Università degli Studi di Padova

Dipartimento di Fisica e Astronomia "G. Galilei"
Corso di laurea in Fisica

Tesi di Laurea

Rank-based Markov chains and self-organized criticality for written communication

Relatore:
Prof. Amos Maritan
Correlatore
Dott. Marco Formentin

Laureando:
Stefano Garlaschi
Matricola: 1069183

Anno Accademico 2015/2016

Università degli Studi di Padova

Dipartimento di Fisica e Astronomia "G. Galilei"
Corso di laurea in Fisica

Tesi di Laurea

Rank-based Markov chains and self-organized criticality for written communication

Relatore:
Prof. Amos Maritan: ....
Correlatore:
Dott. Marco Formentin: ....

Laureando:
Stefano Garlaschi: ....
Matricola: 1069183

Contents

Introduction

To study complex phenomena such as animal and human dynamics it is very helpful the discovery of some hidden universal statistical patterns, removing in this way the dependences from the behavior of single interacting agents, which may be different and very variable from an individual to another. For this reason recently these complex dynamics are studied to find recurrent regularities or common statistical traits. One of the most studied human dynamics and central topic of this work is written communication. The interest in this dynamics is also originated by the large utility of written communication since it allows the human kind to keep in contact with distant people.
So a lot of efforts have been spent in statistical studies of empirical data and in construction and theorical analysis of models in order to find and understand how the universal statistical patterns, that in the case of written corrispondence assume the form of probability distribution of the time interval elapsed between the arrival of a task inbox and its execution, originate.
This thesis tries to be in line with this kind of studies for the chosen topic. In Chapter 1 we will present the collected data used in [1], we will give the definition of the quantities of interest for written communication (inter-event time and response time) and we will introduce the concept of the re-clocking of the time. Then in Chapter 2 we will discuss the consequences of the re-clocking, the rise of the universality in the response time statistics and we will provide a possible explanation of the mechanism that allows this universality to rise. In particular we will present a queuing model with a priority driven dynamics (i.e. the evolution of the system in this model can be seen as a rank-based Markovian chain) with which we try to reproduce the receiveing and answering process of an agent. The features of the model - the existence of a threshold priority above with all the dynamics takes place and the scaling exponent of the power law of the response times probability distribution - that we will obtain in a rigorous way in Chapter 4 also supported by simulations, led us to consider our model as a model that makes the system displaying self-organized criticality. So in Chapter 3 we will deal with self-organized criticality exposing the main points of this theory and we will show queuing model similar to our applied to different fields that display self-organized criticality in order to make evident this relation (priority driven dynamics and self-organized criticality). In Chapter 5 we will describe the first rank-based process trying to explain the written corrispondence, the Barabási model. Although it received several criticism, we will report it for its importance, since it is the first priority-queuing model created for the considered problem.

Chapter 1

The problem of written communication: data and some useful definitions

The aim of this work is to study the main features of an interesting human dynamics such as the written corrispondence. To do this we study the behavior of agents across three different communication media: letters, emails and text messages (sms). To better understand the purpose of the study, we must first show the data used to obtain empirical results in [1] and provide some definitions used in the following.

1.1 Data

Data sets are analyzed in order to find some regularities or some statistical universalities, as said in the Introduction. The written communication can happen in different way, through paper corripondence (i.e letters) or in electronic form (i.e. emails). Recently also the short-text messages (sms) have been studied since it have become an intensly used communication medium.
For the letters the study analyzes the written corrispondence data contained in a database in which are also collected available epistolaries of three famous writers across all their life-time

- Charles Darwin;
- Albert Einstein;
- Sigmund Freud.

As the emails are the most corrent used medium, it is relatively easy to collect a significant amount of data and in the study three different databases are analyzed:

- the first is recently collected and it contains the activities of accounts belonging to, and interacting with, a Department of a large European university over about two years. This database is used by almost 400 active users, each with $\sim 10^3 - 10^4$ emails, making it very interesting from a statistical point of view.
- the second is recently colleced too, but it contains only the email activity of three agents over a long period (from five to nine years).

- the third is a database covering a period of about three months which comprises email data of an European university.

The three considered databases cover periods of different length, this allows us to discern if the statistic of the considered problem change with the period of time which is analyzed.
Finally for the text messages (sms) the data are obtained from a database extending over a one-month period with $\sim 10^2 - 10^3$ messages of the accounts belonging to three Chinese companies.
For all the three media the data are in the form $\{$*sender, receiver, timestamp*$\}$, where time is measured in days for letters and in seconds for email and text messages. For more details on data see [1].

1.2 Definitions

For any agent A of a generic medium it is possible to identify two distinct waiting times, which characterize his/her behavior:

- the inter-event times (IETs) are the time intervals $\tau := \Delta t$ elapsed between two consecutive activities of A, i.e. the time intervals which separate two consecutive written communications of the considered agent;
- the response times (RTs) are the time intervals $\tau = \Delta t$ which separate the arrival in the A's inbox of a message M coming from another agent B and the first message *M'* sent by A to the B's inbox, independently of the contents of the two messages M and *M'*.

For the IETs and the RTs we indicate their probability distributions with $P_I(\tau)$ and $P_R(\tau)$ respectively, when these are clocked using the standard time.
In order to study the response times statistics, we execute a re-clocking of the time [1]: instead of measuring it in seconds for the emails and sms and days for the letters, now the time is clocked through activities.
To do this, let us introduce the parameter s with which we count the number of A's sent messages. After the re-clocking the time interval between the arrival in the A's inbox of a message from B and the A's answer to it becomes the number s of messages sent by A between the two latter events, as we can clearly understand from figure 1.1. It is possible to think at the parameter s as a measure of the agent's written communication activity or, in other words, as the 'proper time' of the agent: in fact with this clocking we count only the activities regardless of the period of time elapsed between the sending of two consecutive messages by A.
The aim of this re-clocking is to study the probability distributions of RTs and IETs using $\sigma = \Delta s$ instead of $\tau = \Delta t$. We indicate the re-clocked probability distributions with $\overline{P}_R(\sigma)$ and $\overline{P}_I(\sigma)$. The latter is, by the definition of s, trivially the same for all agents indipendently from the used medium, since any A's IET increases by one this parameter, so it is evident that $\overline{P}_I(\sigma)$ is concentrated at $\sigma = 1$ assuming the form of a Dirac's delta function centered in 1, $\overline{P}_I(\sigma) \equiv \delta_1$.
With these definitions we want to study the RTs statistics in order to obtain the main features of the underlying dynamics which rules the answering process.

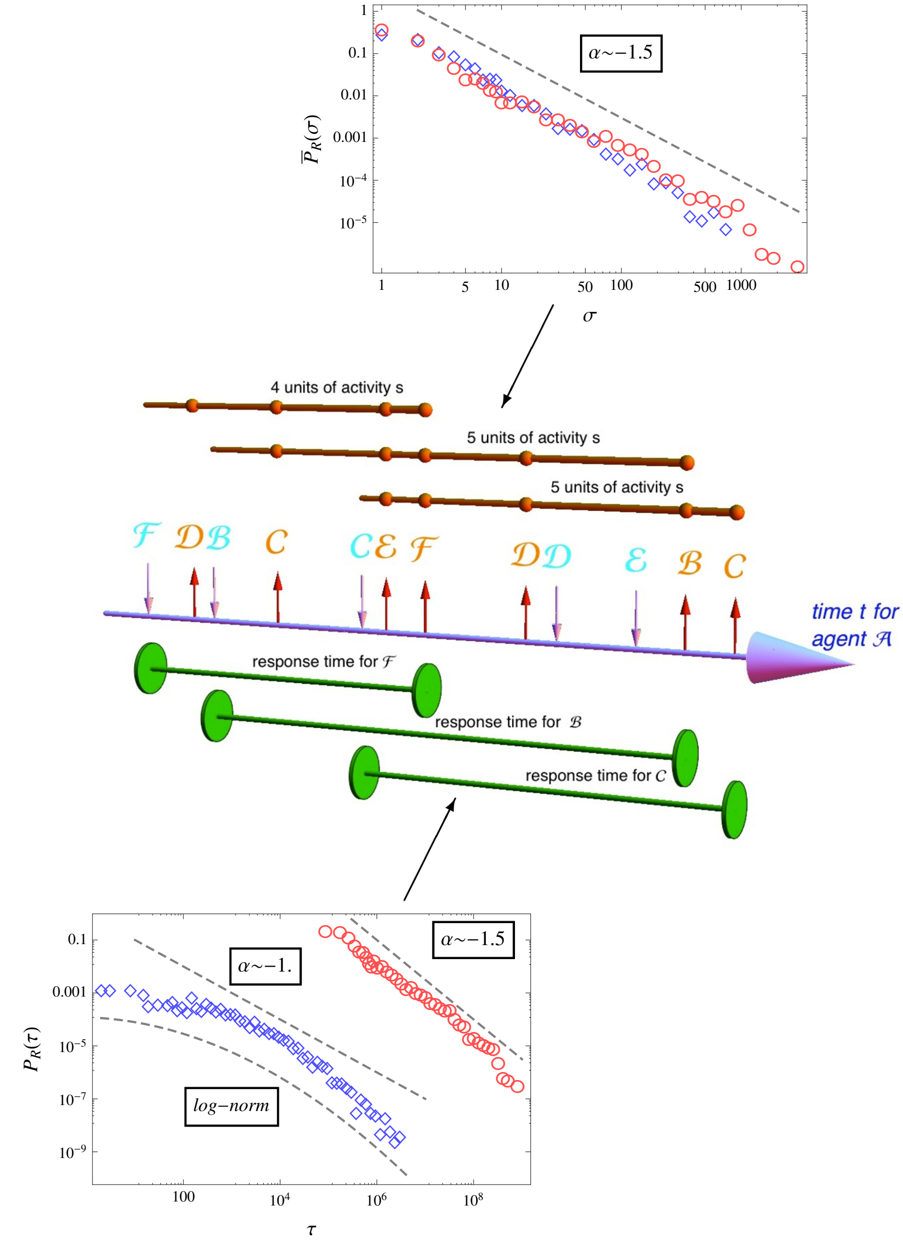

Figure 1.1: The figure shows two possible clocking for written correspondence. For an agent A the communication activity is represented along the axis of time t. Arrows pointing into the axis t symbolize incoming messages from the indicated agents B, C, ..., while the arrows pointing out the axis t symbolize response messages to the same agents. The inter-event times (IETs) of agent A are represented by the time intervals between such arrows. The response times (RTs) of A are defined as shown, either clocked through time t, or through the activity parameter s which counts the number of outgoing messages from A. The two graphics show the RT distributions clocked through standard time and activities of two agents communicating through two different communication media: the blue squares for the one using emails, the red circles for the one using letters. While the two $P_R(\tau)$ are different, the RT distributions clocked through σ assume a common trend.

Chapter 2

Universality after the re-clocking

The re-clocking has an important consequence: not only the IETs present an universal probability distribution as mentioned above, but the RTs too, as we can see from figure 1.1. In fact across all the media in the large majority of agents across all three media the empirical $\overline{P}_R(\sigma)$ are very well fitted by discrete exponentially-truncated power-laws such as:

$$\overline{P}_R(\sigma) \sim \sigma^{\alpha} e^{-\frac{\sigma}{\lambda}}$$

where α is the scaling exponent and λ is the cutoff parameter [2].

To estimate their values it was used the maximum likelihood method: for all media and for all analyzed agents the scaling exponent α has average value close to $-\frac{3}{2}$, showing an hidden common trait of the $\overline{P}_R(\sigma)$. To confirm the found expression of the empirical distribution $\overline{P}_R(\sigma)$ the log-likelihood ratio test is performed on a subset of randomly selected agents and it shows that other distributions do not describe the data as a truncated power-law distribution does.

This fact suggests that all written communications have an intrinsic universal feature. This is partly obfuscated because the two waiting time distributions change across agents and media when the time is clocked through standard time t due to the interactions with the spontaneous IETs, which are media- and agent-dependent. But when the time is expressed in terms of proper time s, this disturb is eliminated and so the two distributions assume a common form for all agents and media, showing the universal behaviour of this human dynamics.

In order to understand how the IETs hide the common form of the probability distribution of the RTs when it is clocked through standard time, we must understand the relation between the probability distribution of the RTs clocked through t and s.

Assuming that the distribution $P_I(\tau)$ of an agent A is known, we can obtain the relation between the two RT distributions $\overline{P}_R(\sigma)$ and $P_R(\tau)$ in a simple way. Let us consider indipendent random variables N and $\rho_I(h)$, $h = 1, 2, \ldots, N$, where N is sampled from $\overline{P}_R(\sigma)$ and it represents the number of activities between the arrival of a message and the response to it, while $\rho_I(h)$ are inter-event times sampled from the given $P_I(\tau)$. So the response time clocked through the standard time is obtained by considering it as the sum of the N inter-event times $\rho_I(h)$ between the N activities which the agent A executes before answering to the considered message, i.e.:

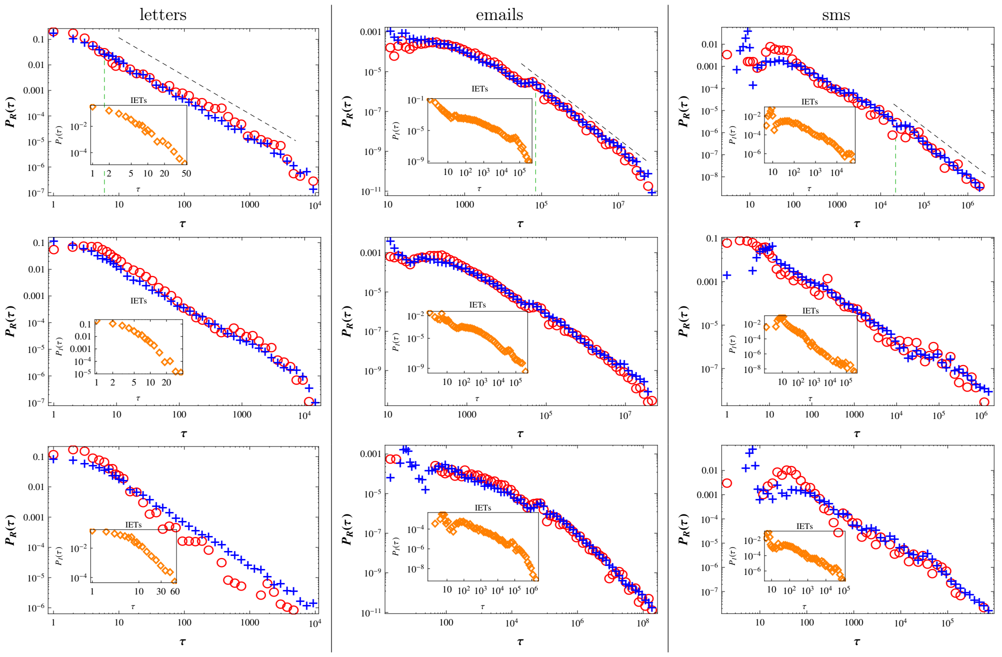

Figure 2.1: Log-log plots of empirical $P_R(\tau)$ clocked through standard time t, measured in days for letters and seconds for emails and sms, for three agents for each different medium. The red circles represent empirical data, the blue crosses represent the computational predictions, while the little box shows the empirical $P_I(\tau)$. For letters the empirical data come from the avaible epistolary corrispondence of (starting from the top) C. Darwin, A. Einstein and S. Freud. The agents considered for the emails are three coming from the two long term databeses and the data for sms are the ones of typical agents in the database collecting data of three Chinese companies. As we can see, for all media and all considered agents we have a quite good agreement between reality and predictions.

$$\rho_R = \sum_{h=1}^{N} \rho_I(h).$$

Since ρ_R is obtained through this compounding process, its probability to be equal to a given time interval τ, which is the probability distribution that we are interested in ($P_R(\tau)$ =Prob($\rho_R = \tau$) by definition), has to be computed by conditioning the probability to have σ activities (between the arrival and the response to the message) and the probability that the sum of σ IETs $\rho_I(h)$ is equal to τ, summed over all the possible value of σ, or:

$$P_R(\tau) = \sum_{\sigma \geq 1} \text{Prob}\left(\sum_{h=1}^{\sigma} \rho_I(h) = \tau \right) \overline{P}_R(\sigma)$$

Numerical simulations confirm the found relation between the empirical distributions $P_R(\tau)$, $P_I(\tau)$ and $\overline{P}_R(\sigma)$ in all media [1] as shown in figure 2.1.
Expressing the $P_R(\tau)$ in this form, it is more evident the reason why when clocked through the standard time the RTs probability distribution loses the universal feature that becomes evident using the activity re-clocking. In fact now we see that this distribution is tied to the spontaneous IETs, which may change between a medium and another and between an agent and another, but when we use the

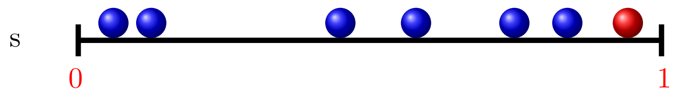

Figure 2.2: A scheme of the queue of unexecuted task in the inbox at a certain time step s: the balls represent tasks in the queue ordered by increasing priority. The red one represents the task with the largest priority, i.e. the task executed at the next time step $s + 1$.

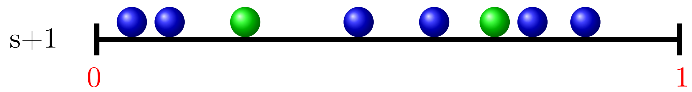

Figure 2.3: A scheme of the queue of unexecuted task in the inbox at the next time step $s + 1$: the $m_{in} = 2$ green balls represent the new tasks just added to the queue. Obviously one of the two new tasks can be the task joining the queue with the largest priority.

parameter s this contribution to the RTs statistic is eliminated because, as we said above, $\overline{P}_I(\sigma)$ assumes an unique form and so the intrinsic behavior of $\overline{P}_R(\sigma)$ can arise.
At this point we introduce a simple queue model capable to describe the statistical patterns in the empirical data we discussed so far. Here we only describe the model and how we use it to fit the data. Its theorical analysis is postponed to Chapter 4.
To modelize the dynamics, since we are interested in the universal pattern of the written communication, i.e in the re-clocked $\overline{P}_R(\sigma)$, we consider discrete time steps. This is done in order to replicate the proper time s.
In our model we assign to each incoming message a number called *priority*. This is an attempt to reproduce the behavior of the average user. In fact not all the messages received have the same importance or require the same efforts to be answered. The priority contains a balance of this two contributions and it tries to quantify and order the answering process. Usually between two consecutive activities, an agent adds m_{in} new tasks to execute, where m_{in} can change step by step. In the model we assign the priority to the incoming message once forever when it is added to the queue sampling a number from a distribution (we use the uniform distribution between 0 and 1).
Now we have all the elements to describe the model dynamics: we start with a given number of tasks in the queue with their priorities. At the first suitable step we response to the message with the largest priority and we remove it from the inbox. Then m_{in} new tasks with indipendent priority join the queue waiting to be executed. This process of answering and receiving goes on until we want to conclude the system dynamics. Figures 2.2 and 2.3 explain this priority driven dynamics reproducing the inbox.
Using this model we have executed simulations to verify its reliability. The results are shown in figure 2.4.

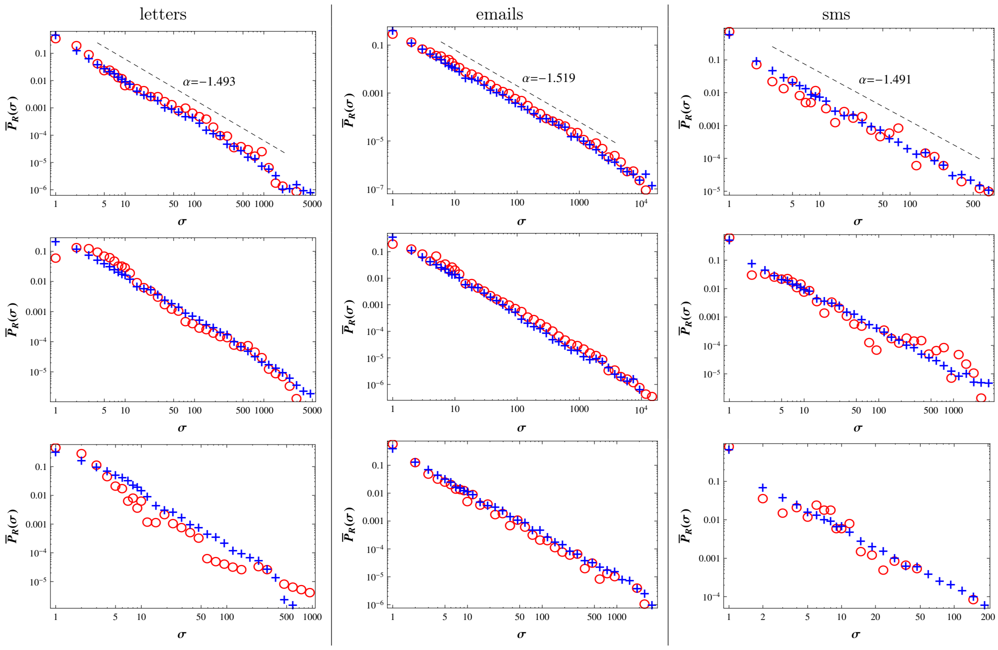

Figure 2.4: Log-log plots of $\overline{P}_R(\sigma)$ for three agents for the three media: letters, emails and text messages. The red circles represent empirical data, the blue crosses represent the predictions of model. This is used in the same way across the three media with the only difference that the number m_{in} of the incoming messages is derived step by step considering the empirical distributions of incoming messages between any two consecutive outgoing messages of each user. The empirical data are relative to the same typical agents as in Figure 2.1. We see that the model is able to reproduce quite well the empirical distributions, which are very well fitted by truncated power-laws with scaling exponent α near to -1.5 in all the considered agents and media.

Chapter 3

Intermezzo: Self-Organized Criticality

The re-clocked RTs probability distribution is described by a power-law for all media and for almost all agents, i.e. the considered dynamics system organizes on its own in a state in which the probability follows a power-law trend, indipendently from all accidental interferences like how much a agent answers to the messages received or the proper time of the medium.
In the described model the answering process is ruled by a priority mechanism: we execute first the tasks with larger priority. For this reason we can see our priority-queueing model dynamics as a Markovian dynamics, since the inbox state at the step $s + 1$ depends only on the inbox state at the step s. We will see that in the inbox we have a threshold, a particular value for the priority under which the tasks can not be executed, i.e. they pile up and join the queue forever. The task execution can take place only above the threshold. This is the main point of the dynamics: all the dynamics happens above the threshold, since only here the messages can be executed. Now the number of these tasks follows a recurrent random walk, i.e. the number evolves stochastically and it can move from a certain value to others with determinated probabilities. After a while the number of tasks above the threshold we will be zero again, i.e this number is a recurrent random walk that always returns to zero. The power-law comes from this feature of the model dynamics: the time of return to zero for a random walk follows a power-law. Thus the system is able to reach on its own a state when the RTs obey to a power-law statistic. In analogy with physics of phase transition (see below for details) we call this state critical.
Such a dynamical mechanism -a recurrent random walk above the threshold- to generate power law statistic is shared with other models that show self-organized criticality. This leads us to consider our modeling framework from the point of view of the *Self-Organized Criticality*. We introduce the concept of self-organized criticality, its characteristics and we try to explain how priority mechanism (i.e queueing models) and the existance of a threshold is bound to SOC showing some models similar to ours but applied to study very different problems.

3.1 Self-Organized Criticality

Self-organized criticality (SOC) is a theory developed in the last couples of decades and it tries to describe very rich phenonemona since it put togheter criticality and self-organization to study their complex behavior.
Self-organized criticality refers to the tendancy of large dynamical systems to organize into a poised state far out of the equilibrium. In this state the system is not chaotic, but phenomena may occur following well defined statistical laws and so we can describe their dynamics using this laws. Another charateristic of this state is that the dynamics of the system develops itself with propagating avalanches of activity of all sizes with defined statistical laws (as we will see in the Bak-Sneppen model taken as example).
To explain what self-organized criticality is, we must explain before the meaning of the key-words *Criticality* and *Self-Organized*.
In physics we speak about criticality for systems which display power-law trend for the quantities of interest at the phase transition point, also called the critical point which marks the transition from a regime to another with different features or behaviors. So we can say that power-laws are the natural mathematical expressions of critical systems. In analogy with physics a generic system/model is said to be critical if the quantities of interest present power-law decay. These power-laws are the statistical laws mentioned above.
Physical systems to be critical have to be tuned, i.e. the control parameters which rule the transition of the system at its critical state have to assume the right value. In SOC the transition to the critical state is fundamentally different as the system organizes on its own and it reaches the critical point without external tuning. This charateristic is summed up in the key-word *Self-Organized*.
So we can say that the systems displayng SOC do not need external tuning of the control parameters, but evolve on its own toward the critical state showing power-law trend for quantities of interest.
This theory was born in the 1987 from Per Bak, Chao Tang and Kurt Wiesenfeld's paper *Self-organized criticality: an explanation of 1/f noise* [3]. It is very powerful since it can be considered as a method with which the complexity of nature can arise (complexity is to meant as there is not any scale to guide the dynamics and the evolution of the considered system).
Within SOC models it is worth to mention the so called *Sandpile model* [4], built up by Bak, Tang and Wiesenfeld as the archetype of this new theory. Let us consider an initial sandpile formed on a horizontal circular base. Adding sand grains we obtain a sandpile of conical shape (which is the steady state of the sandpile). Keeping on adding sand grains, we will see a propagating avalanche of sand on the surface of the pile. In the stationary regime, avalanches are of many different sizes and the authors of the model argued that durations and the sizes of this avalanches would have a power-law distribution. If one starts with an initial uncritical state, initially most of the avalanches are small, but the range of sizes of avalanches grows with time and reaching the critical state of the system, the avalanches can have all sizes and the probability distribution of these assume the form of a power-law.
Some famous model in SOC has a priority driven dynamics that display a threshold and the dynamics in the critical state follows a random walk. We report two examples where priority-queueing arguments and Markovian dynamics are applied to evolutionary biology and economics as SOC examples. This interdisciplinarity

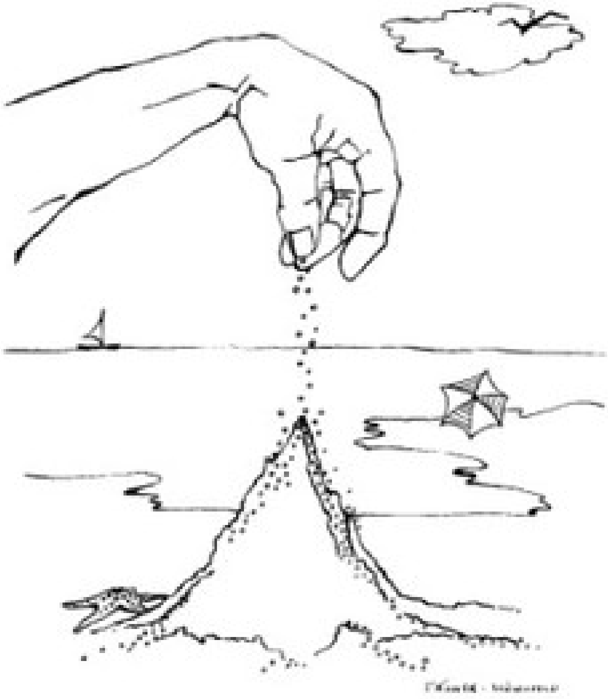

Figure 3.1: Illustration taken from Bak's book *"How nature works"* [4] of the sandpile used to explain the main features of *Self-Organized Criticality.*

can be seen as one of the reasons of the SOC theory success.

3.1.1 Bak-Sneppen model for evolutionary dynamics

With this model, Bak and Sneppen tried to explain the main feaures of the evolutioning process realized by natural selection of the species formulated by Charles Darwin [5]. The main idea is to consider an ecosystem in which N given species live in a cooperative scenario, i.e. each species is not independent from the others and so the environment of any given species is affected by other evolving ones. We assign to each of them a fitness, which is a quantity that tries to measure the capacity of adaptation to the environment and the ability to survive in the habitat. This is of course an oversemplification since we look at a species entirely despite the fact that selection happens on the scale of single individuals. Also there is not an universal formula to compute this parameter, but it can be always choose with a certain degree of arbitrariness. This, however, does not cancel the interesting feature of the model, since this does not want to reproduce the reality, but is aim is to replicate the main mechanisms that rule the process.
Since in nature the most adapted species have smaller probability to change itself than the least adapted one, to take account of the natural selection implemented by the ecosystem, the latter is replaced by a new species with a new fitness (this can be seen as a "macroscopic" mutation of all individuals of that species). In the environment there are interspecies relations, so in order to take account of these relations when a species changes its fitness (by mutating), also the fitnesses of the other species with which it is in relation must change. For example if a species, which is the main prey of another one at the top of the food chain, increase its fitness by mutating, then the adaptation abilities of the other species are now

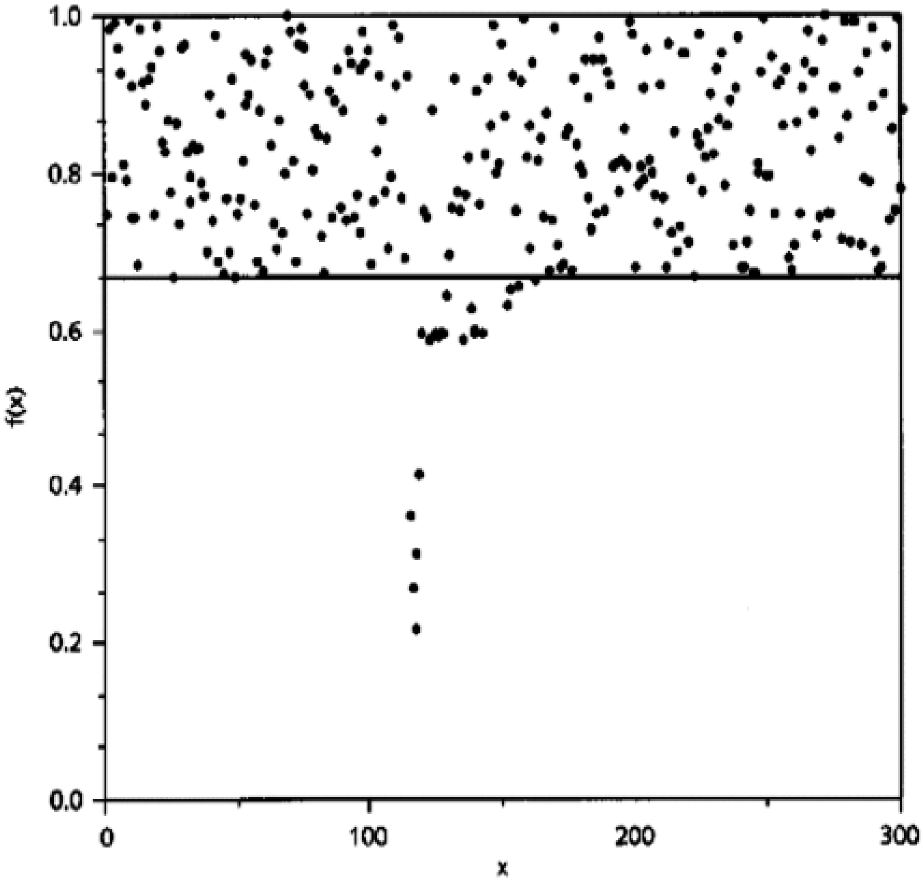

Figure 3.2: Snapshot of the fitnesses in the critical state in Bak-Sneppen model. Almost all the fitnesses in the system have values above f_c. The points below the critical threshold represent an avalanche, as described in text.

different, so its fitness must be changed.
In the model the fitness is assigned by generated a random number sampled from the uniform distribution between 0 and 1. Now we consider the species placed on a ring with increasing fitness (we use periodic condition in the "fitness space").
To simulate the evolution dynamics, at each time step the lowest fitted species is selected and its fitness is replaced with a new randomly assigned one. Notice that the new fitness is randomly generated, so this mutation can bring to an increase or a decrease of the fitness. A finest model can implement a test to evaluate if a trial evolution move increase the fitness, otherwise the simulation reject this trial move.
To take account of the interspecies relations also the fitnesses of the two nearest neighbors of the original species are replaced by new randomly generated ones.
For large N after a long initial transitient, the ecological system organizes on its own in a state where all mutations turns out to take place for fitness below a critical value $f_c \simeq 0.6672$ and the fitnesses are approximately uniform on $(f_c, 1]$ as shown in the figure 3.2.
When there is a fitness under f_c we say that an avalanche (of evolutioning activities) starts, i.e. the system keeps on changing the fitnesses in the discussed way since all of them are above the critical fitness. We measure the size of the avalanche as the number of subsequent mutations below f_c starting at the moment

that all fitnesses, except one, are above the critical fitness and stopping counting at the first next moment when all fitnesses are above f_c. The distribution of the avalanches show a power-law, i.e. if we plot the logarithm of the size versus the logarithm of the probability of finding an avalanche with that size we will obtain a straight line.
These results suggest us that the system displays self-organized criticality, since we obtain power-law behaviour without any external tuning of parameters. In this case SOC can be not mathematically demonstrate, as instead it can be done if we consider a modified version of Bak-Sneppen model [6].
Let us start with N given species, each with its fitness assigned as in the original Bak-Sneppen model. The dynamics of the ecological system in this case is a little bit different: at each time step, the species with the lowest fitness is chosen, togheter with one other, chosen randomly among all the remainings, then we replace their fitnesses with new ones as in the original model.
Now the critical value f_c is exactly equal to $\frac{1}{2}$ and coupling the dynamics of the simulation to a branching process it is possible to demonstrate that the time intervals σ (clocked through activities) between the first moment at which we have species with fitness smaller than f_c and the first moment at which all the species have fitness above the critical value have a power-law distribution:

$$P\big[\sigma > k\big] \sim k^{-\frac{1}{2}}$$

This means that we can have avalanches of all sizes, but the probability displays a power-law decay.
This result does not demonstrate that in the original model we have quantities showing power-law, but at least, only changing the way in which we can interpret the interaction between species, it is possible to obtain that self-organized criticality is tied at the principle of the evolution model invented by Bak and Sneppen.
Then we can say that SOC is able to reproduce the "punctuated equilibrium", which is a characteristic behavior of natural history. With "punctuated equilibrium" we mean the alternation of periods of stasis (in Bak-Sneppen model these periods are given by intervals in which all the fitnesses are above f_c, so we can see the environment as in balance since we have not great significant changes) with avalanches of casually connnected evolutionary activities. So we can also say that, looking at the "punctuated equilibrium" as a self-organized critical phenomenon, evolutionary activities do not require external causes, but since the sizes of avalanches present a power-law decay (i.e. all sizes are theorically possible), extinctions of all sizes, including the largest ones, may be seen as a simple consequence of the ecological dynamics only, without the need of natural disasters.

3.1.2 Stigler model for market dynamics

This model tries to simulate a market in which an asset is traded [7]. To simulate it, in the market there must be an active competition between buyers and sellers. The model deals with the so called "continous double auction" market (CDA), which is a continously operating market for standardized goods in which buyers and sellers can announce offers to trade specified quantities at a specified prices and can also accept such offers.

To trade the given asset, orders of buying or selling are registered in the so called *order book*, which is a list containing offers to buy or sell a specific quantity of the good at a specific price.
The trading orders can be in two form, choosen by the trader at the moment of sending it to the order book. They can be in the form of:

- limit order, i.e. order to buy a specified quantity n of the asset at a price not exceeding some specified maximum, or to sell a specified quantity n of the asset at a price not less than a specified minimum.

- market order, i.e. order to buy or sell a specified quantity n of the asset at the best current price.

The fondamental difference between these two types of trading orders is the fact that market orders are always executed immediately, while limit orders may join the queue of unexecuted orders if they can not be matched with any opposing orders already registered in the order book.
However this model deals only with limit orders: in this model the order book can be represented with an axis of prices going from 0 to ∞. On this axis we can signed in sell or buy orders:

- *sell order* or *ask* at the price α is a commitment to sell at the first opportunity at a price not less than α;

- *buy order* or *bid* at the price β is a commitment to buy at the first opportunity at a price not larger than β.

They are executable at a given price $x \in (0, \infty)$ if $\alpha \leq x \leq \beta$.
We can characterized the model market with the following properties:

- all orders are for single unit. So buyer or seller must specify only the maximum or minimum price that he/she is willing to pay/accept;

- there are large numbers of potential buyers or sellers acting independetly of another and each of them submit only occasionally an order to exchange. So we can say that traders arrive at the market at Poissonian times;

- the supply and demand functions are time-independent during the period of interest;

- traders submit their orders without making use of detailed information about the current state of the order book.

These assumptions allow us to say that traders can not acquire or infer new price-sensitive information during the period of interest, so market partecipants have no reason to revise the price of the order submitted.
The dynamics of the model moves through discrete time steps, so (using the same notation of the re-clocking) at any time s the order book will contain a queue of unexecuted asks at prices $\alpha_1(s), \alpha_2(s), \ldots$ and a queue of unexecuted bids at prices $\beta_1(s), \beta_2(s), \ldots$ waiting to be matched with the incoming orders. With this index notation we mean to order the orders in this way:

$$\ldots \beta_3(s) \leq \beta_2(s) \leq \beta_1(s) < \alpha_1(s) \leq \alpha_2(s) \leq \alpha_3(s) \ldots,$$

so α_1 denotes the lowest ask price (i.e. the best one) and β_1 denotes the highest bid price (i.e. the best one). Following the dynamics of the market, we are interested in the best ask and bid prices registered in the order book at a given time step s, i.e. $\alpha_1(s)$ and $\beta_1(s)$. But it is also necessary to monitor all the other orders because they may become the best orders of their queue.
When a new order is stored at the time s, the contents of the order book is updated as following determining the dynamics of the market model.
If the new order is an ask at the price α, then

- if $\alpha \leq \beta_1(s)$, the new ask is matched with the best current bid. So they are both executed at the bid price $\beta_1(s)$.
- if $\alpha > \beta_1(s)$, we can not find any possible match and so the new order joins the unexecuted asks queue.

If instead the new submitted order is a bid at the price β, then

- if $\beta \geq \alpha_1(s)$, the new bid is matched with the best current ask. So they are both executed at the ask price $\alpha_1(s)$.
- if $\beta < \alpha_1(s)$, we can not find any possible match and so the new order joins the unexecuted bids queue.

Let us consider the steady state of the order book. The transitions are determined by the best sell and buy orders (i.e. the offers with prices α_1 and β_1). To study their distributions we define $A(x)$ as the steady-state probability that the order book contains at least one suitable ask at price x. In a similar way we define $B(x)$ as the probability that the order book contains at least one suitable bid at price x. As consequences of these definitions of $A(x)$ and $B(x)$ we have:

$$A(0) = 0, \qquad A(\infty) = 1, \qquad B(0) = 1, \qquad B(\infty) = 0.$$

A must be non-decreasing and right-continuous, while B must be non-increasing and left-continuous.
Since we are interested in the best sell and buy offers, we can look at $A(x)$ as the probability that the best ask price α_1 is smaller than x and at $B(x)$ as the probability that the best bid price β_1 is larger than x:

$$A(x) = P[\alpha_1 \leq x], \qquad B(x) = P[\beta_1 \geq x].$$

Since the possibilities of having $\alpha_1 \leq x$ and $\beta_1 \geq x$ are mutually exclusive:

$$0 \leq A(x) + B(x) \leq 1 \qquad \forall x \in [0, \infty].$$

Now let us define two particular values (the critical ones) of price:

$$x_{min} \equiv \inf\{x \in [0, \infty] : B(x) < 1\}$$

$$x_{max} \equiv \sup\{x \in [0, \infty] : A(x) < 1\}$$

By the left- and right-continuity of $B(x)$ and $A(x)$ we can say:

$$B(x_{min}) = 1 \qquad A(x_{max}) = 1.$$

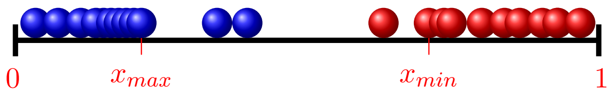

Figure 3.3: The blue balls represent buy orders, while the red ones sell orders. We can see that under x_{min} we have an overconcentrations of unexecuted buy orders and above x_{max} we have an overconcetration of unexecuted sell orders. In this figure the prices are normalized so we consider prices $x \in [0, 1]$.

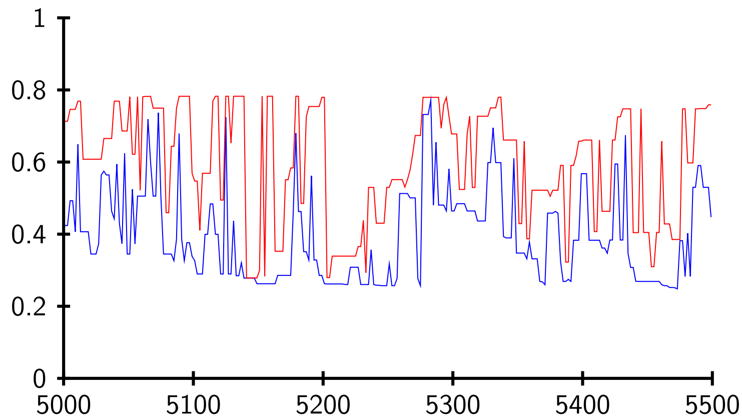

Figure 3.4: Evolution of the bid red and ask blue prices between the arrivals of the 5000th and 5500th trader obtained from a numerical simulation of the model. As we can see, the prices are always bound between two values.

Since $A(x) + B(x) \leq 1 \ \forall x \in [0, \infty]$, we can also say:

$$A(x_{min}) = 0 \qquad B(x_{max}) = 0.$$

Moreover, since A is non-decreasing and $A(x_{min} < A(x_{max})$ we have:

$$x_{min} < x_{max}.$$

For these reasons the best ask price α_1 will always exceed x_{min} but never exceed x_{max} and the best bid price β_1 will always exceed x_{min} but never exceed x_{min}:

$$x_{min} < \alpha_1 \leq x_{max}, \qquad x_{min} \leq \beta_1 < x_{max}.$$

Since every transition takes place at either the best ask price α_1 or the best bid price β_1, so almost all the transitions prices are contained in the price interval (x_{min}, x_{max}). For this reason we call this interval the competitive window, in which all the market activities are bound, while sell orders with price above x_{max} and buy offers under x_{min} remain indefinitely in the order book and never executed.
Obviously this model and its dynamics are oversimplifications of the real situation, but these can allow us to obtain interesting results like the fact that all transitions take place in a restricted window of prices, against the classical economic which predicts the achievement of equilibrium of prices between supply (sell orders) and demand (buy orders) as we can see in the figures 3.3 and 3.4.

Chapter 4

Model and predictions

In the first two chapters we discussed the main features of the considered problem and we built up a model that now we use to obtain these features in a rigorous mathematical way also supported by a simulation of the dynamics of the central problem of this work.
Using this model we can demonstrate the existence of a self-organized critical state and we are also able to predict the theorical scaling exponent of the power-law. In the following we consider for simplicity m_{in} constant and equal to 2.

4.1 Critical priority

In this section we demonstrate that the dynamics organizes on its own into a particular state in which there is a priority threshold. From this point all the dynamics involves only tasks with priority larger than the threshold. The number of tasks above the priority threshold evolves in time as a recurrent random walk. Then it will follow from random walk arguments that RTs statistics show a power-law behavior with scaling exponent equal to $-\frac{3}{2}$.
Starting from a list and fixing a certain value of priority x, we study the evolution of the number of tasks with priority larger than this value at a certain step s. Then we compute the expectation value for the difference of this number at two consecutive steps. At the steady state of the queue the number of tasks above the fixed value does not change, i.e the expectation value of the difference is equal to zero. In the end we solve the equation of the computed expectation value using as fixed priority the value for which we can be sure to have at least one task with priority larger than this. Since at least one priority is above the found value, we will always answer to one of these messages and so this value represents the threshold above which all the dynamics takes place.
Let us suppose to have a list of activities, each with a priority between 0 and 1. At time s let us call the list $Y(s) = \{Y_1(s), Y_2(s), \dots, Y_i(s), \dots\}$, where the Y_i are in order with decreasing priority.
For a given priority x, let us indicate with $N_x(s)$ the number of activities with priority bigger than x at the time step s and define $\Delta N_x(s) := N_x(s+1) - N_x(s)$. Now we can see that $\Delta N_x(s)$ follows an ordinary random walk with indipendent steps and it can be equal to the integer numbers between m_{in} and -1 included, since

$$N_x(s+1) = N_x(s) + j - 1 \qquad \text{if } Y_1(s) \text{ has priority larger than } x,$$
$$N_x(s+1) = j \qquad \text{if } Y_1(s) \text{ has priority smaller than } x,$$

with $j = 0, 1, m_{in}$. We have to distinguish the two cases because $Y_1(s)$ is the next executed task. In fact if $Y_1(s)$ has priority smaller than x, $N_x(s) = 0$ and so at the step $s + 1$ the only tasks with priority larger than x can be the tasks just added to the queue.
We compute the probabilities of $\Delta N_x(s)$ being equal to one of these values starting from $P(\Delta N_x(s) = m_{in})$, which is equal to the probability that Y_1 has priority smaller than x and all the m_{in} new activities have priority bigger than x, i.e.

$$P(\Delta N_x(s) = m_{in}) = P(Y_1(s) < x) \cdot (1 - x)^{m_{in}}.$$

For $0 \leq j \leq m_{in} - 1$, $P(\Delta N_x(s) = j)$ is equal to the probability that Y_1 has priority smaller than x and j new activities have priority bigger than x plus the probability that Y_1 has priority bigger than x and $j + 1$ activities have priority bigger than x, i.e.

$$\begin{aligned} P(\Delta N_x(s) = j) = P(Y_1(s) < x) \cdot \frac{m_{in}!}{j!(m_{in} - j)!} \cdot x^{m_{in}-j} \cdot (1 - x)^j + \\ + P(Y_1(s) \geq x) \cdot \frac{m_{in}!}{(j + 1)!(m_{in} - j - 1)!} \cdot x^{m_{in}-j-1} \cdot (1 - x)^{j+1}. \end{aligned}$$

The last probability computed is $P(\Delta N_x(s) = -1)$, which is equal to the probability that Y_1 has priority larger than x and all the m_{in} activities have priority smaller than x, i.e.

$$P(\Delta N_x(s) = -1) = P(Y_1(s) \geq x) \cdot x^{m_{in}}$$

We can now calculate the expectation value $E(\Delta N_x(s))$, which is equal to:

$$E(\Delta N_x(s)) = \sum_{j=-1}^{m_{in}} j \cdot P(\Delta N_x(s) = j) \tag{4.1}$$

At the equilibrium it is possible to demonstrate that $E(\Delta N_x(s)) = 0$ [8], i.e. at the equilibrium the number of tasks with priority bigger than x, on average, does not change.
Now defining the critical value x_c as the value of priority for which $P(Y_1 \geq x_c) = 1$, from equation (4.1) we obtain:

$$0 = -x_c^{m_{in}} + \sum_{j=1}^{m_{in}} j \cdot \frac{m_{in}!}{(j + 1)!(m_{in} - j - 1)!} \cdot {x_c}^{m_{in}-j-1} \cdot (1 - x_c)^{j+1}$$

Bringing to the first member ${x_c}^{m_{in}}$ and dividing for the same term both members of the equation, we now obtain

$$1 = \sum_{j=1}^{m_{in}} j \cdot \frac{m_{in}!}{(j + 1)!(m_{in} - j - 1)!} \cdot \left(\frac{1 - x_c}{x_c}\right)^{j+1} \tag{4.2}$$

Solving equation (4.2) for x_c we find $x_c = \frac{m_{in}-1}{m_{in}}$. This means that at the equilibrium only the tasks with priority larger than x_c can be executed, while the others stay indefinitely in the queue. In other words the tasks with priority larger than x_c have a finite probability to be executed and the tasks with priority under the critical one have a null probability to be executed. This results is confirmed by a simulation shown in the figure 4.1.

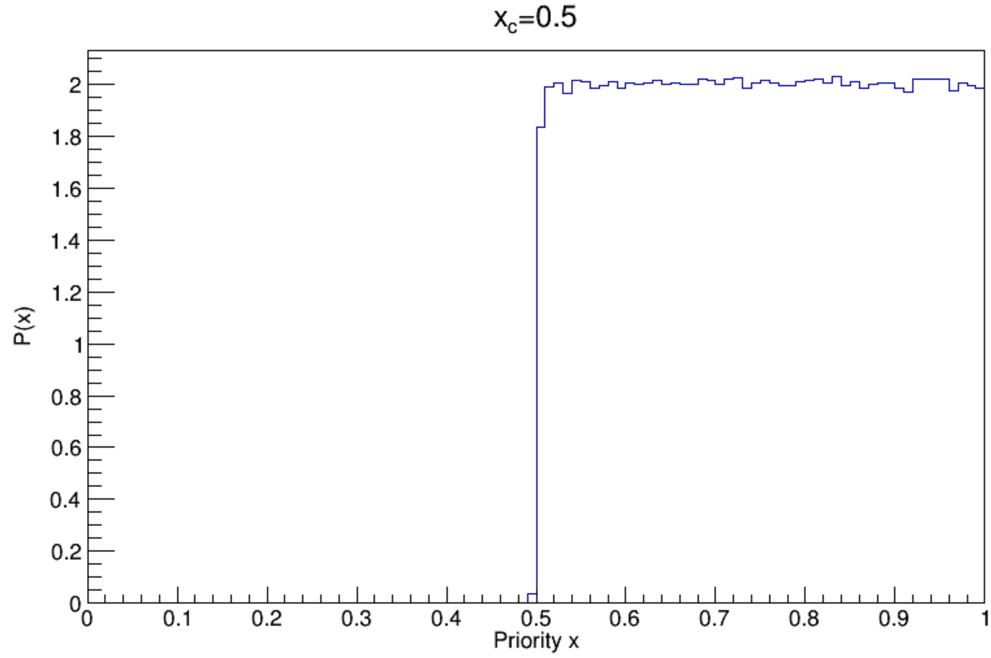

Figure 4.1: This is the result of a simulation of the answering process of an agent A using the queueing model. In the simulation we have discarded the first executions (one thousand) in order to reach the steady state and we start to register the priority of the executed task. Then we have filled up the histogram placing the priority x and the probability distribution $P(x)$ (computed as $P(x) = \frac{\#(\textit{executed tasks with priority x})}{\#(\textit{total executed tasks})\Delta x}$, with Δx the width of the bins in which the histogram is divided) on the axes. Gazing at the figure we can observe that the simulation confirms our predictions: in fact $P(x)$ assumes (approximately) the form of a step function with the discontinuity at $x_c = \frac{2-1}{2} = \frac{1}{2}$.

4.2 Scaling exponent

We have shown above that in the critical state only tasks with priority larger than $x_c = \frac{m-1}{m}$ are executed and their number is a recurrent random walk, so we expect that the waiting time, which is the time of return to zero, follows a power-law. Using random walk and diffusion theory arguments we can demonstrate that the scaling exponent α of the re-clocked $\overline{P}_R(\sigma)$ is equal to $-\frac{3}{2}$ [9].
We start writing the probability distribution of RTs as:

$$\overline{P}_R(\sigma) = \sum_{n=0}^{\infty} \int_{x_c}^{1} \overline{Q}(n,x) G(n,x,\sigma) dx \tag{4.3}$$

where $\overline{Q}(n,x)$ is the probability that at a certain time step at the stationary state there are exactly n tasks with priority $x \geq x_c$ and $G(n,x,\sigma)$ is the conditional probablity that a certain task with priority $x \geq x_c$ which is added to the list at a time step at which there are n other task with priority larger than x in the queue is executed after σ steps. The meaning of equation (4.3) is simple: $\overline{Q}(n,x)G(n,x,\sigma)$ is the probability that a task with priority x added when there are n others with priority larger than x is executed after σ steps. Integrating from x_c to 1 this probability we compute the probability of answering to a message with a generic priority $x_c \leq x \leq 1$ that still depends on n. Summing the computed probability over all the possible values of n we obtaining in this way the global probability due to all the possible contributions of having a waiting time of execution (i.e. the RT in the written communication case) equal to σ.
What we have to do now is to obtain the expressions of $\overline{Q}(n,x)$ and $G(n,x,\sigma)$. Since n evolves following a random walk, we compute the probability of having a certain number n at the step $s+1$ as sum of the probability of having $n+1$, n or $n-1$ at the step s multiplied by the probability that in the inbox comes 0, 1 or m_{in} messages with priority larger than x. Looking for the stationary solution of the obtained equations, we find $\overline{Q}(n,x)$. Instead $G(n,x,\sigma)$ is computed using the continous time t and $n = y$ approximation to write and solve the diffusion equation for the probability $Q(y,x,t)$ and a first-passage process result applied to $Q(y,x,t)$.
We write the master equation for the probability $Q(n,x,s)$ that at the time s there are exactly n tasks with priority larger than x in the queue. The number n changes from a step to the next one for two reasons. Since we are interested in the case $x \geq x_c$, if $n > 0$ moving from a step to another n decreases by an unity because the task with the greatest priority (which is of course greater than x_c) is executed, while if $n = 0$ the executed task has priority larger than x_c but smaller than x. The number n changes also because $m_{in} = 2$ tasks are added step by step. Taking account of these contributions, for $n \geq 3$ we have:

$$Q(n,x,s{+}1) = Q(n{+}1,x,s)x^2{+}Q(n,x,s)2x(1{-}x){+}Q(n{-}1,x,s)(1-x)^2, \tag{4.4}$$

and for $n \leq 2$ we have:

$$\begin{gathered} Q(2,x,s{+}1) = Q(3,x,s)x^2{+}Q(2,x,s)2x(1{-}x){+}Q(1,x,s)(1-x)^2{+}Q(0,x,s)(1-x)^2, \\ Q(1,x,s+1) = Q(2,x,s)x^2 + Q(1,x,s)2x(1-x) + Q(0,x,s)2x(1-x), \qquad (4.5) \\ Q(0,x,s+1) = Q(1,x,s)x^2 + Q(0,x,s)x^2. \end{gathered}$$

where x^2 is the probability that both the new tasks have priority smaller than x, $2x(1-x)$ is the probability that one of the two new task has priority larger than x while the other has smaller priority and $(1-x)^2$ is the probability than both the new tasks have priority larger than x. These probabilities are computed as done in the demonstration of the critical priority.
$\overline{Q}(n,x)$ is the stationary solution of the ricorsive equations (4.4) and (4.5) and solving this system, looking for the well-normalized solutions, we find for $x \geq x_c = \frac{1}{2}$

$$\begin{aligned} \overline{Q}(n,x) &= \frac{2(x-x_c)}{x^2}\left[\frac{(1-x)^2}{x^2}\right]^{n-1} \qquad n \geq 2, \\ \overline{Q}(1,x) &= \frac{2(1-x^2)}{x^2}(x-x_c), \\ \overline{Q}(0,x) &= 2(x-x_c). \end{aligned} \tag{4.6}$$

Now we need to find $G(n,x,\sigma)$. We can demonstrate, using the expressions of the discrete derivatives, that in the continous time t and $n=y$ approximations, the equation (4.4) becomes the diffusion equation

$$\partial_t Q(y,x,t) = c(x)\partial_y^2 Q(y,x,t) + d(x)\partial_y Q(y,x,t) \tag{4.7}$$

with $c(x) = x^2$ and $d(x) = 2x-1$. The solution of equation (4.7) with initial condition $Q(y,x,0) = \delta(y-n)$ is

$$Q(y,x,t) = \frac{1}{\sqrt{4\pi c(x)t}}\left[e^{-\frac{(y+d(x)t-n)^2}{4c(x)t}} - e^{\frac{d(x)n}{c(x)}}e^{-\frac{(y+d(x)t+n)^2}{4c(x)t}}\right]. \tag{4.8}$$

We can see $G(n,x,\sigma)$ as the probability that at the stationarity state at fixed x, starting with $y=n$, the first time we have $y=0$ is after $\Delta s = \sigma$ time steps. Looking at this as a first-passage process [10], we can write

$$G(n,x,t) = -\partial_t\left[\int_0^\infty Q(y,x,t)dy\right]$$

and using (4.8) we find

$$G(n,x,t) = \frac{1}{\sqrt{4\pi c(x)}t^{3/2}}e^{-\frac{(d(x)t-n)^2}{4c(x)t}}. \tag{4.9}$$

Now, returning to consider discrete time s (i.e. clocked through s) and n, we have $\overline{Q}(n,x)$ and $G(n,x,\sigma)$ from equations (4.6) and (4.9), we use them into equation (4.3) to find $\overline{P}_R(\sigma)$. In fact with these results for $\overline{Q}(n,x)$ and $G(n,x,\sigma)$ it is possible to demonstrate that, for large σ, $\overline{P}_R(\sigma) \sim \sigma^{-3/2}$ [11], i.e in the steady state the RTs probability distribution follows a power-law with scaling exponent equal to $-\frac{3}{2}$.
Empirical data confim this result, as shown above in figure 2.4. We have had a confirmation of this result also through a simulation. In this way we have obtained an scaling exponent $\alpha = -1.507 \pm 0.008$ as shown in figure 4.2, in agreement with the prediction of the theory.
The results for $G(n,x,\sigma)$ and $\overline{Q}(n,x)$ obtained from equations (4.9) and (4.6) are independent from the number of incoming tasks m_{in} at each time step. So we expect to have the same result for $\overline{P}_R(\sigma)$ also in the case in which at each

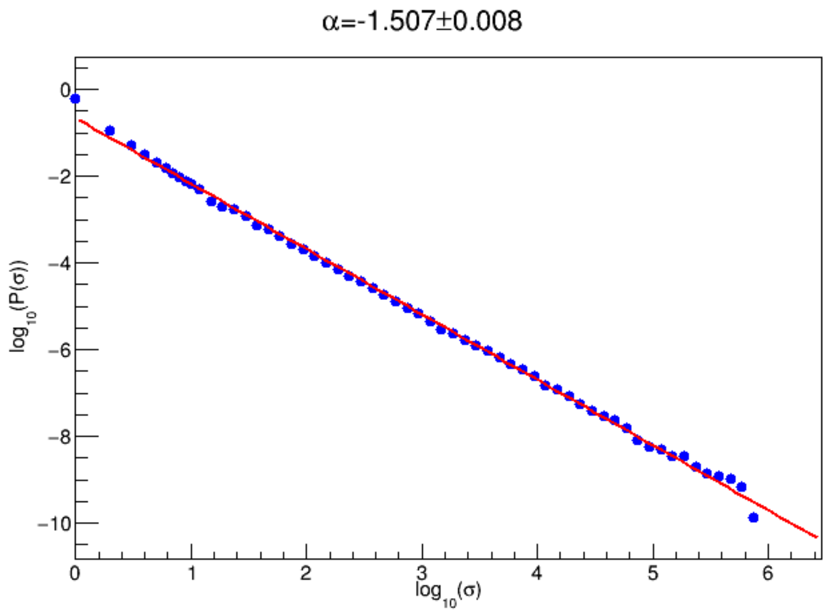

Figure 4.2: Plot of the simulation data. On the axes we have put $\log_{10}\sigma$ and $\log_{10}\mathrm{P}(\sigma)$ in order to obtain a straight line. The slope of this line is the scaling exponent. The quantity of interest is the waiting time Δs that a certain task remains in the queue waiting to be executed. To obtain this, we have registered the time step s_{in} at which the task was added to the list and the time step s_{out} at which the same task was deleted from the list. The waiting time has been computed as $\Delta s = \sigma = s_{out} - s_{in}$. Since the theorical exponent is computed assuming the steady state, we had to wait for the queue to reach it. To ensure this condition, we have neglected the first steps (one thousand). To compute the probability of having a certain waiting time σ we have counted the number of tasks that waited exactly σ steps before being executed and we have normalized it dividing by the total number of executed tasks.

time step $m = m(s)$ is not constant, but fluctuate such that $\langle m \rangle \geq 1$ and finite variance.
To verify this, we can execute a simulation considering a specific agent of a given communication medium. Now the number of tasks added step by step is not constant, but it is derived by considering the empirical distribution of incoming tasks between any two consecutives activities of the considered agent. The results are shown in figure 2.4.As seen also in this case we obtain scaling exponents close to $-\frac{3}{2}$ for the considered agents across the three media.

Chapter 5

Barabási model

Barabási was the first to built up a model that applies rank-based processes to study written communication dynamics [12]. However, his work received many criticisms that involde both data analysis and the model itself. We will discuss briefly at the end of the chapter some flows of the Barabási model.

An user creates a list with L activities that he/she has to do. For each activity a priority $x \geq 0$ is assigned randomly from a probability density function $\rho(x)$.

The initial list has L tasks. At each time step $s > 0$ we have a probability p that the activity with the largest priority is executed, while a random task is executed with probability $1-p$. After having executing a task, a new task with a new indipendent random priority is added to the list. The parameter p is called *control parameter*: if $p = 1$ we are in the situation that the task with the largest priority is always executed, while $p = 0$ means that we always execute a randomly choosen activity.

Let us consider, for semplicity, Barabási's model with initial $L = 2$. We call *new task* the just added one, while the other is called *old task*.

We indicate with $\rho(x)$ the new task priority PDF and with $R(x) = \int_0^x \rho(x)dx$ the distribution function. We assume that these two functions are given. Now we indicate with $\rho_1(x,t)$ and $R_1(x,t) = \int_0^x \rho_1(x,t)\,dx$ the same two functions for the old task at the time step s.

At the step $s+1$ there are two active tasks in the list with their priorities distributed following $R(x)$ and $R_1(x,t)$. After having chosen an activity, the other becomes the old one and its distribution function is

$$R_1(x, s+1) = \int_0^x dx' \rho_1(x', s) q(x') + \int_0^x dx' \rho(x') q_1(x', s), \tag{5.1}$$

where

$$q(x) = p[1 - R(x)] + (1-p)\frac{1}{2} \tag{5.2}$$

is the probability that the new task is choosen if the old one has priority x, and

$$q_1(x, s) = p[1 - R_1(x, s)] + (1-p)\frac{1}{2} \tag{5.3}$$

is the probability that the old task is chosen if the new one has priority x.

Integrating by parts equation (5.1) we obtain

$$
\begin{aligned}
R_1(x,s+1) = {} & \frac{1+p}{2}[R(x)+R_1(x,s)] - \Bigg[R(x)+R_1(x,s)+ \\
& - \int_0^x \rho(x')R_1(x',s)dx' + \int_0^x \rho(x')R_1(x',s)dx'\Bigg] = \\
= {} & \frac{1+p}{2}[R(x)+R_1(x,s)] - \Bigg[R(x)+R_1(x,s)\Bigg].
\end{aligned}
$$

In the stationary state $R_1(x,s+1)=R_1(x,s)$, so we obtain:

$$
\begin{aligned}
R_1(x) = {} & \frac{1+p}{2}\frac{2}{2R(x)+(1-p)}R(x) = \\
= {} & (p+1)\frac{1}{2-\frac{p-1}{R(x)}} = \cdots = \\
= {} & \frac{p+1}{2p}\left[1-\frac{1}{1+\frac{2p}{1+p}R(x)}\right].
\end{aligned}
$$

Let us consider the extremal cases. If $p \to 0$ we obtain:

$$
\lim_{p\to 0} R_1(x) = R(x),
$$

i.e. if we choose randomly the task to execute, the old task priority distribution function is equal to the new task one.
Instead if we have $p \to 1$ we obtain:

$$
\lim_{p\to 1} R_1(x) = \begin{cases} 0, & x=0 \qquad \big(R_1(0)=0\big) \\ 1, & x>0 \end{cases} .
$$

This show us that $R_1(x)$ is different from 1 only when $x \to 0$ and so in this limit $\rho_1(x)$ is concentrated around $x=0$. These two results are shown also in the figure 5.1.
Now we want to find the waiting time distribution according to this model.Let us consider the situation in which a task is added to the queue with priority x. The probability that it waits σ steps before being executed, since the choice of this task is indipendent between two consecutive steps, is equal to the probability that it is not chosen for the first $\sigma-1$ steps multiplied by the probability of chosing it at the step σ. Remembering the meaning of $q_1(x)$ and $q(x)$, the probability that the task is not chosen is $q_1(x)$ for the first step, while for the others is $q(x)$ because it becames the old task. So we can write

$$
\overline{P}_R(\sigma) = \begin{cases} \int_0^\infty dR(x)[1-q_1(x)], & \sigma = 1 \\ \int_0^\infty dR(x)q_1(x)[1-q(x)]q(x)^{\sigma-2}, & \sigma > 1 \end{cases} . \tag{5.4}
$$

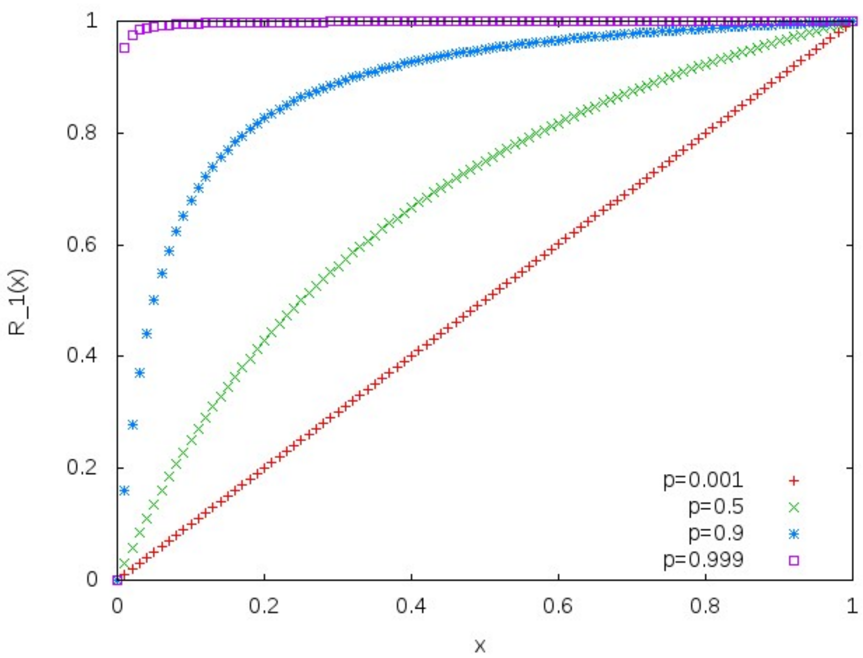

Figure 5.1: We assume $\rho(x)$ equal to the uniform distribution, so $R(x) = x$. In this plot we put the priority $x \in [0, 1]$ and $R_1(x)$ on the axes. When $p \to 0$, we have $R_1(x) \to R(x) = x$, while when $p \to 1$ $R_1(x)$ increases to reach 1 in a small interval around $x = 0$.

Let us consider the case $\sigma = 1$. We obtain:

$$
\begin{aligned}
\overline{P}_R(\sigma) &= \int_0^\infty dR(x)[1 - q_1(x)] = \int_0^\infty dR(x)\left[1 - p[1 - R_1(x,t)] + (1-p)\frac{1}{2}\right] = \\
&= \frac{1-p}{2} + p\int_0^\infty \frac{1+p}{2p}\left[1 - \frac{1}{1 + \frac{2p}{1-p}R(x)}\right] = \\
&= \frac{1-p}{2} + \frac{1+p}{2} - \frac{1+p}{2}\int_0^\infty dR(x)\frac{1}{1 + \frac{2p}{1-p}R(x)} = \\
&= 1 - \frac{1+p}{2}\frac{1-p}{2p}\ln\left(1 + \frac{2p}{1-p}R(x)\right)\Bigg|_0^\infty = \\
&= 1 - \frac{1-p^2}{4p}\left[\ln\left(1 + \frac{2p}{1-p}\right) - \ln(1)\right] = \\
&= 1 - \frac{1-p^2}{4p}\ln\left(\frac{1+p}{1-p}\right).
\end{aligned}
$$

Let us consider the case $\sigma > 1$. We obtain:

$$
\begin{aligned}
\overline{P}_R(\sigma) &= \int_0^\infty dR(x) q_1(x)[1 - q(x)]q(x)^{\tau-2} = \\
&= \int_0^\infty dR(x)\left[p - pR_1(x) + \frac{1+p}{2}\right]\left[1 - p + pR(x) - \frac{1-p}{2}\right]\left[p - pR(x) + \frac{1+p}{2}\right]^{\sigma-2} = \\
&= \int_0^\infty dR(x)\left[\frac{1-p^2}{4} + \frac{p+p^2}{2}R(x) - \frac{p-p^2}{2}R_1(x) - p^2 R_1(x)R(x)\right]\left[p - pR(x) + \frac{1+p}{2}\right]^{\sigma-2} = \\
&= \cdots = \frac{1-p^2}{4p}\left[\left(\frac{1+p}{2}\right)^{\sigma-1} - \left(\frac{1-p}{2}\right)^{\sigma-1}\right].
\end{aligned}
$$

In both cases $\overline{P}_R(\sigma)$ does not depend on $\rho(x)$. In fact all dependences with x of $\overline{P}_R(\sigma)$ in equation (5.4) appear via $R(x)$, since $q(x)$ and $q_1(x)$ depend on x through $R(x)$ as we can see from equations (5.2) and (5.3).
If $p \to 0$ we have

$$\overline{P}_R(\sigma) = \left(\frac{1}{2}\right)^{\sigma}$$

for $\sigma \geq 1$. In the random choice limit we have that an activity is chosen with probability $\frac{1}{2}$ at each step.
Instead if $p \to 1$ we have

$$\lim_{p\to 1} \overline{P}_R(\sigma) = \begin{cases} 1 + \mathcal{O}\left(\frac{1-p}{2}\ln(1-p)\right), & \sigma = 1 \\ \mathcal{O}\left(\frac{1-p}{2}\right)\frac{1}{\sigma-1}, & \sigma > 1 \end{cases}.$$

Now we have that almost all the tasks wait for interval $\sigma = 1$, while the waiting time of the activities not selected at the first step after being added to the queue follows a power-law with scaling exponent $\alpha = -1$.
For large σ we obtain

$$\overline{P}_R(\sigma) = \frac{1-p^2}{4p}\left[\left(\frac{1+p}{2}\right)^{\sigma-1} - \left(\frac{1-p}{2}\right)^{\sigma-1}\right]\frac{1}{\sigma-1} \to \frac{1-p^2}{4p}\frac{1}{\sigma}\left(\frac{1+p}{2}\right)^{\sigma}.$$

For this reason we can write

$$\lim_{\sigma\to\infty} \overline{P}_R(\sigma) = \frac{1-p^2}{4p}\frac{1}{\sigma}\exp\left[\sigma\ln\left(\frac{1+p}{2}\right)\right] = \frac{1-p^2}{4p}\frac{1}{\sigma}\exp\left(-\frac{\sigma}{\sigma_0}\right)$$

with

$$\sigma_0 = \left(-\ln\left(\frac{1+p}{2}\right)\right)^{-1} = \left(\ln\left(\frac{2}{1+p}\right)\right)^{-1}.$$

$\overline{P}_R(\sigma)$ shows an *exponential cut-off* for $\sigma \to \infty$. From expression of σ_0 we can see that the *cut-off* moves to higher σ for $p \to 1$; meanwhile the power-law behaviour $\overline{P}_R(\sigma) \sim \sigma^{-1}$ is more evident.
We have executed a simulation of Barabási model with two different value of p obtaining figure 5.2.
As we reported at the beginning of the chapter the Barabási model received several criticism (see [1] and the references there are in). First of all the inbox queue lenght remaining constant since after executing a task, only one task is added at each step, while we see from empirical data that usually between two consecutive activities an user receives on average a number of messages $\langle m_{in}\rangle > 1$. Another difference from the reality is that only a small part of the probability distribution follows a power-law trend. In fact almost all the tasks wait only $\sigma = 1$ step before being executed and so the waiting times following the power-law behavior are a very little part. Besides it predicts an different scaling exponent $\alpha = -1$. Empirical data and our more realistic model (in which $\langle m_{in}\rangle > 1$ messages enter in the inbox at each step) give us a scaling exponent $\alpha = -\frac{3}{2}$.

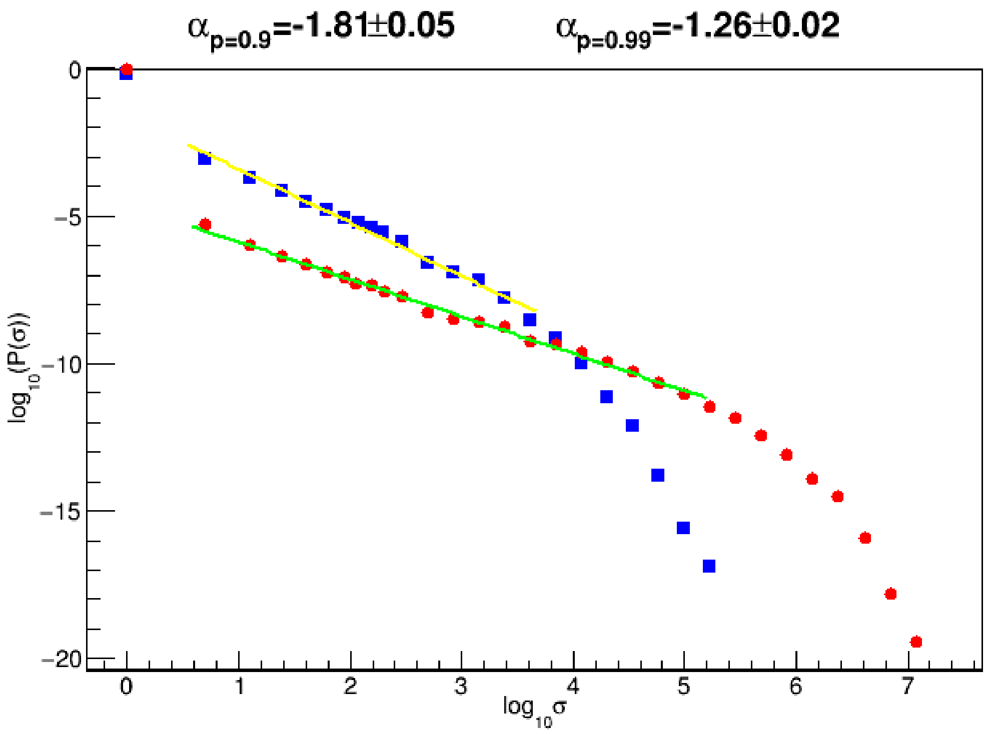

Figure 5.2: In these simulations we have used two different value for the control parameter: blue squares have been obtained using $p = 0.9$ and red circles using $p = 0.99$. We have plotted the results in a log-log scale. From the simulations we can see that almost all the mass of $\mathrm{P}(\sigma)$ is concentrate in $\sigma = 1$. Then the central part of the distribution displays a power-law behavior that turns into an exponential cut-off for large values of σ. The scaling exponent α of the power-law part of the distribution converges to -1 increasing the value of control parameter: in fact we have $\alpha = -1.81 \pm 0.05$ for $p = 0.9$ and $\alpha = -1.26 \pm 0.02$ for $p = 0.99$.

Conclusions

We have introduced the written communication problem giving the concepts of inter-event time (IET) and response time (RT) as the time elapsed between two consecutive activities and the time elapsed between the arrival and the execution of a certain task and then we also have discussed the need of a re-clock of the time, passing from standard time t to time clocked through activities s. In this way we have eliminated the interactions of the RTs with the spontaneous IETs, which are media- and agent-dependent and so, as we saw from the empirical data, the re-clocked RT probability distribution, in which we were interested, followed a common power-law trend and made evident the existence of an universal statistical pattern in the considered answering process.
To make rigorous predictions on the problem and to understand how this power-law behavior originate, we have built up a queuing model to reproduce the dynamics of written communcation. The salient trait of the model is the priority driven dynamics: to each incoming tasks is assigned a number sampled from a distribution, then the tasks with largest priority is executed first. For this reason we have looked at our dynamics as a rank-based Markov chain. This look has suggested us the existence of a priority threshold x_c that has separeted sharply the inbox. In fact the tasks under this threshold could never be executed, while all the dynamics took place above this value. So we have considered only the number of tasks with priority above the critical value and looking at this as a recurrent random walk we have understood the birth of the power law behavior, since the time to return to zero of a random walk follows a power law. To do mathematical demonstration we have assumed to be constant the number of incoming messages m_{in} between two activities and equal to 2, finding $x_c = \frac{m_{in}-1}{m_{in}} = \frac{1}{2}$ and the power law scaling exponent α equal $-\frac{3}{2}$. These two results have been also confirmed by simulations.
The priority mechanism and the presence of a priority threshold with the random walk arguments have led us to consider our model and the problem of written communication as a self-organized critical system. For this reason we have explained what is SOC, looking also to models similar to ours bound to this theory in order to show that there is a relationship between self-organized criticality and queuing models and their implications (i.e. threshold and random walk leading to power-law behaviors).
In the end we have presented Barabási model as the first queuing model applied to written communication and using it we have predicted RTs probability distribution to have a power law behavior with scaling exponent equal to -1 with an exponential cut-off. This is one of the reasons because of the several criticism that Barabási received. We have reported some of these to underline the lacks of the model, without trying to diminuish its historic importance.

Bibliography

[1] M. Formentin, A. Lovison, A. Maritan and G. Zanzotto. Hidden scaling patterns and universality in written communication. *Physical Review Letters* E (Vol.90, No.1), 2003.

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

[3] P. Bak, C. Tang and K. Weisenfeld. Self-organized criticality: an explanation of 1/f noise . *Physical Review Letters*, 38: 364–374, 1987

[4] P. Bak. 1996. *How Nature Works: the Science of Self-organized Criticality*, Springer, Berlin and New York, 1996.

[5] P. Bak, K. Sneppen. Punctuated equilibrium and criticality in a simple model of evolution, 1993, *Physical Review Letters* 74.

[6] R. Meester and A. Sarkar. Rigorous self-organised criticality in the modified Bak-Sneppen model, 2012, *Journal of Statistical Mechanics-Theory and Experiment* 149.

[7] H. Luckock. A steady-state model of the continuous double auction, 2003. *Quantitative Finance* 3(5).

[8] M. Formentin, J.M. Swart. The limiting shape of a full mailbox, 2015. Preprinted.

[9] A. Gabrielli and G. Caldarelli. Invasion percolation and the time scaling behavior of a queuing model of human dynamics. *Journal of Statistical Mechanics-Theory and Experiment*, FEB 2009.

[10] S. Redner. *A Guide to First-Passage Processes*, Cambridge University Press, New York, 2001.

[11] G. Grinstein and R. Linkser. Biased diffusion and universality in model queues. *Physical Review Letters*, 2006.

[12] A. Vázquez. Exact results for the Barabási model of human dynamics. *Physical Review Letters*, 95:248701, 2005.