However, increasing the feature space has the effect of selecting neighbors that are closer in the new, extended, space but not necessarily close in the old, reduced, space. Increasing the dataset size has no effect for the absolute error, but reduces significantly the relative error of the method, a feature that is not present in the AR model.

6.2.4. Summary

From these two simple models we can extract as a conclusion that there is no indication that taking elements in $H_t^{(i)}$ beyond $\tau = 2$ would improve the forecast quality. In particular, the AR model results for $\tau = 2$ suggest further study of a model where the increments dX_t depend on X_{t-1} and X_{t-2} with the particular property $a_{1,t} + a_{2,t} = 1$, which is equivalent to a model where the increments dX_t depend on the previous increment dX_{t-1} ; moreover, the increase of the fluctuations with respect to N indicate the possibility of having fat-tailed fluctuations. These elements are included in the model proposed in the next section.

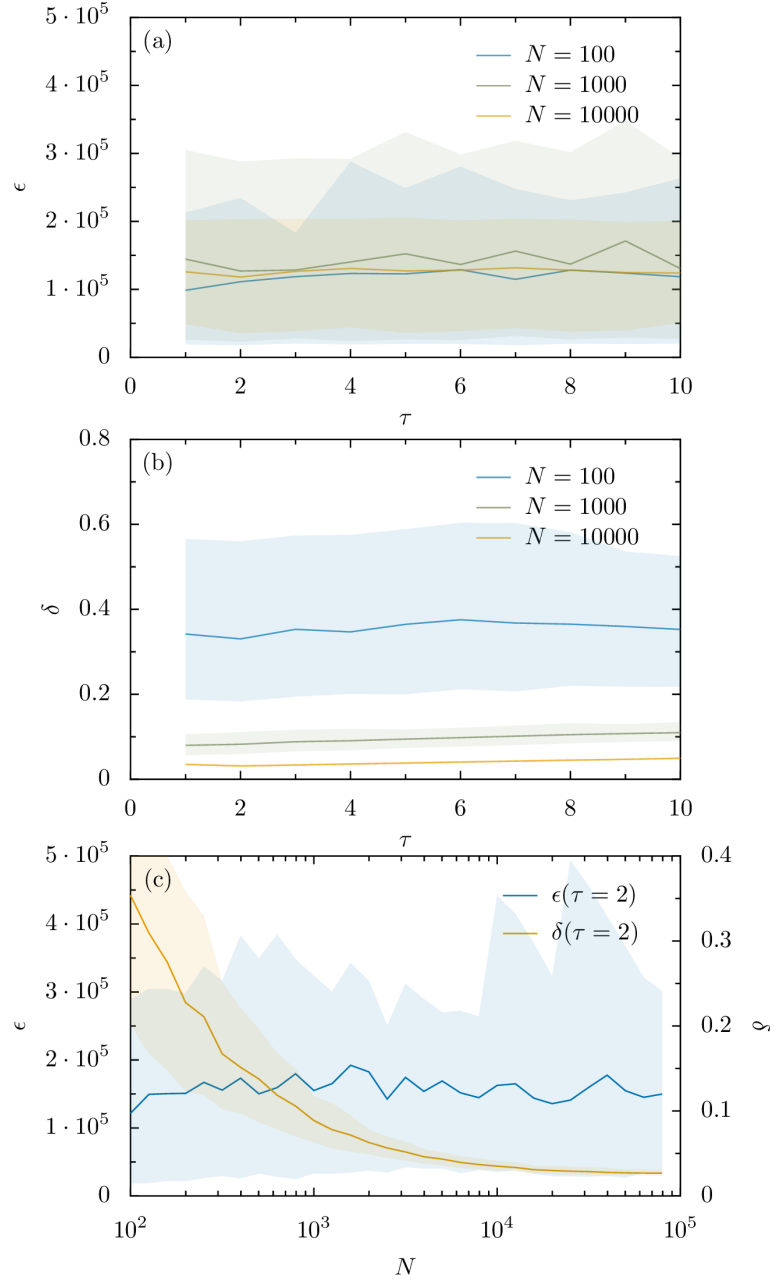


Figure 6.3: Quality of the k -NN forecasting method. Solid lines: average error; shaded area: 90% percentiles of the error. The time of the target is $t = 10$, and information of the previous τ days is used; the number of neighbors was set to $k = 10$ throughout all the analysis, the distance used is the euclidean and the weights given to the neighbors are uniform (i.e. all the neighbors wight the same). (a) Absolute error ϵ as function of τ . (b) Relative error δ as function of τ . (c) Absolute and relative errors as function of N for a fixed $\tau = 2$.

6.3. A new dynamical model: the Daily model

We propose a new dynamical model, where the dependence of the increment with respect to the past is expanded (with respect to the models of Chapter 5) to the pair (X_{t-1}, X_{t-2}) . This equal to the one of X_{t-1}, dX_{t-1} , i.e. the total activity since the release and the previous day activity, two variables that are analogous to the position and the velocity of a particle in a physical system. In particular, based on the results of the previous section, we define a model that only uses dX_{t-1} ;

such daily model, that is denoted with the letter D, should be of the form

$$dX_t = \mu(dX_{t-1}, t)dt + \sigma(dX_{t-1}, t)d\xi_t \quad . \quad (6.21)$$

Note that integrating this equation does not result in a trajectory of total activity X_t , but in a trajectory of daily activity dX_t , which has to be integrated again to get the total activity X_t . So even if the increment only depends on dX_{t-1} , in a simulation it is needed to follow the value of X_{t-1} as well.

The noise term ξ is not specified, but since clearly there are correlations between the total activity X_t and the present activity dX_t (conditioning on one of them preselect values of the other), based on our previous results, it is expected that the distribution of ξ_t is distributed with heavy tails as well. In particular,

$$\mathbb{P}(dX_t|dX_{t-1}) = \sum_{X_t} \mathbb{P}(dX_t|X_t)\mathbb{P}(X_t|dX_{t-1}) \quad , \quad (6.22)$$

meaning that $\mathbb{P}(dX_t|dX_{t-1})$ is a linear combination of the Lévy distributions $\mathbb{P}(dX_t|X_t)$.

6.3.1. Parametric proposal of the Daily model

Here we define the functions μ and σ of Eq. (6.21), by a method analogous to the one of Section 5.4.2. We define histograms $\mathbb{P}(dX_t|dX_{t-1})$ with a minimal amount of items ($N \geq 10^4$), and we compute the quantiles of these distributions, shown in Fig. 6.4, for the quantiles 5%, 25%, 50%, 75% and 95%. The quantiles are linearly related with respect to the condition dX_{t-1} , in a similar fashion as the quantiles of the data conditioned by X_t . In Fig. 6.4 straight lines with slope 1 (linear functions, since the scale is logarithmic in both axes) have also been plotted, which match the quantiles in a large range of values of dX_{t-1} (essentially for $dX_{t-1} > 50$), with a divergence from linearity in the very low values. Therefore, we use as functional forms for μ and σ linear functions, as it was done in the last chapter with the S models.

By the analysis of the quantiles, and in analogy with Eq. (5.29), we therefore propose the model

$$dX_t = (a_t dX_{t-1} + b_t)dt + (c_t dX_{t-1} + d_t)dL_t \quad , \quad (6.23)$$

where dL_t is again a Lévy-stable noise as in Chapter 5.

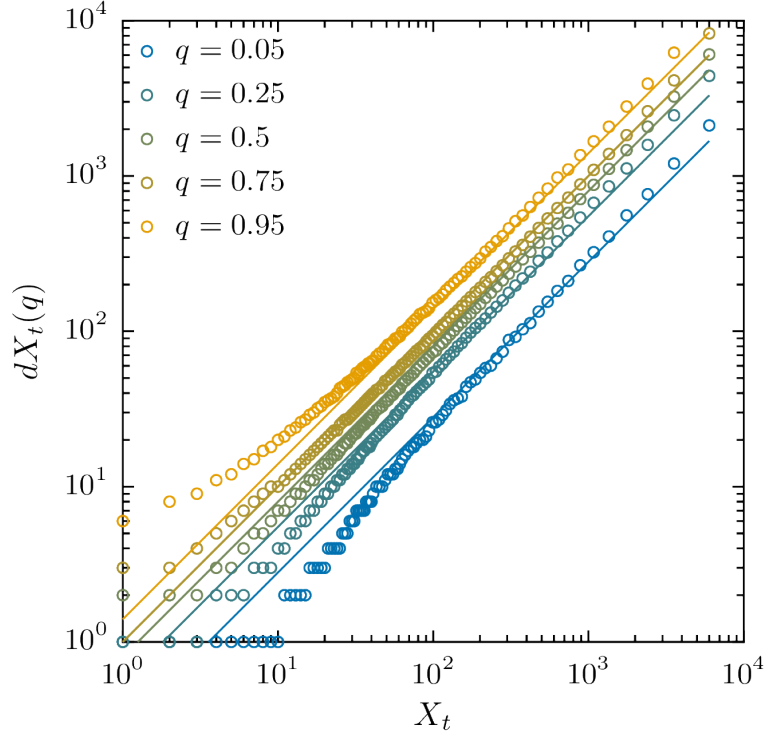


Figure 6.4: Quantiles of the distributions of dX_3 conditioned on dX_2 . The quantiles are computed on groups of items of size $N \geq 10^4$. The lines are linear functions and are shown as references.

6.3.2. Estimation by Maximum Likelihood

The estimation procedure is equal to the one performed for the S model in Section 5.4, where the composite likelihood of dX_t given dX_{t-1} is maximized, i.e. the function ℓ is minimized with

$$\ell = - \sum_{dX_{t-1}} \sum_{dX_t} N(dX_t, dX_{t-1}) \log \mathbb{P}(dX_t | dX_{t-1}) \quad . \quad (6.24)$$

Here $\mathbb{P}(dX_t | dX_{t-1})$ is the probability of dX_t conditioned on dX_{t-1} as given by Eq. (6.23), i.e. a Lévy-stable distribution with location parameter equal to $a_t dX_{t-1} + b_t$ and scale parameter equal to $c_t dX_{t-1} + d_t$. Additionally to α , the parameter of the Lévy distribution β is left free, since the distributions are not necessarily skewed as in the analysis of the S model. The result is a set of six parameters $(\alpha_t, \beta_t, a_t, b_t, c_t, d_t)$, where for most of the values of t , the estimated β is 1 and the estimated b_t is very close to 0, while d_t is again larger than 0. The rescaling of the data by the fitted location and scale parameters is shown in Fig. 6.5. While the bulk of the distribution is notably well adjusted, the tails show some difference with respect to the data. Maximum Likelihood methods basically weight the data points by the logarithm of their probability, so if the tail is not adjusted so well as in the S model, there has to be an improvement in the bulk of the distribution.

The parameters of the model were computed for $t \leq 30$, as in the previous chapter, and are shown in Fig. 6.6. In comparison with the S model, the parameters of change slower with time, with

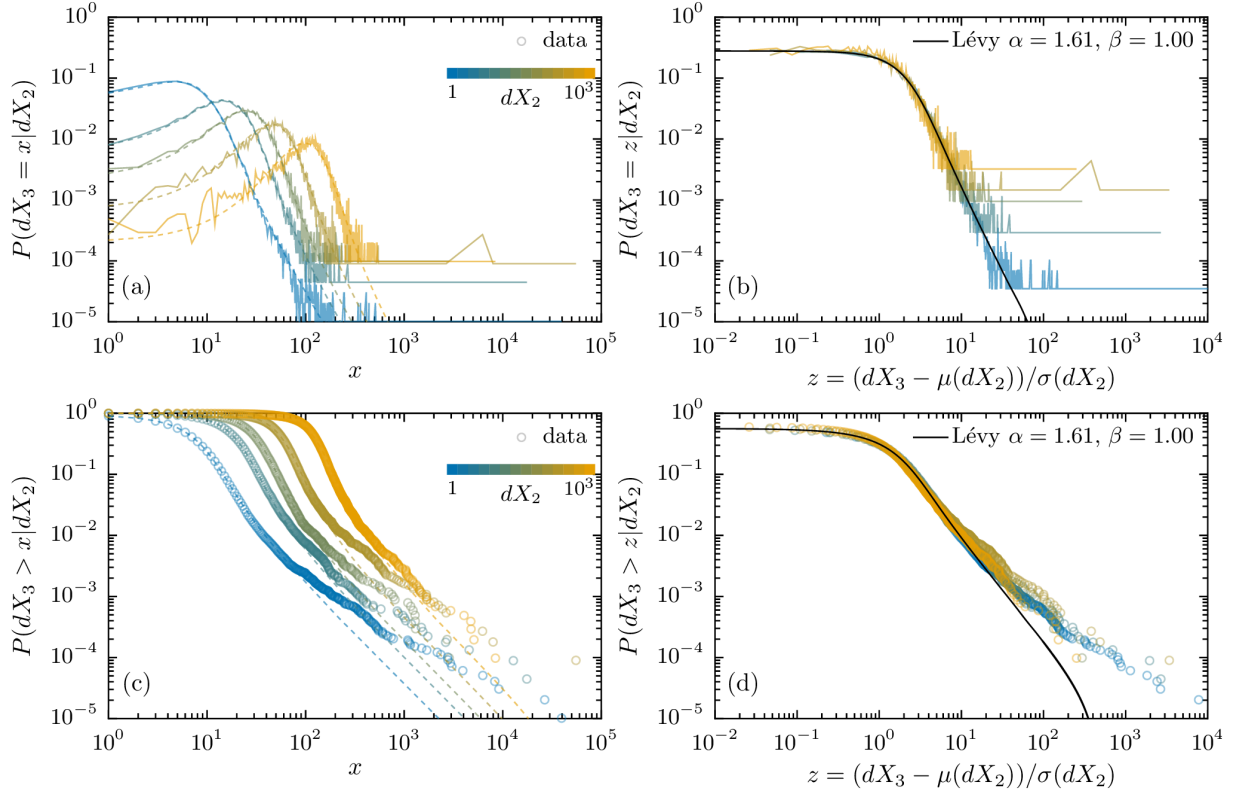


Figure 6.5.: Agreement between data and the D model. Five subsets of points are selected, where each subset is respectively the set of items with $dX_{t=2} \in \{10, 25, 40, 72, 153\}$. (a) PDFs of $dX_{t=3}$ given dX_{t-1} (solid lines), and Lévy PDFs with parameters that correspond to the same dX_{t-1} (dashed lines), based on the global fit of the S model. (b) Rescaling of the PDFs according to the fitted parameters and Lévy PDF with $\mu = 0, \sigma = 1$. (c) CDFs of dX_t given dX_{t-1} (points), and Lévy PDFs with parameters that correspond to the same dX_{t-1} (dashed lines), based on the global fit of the S model. (d) Rescaling of the CDFs according to the fitted parameters and Lévy CDF with $\mu = 0, \sigma = 1$.

a_t , which controls the proportion of activity that persists, tending to 1 as time advances, meaning that for long times the aggregated activity increases linearly, $X_t \propto mt + q$. On the other hand, the change in the parameters stabilizes after the first week; moreover, the index $\alpha_t \rightarrow \alpha_* \approx 1.75$ for $t > 7$. Both observations have resemblance to what obtained for the S model.

Comparison with the S model

To select the best model in terms of descriptive power (between the S and the D model), we employ again the Bayes information Criterion (BIC), which is related to the likelihood of the data being drawn by a model. In Fig. 6.7 the BIC difference is plotted for each time when the analysis was realized, showing overwhelming support for the D model.

Figure 6.6: Evolution of the parameters of the D model with respect to time. The parameters stabilize after roughly a week, as in the aggregate model.

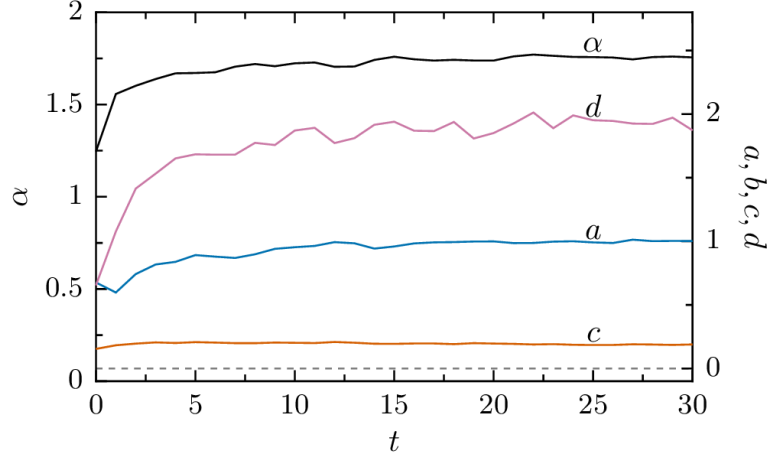
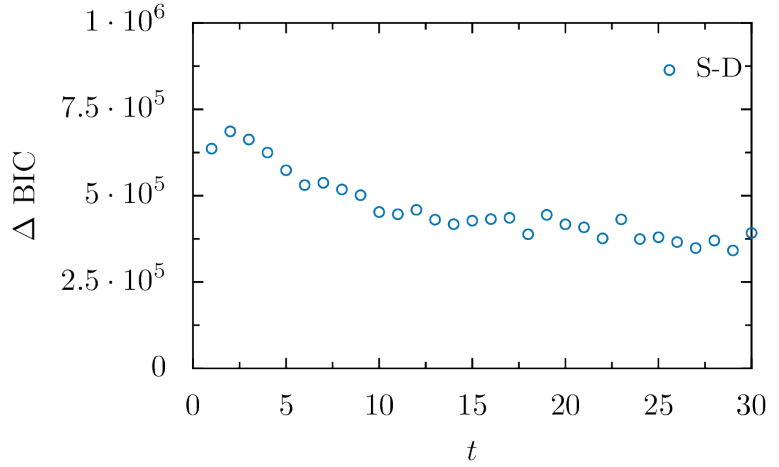


Figure 6.7: Quality of the models. The magnitude (10^5) of the difference in Bayes Information Criterion (ΔBIC) between the D and the S model indicates a very high preference for the D model.



6.4. Forecast of popularity

In order to forecast, as in Section 5.6 we have to compute the probability $\mathbb{P}(X_{t+\tau} = x \mid X_t = x_0)$, i.e. evolving the model from an initial condition x_0 .

The distribution of increments dX_t conditioned on the views X_t is, in the S model, given by the fit of the model itself; this is not the case for the D model, where the increments are related with the previous increments. In order to get the distribution of the increments with a condition on X_t is necessary to include the information of dX_{t-1} , such that

$$\mathbb{P}(dX_t \mid X_t) = \sum_{dX_{t-1}} \mathbb{P}(dX_t \mid dX_{t-1}) \mathbb{P}(dX_{t-1} \mid X_t) \quad . \quad (6.25)$$

If a simulation of the distribution of activity is initialized for items that have the initial activity X_t , for a later time, say $t + \tau$, the sum will run over all the $X_{t+\tau}$ and $dX_{t+\tau-1}$. This requirement is a product of actually using two variables, X_t and X_{t-1} , even if the fit is made only on the condition dX_t . As a consequence, the computational effort is much bigger in comparison with the

S models.

6.4.1. Prediction of big hits

We use the same example as in Section 5.6 to evaluate the quality of the long-term forecast of finding big hits.

The result of evolving the model with the initial condition $X_1 = 100$ is shown in Fig. 6.8, with the result of the S model as well. While in the long term both models seem to misrepresent the real distribution of activity, the D model preserves the overall shape of the data. In particular, the distribution of the S model has at some point a decay faster than data, indicating that items are underdispersed; the D model instead, has a distribution shifted with respect to data, although since the scale in the plot is logarithmic, it should be said actually that the activity is scaled. A common feature in both models is the underestimation of the amount of items without any activity, which can be seen visually by looking where the distributions begin to decay from 1: for $t = 30$, the distribution of the data decreases immediately at the initial value $X_{30} = X_1 = 10^2$, while the models do it at higher values of activity, indicating that very few items are predicted to have small increase in activity. In reality, many items have no activity at all, getting stuck at a certain cumulated value for a certain time, instead of entering and leaving the null activity state constantly. The models we are proposing do not consider a time correlation of this type; the most extreme case is the one where the items have no activity at all. The estimation of the amount of items that have this property is straightforward; if we consider lack of activity for a period of time τ , it is just the product of the probability of having null increments,

$$\mathbb{P}(X_{t+\tau} = x \mid X_t = x) = \prod_{s=1}^{\tau} \mathbb{P}(dX_s = 0) \quad , \quad (6.26)$$

which, if $\mathbb{P}(dX_s = 0)$ has a stable value, would decay approximately as an exponential function with τ . In fact, the exact probability $\mathbb{P}(X_{30} = 100 \mid X_1 = 100)$ is 2×10^{-14} for the S model and 0.0004 for the D model, while for the data is 0.002.

As in Section 5.6, the quality of the model was measured by predicting the amount of extreme events after a certain time, given a condition. The results are shown in Fig. 6.9, where the D model reproduces better the probability of extreme events than the S model. However, as, the S model, the D model shows as well a long-term tendency to deviate from data; while the S model stalls in a certain probability, the D model, instead, overestimate the probability with time. The observations made for Fig. 6.8 can also explain this behavior.

Figure 6.8: Cumulative Distribution Functions of the S and D models, for long-term prediction at $t \in \{2, 10, 30\}$. The D model follows the data for more time than the S model (see $t = 10$), and eventually a shift to higher values of x occurs, visible in $t = 30$; note in particular the lack of decay for low x , which implies an underestimation of such items with no activity.

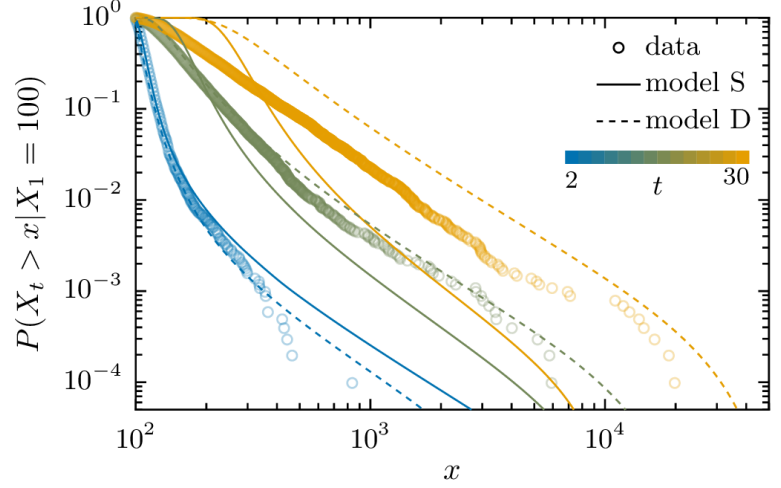
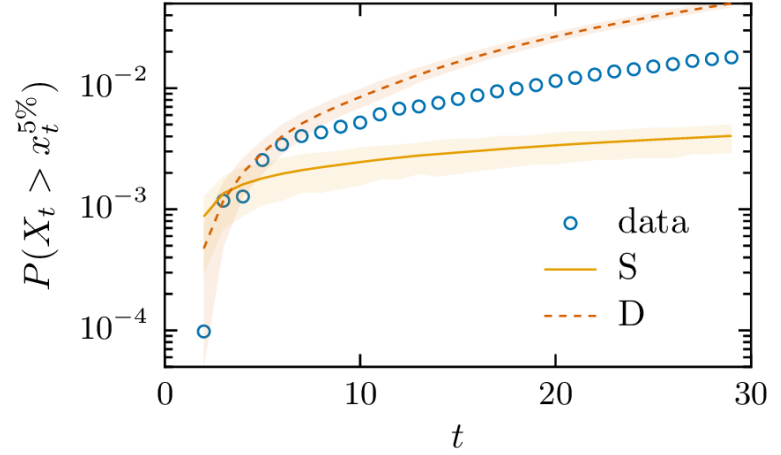


Figure 6.9: Conditional probability of an item to belong to the top 5% quantile, for both S and D models and data. The conditioning is made to select items with $X_1 = 100$.



6.5. Correlations of the models' fluctuations

In this section we compare the correlations measured in the S model with the ones in the D model. The Spearman rank correlation is used over the estimated increments $dL_t^{(i)}$ as in Section 5.6.3. We recall that both models use the simplifying assumption of i.i.d. noise, and therefore predict zero correlation. In Fig. 6.10 the dependence of the correlation with respect to the chosen lag τ is shown. In the D model, the correlation is very low in general, which indicates support to the idea that the D model represent in a more realistic way the dynamics of the items' activity than the S model.

In Fig. 6.11, instead, the results for fixed lags τ are shown. The correlation in the noise of the D model is mostly negligible for $\tau > 1$, while for $\tau = 1$, there is an apparently systematic presence of a small (negative) correlation since the day $t = 12$. The correlation of the D model remains in any case much lower in absolute values than the correlation of the S model.

To summarize, the study of the correlations of the data fluctuations indicated by each model indicates that the D model is better suited to model the data, since these fluctuations appear to

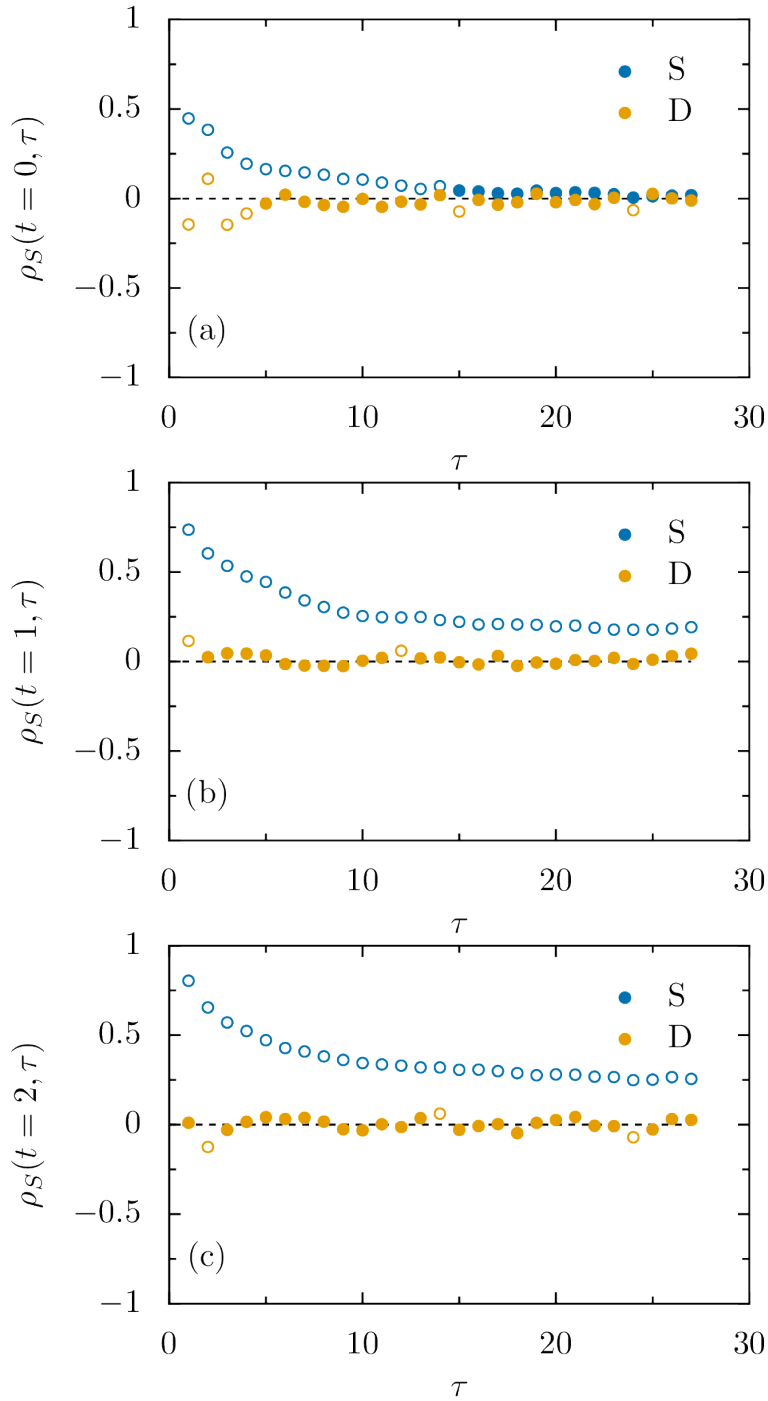


Figure 6.10: Correlation of the estimated Lévy noise with respect to the lag. Different times of the first observation were used, $t = 1, 2, 3$ for (a), (b) and (c) respectively. Where the correlation is not significantly different than zero ($p\text{-value} > 0.05$), the symbols are filled, while where the null hypothesis is rejected are not filled. The correlation in the S model decays to zero only when $t = 0$; this is the correlation between the activity of the first day and the activity of the day τ .

be uncorrelated, one of the hypothesis of stochastic processes to represent faithfully the observed dynamics of activity.

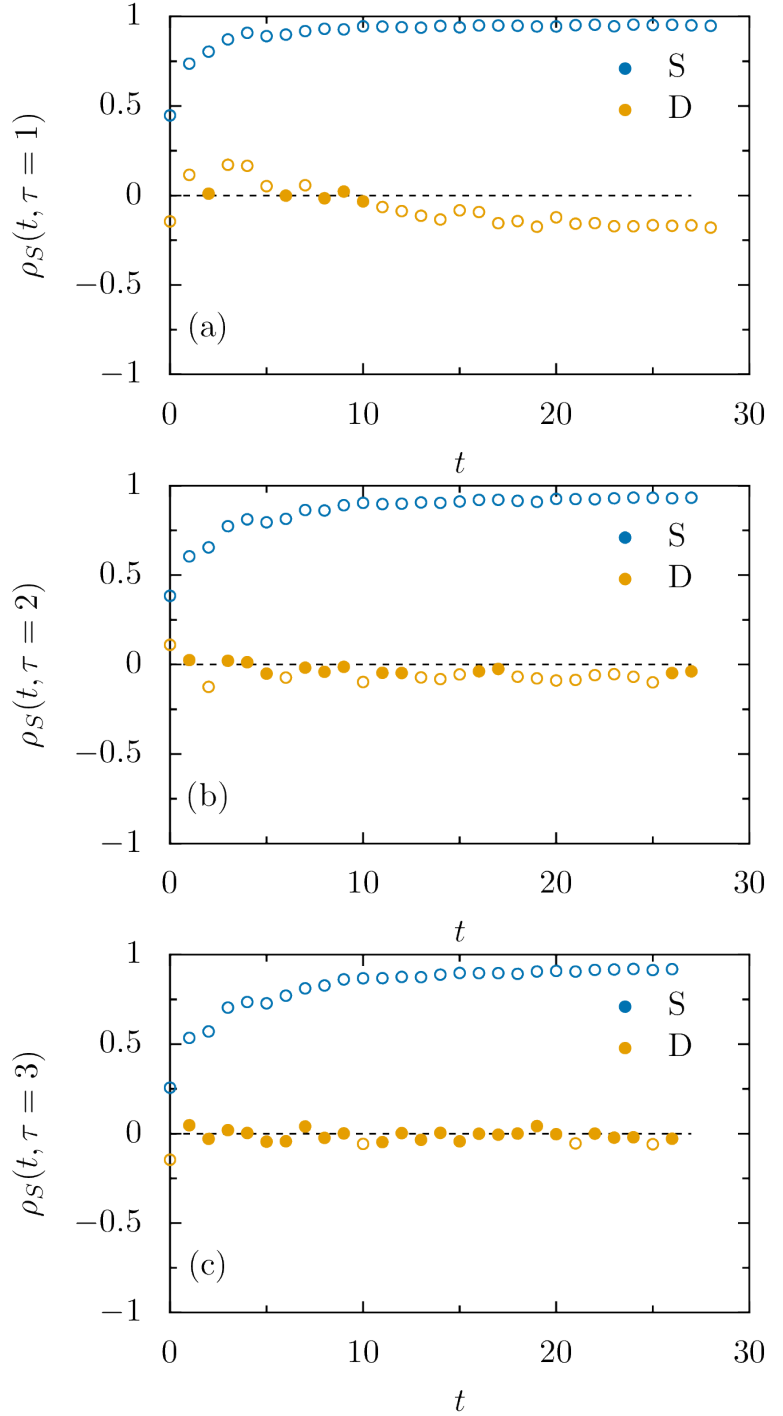


Figure 6.11: Correlation of the estimated Lévy noise with respect to time of the first observation t . Different lags were used, $\tau = 1, 2, 3$ for (a), (b) and (c) respectively. Where the correlation is not significantly different than zero ($p\text{-value} > 0.05$), the symbols are filled, while where the null hypothesis is rejected are not filled.

6.6. Discussion and Conclusions

In this chapter we explored the possibility of having a dynamical model with a longer time dependence than just the last value; this analysis showed that the variable of interest should be instead

the last increment. A stochastic model was constructed with this dependence, which is superior to the model proposed in Chapter 5 both in descriptive power as in estimating the predictability of extreme events, and it practically solves the problem of the correlations of the noise. These results support the idea that fat-tailed fluctuations are inherent to the dynamical process of views accrual, and not an artifact due to the low dimensionality of the model. On the one hand, AR models with longer time span were not better than the D model; on the other hand, potential non-linearities were not uncovered by the k -NN method, where the prediction is based on the behavior of similar items. This is a very important result, because it makes Lévy-stable fluctuations not only suitable for describing the data under proportional effect growth models, but also necessary for a other models of the dynamics of items' activity.

Moreover, the temporal dependence of the parameters offers an additional insight on the dynamics of attention. According to the S model, the expectation of X_{t+1} is

$$\mathbb{E}_S(X_{t+1}) = (1 + a_t)\mathbb{E}(X_t) \quad , \quad (6.27)$$

where a_t is given by $\mathbb{E}(dX_t) = a_t X_t$. According to the D model, it is

$$\mathbb{E}_D(X_{t+1}) = (1 + b_t)\mathbb{E}(X_t) - b_t\mathbb{E}(X_{t-1}) = \mathbb{E}(X_t) + b_t\mathbb{E}(dX_{t-1}) \quad , \quad (6.28)$$

where b_t is given by $\mathbb{E}(dX_t) = b_t dX_{t-1}$. When time increases, both models show the effect of novelty; in the S model a_t decreases approaching 0, while in the D model, b_t increases approaching 1. The fact that $a_t \rightarrow 0$ can lead to the interpretation that attention fades until ultimately nobody is willing to view the item. On the other hand, $b_t \rightarrow 1$ can lead to the interpretation that attention actually stabilizes into a regime where the amount of new viewers is constant, i.e. $\mathbb{E}dX_t = \text{const.} \geq 0$, indicating (i) an average constant interest of the public into the item, or (ii) the sharing of the item is such that a viewer leads to, on average, one other viewer on the next day; to disentangle these two pictures, more detailed information about the process of accruing views is needed, specially on how the item was released. Both possibilities point to a different understanding of how novelty manifests in proportional effect models.

In the D model, however, the increments are not related as a proportional effect in the usual way it is intended. The mentioned trend, $b_t \rightarrow 1$, corresponds asymptotically to the persistence forecasting method, i.e. predicting an increment equal to the last increment [Ahr12]. Nevertheless, proportional effect ($dX_t \propto a_t X_t$) is still being observed; this means that, even very simple models like persistence can superficially look like proportional effect, thus putting into another perspective the ubiquity of this observation.

7. Conclusion

In this Chapter we conclude the Thesis by summarizing and discussing the results. A list of the main specific results is presented in Section 7.1. In Section 7.2 we discuss these results and their relevance in the context of the analysis of systems with fat-tailed distributions and in particular of social activities with proportional effect. In Section 7.3 we present possible extensions of the results and open problems that originate from this work.

7.1. Summary of the results

The main novel results presented in this Thesis are summarized in the following list.

1. Predictability of Extreme Events (Chapter 3 and Chapter 4); these results are published in Ref. [MA14a], while the data is published in Ref. [MA14b].
 - We characterized the distributions of online activity in our datasets by fitting the Generalized Pareto distribution; this was done also for each of the groups in which the data can be categorized. This fit was most of the times statistically significant, indicating that our datasets are representative of the class of fat-tailed social activity.
 - We proposed a measure of Predictability Π , defined as the capacity of forecasting an extreme event by using a given information. This measure is the quality of the best possible forecasting method; since the observable (if the activity of the item will become an extreme event or not), and the predictions are binary, we used the area under the ROC curve as measure of the forecast accuracy. We showed that an optimal strategy exists, i.e. the amount of correct predictions is maximal for a given rate of false positives, and consists in predicting extreme events for all the items in the groups with higher rates of extreme events. This result was shown by using the Simplex algorithm. A straightforward formula is given to compute Π based solely on the probabilities of having an extreme event given the items' group.
 - Applying our proposal for the measure of Predictability, we showed that, in our data, past activity is more informative than any other type of information, in particular release date or simple topical variables.
 - We showed that varying the definition of extreme event, i.e. increasing the threshold by which it is defined, the Predictability increases in all cases, i.e. the more extreme

are the events, the more predictable are. We show that this behavior is a direct consequence of the difference of the tails of the probability of having an event given the used information.

2. Stochastic growth models (Chapter 5); these results are published in Ref. [MKA16], while the program necessary to reproduce the results (in particular the Lévy distributions and their estimation) is publicly available (see Appendix C).

- We computed the exact result of the growth model by allocation with proportional effect described in Section 2.2.2; the result is a probability distribution $\mathbb{P}(dX_t | X_t)$ of a Beta Binomial form, which can be reasonably approximated by a Normal distribution with a certain mean and variance proportional to the mean. We showed that this distribution is not the observed in our data; instead, data must follow much broader distributions.
- Using the framework of Stochastic Differential Equations (SDE) to generalize the allocation model, we showed that the class of linear SDE with Wiener noise dW_t predicts distributions $\mathbb{P}(dX_t | X_t)$ that do not match the observed data, ruling out the classical SDE argument for growth models described in Section 2.2.2. We showed that, instead, Lévy-Stable distributed noise dL_t describes these distributions much better.
- We defined a formal method to estimate the parameters of SDE models, in terms of Maximum Likelihood. We used it to show that the SDE with Lévy-Stable noise as a whole, has a higher Likelihood than the SDE with Wiener noise. Moreover, the evolution in time of the parameters can be assessed, showing a decay in the overall activity after the first week, confirming previous observations of such an effect.
- We showed that the Lévy-Stable SDE model (the "S model") is better than the Wiener SDE models at predicting the amount of extreme events after a longer time than one day. This was done by simulating the distribution through the Chapman-Kolmogorov equation and the conditional distributions predicted by the SDE models. In particular, in the Wiener SDE models the items are very predictable in comparison with empirical data. The S model results in a predictability that is in line with the data, but as time evolves the model departs from data, an effect that we attribute to neglected correlations present in the data.
- We found that, the estimated Lévy-Stable noise dL_t for the S model has correlations (measured by the Spearman Rank correlation coefficient) that are not only statistically significant, but also strong enough to imply a departure in the predictability of the model.

3. Generalized stochastic growth models (Chapter 6).

- We performed an exploratory analysis to understand if a model with more dependencies

was necessary, and if the fat-tailed fluctuations observed are genuine or a spurious effect due to a more complex dynamics. We used the Autoregressive model (AR) and the k -Nearest Neighbors method and we evaluated for both the (out-of-sample) forecast accuracy. The results indicate that a model that includes a dependence of dX_t on both X_t and X_{t-1} can provide better results than a model only dependent on X_t . Moreover, the AR model with these dependencies has coefficients that strongly indicate linear dependence of dX_t with dX_{t-1} .

- We defined a new model, the "D model", where dX_t depends on dX_{t-1} as it depended from X_t in the S model, including Lévy-stable noise. We showed that the Maximum Likelihood estimation technique used before can be used for this model as well, showing a stronger statistical support for the D model. The temporal dependence of the parameters was also analyzed, showing that asymptotically the model tends to a "persistent" state.
- We repeated the test for predicting extreme events after a certain time, and we showed that the D model performs better than the S model. The correlations of the estimated noise was analyzed for the D model and compared with the S model. We found that correlations are greatly reduced in the S model, and, in most cases, not even statistically significant. However, some long-term correlations exist still in the data; in particular, there are items that get systematically 0 views.

7.2. Discussion and outlook

In this Thesis we investigated the dynamics of growth processes with fat-tailed distributions, and in particular online activity, a prominent example of social dynamics, and the consequences that these processes have on the problem of forecasting. The abundance and detail of the datasets, specially the YouTube dataset, allowed us to characterize the temporal evolution of the activity X_t with an improved accuracy when compared with previous works. In particular, we characterized not only the average behavior but also the fluctuations, which we show are fundamental.

We collected datasets from online social media that possess similar characteristics, in particular the presence of fat-tailed distributions of activity, and thus are representative of a larger class of social systems. Fat-tailed distributions require a careful treatment of forecasts, since they can cause lack of robustness in the estimators based on the distributions' moments. Our general approach throughout this Thesis was to use the quantiles of distributions as observables, or, equivalently, if an item would be an extreme event, $E = (X_t > \eta)$, or not. This serves two main purposes. The first is to establish a measure of the predictability of the system, Π given by Eq. (4.3), based solely on the values of $\mathbb{P}(E | g)$, the probabilities of E given that the information being used is g ; we define predictability as the best forecast accuracy attainable by only knowing g , thus it is a

property of the system alone. This means as well that any model that includes that information will be at most as good as Π ; in particular, we found that previous activity is the most informative variable among the available in all datasets, in consonance with similar analysis realized in other social media [MHS⁺16].

The second purpose is the one of measuring how items mix among each other, i.e. how their ranks change. Analyzing this feature of the dynamics is a step beyond the modeling of the average behavior, and has implications on its own, in the context of fat-tailed distributions in society. For instance, it is important to understand the possibility of a person to raise among the ranks of the income distribution (see Chapter 1); in this sense, the problem of measuring economic mobility can be understood as the one of measuring predictability. To predict the amount of extreme events after some time, as done in Section 5.6 and Section 6.4.1, not only tests how appropriate the fluctuations of the model are (as in Chapter 5), but also how much correlation data has if a model is assumed (as in Chapter 6).

The role of models with proportional effect was studied in detail in Chapter 2 and Chapter 5; these are the typical models used to explain fat tails in complex systems [1]. While previous studies focused only in the average behavior due to proportional effect and the observation of fat tails, we try to match the full probabilistic model that leads to such observations to empirical data. We showed that, in our data, even if proportional effect is observed, the fat-tailed distributions of activity $\mathbb{P}(X_t)$ are driven by the fluctuations around this average behavior, $\mathbb{P}(dX_t | X_t)$, which are fat-tailed themselves. This is an important result. On the one hand, having the correct fluctuations in the model allows to predict more accurately the amount of extreme events, as explained before. On the other hand, it indicates that the observation of proportional effect is not enough to explain the statistical properties of the system [RFF⁺10]. Instead, care is needed when linking directly proportional effect and fat tails: we showed that the distribution that results from proportional effect cannot lead to the fat-tailed distributions of increments measured in the data.

We proposed the Lévy-stable distribution as a candidate model for $\mathbb{P}(dX_t | X_t)$, which resulted to be a better fit than other distributions previously proposed. Noise with this distribution, being the most general distribution for sum of random variables, generalizes in a natural way the Wiener noise in the SDE models. The SDE model, given by Eq. (5.29), encompasses both proportional effect and fat-tailed distributions, and proved to be a better description of the data, in comparison with models that do not take into account the leading role of fluctuations in the dynamics. The framework of SDEs for activity dynamics provides a full probabilistic description, and allows to use Maximum Likelihood methods to estimate their parameters; while this framework was previously used in many areas of natural sciences [FPST11], this is one of the first attempts to apply it to growth processes [MSSVK08, MM15], and the first to integrate it with fat-tailed fluctuations. Moreover, the particular setting of the estimation we perform is novel, since we apply it over an ensemble of time series at a certain t , while it is usually done over a single, long time series.

In Chapter 6 models for activity growth that go beyond proportional effect were considered,

and the D model was proposed, given by Eq. (6.23). Besides the fact that its likelihood is higher (in a decisive way) than the S model, this model has also the property of having a practically uncorrelated estimated noise (i.e. the scaled residuals from the average behavior). This observation is decisive to consider the Lévy-stable noise not as a spurious effect from the mixing of items in different dynamical states, but as inherent in the modeling of activity. Notably, the D model can coexist with the observation of proportional effect, which leads to an inferior model. This can be understood in views of the correlation that exists between dX and X (see Eq. (6.25)): it is expected that items with high activity X would be likely to have a bigger increment dX (proportional effect), but at the same time the converse is true, i.e. items with high increments dX will logically have high X . It is needed, therefore, to understand better the nature of the ubiquity of proportional effect, in view of these results.

7.3. Open issues and directions for future work

- We proposed a predictability measure Π and we used to find the most important variables to model and we analyzed how the extreme events threshold affected it. An interesting application beyond these, is to measure how the predictability of the system changes when the fluctuations scale in different ways. Assuming a model where Taylor Law is valid, Eq. (5.23), the variation in β should increase (or decrease) the overall predictability of the system.
- The SDE models make use of a set of different parameters for each time step (i.e. 5 parameters per day); even if the model is not overfitting, it is desirable to find a model with less parameters. In particular, it is possible to replace the particular values of the parameters (α_t , a_t , etc.) with temporal dependent functions ($\alpha(t)$, $a(t)$, etc.). The functional form can be deduced from the shape of the parameters, as seen in Fig. 5.9 and Fig. 6.6, with a given number of parameters. The estimation of these new parameters can be performed in the very same framework of Maximum Likelihood, generalizing Eq. (5.31) by considering a likelihood function for data of all the different times together,

$$\ell = - \sum_t \sum_{X_t} \sum_{dX_t} N(dX_t, X_t) \log \mathbb{P}(dX_t | X_t) \quad , \quad (7.1)$$

where now $\mathbb{P}(dX_t | X_t)$ is a Lévy-stable distribution with $\alpha = \alpha(t)$, $\mu = a(t)X_t + b(t)$, etc. The same can be done for models where dX_t depends on dX_{t-1} . While the method for estimation is clear, it is desirable as well to have an underlying mechanistic model that results in these temporal functions.

- Within the SDE framework, we showed that fluctuations around the average are Lévy-Stable distributed; however, there can be multiple possible sources of these fluctuations.

- One possibility is to consider the noise as the result of bursts of interest in a particular topic; in that case, correlating these sudden peaks in activity with events (from a news database, for instance) may give an insight of how this interest enters into social media [FBA11].
- The problem of the origin of the fat-tailed noise can be also thought as an "internal" process, without the need of external influence. The problem can be first approached by considering the activity of an item as a Branching process [Jac10], where each view of an item by a person triggers some other person to view the same item, a model for sharing (as well as epidemics [BBV08] and information diffusion [IM09]). The total views in a day, for instance, will be the sum of all these views; in this hypothesis the number of persons that can be triggered should be distributed with a fat tail, in order to allow the sum of views to be Lévy distributed. This feature can be naturally introduced if the sharing process occurs on a social network, which usually has fat-tailed degree distribution [BJN⁺02, KW06].
- We analyzed models with Wiener or Lévy noise, which in order to be compared to the data, needed to be discretized. Online activity can also be modeled with Point Processes [DVJ07], where views arrive randomly in time, at a given rate that can be time-dependent. The Poisson process is the most famous model of this class, where the rate of arrivals is constant, but in our case a model that allows for burstiness in the time series is needed. A possibility is to use the self-excited Hawkes process [HO74], which was proposed for the YouTube [CS08, MC09] and papers' citations accrual [SWSB14]; datasets where the timestamp of each action is available (like the Stack-Overflow or the Usenet dataset) would allow for a deeper study of the burstiness issue.
- Another possible approach to modeling the activity is to assign a common dynamics to all the items, but with different parameters for each of them. Such an approach was proposed, for example, for citations accrual [WSB13]; while it is not parsimonious (because a certain number of parameters are needed for each item), it can potentially increase the overall forecast accuracy. A comparison by means of the BIC difference like the one we propose in Section 5.5 and Section 6.3.2 would establish if this approach is better or not; in particular, the presence of fat tails cast doubts on whether the improvement in quality is enough to justify the huge increase in the number of parameters.
- A problem ignored in this Thesis is how an item's content affects its success. Addressing this issue requires to be able to analyze this content, usually a difficult task, specially for video [FABG14]; however, machine-readable formats, like text, are more promising for this type of analysis.
- The mechanism of sharing allows the spreading of an item through the viewers, but at

the same time a viewer can be interested in related content. The user behavior is hard to analyze because users' data is usually protected by privacy regulations; in this case, however, it is possible to analyze the correlations among time series of different items. Additional information can be retrieved using the recommendation systems that online platforms usually have, to find automatically items with related content. An analysis on these lines can also shed light on overall trends of interest from the public [Fig13].

Appendices

A. Computation of the Likelihood

The Maximum Likelihood Estimation proposed needs the computation of $\mathbb{P}(dX_t | X_t)$ for all the pairs (dX_t, X_t) in the data; however, dX_t is always discrete, and $dX_t \geq 0$. For this reason, the continuous distributions $\mathbb{P}(dX_t | X_t)$ are truncated at 0, and discretized. Then, a normalization is applied, since the sum over all the values of dX_t is no longer 1; this requires to compute this sum, for all the values of X_t . Typically, the number of different values that X_t takes, K_X , is around $K_X \sim 55000$ for $t < 10$. With time this value increases because items spread over X_t . This implies that for each evaluation of the log-likelihood ℓ , the distribution function $\mathbb{P}(dX_t | X_t)$ is evaluated over a large number of values (choosing a range of increments up to 50000 is enough for numerical purposes) and then summed, K_X times. ℓ is likely to be evaluated hundreds of times during the minimization process, so it is critical to compute it efficiently in order to avoid very lengthy numerical calculations. The minimization of ℓ was performed using the `fmin` function from the `scipy` Python package (an implementation of the downhill simplex algorithm); the process was repeated 30 times initializing the parameters in random values in order to avoid the algorithm getting stuck in a local minimum.

In order to improve the efficiency of the computation of ℓ , we will perform approximations, mainly to reduce the number of numerical evaluations of $\mathbb{P}(dX_t | X_t)$. This evaluation is straightforward for the case of the Lognormal distribution (and even if more cumbersome, also in the CEV distribution), but not for the Stable distribution, which notably has no explicit formula for its probability density function. The Stable distribution is precomputed numerically through its characteristic function Eq. (2.9) for a grid of values of $\alpha \in [0.5, 2]$ and $\beta \in [0, 1]$ (values of α below cause problems in the numerical integration, and for $\beta < 0$, the distribution can be computed by mirroring the distribution with the correspondent $\beta' = -\beta$). Then, an interpolation is used to compute the values of the distribution for intermediate values of (α, β) . All this process makes the evaluation of the Stable distribution expensive computationally.

We use two approximations for the Stable distribution. The first is to consider a threshold x_* such that $\mathbb{P}(X = x | X > x_*)$ is a power-law. An exact formula for this approximation exists [Nol12],

$$\mathbb{P}(X = x) \approx \alpha \sigma^\alpha c_\alpha (1 + \beta) x^{-(\alpha+1)} \quad , \quad (\text{A.1})$$

(where $c_\alpha = \sin(\pi\alpha/2)\Gamma(\alpha)/\pi$) and we use it to speed up the computation of \mathbb{P} in the tail; moreover, when the normalization of \mathbb{P} is computed, it is possible to use the sum of the full evaluation of \mathbb{P} up to x_* and an exact formula for the sum of the values larger than x_* , since this sum can be expressed in terms of the Hurwitz Zeta function (as the normalization of the Generalized Pareto distribution in Eq. (3.2)). This approximation is implemented in the package `pyLevy` (see Appendix C.2).

The second approximation is by replacing the discrete distribution by the continuous one. This can be done only for high enough values of X_t : this is the region where the importance of truncation is lowered (since the bulk of the distribution of dX_t moves away from the 0), and the same time the discretization is less relevant, since the probability is distributed over many more points. In practice, we select all items such that the number of items that have the same X_t is lower than 10, which is a way of taking the items in the tail of the distribution of X_t . The increments of these items, $dX_t^{(i)}$ are transformed by subtracting the mean of their correspondent Stable distribution, $\mu(X_t^{(i)}, t)$, and scaled by their correspondent scaling factor, $\sigma(X_t^{(i)}, t)$, leaving points that should be distributed according to a Standard Stable distribution ($\mu = 0, \sigma = 1$), thus making possible to evaluate the (Standard) Stable distribution only once for the whole tail. This approximation reduces drastically the amount of evaluations of the Stable distribution, from $K_X \sim 55000$ to 7000.

B. Scaling of fluctuations is incompatible with Wiener noise

The bin size is given by the distribution of X_t , as mentioned; by denoting the borders of the window as x_0, x_1 , they are given by

$$\int_{x_0}^{x_1} \mathbb{P}(X_t = x) dx = \frac{N}{N_{tot}} \quad (\text{B.1})$$

where N_{tot} is the total size of the data set. In order for the method to work, the average of the data has to converge to its expectation value, formally meaning that

$$\langle dX_t | X_t \rangle = \frac{1}{N} \sum_{i=1}^N dX_t^{(i)} \sim \int_{x_0}^{x_1} \mathbb{E}(dX_t | X_t = x) \rho(X_t = x) dx \rightarrow \mathbb{E}(dX_t | X_t) \quad (\text{B.2})$$

when $N \rightarrow \infty$. From the last two equations it can be seen that an increase in N , while reducing the fluctuation of the average, increases at the same time the distance between x_0 and x_1 if N_{tot} is kept constant, mixing variables dX_t that belong to different X_t . Hereafter we consider a model for such mixture and test whether our observations of σ growing with N are explained by it. The expected fluctuation can be estimated from the expectation by considering

$$\mathbb{V}(\langle dX_t | X_t \rangle) = N^{-1} \int_{x_0}^{x_1} \mathbb{V}(dX_t | X_t = x) \mathbb{P}(X_t = x) dx \Big/ \int_{x_0}^{x_1} \mathbb{P}(X_t = x) dx \quad . \quad (\text{B.3})$$

Here it is needed to specify the form of the variance $\mathbb{V}(dX_t | X_t)$; the proposal for the variance scaling of Eq. (5.23), $\mathbb{V}(dX_t | X_t) \propto X_t^{2\beta}$, is therefore used. Additionally, we consider a power-law decay of the X_t distribution, $\mathbb{P}(X_t) \propto X_t^{-\alpha-1}$, and that $x_1 = x_0 + N\gamma$, a first order approximation of the dependence of x_1 with respect to N , with γ as proportionality constant that depends on the exact shape of $\mathbb{P}(X_t)$ and x_0 . The value of the integral in the denominator, N/N_{tot} , is also known. With these hypothesis, we can get an idea of how the estimation of $\mathbb{E}(dX_t | X_t)$ gets better with N . Replacing, we get

$$\mathbb{V}(\langle dX_t | X_t \rangle) \propto N^{-1} \left[\int_{x_0}^{x_0+N\gamma} x^{2\beta-\alpha-1} dx \right] \Big/ (N/N_{tot}) \propto \frac{(x_0 + N\gamma)^{2\beta-\alpha} - x_0^{2\beta-\alpha}}{N^2} \quad . \quad (\text{B.4})$$

This means that the fluctuation of the mean, for very high N , scales with N as

$$\hat{\sigma}(\langle dX_t | X_t \rangle) \propto N^{\beta-\alpha/2-1} \quad (\text{B.5})$$

where the Central Limit Theorem expected relationship is recovered if $\beta = \alpha/2 + 1$, meaning that there is a region of combinations (α, β) for which the error in the estimation does not decay at all, $\beta > (\alpha + 1)/2$, although this condition is rather extreme. On the other hand, from the same calculation the scaling of the expected standard deviation of the window's data points can be

computed, being

$$\hat{\sigma}(dX_t | X_t) \propto N^{\beta-\alpha/2-1/2} \quad . \quad (\text{B.6})$$

Although this relationship predicts that there can be a regime where a substantial difference in the standard deviation exists, we show that for the data taken into account the increase is actually negligible.

In Chapter 3 we showed that the distribution of views in YouTube can be described by a power law with exponent $\alpha \sim 0.7$ for $x > 20$, $N_{tot} \sim 7 \cdot 10^6$; with these values, and using only $X_t = 10^3$, the ratio between the standard deviation of the two values of N can be computed, and is plotted against α and β in Fig. B.1.

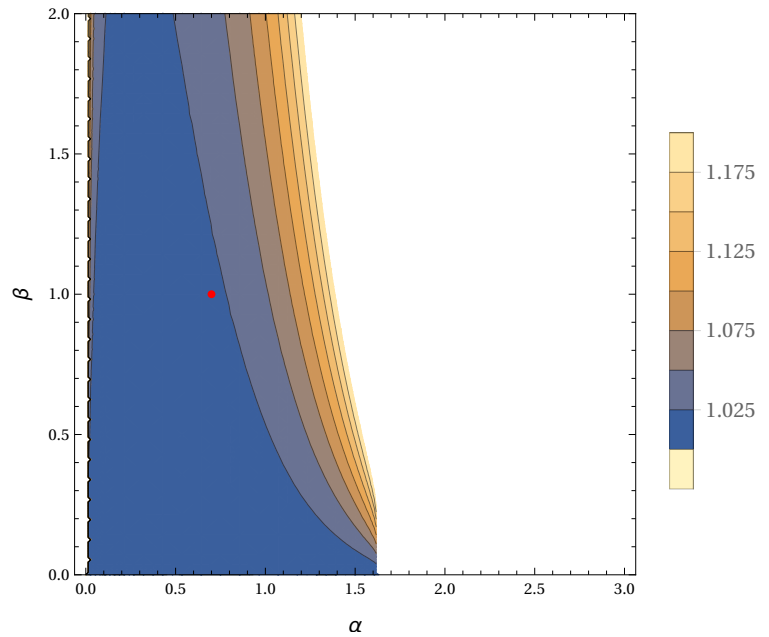


Figure B.1: Ratio of standard deviation with $N = 10^2, 10^4$. The red dot coordinates are $\alpha = 0.7$, $\beta = 1$, roughly the values given for this particular data set.

The ratio for the α, β values given by the data is no more than 2% away from 1. The change measured in the data, instead, is consistently higher, reaching peaks of 850%, therefore is not compatible with the SDE that we are proposing, Eq. (5.23). This can be seen as well in Fig. B.2, where a window of 10^2 videos at $X_{t=3} = 1000$ is enlarged until 10^4 videos are included; the ratio of the standard deviation of the window with size N with respect to the original, of size 10^2 , observed in the data is compared with the ratio expected from the mixture model (the same estimation used in Fig. B.1) and with the expected ratio scaling from a model where the fluctuations are fat-tailed. If the distribution is fat-tailed, and even if the standard deviation is not defined, it is possible to compute an estimate for the fluctuations, by considerations based in extreme value theory [BG90]; the scaling according to this estimate scales with N as $N^{1/\alpha}$, where α is the index of the tail of the distribution. By this analysis it becomes clear that the lack of convergence of the

standard deviation of the bins is not due to the merging of items with different X_t .

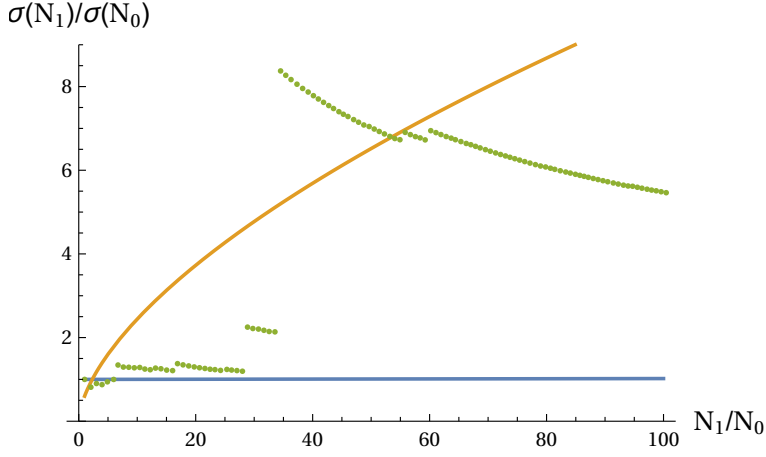


Figure B.2: Ratio of standard deviations with N_0, N_1 sizes, for $X_{t=3} = 1000$. The blue line is the ratio for the mixture model that follows Eq. (5.23) with $\alpha = 0.7$ and $\beta = 1$; the orange line is the ratio for a model where the distribution of $dX_t | X_t$ is Lévy-stable with exponent $\alpha = 1.64$. The points are the values observed in the data for an increasing window at $X_t = 1000$; notice that there are *shocks* given by the appearance of very large jumps in X_t . Although the values of the data points seem close to the unit, already at $N_1/N_0 = 7$ there is a 30% increase in the standard deviation. The maximum is reached at an increase of 850%.

C. Packages released

In this Appendix we mention the main features of two numerical packages in Python language that were mainly created for and used in this Thesis and released to the public. `datagram` is a small package that is useful to handle sparse histograms, as the ones with fat-tailed distributions, and was used throughout the whole Thesis, while `pyLevy` was used to evaluate the Lévy distributions in Chapter 5 and Chapter 6.

C.1. `datagram`

`datagram` features a `Data` class, which uses a dictionary as a container for sparse histograms, avoiding unnecessary requirements. This is particularly useful if the data is discrete, since then the keys of the dictionary are integers. If the data is not discrete, the keys of the dictionary will be rational numbers, appropriately rescaled. The `Data` class can transform the dictionaries, used to aggregate data, to histograms and probability density functions when needed by the user.

The package is available at <https://pypi.python.org/pypi/datagram/0.1>.

C.2. `pyLevy`

`pyLevy` is a package that produces Lévy-distributed random numbers and evaluates its probability density function. The code is based in a package of Paul Harrison (<https://pypi.python.org/pypi/PyLevy>), but it has been modified by adding the tail approximation and implementing the code described in Appendix A.

The computation of the Lévy probability density function is performed by numerical integration of the characteristic function (Eq. (2.9) under parametrization 0 of Ref. [Nol12]), on a grid of values of α and β (every 0.05 for $\alpha \in (0.5, 2)$ and $\beta \in (0.1)$) (values of $\alpha < 0.5$ are very unlikely, and for $\beta < 0$, the distribution can be computed through the one of $-\beta$ using symmetry). For general values of α , β the distribution is computed as an interpolation of the distributions computed already over the grid, through the Catmull-Rom cubic splines.

The package is available at <https://github.com/josemiotto/pylevy>.

List of Figures

1.1. Dynamics of views in YouTube.	2
2.1. Examples of fat-tailed distributions in natural sciences.	7
2.2. Examples of fat-tailed distributions in social sciences.	8
2.3. Proportional effect in citations' dynamics.	16
3.1. Cumulative distribution functions for each dataset.	24
4.1. Quantifying the quality of event-prediction strategies requires measuring both the hit and false alarm rates.	33
4.2. Predictability increases for extreme events.	37
4.3. Predictability of simple stochastic processes.	39
5.1. Mixing of views dynamics	43
5.2. Mean of dX_t conditioned on X_t ; $t = 3$ days, $dt = 1$ day.	44
5.3. Beta Binomial Distribution.	48
5.4. Standard deviation of dX_t conditioned on X_t at $t = 3$ days.	54
5.5. Data and fits of dX_t for a bin of size $N = 10^4$	56
5.6. Quantiles of the distributions of dX_3 conditioned on X_3	56
5.7. Agreement between data and the S3 Lévy-Stable model.	58
5.8. Quality of the models.	60
5.9. Evolution of the parameters of the model, $0 \leq t \leq 20$	60
5.10. Expected distribution of views after 5 days for videos such that $X_1 = 100$	63
5.11. Expected amount of videos in the top 5% among the ones such that $X_1 = 100$	63
5.12. Correlation of the estimated Lévy noise with respect to the lag τ	66
5.13. Correlation of the estimated Lévy noise with respect to time of the first observation t	66
6.1. Quality of the AR forecasting method.	76
6.2. Bootstrap of the AR(2) model parameters.	77
6.3. Quality of the k -NN forecasting method.	78
6.4. Quantiles of the distributions of dX_3 conditioned on dX_2	80
6.5. Agreement between data and the D model.	81

6.6. Evolution of the parameters of the D model with respect to time.	82
6.7. Quality of the models.	82
6.8. Cumulative Distribution Functions of the S and D models	84
6.9. Conditional probability of an item to belong to the top 5% quantile	84
6.10. Correlation of the estimated Lévy noise with respect to the lag.	85
6.11. Correlation of the estimated Lévy noise with respect to time of the first observation t	86
 B.1. Ratio of standard deviation with $N = 10^2, 10^4$	 100
B.2. Ratio of standard deviations with N_0, N_1 sizes, for $X_{t=3} = 1000$	101

Bibliography

- [AG05] L. A. Adamic and N. Glance, *The political blogosphere and the 2004 us election: divided they blog*, Proceedings of the 3rd international workshop on Link discovery, ACM, 2005, pp. 36–43. Cited on page 8.
- [AH02] L. A. Adamic and B. A. Huberman, *Zipf’s law and the Internet*, Glottometrics **3** (2002), no. 1, 143–150. Cited on page 8.
- [Ahr12] C. D. Ahrens, *Meteorology today: an introduction to weather, climate, and the environment*, Cengage Learning, 2012. Cited on page 87.
- [AJK06] S. Albeverio, V. Jentsch, and H. Kantz, *Extreme events in nature and society*, Springer Verlag, 2006. Cited on pages 2 and 29.
- [Aka74] H. Akaike, *A new look at the statistical model identification*, Automatic Control, IEEE Transactions on **19** (1974), no. 6, 716–723. Cited on page 75.
- [ALPH01] L. A. Adamic, R. M. Lukose, A. R. Puniyani, and B. A. Huberman, *Search in power-law networks*, Physical review E **64** (2001), no. 4, 046135. Cited on page 1.
- [APM11] E. G. Altmann, J. B. Pierrehumbert, and A. Motter, *Niche as a determinant of word fate in online groups*, PLoS ONE **6** (2011), no. 5. Cited on pages 10 and 27.
- [BA99] A. Barabási and R. Albert, *Emergence of scaling in random networks*, Science **286** (1999), no. 5439, 509–512. Cited on pages 1, 8, and 14.
- [Bax07] R. J. Baxter, *Exactly solved models in statistical mechanics*, Courier Corporation, 2007. Cited on page 6.
- [BB11] M. I. Bogachev and A. Bunde, *On the predictability of extreme events in records with linear and nonlinear long-range memory: Efficiency and noise robustness*, Physica A **390** (2011), no. 12, 2240–2250. Cited on pages 29 and 41.
- [BBV08] A. Barrat, M. Barthélemy, and A. Vespignani, *Dynamical processes on complex networks*, Cambridge University Press, 2008. Cited on page 94.
- [BDH74] A. A. Balkema and L. De Haan, *Residual life time at great age*, Ann. Probab. (1974), 792–804. Cited on page 11.

- [BG90] J. Bouchaud and A. Georges, *Anomalous diffusion in disordered media: statistical mechanisms, models and physical applications*, Phys. Rep. **195** (1990), no. 4, 127–293. Cited on page 100.
- [BJN⁺02] A. Barabási, H. Jeong, Z. Neda, E. Ravasz, A. Schubert, and T. Vicsek, *Evolution of the social network of scientific collaborations*, Physica A: Statistical mechanics and its applications **311** (2002), no. 3, 590–614. Cited on page 94.
- [BJRL15] G. E. P. Box, G. M. Jenkins, G. C. Reinsel, and G. M. Ljung, *Time series analysis: forecasting and control*, John Wiley & Sons, 2015. Cited on pages 70 and 75.
- [BP95] P. Bak and M. Paczuski, *Complexity, contingency, and criticality*, Proc. Natl. Acad. Sci. USA **92** (1995), no. 15, 6689–96. Cited on pages 13 and 41.
- [Bre96] L. Breiman, *Bagging predictors*, Machine learning **24** (1996), no. 2, 123–140. Cited on page 20.
- [Bri50] G. W. Brier, *Verification of forecasts expressed in terms of probability*, Monthly weather review **78** (1950), no. 1, 1–3. Cited on pages 39 and 40.
- [Brö09] J. Bröcker, *Reliability, sufficiency, and the decomposition of proper scores*, Q. J. R. Meteorol. Soc. **135** (2009), no. 643, 1512–1519. Cited on pages 20, 32, and 40.
- [BRWP03] F. Boettcher, C. H. Renner, H. Waldl, and J. Peinke, *On the statistics of wind gusts*, Boundary-Layer Meteorology **108** (2003), no. 1, 163–173. Cited on page 1.
- [BS95] A. Barabási and H. E. Stanley, *Fractal concepts in surface growth*, Cambridge university press, 1995. Cited on page 6.
- [BS02] D. Brockmann and I. M. Sokolov, *Lévy flights in external force fields: from models to equations*, Chem. Phys. **284** (2002), no. 1, 409–421. Cited on page 67.
- [BYRN99] R. Baeza-Yates and B. Ribeiro-Neto, *Modern information retrieval*, vol. 463, ACM press New York, 1999. Cited on page 32.
- [CD99] J. M. Carlson and J. Doyle, *Highly optimized tolerance: A mechanism for power laws in designed systems*, Physical Review E **60** (1999), no. 2, 1412. Cited on page 13.
- [CFL09] C. Castellano, S. Fortunato, and V. Loreto, *Statistical physics of social dynamics*, Rev. Mod. Phys. **81** (2009), 591–646. Cited on page 41.
- [CG05] F. Clementi and M. Gallegati, *Pareto’s law of income distribution: Evidence for Germany, the United Kingdom, and the United States*, Econophysics of wealth distributions, Springer, 2005, pp. 3–14. Cited on page 7.

- [Cha53] D. G. Champernowne, *A model of income distribution*, Econ. J. (1953), 318–351. Cited on page 14.
- [CL10] R. Chen and C. Lee, *A constant elasticity of variance (CEV) family of stock price distributions in option pricing, review and integration*, Handbook of Quantitative Finance and Risk Management, Springer US, 2010. Cited on page 51.
- [CN86] J. L. Cambier and M. Nauenberg, *Distribution of fractal clusters and scaling in the Ising model*, Phys. Rev. B **34** (1986), 8071–8079. Cited on page 6.
- [Col01] S. Coles, *An introduction to statistical modeling of extreme values*, Springer, 2001. Cited on page 22.
- [COS⁺13] H. L. D. d. S. Cavalcante, M. Oriá, D. Sornette, E. Ott, and D. J. Gauthier, *Predictability and suppression of extreme events in a chaotic system*, Physical review letters **111** (2013), no. 19, 198701. Cited on page 19.
- [CPB97] R. Cont, M. Potters, and J. Bouchaud, *Scaling in stock market data: stable laws and beyond*, Scale invariance and beyond, Springer, 1997, pp. 75–85. Cited on page 52.
- [CS08] R. Crane and D. Sornette, *Robust dynamic classes revealed by measuring the response function of a social system*, Proc. Natl. Acad. Sci. USA **105** (2008), no. 41, 15649–15653. Cited on pages 10 and 94.
- [CSF10] G. Chatzopoulou, C. Sheng, and M. Faloutsos, *A first step towards understanding popularity in YouTube*, INFOCOM IEEE Conference on Computer Communications Workshops, 2010, IEEE, 2010, pp. 1–6. Cited on page 10.
- [CSN09] A. Clauset, C. R. Shalizi, and M. E. J. Newman, *Power-law distributions in empirical data*, SIAM review **51** (2009), no. 4, 661–703. Cited on pages 1, 10, 22, 23, 55, and 56.
- [CST⁺11] A. Chmiel, J. Sienkiewicz, M. Thelwall, G. Paltoglou, K. Buckley, A. Kappas, and J. A. Hołyst, *Collective emotions online and their influence on community life*, PloS one **6** (2011), no. 7, e22207. Cited on page 22.
- [Dan98] G. B. Dantzig, *Linear programming and extensions*, Princeton university press, 1998. Cited on page 34.
- [DB13] T. H. Davenport and J. C. Beck, *The attention economy: Understanding the new currency of business*, Harvard Business Press, 2013. Cited on page 8.
- [DD10] P. S. Dodds and C. M. Danforth, *Measuring the happiness of large-scale written expression: Songs, blogs, and presidents*, Journal of Happiness Studies **11** (2010), no. 4, 441–456. Cited on page 22.

- [DHF07] L. De Haan and A. Ferreira, *Extreme value theory: an introduction*, Springer Science & Business Media, 2007. Cited on page 11.
- [Dit99] P. D. Ditlevsen, *Observation of α -stable noise induced millennial climate changes from an ice-core record*, Geophys. Res. Lett. **26** (1999), no. 10, 1441–1444. Cited on page 67.
- [Don52] M. D. Donsker, *Justification and extension of Doob's heuristic approach to the Kolmogorov-Smirnov theorems*, The Annals of mathematical statistics **2** (1952), 277–281. Cited on page 50.
- [Doo42] J. L. Doob, *The Brownian movement and stochastic equations*, Annals of Mathematics (1942), 351–369. Cited on page 51.
- [DVJ07] D. J. Daley and D. Vere-Jones, *An introduction to the theory of point processes: volume II: general theory and structure*, Springer Science & Business Media, 2007. Cited on page 94.
- [DY00] A. Dragulescu and V. M. Yakovenko, *Statistical mechanics of money*, The European Physical Journal B-Condensed Matter and Complex Systems **17** (2000), no. 4, 723–729. Cited on page 8.
- [EBK08] Z. Eisler, I. Bartos, and J. Kertesz, *Fluctuation scaling in complex systems: Taylor's law and beyond*, Adv. Phys. **57** (2008), no. 1, 89–142. Cited on page 50.
- [EL03] E. Eisenberg and E. Y. Levanon, *Preferential attachment in the protein network evolution*, Physical review letters **91** (2003), no. 13, 138701. Cited on page 16.
- [EP23] F. Eggenberger and G. Pólya, *Über die statistik verketteter vorgänge*, ZAMM-Journal of Applied Mathematics and Mechanics / Zeitschrift für Angewandte Mathematik und Mechanik **3** (1923), no. 4, 279–289. Cited on page 46.
- [FABG14] F. Figueiredo, J. M. Almeida, F. Benevenuto, and K. P. Gummadi, *Does content determine information popularity in social media?: A case study of youtube videos' content and their popularity*, Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, ACM, 2014, pp. 979–982. Cited on page 94.
- [FBA11] F. Figueiredo, F. Benevenuto, and J. M. Almeida, *The tube over time: characterizing popularity growth of youtube videos*, Proceedings of the fourth ACM international conference on Web search and data mining, ACM, 2011, pp. 745–754. Cited on page 94.

- [FC10] J. H. Fowler and N. A. Christakis, *Cooperative behavior cascades in human social networks*, Proceedings of the National Academy of Sciences **107** (2010), no. 12, 5334–5338. Cited on page 8.
- [FHT01] J. Friedman, T. Hastie, and R. Tibshirani, *The elements of statistical learning*, Springer series in statistics Springer, Berlin, 2001. Cited on pages 19, 59, 72, and 77.
- [Fig13] F. Figueiredo, *On the prediction of popularity of trends and hits for user generated videos*, Proceedings of the sixth ACM international conference on Web search and data mining, ACM, 2013, pp. 741–746. Cited on page 95.
- [FL13] M. Fenner and J. Lin, *Cumulative usage statistics for plos papers from plos website*, Available: <http://dx.doi.org/10.6084/m9.figshare.816962>. Accessed 2014 Jun 4., 2013. Cited on page 27.
- [FP97] R. Friedrich and J. Peinke, *Description of a turbulent cascade by a fokker-planck equation*, Physical Review Letters **78** (1997), no. 5, 863. Cited on page 53.
- [FPST11] R. Friedrich, J. Peinke, M. Sahimi, and M. R. R. Tabar, *Approaching complexity by stochastic methods: From biological systems to turbulence*, Physics Reports **506** (2011), no. 5, 87–162. Cited on pages 53 and 92.
- [FR71] E. F. Fama and R. Roll, *Parameter estimates for symmetric stable distributions*, Journal of the American Statistical Association **66** (1971), no. 334, 331–338. Cited on page 57.
- [FRSP02] R. Friedrich, C. Renner, M. Siefert, and J. Peinke, *Comment on "Indispensable finite time corrections for Fokker-Planck equations from time series data"*, Physical review letters **89** (2002), no. 14, 149401. Cited on page 57.
- [GA13] M. Gerlach and E. G. Altmann, *Stochastic model for the vocabulary growth in natural languages*, Phys. Rev. X **3** (2013), no. 2, 021006. Cited on page 7.
- [Gab99] X. Gabaix, *Zipf's law for cities: an explanation*, Q. J. Econ. **114** (1999), no. 3, 739–767. Cited on pages 1 and 14.
- [GGMA14] F. Ghanbarnejad, M. Gerlach, J. M. Miotto, and E. G. Altmann, *Extracting information from S-curves of language change*, J. R. Soc. Interface **11** (2014), no. 101, 20141044. Cited on page 10.
- [Gib30] R. Gibrat, *Une loi des réparations économiques: l'effet proportionnel*, Bull. Statist. gén Fr **19** (1930), 469. Cited on pages 3 and 13.

- [GJY03] A. Göing-Jaeschke and M. Yor, *A survey and some generalizations of Bessel processes*, Bernoulli **9** (2003), no. 2, 313–349. Cited on page 51.
- [GMY04] M. L. Goldstein, S. A. Morris, and G. G. Yen, *Problems with fitting to the power-law distribution*, The European Physical Journal B-Condensed Matter and Complex Systems **41** (2004), no. 2, 255–258. Cited on page 1.
- [GR56] B. Gutenberg and C. F. Richter, *Magnitude and energy of earthquakes*, Annals of Geophysics **9** (1956), no. 1, 1–15. Cited on pages 1 and 6.
- [Gra80] C. W. J. Granger, *Testing for causality: a personal viewpoint*, Journal of Economic Dynamics and control **2** (1980), 329–352. Cited on page 41.
- [Gri93] S. D. Grimshaw, *Computing maximum likelihood estimates for the generalized Pareto distribution*, Technometrics **35** (1993), no. 2, 185–191. Cited on page 23.
- [GS13] M. Golosovsky and S. Solomon, *The transition towards immortality: Non-linear autocatalytic growth of citations to scientific papers*, Journal of Statistical Physics **151** (2013), no. 1-2, 340–354. Cited on page 15.
- [GYH⁺11] M. Ghil, P. Yiou, S. Hallegatte, B. D. Malamud, P. Naveau, A. Soloviev, P. Friederichs, V. Keilis-Borok, D. Kondrashov, V. Kossobokov, O. Mestre, C. Nicolis, H. W. Rust, P. Shebalin, M. Vrac, A. Witt, and I. Zaliapin, *Extreme events: dynamics, statistics and prediction*, Nonlinear. Process. Geophys. **18** (2011), no. 3, 295–350. Cited on page 29.
- [HAHK07] S. Hallerberg, E. G. Altmann, D. Holstein, and H. Kantz, *Precursors of extreme increments*, Phys. Rev. E **75** (2007), no. 1, 016706. Cited on pages 29, 31, and 41.
- [HF11] C. Honisch and R. Friedrich, *Estimation of Kramers-Moyal coefficients at low sampling rates*, Physical Review E **83** (2011), no. 6, 066701. Cited on page 57.
- [HK08] S. Hallerberg and H. Kantz, *Influence of the event magnitude on the predictability of an extreme event*, Phys. Rev. E **77** (2008), no. 1, 011108. Cited on pages 29 and 41.
- [HM82] J. A. Hanley and B. McNeil, *The meaning and use of the area under a receiver operating characteristic (ROC) curve*, Radiology **143** (1982), 29–36. Cited on pages 32 and 33.
- [HO74] A. G. Hawkes and D. Oakes, *A cluster process representation of a self-exciting process*, Journal of Applied Probability (1974), 493–503. Cited on page 94.

- [HT01] D. J. Hand and R. J. Till, *A simple generalisation of the area under the roc curve for multiple class classification problems*, Mach. Learn. **45** (2001), no. 2, 171–186. Cited on pages 32 and 33.
- [IM09] J. L. Iribarren and E. Moro, *Impact of human activity patterns on the dynamics of information diffusion*, Physical review letters **103** (2009), no. 3, 038702. Cited on page 94.
- [Jac10] C. Jacob, *Branching processes: their role in epidemiology*, International journal of environmental research and public health **7** (2010), no. 3, 1186–1204. Cited on page 94.
- [JNB03] H. Jeong, Z. Néda, and A. Barabási, *Measuring preferential attachment in evolving networks*, EPL (Europhysics Letters) **61** (2003), no. 4, 567. Cited on pages 14, 16, and 17.
- [JW93] A. Janicki and A. Weron, *Simulation and chaotic behavior of alpha-stable stochastic processes*, CRC Press, 1993. Cited on pages 52 and 66.
- [Kah73] D. Kahneman, *Attention and effort*, Citeseer, 1973. Cited on page 8.
- [KAH⁺06] H. Kantz, E. G. Altmann, S. Hallerberg, D. Holstein, and A. Riegert, *Dynamical interpretation of extreme events: predictability and predictions*, Extreme Events in Nature and Society (S Albeverio, V Jentsch, and H Kantz, eds.), Springer Verlag, 2006. Cited on pages 19, 29, and 41.
- [Ken48] M. G. Kendall, *Rank correlation methods*, Griffin, 1948. Cited on pages 64 and 65.
- [Kes73] H. Kesten, *Random difference equations and renewal theory for products of random matrices*, Acta Mathematica **131** (1973), no. 1, 207–248. Cited on page 14.
- [KFNP05] D. Kleinhans, R. Friedrich, A. Nawroth, and J. Peinke, *An iterative procedure for the estimation of drift and diffusion coefficients of langevin processes*, Phys. Lett. A **346** (2005), no. 1, 42–46. Cited on page 57.
- [KHRV04] H. Kantz, D. Holstein, M. Ragwitz, and N. K. Vitanov, *Markov chain model for turbulent wind speed data*, Physica A: Statistical Mechanics and its Applications **342** (2004), no. 1, 315–321. Cited on page 1.
- [KPS12] P. E. Kloeden, E. Platen, and H. Schurz, *Numerical solution of SDE through computer experiments*, Springer Science & Business Media, 2012. Cited on pages 14, 15, 46, and 49.

- [KW06] G. Kossinets and D. J. Watts, *Empirical analysis of an evolving social network*, Science **311** (2006), no. 5757, 88–90. Cited on page 94.
- [LAH07] J. Leskovec, L. A. Adamic, and B. A. Huberman, *The dynamics of viral marketing*, ACM Transactions on the Web (TWEB) **1** (2007), no. 1, 5. Cited on page 9.
- [LBK09] J. Leskovec, L. Backstrom, and J. Kleinberg, *Meme-tracking and the dynamics of the news cycle*, Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining, ACM, 2009, pp. 497–506. Cited on page 59.
- [LFL98] T. K. Landauer, P. W. Foltz, and D. Laham, *An introduction to latent semantic analysis*, Discourse processes **25** (1998), no. 2-3, 259–284. Cited on page 22.
- [LMGA16] J. Leitaó, J. Miotto, M. Gerlach, and E. Altmann, *Is this scaling nonlinear?*, in consideration, 2016. Cited on pages 8 and 59.
- [Lor69] E. N. Lorenz, *Atmospheric predictability as revealed by naturally occurring analogues*, Journal of the Atmospheric sciences **26** (1969), no. 4, 636–646. Cited on page 72.
- [MA14a] J. M. Miotto and E. G. Altmann, *Predictability of extreme events in social media*, PLoS ONE **9** (2014), no. 11, e111506. Cited on page 89.
- [MA14b] ———, *Time series of social media activity*. Youtube, Usenet, Stack-Overflow, PLoS ONE., Accessed 2015 May 21. Available at <http://dx.doi.org/10.6084/m9.figshare.1160515>, 2014. Cited on pages 21, 26, 35, and 89.
- [Man53] B. Mandelbrot, *An informational theory of the statistical structure of language*, Communication theory **84** (1953), 486–502. Cited on page 13.
- [Man63] ———, *The variation of certain speculative prices*, J. Bus. **36** (1963), no. 4, 394–419. Cited on page 52.
- [MBK99] R. Metzler, E. Barkai, and J. Klafter, *Anomalous diffusion and relaxation close to thermal equilibrium: A fractional Fokker-Planck equation approach*, Phys. Rev. Lett. **82** (1999), no. 18, 3563. Cited on page 67.
- [MC09] L. Mitchell and M. E. Cates, *Hawkes process as a model of social interactions: a view on video dynamics*, Journal of Physics A: Mathematical and Theoretical **43** (2009), no. 4, 045101. Cited on page 94.
- [McC86] J. H. McCulloch, *Simple consistent estimators of stable distribution parameters*, Communications in Statistics-Simulation and Computation **15** (1986), no. 4, 1109–1136. Cited on page 57.

- [MHS⁺16] T. Martin, J. M. Hofman, A. Sharma, A. Anderson, and D. J. Watts, *Exploring limits to prediction in complex social systems*, Proceedings of the 25th International Conference on World Wide Web, International World Wide Web Conferences Steering Committee, 2016, pp. 683–694. Cited on page 92.
- [Mit04] M. Mitzenmacher, *A brief history of generative models for power law and lognormal distributions*, Internet mathematics **1** (2004), no. 2, 226–251. Cited on page 10.
- [MJG90] D. C. Montgomery, L. A. Johnson, and J. S. Gardiner, *Forecasting and time series analysis*, McGraw-Hill Companies, 1990. Cited on page 19.
- [MKA16] J. M. Miotto, H. Kantz, and E. G. Altmann, *Stochastic dynamics and the predictability of big hits in online videos*, in consideration, 2016. Cited on page 90.
- [MM15] A. Mollgaard and J. Mathiesen, *Emergent user behavior on Twitter modelled by a stochastic differential equation*, PLoS ONE **10** (2015), no. 5, e0123876. Cited on page 92.
- [MMR13] M. Marsili, I. Mastromatteo, and Y. Roudi, *On sampling and modeling complex systems*, J. Stat. Mech. **2013** (2013), no. 09, P09003. Cited on page 13.
- [MS95] R. N. Mantegna and H. E. Stanley, *Scaling behaviour in the dynamics of an economic index*, Nature **376** (1995), no. 6535, 46–49. Cited on page 52.
- [MSSVK08] T. Maillart, D. Sornette, S. Spaeth, and G. Von Krogh, *Empirical tests of Zipf’s law mechanism in open source linux distribution*, Phys. Rev. Lett. **101** (2008), no. 21, 218701. Cited on pages 17 and 92.
- [New01] M. E. Newman, *Clustering and preferential attachment in growing networks*, Phys. Rev. E **64** (2001), no. 2, 025102. Cited on pages 14 and 15.
- [New05] M. E. J. Newman, *Power laws, pareto distributions and Zipf’s law*, Contemporary physics **46** (2005), no. 5, 323–351. Cited on page 7.
- [Nol12] J. P. Nolan, *Stable distributions*, unpublished, 2012. Cited on pages 12, 97, and 102.
- [OED16] *Oxford english dictionary*, Available: <http://www.oed.com/view/Entry/152776>. Accessed 2016 Apr 10., 2016. Cited on page 13.
- [Øks03] B. Øksendal, *Stochastic differential equations*, Springer, 2003. Cited on page 49.
- [ORT10] J. Onnela and F. Reed-Tsochas, *Spontaneous emergence of social influence in online systems*, Proc. Natl. Acad. Sci. USA **107** (2010), no. 43, 18375–18380. Cited on page 41.

- [PAG13] H. Pinto, J. M. Almeida, and M. A. Gonçalves, *Using early view patterns to predict the popularity of youtube videos*, Proceedings of the sixth ACM international conference on Web search and data mining, ACM, 2013, pp. 365–374. Cited on page 70.
- [Par96] V. Pareto, *La courbe de la repartition de la richesse*, Universite de Lausanne, 1896. Cited on pages 1 and 7.
- [PDD13] J. Peterson, P. D. Dixit, and K. A. Dill, *A maximum entropy framework for nonexponential distributions*, Proc. Natl. Acad. Sci. USA **110** (2013), no. 51, 20380–20385. Cited on page 13.
- [Per14] M. Perc, *The Matthew effect in empirical data*, J. R. Soc. Interface **11** (2014), no. 98, 20140378. Cited on pages 10, 13, 14, 15, and 41.
- [PFK98] F. J. Provost, T. Fawcett, and R. Kohavi, *The case against accuracy estimation for comparing induction algorithms.*, ICML, vol. 98, 1998, pp. 445–453. Cited on page 32.
- [PI75] J. Pickands III, *Statistical inference using extreme order statistics*, Ann. Stat. (1975), 119–131. Cited on page 11.
- [PMS13] T. Preis, H. S. Moat, and H. E. Stanley, *Quantifying trading behavior in financial markets using Google Trends*, Sci. Rep. **3** (2013). Cited on page 41.
- [PPP⁺13] O. Penner, R. K. Pan, A. M. Petersen, K. Kaski, and S. Fortunato, *On the predictability of future impact in science*, Sci. Rep. **3** (2013), 3052. Cited on pages 10, 36, and 41.
- [Pri76] D. d. S. Price, *A general theory of bibliometric and other cumulative advantage processes*, J. Ame. So. Inf. Sci. Technol. **27** (1976), no. 5, 292–306. Cited on pages 9, 10, and 41.
- [Red98] S. Redner, *How popular is your paper? An empirical study of the citation distribution*, The European Physical Journal B-Condensed Matter and Complex Systems **4** (1998), no. 2, 131–134. Cited on page 1.
- [RFF⁺10] J. Ratkiewicz, S. Fortunato, A. Flammini, F. Menczer, and A. Vespignani, *Characterizing and modeling the dynamics of online popularity*, Phys. Rev. Lett. **105** (2010), no. 15, 158701. Cited on pages 9, 10, 41, and 92.
- [RK01] M. Ragwitz and H. Kantz, *Indispensable finite time corrections for Fokker-Planck equations from time series data*, Physical Review Letters **87** (2001), no. 25, 254501. Cited on page 57.

- [Ros14] S. M. Ross, *Introduction to probability models*, Academic press, 2014. Cited on page 51.
- [SA94] D. Stauffer and A. Aharony, *Introduction to percolation theory*, CRC press, 1994. Cited on pages 1 and 6.
- [SB94] S. Sukhatme and C. Beam, *Stratification in nonparametric ROC studies*, Biometrics (1994), 149–163. Cited on page 31.
- [SDGA04] D. Sornette, F. Deschâtres, T. Gilbert, and Y. Ageon, *Endogenous versus exogenous shocks in complex networks: An empirical test using book sale rankings*, Phys. Rev. Lett. **93** (2004), no. 22, 228701. Cited on page 10.
- [SDW06] M. J. Salganik, P. S. Dodds, and D. J. Watts, *Experimental study of inequality and unpredictability in an artificial cultural market*, Science **311** (2006), no. 5762, 854–856. Cited on pages 8 and 41.
- [Sim55] H. A. Simon, *On a class of skew distribution functions*, Biometrika **42** (1955), no. 3/4, 425–440. Cited on pages 13 and 14.
- [Sim71] ———, *Designing organizations for an information-rich world*, Computers, communication, and the public interest **37** (1971), 40–41. Cited on page 8.
- [Sor02] D. Sornette, *Predictability of catastrophic events: material rupture, earthquakes, turbulence, financial crashes, and human birth.*, Proc. Natl. Acad. Sci. USA **99** (2002), 2522–9. Cited on pages 29 and 41.
- [Sor06] ———, *Critical phenomena in natural sciences: chaos, fractals, selforganization and disorder: concepts and tools*, Springer Science & Business Media, 2006. Cited on page 6.
- [Sor09] ———, *Dragon-kings, black swans and the prediction of crises*, Swiss Finance Institute Research Paper (2009), no. 09-36. Cited on page 19.
- [Spe04] C. Spearman, *The proof and measurement of association between two things*, The American journal of psychology **15** (1904), no. 1, 72–101. Cited on pages 64 and 65.
- [SRS93] M. Sahimi, M. C. Robertson, and C. G. Sammis, *Fractal distribution of earthquake hypocenters and its relation to fault patterns and percolation*, Physical review letters **70** (1993), no. 14, 2186. Cited on page 6.
- [SSPA10] M. J. Stringer, M. Sales-Pardo, and L. A. N. Amaral, *Statistical validation of a global model for the distribution of the ultimate number of citations accrued by papers*

- published in a scientific journal*, J. Ame. So. Inf. Sci. Technol. **61** (2010), no. 7, 1377–1385. Cited on pages 10 and 41.
- [SWSB14] H. Shen, D. Wang, C. Song, and A. Barabási, *Modeling and predicting popularity dynamics via reinforced poisson processes*, Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence, AAAI, 2014, pp. 291–297. Cited on page 94.
- [SZF95] M. F. Shlesinger, G. M. Zaslavsky, and U. Frisch, *Lévy flights and related topics in physics*, Springer Verlag, 1995. Cited on page 67.
- [Tal07] N. N. Taleb, *The black swan: The impact of the highly improbable*, Random House, 2007. Cited on page 19.
- [VR13] W. N. Venables and B. D. Ripley, *Modern applied statistics with S-PLUS*, Springer Science & Business Media, 2013. Cited on page 30.
- [Wal31] G. Walker, *On periodicity in series of related terms*, Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character **131** (1931), no. 818, 518–532. Cited on page 71.
- [WFVM12] L. Weng, A. Flammini, A. Vespignani, and F. Menczer, *Competition among memes in a world with limited attention*, Sci. Rep. **2** (2012), 335. Cited on pages 10 and 41.
- [WH07a] F. Wu and B. A. Huberman, *Novelty and collective attention*, Proc. Natl. Acad. Sci. USA **104** (2007), no. 45, 17599–17601. Cited on page 9.
- [WH07b] ———, *Novelty and collective attention*, Proc. Natl. Acad. Sci. USA **104** (2007), no. 45, 17599–17601. Cited on pages 50 and 59.
- [WS98] D. Watts and S. Strogatz, *Collective dynamics of small-world networks*, Nature **393** (1998), 440–442. Cited on page 8.
- [WSB13] D. Wang, C. Song, and A. Barabási, *Quantifying long-term scientific impact*, Science **342** (2013), no. 6154, 127–132. Cited on pages 9, 10, 39, 41, and 94.
- [WY22] J. C. Willis and G. U. Yule, *Some statistics of evolution and geographical distribution in plants and animals, and their significance*, Nature **109** (1922), no. 2728, 177–179. Cited on page 7.
- [YHM13] T. Yasseri, S. A. Hale, and H. Magretts, *Modeling the rise in Internet-based petitions*, To be published (2013). Cited on page 10.
- [YK12] S. Yu and S. Kak, *A survey of prediction using social media*, arXiv preprint arXiv:1203.1647 (2012). Cited on page 18.

- [Yul25] G. U. Yule, *A mathematical theory of evolution, based on the conclusions of Dr. JC Willis, FRS*, Philos. Trans. R. Soc. Lond. B Biol. Sci. **213** (1925), 21–87. Cited on page 13.
- [Yul27] ———, *On a method of investigating periodicities in disturbed series, with special reference to Wolfer’s sunspot numbers*, Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character **226** (1927), 267–298. Cited on page 71.
- [ZDK15] V. Zaburdaev, S. Denisov, and J. Klafter, *Lévy walks*, Reviews of Modern Physics **87** (2015), no. 2, 483. Cited on page 67.
- [Zip36] G. K. Zipf, *The psycho-biology of language*, Routledge, 1936. Cited on page 7.
- [Zol86] V. M. Zolotarev, *One-dimensional stable distributions*, American Mathematical Soc., 1986. Cited on page 12.

Acknowledgments

First of all, I want to thank of *Eduardo G. Altmann* for his invaluable supervision. Not only he gave me the opportunity of working in the topic of social dynamics, which was what I was looking for, but he also supported, inspired and guided me in every moment, by continuously discussing the project and other scientific ideas, and by giving me the necessary confidence to carry them forward.

Likewise, I want to thank the rest of the Dynamical Systems & Social Dynamics group; it was a pleasure to share my time at the institute with them, from whom I could learn a lot of things, and with whom shared endless debates on pretty much everything. I found here not only a group of colleagues but also a group of friends.

Moreover, I want to thank my collaborators *Holger Kantz*, who was involved in part of the results presented in this Thesis, and *Martin Gerlach*, *Jorge C. Leitão*, and *Fakhteh Ghanbarnejad*, with whom I worked in different projects.

I acknowledge as well the *Max Planck Institute for the Physics of Complex Systems* for its working conditions, stimulating environment, and very professional staff.

Finally, I want to specially thank my wife *Luz* for her help in all the stages of my doctoral work. The ways in which he supported me are numerous (including moving with me to Dresden), and I could not be grateful enough with her for all the things she does for me, and for having her on my side.

Versicherung

Diese Arbeit wurde am Max-Planck-Institut für Physik komplexer Systeme unter der wissenschaftlichen Betreuung von Dr. Habil. Eduardo Goldani Altmann durchgeführt.

Hiermit versichere ich, dass ich die vorliegende Arbeit ohne unzulässige Hilfe Dritter und ohne Benutzung anderer als der angegebenen Hilfsmittel angefertigt habe; die aus fremden Quellen direkt oder indirekt übernommenen Gedanken sind als solche kenntlich gemacht. Die Arbeit wurde bisher weder im Inland noch im Ausland in gleicher oder ähnlicher Form einer anderen Prüfungsbehörde vorgelegt.

Darüber hinaus erkenne ich die Promotionsordnung der Fakultät Mathematik und Naturwissenschaften der Technischen Universität Dresden vom 23. Februar 2011 an.

José María Miotto

Dresden, 18 Mai. 2016