Emergence of log-normal distributions in avalanche processes, validation of 1D stochastic and random network models, with an application to the characterization of cancer cells plasticity

Stefano Polizzi

Université de Bordeaux

PhD thesis in Physics
École Doctorale des Sciences Physiques et de l'Ingénieur
Spécialité Laser, Matière et Nanosciences

Emergence of log-normal distributions in avalanche processes, validation of 1-D stochastic and random network models, with an application to the characterisation of cancer cells plasticity

PhD Candidate:
Stefano POLIZZI
Defended on the 26th of November 2020

Committee Members:

Françoise ARGOUL, Dir. de recherche CNRS (LOMA, Bordeaux), PhD Supervisor
Martine BEN AMAR, Professeur (ENS Paris), President and Rapporteur
Plamen Ch. IVANOV, Research Professor (Boston University), Rapporteur
Gianluigi MONGILLO, Chargé de Recherche CNRS (Vision Inst., Paris), Examiner
Francisco PEREZ-RECHE, Senior Lecturer (University of Aberdeen), Examiner

3 years PhD - 2017-2020

"Children find everything in nothing, men find nothing in everything"
A complex system theory is a way
to see the universe with the eyes of a child
G. Leopardi

Acknowledgements

Pendant ces 3 ans de thèse nous avons vécu des événements catastrophiques. Le plus malheureux de tous a été la disparition d'Alain Arneodo. Alain n'était pas officiellement mon directeur de thèse, mais je le considérais comme tel, vu les fortes interactions que j'avais avec lui. Alain et Françoise étaient une équipe de direction complémentaire qui était très stimulante pour moi. Ils m'ont montré à quel point la diversité est aussi un facteur important pour faire avancer la science.

J'avais un rapport spécial avec Alain, probablement rare pour un étudiant en thèse avec son directeur. Nous avions souvent des longues discussions soit par rapport à mon travail soit par rapport à la vie, et en même temps on rigolait beaucoup. Mon bureau étant sur le chemin des toilettes il lui était facile de s'arrêter discuter, même si c'était simplement pour commenter le dernier graphique que j'avais fait. Il m'a transmis sa passion pour la recherche et la science. Cette passion était pour lui un défi, le défi de la connaissance. Alain était aussi un passionné de sport, un autre genre de défi. Après sa perte, très douloureuse pour moi sur le plan humain et scientifique, mon travail a dû être complètement réorganisé, je ne pouvais plus compter sur ses conseils. D'un autre côté, cela m'a poussé à acquérir beaucoup plus d'autonomie, nécessaire pour terminer la seconde moitié de ma thèse. Je m'estime heureux d'avoir rencontré quelqu'un comme Alain.

Merci Françoise pour tout ton aide, les discussions qu'on a eu sur les sujets les plus divers, merci pour ton amitié. Tes conseils ont été fondamentaux pour ma maturité en tant que personne et chercheur. Je sais que ces trois ans n'ont pas été faciles pour toi non plus, et ça permit d'apprécier encore plus ta proximité. Je me suis senti intégré dans le groupe dès le début: même quand j'étais un simple stagiaire, je me sentais déjà complètement impliqué et écouté. Dans mon travail, j'ai toujours eu besoin de sentir que les autres me faisaient confiance, ce qui est probablement un défaut, mais avec toi (et Alain aussi) ça a toujours été le cas. Tu m'as aussi appris à aller jusqu'à mes dernières limites, ce qui est nécessaire pour faire de la science, car nous devons toujours aller jusqu'aux limites de nos connaissances, et les pousser encore plus loin. Merci aussi à aux autres membres de l'équipe, Étienne Harté, Alexandre Guillet et Léo Delmarre. J'ai partagé avec vous ces moments compliqués et grâce à votre bonne humeur, aux blagues, aux discussions de science (ou des ...) tout a été un petit peu plus simple. D'ailleurs, Étienne on attend toujours le verre de Sauternes! Merci à vous aussi pour l'assistance lorsque j'ai fait des manips, malheureusement je n'ai pas rentré dans cette thèse le travail expérimental fait avec l'AFM, ce qui m'a cependant pris beaucoup de temps.

I am grateful to all the committee members to have helped in improving my work and have accepted the invitation to be part of the evaluation jury in this complicated period of Covid-19 epidemics. I think that your work, and in general the work of reviewers, is fundamental for the advancing of science. Thanks to Plamen Ch. Ivanov for being rapporteur of this thesis, and to Gianluigi Mongillo and Francisco Pérez-Reche for being examiner. Paco it was a pleasure to work with you, I hope we will still continue our collaboration, and thank you for inviting me in Scotland, it was very stimulating and definitely a big turning point of my PhD. A special thanks to the President of the jury (and rapporteur), Martine Ben Amar, who accepted to come to Bordeaux from Paris,

challenging public transports and viruses, in this way allowing my defence to be held in presence (even though with no public), instead of online. Actually I should also thank Martine for being my professor of Non-Linear Physics in Paris, and for having advised me to come to Bordeaux for this PhD, it was a very good choice.

I like to say that science is an emergent property, and therefore the result of a collective behaviour of a complex system, exactly as the one we will study in this thesis. Even Einstein would not have created General Relativity without his context, without all the apparently useless mathematics developed by his contemporaries. Of course, I am not comparing to Einstein, but still all the people that have in some way interacted with me are necessary for this process, therefore I want to thank them all. Thanks to David Dean, Jean-Pierre Delville (also for sharing your great wine culture, well... and also wine glasses), to Thomas Salez. I am indebted to my present and past officemates, Laura Squarcia, Houssem Kahli, Qingguo Bai, Sotiris Samatas, Harinadha Gidituri, Léo Delmarre. You were all very nice, and there have always been a positive atmosphere between us, helping us each other. The same is true for many other PhDs or post-docs, such as Marcela Rodriguez-Matus, Benjamin Sanchez-Padilla, Hernando Magallanes to name the Mexican team always challenged in table football by our Italian team. I also have to cite Paul Gersberg, Raphaël Saiseau, Lorenzo Mauro, Louis Bellando de Castro, Ludovic Jaubert, Nina Crevettes (ehm Kravets pardon!), Goce Koleski. With many of you I became friend and exchanged much more than everyday working life, sports, drinks, fun. I am happy to have met you, there are no words to say how much you have been important for this PhD. I am grateful to Tommaso Matteuzzi, which besides being an old friend had a direct impact on this work for all exciting discussions about physics and epistemology, I hope that one day we will work on something together.

Thanks to the director of LOMA that allowed me to be here and support my application for the funding and all my practical needs. I wish to thank as well to all the services of LOMA, administration, informatics, optics and electronics, and to the cleaning ladies sweeping our office every day.

Vorrei anche ringraziare, in italiano, anche se so che siete poliglotti, i miei genitori, Gaspare e Gabriella (per gli amici Gasparella!), e la mia sorellina Erika. In fondo tutto è nato da voi, lo so che da «giovane» non sono sempre stato bravo e buono, ma vi ringrazio per avermi sopportato e supportato. Crescere consiste anche in momenti difficili, di solitudine, soprattutto se si fa la scelta di andare lontani, all'estero, per studiare; grazie per avermi aiutato a superarli e ad accettarli. E anche se la vita deve sempre essere un po' un parco giochi, contrariamente a quello che mi dicevate, spero che siate orgogliosi di questo lavoro. Erika sono contento che tu sia arrivata qui, t'immagini a passare tutto il tempo con i genitori! Sono sicuro che anche tu troverai la tua via. Un saluto ed un grazie anche a tutti i miei amici storici di Firenze, che saranno sempre importanti per me. Ogni volta che ci vediamo è come se il tempo non fosse passato, anche se purtroppo oggi vivamo un po' più lontani. Francesco Rinaldo, Cappe, Pala, Jack, Lollo Baldi, Matilde e la piccola Nina, Antonio, Gianluca, Federico, Isabel, Marta e sicuramente mi sto dimenticando di qualcuno, me ne dispiaccio. Vorrei anche ricordare i miei amici di Bordeaux: Lorenzo, che mi ha gentilmente concesso il suo pavimento, Maria, Giulio (figa), Andrea, Giulio, Radu, Hermes, Misha.

Enfin Marie, tu resteras toujours importante pour moi. Tu m'as toujours compris et

aidé quand je travaillais jusqu'à tard, tu as compris ma passion pour la recherche et tu ne m'as jamais fait sentir coupable pour ça, même si cela voulait dire ne pas aller danser la salsa. Merci aussi d'avoir stimulé ma fantaisie, mon originalité, il y a une partie de toi dans cette thèse aussi!

Marie, tu es ma petite goutte de joie crépitante,
sans toi cette aventure n'aurait pas pu être aussi excitante,
peut-être qu'elle n'aurait été du tout,
mais notre aventure n'est pas au bout,
je serai là quand tu auras besoin de moi

Résumé détaillé

Plusieurs matériaux vitreux ont des comportements caractéristiques suite à fractures induites par des contraintes. Ces fractures se présentent comme des processus d'avalanche dont la statistique dans la plus part des cas suit une loi de puissance, rappel de comportements collectifs et critiques auto-organisés. Des avalanches de fractures sont observées aussi dans des matériaux biologiques et des systèmes vivants, qui peuvent être vus, si l'on ne considère pas le remodelage actif, comme un réseau vitreux, avec une structure figée. Le travail accompli dans cette thèse est motivé par ce dernier type de matériaux, en particulier le cytosquelette d'actine (CSK) des cellule vivantes, mais peut être appliqué à une classe de processus plus vaste, qui inclut les matériaux amorphes et d'autres polymères vitreux.

Le cytosquelette d'actine (CSK) forme des structures organisées en microfilaments par un mécanisme dynamique d'assemblage-désassemblage de cross-linkers. L'adaptation rapide de ce réseau de filaments sous compression ou étirement est important pour la survie des cellules en situations réelles, par exemple pour faire face aux distances considérables des organes ou tissus qui les hébergent. De plus, plusieurs pathologies, comme le cancer, modifient significativement les propriétés mécaniques des cellules et du CSK. À l'aide du microscope à force atomique pour sonder les cellules vivantes, nous avons détecté qu'elle répondent à un cisaillement local rapide à travers des cascades d'événements aléatoires de ruptures de leur cytosquelette. Ceci suggère que le CSK se comporte comme un réseau aléatoire quasi-rigide de filaments interconnectés. De cette façon nous visons à reproduire les force et déformations importantes que les cellules subissent en condition physiologiques et/ou pathologiques, qui dépassent souvent la limite visco-élastique. Plus précisément nous analysons des données expérimentales provenant de cellules CD34+ de moelles osseuses saines et leucémiques, cependant ces mêmes comportements ont été depuis observés dans d'autres types de cellules.

Étonnement, les distributions de la force, la taille et l'énergie relâchée lors de ces cascades, ne suivent pas une distribution en loi de puissance typique de phénomènes critiques. En fait la distribution de la taille des avalanches s'avère être log-normale, en suggérant que ces événements de rupture ne suivent pas la statistique en loi de puissance typique des phénomènes critiques et des distributions des tailles d'avalanches dans les matériaux amorphes. Dans le but de donner une interprétation de ce comportement particulier nous proposons d'abord un modèle stochastique minimal (1D). Ce modèle donne une interprétation de l'énergie relâchée dans les cascades de ruptures, au regard d'une somme (étant l'énergie additive) d'un processus multiplicatif de cascade avec une relaxation temporelle. Nous identifions 2 types d'événements de ruptures : des fractures friables susceptibles de représenter des ruptures irréversibles dans un CSK rigide et très connecté, et des fractures ductiles résultant des décrochements dynamiques des cross-linkers pendant la déformation plastique sans perte d'intégrité du CSK. Notre modèle stochastique reproduit quantitativement les distributions expérimentales des énergies relâchées et fournit une compréhension mathématique et mécanique de la statistique log-normale observée dans les deux (friable et ductile) cas. Nous montrons aussi que les fractures friables sont relativement plus importantes dans les cellules leucémiques, en témoignant leur plus grande fragilité et leur différente architecture du CSK, plus rigide et réticulée.

Ce modèle minimal motive la question plus générale de quelles sont les distributions résultantes pour la somme de variables corrélées provenant d'un processus multiplicatif. En conséquence nous analysons la distribution de la somme d'un processus de branchement généralisé évoluant avec un facteur de croissance aléatoire continu. Le processus dépend seulement de 2 paramètres, $\bar{a}$ and $\tilde{a}$: les 2 premiers moments centrés de la distribution du facteur de croissance. Cette question est connectée au problème de la somme de variables aléatoires corrélées suivant une distribution log-normale, d'actualité dans plusieurs domaines, comme par example en finance, épidémiologie, électronique, et en général lorsqu'il y a un processus multiplicatif sous-jacent. La raison est qu'un processus multiplicatif converge asymptotiquement à variables suivant une distribution log-normale. Nos résultats sont exacts dans le cas où le facteur reproductif suit une loi uniforme, mais ils sont généralisables à une classe de distributions plus large, sous certaines conditions de compacité. Nous concevons donc un diagramme de phase en montrant 3 régions différentes : 1) une région où la distribution finale a tous les moments finis et qui est approximativement log-normale. Cette région est caractérisée par des faibles valeurs de $(\bar{a}, \tilde{a})$. 2) Une région où la distribution asymptotique est une lois de puissance, avec un exposant inclus dans l'intervalle $[1; 3]$, dont la valeur est déterminée par les paramètres du modèle. Cette région est obtenue pour des valeurs plus grandes dans le plan $(\bar{a}, \tilde{a})$. 3) Enfin dans la dernière région une distribution exactement log-normale, mais non-stationnaire, juste au dessus de la précédente. Les limites entre les régions sont calculées analytiquement et les résultats sont confirmés par des simulations numériques. Dans les 3 cas, les corrélations se révèlent fondamentales pour la détermination de la distribution asymptotique finale, qui serait une Gaussienne partout sauf en région 1). En outre elles peuvent être une explication pour plusieurs distributions observées dans les systèmes naturels ou artificiels, dont la plupart des exposants mesurés appartient à l'intervalle d'exposants possibles de notre modèle.

En augmentant le niveau de complexité pour la modélisation d'avalanches, nous proposons ensuite un modèle de réseau aléatoire Erdös-Rényi pour modéliser le CSK, en identifiant le nœuds en tant que filaments d'actine et les liens en tant que cross-linkers. Sur cette structure nous simulons la propagation d'avalanches de ruptures. Dans un premier temps nous n'incluons pas la visco-élasticité des cellules vivantes, en supposant que les ruptures sont immédiates (et donc une probabilité p de fracture ne dépendant pas du temps). Ainsi, nous obtenons 3 régimes d'avalanches. (i) Un régime où les avalanches sont interrompues très rapidement, et leur taille suit une distribution décroissant plus vite qu'une loi de puissance ; (ii) un régime de grandes avalanches explosives endommageant le réseau en entier et (iii) un régime où les tailles des avalanches suivent une loi de puissance, qui sépare les 2 régimes précédents. Dans un deuxième temps nous avons introduit une probabilité de fracture dépendant du temps $p(t)$. De cette façon notre modèle de réseau aléatoire adapté reproduit une distribution approximativement log-normale des tailles d'avalanches, similaire à celles observées dans les expériences. Nos simulations montrent que l'on peut reproduire une statistique log-normale avec deux concepts simples : un réseau aléatoire sans échelle d'espace caractéristique et une règle de rupture capturant la visco-élasticité des cellules vivantes, propriété solidement observée. Ce travail ouvre la voie pour des applications futures à plusieurs phénomènes dans les systèmes vivants qui contiennent de larges populations d'éléments individuels, non-linéaires (cerveau, cœur, épidémies,...), où des statistiques log-normales similaires ont été observées.

Contents

1 Introduction **1**
- 1.1 Complex systems: an epistemology point of view . . . 1
- 1.2 What can biologists learn from complex systems? . . . 5
- 1.3 Statistical distributions are important in science: a focus on the log-normal distribution . . . 9
  - 1.3.1 Common statistical distributions . . . 9
  - 1.3.2 The log-normal distribution . . . 12
- 1.4 Interpretation and modelling of avalanche processes . . . 16
  - 1.4.1 The Galton-Watson branching process . . . 17
  - 1.4.2 Avalanches of infections in epidemics . . . 19
  - 1.4.3 The Random Field Ising Model . . . 22
- 1.5 Thesis outline . . . 23

2 A minimal model for living cells plasticity **25**
- 2.1 Cytoskeleton mechanics . . . 26
- 2.2 Atomic Force Microscopy and its use in biology . . . 28
  - 2.2.1 AFM working principle and calibration . . . 29
  - 2.2.2 FICs recording . . . 31
- 2.3 Singular event detection and characterisation . . . 33
- 2.4 Experiments on human cells . . . 36
  - 2.4.1 Ductile *versus* brittle rupture events . . . 37
  - 2.4.2 Log-normal statistics of released energy during rupture events . . . 38
- 2.5 A cascade model for living cells plasticity . . . 41
- 2.6 Comparison with experimental data and model interpretation . . . 45
- 2.7 Discussion . . . 50

3 Phase diagram of avalanche size distributions for processes evolving with a continuous random reproduction rate **51**
- 3.1 The recurrent problem of log-normal variable sums . . . 51
  - 3.1.1 Beaulieu's theorem: presentation and critical analysis . . . 54
  - 3.1.2 Some useful results about the sum of uncorrelated log-normal random variables . . . 57
- 3.2 Model definition and general behaviour . . . 59
  - 3.2.1 Phase diagram removing correlations . . . 63
- 3.3 Beaulieu's theorem approach . . . 64

3.3.1 Comments . . . 69
3.4 Statistical characterisation of the Z distribution . . . 70
3.5 A recurrence point of view . . . 75
3.6 Divergent case . . . 82
3.6.1 Case $\tilde{a} = 1$ . . . 83
3.6.2 Case $\tilde{a} > 1$ and $\bar{a} < 1$ . . . 87
3.6.3 Case $\bar{a} = 1$ . . . 90
3.6.4 Case $\bar{a} > 1$ . . . 93
3.7 The role of correlations: about the surrogate model without correlations . 101
3.8 Discussion . . . 107

4 Emergence of log-normal type distributions in avalanche processes on networks 111
4.1 Avalanches on a network structure: is self-organised criticality exhaustive? 111
4.2 Evidence for log-normal statistics in cell mechanics . . . 114
4.2.1 The fractional rheology of the cytoskeleton network . . . 114
4.2.2 Rupture event statistics from two primary cell lines . . . 116
4.3 A random network model for the cell cytoskeleton . . . 118
4.4 Avalanche statistics with constant probability of breaking . . . 121
4.5 Avalanche statistics by introducing visco-elasticity . . . 124
4.5.1 Avalanche statistics with local restoring of cross-linkers . . . 128
4.6 Avalanche propagation along the weaknesses of the network . . . 128
4.7 Discussion and future directions . . . 130

5 Conclusion and perspectives 132

A Supplemental material of Chapter 2 139

B Table of parameters for simulations of Chapter 3 143

Chapter 1

Introduction

1.1 Complex systems: an epistemology point of view

This is a thesis of physics of complex systems, where the complex system is the cell, with its structure, its motility, its cytoskeleton, its ability to reproduce, in one word living! Therefore it is a multidisciplinary thesis embracing also biophysics, biostatistics and numerical analysis, among others. It is impossible to imagine to deal with such a complex system with tools coming from a unique discipline, which requires in turn a strong team work and necessitated even an epistemological approach.

In this context I think it is important to ask ourselves what we are really doing and what we are looking for, in a general, I would say systemic, way. To tackle these questions we need to keep some distance from the specific work or the particular experiment, and enter in a deeper, kind of philosophical, thought. I believe that this process is important for every physicist, even every scientist, and probably everyone has his own answers, because a «true» answer is maybe impossible to achieve. Today the high specialisation of science makes more difficult to take this distance, but to understand complex systems this is necessary. Indeed, in order to find out the exchange of information taking place in a living organism, within itself and with its environment, and the different behaviours at different scales, we need a *global* point of observation. Here I deal with these problems from a complex systems point of view, giving my personal vision.

First, let us introduce and discuss the definition of complex systems, which is already not an easy task. Historically, we can say that the first appearance of complexity is with the deterministic chaos from Poincaré at the end of the 19th century (Poincaré, 1890). First with the attempt to find a solution of the three-body problem, Poincaré showed that a completely deterministic system can lead to chaotic behaviour, for example via period doubling. The complexity is in the fact that despite the deterministic origin of the system, its behaviour cannot be forecast because of non-linear terms in the ordinary differential equations of Newtonian mechanics. Epistemologically, this raised fundamental questions, because knowing the mathematical formulation of the problem (Newton equations) does not guarantee its prediction, since the system can yield chaotic behaviour. It was therefore an evidence that the scientist dealing with these systems can only do a classification, a phase portrait of the disorder, of the chaotic behaviour (Polizzi, 2013). Later on, Lorenz

(Lorenz, 1963), in the context of meteorology and weather forecasts elaborated on the dependence of a dynamical system with non-linearities on the initial condition, giving rise to the famous butterfly effect. In a century we observed the transition from the simple and well calculable universe of Galileo, Newton and Laplace, to a universe of unexpected unpredictable paths.

If Poincaré introduced the first ideas of complexity in Mathematics, giving origin to the field of dynamical systems, in physics, almost contemporary, complexity appeared at the beginning in the context of neural network modelling, with James (James, 1890) and then later, with more mathematical rigour, with the works of Mcculloch and Pitts (McCulloch and Pitts, 1943). But even though these ideas of complexity were already present since many years in an unstructured way, only in the lasts 20-30 years they were accepted in the physics community as a science (the first research institute of complex systems, the Santa Fe Institute, was founded in 1984), giving rise to the physics of complex systems. A complete historical description of complex systems is not our purpose here, let us only cite an example of complex system that will be useful to introduce the main characteristics of complexity: the Ising model (Ising, 1925). With this example in mind let us move to the definition of complex systems, or, maybe better, to some possible definitions.

In general, we refer to complex systems as systems in which interactions between the objects composing the system, and/or between the system and its environment, are important and give origin to collective behaviour. Complex systems are not necessarily *complicated*: a normal every day pendulum can be considered as a complex system just taking into account the interactions between the pendulum and the environment (friction and an external applied torque), or in interaction with other pendula. Its complexity is given by the fact that varying the control parameter, in this case for example the applied torque, can lead to complex behaviours like period doubling and chaotic oscillations, which are not predictable, in the sense that we cannot have a trajectory of the pendulum indicating the precise position at a given time. This is a complex behaviour that goes out from standard classical mechanics physical tools and therefore needs more adapted statistical and physical instruments to be studied.

At first sight, this can be a good definition, but could lead to the wrong conclusion that, since all the objects are connected with all the other objects of the system, and even with the observer, these complex interactions may lead to an impossibility of a complete knowledge of the system behaviour. The essential fact here, as we will see better later, is that the scientist is himself an *active* part of the system, which builds representations, models, interpretations, and not only a passive observer. As expressed by Licata (Licata, 2011), the theoretical description, built on our choices, is necessary to give a meaning to vague observations.

This makes the definition of complex systems complicated, therefore it is better to discuss some key properties of them. The most important property, that we did not exploit yet, is *emergence*: complex systems are systems in which interactions between a multitude of objects and/or with the environment lead to emergent collective properties which are not directly explainable by the properties characteristic of each element. The phase transition undergoing in the Ising model, for example, is a collective effect not explainable only with individual spins properties. Life itself is an emergent property, try

to mix together 70 kg of hydrogen, oxygen and carbon, shake well, and you will see it starting running around and writing PhD theses.

Let us describe better what emergence is. The first appearance of the idea in the physics world was with P. Anderson with his famous *More is Different*, in 1972 (Anderson, 1972), stating that the formalisms and the concepts needed to understand physical (and in general scientific) phenomena at a given scale are not always linked to the ones at lower scales, and not from them achievable. This was against the dominant reductionist idea (for which everything can be explained starting from a basic, low scale, law) predominating at that time, and even probably nowadays. Besides, he noted a general lower, and in any case different, degree of symmetry while looking at the system at a larger scale. Therefore, the laws of microscopic physics cannot always explain new phenomena emerging at larger scales, for which an adapted theory capturing the essence of the phenomena has to be created. The laws of objects composed by a large number of individuals, in particular living systems, cannot be deduced uniquely from the laws of particle physics, as the reductionist approach would predict. Notably, the lower degree of symmetry observed while increasing the complexity of a system, allows us to say that life can be seen as a breaking of symmetry effect. There are many examples of this, sugar molecules produced by living systems have all a **R** (for right) configuration, while in principle **R** and **S** (*sinister*, latin for left) configurations have the same energy and should be present in the same amount. The same happens for many chiral molecules and cells, like sperm cells, for which chirality is essential for life and which can move in their environment only thank to this symmetry breaking, otherwise the *scallop theorem* would not allow them to move at low Reynolds numbers (*i.e.* at normal life conditions) (Purcell, 1977). In one sentence, emergence is a continuous novelty production in an essentially unpredictable way.

These ideas of emergence were already present at a philosophical level with the idea of *new categories*, ontological entities with a hierarchical organisation needed to describe interactions with different *strata* at least since the late 19th/early 20th century with N. Hartman (Hartmann, 2012), or also J. S. Mill or C. D. Broad, but only in the last 20-30 years were accepted in the physics community (more or less at the same time as the definition of physics of complex systems as a science).

Another key feature of complexity is the definition of the border between system and environment. Here the active choice of the scientist comes into play: to build a model on an aspect of nature we make some assumptions on this border, on the interactions between the system and the environment. These changeable assumptions are the most important active part contribution of the scientist. The definition of them leads to different emergent properties and the modelling of different aspects of the system. In the *Middle Way*, citing the Nobel prize Robert Laughlin, standing between the physics of particles and the cosmological theories, there is the realm of incertitude, of randomness, where nature expresses a game of probability resulting from the competition between freedom and constraints. This does not mean at all that we cannot do science, but in contrast to a classical Newtonian universe, where the observer records events resulting from predefined universal laws, allowing in principle for a full prediction of the system, here the active observer has to look for a global comprehension. He has to do a global picture of the possibilities, without being able to predict which one will be realized. For example the

process of protein folding can happen in a myriad of different fashions with exactly the same energy level and the one finally chosen cannot be predicted. In the same way in an Ising system we cannot predict the exact state of the system at a given time, we cannot say which orientation spin i will have at time t, but only say that at some critical temperature a collective behaviour will arise.

In this realm, reductionist approach cannot explain this diversity, nor these emergent properties, but this is not because there is something wrong in it, simply, in these situations, it does not work. The scientist creates a variety of models, not necessarily all convergent in a unique vision, to describe different levels and different behaviours of the systems. Finally, a complex system is a system which is unpredictable, and not reducible to a single formal model, to a single *theory of everything*.

Now it should be clearer what a complex system is and the issues of a scientist studying it. Let us then focus further on the epistemological side of these issues. A direct and common answer to the epistemological problems settled at the beginning, would be a circular vision between experiment and theory, a kind of experimentalism of Galilean memory: *sensate esperienze e necessarie dimostrazioni* (sensible experiences and necessary proofs), in which the experimental evidence builds the theory, the theory generalises the results, inducing new experiments to verify its consistence. In some cases it is sufficient to stop here, and «keep calculating». But after a deeper epistemological analysis, of relevance in particular for complexity given what we said about the observed/observer interactions, this vision would have at least two problems. First, what would be the starting point of the circle? Theory or experiments? We are tempted to say experiments, since physics is an experimental science, but then there would be another question, can an experiment exist without a theory? The answer is: not really. This leads us directly to the second problem of this circular vision. Is the experiment true independently of the framework in which it is operating, therefore independent of the theory or of the tradition (the social structure)? A theory is an (unstable) equilibrium state between the experiment and the observer, but is not unique and never complete. As we said, in complex systems science we select an aspect of our observation and we model this aspect under a certain hypothesis, a theory of all is here not even conceivable, essentially because of emergency. Therefore maybe a more adapted point of view is the one of Pierre Duhem, who was coincidentally a professor here in Bordeaux. His holistic vision states that experiments and theories are connected to conventional principle which can change during time (Duhem, 2016). The connection to the active observer needed for complex systems is evident, and also the idea that natural phenomena are not pre-existing facts ruled by a unique formula that once discovered will predict everything. A theory and an experiment can be true in a certain set-up, at a certain scale, but could not work at others. So there is no such a thing like a *crucial experiment* allowing us to discern a good from a bad theory. The experiment itself is defined within a set-up, under some hypothesis, ultimately by our cognitive structure.

In this regard you may have thought about the observation of a quantum system, as one of the most evident interaction observed/observer. Therefore, it is very interesting to discuss briefly the idea presented in the nice book edited by I. Licata and A. Sakaji (in which it is worth to mention the articles by Pessa (Pessa, 2008) and by Licata (Licata, 2008)) of a systemic science based on quantum or quantum field theories applications to

phase transition in biological matter, supported by the indissoluble connection between emergent properties and the observer, the scientist himself. Indeed, this connection observer/observed can be thought to have a link with quantum mechanical properties, in which an observation causes the irreversible collapse of the wave function. However, as pointed out by Pessa himself, the success of this quantum biological theory is still very partial, mostly because while the particles in quantum theories are all considered as identical (if, of course they have same charge, mass, etc ...) the variability of living beings is in striking contrast with that. Moreover many complex processes studied from a statistical point of view (like the Moran model for evolution genetics (Moran, 1958) or processes on networks) do not have an evident correspondent Hamiltonian from which one could start a quantum approach, and even if we could build one approximated, we would need an out of equilibrium generalisation of the quantum theories. Also, many biological concepts, like, with the example of evolution models, the fitness, and the environmental effects are, if not impossible, very difficult to be tackled with a quantum field formalism.

The richness of complex systems is given by the fact that they are not linked to a particular physical model, which would be confined in a particular domain of science, like for example gravitation or other physical theories, it is rather a result of physical and mathematical research as a whole. This is related to the fact that complex systems deal mostly with the mesoscopic realm, a realm where physics meets many other disciplines and macroscopic and microscopic descriptions melt together. As a matter of fact, the range of applications of the physics of complex systems is very large. Thinking for example of the theory of deterministic chaos and non-linear theory, which are part of the physics of complex systems, applications go through meteorology, electronics, optics, thermoconvection, chemical reactions, biology and even astrophysics. Its transversal property, creating links, connecting together scientific domains traditionally very far apart from each other, is a unifying factor of science itself and of theory with applications.

To conclude, when we start observing outside well defined ideal conditions, the famous spherical cow, we often have to face complexity. That is because interactions with the environment become important and change themselves as the system evolves, therefore the border itself between the *system* and the *environment* becomes difficult to be defined, leading to an active choice of modelling a particular phenomenon, made by the observer, and leading then to complexity.

In this work we will try to model some aspects of a complex system such as a living system, maybe the system of the highest degree of complexity, and give a global picture of possibilities, without aiming at a deterministic prediction of events.

1.2 What can biologists learn from complex systems?

There are many examples where the physics of complex systems gave important insights on biological systems, and helped to better understand them. We have already mentioned protein folding and chiral motion of cells (as sperm cells, or some bacteria). It is worth to mention, for its historical relevance, also the Lotka-Volterra model, describing the prey-predator competition in simple, but already informative mathematical terms with

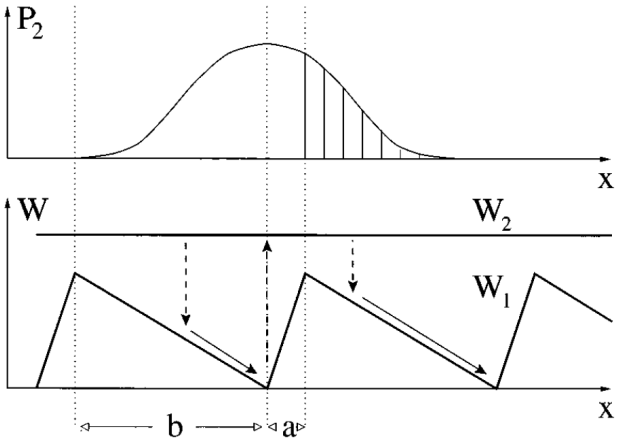

Figure 1.1 – Schematic view of the two states ratchet model, the higher, constant potential (W_2) is the Brownian diffusion state, and W_1 is the asymmetric ratchet periodic potential. Arrows symbolise stochastic transitions between the 2 states, where up pointing arrows need an active energy injection to jump to the upper state. Adapted from (Jülicher et al., 1997).

important implications on ecosystem science (Berryman, 1992). In its simplest version two continuous non-linear differential equations are coupled, to represent the time evolution of both prey and predator populations. Under some assumptions, actually realistic only in ecosystems isolated from other effects and where all the other conditions - weather, temperature, availability of food ... - are constant over the time considered (it is just the simplest version of the model), it can be shown that there are two fixed points of the dynamics. One is the extinction of both populations, and the other is an oscillatory dynamics, with a feedback regulated mechanism: the more prey means the more food for predators, implying a growth of the predator population. Despite its strong assumptions this model was already interesting for the understanding of ecosystems, helping to take decisions on regulatory politics for nature preservation, in particular after human alteration.

More closely related to our system, the living cell, we can cite important works proving the existence of long-range correlations in genomic DNA packaging (Arneodo, Bacry, Graves and Muzy, 1995; Huvet et al., 2007), or works on tissue growth showing self-developed homeostatic stresses (Alford et al., 2008; Ambrosi et al., 2011). The homeostatic stress is the steady stress proper to growing living systems, as biological tissues, arising from the non-equilibrium state of the system, balancing apoptosis and cell division and its regulation is essential in many pathologies, like cancer (Amar and Bianca, 2016). This is an important aspect involved in *mechanosensitivity*, which appears to be a property shared by all cells of the human body and all *phyla*, from mammals to plants, fungi and bacteria (Orr et al., 2006). Diffusion effects in crowded environments, such as cytoplasm or nuclei, are also fields where physics gave a good contribution. Non-standard diffusion exponents have been revealed, different from the standard Brownian motion due to crowding and hydrodynamical back-reflection effects (a molecule moving in a liquid creates a flow which is reflected from other molecules of the same size) (Parry et al., 2014).

Here, I would like to discuss shortly the modelling of molecular motors, active proteins

responsible for transport of vesicles or nutrients along cell cytoskeleton filaments and in general of many other active features of the cell. This is an example of highly out of equilibrium system with an interesting physical interpretation. First, it should be noticed that at the nano-scale (molecular motors typically move of a few tenths of nanometres per step and apply loads of a few pN) viscosity dominates inertia and the relatively high, with respect to molecular motors power, thermal noise makes standard motor motion impossible. In this context a simple symmetric Brownian ratchet, *i.e.* a passive motor subject to thermal fluctuations would not lead to a net directed force. The first step to a solution of the problem is to reproduce the symmetry of the filament in an appropriate potential landscape (see potential W_1 in Fig. 1.1), where we can find the periodical structure of the filament to which the motor is attached and, at the same time, the filament polarisation towards one of its end, giving an asymmetry in the sawtooth. Again, it can be proved (Feynman et al., 1966) that yet it is not sufficient to have a direct movement, as intuitively we could think: a particle falling randomly on this potential landscape could be expected to have a drift to the right. Actually, what is really important is how the system is driven out of equilibrium, therefore how energy (generally hydrolysis of ATP) is used to switch from the state described by potential W_1 to the free diffusion potential described by the constant potential W_2, coupling in this way the 2 states (Jülicher et al., 1997) (up pointing arrows in Fig. 1.1). Within this picture we can find the transition rates between the two states that optimise directed movement, arriving at the conclusion that there should exist active sites localised along the filament which promote transition from state 1 to state 2. This seems to be supported experimentally, by studying the experimental velocity curves with respect to ATP density (Cross et al., 1988; Woehlke et al., 1997).

Another field where complex systems and stochastic processes have successfully contributed, is population genetics. In this field theoretical models have a huge database of information represented by the famous Lenski's experiment on 12 populations of *Escherichia Coli* evolving at constant nutrient-poor conditions since 1988 (Good et al., 2017). The popular Moran model (Moran, 1958) gives a stochastic description of evolution, following the path of the pioneering models from S. Wright (Wright, 1931) and R. A. Fisher (Fisher, 1930), but introducing individual random births and deaths, allowing for a better mathematical description. In the simplest version of it, the most important parameter governing the dynamics of asexually reproducing individuals is the individual fitness. Without going into the details of the model, we can say that the passage to a mathematical description of evolution was essential to the wide acceptance of Darwin and Wallace theories from the scientific community and took more than a century to be partially achieved (there are still important open problems). An important result is that the mean population fitness increases under selection and the rate of fitness increase is proportional to the amount of genetic variability of the population. Furthermore this description helped in the explanation of genetic drift (long-term fluctuations of the genetic expression of the population), genetic fixation (the probability that a genetic feature dominates others) and evolution dynamics. These models are related to a branching process (see Section 1.4.1), which is another class of stochastic models originally created to explain the extinction of a population, but without a genetic point of view.

Finally, I briefly discuss implications of neural network theory to the understanding of

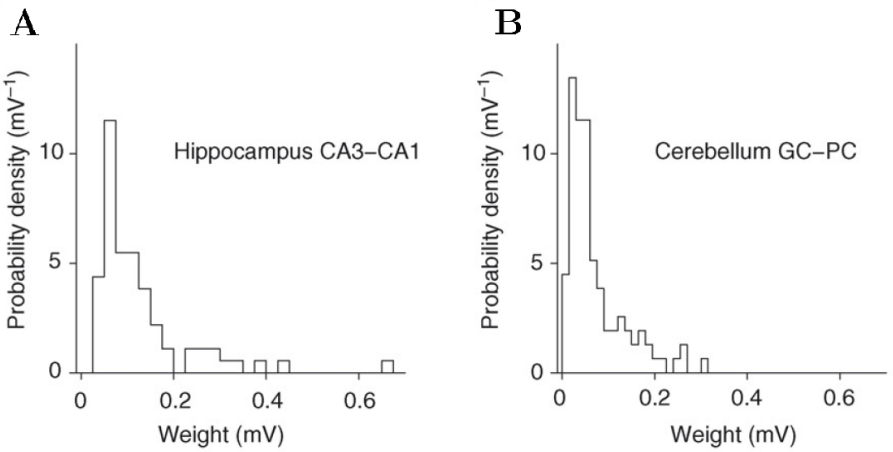

Figure 1.2 – Synaptic weights distributions recorded experimentally in two different areas of the brain: hippocampus (**A.**) and cerebellum, granule cell-Purkinje cell synapses (**B.**), showing strongly asymmetric distributions close to a log-normal. **A.** Adapted from (Sayer et al., 1990) and **B.** adapted from (Isope and Barbour, 2002).

brain mechanisms, a field which I am particularly close to from my experience. This is still a very active subject, since understanding brain mechanisms can be very difficult, but we can cite a few findings obtained by a statistical mechanics/complex systems analysis. A good picture of what is going on in such a complex network is provided by the statistics of some available data, for example the synaptic weights. Since so far it has not been possible to observe dynamically a single synapse weight change, then a theoretical description can help to understand the underlying mechanism and to infer some properties, as storage capacity, of real neural networks. Looking at the distributions in Fig. 1.2 we can see that in different areas of the brain (similar distributions are observed also in cortical networks), the synaptic weight distribution has a skewed form which can be approximated to a log-normal, in fact it has been fitted by a log-normal by Song *et al.* (Song et al., 2005), without considering a large number of silent synapses, up to 60% (Rumpel et al., 1998) (not shown in the figure). Theoretically, this has been associated to memory optimisation in neural network: the condition of maximum storage of memories, together with the constraint of positive synaptic weights (*i.e.* excitatory synapses), leads to a large proportion of silent neurons and to a truncated Gaussian distribution for the synaptic weights (Barbour et al., 2007). Behind such ideas there are mechanisms driving the brain to an attractor where memory would be optimised. If optimality has not yet been reached, the decay of the distribution for large weights could be much slower than Gaussian. Other optimality principles, for instance considering the energetic cost of maintaining excitatory synapses (Varshney et al., 2006), lead to similar conclusions. In both cases it was not necessary to specify any details on the plasticity rule, that could bring to a more precise identification of the final distribution. Caution should be adopted with this evolution-driven optimality and with the idea of evolution itself, remembering what we said in Section 1.1, are not complete unique theories which

can explain everything, we should not forget that we are dealing with a complex system. The quantity to be optimised can change with time and even with the observed scale, we do not face an equilibrium system.

In general, we can say that the point of view of complex systems helps to interpret and explain some observations that otherwise would be considered in biology as unexpected events or noise. We think for example of extreme events, or the observation of asymmetric distributions, considered as atypical with respect to the common normal distribution. Moreover, having a global phase diagram of some aspects of a biological system, helps to understand what are the control parameters that can trigger non-trivial collective behaviours essential for life. To conclude, in all the discussed situations it is now clear that a deterministic description is not even conceivable, because stochastic and out of equilibrium processes are dominant in living systems. Also, we can say that in general we can infer much interesting information on underlying processes just by looking "critically" at statistical distributions or time variations of observable quantities.

1.3 Statistical distributions are important in science: a focus on the log-normal distribution

1.3.1 Common statistical distributions

Probably the most common statistical distribution in science is the normal (or Gaussian) distribution. Its fame is mostly due to the Central Limit Theorem, which states, in the simplest of its formulations, that the sum of small independent random errors results in a Gaussian distributed random variable, for any distribution of the errors, provided finite means and variances. The normality of the errors is also the assumption always made in statistics whenever you want to fit (meaning usual least squares regression) some functions or estimate some parameters, even though this is not always checked and verified. A good part of statistical modelling and machine learning is based on the idea that the errors are produced by the superposition of many small and independent unpredictable events, leading to a Gaussian error.

In physics, the normal distribution is strongly related to thermodynamic equilibrium, being for example the Maxwell-Boltzmann distribution of velocity vectors of molecules in a gas, or the distribution of energy fluctuations in canonical ensemble formulation. Generally speaking, the Gaussian distribution is related to equilibrium physics, because it is the distribution that maximises entropy for a given average energy (or other conserved quadratic quantities), like it is done in the canonical ensemble.

The Gaussian distribution is a two parameters distribution, where the parameters are: the mean, usually noted with Greek letters like μ and the variance, usually noted σ^2. Its support is all the real line, from $-\infty$ to $+\infty$. It has the interesting property to be stable under the sum operation, therefore the sum of independent Gaussian distributed random variables is still a Gaussian, with as mean the sum of the means and as variance the sum of the variances. Mathematically the normal distributions is defined by the following

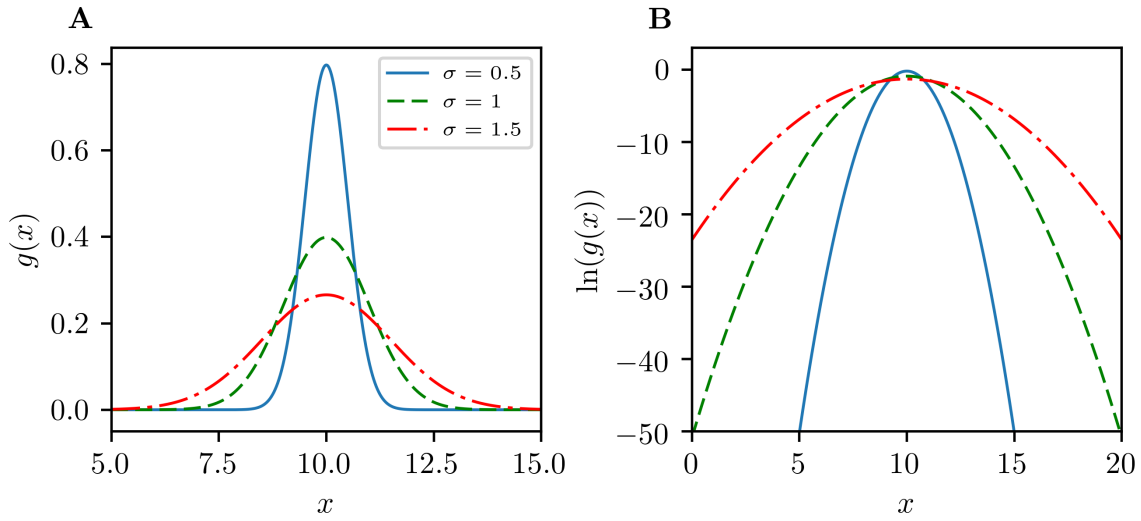

Figure 1.3 – **A.** The Gaussian distribution for different values of the variance σ^2: $\sigma^2 = 0.25$ (solid blue), $\sigma^2 = 1$ (dashed green) and $\sigma^2 = 1.56$ (dot-dashed red). The mean is set to $\mu = 10$ for all plots. Notice that the distribution becomes wider as σ^2 increases. Also the peak is more sharp for small values of σ^2 because of normalisation. **B.** Logarithm of the distribution $g(x)$ for the same parameters values as in panel **A.**, the shape of the normal distribution is a parabola in this representation. Notice that only the $y-$axis is logarithmic.

equation:

$$g(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}. \tag{1.1}$$

In Fig. 1.3 we show the shape of the bell curve for different variances σ^2. The value of μ is set equal to 10, because a change in μ results in a shift on the distribution, but does not change its shape. Therefore, under the change of variable $x' = x + \kappa$ the distribution of x' is just $g(x + \kappa)$, where $g(x)$ is the Gaussian distribution defined above. As far as the scale parameter σ gets larger, the distribution get wider and more peaked. Finally, as shown in panel **B** notice that in a semi-logarithmic representation, the shape of the normal distribution $g(x)$ is parabolic.

Another distribution often encountered in nature is the Pareto (or power-law) distribution, which has the peculiarity to have a long power-law tail, spanning several decades. In physics, the importance of this distribution is its appearance of power law distributions in systems close to criticality, for example close to the ferromagnetic-paramagnetic phase transition of the liquid-gas phase transition. In these cases long-range correlations effects arise and many physical quantities, like the heat capacity, the magnetic susceptibility or the correlation length diverge with a power-law behaviour with respect to the control parameter. More generally the Pareto distribution is closely related to the Zipf's law (the discrete version of it) and to the family of Lévy $\alpha-$stable distributions, which do not have a close analytical expression, but have a Pareto tail for large values of the random variable.

The exponent of the power-law tail has to be negative and smaller than one, to be able to normalise the distribution for $x \to \infty$. Mathematically, this tail cannot span all the

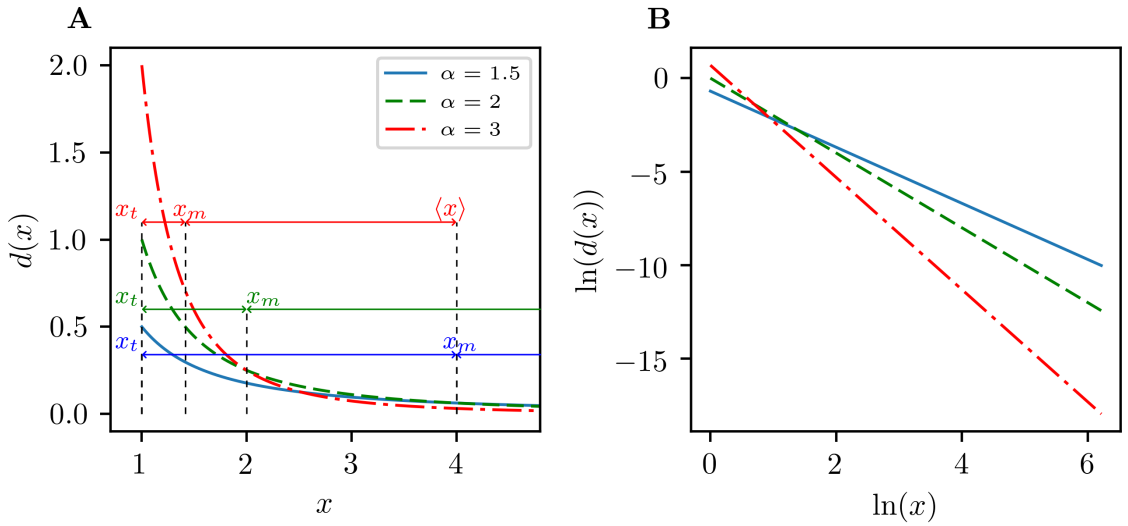

Figure 1.4 – A. The Pareto distribution for different values of the exponent α: $\alpha = 1.5$ (solid blue), $\alpha = 2$ (dashed green) and $\alpha = 3$ (dot-dashed red). The minimum value is set to $x_{min} = 1$ for all plots. Notice that the tail becomes wider as α decreases. The values of the mode (x_t), the median (x_m) and the mean ($\langle x \rangle$) of the distributions are represented. For $\alpha = 1.5$ and $\alpha = 2$ the mean is not defined, the result of the integral being $+\infty$, this is a consequence of the fat tail of the distributions. **B.** Logarithm of the distribution $d(x)$ for the same parameters values as in panel **A.** with respect to $\ln(x)$, the shape of the normal distribution is a straight line in this representation. Notice that both axes are logarithmic, but the plotted distribution is still the distribution of x, and not of $\ln(x)$.

support of the distribution, because otherwise the power law would not be normalisable, therefore the distribution is defined only for all $x > x_{min}$, with x_{min} positive. The shape parameter is the absolute value of the slope of the power law tail α, with $\alpha > 1$.

This leads to the definition of the Pareto distribution as:

$$d(x) = \frac{\alpha - 1}{x_{min}^{1-\alpha}} x^{-\alpha}, \tag{1.2}$$

where the factor $(\alpha - 1)/x_{min}^{1-\alpha}$ is the normalisation constant.

The statistical mode of the Pareto distribution (the maximum of the distribution), is always for $x = x_{min}$. The median, the point dividing in two parts of equal area the distribution is $x_m = x_{min}\ \sqrt[\alpha-1]{2}$ and the average, defined only if $\alpha > 2$, is $\langle x \rangle = \frac{\alpha-1}{\alpha-2} x_{min}$. These quantities can be found by direct integration of the probability density function, the mean for example is just:

$$\langle x \rangle = \int_{x_{min}}^{\infty} x d(x) dx \propto x^{-\alpha+2} \Big|_{x_{min}}^{\infty} < \infty \qquad \text{if} \qquad \alpha > 2, \tag{1.3}$$

where we can see explicitly that if $\alpha \leq 2$ this value diverges. In the same way it can be shown that the variance of the Pareto distribution diverges if $\alpha \leq 3$.

In Fig. 1.4 we can see the shape of the Pareto distribution and its statistics for different values of α. Notice that for $\alpha \leq 2$ the mean value is not represented, since would be $+\infty$. The distribution are all extremely right skewed, since the mean is always largely to the left with respect to the median. In panel **B** we plotted the same distributions as in panel **A** in a log-log plot: the distribution takes now the shape of a straight line, as can be

easily seen by taking the logarithm from both sides of (1.2). For skewed, asymmetric Lévy distributions (with asymmetry parameter $\beta = 1$), x_t is slightly larger than x_{min} and therefore the distribution has a peak separating the power-law tail from the frequent small x values (see (Penson and Górska, 2010)).

Finally, we notice that the power-law distribution is the only function which does not change shape by changing the scale of its variable, thus is also sometimes referred to as scale-free distribution. In mathematical terms this means that $d(x)$ is the only function verifying (Newman, 2005):

$$d(bx) = f(b)d(x), \tag{1.4}$$

for every constant b, where $f(b)$ is just an overall multiplicative factor not depending on x.

1.3.2 The log-normal distribution

In this section we introduce a distribution that will be in the background, if not the protagonist, of the whole thesis: the log-normal distribution. We have already seen how many informations can be deduced just by observing statistical distributions of available data, therefore we can understand how it is important to have a precise idea of the distributions shapes with respect to their parameters and how they can be generated. We have seen that the log-normal distribution can be a model for synaptic weights, but there are many other biological and physical quantities that are observed to follow a log-normal distribution, at least approximately. Positive quantities, when the mean value is low and the variance is large, often follow skewed distributions, that in most cases is either a log-normal or a power-law (Paretian) distribution. Variables reported to closely fit the log-normal distribution are, among others, the length of latent periods of infectious diseases, the species abundance and the amount of mineral ressources in the earth croast (Aitchison and Brown, 1957). Many chemical concentrations also follow the log-normal distribution as the hydroxymethylfurfurol in honey (Limpert et al., 2001), or the triglyceride concentration in human blood (Choi, 2016).In biology as well there are many quantities to be well fitted by the log-normal distribution: the distribution of the population of living organisms, such as bacteria, starting from the same original number of individuals (Limpert et al., 2001), which is related to a multiplicative process, as we will see hereafter; but also the mechanical properties of a population of living cells (Laperrousaz, Berguiga, Nicolini, Martinez-Torres, Arneodo, Satta and Argoul, 2016; Polizzi et al., 2018). Maybe the most famous examples of the log-normal are in the field of finance and economical sciences, starting from the first model of stock-option prices, the Black-Scholes stochastic model (Black and Scholes, 1973). Actually, we can say that the log-normal distribution, judging by its wide presence in nature, is at least as «normal» as the normal distribution.

The log-normal distribution is defined only for positive random variable, as the

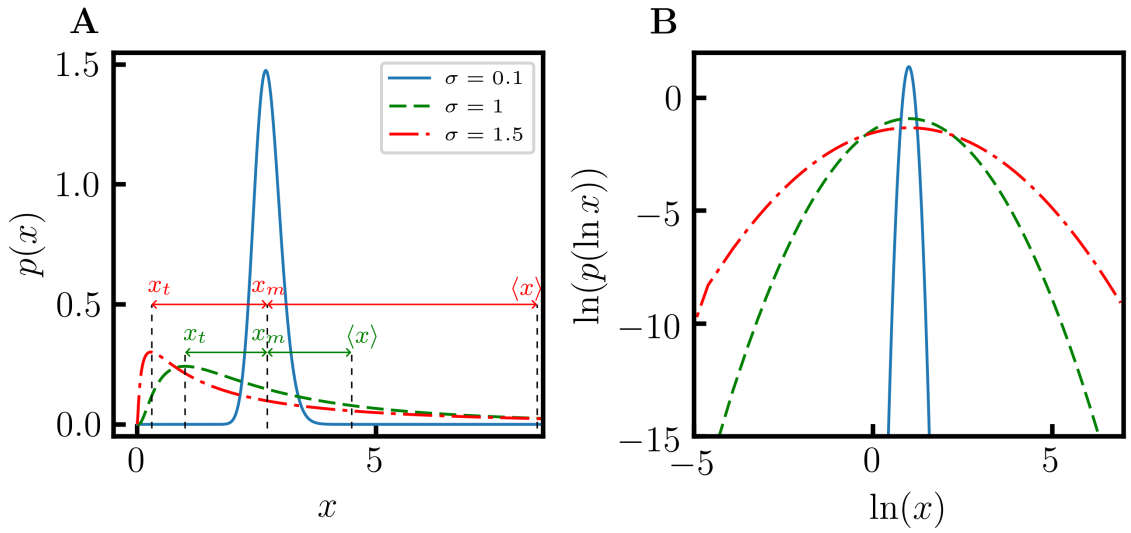

Figure 1.5 – Typical shapes of the log-normal distribution for different values of the parameter σ. For all curves $\mu = 1$. **A.** Log-normal distributions in lin-lin scale, for $\sigma = 0.1$ (solid blue), $\sigma = 1$ (dashed green) and $\sigma = 1.5$ (dot-dashed red). For $\sigma = 1$ and $\sigma = 1.5$ the values of the mode (x_t), the median (x_m) and the mean ($\langle x \rangle$) of the distributions are represented. For the $\sigma = 0.1$ distribution the three statistics coincide with the peak. **B.** Distribution of the logarithm of a log-normally distributed random variable x in a log-log plot. The parameters values are the same as in **A.**

exponential transform of a Gaussian random variable. Therefore, if:

$$\ln(x) \sim \mathcal{N}(\mu, \sigma^2), \tag{1.5}$$

the random variable x follows a log-normal distribution, with parameters μ and σ^2. We can find out the expression of the log-normal distribution by applying the substitution $z = \ln(x)$, implying $dz = dx/x$. Let us define $p(x)$ as the probability density function (p.d.f.) of x (the log-normal distribution) and $g(z)$ as the p.d.f. of z (the normal distribution). By the conservation of probability, we get:

$$p(x) = \frac{g(\ln(x))}{x} = \frac{1}{x\sqrt{2\pi\sigma^2}} e^{-\frac{(\ln(x)-\mu)^2}{2\sigma^2}} \qquad \text{for} \qquad x > 0. \tag{1.6}$$

This can be seen also as the first generating process of a log-normal distribution: the exponential transform of a normal random variable. The shape of this function is plotted in Fig. 1.5 for different values of the shape parameter σ. The value of $\mu = 1$ is fixed, because it can be noted that a change in μ is equivalent to a scaling factor in x, therefore by an appropriate choice of x units we can always set $\mu = 1$. This can be seen by doing the change of variable $x' = \kappa x$, leading to a distribution of x' equal to $p(x')$ (with $p(x)$ defined above), but with a new parameter $\mu' = \mu + \ln(\kappa)$.

Let us now define the statistics of the distribution (Johnson et al., 1994). The mean value $\langle x \rangle$ is given by

$$\langle x \rangle = e^{\mu + \sigma^2/2}, \tag{1.7}$$

the statistical mode (the maximum of the distribution) is

$$x_t = e^{\mu-\sigma^2}, \tag{1.8}$$

and the median, the x value for which the area under the distribution is divided in 2 equal parts, is

$$x_m = e^{\mu}. \tag{1.9}$$

Besides, the variance of the log-normal distribution is:

$$\mathrm{Var}(x) = e^{2\mu+\sigma^2}\left(e^{\sigma^2}-1\right), \tag{1.10}$$

in general all the raw moments can be computed, for the moment of order n we have:

$$\mathbb{E}[x^n] = e^{n\left(\mu+n\sigma^2/2\right)}. \tag{1.11}$$

These quantities can all (apart from x_t, which corresponds to the maximum of $p(x)$) be computed by means of Gaussian integrals, considering that $x = e^z$, with z normally distributed. The last equation implies that all the moments exist and are finite, but the log-normal distribution has the interesting property that the knowledge of all its moments does not define uniquely the distribution: there exists a whole family of distributions with the same moments (Johnson et al., 1994).

Interestingly, we can notice that the equivalent of the standard deviation for a normal random variable, is here $\sigma^* = e^{\sigma}$, meaning that the interval $[x_m/\sigma^*; x_m\sigma^*]$ contains 68.3% of the probability, as the interval $[x_m - \sigma; x_m + \sigma]$ for a normal variable (for a normal variable $x_m = \mu$). This justifies the interesting proposal of (Limpert et al., 2001) to give results or estimations of log-normally distributed random variables in terms of the multiplicative interval as before, and not, as often read (*e.g.* (Stehmann and De Waard, 1996)) in term of the arithmetic mean and standard deviation. It would actually be more representative of data, giving the idea of the multiplicative scheme (detailed hereafter) of the log-normal distribution.

We can observe that the log-normal distribution shares properties of narrow distributions together with properties of broad ones. In fact, We can recognise different characteristics of the distribution with respect to σ:

1 – for $\sigma \ll 1$ the distribution is very narrow, and can be approximated by a normal distribution (see solid blue line in Fig 1.5**A**). Here the mode x_t, the median x_m and the mean $\langle x\rangle$ of the distribution collapse all to the same value (e), as it is for a symmetric distribution;

2 – for $\sigma \lesssim 1$ the distribution starts to be skewed to the right, therefore we have the relations $x_t < x_m < \langle x\rangle$, as we can see from the dashed green curve of Fig. 1.5**A**. The distribution starts having a broad tail, but still, as expected for narrow distributions the peak value $p(x_t)$ decreases while increasing the variance;

3 – by increasing $\sigma > 1$ we can see (dot-dashed red line in Fig. 1.5**A**) two interesting

behaviours: the tail of the distribution becomes very broad ($\langle x \rangle \gg x_t$), and the peak value of the distribution $p(x_t)$ starts increasing again. This is typical of Levy's distributions, which have a very fat power-law tail and a divergence close to 0 (in 0 actually, as the log-normal, they cannot be defined).

In panel **B** we show the distribution shape of $\ln(x)$ for the same parameters as in panel **A** in a log-log plot. We can see that in this representation the distribution looks like a parabola. This is useful to recognise the log-normal distribution, especially from a power-law tailed distribution, that, as we have seen, leads to a straight line at large values of x in the same type of representation. However, we should say that in real data is not always easy to distinguish a log-normal from a power-law tailed distribution, mainly because of finite size effects or of intrinsic cut-off related to the studied system (for example the distribution could have a minimum allowed value cutting the distribution), requiring advanced statistical tools and generating ongoing debates on the most appropriate distribution to model the data (Broido and Clauset, 2019).

Before moving on, we would like to spend some words on the statement done for case 1 ($\sigma \ll 1$), about the Gaussian approximation for small values of σ. We can easily see this by a Taylor expansion of $p(x)$ for small σ. By definition $\ln(x)$ is Gaussian distributed, so we can write it with respect to a standard normal variable $\epsilon \sim \mathcal{N}(0, 1)$:

$$\ln(x) = \mu + \sigma\epsilon \qquad \Rightarrow \qquad x = e^{\mu+\sigma\epsilon} \simeq e^{\mu}\left(1 + \sigma\epsilon\right), \tag{1.12}$$

where the last approximation is valid in the limit $\sigma \ll 1$. We can see thus that x also follows approximately the same statistics as ϵ, therefore a Gaussian with a mean value of e^{μ} and a variance of $e^{2\mu}\sigma^2$. This demonstrates that a log-normal distribution for small σ is very close to a Gaussian distribution, with a certain mean and variance related to the parameters of the log-normal distribution.

It is worth to point out also another different effect that can arise while playing with real data, and which makes also sometimes difficult to distinguish a normal from a log-normal distribution. Starting now from a normally distributed random variable we will show that in some cases its logarithm is also normally distributed, and therefore the original variable distribution can be confused with a log-normal. If we consider the distribution of the logarithm of a normally distributed random variable, again making use of the auxiliary random variable ϵ, we can see that it is still normally distributed if $\sigma/\mu \ll 1$. Consider $z \sim \mathcal{N}(\mu, \sigma^2)$:

$$\ln(z) = \ln(\mu + \sigma\epsilon) \simeq \ln(\mu) + \frac{\sigma}{\mu}\epsilon, \tag{1.13}$$

where the last approximation is not only true if $\sigma \ll 1$ (where we already know that a log-normal and a normal are very close each other), but more generally if $\sigma/\mu \ll 1$. Therefore if $\sigma/\mu \ll 1$ a Gaussian variable can appear as log-normal because the mean value is large with respect to the standard deviation. Remark that the opposite conclusion is not true: a log-normal random variable with a large $\mu \gg \sigma$ independently of the value of σ cannot be confused with a normal random variable.

The most important generating mechanism often cited in the literature to explain

log-normally distributions is the Gibrat's law (Gibrat, 1931; Sutton, 1997), first conceived to model firm growth. The basic idea is that if x_i is a random variable depending on the time step i (for example the size of the firm at time i) in such a way that:

$$x_i = a_i x_{i-1}. \tag{1.14}$$

The process starts from an initial size x_0, with the assumption that the random growth (or shrink) factor of the firm a_i is independent of the size of the firm at any time (a_i and x_i are all independent). The growth factor is thus expressed as a percentage of its size at the previous time step, or in other terms the percentage of growth (shrink) of the firm in one time step ($x_i - x_{i-1}$) is given by $a_i + 1$. If all the a_i are log-normally distributed it is straightforward to see that x_i, being the product of all a_i is also log-normal, because the product of log-normal random variable is a log-normal. This can easily be seen by applying the logarithm from both sides of Eq. (1.14), which can be written as a product, exploiting the recursion of the process:

$$x_i = x_0 \prod_{n=1}^{n=i} a_n \qquad \Rightarrow \qquad \ln(x_i) = \ln(x_0) + \sum_{n=1}^{n=i} \ln(a_n), \tag{1.15}$$

where the last sum is the sum of Gaussian random variables, which is known to be Gaussian. However, the log-normality of x_i is valid more generally, even if the a_i are not log-normal. Indeed, we can apply the Central Limit Theorem to the last sum to prove that $\ln(x_i)$ will converge to a Gaussian, therefore x_i to a log-normal, for $i \to \infty$. In order to use the Central Limit Theorem we need not only the a_i to be independent, but also that they have some regularity, like bounded mean $\bar{a}$ and variance Δa^2 (they do not have to be constant, but we consider this case for simplicity). These conditions are usually verified in applications, leading to x_i log-normally distributed with parameters $\mu_i = i\bar{a}$ and $\sigma_i^2 = i\Delta a^2$. For a better approximation at finite i see (Redner, 1990). We notice that this limit distribution for x_i is not stationary, since its parameters depends on time i, therefore also the shape of the distribution, in general, depend on i, because at the observation time i the distribution may not have converged yet.

1.4 Interpretation and modelling of avalanche processes

In this section we review some useful models of avalanches. In the following chapters we will see what are the differences between these common models and the required characteristics for modelling avalanches in living systems, which will be our main motivation. In particular in the next chapters we will focus on avalanches of mechanical fractures of the cythoskeleton of living cells. Nevertheless, in general, an avalanche process is appropriate to model any mechanism in which there is a jump of information transfer, in systems, typically out of equilibrium, where units have non-linear behaviour beyond a certain threshold. A unit reaching threshold causes other connected units to do the same in turn. We can think again to processes going on in the brain (Beggs and Plenz, 2003), during earthquakes

(Gutenberg and Richter, 1956), during epidemics (Cai et al., 2015) or even simply friction dynamics in granular systems. All these examples have in common to be composed of equal threshold units with the same non-linear behaviour. Usually, the system is driven out of equilibrium by an external input, or by some energy production of the system itself.

1.4.1 The Galton-Watson branching process

In this section we introduce the Galton-Watson branching process of which the model that we will study in Section 2.5 and in (Polizzi et al., 2018) can be seen as a generalization. This is maybe the first basic model for avalanche, or cascade, processes.

A Galton-Watson branching process is a stochastic process named after F. Galton and H.W. Watson (1874) even though the smart solution given by Watson was incomplete and brought him to an erroneous conclusion.

At the origin, this model aimed to explain why so many peer families were extinguishing. Let consider a sequence $\{p_i\}$ of probabilities of having i sons and do the hypothesis that every son reproduces independently with the same sequence of probabilities. The main question the model had to answer was then: what is the probability that a family extinguishes (*i.e.* that the number of sons will be zero) after generation r?

It turned out that the real reason of the peer families extinguishing was the tendency of peers to marry heiress, this is today known to reduce drastically the genetic heritage and to lead to major diseases. This may be the cause of the loss of interest for this model for around 60 years although later in the 20th century it has been successfully applied to many scientific fields such as chemical reaction cascades, survival of populations or genes, gene mutations, nuclear chain reactions and so on.

Let us now define the problem formally, following the treatment of (Harris, 1963) with some extensions. Let us consider a population of reproducing individuals (animals, plants, bacteria, or even computer viruses), starting from a single individual at time $n = 0$. Each individual lives for a single generation, corresponding to one time step. At time $n = 1$ it produces a family of offspring and immediately it dies. The number of offspring is a positive integer number that is called the *number of Young*, Y. Individual i has a number of offspring Y_i. At time $n = 2$ each offspring produces his own family of offspring and immediately dies. And so on for $n = 3, 4, 5 \ldots$. Let us call the size of the population at time n, Z_n, the branching process is then defined by the sequence $\{Z_n\}_{n\in\mathbb{N}}$ and by the discrete probability distribution of Y, $\mathbb{P}\{Y = k\} = p_k$. We now have to do 2 important assumptions in order to solve the problem:

1) All individuals reproduce independently of each other. Therefore the number of offspring an individual has does not depend on how many other individuals are in the system, then there is no interactions between individuals of the same generation;

2) The family sizes Y_i of individual i are identically distributed discrete random variables distributed as the number Y for all i and do not depend on time n. Therefore the size of the generation n, called Z_n only depends on the size of generation $n - 1$ and not on the previous ones. This makes the sequence $\{Z_n\}_{n\in\mathbb{N}}$ a Markov chain.

The more delicate assumption is the first one, mainly in biological systems, where individuals interact both with the environment and within each other. For example thinking about animals, they will tend to have less offspring as far as the population size increases. The branching process can however be a good model also for this kind of systems if we stay far from saturation and we mainly focus on the fast growing (or decreasing) phase in which interactions can be neglected. We shall come back to these remarks later on, since they are necessary also when dealing with our generalization of the branching process.

Remark: Generalization to $Z_0 > 1$ is easily made, since individual's families behave independently of one another. Then $Z_0 > 1$ will just correspond to a time shift of the process.

We will now give the main results of the Galton-Watson process. For first let us introduce the probability generating function f:

$$f(s) = \sum_{k=0}^{\infty} p_k s^k \qquad s \in \mathbb{C} \quad \text{and} \quad |s| \leq 1 \tag{1.16}$$

We can write the branching process as a randomly stopped sum of independent random variables. The size of the population at time n thanks to assumption 1) and 2) can be written as:

$$Z_n = \sum_{i=1}^{Z_{n-1}} Y_i \tag{1.17}$$

Now we can compute the generating function $f_n(s)$ of the randomly stopped sum Z_n noticing that $f(s) = \mathbb{E}\left[s^{Y_i}\right]$. So:

$$f_n(s) = \mathbb{E}\left[s^{Z_n}\right] = \mathbb{E}\left[s^{\sum_{i=1}^{Z_{n-1}} Y_i}\right] = \mathbb{E}_{Z_{n-1}}\left[f(s)^{Z_{n-1}}\right] = f_{n-1}(f(s)), \tag{1.18}$$

where the last step is given by the fact that Z_{n-1} is also a randomly stopped sum. This is actually an effect of the recursion contained in the definition of the model itself. The symbol $\mathbb{E}_{Z_{n-1}}[\,.\,]$ means the average over the random variable Z_{n-1}.

Using the same arguments for f_{n-1} we can find easily that $f_n(s)$ is the n-fold iterate of $f(s)$:

$$f_n(s) = \underbrace{f(f(\dots f}_{n\text{-times}}(s)\dots)) \tag{1.19}$$

With just this simple relation we can already compute the moments of Z_n, let $m = \mathbb{E}[Y]$ and $\sigma^2 = \mathbb{E}[Y^2] - m^2$. By observing the fact that $\mathbb{E}[Z_n] = f_n'(1)$ and $\mathbb{E}[Z_n^2] = f_n''(1) + f_n'(1)$:

$$\mathbb{E}[Z_n] = f_{n-1}'(1) f'(1) = m^n \tag{1.20}$$

by induction.

$$\mathrm{Var}(Z_n) = n\sigma^2 \qquad \text{for } m = 1 \tag{1.21}$$

$$\mathrm{Var}(Z_n) = \sigma^2 m^n \frac{(m^n - 1)}{m^2 - m} \qquad \text{for } m \neq 1 \tag{1.22}$$

by induction on the second derivative of the probability generating function computed in 1.

An other interesting result is about the probability of extinction, that was actually the major goal of the model. This probability is defined as the probability of the sequence $\{Z_n\}$ being 0 starting from a certain n:

$$q = \mathbb{P}\{Z_n \to 0\} = \lim_{n\to\infty} \mathbb{P}\{Z_n = 0\} = \lim_{n\to\infty} f_n(0), \tag{1.23}$$

where the second equivalence follows from the fact that if Z_n is 0 at some n it will be for sure 0 after and the third equivalence comes from the definition of f. There is a theorem (Harris, 1963) that states that q is one if $m \leq 1$, so intuitively if on average each individual is not able at least to replace itself the population will disappear. What is less intuitive (and indeed Watson did not notice it) is that there is a finite probability of extinction even if $m > 1$. From the recursion Equation (1.18) we can see that $f_n(0) = f_{n-1}(f(0)) = f(f_{n-1}(0))$ then taking the limit for $n \to \infty$ from both sides we find the definition of q. We have thus that q is the solution of:

$$s = f(s) \tag{1.24}$$

It can be proved that this equation has a solution even for $m > 1$, intuitively we can observe that $f(s) < s$ if s is slightly lower than 1 and if $p_0 + p_1 < 1$ but $f(0) > 0$, so there must be at least a solution between 0 and 1.

A last important result that is worth to point out here is that the branching process is an unstable process, in the sense that $\lim_{n\to\infty} Z_n$ can either be 0 or infinity, the sequence $\{Z_n\}$ cannot be positive and bounded, even though $m = 1$.

1.4.2 Avalanches of infections in epidemics

Another field where avalanche models have reached an increasing interest and success in the last $20 - 30$ years (with an acceleration recently for the Covid-19 pandemic (World Health Organization, 2020)) is avalanches of infections for epidemics spread. These models can be interpreted as sophisticated branching processes on a network structure. The power of them is that we can have a precise idea of the impact of the network structure on the avalanche and on the effect of the epidemics, helping to identify critical points separating small outbreaks from epidemics (Newman et al., 2001; Newman, 2002; Pastor-Satorras and Vespignani, 2001). This added complexity (the effect of the network structure) can be studied in principle for any type of network under some hypotheses, the most important and difficult to achieve being the absence of loop in the network, *i.e.* taking two random nodes there must be only one possible path of connections joining them.

However, for many applications the presence of loop is negligible, and the generality of these models allow to sketch the general behaviour of the epidemic, with respect to the epidemic parameters, like the average number of infection per unit time, the average

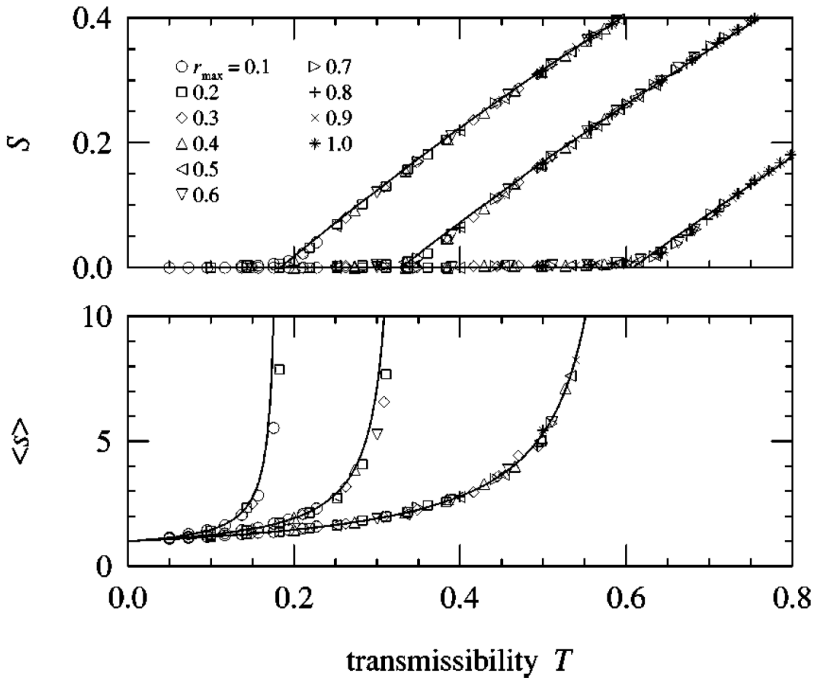

Figure 1.6 – Plot of the epidemic size (top) and of average outbreak size (bottom). The network chosen here has a degree distribution following a power-law of exponent α with an exponential cutoff κ: $p_k = Ck^{-\alpha}e^{-k/\kappa}$. The value of the cutoff are $\kappa = 20, 10, 5$ from left to right. Notice that in the limit of large κ the critical T approaches 0. Each point in the graph corresponds to a different rate of infection r_i and infectious time τ_i. Reproduced with permission from (Newman, 2002) ©(2002) by the American Physical Society.

recovering time, and so on. In particular, the avalanche of infections often relies on the SIR (Susceptible, Infectious, Recovered or Removed) compartmental model of epidemics, which runs on the network structure. SIR model is one of the simplest compartmental epidemics models, in which the population is divided into connected compartment: randomly chosen infected individuals (of course you need at least one infected individual to start the epidemic at the beginning) enter in contact with susceptible individuals at an average rate β per unit time and recover, or die, with an average rate γ per unit time. Although generalisation to more complex compartmental models may be possible and may have been done in the meanwhile, in this introductory section we do not consider this increased complexity. Nonetheless, we point out that such a model (SIR) is adapted only for rapidly spreading diseases which confer immunity upon recovered individuals, as measles. To make the link with avalanches of failures in solids or biologic materials, we can observe that in these materials avalanches can spread very fast as well, meaning with respect to the *restoring* capacity of the material, but it would also be interesting to introduce and study the effect of the capacity of the system to repair the failures on time scales comparable to the avalanche running time.

Let us define the process in more details. The network here is a network of connections, but not of contacts. The network represents the *possibility* to have a contact between two individuals. Therefore on this connection's network a SIR model can propagate in principle with parameters (such as rate of infection or infectious time) varying between individuals. It can be proved (Sander et al., 2002) that even though each pair of individuals $\{i, j\}$

has a different rate of disease causing contacts r_{ij} and each infected individual has a different infectious time τ_i the spread of the epidemics does not depend on the individual details, but on a single general parameters which is basically the integral over the whole network population of these parameters. This parameter, called the "transmissibility" in (Newman, 2002) (T) is the control parameter determining the transition between small outbreaks and epidemics and allowing us to forget about the individuals details and the heterogeneity of the population. What really matter in this modelling is the network structure and the parameter T, representing the probability of a bond in the network to be marked (*i.e.* carrier of infection). This is explicit in Figure 1.6, in which no matter the individual rates of infection and infectious time, the size of the epidemic, if it exists, collapses on the same curve.

Thus the epidemic can be seen as an avalanche process on a network: imagine to mark a bond on the network when an infection between two connected nodes occurs, causing in turn other connected individuals to be infected and so on. At the end you will have a picture with the history of infections resulting in a cluster of infected people, whose size is the size of the outbreak. By means of the theory of generating functions from percolation theory some interesting quantities of the epidemic avalanches can be computed analytically (Newman, 2002). Indeed, it can be proved (Grassberger, 1983; Sander et al., 2002; Newman, 2002) that an epidemic avalanche has some similarities with percolation on a graph, for example for the avalanche size distribution, even though the dynamics is completely different. The most important quantity that can be computed is the critical transmissibility separating the region where small outbreaks die out exponentially fast to the region where an epidemic exists and infect a finite fraction of the population. The critical point T_c is defined with respect to the degree distribution of the network p_k (the distribution of the number of connections k per node):

$$T_c = \frac{\sum_{k=1}^{N} k p_k}{\sum_{k=2}^{N} k(k-1) p_k}, \tag{1.25}$$

where N is the network size. We stress the fact again that this result is valid for networks not containing loops. From Eq. (1.25) it can be seen immediately that if the network is scale free, *i.e.* the degree distribution is a pure power-law, the denominator diverges in the large N limit, for exponents most commonly found in real networks, usually included in the interval $[2; 3]$ (Barabási and Pósfai, 2016; Barabási and Albert, 1999; Barabási et al., 1999). It means that in this case the critical point cannot be defined and all outbreaks will lead to an epidemic, no matter what is the "strength" of the virus. This has been pointed out by (Pastor-Satorras and Vespignani, 2001) in the case of computer viruses spreading on the internet, which always infect a finite fraction of the population. Figure 1.6 illustrates the fraction of the population affected by the infection S in the large N limit (top), and the mean outbreak size $\langle s \rangle$. We can observe that at the critical transmissibility, the mean average size diverges and at the same time the epidemic size increases above zero. We can see as well that for different degree distributions the value of T_c is different (from left to right the exponential cutoff is $\kappa = 20, 10, 5$), going to zero as p_k approaches a pure power-law distribution ($\kappa \to \infty$).

The topology of the network looks then crucial for the spreading of avalanches and

their characteristics. Moreover considering the network topology leads to completely different values of the critical point, with respect to the critical point resulting from standard compartmental epidemic models, which consider a fully mixed population.

1.4.3 The Random Field Ising Model

The Random Field Ising Model (RFIM) is a generalisation of the standard Ising model originally proposed for ferromagnetic to paramagnetic phase transition. At the origin RFIM was proposed by Sethna *et al.* (Sethna et al., 1993) in order to study hysteresis in magnetic systems and to model the Barkhausen effect. The Barkhausen effect is a high frequency noise generated by small jumps in the magnetisation observed for ferromagnets in oscillating magnetic field. It was observed experimentally that magnetisation jumps occur like avalanches phenomena, and can span almost 3 decades of sizes, following a power-law tailed distribution. Similar avalanches mechanisms are revealed also for functional and amorphous materials, like stress-induced plastic deformations in martensites or ferroic materials (Salje et al., 2017). In the latter cases what is propagated along the avalanches is a fracture of the system, whose sizes usually show a power-law behaviour.

The RFIM is a system of two states spins $s_i = \pm 1$ which are situated at the vertices of a given lattice, which could be regular, random, or in general a network of connections. Exact solutions are possible only in few cases, where the network is non recurrent, *i.e.* it does not have loops. For example exact solutions for the avalanche size distribution are possible on Bethe lattices or Cayley trees (Sabhapandit et al., 2000). The RFIM is ruled by the following Hamiltonian:

$$\mathcal{H} = -J \sum_{\langle i,j \rangle} s_i s_j - \sum h_i s_i - H \sum s_i, \tag{1.26}$$

where the first negative term is an interaction between nearest neighbours spins (indicated by the symbol $\langle i, j \rangle$) promoting alignment. The second term is the quenched random field, represented by a random variable h_i drawn from a distribution $p(h_i)$: to every spin site a different realisation of the random field is applied. Finally the last term is the interaction with an external field, for example a magnetic field H, promoting the orientation of spins along the direction of the field.

Usually, the system is driven by quasi-statically changing the external field H, with a dynamics implying a spin flip only if this will lead to a decrease in the Hamiltonian (Sethna et al., 1993). This is also called *nucleation* dynamics. To explain better the dynamics let us consider to drive slowly the field H from $-\infty$ to ∞. With quasi-statically we mean that for each variation of the value of H the system has the time to reach equilibrium, and then all spin flips that would lead to a decrease in the energy have the time to flip. At some value of H therefore the energetic configuration given by (1.26) would be more favourable if some spin flips. The flip of even a single spin yields a local change of the Hamiltonian and therefore may cause other neighbour spins to flip. At this point an avalanche starts, until the system has reached again equilibrium.

In Fig. 1.7 we show the possible shapes of the avalanche size distribution for a RFIM on a cubic 3D lattice. Here the system is driven with a nucleation dynamics and the control

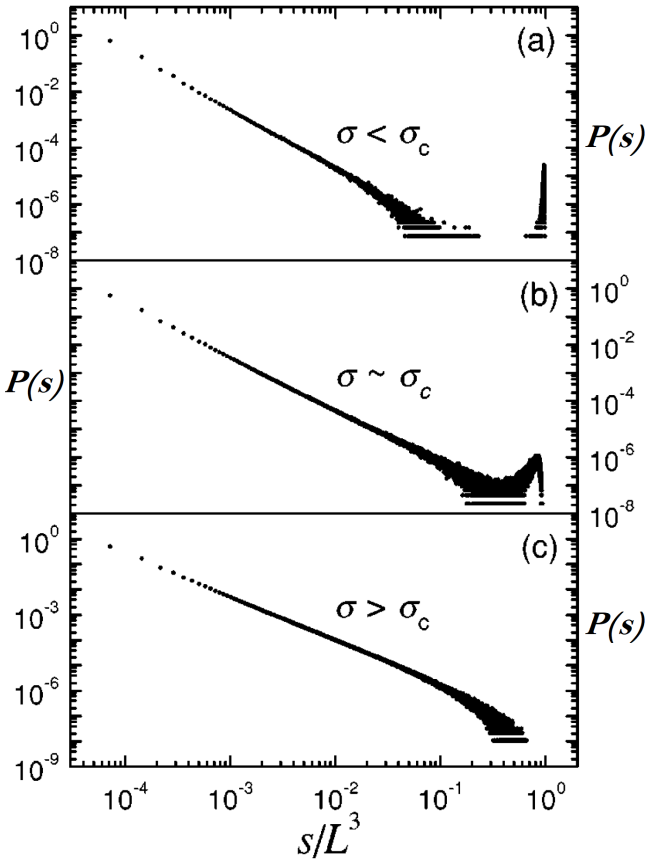

Figure 1.7 – Possible avalanche size distributions $P(s)$ resulting from a RFIM driven with nucleation dynamics. The random field h_i is chosen here as Gaussian distributed with 0 average and standard deviation r. The simulations are run on a 3D cubic lattice. By decreasing r, from panel (a) to panel (c) different avalanches statistics are observed (see text), and when r is close to a critical value r_0 the distribution is a power-law. The value of the critical standard deviation of the noise r_0 is here $\simeq 2.2$. Adapted from (Pérez-Reche and Vives, 2003).

parameter being varied is the noise of the system r. More precisely r is the standard deviation of the distribution $p(h_i)$ of the random field h_i, here considered as Gaussian. As described in (Pérez-Reche and Vives, 2003), for low values of the random field standard deviation $r < r_0$ the system shows big avalanches spanning all the system size (the peak on the right in panel (a)), these avalanches are called *snaps*. While increasing $r > r_0$, the system shows small avalanches of a few spin flips, called *pop* (panel (c)). In between the two regimes there is a critical point, for $r = r_0$ where a power law decay of the avalanche size distribution is observed. This means that the distribution follows a decay of the type $s^{-\alpha}$ at the critical point. The exponent α depends on the type of lattice and the dimension of the system, for the case of Fig. 1.7 is $= 1.6$. On Bethe lattices with different coordination numbers it is found to be exactly $3/2$ (Sabhapandit et al., 2000).

1.5 Thesis outline

The outline of the thesis is as follows. In Chapter 2 we discuss experiments done previously to this work on CD34+ primary blood cells, and we analyse experimental results. We

compare cancer cells with healthy cells mechanical properties. Moreover we give a first minimal (1D) model describing the mechanism of cytoskeleton avalanches of ruptures, which offers an interesting interpretation of the observed differences.

In Chapter 3 we analyse more thoroughly this avalanche 1D model and compare all the possible resulting distributions of the avalanche sizes. We show then a complete phase diagram allowing us to generalise the model to other interesting avalanche processes, more general than the one studied in Chapter 2.

Finally in Chapter 4 we propose a random network model of rupture avalanches reported in Chapter 2. In this case the model is more elaborated, considering avalanches taking place on a network structure, mimicking the cytoskeleton complex network. We also show a phase diagram for the avalanche size distributions, that precisely gives in which parameters region avalanche size distributions similar to experiments are obtained.

The thesis is concluded with a discussion and perspective Chapter, in which we resume our main findings and describe open problems and some possible strategies for a deeper understanding and investigation.

Chapter 2

A minimal model for living cells plasticity

In this chapter, following our publication (Polizzi et al., 2018) we use a nano-indentation (AFM) technique (see Section 2.2) (Radmacher, 2002; Mahaffy et al., 2004; Azeloglu and Costa, 2011; Abidine et al., 2013; Haase and Pelling, 2015) to experimentally investigate the nonlinear mechanical plasticity of single immature $CD34^+$ haematopoietic cells from healthy donors and patients suffering from Chronic Myelogenous Leukaemia (CML) (a blood cancer) in chronic phase (CP) harvested at diagnosis (Laperrousaz, Berguiga, Nicolini, Martinez-Torres, Arneodo, Satta and Argoul, 2016). Classical analysis of force-indentation curves (FICs) aims at estimating an elastic modulus by fitting the curves with linear elastic models (Hertz, 1882; Sneddon, 1965). Here, in contrast, we propose a wavelet-based multi-scale method (Laperrousaz, Berguiga, Nicolini, Martinez-Torres, Arneodo, Satta and Argoul, 2016; Digiuni et al., 2015) to identify singularities in the FICs that likely correspond to rupture events in the CSK. Our study (Polizzi et al., 2018) provides compelling evidence of the existence of two distinct populations of avalanche rupture events, namely *ductile* (corresponding to weakly cross-linked filaments) and *brittle* (corresponding to tightly cross-linked filaments) failures. Both mechanisms display fat-tail distributions of released energy well approximated by log-normal distributions. This is surprising given the ubiquity of power-law statistics for avalanches in solids (Song et al., 2013; Chuang et al., 2013; Salje et al., 2017). We develop a minimal model that reproduces quantitatively the experimental released energy distributions, and provides some mechanistic interpretation of both the ductile and brittle rupture regimes. Despite this phenomenological model does not take into account the visco-elasticity of individual polymer chains constituting the CSK filaments, it sheds a new light on the local unbinding events as a major mechanism underlying the nonlinear response of living cells to large deformations and it further shows that brittle failures are more frequent in CML cells as the signature of their higher mechanical fragility.

2.1 Cytoskeleton mechanics

In this section we introduce the main characteristics and properties of cell mechanics, pointing out the relevant aspects for our future modelling. We focus here on cytoskeleton mechanics, since it was the first motivation of our modelling, but many characteristics are common to other glassy materials, as far as we do not include in the modelling the active characteristic of living systems. Therefore similar conclusions can be applied to amorphous glassy materials, like polymers, metallic glasses or colloidal glasses, which all share slow dynamics, and long mechanical relaxation delays (Ritort and Sollich, 2003; Radhakrishnan et al., 2017).

As the elementary building blocks of living systems, cells are active mechanical machines that constantly remodel their structural organisation to withstand forces and deformations and to promptly adapt to their mechanical environment (Hoffman and Crocker, 2009; Chauvière et al., 2010). This versatility is fundamentally required for many vital cellular functions, such as migration and mitosis, and an alteration of the cell mechanical properties can participate in pathogenesis and disease progression, such as cancer (Brunner et al., 2009; Kai et al., 2016). Concerning cancer, until the 70s of the 20th century, there was a predominating dogma stating that the progression of the pathology was driven by a minority of cancer cells being "leaders", and that the majority of other cancer cells were poorly active and did not play a role in metastasis (Ewing, 1928). This justified dominant therapies focusing on the destruction of these rare aggressive elements. More recently, an alternative model progressively emerged. The latter considers that a fraction of transformed cells is already present at a very early stage of the tumour, and an important proportion of these cells self-specialises and becomes able to divide (Olmeda and Amar, 2019), do a clonal selection and eventually create a tumour. These cell modifications are driven from the properties, including mechanical properties, of the environment, the niche (Peinado et al., 2017). Identifying under which conditions is the mechanical resilience of living cells compromised is therefore a critical issue and a systemic approach is needed.

The cancer type studied in this chapter (CML) arises from a haematopoietic stem cell transformation following the formation of the BCR-ABL oncogene by a single reciprocal chromosomal translocation t(9;22) (Savona and Talpaz, 2008). In CML, BCR-ABL$^+$ cells of the myeloid lineage proliferate uncontrollably, the bone marrow density increases considerably, and their mechanical properties change during disease progression (Jones, 2010). In transformed cells (Laperrousaz et al., 2013), BCR-ABL was shown to bind actin filaments, to inhibit their polymerisation and to disorganise the CSK into punctuate, juxtanuclear aggregates (McWhirter and Wang, 1993; Cheng et al., 2002; Laperrousaz, Drillon, Berguiga, Nicolini, Audit, Satta, Arneodo and Argoul, 2016).

Significant progress in the past decades has provided a rather complete picture of the linear mechanical response to small applied stresses or strains (Fabry et al., 2001; Hoffman et al., 2006; Gardel et al., 2008; Lieleg et al., 2010; Kollmannsberger and Fabry, 2011; Broedersz and MacKintosh, 2014; Rigato et al., 2017). However, cells are often subject to large deformations and reach non-linear regimes that are far from being well understood (Lieleg et al., 2010; Kollmannsberger and Fabry, 2011; Wagner et al., 2006; Broedersz

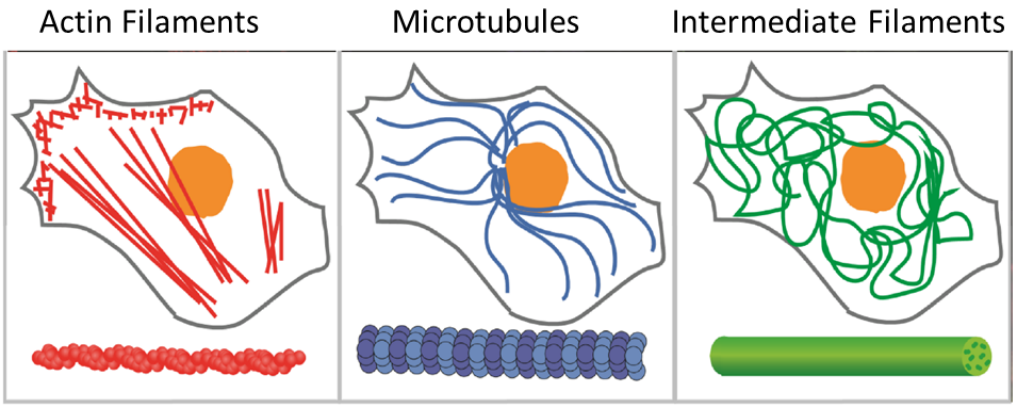

Figure 2.1 – Schematic view of the different cytoskeleton filaments. Courtesy of Ulrike Rölleke.

et al., 2008; Kollmannsberger et al., 2011). The fascinating mechanical properties of living cells are mediated by the cytoskeleton (CSK), a dynamic network of filamentous proteins composed of actin filaments, microtubules, and intermediate filaments (Gardel et al., 2008; Lieleg et al., 2010; Kollmannsberger and Fabry, 2011; Fletcher and Mullins, 2010; Huber et al., 2013; Blanchoin et al., 2014) (see Fig. 2.1). *In vitro* reconstitution of these filaments shows different mechanical properties (Huber et al., 2015; Li and Gundersen, 2008): i) the actin filaments have a persistence length (the length over which the filament can be considered as straight) of about 10 μm and are mostly involved in cell shape, migration and cell mechanics; ii) the microtubules have a persistence length of about 1 mm and so they are considered as rigid filaments at the cell scale. They are involved in mitosis and transport of nutrient, vesicles and organelles between cell compartments; iii) the intermediate filaments are on the contrary less precisely defined, being cell type specific, but their diameters is known to be around 10 nm, and therefore in between actin filaments one (7 nm) and microtubules one (25 nm), from which their name. Their persistence length is larger than 1 μm, they are rod shaped and they are involved in resistance to shear, but their role is not completely clear.

Let us now focus with more details on the actin cytoskeleton. These filaments are the most relevant for our modelling, since they cover the perinuclear zone of cells, which is the cell compartment poked by our experiments. They present a polar structure (Li and Gundersen, 2008), a quite fast, with respect to active processes, polymerisation rate (larger than one per minute) and a depolymerisation rate of the order of one per a few minutes (Tseng et al., 2005). Moreover they can build a cross-linked network whose mechanical properties depend in general on the cross-linker proteins density and on the network structure.

The actin filaments are cross-linked by a wide variety of actin-binding proteins (ABPs), which can be passive or active, *i.e.* activated by ATP (Adenosine Triphosphate) hydrolysis, the latter having a slower dynamics (time scale of tens of minutes) (Gardel et al., 2008; Lieleg et al., 2010; Kollmannsberger and Fabry, 2011; Fletcher and Mullins, 2010; Huber et al., 2013; Blanchoin et al., 2014; Wagner et al., 2006; Ehrlicher et al., 2015). By tuning

the proportions of passive and active ABPs, living cells can control their power-law (scale-free) CSK rheology (Gardel et al., 2008; Juelicher et al., 2007). Interestingly, cells exhibit both solid and liquid-like properties. Solid-like behaviour is associated with strongly cross-linked actin filaments which resist sliding and accumulate tension (fimbrin and fascin are compact cross-linking proteins that create parallel-aligned actin networks (actin bundles) and are found in stiffer membrane protusions of filopodia (Dos Remedios et al., 2003; Tseng et al., 2005)). In contrast, weakly cross-linking proteins produce actin filaments which slide more readily, enabling the network to flow as a liquid (α-actinin and filamin-A are less compact and form networks with more widely spaced and orthogonally aligned actin filaments (Esue et al., 2009)). All these actin cross-linking and/or bundling proteins work cooperatively or competitively, for instance fascin and α-actinin were recently shown to segregate into discrete bundled domains that are specifically recognized by other actin-binding proteins (Winkelman et al., 2016). Under nonlinear loading conditions, living cells can display apparently opposite behaviours ranging from stress stiffening mainly governed by filament and/or cross-linker nonlinear elasticity (Lieleg et al., 2010; Kollmannsberger and Fabry, 2011; Wagner et al., 2006; Broedersz et al., 2008), to stretch softening and fluidisation likely due to force-induced unbinding of ABPs (Lieleg et al., 2010; Kollmannsberger and Fabry, 2011; Kollmannsberger et al., 2011). This paradox can be solved by considering that, upon large deformations, the CSK of a living cell can undergo deep structural transformations such as the unfolding of protein domains, the unbinding of cytoskeletal cross-linkers, and the breaking of weak sacrificial bonds. All these structural changes are inelastic (non-reversible in a strict sense), they dissipate locally the elastic energy of the CSK network (structural damping) (Wolff et al., 2010; Gralka and Kroy, 2015). Unbinding events and bond breakings confer living cells with the unique ability to adapt to different mechanical situations by actively controlling the amount of stress stiffening and fluidisation (Fredberg et al., 1997; Wolff et al., 2010). At the same time, such events reduce the connectivity of the CSK and may result in permanent plastic deformations or even more dramatic irreversible failures (Lieleg et al., 2010; Kollmannsberger and Fabry, 2011) which, for instance, could be at the origin of the recently observed incomplete shape recovery of living cells after repeated creep (Bonakdar et al., 2016). These effects are reminiscent of those in cyclically loaded solids which can lead to fatigue-induced failure (Song et al., 2013; Chuang et al., 2013; Salje et al., 2017).

2.2 Atomic Force Microscopy and its use in biology

We describe here the experimental technique used for experiments on living cellsThe technique is called Atomic Force Microscopy (AFM). The most common use of this technique is for imaging small living systems, ranging from bacteria and isolated cells to tissues, with a nanometre resolution (Morris et al., 1999). In this chapter we are interested in a less common, but still quite spread way of using AFM: for doing force spectroscopy. This mode of usage of the AFM, is probably quite far from the usual idea that people have of microscopy, because it does not give an image of a sample. In contrast, it gives a measure of the mechanical properties of an object under a constraint, as for example a

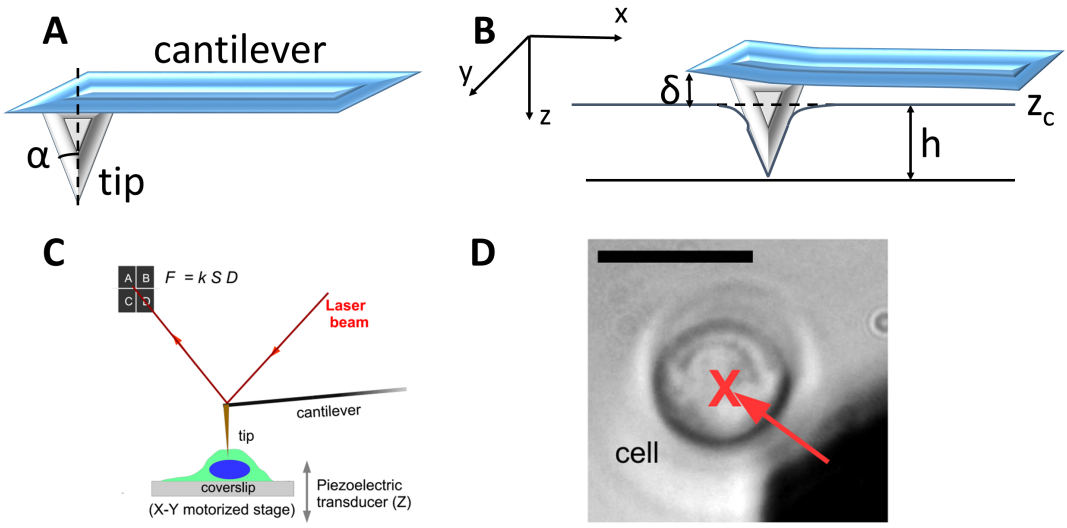

Figure 2.2 – **A.** Schematic view of cantilever and sharp tip of the AFM, with a half-opening angle α. **B.** Schematic view of an indenting tip, with a the sample deformation h and a cantilever deflection δ. **C.** AFM working principle and force transduction mechanism (see text for details). **D.** Image from transmitted light optical microscopy (coupled with AFM) of a CD34+ cell during the experiment. Scale bar 10μm.

constant strain rate. Doing a parallel with human sensing, if common microscopy can be associated with the view, force spectroscopy can be associated with the touch. The mechanical properties of the sample are deduced (in some way we will explain hereafter) from the Force Indentation Curves (FICs), *i.e.* the curves of the force exerted on the object in function of the indentation length.

2.2.1 AFM working principle and calibration

Let us now describe shortly the working mechanism of AFM. In our experiments we only focused on what are called in rheology relaxation measurements, or, in other words, we controlled the strain exerted on the material and let the material relaxing. Usually in relaxation experiments the strain applied at an instant $t = 0$ is constant, here we consider the case where the strain rate is constant, *i.e.* the strain is linear with time. Other rheological experiments are possible with the AFM, such as creep measurements or relaxation with different setups of the strain rate, but we will not describe them here. The probing mechanism of an AFM is composed, as represented in Figure 2.2**A** by a rectangular cantilever with stuck at one of its extremities a conical (or pyramidal) sharp tip with a half-opening angle α (of between $12° - 18°$ degrees). Of course, other shapes and configurations are possible but here our purpose is to indent significantly cells, and therefore we consider only sharp tips.

In Figure 2.2**B** is shown a cantilever indented on a deformed sample. The deformation length is h, corresponding to the non normalised strain on the material. The contact point is z_c, corresponding to the z at which the tip and the sample start being in contact. The response of the object to the applied strain causes a deflection of the cantilever δ. This deflection deviates the laser beam reflected on the top side of the cantilever extremity

corresponding to the tip (see Fig. 2.2**C**). At the reflected side of the laser beam there is a CCD fours quadrant photodiode which measures the deviation of the beam with respect to the rest position, giving a resulting signal D (as deflection) in Volts. At the same time a measure of the position of the base of the cantilever (the right side in Fig. 2.2**B**) is made by a piezoelectric transducer with a nanometre precision, giving the distance z, oriented towards the sample. At this point we have a curve of D against z: further delicate steps are still needed to have the force against the sample deformation.

First, the deflection D needs to be translated in a length. This is done previously to every experiment by indenting an infinitely (with respect to the cantilever) rigid surface (Meyer and Amer, 1988), such as glass. In this way we can safely suppose that the sample did not deform and then that all the measured deformation is due to the cantilever bending. Then we can obtain by the inverse of the slope of the D *vs* z curves the sensitivity S (typically in nm/ V). There exist other methods to estimate S (see Ch. 9 of (Morris et al., 1999) and (Higgins et al., 2006)) without putting the tip in contact with hard surfaces, possibly destroying its functionality. However, in our case this was done carefully and it was the most adapted and accurate way, since we did not need a particular functionalisation of the tip.

Once we have S we still need to transform the cantilever deflection distance into a force. Therefore we need the cantilever spring constant k. With the assumption that the cantilever behaves as a pure elastic material it can be computed by means of the thermal fluctuations method (Sader et al., 2014, 1999; Hutter, 2005). This method consists in measuring the deflection of the cantilever at a fixed z (far from the sample) only due to thermal random noise. In this way, using the equipartition theorem, the cantilever stiffness reads (Sader et al., 2014):

$$k_s = \frac{1}{\chi^2 \xi} \frac{k_B T}{\langle z_s^2 \rangle}, \tag{2.1}$$

where $\chi^2\xi$ is a factor taking into account the optical and mechanical differences between dynamic and static sensitivity and dynamic and static stiffness, depending both on the cantilever geometry. Indeed, to be precise, since in our experiments the cantilever tip will come in contact with the sample, the k needed is the static one, which is the one of Eq. (2.1), but, while measuring the thermal fluctuations, we are not in contact, therefore we need this correction factor. In the case of rectangular cantilevers, the ones we used in all the experiments, $\chi^2\xi = 1.224$, as has been computed by finite element calculations by (Sader et al., 2014). Besides, in Eq. (2.1) k_B is the Boltzmann constant and T is the temperature. Finally, $\langle z_s^2 \rangle$ is the mean squared position of the cantilever with respect to the neutral position. It is usually computed by the Fourier power spectrum of the cantilever deflection, which by Parseval's theorem is equivalent to $\langle z_s^2 \rangle$. Again the subscript s means static, since we computed the deflection distance by using the static sensitivity S. From now on we will avoid to put the subscript s, since there will be no possible misunderstanding, being all quantities consistent. At this point the force F can be computed:

$$F = kSD. \tag{2.2}$$

It is important to point out that all this calibration procedure (computing S and k) has to be done in the same medium (liquid or air) where the final measures are done, because the refractive index of the medium affect significantly the optical lever detection properties (affecting S) and the viscosity of the medium affects the mechanical deflection of the cantilever.

Once the calibration is done, we have to pay attention to a few further corrections. The first one is to correct z by the actual indentation length. This is needed, because the actual deformation of the material is h (see Fig. 2.2**B**) but the measured z includes also the cantilever deflection δ. Therefore to find h, we have to correct z according to (Cappella and Dietler, 1999):

$$h = z - z_c - \frac{F}{k}, \tag{2.3}$$

where F/k is equal to δ, *i.e.* to the deflection of the cantilever at distance z. From this equation we notice that the choice of the cantilever has to be a compromise between the force resolution (pushing towards a small k) and a small effect of the correction F/k, to be able to actually deform enough the sample (pushing towards a large k).

Moreover, a correction for the viscous drag is also performed, by estimating it in a far from contact region (Sader, 1998).

2.2.2 FICs recording

After all these steps we are able to plot a FIC. As we said previously, a constant strain rate relaxation experiment is done, meaning that the indentation velocity v_0, is constant. For the experiments we will analyse in the following sections $v_0 = 1\,\mu$m/s, but the effect of v_0 can be studied and was studied on fibroblasts in (Streppa, 2017). Starting from far (a few μm) from the sample the tip moves downward until the distance z_c is reached, where the tip-sample contact is reached. At this point the cantilever starts being deflected from the applied force, and the indentation lasts until a set-point (in Volts) is attained, therefore until a certain force is achieved. Then, instantaneously, the piezoelectric motor inverts its motion and starts moving upward. During this movement the cantilever is deflected because is pulled toward the sample by the sample-tip interaction. Both, the approach and the retract curves are recorded and can be informative. In Figure 2.3 we show an example of recorded FICs on a living myoblast taken from a previous PhD thesis of my hosting team (Streppa, 2017). In red are shown the approach curves and in green the retract curves, for different v_0.

By looking at Figure 2.3 we can introduce two important features for the analysis of the Force Indentation Curves. We can notice that at least for panels (a) and (b) the approach curves look parabolic. This is what is predicted from the Sneddon model, a model giving the force response to a given strain for a conical axisymmetric indenter (the tip) and for elastic isotropic materials (Sneddon, 1965):

$$F = \frac{4\tan(\alpha)}{\pi(1-\nu)} G h^2, \tag{2.4}$$

where α is the semi-opening angle and G is the shear relaxation modulus, related to the

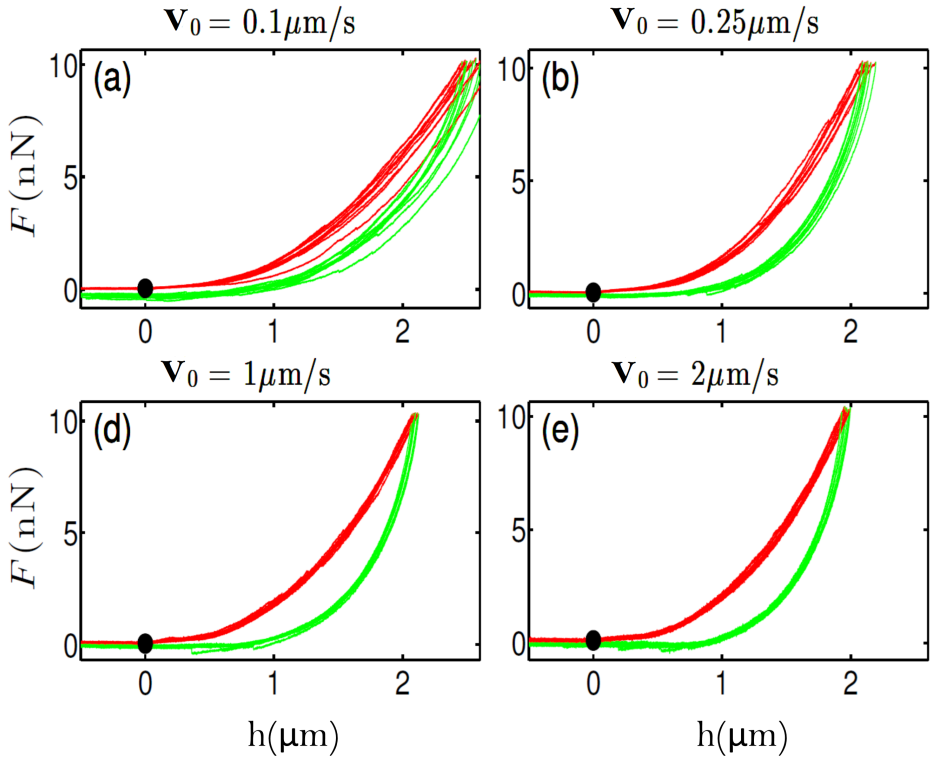

Figure 2.3 – Examples of indentation curves performed on a myoblast. For each panel 10 approach curves (red) and 10 retract curves (green) are done sequentially on the same cell. The strain rate (*i.e.* indentation velocity) is impressed above each panel. The black dot represents the contact point. Adapted from (Streppa, 2017).

1D Young modulus of the sample by the equation $G = \frac{E}{2(1+\nu)}$, with ν Poisson's ratio of the material. For cells, usually considered as incompressible and isotropic materials, $\nu \simeq 0.5$ (Sirghi et al., 2008), and therefore $G = E/3$. Classical analysis of FICs usually done is based on this modelling, allowing by a parabolic fit to find an estimate of the modulus G, or E.

In general visco-elastic materials respond to a time varying strain rate via a *hereditary* integral representation (Cheng and Cheng, 2004):

$$F(t) = \frac{4}{1-\nu} C_n \int_0^t G(t-\tau) \frac{dh^{(n+1)/n}(\tau)}{d\tau} d\tau \tag{2.5}$$

meaning that the actual state of the system at time t depends on all its history from the starting point of the deformation (at $t = 0$). The power n and the constant C_n depend on the particular shape of the indenter. For a conical indenter $n = 1$ and $C_1 = \tan(\alpha)/\pi$. In this chapter, we directly used this relation to model the response of cells, without needing to do the strong assumption of quadratic dependence of the force against the deformation as in Eq. (2.4). Recently, a more complete model taking into account possible holes in the membrane has been proposed (Jia and Amar, 2020), but here this effect is not important because we will focus on singular points of the FICs.

A second useful physical parameter can be introduced by observing Figure 2.3. You can notice that the approach and the retract curves do not overlap. This is the effect

of energy dissipation and it is a signature that cells are not purely elastic materials, in which case the two curves would overlap. This dissipation can be due either to viscous dissipation or plastic deformations, or a mixture of both. We can simply estimate the dissipation by computing the integral above the approach curve, W_a, which is the total work injected into the system by the indentation, and the integral above the retract curve, W_r, which is on the contrary the work given back from the system. The dissipation D in units of the injected work is then:

$$D = \frac{W_a + W_r}{W_i}. \tag{2.6}$$

Note that since the motion is inverted during the retract curve, W_r is negative, while W_a is positive.

As discussed in (Streppa, 2017) there are no adapted mechanical models combining viscous and elastic elements (such as the Maxwell, Kelvin-Voigt or SLS model) that are able to capture the visco-elastic response of living cells, as far as the velocity increases (as in panels (d) and (e) of Fig. 2.3), suggesting that an important role for dissipation is played by plastic deformations. It is therefore very likely that, under these conditions, myoblasts (and, as we will see soon, also other types of living cells) undergo local fractures in their cytoskeleton filamentous network, as far as the tip increases the load on the cell.

All the experiments of this chapter are done with a JPK AFM equipped with a CellHesion system with a $15 - 200\,\mu$m motorised stage, from JPK Instruments AG (www.jpk.com). The cantilever used in this chapter is a triangular SNL-10 (Bruker) equipped with a pyramidal tip with nominal spring constant $k = 0.06\,\mathrm{N/m}$, which is large enough to make correction (2.3) negligible.

2.3 Singular event detection and characterisation

FICs are analysed by using the wavelet transform mathematical microscope. Starting from Equation (2.5) for a conical indenter injecting the condition of constant velocity indentation $h = v_0\tau$ we obtain:

$$F(t) = \frac{8\tan(\alpha)v_0^2}{\pi(1-\nu)}\int_0^t G(t-\tau)\tau d\tau. \tag{2.7}$$

Then, deriving with respect to t by Leibniz integral rule (considering that $G(0) = 0$; we get:

$$\frac{dF}{dt} = \frac{8\tan(\alpha)v_0^2}{\pi(1-\nu)}\int_0^t \frac{dG(t-\tau)}{dt}\tau d\tau = -\frac{8\tan(\alpha)v_0^2}{\pi(1-\nu)}\int_0^t G(t-\tau)d\tau, \tag{2.8}$$

where we integrated by parts in the last step. Finally, by deriving once again with respect to t we find the expression for $G(t)$:

$$G(t) = \frac{\pi(1-\nu)}{8\tan(\alpha)v_0^2}\frac{d^2F}{dt^2}, \tag{2.9}$$

which can be written equivalently in term of the indentation h:

$$G(h) = \frac{\pi(1-\nu)}{8\tan(\alpha)}\frac{d^2F(h)}{dh^2}. \tag{2.10}$$

This equation shows that the change of F with h is an important quantity that is stored in the whole deformation story. The shear relaxation modulus can be obtained therefore directly from the second derivative of the force indentation curve with respect to h, without assuming *a-priori* a particular viscoelastic cellular model. Moreover this setup is prone to be used for FICs wavelet analysis. We used a time-frequency adaptative wavelet-based method to compute $G(h)$ from FICs. Within the norm $\mathcal{L}^1$, the one-dimensional WT of a signal $F(h)$ reads (Grossmann and Morlet, 1984; Misiti et al., 2007; Daubechies, 1992; Arneodo, Bacry and Muzy, 1995; Mallat, 1999):

$$W_\psi[F](b,s) = \frac{1}{s}\int_{-\infty}^{\infty} F(h)\psi^*(\frac{h-b}{s})dh, \tag{2.11}$$

where b is a spatial coordinate (homologous to h) and s (>0) a scale parameter and ψ is the analysing wavelet. In the context of this study, we concentrated on the family of analyzing wavelets obtained from the successive derivatives of the Gaussian function (Arneodo, Bacry and Muzy, 1995; Mallat, 1999; Arneodo et al., 2002, 2008, 2011) (see Fig. 2.4):

$$g^{(0)}(z) = e^{-z^2/2}\,. \tag{2.12}$$

Let us define the first derivative of the Gaussian function (see Fig. 2.4):

$$g^{(1)}(z) = -\frac{d}{dz}g^{(0)}(z) = ze^{-z^2/2}, \tag{2.13}$$

and its second derivative, also called the Mexican hat wavelets (see Fig. 2.4):

$$g^{(2)}(z) = -\frac{d^2}{dz^2}g^{(0)}(z) = e^{-z^2/2}(1-z^2). \tag{2.14}$$

Via one (resp. two) derivative under the integral operator, it is straightforward to demonstrate (Arneodo, Bacry and Muzy, 1995; Mallat, 1999; Arneodo et al., 2008, 2011; Muzy et al., 1994) that the WT of F with the first (resp. the second) derivative of a Gaussian wavelet at scale s, $W_{g^{(1)}}[F](b,s)$ (resp. $W_{g^{(2)}}[F](b,s)$) is precisely the first (resp. second) derivative with respect to b of a smooth version $W_{g^{(0)}}[F](b,s)$ of F, filtered by a Gaussian function at the same scale s:

$$W_{g^{(1)}}[F](b,s) = s\frac{d}{db}W_{g^{(0)}}[F](b,s), \tag{2.15}$$

and

$$W_{g^{(2)}}[F](b,s) = s^2\frac{d^2}{db^2}W_{g^{(0)}}[F](b,s). \tag{2.16}$$

The interest of the WT method is two-fold. The first advantage is to use the same smoothing function to filter out the experimental background noise and to compute

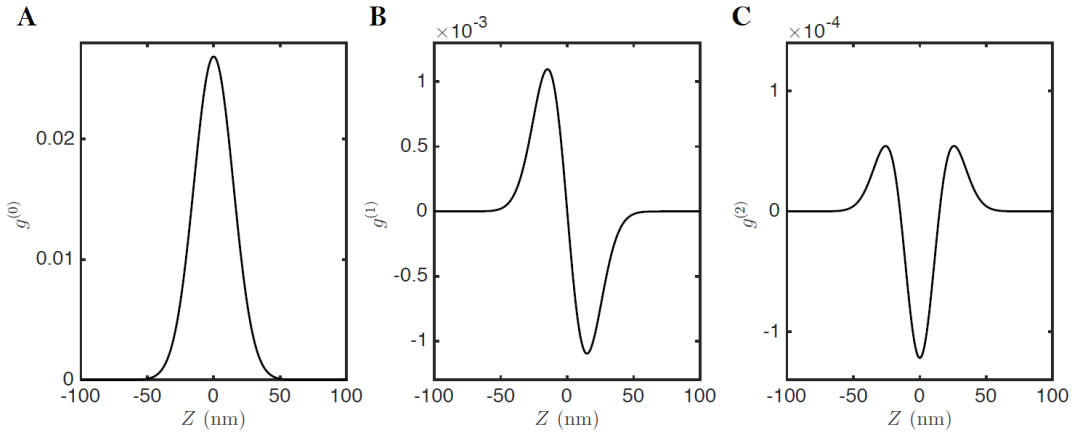

Figure 2.4 – Analyzing wavelets constructed from a Gaussian function. **A.** Gaussian function $g^{(0)}(Z)$ (see Equation (2.12). **B.** First derivative of a Gaussian $g^{(1)}(z)$ (see Equation (2.13)). **C.** Second derivative of a Gaussian $g^{(2)}(z)$ (see Equation (2.14)).

higher-order derivatives (for instance up to second-order in this study) at a well defined smoothing scale s_g. The second advantage relies on the power of the WT to detect and characterise local singularities (including rupture events in the FICs) (Muzy et al., 1991; Arneodo, Bacry and Muzy, 1995; Bacry et al., 1993). In this study, we used modified versions (Laperrousaz, Drillon, Berguiga, Nicolini, Audit, Satta, Arneodo and Argoul, 2016) of the definition of the WT (Eq. (2.11)) to get a direct measure of F in nN ($T_{g^{(0)}}[F](b,s)$), dF/dZ in nN/nm ($T_{g^{(1)}}[F](b,s)$) and d^2F/dZ^2 in Pascal ($T_{g^{(2)}}[F](b,s)$), once smoothed by a Gaussian window ($g^{(0)}(Z)$) of width s:

$$T_{g^{(0)}}[F](b,s) = W_{g^{(0)}}[F](b,s), \tag{2.17}$$

$$T_{g^{(1)}}[F](b,s) = \frac{1}{s} W_{g^{(1)}}[F](b,s), \tag{2.18}$$

$$T_{g^{(2)}}[F](b,s) = \frac{1}{s^2} W_{g^{(2)}}[F](b,s). \tag{2.19}$$

Figure 2.5 illustrates on a single FIC (approach and retract curves) the computation of the wavelet-based force derivatives. Let us point out that from Equation (2.10), after application of the wavelet transform, we get the following expression for the shear relaxation modulus (Digiuni et al., 2015; Laperrousaz, Berguiga, Nicolini, Martinez-Torres, Arneodo, Satta and Argoul, 2016; Laperrousaz, Drillon, Berguiga, Nicolini, Audit, Satta, Arneodo and Argoul, 2016):

$$G(Z) = \frac{\pi(1-\nu)}{8\tan\theta} T_{g^{(2)}}[F](Z,s). \tag{2.20}$$

From now on, we will use the capital letter Z instead of h, considering that the correction (2.3) has been done, or can be neglected and therefore we can indistinctly use z as h. Notice that by using property (2.16), the second derivative can be computed numerically at the same time as smoothing the FIC, instead of deriving noisy experimental data.

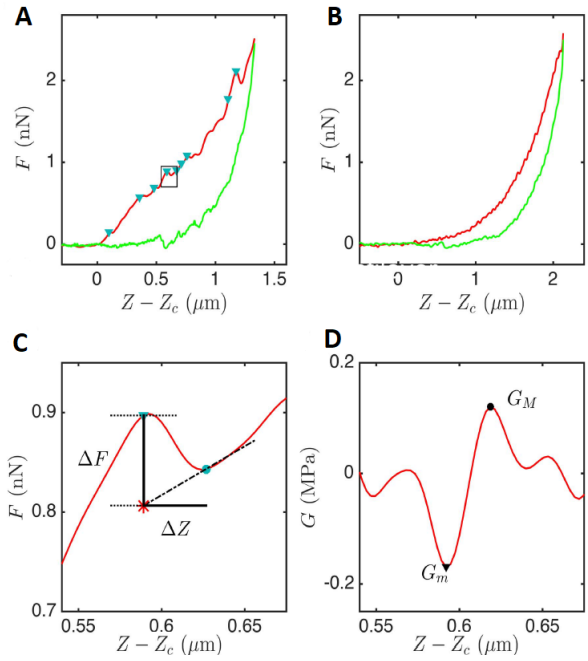

Figure 2.5 – **A.** Typical approach (red) and retract (green) FICs collected on a CML haematopoietic cell (CP-CML patient). **B.** Same as **A** but for a normal haematopoietic cell (healthy donor). **C.** Zoom on **A** around a disruption event indicated by a square. **D.** Corresponding second-order derivative $G(Z)$ of the FIC (see equation (2.20)) computed with a wavelet of size $w = 2\sqrt{2}s = 42\,\text{nm}$; the local minima G_m (resp. maxima G_M) of d^2F/dZ^2 corresponding to a strong negative (resp. positive) curvature of the FIC are marked with black triangles (resp. dots). In a close neighbourhood of G_m and G_M, the local maxima and minima of the FIC are detected and marked with blue triangles and dots in **A** and **C**. The force drop ΔF of a disruption event is corrected by taking into account the increasing behaviour of the FIC (linear dashed-dotted line in **C**).

Finally, once we have computed (2.20), (see Fig. 2.5**D**), we can detect maxima and minima of $G(z)$ exceeding the background thermal fluctuations, filtered out by a Gaussian wavelet of size $w = 2\sqrt{2}s = 42\,\text{nm}$. In this way we can compute the size ΔZ and the strength ΔF of these singular events: the starting point of the event is the local minimum G_m and the ending point is the local maximum G_M (see Fig. 2.5**C** and **D**). In the computation of the strength ΔF we have corrected by the local slope in the neighbour of the singular event, taking into account the average increase of the FICs. Therefore for each disruption event we can compute the energy released along the fracture:

$$E = \Delta F \cdot \Delta Z \tag{2.21}$$

2.4 Experiments on human cells

We are now going to analyse and discuss AFM experiments done in my hosting team by B. Laperrousaz (Laperrousaz, Berguiga, Nicolini, Martinez-Torres, Arneodo, Satta and Argoul, 2016) retracing the main lines of our paper (Polizzi et al., 2018). All the figures from this and the next section are taken from the same publication (Polizzi et al., 2018).

We used AFM, as explained previously, to indent single haematopoietic purified $CD34^+$ cells from CP-CML patients at diagnosis and healthy donor bone marrows (Laperrousaz,

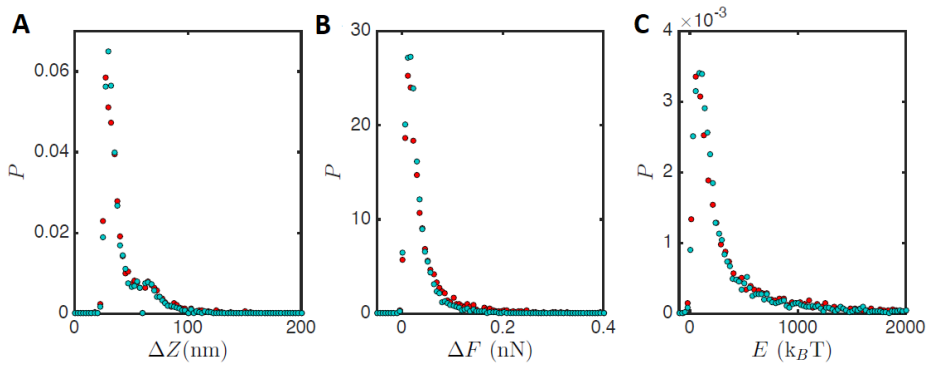

Figure 2.6 – **A.** Experimental probability density function of cascade size ΔZ. **B.** Experimental probability density function of cascade force drop ΔF. **C.** Experimental probability density function of cascade released energy E. For all plots red dots correspond to data from CML cells and blue from data normal ones.

Berguiga, Nicolini, Martinez-Torres, Arneodo, Satta and Argoul, 2016). Approach and withdraw are done at the same constant velocity of $1\,\mu$m/s. Thank to the wavelet based method we detected singular cascade rupture events. As noted in (Laperrousaz, Berguiga, Nicolini, Martinez-Torres, Arneodo, Satta and Argoul, 2016; Streppa, 2017; Polizzi et al., 2018) this indentation caused locally a strain stiffening, signature of an increased tension in the network, eventually resulting in local singularities in the force indentation curves, interpreted as avalanches of fractures.

Globally, the experimental sample was composed by two large sets of single cell FICs from 5 CP-CML patients (1301 FICs - 49 cells) and 5 healthy donors (1671 FICs - 60 cells). We detected 6161 singular rupture events distributed on 1153 FICs in CP-CML cells as compared to 6765 rupture events distributed on 1111 FICs in normal cells. Thus only 11.4% (148/1301) of FICs do not display rupture events for CML cells which is significantly lower than 33.5% (560/1671) for normal cells.

2.4.1 Ductile *versus* brittle rupture events

The computation of the normalized histograms (p.d.f. for probability density function) of ΔZ, ΔF, and E of these rupture events, revealed rather wide distributions with a fat tail on the right (see Fig. 2.6).

When focusing on ΔZ (nm), which can also be interpreted as the time duration $\Delta t = \Delta Z/V_0$ (ms) of the rupture event, we got very satisfactory fits (non-linear least square fit method) of the p.d.f. of $\log_{10}(\Delta Z)$ with the sum of two distinct Gaussian distributions, for both normal and CML cells (Figures 2.7**A** and **B**):

$$P(Y) = \frac{(1-\alpha)e^{\frac{-(Y-\mu_1)^2}{2\sigma_1^2}}}{\sqrt{2\pi}\sigma_1} + \frac{\alpha e^{\frac{-(Y-\mu_2)^2}{2\sigma_2^2}}}{\sqrt{2\pi}\sigma_2}, \tag{2.22}$$

where the indices 1 and 2 refer to the first Gaussian population and the second (for larger Z) Gaussian population of rupture events respectively. μ_1 (resp. μ_2) and σ_1 (resp. σ_2) are the mean and root-mean-square (r.m.s.) of the random variable $Y = \log_{10}(X)$. The arithmetic

and the geometric means of X are $\overline{X} = \mathbb{E}[X] = 10^{\mu+\ln(10)\sigma^2/2}$ and $\tilde{X} = \mathbb{GM}[X] = 10^{\mu}$ respectively. The parameter α in Equation (2.22) quantifies the statistical contribution (%) of events from population 2 as compared to the contribution $1-\alpha$ (%) of rupture events from population 1.

In this way, we identified two populations of rupture events, a subpopulation 1 of rather short duration ($\overline{\Delta t_1} \sim 30\,\text{ms}$) and weakly penetrating ($\overline{\Delta Z_1} \sim 30$ nm) failures, and a subpopulation 2 of longer ($\overline{\Delta t_2} \sim 50\,\text{ms}$) and deeply penetrating ($\overline{\Delta Z_2} \sim 50$ nm) failures. But what distinguishes CML from normal cells is the higher percentage ($\alpha = 0.34$ compared to 0.26) of the larger rupture events (subpopulation 2) in the CP-CML cells, as an indication of their greater mechanical brittleness.

When investigating the correlations between ΔZ, ΔF and E, we found a rather strong correlation between $\log_{10}(\Delta F)$ and $\log_{10}(\Delta Z)$ for both the CML (Pearson's correlation coefficient $r = 0.68$) and the normal ($r = 0.62$) cells. The variables $\log_{10}(E)$ and $\log_{10}(\Delta Z)$ were also found significantly correlated in CML ($r = 0.83$) and in normal ($r = 0.80$) cells. These correlations reveal themselves as of two clouds of points in the respective scatter plots (Figures 2.7**C** and **D**), corresponding to subpopulation 1 of small indentation depth, short duration, small force drop and low released energy failures, and to subpopulation 2 of large indentation depth, long duration, large force drop and high released energy failures. We will classify the former as *ductile* and the latter as *brittle* rupture events.

2.4.2 Log-normal statistics of released energy during rupture events

The pertinence of the fitting of the p.d.f. of $\log_{10}(E)$ (Fig. 2.7**E**) by the sum of 2 Gaussians is compelling when using a logarithmic representation (Fig. 2.7**F**). For the CML cells, when fixing the relative percentages of ductile ($1-\alpha = 0.66$) and brittle ($\alpha = 0.34$) rupture events as previously estimated, we obtained the parameter values reported in Table 2.1 with the following arithmetic and geometric means $\overline{E_1} = 216\,\text{k}_\text{B}\text{T}$ (resp. $\overline{E_2} = 1547\,\text{k}_\text{B}\text{T}$), and $\tilde{E}_1 = 141\,\text{k}_\text{B}\text{T}$ (resp. $\tilde{E}_2 = 776\,\text{k}_\text{B}\text{T}$). For normal cells, when fixing $\alpha = 0.26$ ($1-\alpha = 0.74$), we ended with consistent parameter values (Table 2.1) with $\overline{E_1} = 188\,\text{k}_\text{B}\text{T}$ (resp. $\overline{E_2} = 1158\,\text{k}_\text{B}\text{T}$), and $\tilde{E}_1 = 138\,\text{k}_\text{B}\text{T}$ (resp. $\tilde{E}_2 = 741\,\text{k}_\text{B}\text{T}$). Note that the standard deviations of the Gaussian distributions for $\log_{10}(E)$ are slightly larger for CML than for normal cells and this for both ductile and brittle events.

When comparing the mean released energies during ductile ($\sim 200\,\text{k}_\text{B}\text{T}$) and brittle ($\sim 1300\,\text{k}_\text{B}\text{T}$) rupture events to the dissociation energies of α-actinin/actin binding ($\sim 4.3\,\text{k}_\text{B}\text{T}$) and of filamin/actin binding ($\sim 3.6\,\text{k}_\text{B}\text{T}$) (Ferrer et al., 2008), we obtain rough estimates for the number of ABP unbindings during ductile (~ 50) and brittle (~ 325) rupture events. Thus, more than 6 times the amount of energy is released during brittle failures that last ($\sim 50\,\text{ms}$) not more than twice the duration ($\sim 30\,\text{ms}$) of ductile failures, meaning that the mean rate of released energy is significantly higher during brittle failures.

Interestingly, the computation of the global G (found by Eq. (2.4)) and initial G_i (found by Eq. (2.4) only over the first 500 nm after contact) shear moduli confirms that CML cells are stiffer ($\overline{G} = 1.13 \pm 0.31\,\text{kPa}$, $\overline{G_i} = 0.77 \pm 0.26\,\text{kPa}$) than normal ones

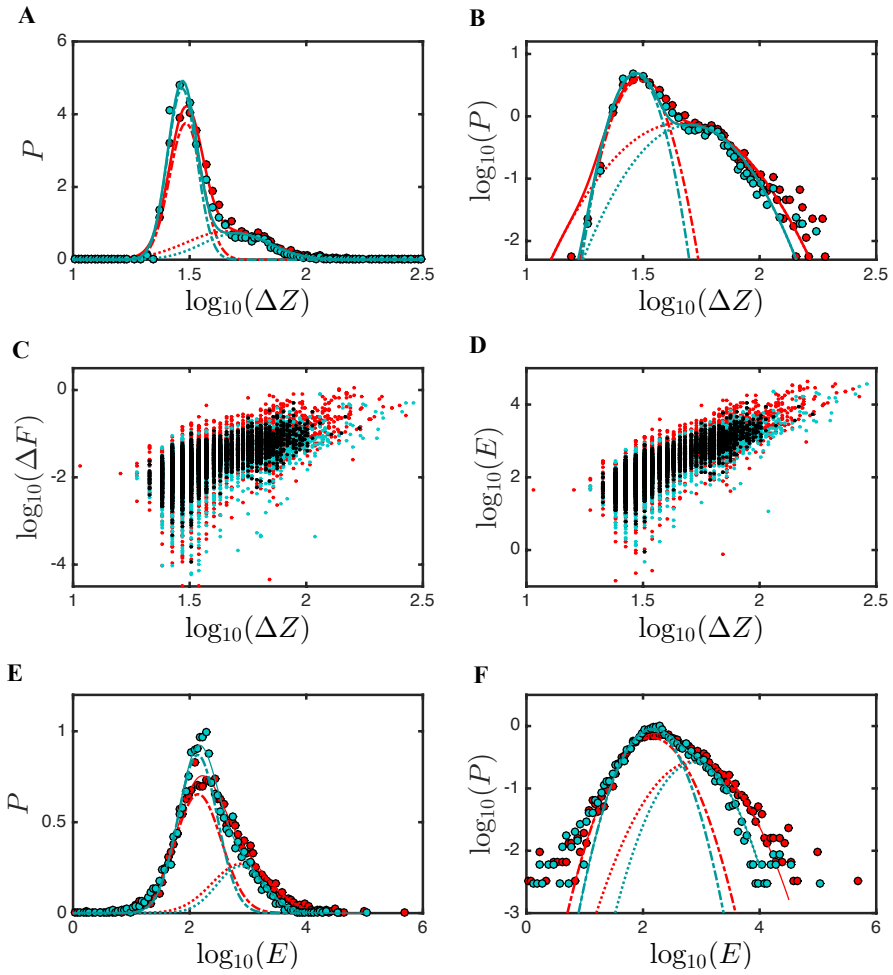

Figure 2.7 – Statistical analysis of the indentation depths (nm), force drops (nN) and released energies (k_BT) of disruption events. These parameters were obtained from local disruption events collected from the FICs of the two sets of CML (red) and normal (blue) cells. **A.** Normalized histograms of $\log_{10}(\Delta Z)$; the dots represent the experimental data and the continuous lines the corresponding fits by the sum of 2 Gaussian distributions (see equation (2.22)) with parameters $\alpha = 0.34$, and $\mu_1 = 1.49$, $\sigma_1 = 0.07$ corresponding to the following means $\overline{\Delta Z_1} = 31\,\text{nm}$ and $\widetilde{\Delta Z_1} = 31\,\text{nm}$ (red dotted-dashed line), $\mu_2 = 1.67$, $\sigma_2 = 0.18$, *i.e.*, $\overline{\Delta Z_2} = 51\,\text{nm}$ and $\widetilde{\Delta Z_2} = 47\,\text{nm}$ (red dotted line) for CML cells, and $\alpha = 0.26$, and $\mu_1 = 1.47$, $\sigma_1 = 0.06$, *i.e.*, $\overline{\Delta Z_1} = 30\,\text{nm}$ and $\widetilde{\Delta Z_1} = 29\,\text{nm}$ (blue dotted-dashed line), $\mu_2 = 1.70$, $\sigma_2 = 0.15$, *i.e.*, $\overline{\Delta Z_2} = 53\,\text{nm}$, and $\widetilde{\Delta Z_2} = 50\,\text{nm}$ (blue dotted line) for normal cells. **B.** Same as **A** but in a logarithmic representation where Gaussians become parabolae. **C.** Scatter plot of $\log_{10}(\Delta F)$ *vs* $\log_{10}(\Delta Z)$; each red (resp. blue) dot corresponds to a rupture event in a CML (resp. normal) FIC; whenever a red dot and a blue dot superimpose they were turned into black. **D.** Scatter plot of $\log_{10}(E)$ vs $\log_{10}(\Delta Z)$. **E.** Normalized histograms of $\log_{10}(E)$; same representation as for $\log_{10}(\Delta Z)$ in **A**; the parameters of the fit with the sum of 2 Gaussian distributions are for CML cells: $\alpha = 0.34$ and $\mu_1 = 2.15$, $\sigma_1 = 0.40$, $\mu_2 = 2.89$, $\sigma_2 = 0.51$, and for healthy cells: $\alpha = 0.26$, and $\mu_1 = 2.14$, $\sigma_1 = 0.34$, $\mu_2 = 2.87$, $\sigma_2 = 0.41$ (Table 2.1). **F.** Same as **E** but in a logarithmic representation.

Table 2.1 – Distributions of energy released during ductile and brittle rupture events in normal and CML cells: simulations versus experiments. Characteristics ($\overline{N}$, σ_N, $\mu = \overline{\log_{10}(E)}$, $\sigma = \sigma_{\log_{10} E}$, $\overline{E}$, $\tilde{E}$) of the numerical cascades simulated with Equation (2.24) for parameters values a_0, $\hat{a}$, $\Delta E_0 = 12\,\mathrm{k_B T}$, $\Delta E^* = 4\,\mathrm{k_B T}$ versus experimental data (Figs. 2.7**E** and 2.7**F**)

	Normal Cells	**CML Cells**
Ductile		
Model Parameters	$a_0 = 1.3$, $\hat{a} = 15$, $\Delta a = 0.39$	$a_0 = 1.3$, $\hat{a} = 15$, $\Delta a = 0.48$
Simulations	$\overline{N} = 9$, $\sigma_N = 3$ $\mu_1 = 2.17$, $\sigma_1 = 0.31$ $\overline{E}_1 = 191\,\mathrm{k_B T}$, $\widetilde{E}_1 = 148\,\mathrm{k_B T}$	$\overline{N} = 8$, $\sigma_N = 4$, $\mu_1 = 2.13$, $\sigma_1 = 0.38$ $\overline{E}_1 = 198\,\mathrm{k_B T}$, $\widetilde{E}_1 = 135\,\mathrm{k_B T}$
Experiments	$\mu_1 = 2.14$, $\sigma_1 = 0.34$ $\overline{E}_1 = 188\,\mathrm{k_B T}$, $\widetilde{E}_1 = 138\,\mathrm{k_B T}$	$\mu_1 = 2.15$, $\sigma_1 = 0.40$ $\overline{E}_1 = 216\,\mathrm{k_B T}$, $\widetilde{E}_1 = 141\,\mathrm{k_B T}$
Brittle		
Model Parameters	$a_0 = 1.3$, $\hat{a} = 46$, $\Delta a = 0.35$	$a_0 = 1.3$, $\hat{a} = 53$, $\Delta a = 0.38$
Simulations	$\overline{N} = 21$, $\sigma_N = 7$ $\mu_2 = 2.84$, $\sigma_2 = 0.42$ $\overline{E}_2 = 1104\,\mathrm{k_B T}$, $\widetilde{E}_2 = 692\,\mathrm{k_B T}$	$\overline{N} = 22$, $\sigma_N = 8$ $\mu_2 = 2.92$, $\sigma_2 = 0.47$ $\overline{E}_2 = 1494\,\mathrm{k_B T}$, $\widetilde{E}_2 = 832\,\mathrm{k_B T}$
Experiments	$\mu_2 = 2.87$, $\sigma_2 = 0.41$ $\overline{E}_2 = 1158\,\mathrm{k_B T}$, $\widetilde{E}_2 = 741\,\mathrm{k_B T}$	$\mu_2 = 2.89$, $\sigma_2 = 0.51$ $\overline{E}_2 = 1547\,\mathrm{k_B T}$, $\widetilde{E}_2 = 776\,\mathrm{k_B T}$

($\overline{G} = 0.42 \pm 0.05\,\text{kPa}$, $\overline{G_i} = 0.28 \pm 0.05\,\text{kPa}$). In particular, about 37% of FICs bears mean shear relaxation moduli (G) larger than 1 kPa in CML cells with very low (7%) counter part in normal cells (see Fig. 2.8). Altogether, these observations show that CML cells display a significant proportion of highly tensed perinuclear zones propitious to localized brittle failures by disruption of cross-linked CSK domains impeding complete shape recovery after deformation. This interpretation is strengthened by the experimental observation of an important structural alteration of the actin CSK of immature TF1 cells consecutive to BCR-ABL oncogene transduction. In particular, confocal fluorescence microscopy revealed that in this model of human CML (Savona and Talpaz, 2008; Jones, 2010; Laperrousaz et al., 2013; McWhirter and Wang, 1993; Cheng et al., 2002; Laperrousaz, Drillon, Berguiga, Nicolini, Audit, Satta, Arneodo and Argoul, 2016), juxtanuclear actin aggregates were found in almost 30% of the BCR-ABL-transduced TF1 cells at the expense of the cortical F-actin (Laperrousaz, Drillon, Berguiga, Nicolini, Audit, Satta, Arneodo and Argoul, 2016; McWhirter and Wang, 1993) (Fig. 2.9b). Very likely, these solid structures were induced by the oncogene since they were rarely observed in the parental TF1 cell line (Figure 2.9a).

2.5 A cascade model for living cells plasticity

To account for the observed log-normal released energy distributions (Figs. 2.7**E** and **F**), we propose a multiplicative cascade description of CSK failure events that is inspired from pioneering works on population growth dynamics (Kesten, 1973; Malevergne et al., 2011).

Probability density functions with heavy tails have been widely observed in various domains of fundamental and applied sciences (Meyers , Eds). The most popular fat-tail distributions are power-law and log-normal distributions (see Section 1.3) that often have been considered as competing models of experimental data (Mitzenmacher, 2004; Malevergne et al., 2011). Power-law distributions are commonly thought to be a statistical characteristics of systems that display space and/or time scale invariance properties (Meyers , Eds) with as notable examples scale-free networks (Barabási and Albert, 1999) and self-organized critical systems (Bak, 2013; Sornette, 2006; Perez-Reche et al., 2016). Log-normal distributions are paradigmatic fat-tail distributions generated by self-similar fragmentation and more generally multiplicative processes (Gibrat, 1931; Mandelbrot, 1982; Arneodo et al., 1998), with as a historical example the energy cascade model of fully developed turbulence (Frisch and Kolmogorov, 1995). But, as originally proposed by Kesten (Kesten, 1973), power-law and log-normal distributions are indeed intrinsically connected when combining a multiplicative and an additive random processes:

$$S_t = a_{t-1} S_{t-1} + b_{t-1} \,, \tag{2.23}$$

where S_t is the size of a population at time t, a_t the random positive growth factor, and b_t a small positive random increment. In the absence of the additive b_t term, one recovers the multiplicative Gibrat's law (Gibrat, 1931) of proportional growth which is nothing but a random walk in log-size leading to a log-normal distribution of S_t. But this distribution is not stable since when the time increases the process S_t either asymptotically shrinks

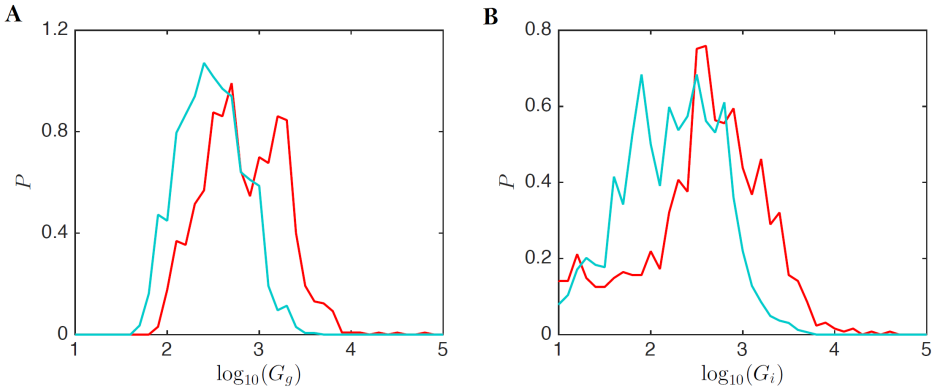

Figure 2.8 – Statistical analysis of the shear relaxation moduli of CML and normal cells. The parameters were estimated from the FICs of the two sets of CML (red) and normal (blue) cells. **A.** P.d.f. of the global elastic modulus G_g. **B.** P.d.f. of the initial elastic modulus G_i, fitted only over the first 500 nm after contact, always with Formula (2.4). G_g and G_i are given in Pa.

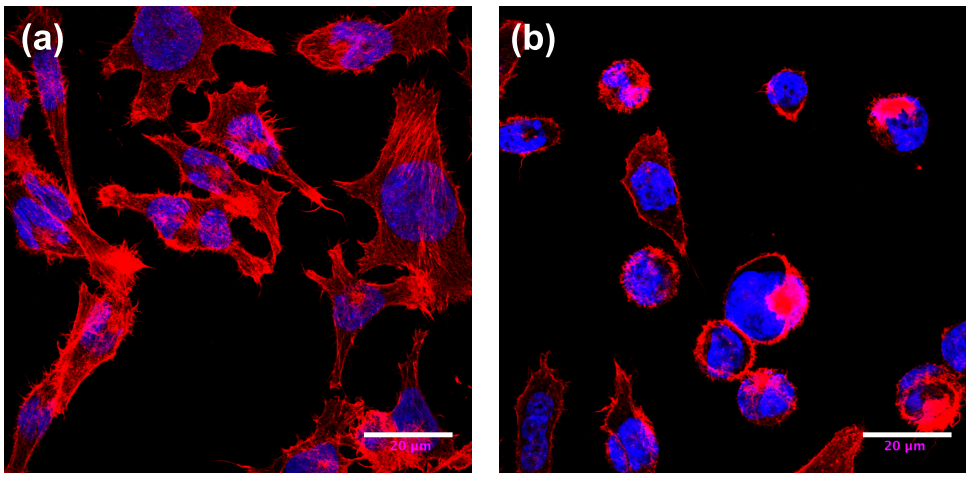

Figure 2.9 – Cytoskeleton structure of TF1 cells revealed by confocal microscopy. (a) Parental TF1 cells. (b) TF1 haematopoietic cells after transfection by the CML BCR-ABL oncogene. F-actin was labeled with phalloidin -rhodamin (red), and the nuclei were labeled with DAPI (blue). Immunofluorescence images were taken using a confocal microscope on fixed TF1 and TF1-BCR-ABL adherent cells on fibronectin. Scale bar: 20 μm.

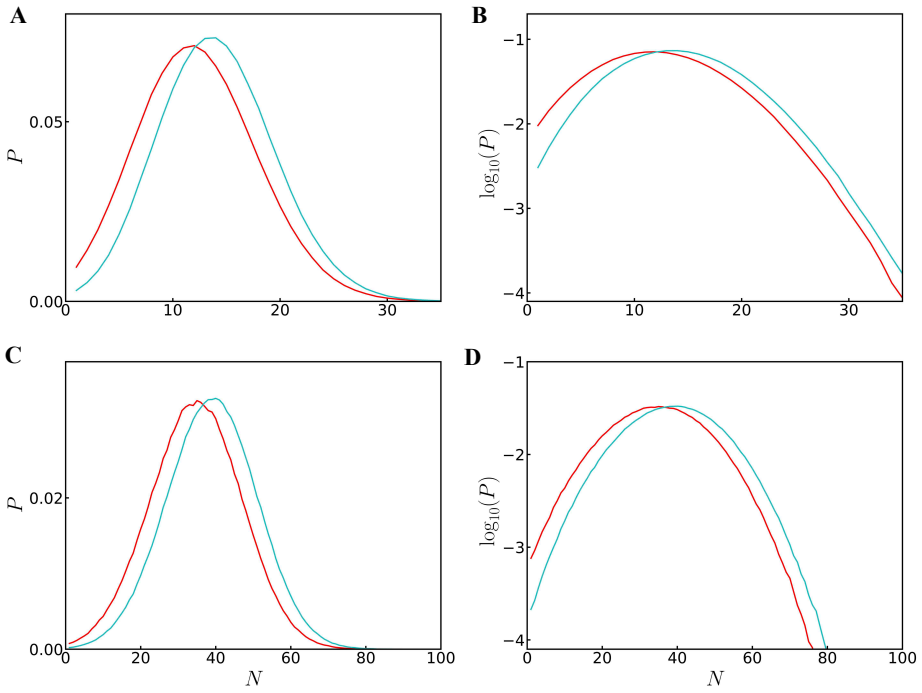

Figure 2.10 – Distributions of the number of steps N of the log-normal rupture cascade model. P.d.f. of N computed from 1.2×10^6 realisations of the multiplicative process defined by Equation (2.24), for parameters values given in Table A.1 in Appendix A. **A.** Ductile rupture events in normal (blue) and CML (red) cells. **B.** semi-log representation. **C.** Brittle rupture events in normal (blue) and CML (red) cells. **D.** semi-log representation.

stochastically to zero (if $\overline{\ln(a_t)} < 0$) or diverges to infinity (if $\overline{\ln(a_t)} > 0$). Kesten (Kesten, 1973) showed that provided $\overline{\ln(a_t)} < 0$, adding the random variable b_t prevents the annihilation of S_t which progressively switches and converges to a stationary power-law distribution with exponent α given by the strictly positive solution of the equation $\overline{a_t^\alpha} = 1$.

Our model of CSK cascading rupture events is deliberately reductionist with a minimal number of parameters. At the cascade step $n+1$, the energy released ΔE_{n+1} satisfies a Gibrat's multiplicative law (Gibrat, 1931), meaning that it can be expressed as a percentage of the energy ΔE_n released at the previous cascade step:

$$\Delta E_{n+1} = a_n \Delta E_n \ , \quad \text{with} \ \ a_n \sim \mathcal{N}\left(a_0 e^{-n/\hat{a}}, \Delta a^2\right), \tag{2.24}$$

where a_n is a Gaussian random variable with initial mean value $a_0 > 1$ to allow the cascade to start. The mean value of the random growth factor $\overline{a_n}$ decreases exponentially (CSK structural relaxation) with a characteristic time $\hat{a}$ and a fixed standard deviation $\sigma_{a_n} = \Delta a$. Note that the relationship between the cascade index n and time is not defined in our model since there is no reason *a priori* to assume that the CSK rupture cascade steps occur at regular time intervals. Starting from some initiation threshold ΔE_0, ΔE_n will asymptotically tend to zero because the exponentially decreasing multiplicative constant a_n becomes smaller than 1 after some finite number of steps. The cascade stops when $\Delta E_{N+1} < \Delta E^*$, where $\Delta E^* \geq 0$ is a cascade arrest energy cut-off (to be interpreted as the lowest possible discrete amount of energy to be released in the system). During

the N steps of the CSK rupture cascade, the total released energy is:

$$E = \sum_{n=1}^{N} \Delta E_n. \tag{2.25}$$

Our random cascade model mainly depends on three parameters: a_0, $\hat{a}$ and Δa. N is a random variable that depends on the realization of the cascade. Its p.d.f. $P(N)$ is well approximated by a Gaussian distribution (Figure 2.10). This total energy released, given by Eq. (2.25) is the quantity accessible via experimental results, and therefore it is the quantity on which we focus our modelling.

To account for the experimental data, all our simulations of the released energy cascade model defined by equation (2.24) were performed for rather small values of $\Delta a \ll 1$. In this limit, E (equation (2.25)) can be approximated by:

$$E \simeq E_0 \sum_{n=1}^{N} a_0^n \, \Pi_{i=1}^{n} e^{-i/\hat{a}} \; = E_0 \sum_{n=1}^{N} a_0^n \; e^{-\sum_{i=1}^{n} i/\hat{a}} \; . \tag{2.26}$$

From the well known identity $\sum_{i=1}^{n} i = n(n+1)/2$, we get

$$E \simeq E_0 \sum_{n=1}^{N} a_0^n \; e^{-\frac{n(n+1)}{2\hat{a}}} \; . \tag{2.27}$$

Thus, a good approximation of the mean of E can be obtained by replacing N by the nearest integer $[\overline{N}]$ of its mean $\overline{N}$ in Equation (2.27):

$$\overline{E} \simeq E_0 \sum_{n=1}^{[\overline{N}]} a_0^n \; e^{-\frac{n(n+1)}{2\hat{a}}} \; . \tag{2.28}$$

We recall that the p.d.f. of the random variable N was shown to be well approximated by a Gaussian (Figure 2.10). Now from the Taylor expansion of $\ln(E) = \ln\left(\overline{E}\right) + (E - \overline{E})/\overline{S}$ at first order, we can show by taking the ensemble average on both sides of this equation that $\ln\left(\overline{E}\right)$ is well approximated by

$$\overline{\ln(E)} \approx \ln\left(\overline{E}\right) \; . \tag{2.29}$$

We have confirmed numerically the pertinence of this approximation in all our simulations with a rather good accuracy ($< 5\%$ error). Now, if we assume that the energy cascade ends when the difference between $\overline{\Delta E_n}$ and the threshold ΔE^* is of the order of Δa, *i.e.* $ae^{-\overline{N}/\hat{a}} \simeq K\Delta a$, then

$$\overline{N} \simeq \hat{a} \; \ln\left(\frac{a_0}{K\Delta a}\right) \; , \tag{2.30}$$

where K is a constant that has been numerically estimated, of order 1 (*e.g.* $K = 1.52 \pm 0.05$ for $a_0 = 1.3$, $\hat{a} = 45$ and $\Delta a = 0.36$ see Figure 2.12). Equation (2.30) shows that the mean size $\overline{N}$ of the energy cascade increases as expected when increasing a_0 and $\hat{a}$, but decreases when increasing Δa, since the probability to have a $\Delta E_n < \Delta E^*$ after a limited

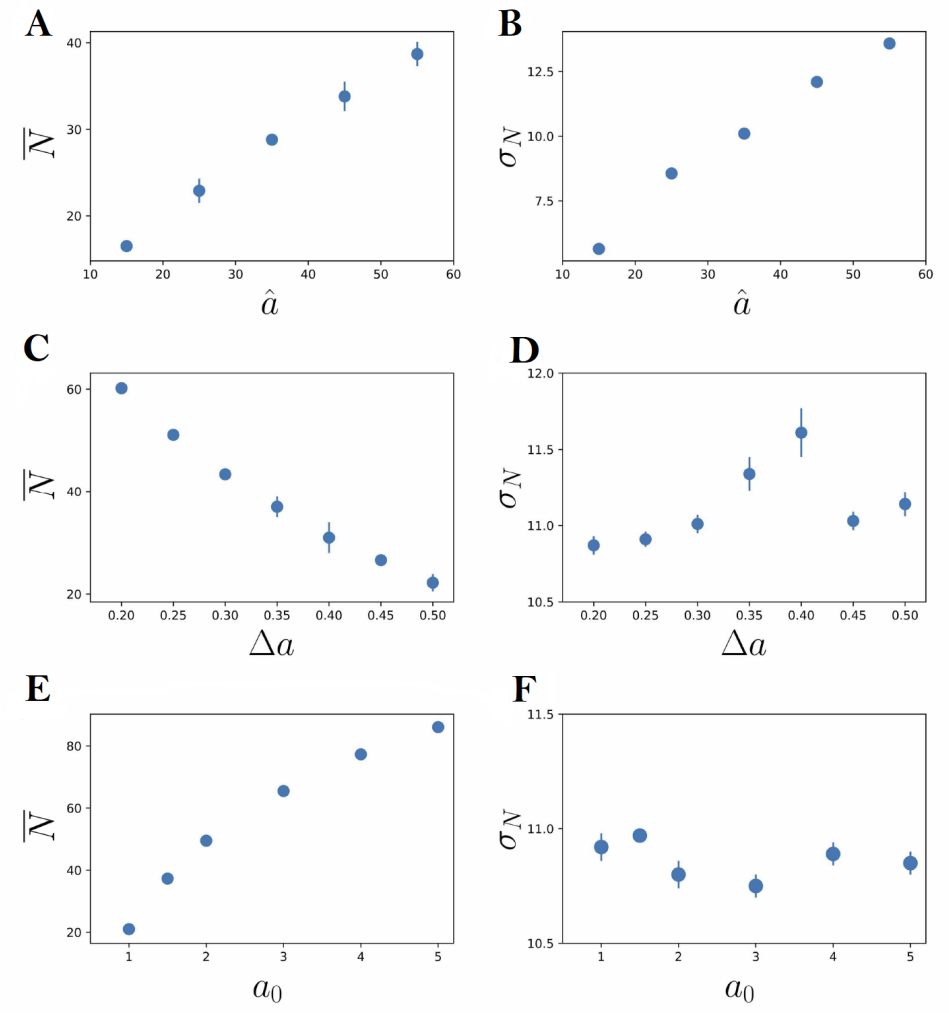

Figure 2.11 – Dependence of the average number of steps N and its standard deviation on the model parameters parameters $\tilde{a}$, Δa and a_0. Fixed parameter values are: $\Delta E_0 = 12\,\mathrm{k_B T}$, $\Delta E^* = 0$. **A.** $\overline{N}$ *vs* $\hat{a}$. **B.** σ_N *vs* $\hat{a}$, for fixed $a_0 = 1.3$, $\Delta a = 0.4$. **C.** $\overline{N}$ *vs* Δa. **D.** σ_N *vs* Δa, for fixed $a_0 = 1.3$, $\hat{a} = 40$. **E.** $\overline{N}$ *vs* a_0. **F.** σ_N *vs* a_0, for fixed $\hat{a} = 40$, $\Delta a = 0.4$.

number n of cascade steps increases (see Fig. 2.11).

2.6 Comparison with experimental data and model interpretation

In our simulations, we fixed $a_0 = 1.3$ and started iterating Equation (2.24) with the initial threshold $\Delta E_0 = 12\,\mathrm{k_B T}$. Note that ΔE_0 is a multiplicative factor that only shifts the p.d.f. of $\log_{10}(E)$ without affecting its shape. When using the cut-off $\Delta E^* = 0$ ($\Delta E_n > 0, \forall n$), after some transient ($n \gtrsim 5$), the p.d.f. of ΔE_n obtained from 1.2×10^6 realizations with parameter values $\hat{a} = 45$ and $\Delta a = 0.36$ is well approximated by a log-normal Gibrat's law (Gibrat, 1931), as expected from multiplicative process (Figures 2.12**A** and **B**). What is more surprising is the fact that the sum E of these log-normal variables turns out also to be well approximated by a log-normal distribution and not by a normal distribution as expected from the central limit theorem. Indeed, this theorem does not apply since the

random variables ΔE_n are not only not identically distributed but also correlated.

Beaulieu (Beaulieu, 2012) has recently proved an extended log-normal limit theorem that states that the limit distribution of the sum of non-identically distributed correlated (with a particular structure of correlation) log-normal random variables is a log-normal distribution. When investigating how the energy $E_{n^*} = \sum_{n=1}^{n^*} \Delta E_n$, released during the first n^* steps of the cascade converges to the total released energy E_N, we confirmed a rather fast convergence to an asymptotic log-normal distribution (Figures 2.12**C** and **D**). When numerically simulating surrogate uncorrelated released energy cascades where at each step, ΔE_n is randomly generated with the p.d.f. $P(\Delta E_n)$ (Fig. 2.12**A**) but independently of the previous cascade steps, then the sum E_{n^*} no longer converges to a log-normal distribution but eventually (if n^* is large enough) to a normal distribution as expected for uncorrelated random variables (Figures 2.12**E** and **F**). This theoretical argument suggests that the log-normal distribution observed experimentally (Figures 2.7**E** and **F**) are evidence that the different cascade steps of ABP unbinding are correlated. These conclusions apply to all the numerical simulations reported in this study, for ductile and brittle rupture events, in normal as well as in CML cells (Table 2.1, Figs. 2.13 and 2.14). In Chapter 3 we will see in more details what are the effects of correlations on the sum of random variables resulting from a multiplicative process and we will discuss the application of Beaulieu's theorem (Beaulieu, 2012) to our particular case.

We performed numerical simulations of the released energy cascade model to reproduce quantitatively the p.d.fs of E obtained for both normal and CML cells (Figures 2.7**E** and **F**). We have generated 1.2×10^6 realizations of the multiplicative process defined by Equation (2.24) using $\Delta E_0 = 12\,\mathrm{k_B T}$ as initial threshold and $\Delta E^* = 0$ (Figures 2.10, 2.11, 2.12 and also Figures A.1 and A.2 and Table A.1 in Appendix A) or $DeltaE^* = 4\,\mathrm{k_B T}$ (Figure 2.13, Table 2.1) $DeltaE^*$ being the energy cascade arrest cut-off ($\Delta E_{N+1} < \Delta E^*$). For $\Delta E^* = 0$, we have systematically investigated the dependence of the p.d.fs of the number of cascade steps N and of the corresponding released energy E on the three parameters a_0, $\hat{a}$ and Δa. As long as $\overline{N}$ is not too small, $P(N)$ is well approximated by a Gaussian distribution (Figure 2.10) whose mean increases when increasing a_0 or $\hat{a}$, or when decreasing Δa, fixing the two other parameters (see Fig. 2.11), in good agreement with the theoretical prediction (Eq. (2.30)). Interestingly, the corresponding $P(E)$ turns out to be well approximated by a log-normal distribution (Fig. 2.12**C** and **D**). When using the parameter dependence of $\overline{\log_{10} E}$ and $\overline{\sigma}_{\log_{10}(E)}$ versus a_0, $\hat{a}$ and Δa (see Fig. A.1 in Appendix A), we realized that we could fit the experimental log-normal distributions by fixing $a_0 = 1.3$ (> 1) and adjusting $\hat{a}$ and Δa to match the parameters μ_i and σ_i estimated from log-normal fits of the data for ductile and brittle events in both normal and CML cells (see Fig. A.2 and Table A.1 in Appendix A). When using a finite energy cut-off $\Delta E^* = 4\,\mathrm{k_B T}$ consistent with the experimental estimate of ABP unbinding energy (Ferrer et al., 2008), we still were able to quantitatively reproduce the log-normal distributions of released energy observed experimentally for parameter values reported in Table 2.1 (Fig. 2.13). Note that with this finite ΔE^* cut-off, some departure from Gaussian tail is observed in $P(N)$ for small N (see Fig. A.3 in Appendix A), and in turn in $P(\log_{10}(E))$ for small $\log_{10}(E)$ values (Fig. 2.13), as the signature of a lack of statistical convergence of the shortest rupture cascades. Interestingly this departure seems also to be present in

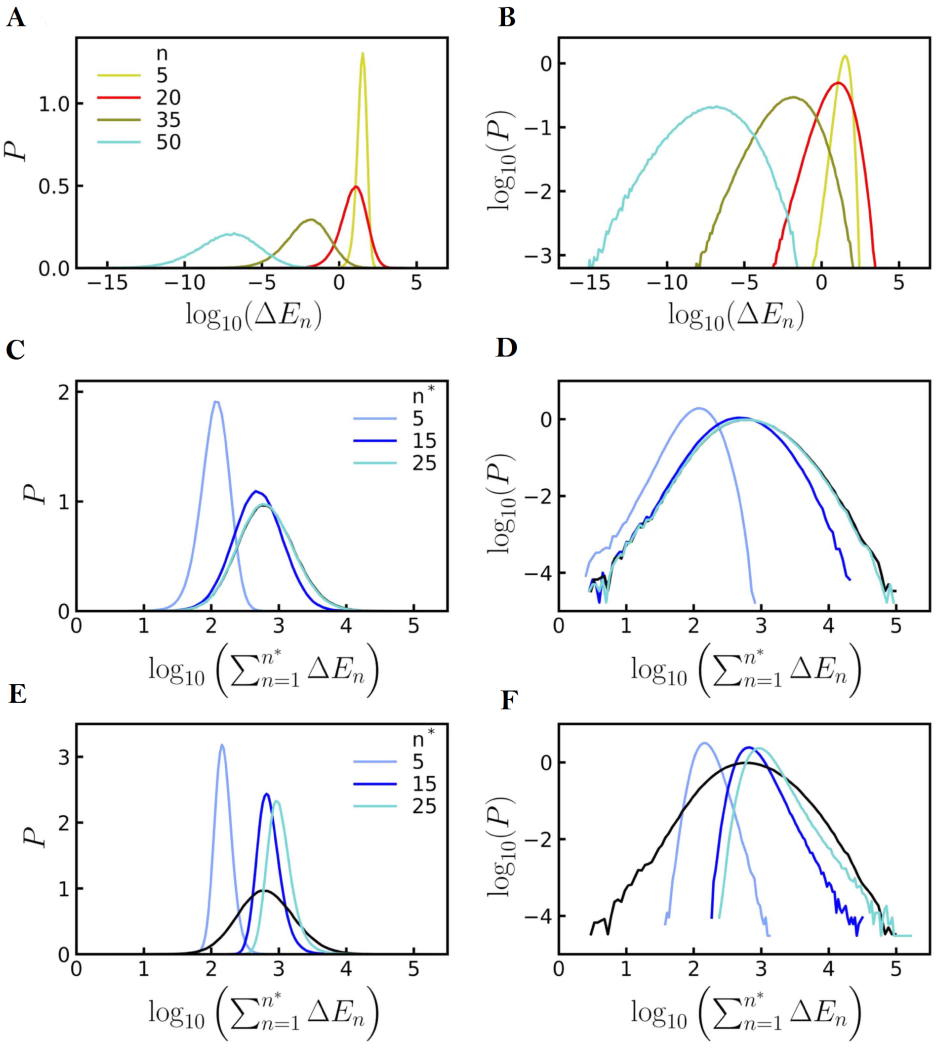

Figure 2.12 – Released energy (k_BT) distributions simulated with the log-normal rupture cascade model. **A.** P.d.f. of $\log_{10}(\Delta E_n)$ computed from 1.2×10^6 realisations of the multiplicative process defined by equation (2.24), for parameters values: $a_0 = 1.3$, $\hat{a} = 45$, $\Delta a = 0.36$, $\Delta E_0 = 12\,k_BT$ and $\Delta E^* = 0$. The colours correspond to $n = 5$ (yellow), 20 (red), 35 (dark green), 50 (cyan); $P(E)$ (black) is shown for comparison. **B.** Same as **A** but in a logarithmic representation. **C.** P.d.f. of $\log_{10} E_{n^*}$, where $E_{n^*} = \sum_{n=1}^{n^*} \Delta E_n$ for $n^* = 5$ (light blue), 15 (dark blue), 25 (cyan) and N (black, overlapped with the $n^* = 25$ curve). **D.** Same as **C** but in a logarithmic representation. **E.** Same as **C** but for surrogate uncorrelated released energy cascades (see text). **F.** Same as **E** but in a logarithmic representation. In **E** and **F**, $P(E_N)$ (black) has the same meaning as in **C** and **D** respectively.

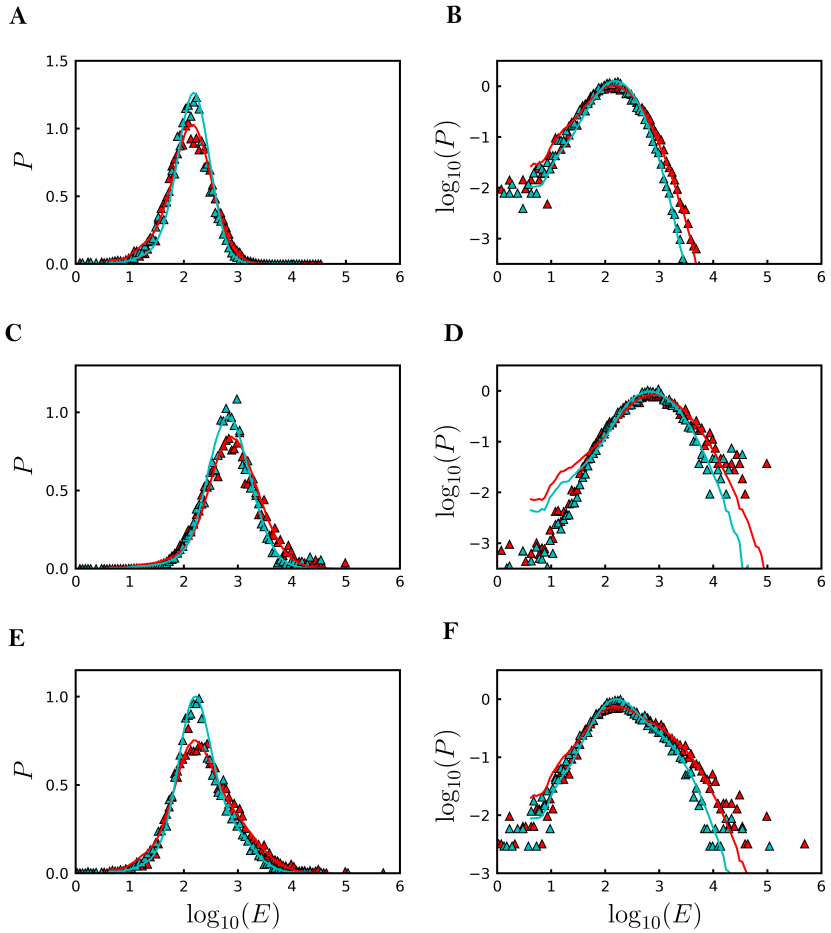

Figure 2.13 – Computed model simulations of the p.d.fs $P(E)$ of energy (k_BT) released during rupture events in normal (blue) and CML (red) cells. Ductile rupture events: **A.** semi-log representation ; **B.** log-log representation. Brittle rupture events: **C.** semi-log representation ; **D.** log-log representation. All rupture events: **E.** semi-log representation ; **F.** log-log representation. The triangles (△) represent the experimental data (Figures 2.7**E** and **F**); ductile and brittle rupture event log-normal p.d.fs were disentangled using a classical two-component Gaussian mixture model (see Eq. (2.22)). The curves represent the prediction of the released energy cacade model defined by Equation (2.24); the corresponding sets of model parameters are given in Table 2.1.

the experimental $P(\log_{10}(E))$ distributions (Fig. 2.13**A** and **B**).

To this end, the p.d.f.s corresponding to ductile and brittle rupture events were disentangled with a two-component Gaussian mixture model (see Eq. (2.22)). When fixing $\Delta E_0 = 12\,k_BT$ and $\Delta E^* = 4\,k_BT$ (characteristic unbinding energy of ABPs (Ferrer et al., 2008)), sets of parameters (a_0, $\hat{a}$, Δa) values (Table 2.1) were found providing quite satisfactory fits of the experimental released energy distributions (Figure 2.13).

For ductile rupture events, the mean number of cascade steps is rather limited for normal ($\overline{N} = 9$) as well as CML ($\overline{N} = 8$) cells and correspond to low mean released energy values $\overline{E}$ and likely to a few tens of ABP unbindings ($\overline{N}_{\text{unb}} \sim \overline{E}/(4\,k_BT)$) for normal ($\overline{E} = 191$ k_BT, $\overline{N}_{\text{unb}} \sim 48$) and CML ($\overline{E} = 198$ k_BT, $\overline{N}_{\text{unb}} \sim 50$) cells. For brittle failures, our model predicts more expanded cascades with significantly larger values of $\overline{N}$, $\overline{E}$ and $\overline{N}_{\text{unb}}$ for both normal ($\overline{N} = 21$, $\overline{E} = 1104$ k_BT, $\overline{N}_{\text{unb}} = 276$) and CML ($\overline{N} = 22$,

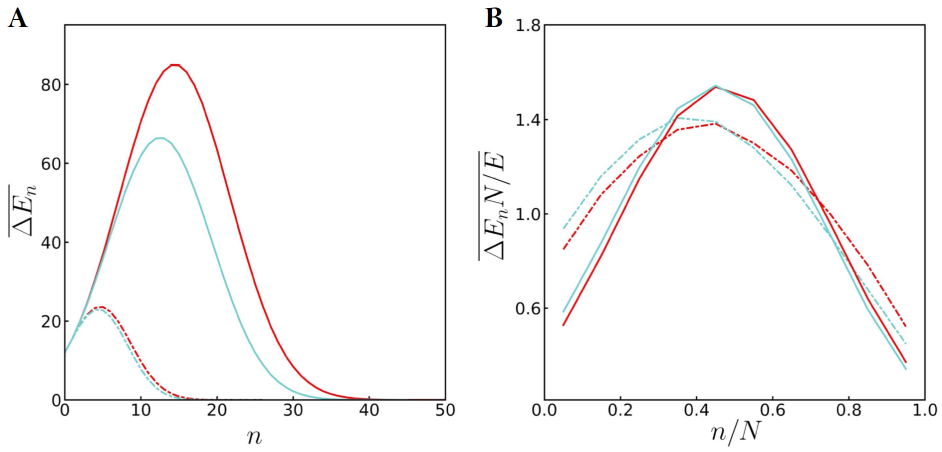

Figure 2.14 – Model predictions for the energy released rate per cascade step (k_BT/step). **A.** $\overline{\Delta E_n}$ *vs* n. **B.** $\overline{\Delta E_n/(E/N)}$ vs n/N. The continuous (resp. dotted-dashed) lines correspond to brittle (resp. ductile) rupture events. The colours correspond to normal (blue) and CML (red) cells. The corresponding sets of model parameters are given in Table 2.1.

$\overline{E} = 1494$ k_BT, $\overline{N}_{unb} \sim 373$) cells (Table 2.1).

Besides confirming that brittle rupture events involve more ABP unbindings than ductile ones, our rupture cascade model also predicts that the rate of energy released during the first cascade steps $\overline{\Delta E_n}$ and the maximum reached before the exponential decrease to zero for large n are much higher for brittle than for ductile rupture events (Figure 2.14**A**). Interestingly, when adimensionalizing ΔE_n by a mean released energy E/N per step and n by the total number of cascade steps (Figure 2.14**B**), the numerical data obtained for ductile events in normal and CML cells superimpose and display a slow increase from $\sim$0.9 to $\sim$ 1.3 during the first part of the cascade before decaying toward the energy cut off. For brittle events, the numerical data for normal and CML cells again superimpose (Fig. 2.14**B**), but now exhibit an initial three-fold increase from $\sim$0.5 to $\sim$1.5. This initial acceleration of ABP unbindings from $\sim$2-3 to reach values as high as $\sim$15-20 unbindings per cascade step in brittle events confirms the existence of some correlation between the ABP unbindings of successive cascade steps that likely results in a collective local disorganization and possible disintegration of the CSK. In comparison, the smoother released energy cascade with only a few ABP unbindings (up to 5) at each cascade step strengthens our interpretation of ductile rupture events as more dynamical and reversible stress-induced cross-linker unbindings that would confer to the cell ductile plasticity to large deformations.

We have simplified the discussion here by considering only one energy $\simeq 4\,k_BT$ for the actin cross-linker unbinding events, and grouping the brittle and ductile failures into two groups involving a different number of cross-linker unbinding events. Actually we could have also considered that the actin cytoskeleton network includes two populations of cross-linking proteins, namely weak cross-linking proteins that give a ductile failure and tight cross-linking proteins leading to brittle failures. The important physical quantity that is quantitatively predicted by the model is the total energy E released during a cascade of rupture events.

Numerical simulations performed with different threshold ΔE_0 (8,12 and 16 k_BT) and

arrest cut-off ΔE^* (0 and 4 k_BT) confirm the robustness of our rupture cascade model predictions.

2.7 Discussion

By developing a minimal rupture cascade model, we have identified two distinct populations of singular events in FICs that both correspond to mechanical failures of the CSK with correlated log-normal statistics of released energy. A first type of singular events is associated with rather moderate released energy ($\overline{E} \sim 200\,k_BT$) and likely corresponds to dynamical stress-induced cross-linker unbindings that confer to the cell a ductility preserving the perinuclear CSK architecture. In contrast, the second type of singular events is associated with more dramatic failures that release significantly higher energy ($\overline{E} \sim 1300\,k_BT$) as the signature of irreversible brittle disruption of the CSK integrity in highly tensed perinuclear zones. Besides providing quantitative robust modelling of the observed log-normal statistics, this model predicts that (i) the number of cascade steps, and in turn the number of cross-linker unbindings, is significantly higher in brittle than in ductile failures, (ii) the released energy increases during the first cascade steps until a maximum value that is significantly higher in brittle than in ductile failures, and (iii) the rate of released energy during these first cascade steps is also significantly higher in brittle than in ductile rupture events. This model further confirms that brittle failures are more frequently observed in CP-CML than in healthy cells as the signature of their higher mechanical fragility under large and fast strain. We anticipate that the mechanistic description provided by our minimal rupture cascade model will apply quite generally to other cell types in physiological and pathological situations (Streppa et al., 2018) and to other nonactive soft matter or solid systems such as biopolymer gels and glassy materials (Sharon and Fineberg, 1996). This minimal model is a very promising first attempt that will be used as a guide for future 2D or 3D simulations, aiming at elucidating the impact of CSK network architecture on the rupture mechanics of living cells (see Chapter 4).

Chapter 3

Phase diagram of avalanche size distributions for processes evolving with a continuous random reproduction rate

3.1 The recurrent problem of log-normal variable sums

In this chapter we study in more details a model strongly related to the model introduced in Section 2.5 and in (Polizzi et al., 2018). The goal is to understand how and under which conditions we obtain log-normal distributions as result of a sum of random variables, as was the case for the model used for cytoskeleton plasticity. We analyse a simplified version of that model, in order to better identify these conditions and be able to answer these questions. Since our modelling of ruptures cytoskeleton avalanches, together with experimental data, has shown log-normal distributions, we are particularly interested in the case where the resulting sum is log-normal, but, more generally, different sum distributions can be interesting for other applications. Therefore, understanding the complete behaviour of this model, along with the conditions causing possibly different distributions and different scenarios is important. Moreover, this can also add some information to our understanding of cytoskeleton avalanches.

We have seen that the summed variables z_n (see Section 2.5) converge rapidly to log-normals, as a result of the multiplicative process that generates them. Hence we can rephrase the original question in the connected one: what is the distribution of a sum of log-normal random variables? This can extend even more the interest of this study, since it is not only limited to the sum of cascade process events, but can also give some insights about the general problem of the sum distribution of log-normal variables.

In nature, many processes are based on cascade processes, that can be continuous, like geometric Brownian motion, or discrete, like branching processes, as the Galton-Watson branching process introduced in Section 1.4.1. Actually, all the processes involving a

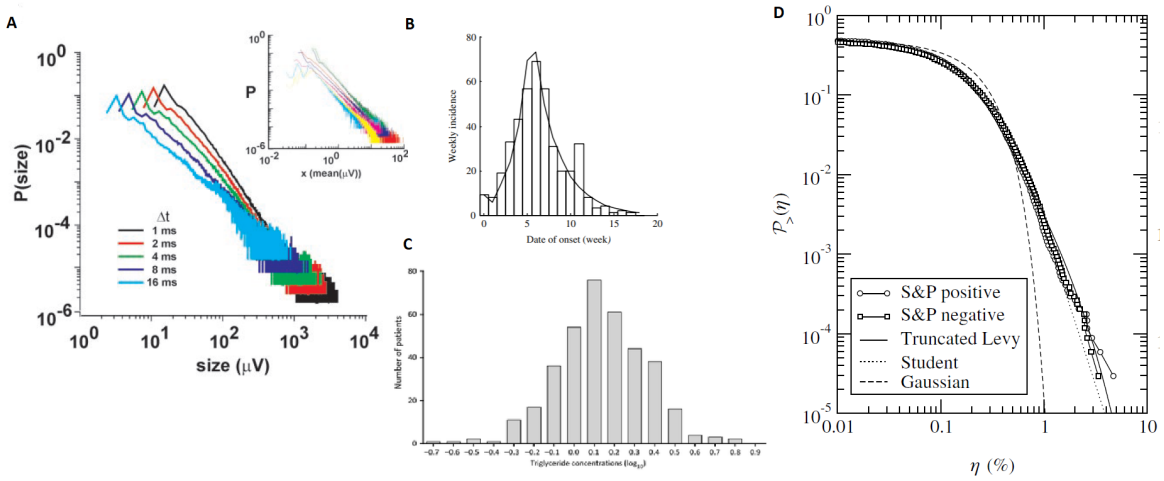

Figure 3.1 – Some examples of cases following multiplicative processes and/or log-normal distributions. **A.** Probability distribution of avalanches sizes in neocortical circuits, for different experimental time binning, showing a power law tail, and distribution of mean voltage for each separated culture (inset), adapted from (Beggs and Plenz, 2003). **B.** Epidemic size distribution for Ebola haemorrhagic fever of 2000 Uganda outbreak. Histogram of observed data and fitted model by a chosen compartmental epidemic model, adapted from (Legrand et al., 2007). **C.** Distribution of the $\log_{10}$ of the triglyceride concentration in human blood, showing a log-normal behaviour, adapted from (Choi, 2016). **D.** Stock price returns cumulative distribution of the S&P500 index computed over a time interval of 30 min. showing a fat tail distribution (truncated Lévy distribution), compared to other typically used distributions (like Gaussian), adapted from (Bouchard and Potters, 2003).

transfer of information and threshold processes are based on this multiplicative idea, when one unit exceeding threshold causes other units to do so in turn, generating a cascade over the large system of the full population. Examples range from microscopic world, like in biology avalanches of failures of cells cytoskeleton (Polizzi et al., 2018), avalanches in neuronal circuits (Lo et al., 2002; Beggs and Plenz, 2003; Lo et al., 2004; Dvir et al., 2018) or virus reproduction; to the social world, like epidemics spread over a population, news spread (or even fake news!) on social networks or cities growth; to electricity transmission: in electronic systems power transmission failures is a multiplicative process (Carreras et al., 2002); to the price evolution of stock options in finance (Black and Scholes, 1973); to the geophysical world, like the size of a rock falling apart, or to model the distribution of earthquake energies (Gutenberg and Richter, 1942, 1956); or even to, at larger scales, astrophysics, as a model of gravitational clouds growth (Sheth, 1996) (see Fig. 3.1 for examples of distributions resulting from these multiplicative processes).

Sometimes interesting quantities of these processes can be the sum of the singular events, thinking for example to the spread of the epidemics, one could be interested in the link between the distribution of how many people are infected at a given time and the distribution of the total size of the epidemic since the beginning up to that given time. Other times, an accessible physical quantity of these processes is the resulting sum of each singular event, as it is the case in electronics for the noise fluctuations of the power in a FM communication system (Fenton, 1960) and for most of applications regarding the microscopic world, where experimental limits often do not allow to reveal every singular

event. Naturally, the sum distribution of each cascade process event becomes of interest and is what we study here.

The problem of the sum of log-normal random variables is important itself, as attested by the number of publications (see for a review (Dufresne, 2008)), and it is relevant not only when there is an underlying multiplicative process, but also whenever the log-normal is a good model for experimental data. It is the case for instance for creating semi-conductor junctions for spin electronics applications (Kelly et al., 1999), where an increase of the size of the junction results in summing an increasing number of log-normal random variables, since in a single tunnel junction current fluctuations are well described by log-normal distributions (Da Costa et al., 2000). Indeed, the log-normal distribution is a good model in many fields (see also Section 1.3.2) and not always with an evident underlying multiplicative process, as in survival time distribution of certain cancers (pleural mesothelioma (Mould et al., 2004) and of larynx cancer (Mould et al., 2002)) or triglyceride concentration in human blood (Choi, 2016) and many other examples in biology (an example is shown in Fig. 3.1).

The fundamental problem of the distribution of the sum of log-normal random variables has been addressed intensively and became famous since the 70s, essentially when the Black Scholes model (Black and Scholes, 1973), in 1973, started being used for modelling option prices, placing the log-normal distribution at the center of interest in financial and economic sciences. Actually, it was a known problem in the field of electronic engineering and electronics already since the 60s, because the noise power of a radio jumps, shadowing of both, interfering and useful signals, and long-term receiver power fluctuations bringing to signal loss, follow all log-normal distributions. In these years there were the first unsuccessful attempts to have a closed form of the sum of either *independent identically distributed* log-normal random variables (Fenton, 1960) or by two *correlated* log-normal random variables (Naus, 1969; Hamdan, 1971). The former being based on the first two moments matching and the latter on the computation of the moment generating function. The basic problem of these attempts was to try to identify the distribution of the sum by methods involving the moment generating function, while the log-normal distribution, as has been proved about ten years later by Feller (Feller, 1971), is not uniquely defined by its moments. Therefore there are infinitely many other distributions with the same moments as the log-normal. For this reason, even if there was the idea that under some circumstances the sum of independent log-normal distributions is close to a log-normal, this was very hard to prove by all the methods involving moments. Since then anyway, finding a closed form of the distribution of the sum of log-normally distributed random variables is still an unsolved problem, even though today we have accessible and reliable numerical solutions. Furthermore, the basic approaches, with improved analytical and statistical (fitting) techniques, are in substance the same as in the first papers: mainly based on moments matching (see *e.g.* (Wu et al., 2005)) in order to find a good approximation of the distribution.

Usually, general results about the sum of random variables are obtained for independent random variables, often identically distributed. Here the novelty is that our sum is not only non-identically distributed, but also correlated, since each cascade event can be strongly correlated with the previous ones (for example, the number of today infected

people is strongly related to the number of yesterday infected people).

Of course dealing with independent random variables is simpler, and if sometimes is justified by experimental or theoretical observations, it is not always a good model for real life processes, even for options prices for instance, where the behaviour of low valued equities (called in financial jargon *penny stocks*) is essentially different from high and well affirmed stocks. This evokes some correlations between the increments of option prices and the prices themselves which is not taken into account in common financial models (Bouchard and Potters, 2003). Another situation where correlations have revealed their importance is the Asian option pricing model (Milevsky and Posner, 1998). Correlations between the log-normal random variables that compose the sum in not important only in finance, it arises also in other contexts where a cascade process well reproduces the mechanism of observed data, like, as we discussed above in biology, epidemics and physics, because in general in avalanche processes correlations among the summed variables are inevitable and built in the process itself.

The role played by correlations will be studied, although the problem of summing independent log-normal variables is itself interesting, as we said, and not solved up to now, since the log-normal distribution is not closed under convolution, on the contrary of normal distributions, and then only approximated results exist. Talking about sums of random variables, it is necessary to mention Lévy $\alpha-$stable distributions (Mandelbrot, 1960) for which a generalised central limit theorem exists (Gnedenko et al., 1954). In general $\alpha-$stable distributions are distributions such that summing variables with the same distributions gives as result a random variable having the same distribution shape, shifted and/or rescaled. The generalised central limit theorem gives an important asymptotic result for the limit where the number of summed variables tends to infinity, saying that the distribution of the sum will converge towards a Lévy $\alpha-$stable distribution. This is true only for non-correlated identically distributed random variables following a power-law tailed distribution, otherwise only the classical central limit theorem applies (with Lyapunov extension for not too dissimilar distributions). If the variables are not independent identically distributed this result does not apply and a general result for the sum of random variables does not exist.

3.1.1 Beaulieu's theorem: presentation and critical analysis

With this background we thought first that the reason why we found log-normally distributed sum was because of correlations. It turned out that a result about the sum of correlated log-normal random variables has been published recently by Beaulieu (Beaulieu, 2012). Since we will try to apply this theorem in the following, we give here the main lines of the proof.

Let $\{z_i\}_{i=1}^N$ be correlated log normal random variables with underlying Gaussian correlated random variables $\{y_i\}_{i=1}^N$, therefore $z_i = e^{y_i}$. Let $\mathbb{E}[z_i] = m_i$, $\mathbb{E}[(z_i - m_i)^2] = s_{z_i}^2$ and $\mathbb{E}[y_i] = \mu_i < \infty$ and $\mathbb{E}[(y_i - \mu_i)^2] = \beta_i^2$ with $0 < \beta_i^2 < \infty$. If the variables y_i are

correlated in the following particular way:

$$\rho_{ij} = \frac{\mathbb{E}[(y_i - \mu_i)(y_j - \mu_j)]}{\beta_i \beta_j} = \alpha_i \alpha_j, \qquad 0 < \alpha_i < 1 \qquad \forall i. \tag{3.1}$$

a correlation coefficient written in this form means essentially that it can be separated in a part depending only on i and a part depending only of j. With these assumptions the theorem states that with probability 1 the distribution of the normalised sum $\overline{Z} = \sum_{i=1}^{N} z_i/N$ of the correlated log-normal random variables has a limit distribution which is log-normal with mean m_Z and variance s_Z^2 given by:

$$\begin{aligned} m_Z &= \lim_{N\to\infty} \frac{\sum_{i=1}^{N} m_i}{N} \\ s_Z^2 &= m_Z^2 \left(e^{\lambda^2} - 1\right) \end{aligned} \tag{3.2}$$

where

$$\lambda^2 = \beta_i \beta_j \alpha_i \alpha_j. \tag{3.3}$$

This is how the theorem was stated by Beaulieu, in his paper (Beaulieu, 2012), but actually some hypotheses of the theorem can be made lightly more general, without impacting too much the proof. The condition on the variance of y_i, $0 < \beta_i^2 < \infty$ can be released, leading to possibly infinite variance, if the average converges fast enough as will be specified hereafter. Also, the condition on the correlation (3.1) can be stated by just saying that the covariance matrix has to be constant and strictly positive (> 0), without necessarily writing it in the form $\alpha_i \alpha_j$. Indeed, the parameter λ^2 is nothing else that the covariance matrix for the variables y_i, which is therefore composed by positive elements all equal to $\lambda^2 \neq 0$.

The heart of the proof is to be able to isolate the covariance while summing the y_i. This is possible only if we can write

$$y_i = \lambda X_0 + X_i. \tag{3.4}$$

Since y_i are Gaussian random variables so has to be the $\{X_i\}_{i=0}^{N}$. We impose also that the $\{X_i\}_{i=0}^{N}$ are independent, their means are $\mathbb{E}[X_i] = \mu_i$ and define their variances as $\sigma_i^2 = \mathbb{E}[(X_i - \mu_i)^2]$. The variable X_0 is chosen such that it has a mean $\mu_0 = 0$ and a variance $\sigma_0^2 = 1$. Thank to Eq. (3.1) the variables y_i can be expressed as a linear transformation of independent Gaussian random variables and the covariance between the y_i is all contained in the term λX_0. We can easily check that with the previous definition of μ_0 and σ_0^2 an explicit calculation of the covariance gives the correct λ^2 value because the only term which survives of the product $(\lambda X_0 + X_i)(\lambda X_0 + X_j)$ after the expected value operator is $\mathbb{E}[\lambda^2 X_0^2] = \lambda^2$.

Leaving aside some details, we are interested in the distribution of the sum, $\overline{Z}$, that we can now write:

$$\overline{Z} = \frac{1}{N} e^{\lambda X_0} \sum_{i=1}^{N} e^{X_i}. \tag{3.5}$$

Here is where it is important to have a constant covariance matrix, *i.e.* λ^2 not depending on the index i.

Now by using the strong law of large numbers due to Kolmogorov (Serfling, 2002), defining $V_i = e^{X_i}$, with variance $\sigma^2_{V_i}$, one has:

$$\lim_{N\to\infty} \frac{\sum_{i=1}^{N} e^{X_i}}{N} = \frac{\sum_{i=1}^{N} \mathbb{E}[V_i]}{N} = \bar{v} \tag{3.6}$$

with probability 1 if the variables V_i are independent, which is true by definition of X_i and if the series $\sum_{i=1}^{N} \frac{\sigma^2_{V_i}}{i^2}$ converges. It can be proved that this series converges provided that the sequence of variances $\{\sigma_{V_i}\}_{i=1}^{N}$ is bounded by a maximum value called

$$\sigma^2_{max} = \max_i \left\{\sigma^2_{V_i}\right\} < \infty \tag{3.7}$$

by the formula of the variance of a log-normal random variable, like V_i, this condition can be written in term of the mean and the variance of X_i:

$$\sigma^2_{max} = \max_i \left\{e^{2\mu_i+\sigma_i^2}\left(e^{\sigma_i^2}-1\right)\right\} < \infty. \tag{3.8}$$

We note that this inequality is verified in the hypotheses of the theorem, because μ_i and β_i are bounded, σ_i^2 being, from a short calculation, $\sigma_i^2 = \beta_i^2 - \lambda^2$. Actually, from Equation (3.8) we can see that what needs to be bounded for the theorem to be true is the sum $2\mu_i + 2\sigma_i^2 = 2\mu_i + 2\beta_i^2 - 2\lambda^2$, therefore as far as

$$\mu_i < \frac{\lambda^2}{\beta_i^2} \tag{3.9}$$

condition (3.8) is verified and the proof of the theorem is valid. This allows the sequence β_i to tend to ∞, if the corresponding μ_i is negative. To conclude the proof combining (3.5) and (3.6) yields:

$$\lim_{N\to\infty} \overline{Z} = \lim_{N\to\infty} e^{\lambda X_0}\bar{v} = B \tag{3.10}$$

where B is clearly a log-normal random variable being the exponential of the Gaussian variable X_0 rescaled by a constant factor. Continuing the calculation gives a mean and a variance of B as stated in the theorem.

To conclude we would like to state a few remarks:

Remark 1: We point out that here variables do not need to be identically distributed, however the condition on the constant covariance matrix is quite strict, and in many real life applications it may be not verified.

Remark 2: The limit distribution of $\overline{Z}$ is of course the same as the distribution of the non normalised sum $Z = \sum_{i=1}^{N} z_i$, with the difference of an average and a variance rescaled respectively by a factor N and N^2.

Remark 3: The Central Limit Theorem is not in danger, because if the variable are uncorrelated $\lambda^2 = 0$ and then from (3.4) the y_i are also uncorrelated, and the sum should converge towards a Gaussian. This is what is stated by Beaulieu, but probably the consequence of taking first the limit in (3.6) than the limit in (3.10) should have

been discussed more, appearing crucial to have eventually a log-normal. This particularly jumps out in the $\lambda^2 = 0$ case, when doing the same as in the proof would give $\lim_{N\to\infty} \overline{Z}$ equal to a constant, while we know it has to be Gaussian.

3.1.2 Some useful results about the sum of uncorrelated log-normal random variables

In this section we give few results on the sum of uncorrelated and identically distributed log-normal random variables, that will help to understand the following sections. The results given here are, hopefully, given in an intuitive and qualitative way without aiming mathematical rigour, for details see given references and references thereon.

A first point to discuss is the behaviour of the sum with respect to its deviation from the Central Limit Theorem and the law of large numbers (Serfling, 2002), saying that a sum of a number N of random variables will converge to the sum of their average values for large n. It is easy to see that a log-normal distribution

$$p(x) = \frac{1}{\sqrt{2\pi}\sigma x} e^{-\frac{(\ln(x)-\mu)^2}{2\sigma^2}} \tag{3.11}$$

in the limit $\sigma \ll 1$ is very close to a Gaussian distribution with average e^{μ} and variance $(\sigma^2)\, e^{2\mu}$, that is essentially like saying that $\lim_{\sigma\to 0} e^{(t\sigma)} = 1 + t\sigma$, being t the statistical fluctuation around the mean of $\ln(x)$, and so taking the exponential of a log-normal random variable gives the same distribution of the variable $\ln(x)$, that is by definition a Gaussian (see Section 1.3.2).

From this, it follows easily that since the log-normal is close to a Gaussian, the sum will also be very close to a Gaussian. In turn, the sum of Gaussian random variables gives Gaussians. Moreover being the distributions in the limit $\sigma \ll 1$ very narrow, the law of large numbers is also verified, even for a small number of terms.

As far as the variance σ^2 increases things become more complex. In an article by Romeo *et al.* (Romeo et al., 2003) two main regimes are identified, for a finite number of summands N, since the limit $N \to \infty$ is known as Gaussian for the Central Limit Theorem, having the log-normal finite moments of all orders:

- a regime where $\sigma^2 \lesssim 1$ in which the sum is supposed to be log-normal for every N, with parameters μ_N and σ_N^2 depending on N (it is an assumption not supported by any analytical calculation, which does not exist so far). By identifying the first two moments of the distribution of the sum, $N\langle x\rangle$ and $N\mathrm{Var}(x)$, with the first two moments of the log-normal distribution (moments matching), they could therefore find a dependence of the parameters and in turn of the typical value of the sum on N. This dependence turns out to be different from the law of large numbers expected one, $(N\langle x\rangle)$ by a factor $1/\left(1 + C^2/N\right)^{3/2}$, C being a constant.

- a regime where $\sigma^2 \gg 1$ in which the behaviour is very complex. The reasoning is that the sum is dominated by the largest term and therefore the cumulative distribution of the sum is approximated by the cumulative distribution of the maximum value composing the sum. From this they could find an unsolvable equation (Eq (51) of

(Romeo et al., 2003)) for the typical value of the sum. Adding to this heuristically a moment matching condition, supposing again the distribution of the sum as log-normal with parameters μ_N and σ_N^2, they could find a value even more different from $N\langle x\rangle$ than in previous case (the corrective factor being here $e^{-\frac{3\sigma^2}{2N^{\ln(3/2)/\ln 2}}}$). This last factor was obtained by purely heuristic and empirical arguments adjusting the approximate calculations, since they led to some striking problems in the limit $N \to \infty$, not recovering the law of the large numbers limit.

However the study is focused on the effect of the finite sum on the typical sum, *i.e.* the maximum of the distribution, but nothing is said (probably it was not the aim) on the distribution of the sum and on the error done by the approximating it with log-normals.

The analysis carried on by this article should teach us mainly two things. First that the problem of the sum of log-normals, can be very complex, apart from the not very useful $\sigma \to 0$ case. That is because log-normal distributions have at the same time properties of broad fat tail distributions and of narrow distributions, being all the moments finite. Therefore often, the best possible solutions are approximated and/or only empirically supported, thing that makes difficult a generalisation. This brings us directly to the second point. The convergence to the Central Limit/law of large numbers result can be very slow and intermediate situations for finite N can be of very significant impact for natural processes, whenever N cannot be too large. These limitations exist even in the simplest case of **uncorrelated** and **identically distributed** random variables, we will see how correlations can modify the scenario, along with sifting and/or rescaling summands distributions, the latter potentially mixing the effects of the different regimes.

Another result which is in some sense related this, is the one from Mouri's article (Mouri, 2013). The reasoning is to write the sum as a product, obtaining as a final distribution a log-normal if some conditions are verified. The result is obtained in the large number of summands N limit, and again for non-correlated random variables. Finally, under these conditions the distribution will converge to the case when the log-normal is not distinguishable from a Gaussian and then the Central Limit Theorem is verified.

The main idea of the paper (Mouri, 2013) is to consider the distribution of the average $\overline{Z} = 1/N \sum_{i=1}^{N} z_i$ in terms of the fluctuations ϵ_i around the mean $\langle z_i\rangle = \langle z\rangle (1+\epsilon_i)$, considering z_i as identically distribute. Then, $\overline{Z}$ is written as a multiplicative process:

$$\overline{Z} = \frac{1}{N}\sum_{i=1}^{N}\langle z_i\rangle (1+\epsilon_i) = \langle z\rangle \left(1 + \frac{\epsilon_1}{N} + \frac{\epsilon_2}{N} + \cdots + \frac{\epsilon_N}{N}\right) \simeq \langle z\rangle \left[\prod_{i=1}^{N}(1+\epsilon_i)\right]^{1/N}. \quad (3.12)$$

Therefore the arithmetic average is approximated as the geometric average and the logarithm of $\overline{Z}$ can be written as a sum of uncorrelated random variables:

$$\ln \overline{Z} = \frac{1}{N}\sum_{i=1}^{N} \ln z. \quad (3.13)$$

All the heart of the question now is to understand if the convergence to the Gaussian Central Limit Theorem result is faster for $\ln\left(\overline{Z}\right)$ or for the original average $\overline{Z}$, which

both converge to a Gaussian for $N \to \infty$. It turns out that for $\overline{Z}$ to converge faster to a log-normal distribution than to a normal one, there are two necessary conditions:

1 − the skewness of $\ln(z)$ is smaller that the skewness of z:

$$\left| \frac{\langle (\ln z - \langle \ln z \rangle)^3 \rangle}{\langle (\ln z - \langle \ln z \rangle)^2 \rangle^{1.5}} \right| < \left| \frac{\langle (z - \langle z \rangle)^3 \rangle}{\langle (z - \langle z \rangle)^2 \rangle^{1.5}} \right| \tag{3.14}$$

2 − the kurtosis of $\ln(z)$ is smaller that the kurtosis of z_i:

$$\left| \frac{\langle (\ln z - \langle \ln z \rangle)^4 \rangle}{\langle (\ln z - \langle \ln z \rangle)^2 \rangle^2} \right| < \left| \frac{\langle (z - \langle z \rangle)^4 \rangle}{\langle (z - \langle z \rangle)^2 \rangle^2} \right| \tag{3.15}$$

This two conditions are only necessary, and may be sufficient if we neglect larger moments, this anyway should be done with care. Intuitively, they just mean the skewness and the kurtosis of $\ln\left(\overline{Z}\right)$ are closer to the Gaussian value of 0 than skewness and kurtosis of $\overline{Z}$. However, before converging to log-normals, the distribution of the sum can be very far from both, normal and log-normal, and will strongly depend on the initial distribution and on N (we could deduce this from Fig. 1 of (Mouri, 2013)). As a comment, we tried to do some Monte-Carlo simulations of the example given in paper (Mouri, 2013), about the gamma distribution, and practically the difference between log-normal and Gaussian distribution for the sum was not appreciable. The goodness of fit was statistically comparable between both distributions, for different N, meaning that essentially we are in the case where Gaussian and log-normal are very close, either because $\sigma \to 0$ or because the ratio $\sigma/\mu \to 0$ (see Section 1.3.2).

In any case there are situations, again for finite values of N, that can be important in nature, where the behaviour of the sum can be log-normal before being Gaussian because of the Central Limit Theorem, and the convergence to a Gaussian can be very slow. It is therefore relevant to understand not only the asymptotic behaviour, but also the finite N range. This is the main conclusion which unifies both discussed articles.

3.2 Model definition and general behaviour

In this section we define the model in a formally precise way and we give a schematic view of the most important results, which are derived more rigorously in the following.

As a reminder, we try to give some insights about our main question, that is why and under which conditions the sum of random variables can be approximated by a log-normal distribution. What are the key ingredients to put for this to happen? Is it a limiting property or a consequence of a finite sum of terms? Is it a consequence of correlations between the summed random variables or not?

Let us call Z the sum of random variables z_i, which are resulting from a stochastic multiplicative process, strongly related to the one of (Polizzi et al., 2018) and of Section 2.5.

Defining the z_i formally we have:

$$z_{i+1} = a_i z_i = a_i z_0 \prod_{j=0}^{i-1} a_j, \tag{3.16}$$

where the starting point will always be $z_0 = 1$, as usual for branching processes (a different one would only lead to a time shift). The random variables a_i are the growth (or reproduction) factors and are independent each other, they have mean $\bar{a}$ and variance Δa^2. For the moment we do not specify their distribution, but we can say that for the model for cytoskeleton ruptures avalanches (Polizzi et al., 2018) it was a Gaussian with an exponentially decreasing mean: $a_i \sim \mathcal{N}(a_0 \exp(-i/\hat{a}), \Delta a^2)$. All the a_i are positive and therefore also the z_i are positive. This is imposed by the physical phenomenon that we are modelling, *i.e.* in general avalanche or growth processes. Moreover the process generating the z_i can be seen as if all the z_i random variables were the result of a generalised branching process stopped at time i. The generalisation with respect to the Galton-Watson branching process introduced previously (see Section 1.4.1) comes from the fact that here our sequence $\{z_n\}$ is a sequence of continuous, still positive, random variables. This is because the number of Young, called now a, here has a continuous distribution with constant average, $\bar{a}$ and variance Δa^2. In principle as in the cascade model of (Polizzi et al., 2018) the average $\bar{a}$ could also depend on time, but here we do not consider this case. We are interested in the distribution of the sum:

$$Z = \sum_{i=1}^{N} z_i. \tag{3.17}$$

It is useful to define also the variables y_i:

$$y_i = \ln(z_i) = \sum_{j=0}^{i-1} \ln a_j. \tag{3.18}$$

The link with the general problem of the sum of log-normal random variables is the argument that the variables z_i, as shown in Section 2.5 (for more details see also Section 1.3.2), converge quite fast too log-normally distributed random variables. Therefore the z_i are such that when i reaches a certain value, they can be written as:

$$z_i \sim \text{Log-Normal}(\mu_i, \beta_i^2) \tag{3.19}$$

The variables z_i have mean $m_i = \mathbb{E}[z_i]$ and variance ${s_{z_i}}^2 = \mathbb{E}[(z_i - m_i)^2]$. As a memorandum, we say that a variable X follows a log-normal distribution with parameters μ and σ^2 if $\ln(X)$ follows a Gaussian distribution with mean μ and variance σ^2. The log-normal probability density function for the variables z_i is:

$$p(z_i) = \frac{1}{\sqrt{2\pi}\beta_i z_i} e^{-\frac{(\ln(z_i)-\mu_i)^2}{2\beta_i^2}}$$

Then, the variables $y_i = \ln(z_i)$ follow approximately a normal distribution:

$$y_i = \ln(z_i) \qquad y_i \sim \mathcal{N}(\mu_i, \beta_i^2) \tag{3.20}$$

The parameters μ_i and β_i are linked to m_i and s_{z_i} in some way that will be specified afterwards, but for the moment that is not important for our purposes.

The sum Z is therefore a random variable given by the sum of other random variables (approximately log-normal) and our goal is to understand something about the asymptotic probability distribution of Z in the case where the z_i are correlated in a specific way, as result of Eq. (3.16), and when we sum a large number of them. In this case the Central Limit Theorem does not apply and we cannot state that the distribution of Z is Gaussian (see Fig. 3.3 and Section 3.7 and for a comparison with the non-correlated case).

We discuss now the main results of the model. Here we consider the case where a_i are all equally distributed and follow a uniform distribution of first and second moment respectively $\bar{a} = \mathbb{E}[a]$ and $\tilde{a} = \mathbb{E}[a^2]$. We can thus write:

$$a_j \sim \mathcal{U}_{[b;c]} \qquad \text{with} \qquad b, c > 0, \tag{3.21}$$

where b and c are the lower and the upper bounds of the uniform distribution. Since the uniform distribution is a two parameters distribution, b and c are related to $\bar{a}$ and $\tilde{a}$, but the results are given with respect to these two last parameters because they are more easily generalisable to other distributions and gives an immediate and intuitive idea of the growth parameter value.

The choice of the distribution of a is discussed in further sections, but the uniform distribution allows us to have some analytical results and captures the essence of the model. Other distributions, called *admissible* in the following, can be chosen (for example a Gaussian), if they respect some assumptions (basically regarding symmetry and width), without changing significantly the global picture. Also, many results are completely independent of this choice.

Figure 3.2 gives a schematic picture of the results. The coloured region is the one allowed for the parameters of a uniform distribution. The lower bound $\tilde{a} = \bar{a}^2$ is simply the limit condition where the variance of the variable a is zero. This bound is common for whatever choice of the distribution of a. On the other hand the upper bound $\tilde{a} = \frac{4}{3}\bar{a}^2$ is a result of the positivity constraint and is proper to the uniform distribution choice, but with a different choice it would be slightly different, even though it may not be exactly analytically computed (already for a Gaussian is not). These two red lines give therefore the 2 opposite limit behaviour: the first one, for $\tilde{a} = \bar{a}$ gives as asymptotic distribution for Z a Dirac δ function, being in this case the process deterministic; the second one, $\tilde{a} = \frac{4}{3}\bar{a}^2$ gives a distribution that decreases like Z^{-1} al large Z, if Z is not bounded this will yields in the limit $N \to \infty$ the asymptotically flat distribution. However, if Z is bounded, as we will see, a full analytic result can be found, and even if it is not bounded the convergence to a flat distribution can be very slow and for finite N the distribution will be the same as in the bounded case.

The region in blue, delimited above by the line $\tilde{a} = 1$,is the Finite Scale one. With this

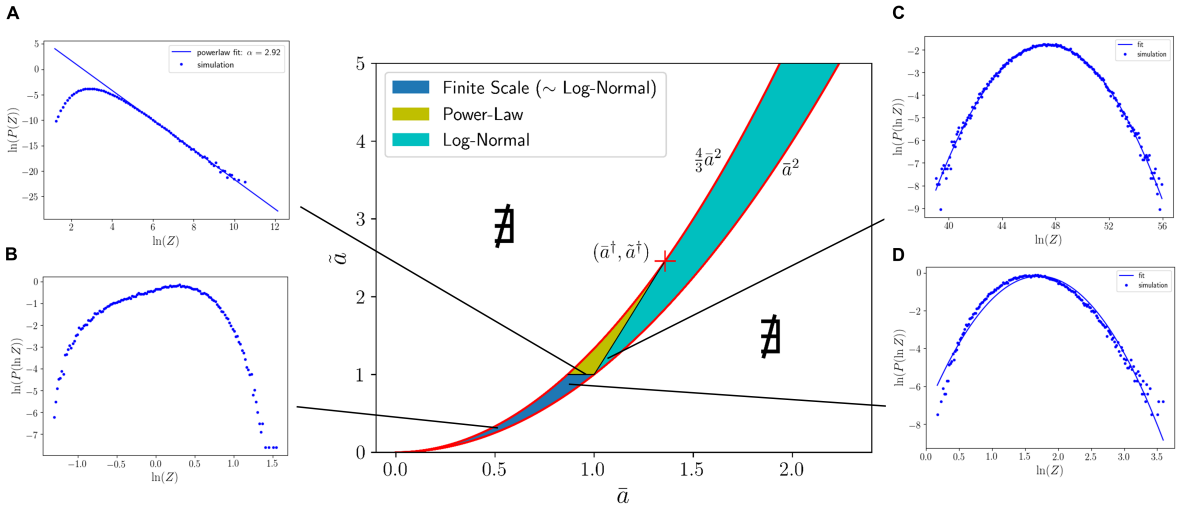

Figure 3.2 – Schematic visualisation of the results of the model defined in the text. The blue region (Finite Scale) is the region where both $\tilde{a}$ and $\bar{a}$ are < 1 and then the mean and the variance of Z converge both to a finite value. The yellow region (Power Law) is the region where Z follows a power law tailed distribution, with a tail exponent going from 3 to 1, at the critical point $(\bar{a}^\dagger, \tilde{a}^\dagger)$. The cyan region (Log-Normal) is the region where the asymptotic distribution is log-normal. The red delimiting lines are the two limit solutions of asymptotically flat and δ distribution (see text). Outside of this coloured region the parameters are not allowed, because of the uniform distribution choice (for other admitted choices the bottom border would be the same, but the upper one could be slightly different and may not be analytically solvable). Examples of resulting distributions are shown for each region (see text for details). For each panel the coordinates in the $(\bar{a}, \tilde{a})$ plane are: panel **A.** $(0.98, 1)$; panel **B.** $(0.55, 0.34)$; panel **C.** $(1.1, 1.2)$; panel **D.** $(0.85, 0.76)$.

we mean that the distribution of Z has a scale, in the sense that all its moment are finite, as opposed to *scale free* distributions which does not have a typical scale, like power laws. Here Z is given by the sum of a finite number of terms, even if $N \to \infty$, because the z_i converge fast to zero and after a certain N^* they do not contribute to Z. In this region the distribution is dependent on N^* but can be very close to a log-normal, as supported by the uncorrelated sum of log-normal random variables (see Section 3.1.2). In panel **B** an example of distribution for coordinates $(0.55, 0.34)$ is shown. The distribution of the $\ln(Z)$ is plotted to show the similarity to a log-normal distribution (a parabola here), but we can see a discrepancy between the parabolic fit and the distribution for values of $\ln(Z) < 0$. This shoulder disappears in panel **D** (coordinates $(0.85, 0.76)$), where the distribution is closer to a log-normal, but still with statistically significant differences. In both cases all the moments of the distribution are finite.

The region in yellow (Power-Law) is the region where the asymptotic Z distribution has a power-law (or Paretian) tail $Z^{-\alpha}$, with a decreasing exponent α ranging from 3 (when $\tilde{a} = 1$) to 1 (at the critical point $\left(\bar{a}^\dagger, \tilde{a}^\dagger\right)$). This point is the largest point in the parameters plane for which the Z distribution is a power-law, and its coordinates have an analytical expression. In panel **A** is shown the distribution for coordinates $(0.98, 1)$, along the critical line between the blue and the yellow region. The distribution of Z has an evident power-law tail.

Finally, the region in cyan (Log-Normal) is where the asymptotic Z distribution is a log-normal with a very solid statistical and analytic agreement. Of course the region

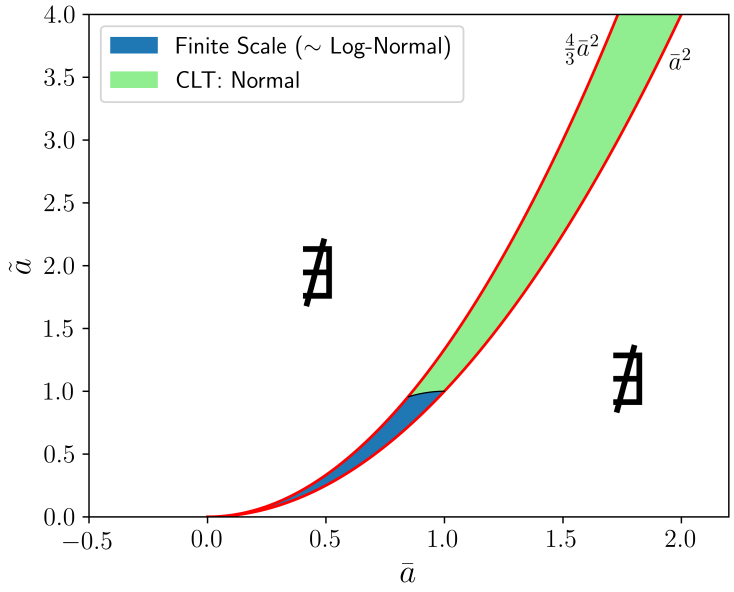

Figure 3.3 – Schematic view of the asymptotic distribution for the surrogate model. The upper and lower bounds are the same as in Fig. 3.2. The blue region (Finite Scale) is where we cannot apply the Central Limit Theorem, and therefore the asymptotic Z distribution can be similar to a log-normal. The green region (CLT) is where the Central Limit Theorem (Lyapunov version) can be applied and the asymptotic Z distribution is normal. The line delimiting the two regions is defined analytically.

continues out of the plot for larger $\bar{a}, \tilde{a}$. Panel **C** shows the distribution just after the critical line between the yellow and the cyan region, for coordinates $(1.1, 1.2)$. The distribution of $\ln(Z)$ is perfectly parabolic and therefore Z is log-normal, with solid statistical evidence. However the distribution of $\ln(Z)$ is no more stationary and will shift to the right and increase its width while N increases. We can still define an asymptotic distribution since the shape does not change with N.

3.2.1 Phase diagram removing correlations

We summarise here the results for a surrogate model where the z_i have the same distributions as in our model (defined by Eq. (3.16)), which are approximated by log-normals, with parameters μ_i and β_i. In the surrogate model we remove correlations, therefore the problem reduces to a sum of independent log-normal random variables, with parameters depending on the time step of the avalanche. Even by removing correlations, as we will see with more details in Section 3.7 the fact that the parameters depend on time makes the problem not as trivial as we could firstly think.

In Figure 3.3 we observe a phase diagram of the asymptotic Z distribution. We can see that the diagram is simpler: only two different regions are represented. The blue region is similar to the one of Fig. 3.2 in the sense that here also the resulting distribution can be close to a log-normal and all its moments are finite, but removing correlations makes the distributions much more narrow. A difference with respect to the blue region of Fig. 3.2 is that in all the region, no matter the parameters value the Z distribution is very close to a log-normal. This is a consequence of the definition of the z_i which are

here defined as exact log-normal random variables and do not have a transient before convergence. On the other hand the green region is where the Lyapunov version of the Central Limit Theorem (version for *moving* distributions) can be applied, and therefore the final distribution of Z will be a Gaussian. This can be proved analytically, but numerically, since the convergence happens to be very low, it cannot be seen. In any case the distributions obtained numerically are totally different from the ones obtained previously in the correlated model.

Finally the delimiting line between the blue and the green region can be computed analytically and is not exactly, as in Fig. 3.2 the horizontal line $\tilde{a} = 1$. The difference comes from the approximation of z_i as log-normals, which is true within some error bounds. However this line is exact if we consider, as we do here, the distribution of the sum of log-normal random variables.

3.3 Beaulieu's theorem approach

In this section we show an attempt to generalise the proof of Beaulieu's theorem (see Section 3.1.1) to our model, to try to justify the rise of log-normal distributions for the sum Z, as was for the model of ruptures cascades of the cytoskeleton (see (Polizzi et al., 2018)). We also give a definition of *admissible* distribution and compute the parameters of the y_i distributions and their correlations. It turns out that the theorem cannot be used to prove the convergence to log-normals, therefore the reasons need to be found somewhere else.

In order to approximate the average of a function of a random variable and its square, we use a Taylor expansion around the mean value and then we take the expected value. So in general if X is a random variable and $f(X)$ is its transformed variable we can approximate $\langle f(X)\rangle = \mathbb{E}[f(X)]$ and $\langle f^2(X)\rangle$ in the following way. The second order Taylor expansions are:

$$f(X) = f(\overline{X}) + f'(\overline{X})(X-\overline{X}) + f''(\overline{X})\frac{(X-\overline{X})^2}{2} + o((X-\overline{X})^2) \tag{3.22}$$

$$f^2(X) = f(\overline{X})^2 + 2f(\overline{X})f'(\overline{X})\left(X-\overline{X}\right) + \left(f(\overline{X})f''(\overline{X}) + f'(\overline{X})^2\right)\left(X-\overline{X}\right)^2 + \\ +o((X-\overline{X})^2) \tag{3.23}$$

We can apply to both sides of both equations the expected value operator and, considering that $\mathbb{E}[X-\overline{X}] = 0$ and that $\mathbb{E}\left[\left(X-\overline{X}\right)^2\right] = \mathrm{Var}(X)$, we get up to second order :

$$\langle f(X)\rangle = f(\overline{X}) + f''(\overline{X})\frac{\mathrm{Var}(X)}{2} + \mathbb{E}[o((X-\overline{X})^2)] \tag{3.24}$$

$$\langle f^2(X)\rangle = f(\overline{X})^2 + \left(f(\overline{X})f''(\overline{X}) + f'(\overline{X})^2\right)\mathrm{Var}(X) + \mathbb{E}[o((X-\overline{X})^2)] \tag{3.25}$$

This approximation is valid if the expected value operation is taken over an interval that contains all the statistic relevance of X. In other words it breaks down if the function $f(X)$ is significantly non-quadratic (2nd degree polynomial approximation) in a region around the mean $\overline{X}$ of a size comparable to the standard deviation of X, $\sqrt{\mathrm{Var}(X)}$. We need also to require that we can control (*i.e.* they don't diverge) **all** the central moments of X.

We can write the error of the second order approximation as:

$$\mathcal{E}_2 := \left| \sum_{n=3}^{\infty} \frac{f^{(n)}(\overline{X})}{n!} \mathbb{E}[\left(X - \overline{X}\right)^n] \right| \leq \sum_{n=3}^{\infty} \frac{\left| f^{(n)}(\overline{X}) \right|}{n!} \mathbb{E}\left[\left| X - \overline{X} \right|^n \right] \tag{3.26}$$

We can then see explicitly that all the moments have to go to zero and the support of the random variable has to be strictly in the radius of convergence of the Taylor series of the function f.

For most common distributions (Poisson, exponential, Gaussian, gamma etc.) and for function f not having singular points (like exponential, polynomials, etc.) this means that $(\mathrm{Var}(X))^n \to 0$ as $n \to \infty$, and an estimate of the error if the distribution is symmetric is of the order of $(\mathrm{Var}(X))^2 f^{(4)}(\overline{X})/4!$ if we go up to third order (since the third central moment of a symmetric distribution is zero).

Let us now consider to apply this to our model. As we said, at big i, the distribution of z_i converges to a log-normal because of central limit theorem, if $\langle \ln(a) \rangle$ and $\mathrm{Var}(\ln(a))$ are finite. This will always be true if we want the error in (3.26) to converge to 0 for a function $f = \ln(\cdot)$. Therefore if we respect condition (3.26) we automatically have the z_i convergent toward a log-normal. That is because if a has a compact support (it has to be inside the radius of convergence of the Taylor series) also the average and the variance of $\ln(a)$ will be finite, even for a close to 0 because of the derivative $e^{\ln(a)}$ that comes out with the change of variable that leads to an exponential cut-off. This is the reason why we can consider the z_i log-normally distributed.

As we pointed out before for a standard branching process (see Section 1.4.1), the sequence $\{z_n\}$ is unstable also in this generalised case; we will then choose for now the distribution of a in such a way that on average the process will die out, *i.e.* on average $z_n \to 0$. But if in the Galton-Watson process the condition $m \leq 1$ was sufficient for the process to go to zero, here the condition $\bar{a} < 1$, that is still necessary for convergence, does not exclude that for some realisation of the process z_n can diverge, even though this will happen with a time exponentially decreasing probability.

We can now apply the approximations stated before to the variables $X = a_j$ taking as function $f(X)$ the logarithm $\ln(X)$. Before let us notice that in general we do not have any restriction on the distribution of a_j, as far as we verify the condition stated above (Eq. 3.26), so all the results should work for several different distribution of a, as for example the uniform, Gaussian or Gumbel distributions (even if not symmetric in general if they fulfil some conditions of small width they are admissible). For the logarithm and its square the condition to be fulfilled can be translated in:

$$\frac{\Delta a^4}{\bar{a}^4} \ll 1, \tag{3.27}$$

where the 4-th power has to be replaced by a 3-rd power if the distribution is not symmetric. This comes out from the fact that the n-th derivative of the logarithm is of the order of $1/x^n$. Condition (3.27) implies also that if $\Delta a/\bar{a}$ is less than one the error in (3.26) converges as far as the a_j is strictly in the interval $]0, 2\bar{a}]$, corresponding to the radius of convergence of the Taylor series for the logarithm.

In this section a_j will follow a Gaussian distribution $a_j \sim \mathcal{N}(\bar{a}, \Delta a^2)$, for which condition (3.27) is verified taking $\Delta a < \bar{a}/4$, giving an error at most of the order of 10^{-3} and assuring that a will stay within the radius of convergence almost with probability one (we can always truncate the Gaussian if this is not the case).

Then developing the computations we obtain these 2 formulas that will be useful in the following:

$$\langle \ln a_j \rangle = \ln \bar{a} - \frac{1}{2}\frac{\Delta a^2}{\bar{a}^2} + o\left(\frac{\Delta a}{\bar{a}}\right)^3 \qquad \forall\, j \tag{3.28}$$

$$\langle \ln^2 a_j \rangle = \ln^2 \bar{a} + \left(\frac{1}{\bar{a}^2} - \frac{\ln \bar{a}}{\bar{a}^2}\right)\Delta a^2 + o\left(\frac{\Delta a}{\bar{a}}\right)^3 \qquad \forall\, j. \tag{3.29}$$

Notice here that we earn an order of approximation thanks to the properties of the Gaussian distribution, for which the third central moment is zero (since $\mathbb{E}[(a_j - \bar{a})^3] = 0$).

We can now compute the average for y_i (we work here on the logarithms of z_i):

$$\mu_i = \mathbb{E}[y_i] = \sum_{j=0}^{i-1} \mathbb{E}[\ln(a_j)] = i\left(\ln \bar{a} - \frac{\Delta a^2}{2\bar{a}^2}\right) + io\left(\frac{\Delta a}{\bar{a}}\right)^3. \tag{3.30}$$

Let us now perform a little bit more tricky calculation, the variance of y_i:

$$\beta_i^2 = \mathbb{E}[(y_i - \mu_i)^2] = \mathbb{E}[y_i^2] - \mu_i^2 = \sum_{j=0}^{i-1} \mathbb{E}[\ln^2 a_j] + \sum_{j\neq k}^{i-1}\sum_{k}^{i-1} \mathbb{E}[\ln a_j]\mathbb{E}[\ln(a_k)] - \mu_i^2, \tag{3.31}$$

because a_j and a_k are independent random variables if $j \neq k$. We can then apply (3.28) to the first sum and (3.29) to the second one and then count how many terms there are in each sum.

$$\begin{aligned} \beta_i^2 = i\left[\ln^2(\bar{a}) + \left(\frac{1}{\bar{a}^2} - \frac{\ln \bar{a}}{\bar{a}^2}\right)\Delta a^2\right] + \\ +(i^2 - i)\left[\left(\ln(\bar{a}) - \frac{\Delta a^2}{2\bar{a}^2}\right)\left(\ln \bar{a} - \frac{\Delta a^2}{2\bar{a}^2}\right)\right] - \mu_i^2 + io\left(\frac{\Delta a}{\bar{a}}\right)^3 \end{aligned} \tag{3.32}$$

Then after simplification we get:

$$\beta_i^2 = i\frac{\Delta a^2}{\bar{a}^2} + io\left(\frac{\Delta a}{\bar{a}}\right)^3 \tag{3.33}$$

Let us now compute the ik-element of the covariance matrix of the y_i (being the z_i

correlated also the y_i are). By following the same method as before, we get:

$$\begin{aligned}\mathbb{E}[(y_i-\mu_i)(y_k-\mu_k)] = \mathbb{E}[y_i y_k] - \mu_i\mu_k = \sum_{j=0}^{l-1}\mathbb{E}[\ln^2 a_j] + \\ + \sum_{j\neq n}^{i-1}\sum_{n=0}^{k-1}\mathbb{E}[\ln a_j]\mathbb{E}[\ln(a_n)] - \mu_i\mu_k = \\ = l\left[\ln^2\bar{a} + \left(\frac{1}{\bar{a}^2} - \frac{\ln\bar{a}}{\bar{a}^2}\right)\Delta a^2\right] + (ik-l)\left[\left(\ln\bar{a} - \frac{\Delta a^2}{2\bar{a}^2}\right)^2\right] + \\ -\mu_i\mu_k + io\left(\frac{\Delta a}{\bar{a}}\right)^3\end{aligned} \tag{3.34}$$

where $l = \min(i,k)$. Let us say without loss of generality that $l = k$. So:

$$\lambda_l^2 = \mathbb{E}[(y_i-\mu_i)(y_l-\mu_l)] = l\frac{\Delta a^2}{\bar{a}^2} + io\left(\frac{\Delta a}{\bar{a}}\right)^3 \tag{3.35}$$

By following the same notations as in Beaulieu's theorem ((Beaulieu, 2012) and Section 3.1.1), we have:

$$\lambda_l^2 \simeq l\frac{\Delta a^2}{\bar{a}^2} \qquad \text{with} \qquad l = \min(i,k) \tag{3.36}$$

and for the correlation coefficient:

$$\rho_{ik} = \frac{\lambda_k^2}{\beta_i\beta_k} \simeq \sqrt{\frac{k}{i}}. \tag{3.37}$$

We consider now the partial sum of the z_i:

$$S_m = \frac{1}{N}\sum_{j=m+1}^{N} z_j, \tag{3.38}$$

therefore Z averaged will be:

$$\overline{Z} = \frac{1}{N}\sum_{j=1}^{N} z_j \Rightarrow \overline{Z} = S_m + \frac{1}{N}\sum_{n=1}^{m} z_n \tag{3.39}$$

the idea is to be able to say that $S_m \sim \text{Log-normal}(\hat{\mu}_m, \hat{\sigma}_m^2)$, because of Beaulieu's theorem (Beaulieu, 2012). Indeed, S_m is log-normal if we can write:

$$y_j = \lambda_m W_m + X_j \qquad \forall j \in [m+1, N] \tag{3.40}$$

where λ_m is as defined above (here m is always the smaller of the two indices) and W_m and $\{X_j\}_{m+1}^N$ are independent Gaussian random variables with $\mathbb{E}[X_j] = \mu_j$ and we require also, in order to get rid of mixed terms of the square of y_j, $\mathbb{E}[W_m] = 0$. This is actually an essential condition for the proof.

Let us check then that everything is correct by computing the average and the variance of y_j.

We found that:

$$\mathbb{E}[y_j] = \mu_j = \lambda_m \mathbb{E}[W_m] + \mathbb{E}[X_j] = \mathbb{E}[X_j] \simeq -jb^2 \qquad \text{with} \qquad b^2 = -\ln \bar{a} + \frac{\Delta a^2}{\bar{a}^2} \tag{3.41}$$

since $\mathbb{E}[W_m] = 0$. We notice that μ_j has to be negative because otherwise $m_j = \mathbb{E}[z_j]$ would not be bounded and Z would diverge. Indeed it is known that $m_j = \exp\{\mu_j + \beta_j^2/2\}$, that plugging in the approximations (3.30) and (3.33) becomes:

$$m_j \simeq e^{j \ln \bar{a}} = \bar{a}^j \tag{3.42}$$

For the variance we have:

$$\mathrm{Var}(y_j) = \lambda_m^2 \mathrm{Var}(W_m) + \mathbb{E}[X_j^2] - \mu_j^2 \simeq \frac{\Delta a^2}{\bar{a}^2} + \mathrm{Var}(X_j) \tag{3.43}$$

Where we imposed

$$\mathrm{Var}(W_m) = \frac{1}{m} \tag{3.44}$$

This condition is fundamental to be coherent with the definition of y_j. As a matter of fact, if we compute $\mathrm{Var}(y_j)$ we should get what we found in (3.33), so the variance of y_j cannot depend on m.

We have then a condition on $\mathrm{Var}(X_j)$, which satisfies (3.33):

$$\mathrm{Var}(X_j) = j\frac{\Delta a^2}{\bar{a}^2} - \frac{\Delta a^2}{\bar{a}^2} \tag{3.45}$$

Let us now check the condition for $s_{z_j}^2$ to be bounded.

$$\begin{aligned} s_{z_j}^2 = \left(e^{\beta_j^2} - 1\right) e^{2\mu_j + \beta_j^2} \simeq \left(e^{j\frac{\Delta a^2}{\bar{a}^2}} - 1\right) e^{2j\left(\ln \bar{a} - \frac{\Delta a^2}{2\bar{a}^2}\right) + j\frac{\Delta a^2}{\bar{a}^2}} = \\ = \left(e^{j\frac{\Delta a^2}{\bar{a}^2}} - 1\right) e^{2j \ln \bar{a}} = e^{j\left(\frac{\Delta a^2}{\bar{a}^2} + 2\ln \bar{a}\right)} - e^{2j \ln \bar{a}} \end{aligned} \tag{3.46}$$

The last term tends to zero because of the condition on the mean we found before. To have a bounded $s_{z_j}^2$ the following condition has to be fulfilled:

$$j\left(\frac{\Delta a^2}{\bar{a}^2} + 2\ln \bar{a}\right) < 0 \tag{3.47}$$

This is a transcendental equation that can be solved numerically and that defines the interval in the parameter Δa^2 and $\bar{a}$ space in which we get a log-normal distribution for Z. For example in the simulations we chose $\Delta a = 0.2$ that gives from this condition a critical point of $\bar{a} = 0.98$ over which the distribution for Z would not be log-normal anymore. Finally we notice that as far as Δa approaches to zero the critical value for $\bar{a}$ tends to 1, which is the critical point for divergence in standard branching processes. Curiously the variance of y_j diverges with j while the variance of z_j goes to zero.

Since both m_j and $s_{z_j}^2$ are bounded we verify the hypothesis of Beaulieu's theorem for

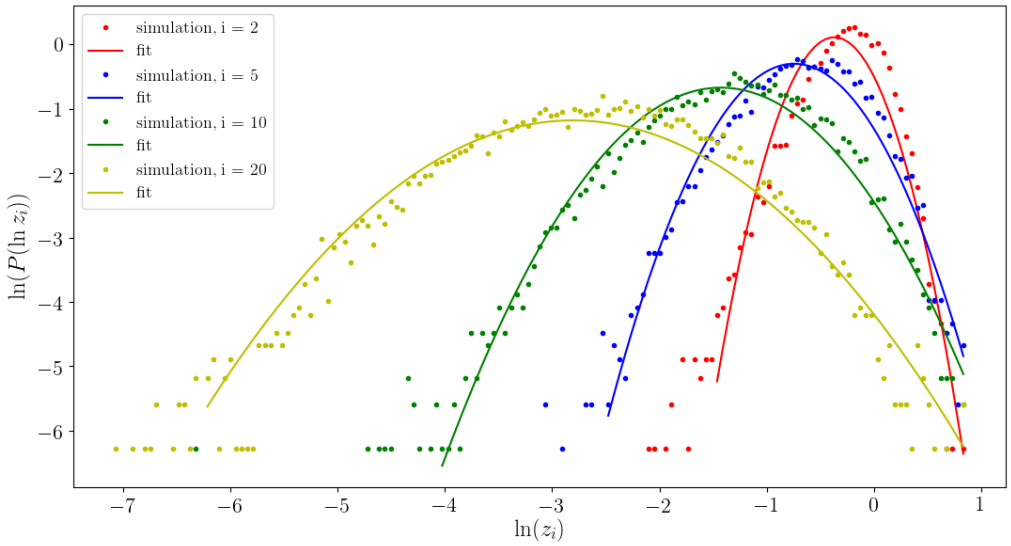

Figure 3.4 – Distribution of $\ln(z_i)$ in a log-log plot at the times $i = 2, 5, 10, 20$ and corresponding parabolic fit. Here the distribution parameters are $\bar{a} = 0.9$ and $\Delta a^2 = 0.04$ and the number of repetitions for the statistics is 10000.

S_m. Then once we have shown that S_m is log-normal we can write:

$$\overline{Z} = \lim_{m \text{ small}} S_m \tag{3.48}$$

where m small enough for z_j to be log-normal. If we can go to this limit we get that the sum Z and $\overline{Z}$ are approximately log-normal. Indeed the distribution for $\overline{Z}$ is the same as the distribution of Z since they differ only by a constant factor.

We notice also that m of the order of few unities is fine enough (and was the case in simulations, where already at $m = 4 - 5$ the shape of the z_m was very close to a log-normal) (see Fig. 3.4), since the beginning of the process is where the variance s_{z_i} is very small (for $i = 0$ it is zero) and then for small i the variables z_i have a very small contribution to the global variance, *i.e.* to the shape, of the final distribution for Z.

3.3.1 Comments

This was our first attempt trying to generalize Beaulieu's theorem. If at first glance the proof seems satisfying, afterwards we realized an incoherence. In fact if we take condition (3.44) in order to have the correct variance we cannot have the correct correlation at the same time, because we have a symmetry in the indices j and m that we have in some way to break to be able to generalized correctly the theorem. We thought first that the free parameters μ_j, $\mathrm{Var}(X_j)$, $\mathrm{Var}(W_m)$ and $\mathbb{E}[W_m]$ were enough to overcome the problem but we can in reality prove that it is impossible to find a set of parameters for which everything is coherent (good variance and good correlations for the y_i). Actually, for some of them we do not have choice, as for $\mathbb{E}[W_m] = 0$. We show this with short reasoning. Let us start by showing that $\mathbb{E}[W_m] = 0$ is forced. To be consistent for the value of $\mathbb{E}[y_j]$ from Eq. (3.41) we need to impose $\mathbb{E}[W_m] \sim 1/\sqrt{m}$, in order to eliminate the dependence

on m. If we write then the full formula (without any assumption on the values of the parameters) of the covariance we have:

$$\begin{aligned}\mathrm{Cov}\,(y_j, y_k) = \lambda_m^2\mathbb{E}[W_m^2] + \lambda_m^2\mathbb{E}[W_m]\mathbb{E}[X_k] + \lambda_m\mathbb{E}[W_m]\mathbb{E}[X_j] + \mathbb{E}[X_j]\mathbb{E}[X_k] + \\ -\lambda_m\mathbb{E}[W_m]\mu_k - \lambda_m\mathbb{E}[W_m]\mu_j + \mu_j\mu_k - \mu_j\mathbb{E}[X_k] - \mu_k\mathbb{E}[X_j] \equiv \min(j,k)\frac{\Delta a^2}{\bar{a}^2}.\end{aligned} \tag{3.49}$$

By looking at this formula wa can deduce that the cross product $\lambda_m^2\mathbb{E}[W_m]\mathbb{E}[X_k]$ will depend on both index and this is not allowed. We are therefore forced to set $\mathbb{E}[W_m] = 0$. At this point by Eq. (3.43) we need to fix as well $\mathbb{E}[W_m^2]$, which has to go like $\sim 1/m$. Now coming back to Equation (3.49) we can see that we cannot get rid of the largest of the 2 indexes, since the expression is now symmetric in j, k, being the only non-symmetric term $\lambda_m^2\mathbb{E}[W_m]\mathbb{E}[X_k]$. This actually proves that the theorem is surely not generalisable if λ depends only on one of the 2 indexes.

We tried to do the same procedure taking instead of λ_m as a scalar, a matrix, namely the covariance matrix, as it is defined in (3.35). But even in this way we were not able to select only the smaller of the 2 indices, *i.e.* to break the symmetry between the two indices. Here we write the covariance matrix of the process for clarity:

$$\begin{bmatrix} \lambda_1^2 & & \dots & & \lambda_1^2 \\ & \ddots & \dots & & \vdots \\ \vdots & \vdots & \lambda_i^2 & \dots & \lambda_i^2 \\ & & \vdots & \ddots & \vdots \\ \lambda_1^2 & \dots & \lambda_i^2 & \dots & \lambda_N^2 \end{bmatrix}$$

In the case of Beaulieu's theorem this matrix is a constant matrix where all the elements are λ^2. This apparently small difference is enough to make the use of the theorem in this case pointless and not able to explain the convergence toward a distribution close to a log-normal for Z, as shown in the panel **A** of Fig. 3.5. To remove any doubts we also try to not impose any condition on the free parameters and see what should look like the variable $\lambda_m^2 W_m$ and we found that it necessarily depends on j (see Fig. 3.5**A**). We can see that, while the mean gives the correct value of 0, the variance is not constant with respect to j, as it should be for the theorem to work. This behaviour is not dependent on the choice of the parameters and of the distribution of a.

3.4 Statistical characterisation of the Z distribution

In this section we give some results about the distribution of Z. We work out, for example, the mean and the variance of the Z distribution for any choice of parameters $\bar{a}$ and $\tilde{a}$ and regardless of the distribution of a. These results are general and are found by directly working on the z_i and their definition.

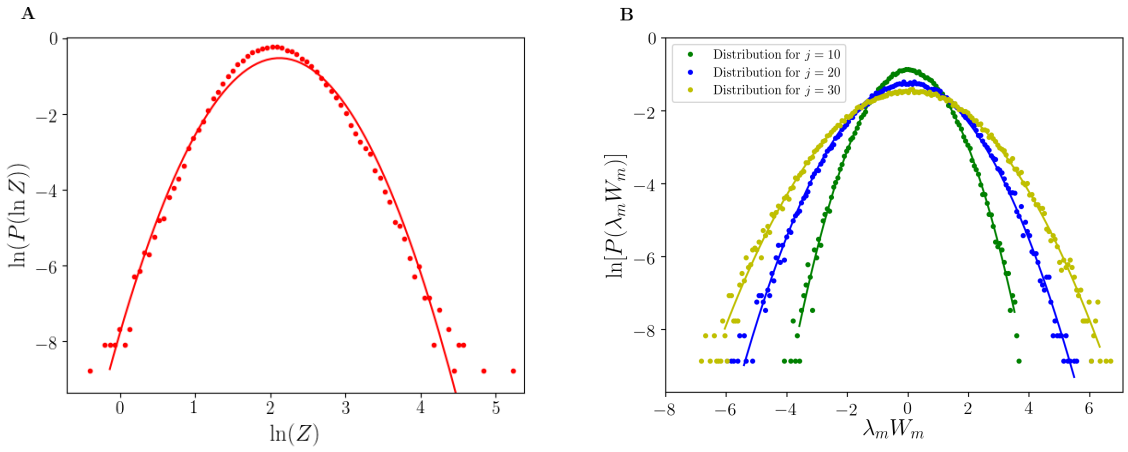

Figure 3.5 – A. Logarithm of the probability density function of $\ln(Z)$ for the same parameters as in Fig 3.4 (*i.e.* $a \sim \mathcal{N}(0.9, 0.04)$) and corresponding parabolic fit. The distribution is well approximated by a parabola, meaning that the distribution of Z is close to a log-normal. **B.** Logarithm of the distribution of $\lambda_m^2 W_m$ for different values of j and corresponding parabolic fit. The variable $\lambda_m^2 W_m$ is defined by Eq. (3.40) and it is computed by generating random numbers (the X_j) with normal distributions having the parameters defined in the text, which depend on j. We clearly see that the distributions are Gaussian (here a parabola stands for a Gaussian distribution) for all j, but their variance increases with j. For both plot the number of repetitions for the statistics is 100000 ad the time length $N = 500$, which was largely enough for convergence.

Let us write down explicitly the sum Z using Equation (3.16):

$$Z = a_1 + a_1 a_2 + \cdots + a_1 \dots a_N. \tag{3.50}$$

Directly from this equation we can already compute the mean and the variance of Z exactly. So:

$$m_Z = \mathbb{E}[Z] = \sum_{i=1}^{N} \left(\mathbb{E}\left[a\right]\right)^i = \frac{\bar{a}}{1-\bar{a}} \tag{3.51}$$

since $\mathbb{E}[.]$ is linear no matter if the variables are independent or not and $\mathbb{E}[a_i a_j] = \mathbb{E}[a_i]\mathbb{E}[a_j]$, since variables with different indices are independent. This formula is valid if $0 \leq \bar{a} < 1$ since we took the limit $N \to \infty$.

We can apply the same principles to compute $\mathbb{E}[Z^2]$, let us call $\tilde{a} = \mathbb{E}[a^2]$:

$$\begin{aligned} \mathbb{E}[Z^2] = \sum_{j=1}^{N} \tilde{a}^j + 2 \sum_{j=1}^{N-1} \sum_{k=1}^{N-j} \tilde{a}^j \left(\mathbb{E}[a]\right)^k = \frac{\tilde{a}}{1-\tilde{a}} + 2 \sum_{j=1}^{N-1} \tilde{a}^j \left(\frac{\bar{a} - \bar{a}^{N-j+1}}{1-\bar{a}}\right) = \\ = \frac{\tilde{a}}{1-\tilde{a}} \left(1 + \frac{2\bar{a}}{1-\bar{a}}\right) \end{aligned} \tag{3.52}$$

For the limit $N \to \infty$ to converge we need as before that $0 \leq \bar{a} < 1$ but also that $\tilde{a} < 1$. Then for the variance of Z we get:

$$s_Z^2 = \mathrm{Var}(Z) = \frac{\tilde{a}}{1-\tilde{a}} \left(1 + \frac{2\bar{a}}{1-\bar{a}}\right) - \frac{\bar{a}^2}{(1-\bar{a})^2}. \tag{3.53}$$

We checked these formulae in the numerical simulations and we should consider them as

exact, since, in simulations with $\tilde{a} < 1$ and $< \bar{a} < 1$, we always take care that the time N is large enough to have all the z_i that have already converged to zero. We can notice that the condition for a limited variance ($\tilde{a} < 1$) is coherent with the condition we found in (3.46) starting from the variables y_j. In fact the limit of the variances of z_j as $j \to \infty$ has to be zero, for the variance of the sum s_Z^2 to be bounded. Solving numerically the condition for Equation (3.46) to be bounded gives something very close to the condition $\tilde{a} < 1$, meaning that everything is consistent.

In the case where $\bar{a} > 1$ (and then for sure also $\tilde{a} > 1$), for a given N this is what we obtain for the average and the variance of Z:

$$\mathbb{E}[Z] = \frac{\bar{a} - \bar{a}^{N+1}}{1 - \bar{a}} \tag{3.54}$$

$$\mathrm{Var}(Z) = 2\frac{\bar{a}}{1-\bar{a}}\left[\frac{\tilde{a} - \tilde{a}^N}{1-\tilde{a}} - \frac{\bar{a}^{N+1}}{\bar{a}-\tilde{a}}\left(\frac{\tilde{a}}{\bar{a}} - \frac{\tilde{a}^N}{\bar{a}^N}\right)\right] + \frac{\tilde{a} - \tilde{a}^{N+1}}{1-\tilde{a}} - \left(\frac{\bar{a} - \bar{a}^{N+1}}{1-\bar{a}}\right)^2 \tag{3.55}$$

The previous formulae are true of course also in the convergent case, and are useful to check if the simulations converged numerically, by checking m_Z and s_Z^2 in function of the time step N.

There are 3 other cases that is worth to mention here. The first is when $\bar{a} = 1$, and then necessarily $\tilde{a} > 1$, since $\Delta a^2 > 0$. In this case we get:

$$\mathbb{E}[Z] = N. \tag{3.56}$$

For the computation of the variance we have before to compute the series:

$$s_N = \sum_{j=1}^{N} j\tilde{a}^j, \tag{3.57}$$

this can be solved exactly by noting that:

$$s_N - \tilde{a}s_N = \sum_{j=1}^{N} \tilde{a}^j - N\tilde{a}^{N+1}, \tag{3.58}$$

which, if we solve for s_N, gives:

$$s_N = \frac{\tilde{a} - \tilde{a}^{N+1}\left(1 + N(1-\tilde{a})\right)}{(1-\tilde{a})^2}. \tag{3.59}$$

Now we are able to compute $\mathbb{E}[Z^2]$:

$$\begin{aligned} \mathbb{E}[Z^2] &= \frac{\tilde{a} - \tilde{a}^{N+1}}{1-\tilde{a}} + 2\sum_{j=1}^{N-1} \tilde{a}^j (N-j) = \\ &= \frac{\tilde{a}^{N+2} + \tilde{a}^{N+1} - \tilde{a}^2(1+2N) + \tilde{a}(2N-1)}{(1-\tilde{a})^2}, \end{aligned} \tag{3.60}$$

and therefore the variance:

$$\mathrm{Var}(Z) = \frac{1}{(1-\tilde{a})^2}(\tilde{a}^{N+2} + \tilde{a}^{N+1} - \tilde{a}^2\,(1+N)^2 + \tilde{a}\left(2N^2+2N-1\right) - N^2). \tag{3.61}$$

The second case is when $\tilde{a} = 1$ and then for sure $\bar{a} < 1$. In this case $\mathbb{E}[Z]$ is the same as in Equation (3.51), but the variance becomes:

$$\mathrm{Var}(Z) = N + 2\,(N-1)\,\frac{\bar{a}}{1-\bar{a}} - \frac{2\bar{a}^2}{1-\bar{a}}\left(\frac{\bar{a}^{N-1}-1}{\bar{a}-1}\right) - \left(\frac{\bar{a}-\bar{a}^{N+1}}{1-\bar{a}}\right)^2, \tag{3.62}$$

which for $N \to \infty$ becomes:

$$\mathrm{Var}(Z) = N + 2\,(N-1)\,\frac{\bar{a}}{1-\bar{a}} - \frac{2\bar{a}^2}{(1-\bar{a})^2} - \left(\frac{\bar{a}}{1-\bar{a}}\right)^2. \tag{3.63}$$

The last case is for $\bar{a} < 1$ and $\tilde{a} > 1$. In this case the average $\mathbb{E}[Z]$ is again given by Formula (3.51). The variance is:

$$\mathrm{Var}(Z) = \frac{\tilde{a}-\tilde{a}^{N+1}}{1-\tilde{a}} + 2\frac{\bar{a}}{1-\bar{a}}\frac{\tilde{a}-\tilde{a}^N}{1-\tilde{a}} - \frac{2\bar{a}^2}{1-\bar{a}}\left(\frac{\tilde{a}\bar{a}^{N-1}-\tilde{a}^N}{\bar{a}-\tilde{a}}\right) - \left(\frac{\bar{a}-\bar{a}^{N+1}}{1-\bar{a}}\right)^2, \tag{3.64}$$

that for $N \to \infty$ becomes:

$$\mathrm{Var}(Z) = \frac{\tilde{a}-\tilde{a}^{N+1}}{1-\tilde{a}} + 2\frac{\bar{a}}{1-\bar{a}}\frac{\tilde{a}-\tilde{a}^N}{1-\tilde{a}} + \frac{2\bar{a}^2}{1-\bar{a}}\left(\frac{\tilde{a}^N}{\bar{a}-\tilde{a}}\right) - \left(\frac{\bar{a}}{1-\bar{a}}\right)^2. \tag{3.65}$$

Therefore in order to resume all the possible behaviours with respect to $N \to \infty$, there are 5 possibilities:

1 − both $\mathbb{E}[Z]$ and $\mathrm{Var}(Z)$ converge, for $\bar{a} < 1$ and $\tilde{a} < 1$;

2 − $\mathbb{E}[Z]$ converges and $\mathrm{Var}(Z)$ diverges linearly, for $\tilde{a} = 1$;

3 − $\mathbb{E}[Z]$ converges and $\mathrm{Var}(Z)$ diverges exponentially, for $\tilde{a} > 1$ and $\bar{a} < 1$;

4 − $\mathbb{E}[Z]$ diverges linearly with N and $\mathrm{Var}(Z)$ diverges exponentially, for $\bar{a} = 1$;

5 − both $\mathbb{E}[Z]$ and $\mathrm{Var}(Z)$ diverge exponentially, for $\bar{a} > 1$;

We could expect that in general the distribution of the sum Z, and in particular its asymptotic properties, like a power-law tail, will depend on these particular behaviours.

Another thing we can compute directly on the variables z_j rather than on its logarithm, is the covariance, $\mathrm{Cov}\,(z_i, z_k) = \mathbb{E}[(z_i - m_i)(z_k - m_k)]$. To do so we have to group by

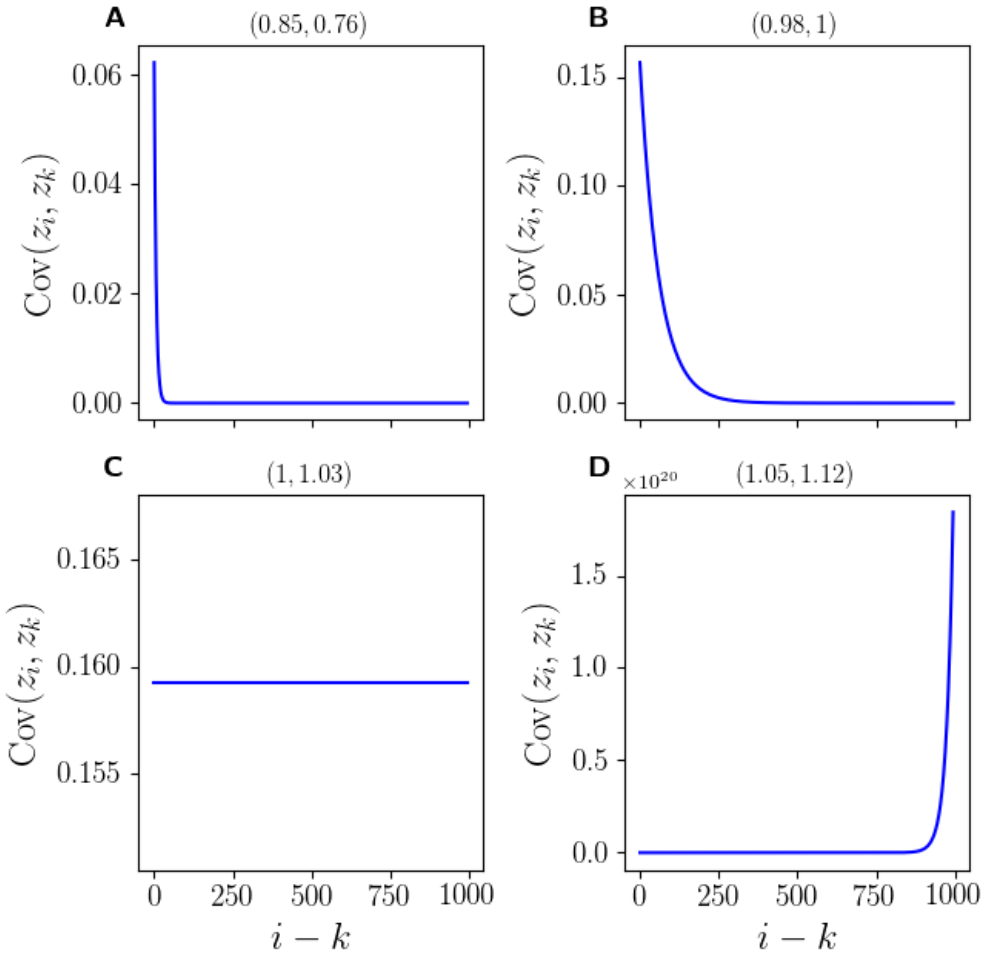

Figure 3.6 – Behaviour of the covariance function with respect to the time distance $i-k$ for different values of the parameters $\bar{a}$, $\tilde{a}$. The parameters coordinates $(\bar{a}, \tilde{a})$ chosen are given above each plot. The particular value of $k = 5$ is chosen for all the plots.

independent variables, a_i, again noticing that if the indices are different the variables are independent and then count how many terms we have. At the end the covariance is:

$$\mathrm{Cov}\left(z_i, z_k\right) = \mathbb{E}\left[a^2\right]^k \bar{a}^{i-k} - \bar{a}^{k+i} \qquad \text{if} \qquad k < i \tag{3.66}$$

so now is depending on both i and k, in a way that depends on the values of $\bar{a}$ and $\tilde{a}$. In Fig. 3.6 we show all possible behaviours of the covariance with respect to the values of the parameters. As far as $\bar{a}$ is smaller than 1 the covariance decreases with time (panels **A** and **B**). When $\bar{a} = 1$ the covariance is constant for all times and the constant will depend on the particular value of k chosen, here $k = 5$ (see panel **C**). When $\bar{a} > 1$ the covariance shows, instead, an increase with time (see panel **D**). What really makes the type of covariance curve behaviour is the value of $\bar{a}$. The value of $\tilde{a}$ only changes the particular increase/decrease rate or the value of the constant for $\bar{a} = 1$. We can easily see also that the covariance is always positive, since $\mathbb{E}\left[a^2\right]^k > \bar{a}^{2k}$, since $\mathrm{Var}(a)$ is definite positive.

We stress the fact that all the relations that we found in this section do not depend on the particular distribution of a.

3.5 A recurrence point of view

In this section we are going to set the problem from an other point of view exploiting the recurrence hidden already in the definition of the model itself. We also work out a complete analytical solution for the specific case where the lower bound of the uniform distribution is 0. Finally we show some typical distributions resulting from the Finite Scale region (see Fig. 3.2).

We are now going to focus only on the convergent case (case 1), that is actually the most relevant for the modelling, at least for our original motivation about ruptures avalanches in cells cytoskeleton, since in experiments avalanches do not diverge. The divergent case will be studied in a further section.

First, we can notice that supposing Z log-normally distributed we can relate the first two moments of Z previously found (Eqs. (3.51), (3.53)), to the parameters of the log-normal. Starting from the expressions of the mean $e^{\mu+\sigma^2/2}$ and the variance $\left(e^{\sigma^2}-1\right)e^{2\mu+\sigma^2}$ of a log-normal random variable, we can solve a system of equations for μ_Z and σ_Z^2 giving:

$$\mu_Z = \ln\left(\mathbb{E}[Z]\right) - \frac{1}{2}\ln\left(\frac{\mathrm{Var(Z)}}{\mathbb{E}^2[Z]}+1\right) \tag{3.67}$$

and

$$\sigma_Z^2 = \ln\left(\frac{\mathrm{Var(Z)}}{\mathbb{E}^2[Z]}+1\right) \tag{3.68}$$

This system has a finite solution meaning that the hypothesis of log-normality for Z is consistent. Of course this does not prove that Z is log-normal, but it says that it can be, or at least it can be something close to a log-normal. Moreover, it allows a comparison between the log-normal approximation and numerical estimates of μ_Z and σ_Z^2: if the values are very different the log-normal approximation is bad.

Let us now write again Z in an other way, putting in evidence the recursion, always considering $N \to \infty$:

$$Z = a_1(1+\underbrace{a_2 + a_2a_3 + \cdots + a_2\ldots a_N}_{Z_1}) = a_1\left(1+Z_1\right). \tag{3.69}$$

We notice that on the right, within parentheses, we have 1+ a variable Z_1 that has to be distributed like Z. In this way, we write Z as the product of 2 independent random variables, since a_1 is not contained in Z_1. We could also continue the recursion and write $Z_1 = a_2(1+Z_2)$, this may be useful but already with this relation we can get something interesting.

We can now write the distribution of Z, $g(Z)$, as the Mellin convolution between the distributions of the 2 independent variables a_1 (here distributed with a probability density function as $f(a)$) and $1+Z_1$, noticing that the latter one has the same distribution as Z shifted by 1. So:

$$p(Z) = \int_1^{+\infty} p(y-1) f\left(\frac{Z}{y}\right)\frac{dy}{y}, \tag{3.70}$$

where we used the information that Z is a positive random variable, so $g(Z)$ can be written as $g(Z) = \theta(Z)p(Z)$, making use of the Heaviside step function $\theta(Z)$. This is a Fredholm Integral Equation (FIE) of the second type, that, if solved for the function p, would give us all the information about the process. It can be proved by Fredholm's theorem on Fredholm integral operators (for examples see theorem 3.1 in (Hasan et al., 2019)) that a solution exists and is unique, if f is defined on a compact interval and is bounded. However, finding a general solution was impossible with actual mathematical and numerical techniques.

If we now remember that we are not restricted with respect to the distribution of a, $f(a)$, we can simplify the computations taking a well defined (*i.e.* with $\bar{a}$ and $\tilde{a} < 1$) uniform distribution. In order to convince you that the result is still approximately log-normal we can observe Fig. 3.7 (comparing the $\ln(Z)$ distribution of **C** with **D** and **E** with **F**). We can also wonder, again, how fast the convergence of z_i to log-normal variables is. From Figure 3.8 it seems that the convergence is at least as fast as when the distribution of a is Gaussian. The conjecture about the shape distribution of a not being important seems then verified. But still the parameters of this distribution can play a role. Before moving to the analytic understanding, let us describe more qualitatively the results of this region. Observing Fig. 3.7, where we plot the distributions of the logarithm of partial sums $Z_p = \sum_{i=1}^{N^*} z_i$ together with the final distribution of $\ln(Z)$, we can first say that in most of the region the final distribution is quite well approximated by a log-normal distribution, either for a following a uniform distribution or a Gaussian. In fact plots **C**, **D**, **E** and **F** show a good agreement with the parabolic fit, which, we remember, in a log-log plot corresponds to a log-normal distribution. These are examples, but we did simulations systematically and Fig. 3.7 gives a summary typical results. We notice also that the convergence to the final distribution is quite fast and therefore the approximately log-normal distribution is given by the sum of a very limited number of variables (in the examples $N^* = 25$ or $N^* = 39$). We can also see that apart from the first $N^* = 5$ distributions, all the others, even though partial, seem already well parabolic. This suggests that varying the parameters around the chosen values will shift and change the width of the distribution, changing the maximum N^*, but will not change its shape. We notice still that in plots **E** and **F**, *i.e.* for a bigger maximum N^*, we start seeing a tail on the right, for larger values of Z. The top line of Fig. 3.7, on the other hand, shows an other different possible outcome. In this case however the width of the a distribution is comparable with its mean, $(\bar{a}, \tilde{a}) = (0.55, 0.34)$ and therefore we do not respect anymore the narrow condition (3.27) and we explicitly see the effect of the fact that zero is an attractor of the product operation, since all (for **A** and most for **B**) of the a_j are lower than 1. The sum is therefore stopped very fast, even before the z_i distributions can converge toward log-normals (see 3.8). For this reason the distribution is here very dependent on the particular distribution of a and a direct computation of the distribution by convolution could be done, at least numerically. In the examples we see a shoulder for values of Z lower than 1 in **A** (uniform case) and a long tail for very small values of Z in **B** because the Gaussian distribution allows for values of a going to zero, giving results similar to Fig. 3.10 (discussed later). We can conclude that what matters the most, in this region, for the final shape of the distribution, is the value of the

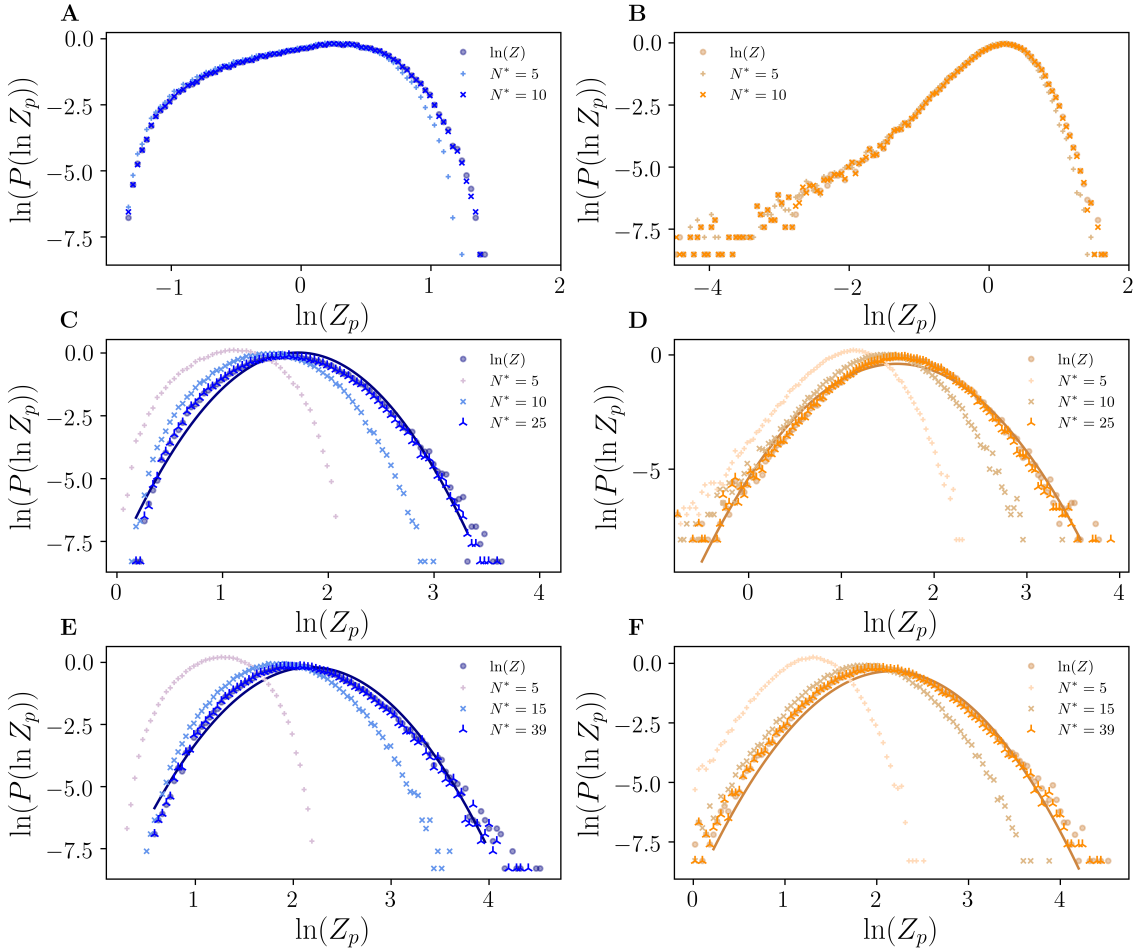

Figure 3.7 – Distribution of the logarithm of the partial sum $Z_p = \sum_{i=1}^{N^*} z_i$ for different points in the parameters plane and different N^* (see legends), until convergence toward the final distribution of $\ln(Z)$. For all the plots the number of realization for the statistics is 10^5 and the time length for the final distribution $N = 500$. The blue shaded colour map (on the left) corresponds to $a_j \sim \mathcal{U}$, while the orange shaded colour map (on the right) corresponds to $a_j \sim \mathcal{N}$. For all the plots we draw the logarithm of the resulting distribution. **A.** and **B.** Coordinates in the parameters $(\bar{a}, \tilde{a})$ plane are $(0.55, 0.34)$ for $a_j \sim \mathcal{U}_{[0.2;0.9]}$ (**A**) and $a \sim \mathcal{N}(0.55, 0.037)$ (**B**). Distribution of $\ln(Z_p)$ for $N^* = 5$ (plus signs), $N^* = 10$ (multiplication signs) and final distribution of $\ln(Z)$ (rounds). Notice the fast convergence to the final distribution, already for $N^* = 10$ the distributions overlap. **C.** and **D.** Coordinates in the parameters $(\bar{a}, \tilde{a})$ plane are $(0.85, 0.76)$ for $a_j \sim \mathcal{U}_{[0.5;1.2]}$ (**C**) and $a \sim \mathcal{N}(0.85, 0.04)$ (**D**). Distribution of $\ln(Z_p)$ for $N^* = 5$ (plus signs), $N^* = 10$ (multiplication signs), $N^* = 25$ (triangle signs), final distribution of $\ln(Z)$ (rounds) and parabolic fit of the final distribution (solid line). Remeber that a parabolic fit in a log-log plot corresponds to a log-normal distribution. **E.** and **F.** Coordinates in the parameters $(\bar{a}, \tilde{a})$ plane are $(0.9, 0.85)$ for $a_j \sim \mathcal{U}_{[0.57;1.23]}$ (**E.**) and $a \sim \mathcal{N}(0.9, 0.04)$ (**F.**). Distribution of $\ln(Z_p)$ for $N^* = 5$ (plus signs), $N^* = 10$ (multiplication signs), $N^* = 39$ (triangle signs), final distribution of $\ln(Z)$ (rounds) and parabolic fit of the final distribution (solid line).

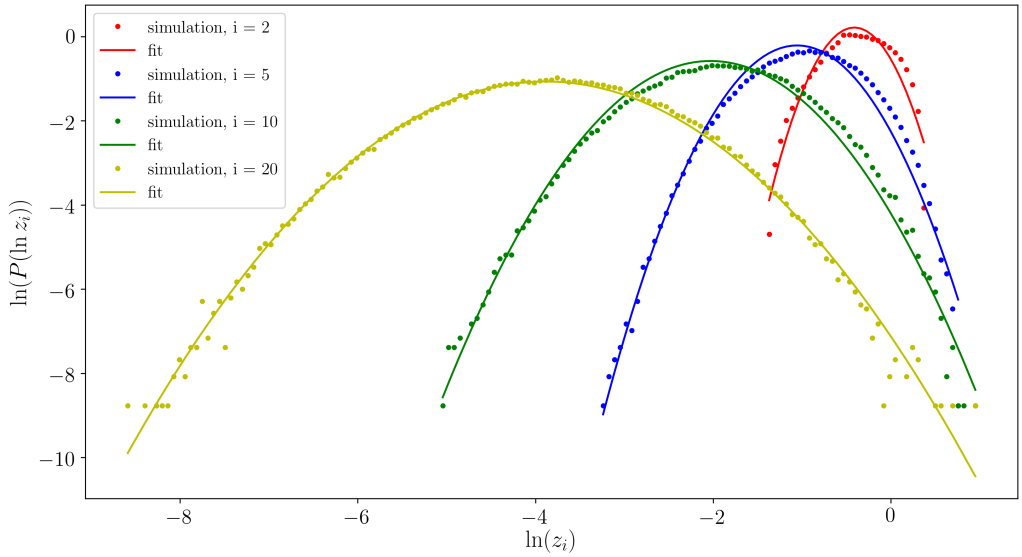

Figure 3.8 – Distribution of $\ln(z_i)$ in a log-log plot at the times $n = 2, 5, 10, 20$ and corresponding parabolic fit. Here the distribution for a a uniform distribution between $b = 0.5$ and $c = 1.2$.

maximum N^*, which in turn depends on the model parameters.

As a final remark, the details of the distributions in this region depend also on the value of the number of repetitions n (in Figs. 3.7 and 3.8 it was 10^5), which play a role in the fine definition of the distribution and in the shape of the tails. This number, if in simulations is tunable, in experiments is often limited by experimental conditions and with respect to the available statistics, a given experiment can exhibit a distribution with a more or less precise log-normal shape. For instance with a statistics as in our experiments on cells (see Section 2.4 (around 10000 of events), distributions like the ones in Fig. 3.7 would look definitely log-normals and the small defects would be cancelled out by the statistical noise.

Let us now continue our analytical understanding and write down Equation (3.70) for a uniformly distributed between b and c:

$$p(Z) = \int_1^{+\infty} p(y-1) \frac{\theta\left(\frac{Z}{y} - b\right) - \theta\left(\frac{Z}{y} - c\right)}{c-b} \frac{dy}{y}, \tag{3.71}$$

where $\theta(x)$ is the Heaviside step function, used to represent the uniform distribution: $\mathcal{U}_{[b,c]} = [\theta(x-b) - \theta(x-c)]/(c-b)$. This equation leads to:

$$p(Z) = \int_{Z/c}^{Z/b} \frac{p(y-1)}{c-b} \frac{dy}{y} \qquad \text{for } Z \geq c \qquad , \tag{3.72}$$

and:

$$p(Z) = \int_1^{Z/b} \frac{p(y-1)}{c-b} \frac{dy}{y} \qquad \text{for } Z < c. \tag{3.73}$$

We can then differentiate (3.72) with respect to Z in order to get a differential equation

for p.

$$p'(Z) = \frac{p\left(\frac{Z}{b}-1\right)}{c-b}\frac{1}{Z} - \frac{p\left(\frac{Z}{c}-1\right)}{c-b}\frac{1}{Z} \qquad \text{for } Z \geq c, \tag{3.74}$$

while for $Z < c$ the only difference is that the second term of the RHS vanishes. Since this equation is at the origin of many of the results of this chapter, it is worth to stress again the fact that this equation is exact for $a \sim \mathcal{U}$, but we can see that for all distributions called *admissible* in Section 3.3 we expect similar results. That is because, even though we do not directly use the approximation as in Section 3.3, if we respect condition (3.27) the distribution is close to have a compact support and Eq. (3.74) will be similar. Equation (3.74) is a functional (in the sense of non local, called also delayed) differential equation is complicated to solve either analytically or numerically. We tried to solve it with Wolfram Mathematica v11.3 and piecewise analytically or by power series, without success. It may be possible in some way to prove that for some b and c, p can be approximated by a log-normal, but it would not add much to the simulation result.

What we can say for sure is that we can solve Eq. (3.74) analytically when $b = 0$ and $c = 1$. In this case we notice that the first term in the RHS of Equation (3.74) goes to zero, since $\lim_{Z\to\infty} p(Z) = 0$, being p a probability density function. Then Eq. (3.74) becomes:

$$p'(Z)\theta(Z) + p(Z)\delta(Z) = -p(Z-1)\theta(Z-1)\frac{1}{Z}. \tag{3.75}$$

Here we injected in the equation the global definition for $g(Z)$ where we can see explicitly that Z is positive and then considering that the derivative of $\therefore(Z)$ is $\delta(Z)$. We can now solve Equation (3.75) per intervals of Z, starting from $Z \in (0, 1)$. Since Z has to be positive:

$$p'(Z) = 0 \Rightarrow p(Z) = k_1 \qquad \text{for} \qquad Z \in (0, 1) \tag{3.76}$$

with k_1 a constant to be imposed by the normalization and $\delta(Z)$ is the Dirac delta distribution. Let us notice that in $Z = 0$ we would have $0 = p'(0) = -p(0)\delta(0)$ pointing out that p is not derivable in zero. We can still for continuity extend the definition of p as $p(0) = k_1$. Then in the following interval $(1, 2)$:

$$p'(Z) = -\frac{k_1}{Z} \Rightarrow p(Z) = -k_1 \ln(Z) + k_2 \qquad \text{for} \qquad Z \in (1, 2) \tag{3.77}$$

By imposing continuity in $Z = 1$ we get $k_2 = k_1$. Starting from interval $(2, 3)$ the calculation becomes more complicated because of the non locality of the differential equation, but still possible and implies the polylogarithm function defined as below:

$$\mathrm{Li}_{s+1}(x) = \int_0^x \frac{\mathrm{Li}_s(t)}{t}dt \qquad \text{and} \qquad \mathrm{Li}_1(x) = -\ln(1-x)$$

So in the interval $(2, 3)$ we have:

$$\begin{aligned} p'(Z) &= k_1 \frac{\ln(Z-1)}{Z} - \frac{k_1}{Z} \\ \Rightarrow p(Z) &= -k_1 \ln(Z) - k_1 \mathrm{Li}_2(Z) + k_1(\ln(Z-1) - \ln(1-Z))\ln(Z) + k_3 \end{aligned} \tag{3.78}$$

Again from continuity arguments we could find an expression for k_3 that only depends on k_1, that is actually the only normalization constant of the probability $p(Z)$.

For the following interval the expression becomes very long, but that still exists, involving again the polylogarithm function. We write then, for the interval $(3, 4)$:

$$\begin{aligned} p(Z) = &-k_3 \ln(Z) - k_1 \mathrm{Li}_2(Z) + k_1(\ln(Z-1) - \ln(1-Z))\ln(Z) + \\ &+k_1[-\mathrm{Li}_3(2-Z) - \mathrm{Li}_3(Z) + \mathrm{Li}_3\left(\frac{Z}{2-Z}\right) - \mathrm{Li}_3\left(\frac{Z}{Z-2}\right) + \\ &+\ln\left(\frac{Z}{2-Z}\right)\left(\mathrm{Li}_2\left(\frac{Z}{Z-2}\right) - \mathrm{Li}_2\left(\frac{Z}{2-Z}\right)\right) + \mathrm{Li}_2(2-Z)\left(\ln(Z) - \ln\left(\frac{Z}{2-Z}\right)\right) + \\ &+\mathrm{Li}_2(Z-1)\ln(Z) + \mathrm{Li}_2(Z)\left(\ln(2-Z) + \ln\left(\frac{Z}{2-Z}\right)\right) + \\ &+\frac{1}{2}\left(\ln\left(-\frac{2}{Z-2}\right) + \ln(Z-1) - \ln\left(\frac{2(Z-1)}{Z-2}\right)\right)\ln^2\left(\frac{Z}{2-Z}\right) + \\ &+(\ln(1-Z) - \ln(Z-1))\ln(Z)\ln\left(\frac{Z}{2-Z}\right) + \ln(2-Z)\ln(Z-1)\ln(Z) + \\ &+\frac{1}{2}(\ln(Z-1) - \ln(1-Z))\ln(Z)(\ln(Z) - 2\ln(2-Z)) + \\ &\mathrm{Li}_3(1-Z) + \mathrm{Li}_3\left(1-\frac{Z}{2}\right) + \mathrm{Li}_3\left(\frac{Z-1}{Z-2}\right) - \mathrm{Li}_3\left(\frac{2(Z-1)}{Z-2}\right) - \\ &+\ln\left(\frac{2(Z-1)}{Z-2}\right)\left(\mathrm{Li}_2\left(\frac{Z-1}{Z-2}\right) - \mathrm{Li}_2\left(\frac{2(Z-1)}{Z-2}\right)\right) + \\ &-\mathrm{Li}_2(1-Z)\left(\ln(Z-2) + \ln\left(\frac{2(Z-1)}{Z-2}\right)\right) + \\ &-\mathrm{Li}_2\left(1-\frac{Z}{2}\right)\left(\ln(Z-1) - \ln\left(\frac{2(Z-1)}{Z-2}\right)\right) + \\ &-\frac{1}{2}\left(\ln\left(\frac{1}{4-2Z}\right) + \ln(Z) - \ln\left(\frac{Z}{2-Z}\right)\right)\ln^2\left(\frac{2(Z-1)}{Z-2}\right) + \\ &-\ln(2)\ln(1-Z)\ln\left(\frac{2(Z-1)}{Z-2}\right) + \\ &+\frac{1}{2}\ln(2)\ln(1-Z)(\ln(1-Z) - 2\ln(Z-2)) - \ln(Z-2)\ln(Z-1)\ln\left(\frac{Z}{2}\right) + \\ &+\mathrm{Li}_3(Z) - \mathrm{Li}_2(Z)\ln(-Z) + \frac{1}{2}(\ln(Z-1) - \ln(1-Z))\ln^2(-Z)] + k_4 \end{aligned} \tag{3.79}$$

As you can see the solution becomes quick very complicated, but in principle can be found for all intervals and will involve all degrees of the polylogarithm function. More pragmatically we can observe that the solution for large Z is easy to be calculated and

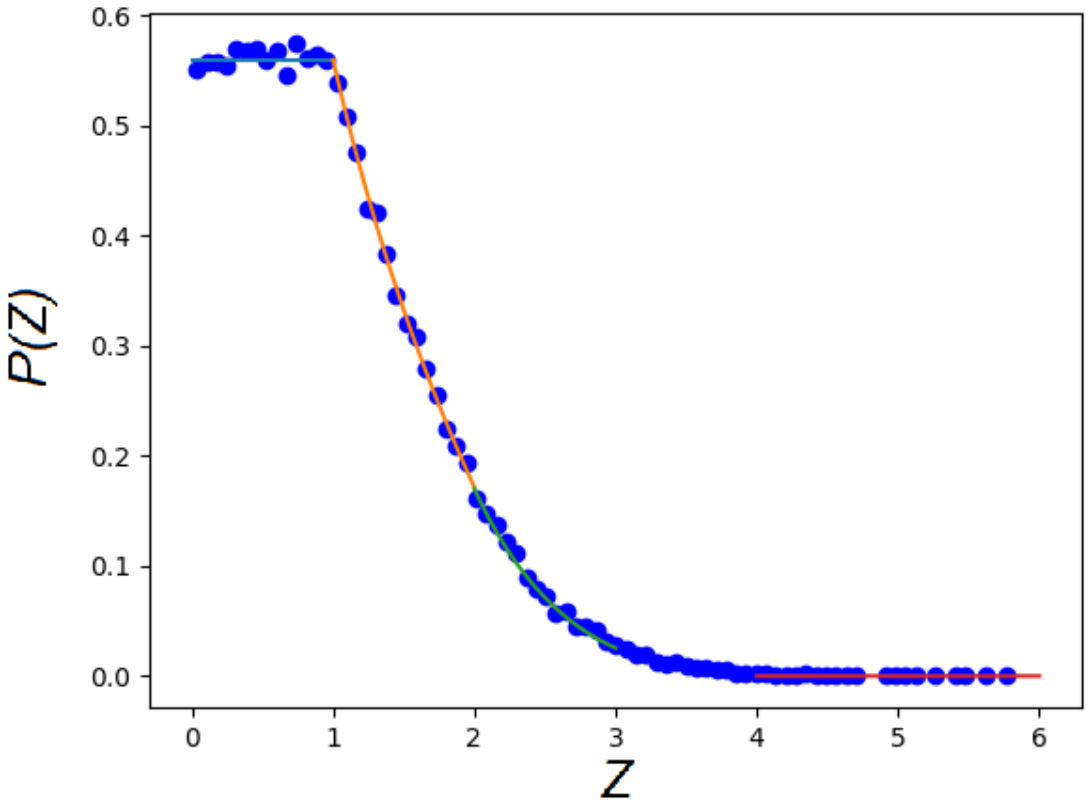

Figure 3.9 – Distribution of Z and corresponding analytical solution up to 3. Here the distribution for $a$is a uniform distribution between $b = 0$ and $c = 1$ and the number of repetitions for the statistics is 10000 and $N = 1000$. Starting from $Z = 4$ is plotted the asymptotic solution which fits already very well the numerical distribution. The interval between $(3, 4)$ was too complicated to be plotted and anyway does not really add something to the solution.

gives us the asymptotic behaviour of $P(Z)$. That is:

$$p'(Z) = -\frac{p(Z)}{Z} \Rightarrow p(Z) = \frac{k_\infty}{Z} \qquad \text{for} \qquad z \gg 1. \tag{3.80}$$

One could wonder that this asymptotic solution is not normalizable, since $\int_a^\infty 1/Z$ diverges, but we do not have to forget what Z is, *i.e.* that is for a uniform in the interval $[0, 1)$ (numerically 1 is not included, but even if included the same would be true), Z will always stop at some finite value. This solution is immediately generalisable to the case where the upper bound c is any real positive number. However, in this case, we could pay attention if $c > 1$: Z is allowed to go to infinity and the normalisation will impose the asymptotic solution to be flat and going to 0 for all the positive real axis. We can notice that this flat solution is also a solution of Eq. (3.75), the convergence to this solution nevertheless turned out to be very slow and then before convergence, as far as Z is bounded, the observed solution will be still the same as for $c \leq 1$.

The global solution and the corresponding simulation are given in Figure 3.9 where for $Z > 4$ the asymptotic solution has been employed with a very good agreement with the simulations. Of course, we can see that for these parameters of the distribution

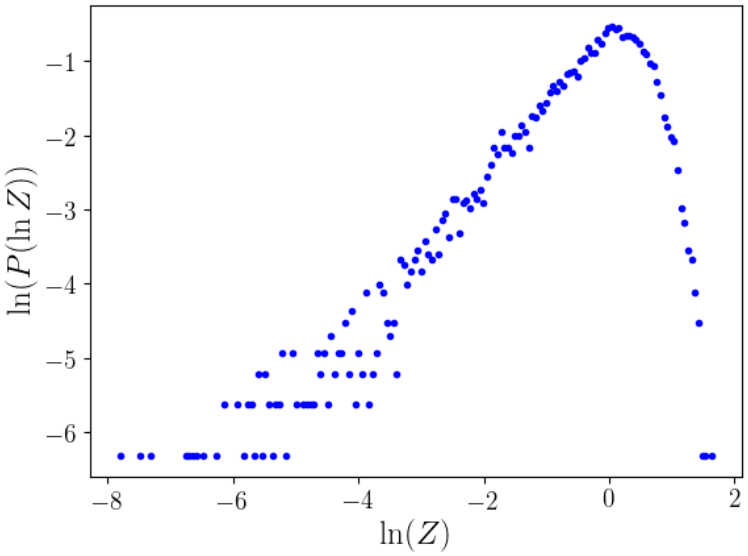

Figure 3.10 – Distribution of $\ln(Z)$ in a log-log plot. All the parameters are the same as in Fig. 3.9.

$p(Z)$ is far from being log-normal, but solving Equation (3.70) for other cases, even only approximately can give us some good analytical insights about the final distribution. For curiosity, to compare it with distributions obtained in other sections and in Fig. 3.7**B** we plot in Figure 3.10 the distribution for $\ln(Z)$ for the same parameters as in Fig. 3.9, in a log-log plot.

3.6 Divergent case

In this section we are going to analyse what happens in all the other cases where $\tilde{a}$ and $\bar{a}$ are not smaller than 1. All these cases are called here divergent cases, because either the variance or both the variance and the mean of Z diverge with the time steps number, N. As we will see and as it is shown in the schematic representation in Fig. 3.2, a different type of divergence (with reference to the cases listed in the previous section) brings to different types of distributions of the sum Z. These divergent cases are considered less important for the modelling, since, in the experiments we did and discussed, there are no such things like divergent avalanches or avalanches with divergent variance. Nonetheless they are still of interest, first to give a global picture about the model and secondly because this could be useful in general to understand the phenomenology of the sum of correlated (log-normal) random variables.

Moreover, all right, it is true that in experiments things are not diverging, but the reason of this may be simply because of the finite size of real and living objects. Therefore it can still be possible that in experiments we see similar distributions to the ones presented in this section with a certain cutoff given by the finite size. Thinking for example about other models discussed in Chapter 1 or 4 from the book (Salje et al., 2017) to model avalanches in functional materials and geophysics, or in phase transitions, or even in brain signals (Lo et al., 2002; Beggs and Plenz, 2003) they are all fat tail distributions for which

this model and this theory could be applied. In general in physics near criticality the appearance of fat tail (power-law) distributions is typical. If the model is versatile and shows to be able to reproduce different phenomena and predicts transitions from one to another it just means that it gives a more complete understanding of the process. With a view to the global problem of modelling processes relying on sums of random variables (correlated or not), like for instance extensive physical quantities as the energy, divergent cases become relevant.

3.6.1 Case $\tilde{a} = 1$

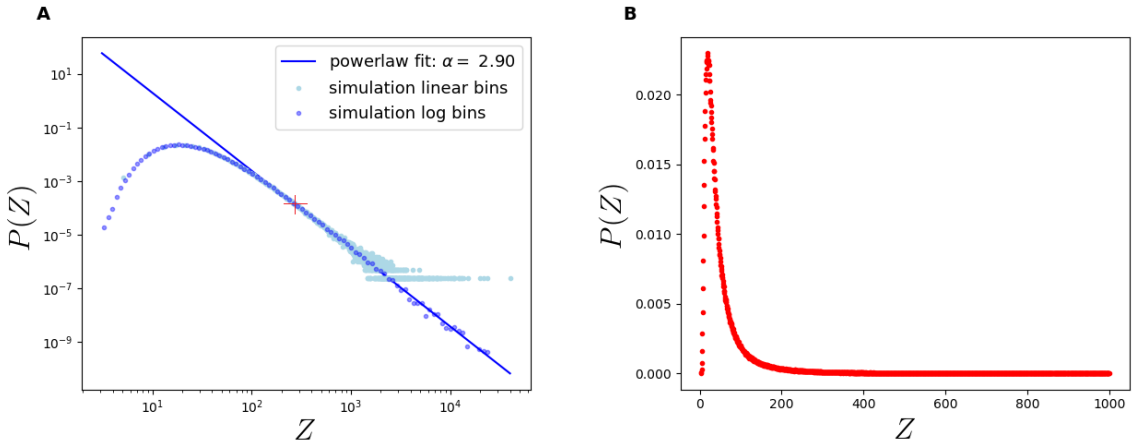

Figure 3.11 – A. Distribution of $\log(Z)$ in a log-log plot. The simulation is generated with $a_j \sim \mathcal{U}_{[0.67;1.3]}$, where the lower bound $b \simeq 0.67$ has been chosen in order to have $\tilde{a} = 1$. The number of realizations for the statistics is 10^6 and the time length $N = 500$. The binning of the dark blue curve is done with an exponentially increasing width, while the shaded dots represent the same distribution with linearly equally spaced bins. The straight line is a fit of the tail with the method explained in the text, starting from the optimal value of 254 (the red cross in the plot). We can see that the tail has a good power law decay with an exponent close to 3 ($\alpha = 2.90$) which contains the asymptotic properties of the Z distribution corresponding to $N \to \infty$. **B.** Distribution of Z in a lin-lin plot, for the same parameters as before, with a linear standard binning.

This is the weakest degree of divergence: the expected value of Z converges (Eq. (3.51)) and its variance diverges linearly with N (Eq. (3.63)). This is the first critical point we encounter while increasing the parameter $\bar{a}$ (cf. Fig. 3.2), that marks the passage from a finite scale distribution to a scale free distribution. To follow the schematic picture in Fig. 3.2, here we are going to take as an example the case where we fix $c = 1.3$ and we will move the other parameter b, in order to visit all the possibilities in the $(\bar{a}, \tilde{a})$ space. Then when $\tilde{a} = 1$ we have $b \simeq 0.67$. This value is simply computed by inverting the relation between the two bounds of the uniform distribution and its second central moment: $1/3\,(b^2 + c^2 + bc) = 1$. The average of the distribution $\bar{a}$ is here 0.98. We can see from Figure 3.11 the appearance of a fat tail distribution, replacing the less skewed simile log-normal decay we discussed in previous sections. The distribution has a peak, the mode, of around 17 and an average of 58.2 (Formula (3.51)). We can estimate the exponent of the decay to be $\alpha = 2.90$. Here and everywhere in this section the error on the statistical estimates is on the last digit).

In the plot **A** of Fig. 3.11 we have drawn the distribution of Z in equally spaced bins on the logarithmic axis (exponentially increasing widths). This allows us to give a reliable representation of the tail of the distribution, that otherwise would appear like a plateau at high values (as we can see in the plot in light blue, where the standard linear binning is taken), due to the very low probability of observing these values. An exponentially increasing binning, instead increases the likelihood to observe large values in each bin, since bins corresponding to large values are larger, and then after proper renormalisation resulting in a larger interval for the visualisation of the tail, improving the appearance of the plot. The price to pay for this is a smaller resolution all over the Z range.

The mathematical procedure to fit the distribution tail is the one described in (Clauset et al., 2009) and the software used is a modified version of the python module described here (Alstott et al., 2014) (with some added features, in particular a corrected definition of the cumulative distribution used to perform the fit). Briefly the fit procedure consists in identifying which is the best interval of the domain to be appropriately fitted by a power law and it works like that:

- the first step of the algorithm is to order the data and to consider all the possible minimum values Z_{min} of the interval, starting from the minimum Z in the distribution, while the maximum value, being the power law a tail distribution, is considered here the maximum of Z;
- for each of the Z_{min} a fit by maximum of likelihood method is performed (α can be computed analytically every time via the formula $\alpha = 1 + n/\sum_{i=1}^{n} \log(Z_i/Z_{min})$, with n size of the subset $Z \geq Z_{min}$) and the Kolmogorov-Smirnov (KS) distance (Frank and Massey, 1951) between the cumulative distribution of the fit and the cumulative distribution of the simulated data (the KS distance is the absolute maximum distance between the 2 distributions), is computed;
- the Z_{min} which has the smallest KS distance is taken as the best starting point of the power-law tail and the corresponding value of the exponent α is considered as the best fit of the tail.

Of course the fit fails if the tail is not sufficiently a power law function because in this case the best Z_{min} chosen would be the point just before the maximum Z, resulting in a 2 point fit that has to be discarded. This is not the case here, where the best Z_{min} resulted in $Z_{min} \simeq 254$, corresponding to the red cross in plot **A**. The exponent α found is coherent with the fact that the variance of Z starts to diverge, since the second central moment

$$\int_{Z_{min}}^{\infty} Z^{-\alpha} Z^2 dZ \tag{3.81}$$

is a divergent integral for $\alpha < 3$. At the same time for this value of α the mean value converges as we found with the full calculation. The convergence of the mean value is assured as far as $\alpha > 2$. Of course the value of the exponent reported here is given for a fixed $N = 500$, hence the variance is not infinite. However the value chosen for N,

and the same will be true in the following, is big enough for the variable Z to be in the asymptotic behaviour.

This can be verified simply by increasing the value of N and comparing the distributions at different values of N: after a certain N the distribution does not change, only the tail is better defined since the probability to have large values is increased. An exhaustive analysis on the effect of N will be shown in the following, here we point out this by performing a statistical analysis using a nice feature of the module (Alstott et al., 2014). A likelihood-ratio test (Severini, 2001) (a test based on the ratio of the likelihood of the model fitted by two different distributions) can be done to compare a power law distribution with a truncated power law, with distribution $kx^{-\alpha}e^{-\tau x}$ (a power law with an exponential cutoff with τ decay rate and k normalisation constant).

The truncated power law distribution is often used to model real systems that show a power law behaviour, like processes on networks, that are limited in the sizes giving rise to a cutoff corresponding to the largest possible system size, that is what is referred to as *high-degree cutoff* in the theory of scale free networks (see Ch. 4 of (Barabási and Pósfai, 2016)). The result of the test gives a p-value of 0.001, with a likelihood ratio of 2.26, indicating that the truncated power law is a better fit for the tail distribution, considering, as usual, a statistical significance threshold of 0.05. The maximum likelihood estimates of the parameters for the truncated power law distribution are $\alpha \simeq 2.85$ and $\tau \simeq 4.7 \cdot 10^{-5}$, hence values very close to the non truncated frame (τ is close, but still significantly different, to zero). This is actually just a signature of the cut off given by the chosen finite N. Indeed if we take $N = 5000$ the p-value becomes 0.09 all other parameters being equal, saying that the truncated power law is not a better fit of the distribution, but the power law exponent stays approximately the same ($\alpha \simeq 2.90$).

We can thus say that the tail of the asymptotic (for $N \to \infty$) distribution has a power-law decay with an exponent close to 3, therefore close to the smallest value of α to have a divergent variance (a value larger than 3 would give a finite variance) confirming that the variance is diverging less fast than an exponential with N.

We can have an analytical insight of this going back to the basic differential Equation (3.74), in the large Z limit, to see the behaviour of the tail, this equation becomes (neglecting the -1 in the RHS):

$$p'(Z) = \frac{p\left(\frac{Z}{b}\right) - p\left(\frac{Z}{c}\right)}{(c-b)\,Z}. \tag{3.82}$$

Injecting in this equation $p \sim x^{-\alpha}$ yields an equation for α depending on the bounds of the uniform distribution (b, c):

$$-\alpha = \frac{c^{-\alpha} - b^{-\alpha}}{b^{-\alpha}c^{-\alpha}\,(c-b)}, \tag{3.83}$$

which has a trivial solution $\alpha = 0$ that is not compatible neither with the definition of a probability distribution other than the one being zero after a certain value, nor with the observed one. There exists also an other easy solution for $\alpha = 1$, which does not have the properties and the shape of the observed one, and it would not be normalisable in the limit

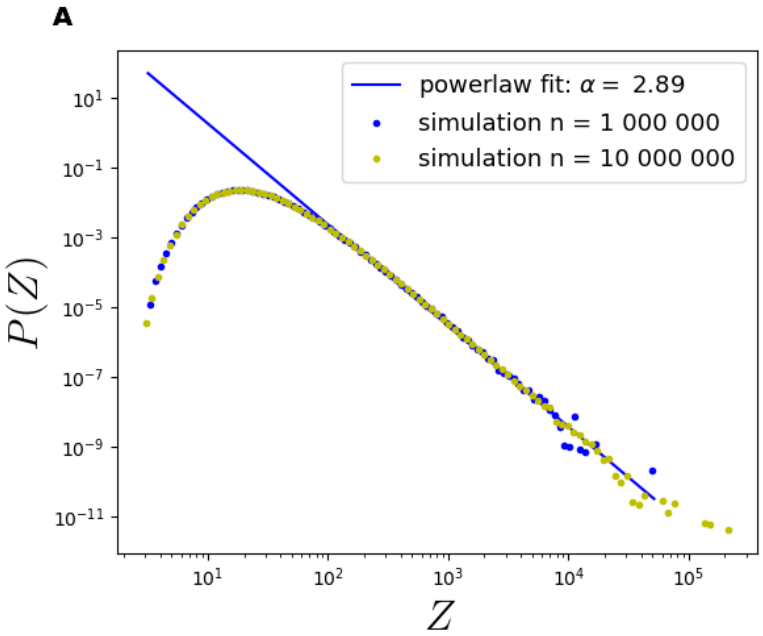

Figure 3.12 – In this figure we compare the distribution of Z in a log-log representation for 2 different values of the number of repetitions n, $n = 10^6$ (the value chosen all over the section, blue dots) and $n = 10^7$ (yellow dots). The power law fit following the procedure described in the text for the value $n = 10^6$ is shown, the small variation of α is explained by the fact it is simulation dependent within the error bounds.

$Z \to \infty$ and then again not compatible with the definition of a probability distribution. Notably this equation has also another solution that can be found numerically and gives as result $\alpha = 3$ for $b \simeq 0.67$ and $c = 1.3$. Therefore the asymptotic behaviour of $p(Z)$ is proved to follow a power-law tailed distribution, which is an asymptotic solution of the differential equation.

This result is universally true in the case $\tilde{a} = 1$ and does not depend on the particular bounds b and c: it always gives the same result of $\alpha = 3$ defining uniquely the tail power law decay of the distribution of Z. This universality can be seen replacing in Equation (3.83) the expression of b and c written in function of $\tilde{a}$ and $\bar{a}$ computed by inverting the expressions of the first and second central moment of a uniform distribution ($\bar{a} = \frac{b+c}{2}$ and $\tilde{a} = \frac{1}{3}(b^2 + bc + c^2)$):

$$\begin{aligned} b &= \bar{a} - \sqrt{3}\sqrt{\tilde{a} - \bar{a}^2} \\ c &= \bar{a} + \sqrt{3}\sqrt{\tilde{a} - \bar{a}^2}, \end{aligned} \tag{3.84}$$

and hence Equation (3.83) becomes:

$$-\alpha = \frac{\left(\bar{a} - \sqrt{3}\sqrt{\tilde{a} - \bar{a}^2}\right)^{\alpha} - \left(\sqrt{3}\sqrt{\tilde{a} - \bar{a}^2} + \bar{a}\right)^{\alpha}}{2\sqrt{3}\sqrt{\tilde{a} - \bar{a}^2}}.. \tag{3.85}$$

As a side remark the argument of square root is always definite positive, because is the variance of variable a. Injecting $\alpha = 3$ in the RHS of (3.85) yields:

$$-\alpha = -3\tilde{a}. \tag{3.86}$$

Observing that the dependence on $\bar{a}$ vanishes, we conclude the power-law decay is not dependent on the model parameters, provided $\tilde{a} = 1$ and it is thus equal to 3 for whatever bounds b and c we choose for the uniform distribution. This is true only in this particular case where $\tilde{a} = 1$, in general the exponent depends on both parameters $\bar{a}$ and $\tilde{a}$, as in Eq. (3.85). We arrived at the same conclusions also with other simulations (not shown here), measuring the same asymptotic power-law decay with the statistical method described above, for different b and c keeping $\tilde{a} = 1$ but then necessarily with different $\bar{a}$.

Finally we want to point out the effect of the number of repetitions n for the statistics (see Fig. 3.12). Increasing n does not change the distribution $p(Z)$, α is the same (within the statistical error bounds here 0.02), the tail properties are the same, *i.e.* the truncated power law is still a better fit of the tail. The 2 probability density functions overlap meaning that the chosen number of repetitions, 10^6 is large enough for the statistical representation of the process. Nevertheless the tail looks lightly larger (the yellow distribution spreads more to the right), the reason of that is that outcome values which are very large has an exponentially small probability to be drawn because all the z_i terms have to be drawn together on the right tail of the respective log-normal distributions. Therefore increasing n results in a appreciable probability for these rare values.

3.6.2 Case $\tilde{a} > 1$ and $\bar{a} < 1$

In this section we analyse the case $\tilde{a} > 1$ and $\bar{a} < 1$, which is characterised by a convergence of the average value of Z (always following Formula (3.51)) and a variance which is exponentially divergent with N (Eq. (3.65)). Just considering this we would expect a behaviour similar to the one described above, but with a power law decay with a stronger divergence in Z, *i.e.* an α smaller than the value of 3 found above.

We will consider the generating distribution for the simulation to be $a_j \sim \mathcal{U}_{[0.68;1.3]}$, where the choice of $b = 0.68$, is just to have a $\tilde{a} > 1$ and at the same time $\bar{a} < 1$. The upper bound is the same as before to show the differences given only by the change of $\tilde{a}$. The resulting values of $\bar{a}$ and $\tilde{a}$ are respectively 0.99 and 1.01. Anyway as we showed previously and as we will confirm also in this section, the particular choice of the bounds of the distribution does not influence the behaviour of the tail (in the sense that we still have a power-law decay), which in general only depends on the 2 parameters $\bar{a}$ and $\tilde{a}$.

We follow again the same statistical procedure as before, based on the KS distance to find the best Z_{min}. We find here as best $Z_{min} = 350$ for the starting point of the power-law tail. Observing Figure 3.13 we confirm the hypothesis of a power-law decay, even though doing the likelihood ratio test we find again a better fit by a truncated power-law, with a p-value very significant, 0.004. This is again a signature of the finite size of the simulations. The resulting exponent of the maximum of likelihood fit is $\alpha = 2.56$. Apart from the differences in the measured exponent α, there are no differences in the shape of the Z distribution with respect to the case in the previous section: it is a right skewed distribution with average value given by Formula (3.51) of 99 and a peak (mode), around 22.

By solving numerically Eq. (3.85) we found a value $\alpha \simeq 2.62$, very close to our statistical estimation. We can thus see from the general Equation (3.85) that here the

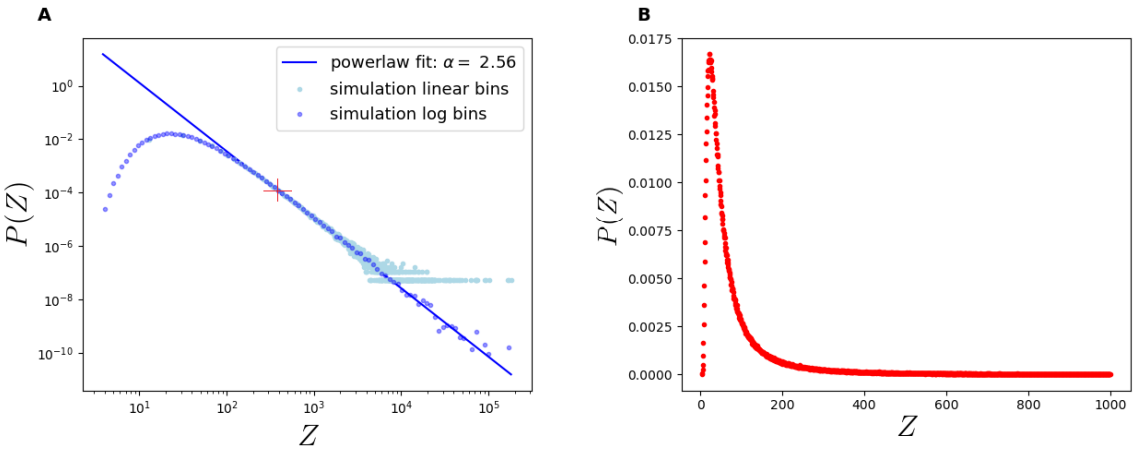

Figure 3.13 – **A.** Distribution of Z in a log-log plot. The simulation is generated with $a_j \sim \mathcal{U}_{[0.68;1.3]}$, where the lower bound $b = 0.68$ has been chosen in order to have $\tilde{a} > 1$ and $\bar{a} < 1$. The number of realizations for the statistics is 10^6 and the time length $N = 500$. The binning of the dark blue curve is done with an exponentially increasing width, while the shaded dots represent the same distribution with linearly equally spaced bins. The straight line is a fit of the tail with the method explained in the text, starting from the optimal value of $Z \simeq 350$ (the red cross in the plot). We can see that the distribution has a good power law tail with an exponent estimated of $\alpha = 2.56$) which contains the asymptotic properties of the Z distribution for $N \to \infty$. **B.** Distribution of Z in a lin-lin plot, for the same parameters as before, with a linear standard binning.

value of the exponent depends on both parameters $\bar{a}$ and $\tilde{a}$, in particular it can be checked that increasing $\bar{a}$ towards 1 (that will also consequently increase $\tilde{a}$) causes a decrease of α towards 2. The range of values of the exponent α is then between 3 and 2, within this interval the variance s_Z^2 of $p(Z)$ is power-law divergent for large Z, $s_Z^2 \sim Z^{-\alpha+3}$ (see Eq. (3.81)), but since $\alpha > 2$ the mean value is convergent at large Z, as expected.

As a final remark of these observations, we can conclude that in all the power-law decaying cases, the particular finite value of N, necessarily chosen for computer simulations, is not important for the final distribution $p(Z)$ which has the same shape regardless of N, as far as it is big enough, in a sense we are going to clarify later on. We analyse here in more details the effect of a finite time size N on the final results, but the same picture applies also to the other power-law decaying cases.

The average of $p(Z)$, m_Z, varies with N (Eq. (3.54) or (3.56) for $\bar{a} = 1$) and we can see in Fig. 3.14**A** the fast rate of convergence to the asymptotic value given by Eq. (3.51), at the chosen $N = 500$, the error between the result of Eq. (3.51) is only 1% meaning that the average will not shift for greater time sizes $N > 500$, and the asymptote can be considered as achieved. Also, we can observe that the yellow curve, representing for simulations m_Z *vs* N, overlaps the theoretical one (in blue) within a maximum error of 0.3%. This is not true by looking at plot **B**, where the variance of Z, s_Z^2, starting from $N \simeq 200$ becomes quite different from the expected one (on the y-axis is plotted the $\log s_Z^2$). We can observe that after a certain value of N, around 110 the blue curve becomes linear, showing an exponential asymptotic growth with N. Indeed a linear curve in a log-lin plot means an exponential growth.

The discrepancy between the yellow and the blue curve is due to a numerical error. More precisely the number of repetitions for the statistics, n, as described in Section 3.6.1,

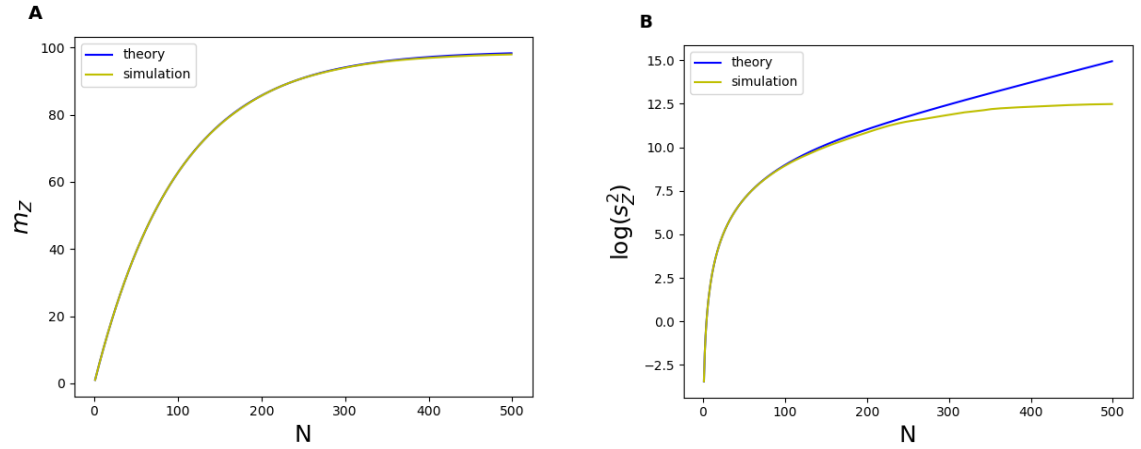

Figure 3.14 – **A.** Theoretical average value given by Formula (3.54) in blue and average value of Z obtained from the simulation (yellow) with the same parameters as in Fig. 3.13 versus time step N. Notice here the rate of convergence to the final value of (3.51), here 99, which can be considered as achieved for $N > 300$ (with an error of 5%, while for $N = 500$ it is 1%). The two curves overlap almost perfectly. **B.** Theoretical variance of Z given by Formula (3.64) in blue and the simulation (in yellow) result with the same parameters as before, both in log scale, versus time step N. After a certain value of N, around 110 the blue curve becomes linear, showing an exponential asymptote in N. The discrepancy between the blue and the yellow curve is due to numerical errors and by the finite number of repetition for the statistics, which should also be increased exponentially to represent correctly the large values of Z. Its impact on the distribution is discussed in the text.

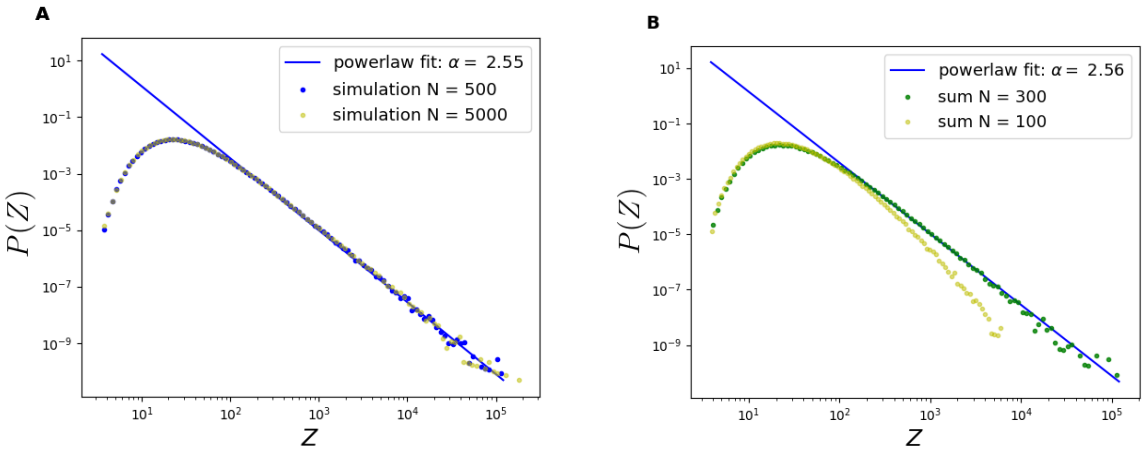

Figure 3.15 – **A.** The distribution $p(Z)$ in a log-log plot for 2 different values of the time size N, 500 (blue dots) and 5000 (yellow dots). The other parameters of the distribution are the same as in Fig. 3.13. The blue straight line is the a fit with a power-law decay with the method explained in the text for the $N = 500$ simulation. **B.** Distributions of Z for $N = 100$ (yellow dots) and $N = 300$ (green dots). The other parameters of the distribution are the same as in Fig. 3.13. The blue line is the power law fit of plot of Fig. 3.13, *i.e.* the $N = 500$ simulation. We can observe that already for $N = 300$ (green dots) the final distribution of Z is very close to the asymptotic one, *i.e.* the blue one.

should increase exponentially to have the expected variance, which depends strongly on large outliers of Z. This was not possible for computer memory size reasons. Observing Figure 3.15**A** we can ascertain that nonetheless the difference on s_Z^2 does not appreciably change the distribution $p(Z)$, because for different N, 500 and 5000, $p(Z)$ stays the same, while the error becomes much larger. On the other hand in Figure 3.15**B** we compare $N = 100$ (yellow dots) and $N = 300$ (green dots), and we note that the yellow curve is not following the asymptotic power-law decay represented by the blue line, which is the fit of the $N = 500$ simulation. We can conclude that choosing a value of N bigger than the value where the crossover to the asymptotic increase of s_Z^2 takes place (here for $N \simeq 110$, after which the increase is exponential), should be enough. The analytical results allow us to quantify the effect of this error, in fact this is what creates the small difference between the estimated exponent α and the theoretical one. At the end the effect of this error is not very important for the final results, as we could see previously, apart from a very small truncation effect at large values (not even visible but only statistically detectable) and consequently a small error on the estimation of the exponent.

Empirically we always check that if increasing N the distribution stays the same, like in Fig. 3.15, that means we reached the asymptotic shape of the distribution, in the large N limit. In this sense we can say that the shown probability distribution is the asymptotic one, since the distribution becomes stable after a certain value of N, which has to be large enough in order to capture the *asymptotic tendency* of the distribution.

3.6.3 Case $\bar{a} = 1$

We consider here the case $\bar{a} = 1$ and we take as usual $c = 1.3$ and consequently $b = 0.7$, therefore $\tilde{a} \simeq 1.03$. The point $\bar{a} = 1$ is the critical point for the average of Z to start to be divergent. Indeed the average follows Eq. (3.56), which is linearly diverging in N. At the same time also the variance diverges and this happens again exponentially in N, because even though in Eq. (3.61) there are polynomial terms in N, the asymptotic behaviour for large N is led by the exponential terms with base $\tilde{a}$, since $\tilde{a} > 1$. Following the reasoning developed so far in this section we would expect a power-law tail with en exponent close to 2.

This is actually what appears from Fig. 3.16, where after applying the adapted statistical procedure described in Section 3.6.1, we find an exponent $\alpha \simeq 2.03$ and an optimal $Z_{min} \simeq 456$. The shape of the distribution is similar to the previous ones, with a statistical mode of around 29 and an average value going to ∞ (equal to 453 for the given $N = 500$). Exactly in the same way as before we can compare the distribution of the tail to other possible distributions via the likelihood ratio test. We encounter the effect of the finite size N on the tail of the distribution: the truncated power-law here is preferred with an extremely significant p-value of 0. Unlike previous cases the average also depends on N, with the effect of a heavier tail influencing the mean value of Z, but, as before, the distribution becomes stable for large N and after a certain N, does not change any longer. This is verified numerically for the chose $N = 500$, but also analytically: starting from the same Equation (3.85) we find an asymptotic exponent for large Z exactly equal to 2. Also in this case then, with $N \to \infty$ the distribution confirms to have an asymptotic

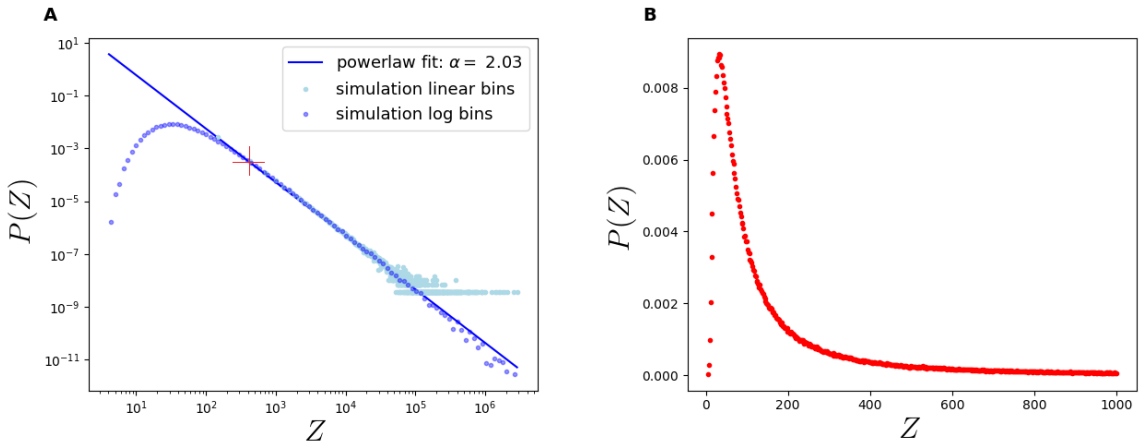

Figure 3.16 – A. Distribution of Z in a log-log plot. The simulation is generated with $a_j \sim \mathcal{U}_{[0.7;1.3]}$, where the lower bound $b = 0.7$ has been chosen in order to have $\bar{a} = 1$. The number of realizations for the statistics is 10^6 and the time length $N = 500$. The binning of the dark blue curve is done with an exponentially increasing width, while the shaded dots represent the same distribution with linearly equally spaced bins. The straight line is a fit of the tail with the method explained in the text, starting from the optimal value of $Z_{min} \simeq 456$ (the red cross in the plot). We can see that the tail has a good power law tail with an exponent estimated of $\alpha = 2.03$ which contains the asymptotic properties of the Z distribution for $N \to \infty$. **B.** Distribution of Z in a lin-lin plot, for the same parameters as before, with a linear standard binning.

power law shape with an exponent $\alpha = 2$.

Once we have computed the exponent α, if we re-inject the obtained value ($\alpha = 2$) in the RHS of Eq. (3.85) we notice that the dependence on $\tilde{a}$ vanishes and thus α is always equal to 2 for any values of $\tilde{a}$, *i.e.* of the bounds b and c, provided $\bar{a} = 1$. We could say that this is the mirror case of case $\tilde{a} = 1$, but here the tail is fatter, indeed with an exponent of 2 even the average diverges going as $\log(Z)$ for large Z, by solving $\int_{Z_{min}}^{\infty} Z^{-\alpha} Z dZ$.

At this point we can conclude that the variance and the average show a correspondence between the dependence on N and the behaviour of the tail of the distribution and therefore also its asymptotic dependence on Z, for large Z. Indeed, for large N and large Z, a linear dependence on N, for the average and/or the variance of Z, is translated as a logarithmic dependence on Z, while an exponential increase with N is translated as a power dependency on Z with an exponent $\zeta > 0$. The divergence with ζ comes from the fat power law tail, with an exponent α causing the divergence of the considered statistic, *i.e.* of integral $\int_{Z_{min}}^{\infty} Z^{-\alpha} Z dZ$ for the average and $\int_{Z_{min}}^{\infty} Z^{-\alpha} Z^2 dZ$ for the variance. Instead, if the average and/or the variance of Z converge to a finite value at large N, the corresponding exponent at large Z will be negative, $\zeta < 0$. Going to the limit $N \to \infty$ reflects then so far the limit $z \to \infty$.

In order to compare with the distributions shown in other sections, where we find log-normal like behaviour, we show in Figure 3.17 the distribution of $\log(Z)$, which is different from the distribution of Z of Fig 3.16**A**: changing variable from Z to $t = \log(Z)$, the distribution is multiplied by the factor of e^t, and then the shape is modified. In particular the tail (the distribution for large t, corresponding to large Z) is distributed

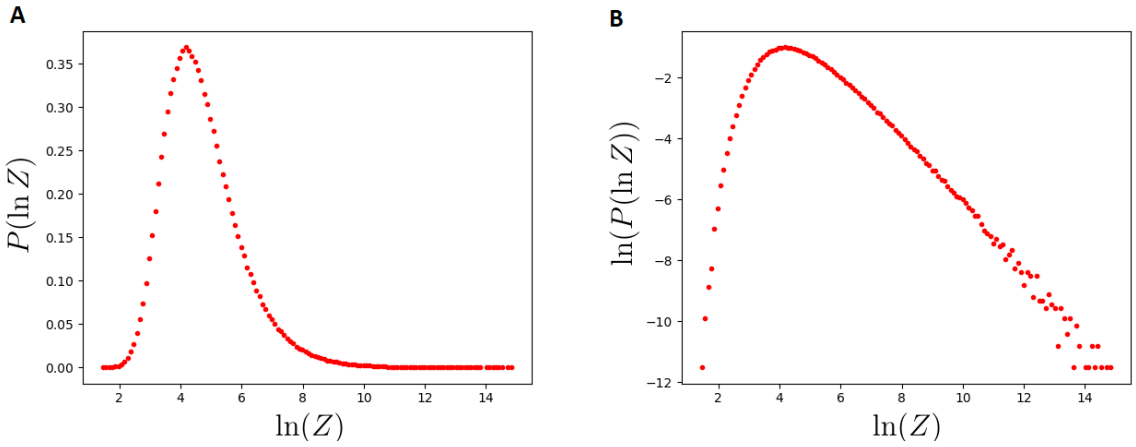

Figure 3.17 – **A.** Distribution of $\log(Z)$ (base e). The simulation is generated with $a_j \sim \mathcal{U}_{[0.7;1.3]}$. The number of realizations for the statistics is 10^6 and the time length $N = 500$. We can see that the tail is heavier than a log-normal tail, resulting in a distribution that is more skewed to the right than a log-normal (a log-normal would appear as a Gaussian in this plot). **B.** Distribution of $\log(Z)$ in a log-lin plot, for the same parameters as before. For purposes of comparison a log-normal distribution here would appear as a parabola. This plots are given as a comparison with other sections where the distribution is very close to log-normal.

by:

$$g(t) \sim e^{(1-\alpha)t}, \tag{3.87}$$

resulting in a distribution more skewed to the right than a log-normal. In fact in both Fig. 3.17**A** and **B**, we can see how the distribution changes and the differences with respect to a log-normally distributed random variable which logarithm would be Gaussian, *i.e.* a parabola in plot **B**. Instead, the tail follow an exponential decay with $\log(Z)$ which results in a straight line with slope -1 (being $\alpha = 2$), if we plot the logarithm of the distribution as in Fig. 3.17 **B**.

Remarkably, we notice that from the tail exponent we can go back to the parameters of the underlying distribution of the process governing the avalanche, which is in this case a uniform distribution, but can actually be any other distribution respecting the properties discussed in Section 3.3. This is important for instance if we want to compare the model with experiments, we can imagine to estimate the power law exponent from observed data and with this estimate infer the parameters $\bar{a}$ and $\tilde{a}$, and therefore also the variance of the distribution which governs the avalanche.

Finally the range of exponents α found in the power law decaying cases analysed in this chapter falls in the range of the most common exponents found in the degree distributions of scale free real network (for example the protein interaction network or the actor network (Barabási and Pósfai, 2016)) and in most solid and known examples of power-law tailed distributions, as the distribution of frequency of use of English words or number of hits on web sites (Newman, 2005).

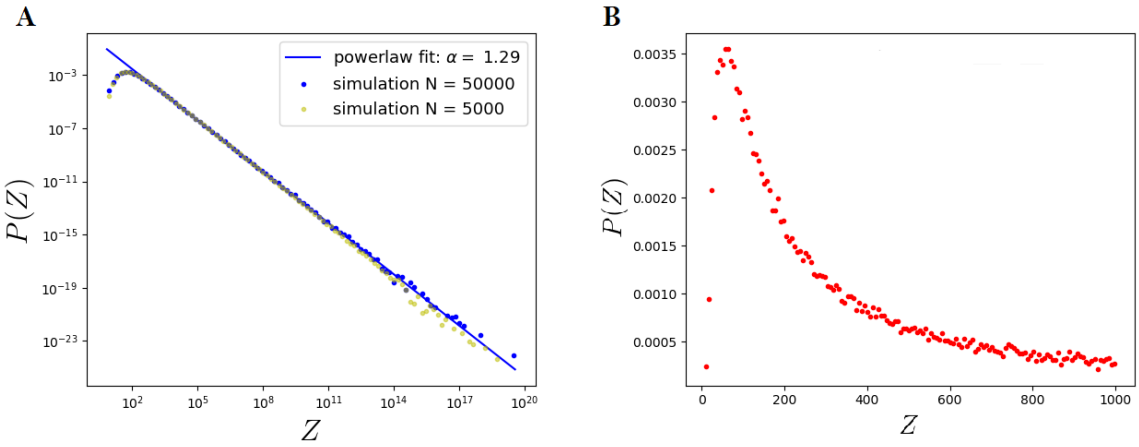

Figure 3.18 – A. Distribution of Z in a log-log plot. The simulation is generated with $a_j \sim \mathcal{U}_{[0.72;1.3]}$, where the lower bound $b = 0.72$ has been chosen in order to have $\bar{a} > 1$, but also $\alpha > 1$. The number of realizations for the statistics is 10^6 for the time length $N = 5000$ (yellow dots) and 100000 for the time length $N = 50000$, because of computer memory size limits. The binning is done with an exponentially increasing width. The straight line is a fit of the tail of the $N = 50000$ simulation, with the method explained in the text, starting from the optimal value of $Z_{min} \simeq 4 \cdot 10^5$. We can see that the tail has a good power law tail with an exponent estimated of $\alpha = 1.29$, which contains the asymptotic properties of the Z distribution for $N \to \infty$. **B.** The distribution of Z in a lin-lin plot, for the same parameters as before ($N = 50000$), with a linear standard binning.

3.6.4 Case $\bar{a} > 1$

This is maybe the most interesting case between the divergent ones. For $\bar{a} > 1$ for sure also $\tilde{a} > 1$ and then both the variance and the average of Z diverge exponentially with N. Following the parallelism between the behaviour for large N and the power law decay for large Z, we would expect a power law exponent $\alpha < 2$. At the same time we can imagine that as far as $\bar{a}$ and $\tilde{a}$ increase, the exponential divergence with N will be too fast to be «contained» in the tail of $p(Z)$, and the distribution will start to shift to the right. Intuitively, this is what happens and we are now going to analyse this transition with more details.

Taking as a guide the case $c = 1.3$, for all $b > 0.7$ we respect the condition $\bar{a} > 1$. Let us take $b = 0.72$, corresponding to $\bar{a} = 1.01$ and $\tilde{a} \simeq 1.05$; the reason why we take this value will be clearer in the following, but it is a good starting point to understand what happens. In Fig. 3.18 we draw the distribution obtained from simulations for $N = 5000$ (yellow dots) and $N = 50000$ (blue dots), in order check we reached the asymptotic tail distribution, as we described previously. The distribution has a very large power-law tail, starting from $Z_{min} \simeq 400000$, and an exponent $\alpha \simeq 1.29$, both estimated with the statistical method described in Section 3.6.1 applied to the $N = 50000$ simulation. The distribution looks then very similar to the previous ones, but with a much heavier tail, given by an exponent $\alpha < 2$, causing the divergence of both the average and the variance of Z for large Z. The peak of the distribution is around $Z = 56$. Let us remind now the analytical formula used to compute α (Eq. (3.83)). If we solve it numerically we get $\alpha \simeq 1.29$, perfectly consistent with the numerical result.

This is not all, having a look at Figure 3.19 we observe the general behaviour of the

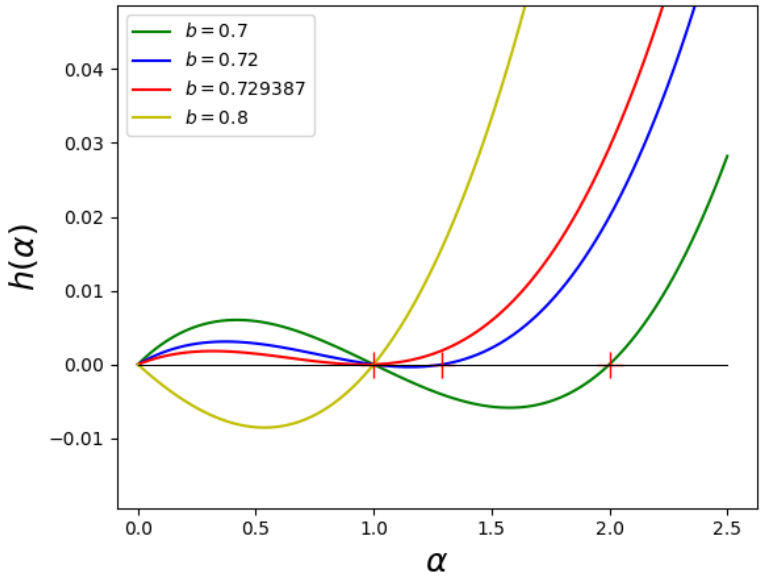

Figure 3.19 – In this figure we show the behaviour of the roots of Eq. (3.83). The roots are represented with red crosses. Each curve is the function $h(\alpha)$ defined in Eq. (3.88) for different b as specified in the legend of the plot, $c = 1.3$. The black horizontal line is the zero reference.

roots Equation (3.85). On the y axis we plot the function of α:

$$h(\alpha) = -\alpha - \frac{c^{-\alpha} - b^{-\alpha}}{b^{-\alpha}c^{-\alpha}\,(c-b)}, \tag{3.88}$$

which roots are the solutions of Eq. (3.83). As a reference the black horizontal line gives the value of 0 and the red crosses are the accepted solutions (only those solutions which correspond to the real behaviour of the tail, excluding then $\alpha = 1, 0$) for each one of the chosen parameters b. As we know, the accepted solution for $b = 0.7$ is $\alpha = 2$; by increasing b (and then also both $\bar{a}$ and $\tilde{a}$) we have a region where the solution α goes from 2 to 1, 1 being the smallest possible value of the exponent, obtained for $b \simeq 0.73$. Therefore there is a region in the $(\bar{a}, \tilde{a})$ space, with $\bar{a}$ and $\tilde{a}$ lightly larger than 1, where the tail is still a heavy power-law, causing the divergence of both the average and the variance of Z, with an exponent $\zeta > 0$ but always smaller than 1 for the average and than 2 for the variance. We notice that as b increases, starting from 0.7, the curve flattens around 0, until the value where the function $h(\alpha)$ has two coincident roots at point $\alpha = 1$, therefore having $h(\alpha)$ a local minimum in $\alpha = 1$. This is the critical point that brings the distribution to change completely shape. It is like if the process wanted an α smaller than 1, and then a faster divergence of m_Z and s_Z^2, but this is forbidden by the definition itself of probability density function. Indeed, as we noticed in the discussion about Eq. (3.83) the solution $\alpha = 1$ is not consistent with the normalisation of the probability density function $p(Z)$. With this observations we can find a method to compute the critical point, called b^*, and then translate it in the $(\bar{a},\tilde{a})$ plane. To do so, we look for the b which implies a

local minimum, *i.e.* a null derivative of function $h(\alpha)$ at $\alpha = 1$. Deriving h gives:

$$h'(\alpha) = -1 - \frac{1}{c-b}\left[c^\alpha \ln(b) - b^\alpha \ln(c) + \left(c^{-\alpha} - b^{-\alpha}\right)\left(b^\alpha c^\alpha \ln(b) + b^\alpha c^\alpha \ln(c)\right)\right]. \quad (3.89)$$

Imposing this derivative equal to 0 and replacing $\alpha = 1$ gives us the critical b^* for a given c:

$$c\,(1 - \ln(c)) - b^*\,(1 - \ln(b^*)) = 0. \quad (3.90)$$

Unfortunately, this relation does not have a closed form solution, but can be easily numerically solved, with standard numerical methods. Going further we replace in Eq. (3.90) the relations (3.84) and obtain:

$$\frac{\left(\bar{a} + \sqrt{3\,(\tilde{a} - \bar{a}^2)}\right)\left(1 - \ln\left(\bar{a} + \sqrt{3\,(\tilde{a} - \bar{a}^2)}\right)\right)}{\left(\bar{a} - \sqrt{3\,(\tilde{a} - \bar{a}^2)}\right)\left(1 - \ln\left(\bar{a} - \sqrt{3\,(\tilde{a} - \bar{a}^2)}\right)\right)} = 1. \quad (3.91)$$

The $\bar{a}$ and $\tilde{a}$ that verify this relation define a critical curve above which the distribution is no longer a power law. We notice that since an increase of $\bar{a}$ causes an increase of $\tilde{a}$, after a certain $\bar{a}$, we call $\bar{a}^\dagger$, the numerator will increase faster than the denominator and the equation will not have anymore a solution (other that, of course, the trivial zero variance one: $\tilde{a} = \bar{a}^2$), then this critical point will not exist and the distribution $p(Z)$ will not have a power-law shape for any possible choice of $\tilde{a}$. To better define this region we point out that in general for a given $\bar{a}$, $\tilde{a}$ is bounded. For a general distribution of a the only bound is $\tilde{a} > \bar{a}^2$ (positive variance condition), but the positivity of a implies also an upper bound for $\tilde{a}$. This upper bound depends on the particular choice of the distribution of a (for a Gaussian cannot be computed analytically), for the choice of the uniform distribution the bounds are found by first noticing that:

$$\bar{a}^2 = \frac{3}{4}\tilde{a} + \frac{1}{4}bc. \quad (3.92)$$

Therefore for a given $\bar{a}$ the upper bound of $\tilde{a}$ is where $b = 0$, resulting in $4/3\bar{a}^2$. The lower bound is given by the maximum value taken by the product bc, keeping $\bar{a} = (b + c)/2$ fixed. With this constraint we have $bc = (2\bar{a} - c)c$, which has a maximum for $c = \bar{a}$. Injecting this value in Eq. (3.92) we find as lower bound for $\tilde{a}$ the value $4/3\bar{a}^2 - \bar{a}^2/3$, being this the zero variance limit case. We have then for a given $\bar{a}$:

$$\tilde{a} \in \left[\bar{a}^2; \frac{4}{3}\bar{a}^2\right] \quad (3.93)$$

This is true in the region $\bar{a} > 1$, but it is also true in all the parameters plane. Within this region is therefore located the critical line discussed above. The value of $\bar{a}^\dagger$ can be obtained numerically and it is $\simeq 1.36$, corresponding to $\tilde{a}^\dagger \simeq 2.46$.

We now focus on the distribution shape of $p(Z)$ above the critical line, for the cyan region, referring to Fig. 3.2. As we noticed the distribution will start to shift to the right, since both the average and the variance of Z diverge exponentially with N in a fast way

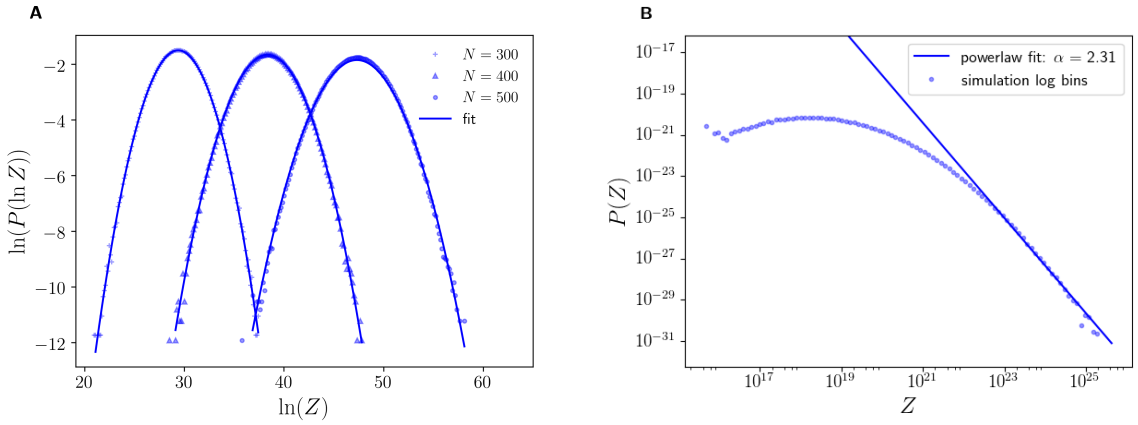

Figure 3.20 – A. The logarithm of the probability density function of $\ln(Z)$ for the process generated by a uniform distribution $a_j \sim \mathcal{U}_{[0.9;1.3]}$ where the bound $b = 0.9$ is chosen to be above the critical line defined by Eq. (3.91). We show the resulting distributions for different values of N: $N = 300$ (crosses), $N = 400$ (triangles), and $N = 500$ (dots). The solid lines are the quadratic fits of each distribution. **B.** The distribution of Z in a log-log plot, with logarithmically spaced bins, as a comparison with previous sections. We can see that the power law tail is completely lost. For both plots the number of repetitions for the statistics is 10^6 and $N = 500$.

that cannot be represented by the tail of the distribution. This means that the mode of the distribution will no longer be stable, as before, but will also move to the right. Increasing N, the distribution will have a transient towards an exact log-normal distribution. Of course it will not be a properly stationary distribution, differently from previous cases, because the mode of the distribution is depending on N, but can still be considered as an asymptotic distribution, because the shape does not change while increasing N, apart from a shift and an increase in the width. In this sense the distribution is stable. In order to show this we take $b = 0.9$, in such a way to be above the critical line, and, as usual, $c = 1.3$. The resulting parameters values are $\bar{a} = 1.1$ and $\tilde{a} \simeq 1.22$.

In Figure 3.20**A** we plot the distribution of $\ln(Z)$ and the corresponding parabolic fit, for different values of N. Since on the y axis we plot the logarithm of the distribution, a parabola corresponds to a Gaussian distribution. If the logarithm of Z is normally distributed, Z is log-normally distributed. Already looking at the parabolic fits this can be confirmed, since the fits follow very well the simulated distribution. We observe also that the distribution is a very good log-normal already for $N = 300$ (and even $N = 100$, not shown here) and that, as we were saying, it shifts to the right and broadens as far as N increases. We also applied the statistical method of Section 3.6.1 to determine the tail properties of the distribution, but we obtained a bad estimate of the exponent ($\alpha = 2.3$), for the obvious reason that the tail does not follow at all a power law decay. We compared via the loglikelihood ratio test the tail distribution to a log-normal tail, and the p-value of the test was $4 \cdot 10^{-6}$, indicating that the tail follows a log-normal decay, instead than a power-law one, with very good statistical significance. But we did not stop here, we used a powerful statistical technique to check how much the distribution is close to a log-normal, or if this result is only approximated, as in the convergent case (see Section 3.5). To understand that, the statistical procedure used to determine the tail

exponent is not sufficient, because we are looking for some statistical information regarding the full distribution, not only the tail. This technique is the bootstrap, described for instance in (Davison and Hinkley, 1997). We developed a python code to apply parametric bootstrap to our case, the bootstrap being parametric since we are supposing that the distribution is a log-normal and then it depends on 2 parameters, μ_Z and σ_Z^2. The goal is to see if the difference between the Z distribution and the supposed one (the log-normal) is comparable with random generated log-normal synthetic values. The algorithm works like that:

1 – we do a maximum likelihood estimate of μ_Z and σ_Z^2 from the original data (the simulation) and we compute the KS distance, D^* (quantifying the distance between the expected distribution and the one of the data, see Section 3.6.1);

2 – we generate new random data with the same size as the original one (the simulation), following a log-normal distribution with the estimated parameters μ_Z and σ_Z^2;

3 – we estimate by maximum of likelihood the parameters μ and σ of the random data from point 2 and the distance D between the distribution of the generated random data and the log-normal distribution having as parameters the new μ and σ;

4 – we repeat 2 and 3 a statistically large number of times (of the order of at least 1000-2000), in order to have a good statistics for the distribution of the distance D.

At the end we have a large number of D values and we can thus plot their distribution. Therefore we can compute the probability that the D obtained with the bootstrap process is bigger than the D^* of the simulation. This gives us a p-value under the null hypothesis: Z follows a log-normal distribution, the p-value will then tell us the probability that Z follows a log-normal. With a typical significance threshold of 5%, we can then conclude about the null hypothesis. Basically all this method consists in checking if the difference between the distribution of Z and the log-normal distribution with parameters μ_Z and σ_Z can be explained by random fluctuations only.

It turns out that the distribution for large N (here is $N = 500$, but already for $N = 300$ the p-value is non significant) in this parameters space region is a log-normal with a non significant p-value > 0.05, meaning that the null hypothesis can be statistically accepted. For the considered parameters the bootstrap p-value is 0.31. This gives us a solid statistical conclusion about the log-normality of the distribution, that does not look like an approximation, and a quantification of the error we make by considering it log-normal. It is worth to mention that the p-value tends to increase with N, confirming an asymptotic convergence to the log-normal distribution, which can nevertheless be very fast and observable also for finite N. In general as we move toward the top-right of Fig. 3.2 the convergence is faster.

As a reminder we stress the fact that in all these divergent cases we still respect condition (3.27) (being $\Delta a^4/\bar{a}^4$ of the order of 10^{-4}, for these particular parameters $\simeq 1.5 \cdot 10{-}4$) and all approximations of Section 3.3 are still valid. Among them it is also still true that we are dealing with the sum of z_i which distributions are quickly converging

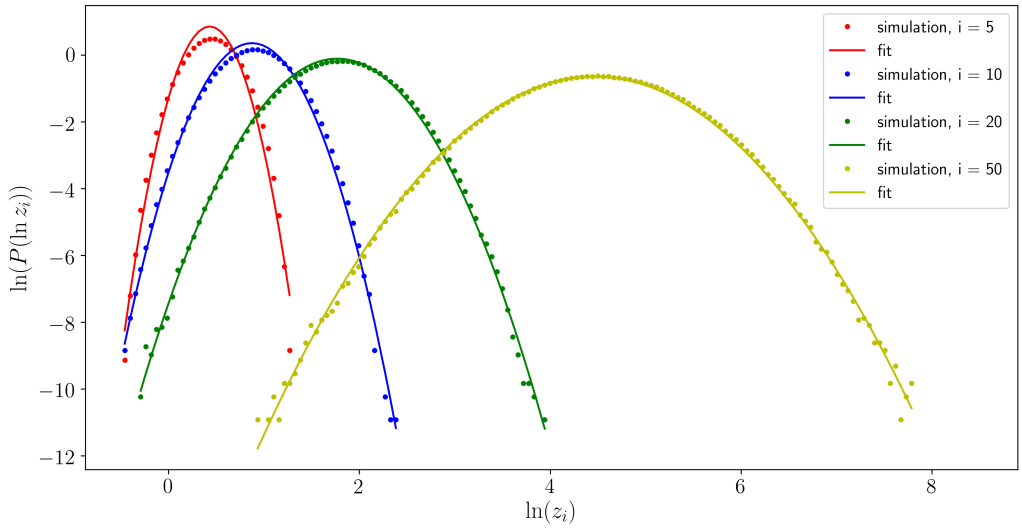

Figure 3.21 – Distributions of the logarithm of the addends z_i in a log-log plot, for $i = 5$, $i = 10$,$i = 20$,$i = 50$, and corresponding parabolic fits. We can see that already for $i = 10$ the distribution is close to a log-normal, since a parabola is a good fit for the distribution.

towards log-normals. As an example we show in Fig. 3.21 the distributions of z_i at different i for these particular parameters, but the same is valid for all divergent cases (also for the power-law decaying ones). We can see that the parabolic fit, representing a log-normal distribution, is a good fit already starting from $i = 10$, and that the parameters μ_i and β_i^2 diverge as expected, the distribution moving to the right and broadening for large i (for a remind of the definition of μ_i and β_i^2 see Section 3.3). Even though the variance and/or the average diverge we are therefore dealing we the same underline problem: a sum of correlated log-normally distributed random variables.

Let us now try to have an analytical understanding of the problem. As a basic consideration if we remind Eq. (3.69) we can see that the process can be approximated by a simple multiplicative process:

$$Z_N = a_1 Z_{N-1}, \tag{3.94}$$

since Z_1 in the RHS is $\gg 1$ and Z_1 can be written as Z_{N-1}, since it follows the same distribution as Z if the process had stopped at time $N - 1$. For the same reason we wrote Z as Z_N, since the mode of $p(Z)$ depends here on N. Hence going further with the recurrence we can see the process as a Gibrat's law and taking the logarithm from both sides of Eq. (3.94) will result in a sum of $\ln(a_i)$ which are independent and with finite average and variance. The log-normal distribution is then justified by the Central Limit Theorem. However we can go deeper, we can do the hypothesis, verified numerically, that if we write:

$$Z = a_1(1 + Z_1), \tag{3.95}$$

as in (3.69), but with the small difference that Z_1 does not follow the same distribution as Z, like in the convergent case, but a modified version of it, which corresponds to a shift

to the left on the distribution of $t = \ln(Z)$: $p(t + \epsilon)$. The amount of the shift depends on $\tilde{a}$ and $\bar{a}$ for a difference of one unit in N and can be computed for given parameters. If the distribution did not «move» with N, Eqs. (3.74) and (3.82) for the tail decay would still be valid, but this we already know that it is not true, then $p(Z)$ has to move. Taking the logarithm of Eq. (3.95), since we are interested in proving that the distribution of $\ln(Z)$ is a Gaussian, yields:

$$\ln(Z) = \ln(a_1) + \ln(1 + Z_1) \simeq \ln(a_1) + \ln(Z_1) \qquad \text{for} \qquad Z_1 \gg 1. \tag{3.96}$$

We can thereby compute the Fourier convolution of the distributions to find the distribution of the sum of the two independent variables $\ln(a_1)$ and $\ln(Z_1)$. Calling $t = \ln(Z)$, we have:

$$p(t) = \int_{-\infty}^{+\infty} p(y + \epsilon) \frac{[\theta\,(t - y - \ln(b)) - \theta\,(t - y - \ln(c))]}{c - b} e^{t-y} dy, \tag{3.97}$$

where $\theta(\cdot)$ is the Heaviside step function and the exponential factor e^{t-y} comes from the change of variable from the uniformly distributed variable a to $\ln(a)$. We are going to solve the integral by a new change of variable: $v = e^{-y}$, $dv = -e^{-y}dy$, leading to:

$$p(t) = -e^t \int_0^{+\infty} p(-\ln(v) + \epsilon) \frac{[\theta\,(t + \ln(v) - \ln(b)) - \theta\,(t + \ln(v) - \ln(c))]}{c - b} dv, \tag{3.98}$$

Using now the properties of the Heaviside function we can write:

$$p(t) = e^t \int_{be^{-t}}^{ce^{-t}} \frac{p(-\ln(v) + \epsilon)}{c - b} dv, \tag{3.99}$$

and finally deriving with respect to t, using Leibniz differentiation rule, we get a delayed differential equation similar to (3.75):

$$p'(t) = p(t - \ln(b) + \epsilon) \frac{b}{c - b} - p(t - \ln(c) + \epsilon) \frac{c}{c - b} + p(t) \qquad \text{with} \qquad t = \ln(Z) \tag{3.100}$$

We tried to solve this delayed differential equation with Wolfram Mathematica v11.3, but finding a general solution was not possible. However we could inject in the equation the distribution found with simulations, a log-normal with the estimated parameters μ_Z and σ_Z (respectively 47.4 and 2.32 for the case under study, where since the distribution is a very good log-normal the estimates by maximum likelihood and by a quadratic fit coincides within error bounds). In order to compute ϵ we use Eqs. (3.67) and (3.68), injecting in them the expression for the variance s_Z^2 (Eq. (3.55)) and m_Z (Eq. (3.54)), obtaining $\epsilon = 0.0898$. In Fig. 3.22 we show the LHS and the RHS of equation (3.100) where we introduced as $p(t)$ a Gaussian distribution, since we want to check if Z is log-normally distributed. Looking at the perfect overlap of the 2 plots, we can conclude that $p(t)$ equal to a Gaussian distribution is a solution of Eq. (3.100) and therefore Z is truly log normally distributed.

Moreover we can notice that in this region of the phase space the simulated and the

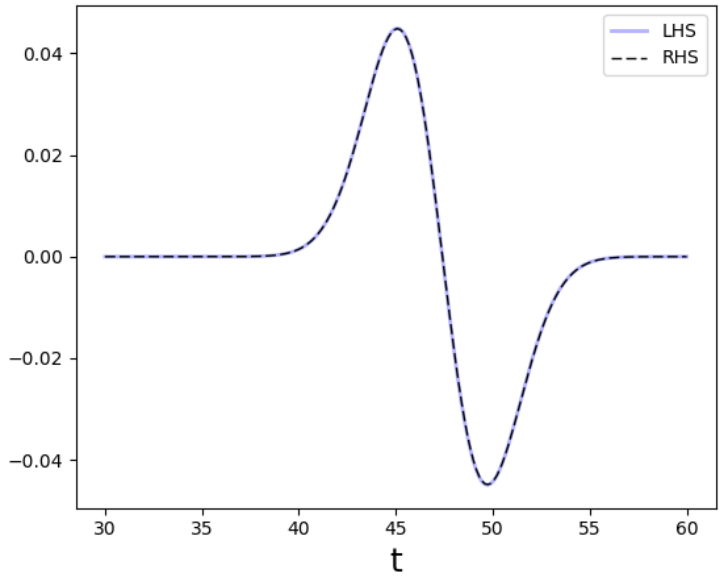

Figure 3.22 – Right hand side of Eq. (3.100) (dashed black line) with $p(t)$ chosen as a Gaussian distribution and left hand side of the same equation (blue full line), t being $\ln(Z)$. The curves overlap very well.

theoretical values of m_Z and s_Z *versus* N show a very good agreement, indicating that there is no longer the numerical effect there was in other divergent cases, because Z diverges very fast. As we can see from Fig. 3.23 the maximum error between the curves is of 0.2% the the $\ln(m_Z)$ and of 0.4% for $\ln(s_Z^2)$. Both growing linearly in the plots, hence exponentially with N, for large N. Therefore, the asymptotic growth at large N is governed by the largest exponential terms in Equations (3.54) and (3.55) which are respectively:

$$m_Z \simeq \bar{a}^{N+1} / (\bar{a} - 1) , \tag{3.101}$$

and

$$s_Z^2 \simeq \frac{2\bar{a}\bar{a}^{N+1}\tilde{a}^N}{(1-\bar{a})(\bar{a}-\tilde{a})\bar{a}^N}. \tag{3.102}$$

Injecting these equations in (3.67) and (3.68) we can have μ_Z and σ_Z^2 with respect to the phase space parameters $\bar{a}$ and $\tilde{a}$:

$$\mu_Z \simeq \ln\left(\bar{a}^{N+1}/(\bar{a}-1)\right) - \frac{1}{2}\ln\left(\frac{2\tilde{a}^N(\bar{a}-1)}{(\tilde{a}-\bar{a})\bar{a}^{2N}} + 1\right) \tag{3.103}$$

$$\sigma_Z^2 \simeq \ln\left(\frac{2\tilde{a}^N(\bar{a}-1)}{(\tilde{a}-\bar{a})\bar{a}^{2N}} + 1\right). \tag{3.104}$$

These relations, first allow us to compare the values of μ_Z and σ_Z^2 to the maximum likelihood estimated ones (which estimates are independent from Eqs. (3.103) and (3.104) and underlying hypothesis), and if the values from Eqs. (3.103) and (3.104) coincide with the maximum likelihood estimates, within the error bounds, it means that Z converged

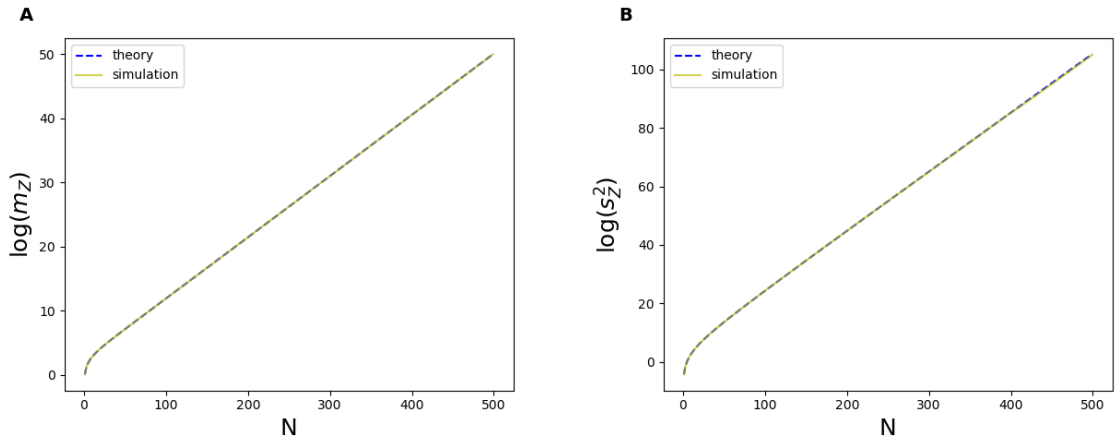

Figure 3.23 – A. Logarithm of the theoretical average value given by Formula (3.54) in dashed blue and of the average value of Z obtained from the simulation (yellow) with the same parameters as in Fig. 3.20 versus time step N. The two curves overlap almost perfectly, with a maximum error of 0.2%. **B.** Theoretical variance of Z given by Formula (3.55) in dashed blue and the simulation (in yellow) result with the same parameters as before, both in log scale, versus time step N. The overlap is very good with a maximum error of 0.5%.

to a log-normal distribution, since the equations are true under the hypothesis that Z is log-normal. Secondly, they allow us to conclude that in the large N limit μ_Z and σ_Z^2 have a linear dependence with N and to define the rate of convergence with respect to parameters $\bar{a}, \tilde{a}$. Hence, we can conclude that the standardized variable

$$\frac{(\ln(Z) - \mu_Z)}{\sigma_Z} \sim \mathcal{N}(0,1)\,,$$

in the large N limit.

Finally, the log-normal distribution for Z, which appears immediately on the right of the critical line of the $(\bar{a},\tilde{a})$ plane and for $\bar{a} > \bar{a}^\dagger$ (see diagram in Fig. 3.2), and all the distributions showed in all divergent cases (power-laws), are essentially due to correlations, without which the distribution of Z would be completely different even at finite N, and eventually converging for $N \to \infty$ to a Gaussian as we will prove in Section 3.7.

3.7 The role of correlations: about the surrogate model without correlations

For better understanding why log-normal distributions of a sum of random variables may occur, we could wonder if this also happens without correlations. The question behind this is to put in evidence the role of correlations in the observed distributions. In order to accomplish this, we are going to keep the same (approximated) statistics for the single random variables z_i, that we characterized for the minimal model (see Section 3.3), *i.e.* we will take log-normal random variables with parameters like in Eqs. (3.30) and (3.33), but removing correlations. Let us observe Figure 3.24. We can see that the resulting distribution for Z is still a good log-normal, at least in the right plot (R^2 of the fit 0.993),

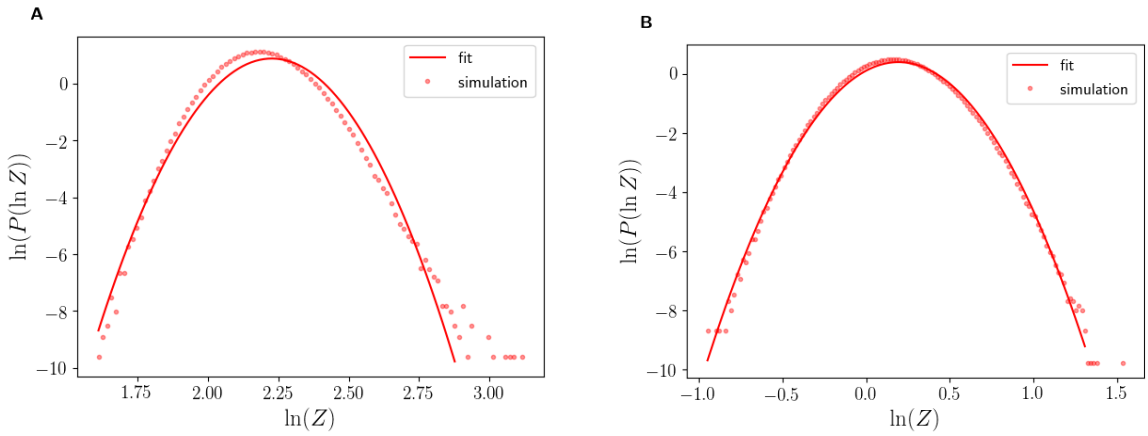

Figure 3.24 – Distributions of $\ln(Z)$ in a log-log plot and corresponding parabolic fit. On the left: the distribution of $\ln(Z)$ for the surrogate model where $a_j \sim \mathcal{N}(\bar{a} = 0.9, \Delta a^2 = 0.04)$, following (3.30) and (3.33) with parameters $\bar{a} = 0.9, \Delta a^2 = 0.04$, without correlations. The parabolic fit is taken excluding the last points on the right for which the statistics is not enough to determine the shape. On the right: the same for the surrogate model with $a_j \sim \mathcal{U}_{[0.2;0.9]}$, then following (3.30) and (3.33) with parameters $\bar{a} = 0.55, \Delta a^2 \simeq 0.04$, also without correlations. Remark that here the number of repetitions for the statistics is 10^6. The time steps are again $N = 1000$, that is largely sufficient fot the z_i to converge to zero.

where a is a uniform variable in the interval $[0.2; 0.9]$. In the left plot we start seeing a power law tail on the right due to the divergent moments of the variables z_i. We can compare these plots to the first line of Fig. 3.7 seeing that here the distributions are already log-normal even for such small parameters, supporting the idea that in Fig. 3.7 N^* was too small to get the z_i converged to log-normals. On the other hand here the z_i are log-normal since the beginning, by definition.

We are now going to try to understand this. We know that the log-normal distribution is typically related to multiplicative processes, because if we multiply random variables with finite averages and variances and then we take the log of the product we can apply the Central Limit Theorem to the sum of the logarithms of the variables. This gives a normal distribution for the log of the product, then the distribution of the product will approach asymptotically to a log-normal. This works only if the summed variables are not correlated (cf. (Feller, 1971) (Mitzenmacher, 2004)).

Following this idea we could think that our sum Z, can be in some way written as a product of random variables.

$$Z = \sum_{j=1}^{N} z_j = \sum_{j=1}^{N} e^{\left(\frac{\Delta a^2}{2\bar{a}^2} - b^2\right)j}(1 + \epsilon_j), \tag{3.105}$$

where b^2 is defined as in (3.41) and ϵ_j are the fluctuations from the mean, considered as small. This means that we are neglecting large deviations of ϵ_j from the mean $\langle \epsilon_j \rangle = 0$. This is true if $s_{z_i}^2 \ll m_i$, and all the central moments do not diverge. This is generally verified in the cases we studied, for example in the cases in Figure 3.24, even though for the uniform case ($a_j \sim \mathcal{U}_{[0.2;0.9]}$), the ratio $s_{z_i}^2/m_i$ is much smaller than in the Gaussian case. This may be the reason why the log-normal is a better approximation for the sum

Z in the second case. About the second requirement of finite central moments we can notice that the n-th *row* moments diverge with i as $n > n^*$ for some n^*, since they are given by formula $\mathbb{E}[z_i^n] = e^{ni\left(\ln\bar{a} - \frac{\Delta a^2}{2\bar{a}^2} + \frac{n\Delta a^2}{2\bar{a}^2}\right)}$. Even though, the *central* moments are all finite as we will see in the following by (3.117) (and related reasoning), for $\bar{a} < 1$. In any case, the bigger n the smaller is the effect of the n-th moment on the tails. This effect becomes soon negligible and unobservable in real life, unless you have a huge amount of data. With these assumptions we can think to write (3.105) as a product:

$$Z = \sum_{j=1}^{N} \bar{a}^j(1+\epsilon_j) = \bar{a}\,(1+\epsilon_1) + \bar{a}^2\,(1+\epsilon_2) + \cdots + \bar{a}^N\,(1+\epsilon_N) = \frac{\bar{a}}{1-\bar{a}} + \sum_{j=1}^{N} \bar{a}^j\epsilon_j \quad (3.106)$$

Then we can factorise:

$$Z = \frac{\bar{a}}{1-\bar{a}}\left(1 + \sum_{j=1}^{N} \bar{a}^j\epsilon_j \frac{1-\bar{a}}{\bar{a}}\right) \simeq \frac{\bar{a}}{1-\bar{a}} \prod_{j=1}^{N}\left(1 + \bar{a}^j\epsilon_j \frac{1-\bar{a}}{\bar{a}}\right) \quad (3.107)$$

Now we can take the logarithm and get:

$$\ln(Z) \simeq \ln\left(\frac{1-\bar{a}}{\bar{a}}\right) + \sum_{j=1}^{N} \ln\left(1 + \bar{a}^j\epsilon_j \frac{1-\bar{a}}{\bar{a}}\right) \quad (3.108)$$

This shows that Z is close to a log-normal random variable, since all the statistics of z_j is contained in ϵ_j and since the z_j are log-normal by construction the random variables inside the sum sign are Gaussian. It is known that the sum of Gaussian random variables is still Gaussian, so $\ln(Z)$ has to be close to a Gaussian, meaning Z log-normal. This natural idea of writing a sum as a product is actually similar to the one stated in a paper by Mouri (Mouri, 2013), which comes to supports our independently achieved reasoning.

We could wonder why the sum is not Gaussian, as the central limit theorem would predict. In fact there is a version of the central limit theorem for non identically distributed independent random variables, which is the case for our problem. This is a version stated by Lyapunov (there is a similar one by Lindeberg also, see for instance (Feller, 1971)), in this version the limiting distribution for the average of the z_i is a Gaussian if the moments of the distributions of the z_i are not too much different each other, so that the increments are not too large. More precisely, if $n \to \infty$:

$$\frac{\sum_{i=1}^{n}(z_i - m_i)}{s_n} \to \mathcal{N}(0,1) \qquad \text{with} \qquad s_n^2 = \sum_{j=1}^{n} s_{z_j}^2 \quad (3.109)$$

if for some integer $\delta > 0$ the following condition is verified:

$$\lim_{n\to\infty} \frac{1}{s_n^{2+\delta}} \sum_{i=1}^{n} \mathbb{E}[|z_i - m_i|^{2+\delta}] = 0 \quad (3.110)$$

It turns out that this condition can't be verified for our z_i. We are going to show this in the following.

We know the n-th row moment of a log-normal variable (Johnson et al., 1994), if we

apply this to our z_i, using Formulas (3.30) and (3.33) we get:

$$\mathbb{E}[z_i^n] = e^{ni\left(\ln \bar{a} - \frac{\Delta a^2}{2\bar{a}^2} + \frac{n\Delta a^2}{2\bar{a}^2}\right)}. \tag{3.111}$$

Now we have to do some reasoning to apply condition (3.110). We also know the central moments of the log-normal distribution, that actually is just an application of the binomial formula to (3.111):

$$\begin{aligned} \mathbb{E}[(z_i - m_i)^{2+\delta}] &= \sum_{j=0}^{2+\delta} \binom{2+\delta}{j} \mathbb{E}[z_i^j](-m_i)^{2+\delta-j} = \\ &= e^{(2+\delta)\mu_i + \frac{1}{2}(2+\delta)\beta_i^2} \sum_{j=0}^{2+\delta} (-1)^{2+\delta-j} \binom{2+\delta}{j} e^{\frac{1}{2}j\beta_i^2(j-1)} \end{aligned} \tag{3.112}$$

Now we can notice that the sum over j is an oscillating series which terms that grow exponentially with j are multiplied by a binomial coefficient. This makes the sum always positive, even for odd moments. This is not surprising, because the log-normal distribution is skewed to the right, so all the central moments have to be positive. Let us anyway point out what happens to the series. First, if $2+\delta$ is odd the number of terms of the series is even and the first term of the series is negative. Because of the property of the binomial coefficient,

$$\binom{n}{n-k} = \binom{n}{k}$$

every binomial coefficient comes twice in the series, once negative and once positive, at a symmetric position in the series (the first and the last, the second and the one before the last ...). But, because of the exponentially increasing term, the sum (with sign) between the 2 most external terms (0 and n) is bigger than the difference of the terms immediately inside (1,$n-1$) and so on, and since the sum (with sign) of the most external terms is always positive the total sum will be positive. Secondly, if $2+\delta$ is even the number of terms is odd and then there will be one positive term more then the negative terms. So because of the same property of the binomial coefficient and noticing that the first term of the series is positive, the sum taking out the exponential term would be always $= 0$. Because of the exponentially increasing term the series will always be greater than 0, if $\beta_i^2 > 0$.

What matters for the Lyapunov condition (3.110), is the absolute value of this series, then we can write:

$$\mathbb{E}[|z_i - m_i|^{2+\delta}] \geq \mathbb{E}[(z_i - m_i)^{2+\delta}] \geq e^{(2+\delta)\mu_i + \frac{1}{2}(2+\delta)\beta_i^2}\epsilon \tag{3.113}$$

where $\epsilon > 0$ is the lower bound of each term of the series in (3.112). Notice that ϵ would depend on i, but the sum grows as far as β_i increases, since we remember that $\beta_i = i\Delta a^2/\bar{a}^2$, and then taking the smallest β_i ($\beta = \Delta a^2/\bar{a}^2$) the dependence on i disappears and the last inequality is true. Now let us check the sum of variances s_n, in order to apply Lyapunov

condition.

$$s_n^{2+\delta} = \left(\sum_{j=1}^{n} s_{z_j}^2\right)^{\frac{2+\delta}{2}} = \left(\frac{e^{\frac{\Delta a^2}{\bar{a}^2}+2\ln\bar{a}}}{1-e^{\frac{\Delta a^2}{\bar{a}^2}+2\ln\bar{a}}} - \frac{e^{2\ln\bar{a}}}{1-e^{2\ln\bar{a}}}\right)^{\frac{2+\delta}{2}}, \tag{3.114}$$

where we injected Equation (3.46) that in this case is not an approximation, since we imposed the variables z_j to follow a log-normal distribution. This is true in the region of the parameters space where $\frac{\Delta a^2}{\bar{a}^2} + 2\ln\bar{a} < 0$ and $\bar{a} < 1$, where both s_n^2 and the sum $\sum_{i=1}^{n} m_i$ converge as far as $n \to \infty$. The divergent case will be studied later on. Then we can see that in this case $s_n^{2+\delta}$ does not depend on n, giving as result a $s_n^{2+\delta}$ that is a finite number different from 0 for any δ.

Let us now check what the sum in Lyapunov condition (3.110) does. So, injecting the lower bound found in (3.113) this becomes:

$$\sum_{i=1}^{n} \mathbb{E}[|z_i - m_i|^{2+\delta}] \geq \sum_{i=1}^{n} e^{(2+\delta)\mu_i + \frac{1}{2}(2+\delta)\beta_i^2}\epsilon = \epsilon\frac{\bar{a}^{2+\delta}}{1-\bar{a}^{2+\delta}}, \tag{3.115}$$

the series in the middle, in the same hypothesis of convergence as before, also converges and does not depend on n. Since this is a lower bound and the denominator does not go to infinity, the limit as $n \to \infty$ in (3.110) can't go to zero.

This means that Lyapunov condition is not valid for this exponentially variating m_i and s_{z_i}, because the changes are too fast. This is also coherent with what we observe in simulations and what we discussed before, *i.e.* the observation of log-normal, instead of Gaussian, distributed sums even without correlations.

Let us now consider the case when $\sum_{i=1}^{n} m_i$ diverges. We note that if $\sum_{i=1}^{n} m_i$ diverges for sure also s_n^2 diverges and also $\mathbb{E}[z_i^n]$ diverges for $n > 3$. In principle there can be also a case where $\sum_{i=1}^{n} m_i$ converges and where s_n^2 diverges, but this can be seen as a particular case of the previous one, as we will see.

For this purpose we are going to remind here an inequality from probability theory (Feller, 1971). For random variables X and Y, Minkowski inequality states:

$$\mathbb{E}[|X+Y|^r] \leq \left[\mathbb{E}[|X|^r]^{1/r} + \mathbb{E}[|Y|^r]^{1/r}\right]^r \qquad \text{for} \qquad r \geq 1 \tag{3.116}$$

We this in mind we can put an upper bound to the sum in (3.110). So if we take as $Y = -m_i$, since $\mathbb{E}[|-m_i|^r]^{1/r} = m_i$, and $m_i = \bar{a}^i$, we can write:

$$\sum_{i=1}^{n} \mathbb{E}[|z_i - m_i|^{2+\delta}] \leq \sum_{i=1}^{n} \left(e^{i\left(\ln\bar{a} - (1+\delta)\frac{\Delta a^2}{2\bar{a}^2}\right)} + \bar{a}^i\right)^{2+\delta} \tag{3.117}$$

Now we notice that if $\bar{a}$ is bigger than one the exponential term $e^{\left(\ln\bar{a} - (1+\delta)\frac{\Delta a^2}{2\bar{a}^2}\right)}$ in the sum is always smaller than $\bar{a}$. Therefore the term that will dominate the series as $n \to \infty$ is $\bar{a}^i$ and the sum on the right then goes as $\bar{a}^{n(2+\delta)}$. We now have to compare this to the denominator of (3.110). The sum s_n^2 for large n goes as:

$$s_n^2 \simeq \left(e^{n\frac{\Delta a^2}{\bar{a}^2}} - 1\right) e^{n2\ln\bar{a}} \simeq e^{n\left(\frac{\Delta a^2}{\bar{a}^2} + 2\ln\bar{a}\right)}, \tag{3.118}$$

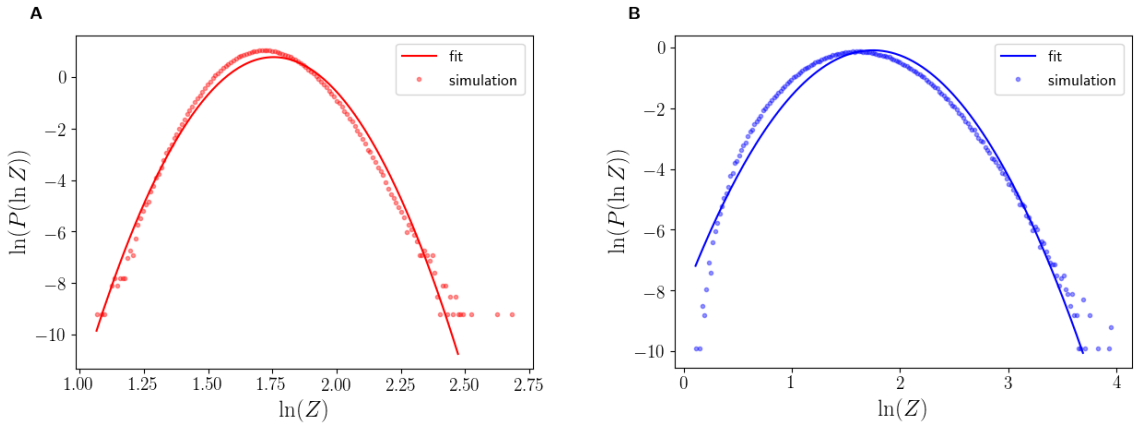

Figure 3.25 – Log-log plots of the distribution of $\log(Z)$, in both plots the number of repetitions is 10^6 and N is large enough for convergence ($N = 1000$). **A.** surrogate model without correlations with $a_j \sim \mathcal{U}_{[0.5;1.2]}$, parameters $\bar{a} = 0.85, \Delta a^2 \simeq 0.04$. **B.** correlated model with the same distribution for a_j. Notice that the distribution for the correlated model is much larger than the one without correlations, even though the average is approximately the same.

and the denominator of Lyapunov condition then goes as:

$$s^{2+\delta} \simeq e^{\frac{2+\delta}{2} n \left(\frac{\Delta a^2}{\bar{a}^2} + 2 \ln \bar{a} \right)}. \tag{3.119}$$

We notice that the denominator of (3.110) in the large n limit will always be bigger than the numerator, giving then, as result of the limit $n \to \infty$, 0 as required by the condition (remember that $\delta > 0$):

$$\bar{a}^{2+\delta} < e^{\frac{2+\delta}{2} \left(\frac{\Delta a^2}{\bar{a}^2} + 2 \ln \bar{a} \right)}, \tag{3.120}$$

and considering that $\lim_{n\to\infty} \bar{a}^{n(2+\delta)} \geq \lim_{n\to\infty} \sum_{i=1}^{n} \mathbb{E}[|z_i - m_i|^{2+\delta}]$ the limit has to be zero and Lyapunov condition (3.110) is verified. Let us point out here that if $\bar{a} \leq 1$, so in the case where $\sum_{i=1}^{n} m_i$ converges but s_n^2 diverges, the Lyapunov condition is also verified. Indeed, the right member of (3.117) is, in this case, a convergent series, since all the terms are geometric series with a common ratio lower than 1. The series will then converge to some finite number we call η. Still, the inequality $\eta \geq \sum_{i=1}^{n} \mathbb{E}[|z_i - m_i|^{2+\delta}]$ is valid and then the limit of (3.110) will again be zero, since the denominator is divergent in the same way as in (3.120). In this sense this is a particular case of the case treated just above. The results of all this discussion are summarised in Fig. 3.3.

Finally, we remark that Equation (3.117) if we take the case $\bar{a} < 1$ proves also, as we said previously, that all the central moments are finite, for every i. The series on the right in Equation (3.117) is convergent in this case and then the central moments are finite (the absolute value does not impact this conclusion). This finally justifies completely the reasoning at the beginning.

In conclusion we proved that the sum of uncorrelated random variables with the statistics defined in (3.30) and (3.33) can converge to something different than a Gaussian variable when both s_n^2 and $\sum_{i=1}^{n} m_i$ converges. In particular this opens the possibility to

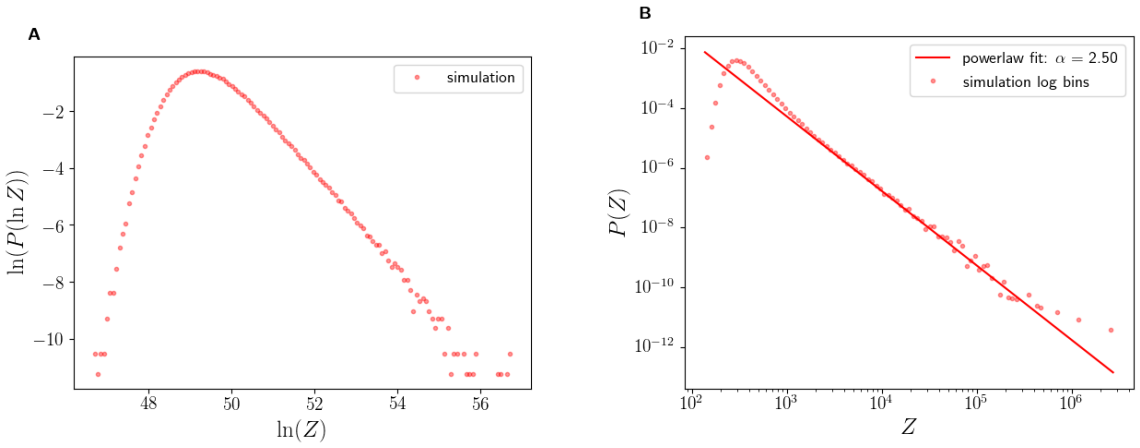

Figure 3.26 – A. Distribution of $\ln(Z)$ in a log-log plot, for the surrogate model without correlations with $a_j \sim \mathcal{U}_{[0.9;1.3]}$, parameters $\bar{a} = 1.1, \tilde{a} \simeq 1.22$. **B.** Log-log plots of the distribution of Z, for the surrogate model without correlations with $a_j \sim \mathcal{U}_{[0.7;1.3]}$, parameters $\bar{a} = 1, \tilde{a} \simeq 1.03$. And better power-law fit. In both plots the number of repetitions is 10^6 and $N = 500$.

have log-normal (or close to log-normal) distributions for the sum of log-normal variables, as observed in simulations. The observation of log-normal distributions shows up also in the stochastic model with correlations, in the same region of parameters (convergent region). The only difference between the two cases, as we can see in the comparison in Fig. 3.25, is that in presence of correlations the variance of the sum is significantly bigger. This can be understood by the general fact (Feller, 1971) that the variance of the sum of 2 correlated variables is the sum of the two variances plus a term involving correlation, which, as we showed in Section 3.4, is positive.

In addition we show in Fig. 3.26 the resulting distribution for the surrogate model where all parameters are the same as in Figs. 3.16 and 3.20. We can see that the distributions are very different for the correlated case, even though we cannot appreciate the convergence toward a Gaussian distribution expected from the reasoning of this section. However the convergence, even for the sum of identically distributed log-normal random variables is known to be very low (see for example (Da Costa et al., 2000)), therefore for moving distributions is probably even slower.

To conclude, all this explains the difference between what is observed for the stochastic model and what is observed here in the divergent region. The central limit theorem, in its generalized version, is valid in the case of the absence of correlations, but for the stochastic model this is not true any more, giving rise in the first case to distributions that have to converge to Gaussian, even though the number of steps is not sufficient to see it (because the single variables z_i increase too fast and they «saturate» numerically), and in the second case to power-law or log-normal distributions.

3.8 Discussion

In this chapter we have shown that the sum of values resulting from a multiplicative cascade (branching) process can give rise to different types of distributions and very rich

behaviours, which depend on both parameters of the *reproduction rate* (also called *growth factor* or *branching ratio*) distribution (the average $\bar{a}$ and the second central moment $\tilde{a}$). Detailed results are obtained for a uniformly distributed reproduction rate a, but the phenomenology of the results is the same as long as the distribution has a compact support (for example excluding heavy tailed distributions). More precisely all results are expected the same if the branching ratio distribution belongs at least to the symmetric exponentially bounded class of distribution (but the symmetry condition can be released if the distribution has compact support). Actually as long as the error (3.26) is small the main picture should be the same. In fact, the same behaviour described hereafter was observed in simulations considering a normally distributed reproduction rate.

In a first region, where both the average and the variance of Z converge, we get finite scale distributions, whose moments are all finite. These distributions can be very close to log-normals depending on how fast the decrease of the average of Z is and therefore on how many variables are important for the sum (after a certain N^* all the summed z_i become negligible). For very low values of N^* the distribution moves away from a log-normal showing two bumps separating the probability to have normal size avalanches and short ones, which have both comparable probabilities. The position of this region in the phase diagram is similar when removing correlations, essentially preserving the shape of the resulting distributions, with the difference that they are significantly shrunk. The main effect of correlations is thus to importantly broaden the distributions.

In the second region we observe a power-law tailed distributions, which look very close to Lévy $\alpha-$stable distributions with a Paretian tail. The exponent of the tails varies from 3 to 1, by increasing $\tilde{a}$ and $\bar{a}$ (see Fig. 3.2). For $\tilde{a} = 1$ the exponent is equal to 3 and does not depend on the value of $\bar{a}$, on the other hand, for $\bar{a} = 1$ the exponent is equal to 2 and does not depend on the value of $\tilde{a}$. This can be the reason of the variation of the exponent α in (Beggs and Plenz, 2003) with respect to the time interval under consideration (see Fig. 3**A** in the article and Figure 3.1 of Section 3.1). Indeed, varying the time interval can be thought of as aggregating together several steps of the avalanche which results in a change in the growth factor, *i.e.* $\bar{a}$ and $\tilde{a}$.

The third region showed a very solid statistical behaviour for Z, whose distribution converges to an exact log-normal distribution. The velocity of convergence depends on the choice of the particular parameters (the bigger $\bar{a}$ and $\tilde{a}$, the faster the convergence), but the log-normal may be appreciated even in experimentally reasonable time lengths (N) of the process. Nevertheless the distribution here is not stationary, since both parameters of the log-normal depend on N. However, with a standardisation of the variable this dependence disappears and the asymptotic distribution is well defined.

A particular case is when the uniform distribution from which the variable a is drawn has the lower bound $b = 0$ (the upper bound of Fig. 3.2). In this case the complete distribution of Z can be analytically computed and gives as result a decay as Z^{-1} for finite N. Since this distribution in the $N \to \infty$ limit is not normalisable, the asymptotic distribution will therefore converge to a flat distribution if Z is not bounded, but for any finite N it will have the same well defined analytical shape. Another, opposite, particular case is the lower bound of Fig. 3.2, where the distribution will be a Dirac δ function, since all summands z_n will have zero variance.

Correlations are important in all the regions, but they are even crucial in the second and third region and cannot be neglected. Indeed in the last two regions by removing them we obtain strongly different behaviours, proving that the Central Limit Theorem cannot be applied. That is even considering Lyapunov version (for not too different distributions of the summands) or the version regarding weakly correlated random variables. Notice that, by adding correlations, it is possible that the sum of log-normal random variables (the z_i) results in a log-normal random variable, and therefore the log-normal distribution can be as well a signature of scale invariance of a system.

The randomness of the reproduction factor actually resumes all the aleatory processes that can rise in reality and can be thought as an *effective* random reproduction rate. For example, for epidemics it would contain in itself the randomness of the contacts of an individual and the randomness of transmitting the infection. Remarkably, for the avalanche size to be bounded, not only the average reproduction rate $\bar{a}$ has to be lower than one, but also its second moment, $\tilde{a}$. Again with the example of the epidemics, this two conditions would be necessary for avoiding the epidemic to spread massively.

Notably, if the accessible physical quantity is Z, by plotting its distribution we can have some information about the often inaccessible parameters of the underlying avalanche, like the growth factor $\bar{a}$ and $\tilde{a}$. Moreover a value of the tail exponent of the Z distribution is uniquely related to a line in the parameter plane, whose coordinates can be analytically computed. Therefore if we have power-law distributed experimental avalanche data, by measuring with a suitable statistical technique the exponent of the tail, the parameters of the corresponding branching ratio distribution can be determined.

Another important result is that our model is a possible explanation of some of the power-law distributions arising in nature and human systems. Most of their exponents (see for a list (Newman, 2005)) are in the range of the possible outcome exponents of the model, and many of them can actually be interpreted as the result of a sum of an underlying multiplicative process. In addition, this connects the power-law measured exponent to the growth factor of the multiplicative process, giving an immediate reason for a change in the exponent. Preliminary results of this model show also that allowing the growth factor a to have negative values (which can make sense in some cases, like stock pricing), can explain observed distributions that have two different power-law regimes with different exponents, like proposed for USA family names and web hits distributions in (Newman, 2005), and could reproduce well also the abundance of species of birds distribution shown in the same reference.

We point out that this avalanche process can also be used as a guide for building a network with a degree distribution as one of the possible avalanche size distributions resulting from this study. Indeed, starting from a network of $\mathfrak{N}$ nodes, at each node we could associate an avalanche process as the one described here (actually a discrete version of it, the number of connections being an integer number). Hence, by setting the suitable parameters needed for the desired network degree distribution, we could attribute to each node a number of connections equal to the result of the corresponding avalanche process. After all, the growth of a network can be seen itself as an avalanche process, thinking for example of the preferential attachment rule (Albert et al., 1999). This model can therefore be proposed as an interpretation for some evidence of not scale-free networks

(Broido and Clauset, 2019).

Finally, coming back to the original motivation of this study, this analysis suggests that the reason why we have seen close to log-normal distributions for the size of the avalanches for the model in Chapter 2 and (Polizzi et al., 2018), is the short duration of the avalanches. Actually the observed distributions were not exact log-normals and therefore could not be interpreted as outcome of the third region, even though deviations from log-normality could also be due to other reasons. Nevertheless this was consistent with the large width of the observed distributions and can be interpreted in an effective average reproduction rate lower than 1.

Chapter 4

Emergence of log-normal type distributions in avalanche processes on networks

4.1 Avalanches on a network structure: is self-organised criticality exhaustive?

Avalanche processes are very common in living systems, such as firing rates in brain (Beggs and Plenz, 2003), fractures in living cells cytoskeleton (CSK) (as we have seen in Chapter 2), but also in amorphous and random media (Salje et al., 2017), or earthquakes. Actually, all the systems that can be considered as composed by elementary threshold units, which are connected and can transfer information to each other, are subject to avalanches. Indeed, as soon as a unit reaches its threshold it can cause other units to do the same in turn. In this chapter we focus in particular on avalanches on networked systems.

In statistical physics, avalanches are often treated in the framework of critical systems. This is justified by the observation of critical behaviour of avalanche statistics, with distributions usually approximated by power-laws, at least for some length or energy scales. Power-laws are reminiscent of self-organised criticality (Bak, 2013), and are ubiquitous for avalanches in solids and amorphous materials (Song et al., 2013; Chuang et al., 2013; Perez-Reche, 2017), describing for example the avalanche sizes of a granular material falling apart. Self-organised criticality has also been associated to neural dynamics (first in (Lo et al., 2002) but also in (Lo et al., 2004; Klaus et al., 2011; Dvir et al., 2018)) showing widely spread power-law statistics with a typical exponent looking constant among different species (Dvir et al., 2018) and in different environmental situations. In all these examples, fluctuations, and thus their distribution, play a major role in triggering the system to the critical state. It is worth to mention also the theory of marginal stability, which considers avalanches as arising from marginally stable states in which the system is stuck because of their very slow glassy dynamics, generating also power-law avalanches (Müller and Wyart, 2015). Recently, some doubts about the ubiquity of power-law distributions have been advanced (Broido and Clauset, 2019), pointing out that

many real system's data are actually better fitted by other skewed, fat-tailed distributions, such as the log-normal. This is true for network degree distributions, in social, biological or technological networks (Broido and Clauset, 2019), but also for avalanche statistics, in neural networks (Ribeiro et al., 2010; Buzsáki and Mizuseki, 2014) or living cells cytoskeleton (Polizzi et al., 2018). Indeed, the growth of a network can be seen as an avalanche process in itself, thinking for example of the preferential attachment rule (Albert et al., 1999). In general, power-law distributions are difficult to irrefutably hold for any kind of real data. Remarkably, power-law and log-normal distributions are closely related to each other and small variations in the generative mechanisms can lead to one or the other as noticed by (Kesten, 1973) and as we have deeply seen in Chapter 3.

Nevertheless the power-law, with some adaptations, has been the largely dominant model in the last decades for all data showing a fat-tailed distribution, either for network theory or for avalanche processes. There are mainly two reasons for the large predominance of the power-law modelling up to now: i) because much is known about the modelling of power-laws, thanks to the statistical physics of phase transitions, leading to an ease of data treatment and interpretation. Concerning network science theory, the same is true for scale free networks (Barabási et al., 1999), justifying a power-law interpretation of real data distributions, while models for other fat tail distributions do not exist so far (Broido and Clauset, 2019; Barabási and Pósfai, 2016); ii) due to the difficulty to fit real data with fat-tailed distributions, because a power-law tail is only visible beyond a given threshold of the random variable, and not in the whole domain. Actually, any skewed distribution can be approximated by a power-law, even if only in a finite and small scale range (see Fig. 4.1**C**). In Figure 4.1, comparing the distributions of hippocampal firing rates in a linear scale (Fig. 4.1**A**), in a log-lin scale (Fig. 4.1**B**) and in a log-log scale (Fig. 4.1**C**), we realize that focusing on the tail of the distribution in a log-log representation increases the risk of missing some essential features of the underlying process (Buzsáki and Mizuseki, 2014).

The point ii) led to an ongoing debate about what is the most appropriate fitting distribution for the considered data. Some concepts such as *low-degree saturation* and *large-degree cutoff* have been developed for example in network-science theory, in order to account for important characteristics of real systems, such as their finite size (Barabási and Pósfai, 2016). This solves point ii), adapting the power-law distribution to be applied to real systems. On the other hand, point i) reveals the need of an alternative modelling framework which could validate the choice of other fat-tail distributions, in order to understand the underlying mechanisms leading to one distribution or the other. In the same way as degree distributions of networks, so far there are no models accounting for log-normal distributions of avalanches on networks, while there is strong evidence that log-normally distributed avalanches exist. In Figure 4.1, we show a few examples of processes related to an avalanche dynamics showing log-normal distributions: the distribution of the spine (small dendritic protrusion crucial in the transmission of electrical signals) sizes in neural networks (Loewenstein et al., 2011) (Fig. 4.1**D**); the distribution of new numbers of Ebola cases per week (Legrand et al., 2007) (Fig. 4.1**E**); the distribution of lags between 2 consecutive vehicles to cross the stop line at a crossroad (Jin et al., 2009) (Fig. 4.1**F**).

Beyond power-laws and log-normals, other skewed distributions are sometimes used

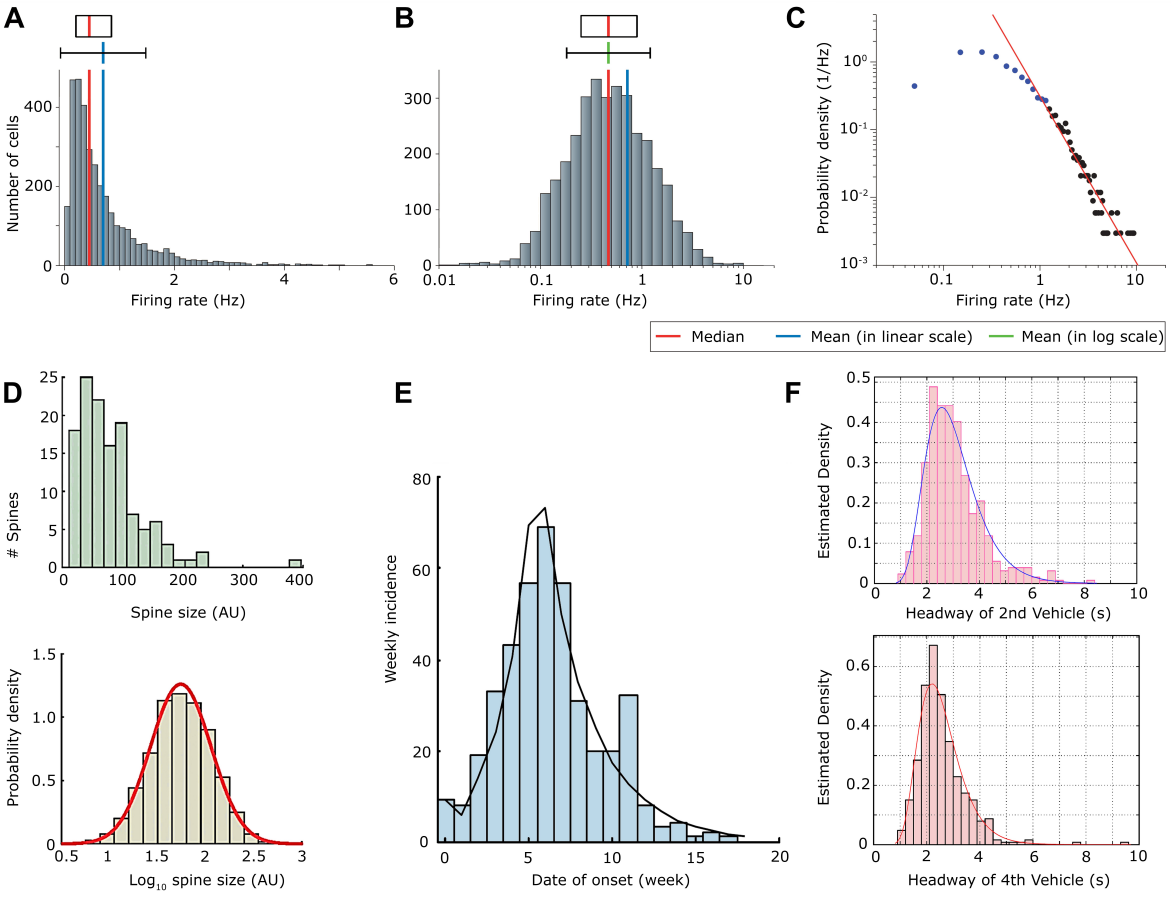

Figure 4.1 – **A.** Non-normalised distribution of firing rates of hippocampal CA1 pyramidal neurons during slow-wave sleep. The firing rate is measured in Hz. **B.** Distribution (non-normalised) of the logarithm of the firing rate as in **A**. **C.** Distribution of the firing rate as in **A** in a log-log plot. Notice that by taking into account only the tail of the distribution (black dots), the distribution could be fitted by a power-law, even though we may be missing some important information. **D.** Distribution of the spine size in arbitrary units (AU) (top), and probability density of the logarithm of the spine size (bottom). **E.** Distribution of the new number of Ebola cases per week (week incidence). **F.** Probability density of the lag times at which 2 vehicles waiting at rest at a traffic light pass the stop line. This is what is called headway. The distributions are reported for the lag between two successive vehicles considering the second position (top) and the fourth position (bottom) of the line: for instance the bottom plot describes the distribution of the difference of times at which the 3rd and the 4th vehicle of the queue cross the stop line. All times are given in seconds. **A**, **B** and **C** are adapted with permission from (Buzsáki and Mizuseki, 2014), **D** is adapted with permission from (Loewenstein et al., 2011), **E** is reproduced with permission from (Legrand et al., 2007) and **F** from (Jin et al., 2009).

to model biological data. For example, it is worth mentioning the Gamma distribution used to model inter-spike intervals of neurons (Gerstein and Mandelbrot, 1964) or the inter-beat interval variation of heartbeats (Ivanov et al., 1996, 1998). We should notice that statistically it is not always easy to distinguish log-normal from gamma distributions. In any case both are alternative models to the power-law framework.

In this chapter, with respect to previous models in Chapter 2 and 3, we increased the dimensionality and the complexity of our avalanche modelling, by introducing a network structure. Nevertheless the original motivation is still given by the experimental observation of log-normal statistics in avalanches of fractures of the CSK of living cells (Laperrousaz, Drillon, Berguiga, Nicolini, Audit, Satta, Arneodo and Argoul, 2016; Polizzi et al., 2018) analysed in Chapter 2. Our model, besides being inspired by our previous (mean field) models (see Chapter 2 and 3), was also inspired by works on epidemics spreading on networks (Newman et al., 2001; Newman, 2002). In our model the analogue of the epidemic population is represented by actin filaments, being in two possible states: cross-linked or not. The network structure models the CSK structure and the avalanche is a model for rupture mechanics in living cells, when plastic deformations are considered. Given that we do not consider here active cross-linking carried out from myosin filaments, the same framework can be applied to other cross-linked polymers with glassy dynamics. Therefore conclusions similar to the ones that will be described in this chapter can be applied to amorphous glassy materials, like polymers, metallic glasses or colloidal glasses, which all share slow dynamics, and long mechanical relaxation delays (Ritort and Sollich, 2003). This can be useful for understanding all processes not showing good power-law statistics, and in general what makes a distribution shifting from power-law to log-normal, highlighting the characteristics that a process needs to have to deviate from the most common power-law modelling.

Moreover our model will also provide an interpretation of that log-normal noise often observed in living systems, such as brain, which could be interpreted as an avalanche of events.

4.2 Evidence for log-normal statistics in cell mechanics

4.2.1 The fractional rheology of the cytoskeleton network

In this section we shortly resume some characteristics of the cytoskeleton mechanics useful for this particular chapter, some of them have been more thoroughly described in Section 2.1.

Among all the fascinating properties of living cells, we must emphasize here their ability to constantly remodel their structural organization to withstand forces and deformations and to promptly adapt to their mechanical environment (Hoffman and Crocker, 2009; Chauvière et al., 2010). This versatility is fundamentally required for many vital cellular functions, such as migration, mitosis, apoptosis or wound healing, and an alteration of the cell mechanical properties can participate in pathogenesis and disease progression,

such as cancer (Brunner et al., 2009; Kai et al., 2016).

We have already seen (see Section 2.1) that the CSK is composed by intertwined filaments of different type. Remember that for us the the most relevant for our modelling are the actin filaments, which cover the perinuclear zone, the cell compartment poked by our micro-indenting tips (see Section 2.2). As a reminder, they present a polar structure (Li and Gundersen, 2008), a quite fast, with respect to global active reorganization processes, polymerization rate (larger than one per minute) and a depolymerization rate a little slower, of the order of one per a few minutes (Tseng et al., 2005). Moreover they can build a cross-linked network whose mechanical properties depend in general on the cross-linker proteins density and on the network structure.

Actin filaments networks are designed by a wide variety of actin-binding protein cross-linkers, which can be passive or active, *i.e.* activated by ATP (Adenosine Triphosphate) hydrolysis, the latter having a slower dynamics (time scale of tens of minutes) (Gardel et al., 2008; Lieleg et al., 2010; Kollmannsberger and Fabry, 2011; Fletcher and Mullins, 2010; Huber et al., 2013; Blanchoin et al., 2014; Wagner et al., 2006; Ehrlicher et al., 2015). This cross-linked network gives to living cells some properties of soft-glassy materials (Fabry et al., 2001; Sollich, 1998), such as the weak power-law dependence of the shear relaxation modulus G with both time and frequency. This behaviour was first modelled by an empirical law known as structural damping from material engineering (Hildebrandt, 1969) and later on associated to the fractional visco-elastic Kelvin-Voigt model (a springpot and a dashpot in parallel), see *e.g.* Bonfanti et al. (2020). The power-law decay (in time) is quite impressively not depending much on the particular experimental technique, neither on the different type of cell: the decay exponent α mostly belongs to the interval $[0.2 - 0.4]$ (Fabry et al., 2001; Streppa et al., 2018; Trepat et al., 2005; Gerasimova-Chechkina et al., 2018; Khalilgharibi et al., 2019). Notice that $\alpha = 0$ corresponds to a purely elastic response and $\alpha = 1$ to a purely Newtonian fluid response. Other models for cell rheology have been used (see for instance (Bonfanti et al., 2020; Carmichael et al., 2015)), but they all require a fractional visco-elasticity to capture the cell response, resulting in a power-law decay $G(t) \sim (t/\tau)^{-\alpha}$ or in more complicated functions such as the Mittag-Leffler function (Gorenflo et al., 2002), which can be approximated for $t/\tau \to 0^+$ by the stretched exponential (Baró and Davidsen, 2018):

$$G(t) \sim e^{-(t/\tau)^{\alpha}/\Gamma(\alpha+1)}. \tag{4.1}$$

For our purposes the three functional forms will be considered as equivalent, since they all account for a slow relaxation dynamics given by the glassy structure and thus a memory of the past deformation.

Considering cells as soft glassy materials, we can extrapolate that they are constructed from a disordered structure of connected discrete elements by weak attractive interactions. Each of these elements would be in a metastable state (Sollich, 1998), allowing cells to flow, and therefore prone, for instance by an external forcing, to generate avalanches of fractures, which are typically out of equilibrium processes. Indeed, by tuning the proportions of passive or active actin-binding proteins, and reorganizing the network structure, living cells can control their power-law (scale-free) CSK rheology (Gardel et al., 2008; Juelicher

et al., 2007). Interestingly, cells exhibit both solid and liquid-like properties, as we have observed in Chapter 2. Solid-like behaviour is associated with strongly cross-linked actin filaments which resist sliding and accumulate tension (Dos Remedios et al., 2003; Tseng et al., 2005), while weakly cross-linking proteins produce actin filaments which slide more readily, enabling the network to flow as a liquid (Esue et al., 2009).

This paradox can be solved within the theory of soft glassy materials, by considering that, upon external deformations, the CSK of a living cell can undergo deep structural transformations such as the unfolding of protein domains, the unbinding of cytoskeletal cross-linkers, and the breaking of weak sacrificial bonds. All these structural changes are inelastic (non-reversible in a strict sense), they dissipate locally the elastic energy of the CSK network (structural damping) (Wolff et al., 2010; Gralka and Kroy, 2015).

The ability of cells to switch quickly from fluid-like responses to more brittle solid-like responses (Polizzi et al., 2018) is directly linked to the interplay between stability/rigidity and flexibility, in a way similar to what happens in neural networks (Badre and Wagner, 2006). This fast switch is likely driven by avalanche processes, which allow fast transfer of information. At the same time, such events reduce the connectivity of the CSK and may result in permanent plastic deformations or even more dramatic irreversible failures (Lieleg et al., 2010; Kollmannsberger and Fabry, 2011) which, for instance, could be at the origin of the recently observed incomplete shape recovery of living cells after repeated creep (Bonakdar et al., 2016). These effects are reminiscent of those in cyclically loaded solids which can lead to fatigue-induced failure (Song et al., 2013; Chuang et al., 2013; Salje et al., 2017).

The observation of universal relationships governing cell (and not only) rheology is evocative of the universality of statistical mechanics systems, such as the Ising model (Ising, 1925), in which individual details of the filaments and particular molecular interactions are unimportant to identify global behaviours.

4.2.2 Rupture event statistics from two primary cell lines

Experiments described in Chapter 2 gave some evidence of log-normal avalanches of fractures, but also of log-normally distributed global moduli. As far as we are concerned in this chapter, we only recall that it was important to carry out AFM indentation experiments (Laperrousaz, Berguiga, Nicolini, Martinez-Torres, Arneodo, Satta and Argoul, 2016; Streppa, 2017; Polizzi et al., 2018) with very sharp tips (pyramidal or conical) with a tip curvature radius of only a few nanometers, allowing their penetration inside the meshes of the cross-linked network. Additionally, the indentation was performed at constant velocity ($1\,\mu$m/s), and therefore constant strain rate, so that one indentation-retract experiment lasted for only a few seconds

Some examples of the statistics of these singular fracture events, together with global shear relaxation moduli distributions are shown in Fig. 4.2. We can see that both the force drop F_d (Fig. 4.2**A**), caused by the avalanche of failures, and the indentation length ΔZ (Fig. 4.2**C**) over which the avalanche takes place are log-normally distributed (the plots show the distribution of the logarithm of the variable, then a log-normal in a log-log representation is a parabola). For the indentation length ΔZ we could separate 2 different

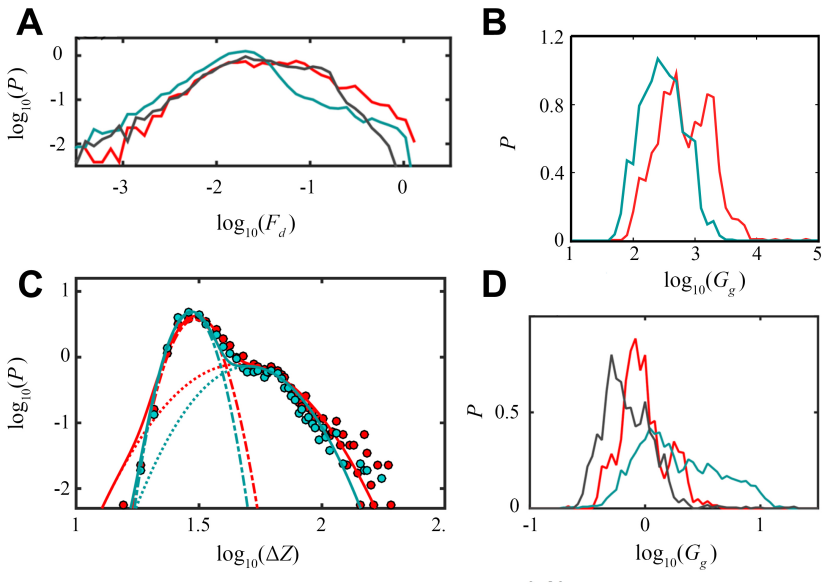

Figure 4.2 – A. Probability distributions of the logarithm of the local force drops F_d(nN) estimated from local disruption events collected from sets of myoblasts (red), myotubes (blue) and ATP depleted myoblasts (black). **B.** Probability distributions of the logarithm of shear relaxation modulus G_g(kPa) estimated by a parabolic fit of the Force-Indentation Curves (Sneddon model (Sneddon, 1965)), from healthy (blue) and leukemic (red) immature CD34$^+$ hematopoietic cells. **C.** Distribution of the indentation lengths of the rupture events detected from healthy (blue) and leukemic (red) immature CD34$^+$ hematopoietic cells. **D.** Probability distributions of the logarithm of shear relaxation modulus G_g(kPa) on the same sets of cells as in **A.**. Note that plots **A** and **C** are in log-log scale and plots **B** and **D** are in semi-log scale. All the logarithms are here in base 10. Panels **A** and **D** are adapted from (Streppa, 2017), **B** from supplemental material of (Polizzi et al., 2018) and **C** from (Polizzi et al., 2018).

regimes of avalanches both of which approximately log-normally distributed (see again Ch. 2). This log-normal statistics was found with different cell lines: in Figure 4.2**A** are shown myoblasts (red), myotubes (blue) and ATP depleted myoblasts (black), while in Figure 4.2**C** are shown immature CD34$^+$ hematopoietic cells (blood cells) healthy (blue) and leukemic (red). It follows that the energy released $E_d = F_d \Delta Z$, not shown here, is also log-normally distributed. The same log-normal distribution is also observed on global quantities such as the global shear relaxation modulus G_g (see Fig. 4.2**B** and **D**), extracted by the identification of the whole force indentation curve with the Sneddon model (Sneddon, 1965).

We speculate that the log-normal statistics observed in different living cell types on macroscopic elastic moduli (see also results on breast tissue cells from (Carmichael et al., 2015)) is strictly connected to the microscopic processes taking place in the polymer network, *i.e.* avalanches of reorganization fractures. Indeed, the log-normal statistics of avalanches would reflect in a log-normal noise (skewed fluctuations are observed also *e.g.* in (Carmichael et al., 2015)) biasing the estimation of the global elastic moduli. We can thus interpret the shear-relaxation modulus as the response of a sum of microscopic avalanches, triggered by a dynamical reorganization of the network structure. This interpretation is supported by Sollich's theory of soft glassy rheology (Sollich, 1998), in which the glassy polymer is interpreted as composed of individual units in metastable energetic states (with

different energy depths) and in which rearrangements are due to disordered interactions summarized by an effective temperature.

4.3 A random network model for the cell cytoskeleton

Let us now introduce our random network model for the cell CSK. We consider a network with N nodes, being identified as the actin filaments, connected by randomly assigned (in a way explained hereafter) links, identified as the cross-linker proteins. In light of what we said previously about CSK mechanics (see Section 4.2.1), if we want to build a random network model for the cytoskeleton we need the network to be in a percolating regime. Indeed, since a cell can be seen as a soft glassy material there needs to be a giant, percolating cluster of connected nodes, allowing for the transition between a fluid material and a glassy one. A second important condition that we ask the network to satisfy is the correct degree distribution (*i.e.* the distribution of cross-links per filament) as observed in real living cells CSK. Unfortunately, precise data on this distribution do not seem to be available to our knowledge. However, from the literature (Eghiaian et al., 2015), we can deduce that the degree distribution would not be well approximated by a power-law and therefore the CSK could not be modelled as a scale-free random network. There is also another reason why the scale-free network may not be a good model for cell CSK: it has been proved (Pastor-Satorras and Vespignani, 2001) that on scale-free networks the critical threshold for an epidemic is exactly 0, meaning that avalanches would always affect a finite proportion of the population (in the large N limit), thus implying irreversibly a large proportion of cross-linkers, scaling as the size of the system. This is not what is observed for avalanches of fractures in cells, since their distributions are of finite size.

For all these reasons, we propose a Gilbert (Gilbert, 1959) random graph to model the CSK structure. A Gilbert graph is a variant of the Erdős-Rényi graph (Erdös and Rényi, 1959), and the two graphs are equivalent in the large N limit (our case). The network is constructed in the following way: i) we define a network of N isolated nodes, with N fixed; ii) we connect each possible pair of nodes with probability p_l.

This process creates a network with a binomially distributed degree of connections:

$$p_k = \mathcal{B}\left(N-1, p_l\right) = \binom{N-1}{k} p_l^k (1-p_l)^{N-1-k}, \tag{4.2}$$

where the binomial factor takes into account all the possible pairs of nodes having a degree of k, out of all possible $N-1$ links coming out from a node. The probability p_k indicates the probability that a randomly picked node has k links. The average degree of the random network can then be computed:

$$\langle k \rangle = \sum_{k=1}^{N-1} k p_k = (N-1) p_l. \tag{4.3}$$

In the limit of $N \to \infty$, keeping $\langle k \rangle$ fixed, the binomial degree distribution is well

approximated by the simpler (one parameter) Poisson distribution, noted $\mathcal{P}(\lambda)_k$, with parameter $\lambda = \langle k \rangle$. The average connectivity $\langle k \rangle$ will be a control parameter of our model, since its value is not well-known from experiments.

The average global number of links of the network, and thus of cross-linkers available for the avalanche, can also be computed by multiplying the probability of connection of pair of nodes p_l with the number of all the possible pairs (by excluding double counting):

$$\langle N_l \rangle = p_l \frac{N}{2}(N-1), \tag{4.4}$$

and then for a network of $N = 10^4$ nodes, and $p_l = 5 \cdot 10^{-4}$, the average number of links is $\langle N_l \rangle \simeq 50000$. The distribution of the number of links is also a binomial.

Before moving on and describing the avalanche process, we would like to observe some characteristics of this network useful for our modelling. From probabilistic arguments it can be noticed (see (Erdös and Rényi, 1959)) that, in the limit of $N \gg \langle k \rangle$, with fixed $\langle k \rangle$ (as in all the simulations of this article), the fraction u of nodes not belonging to the largest connected component of the network is given by the transcendental equation:

$$1 - u = 1 - e^{-\langle k \rangle (1-u)}. \tag{4.5}$$

This can be found by considering that the fraction of nodes u is composed by the sum of two separated events. A node i not belonging to the giant component is: i) either disconnected to an other node j because the link $i - j$ does not exist, and this happens with probability $1 - p_l$, ii) or connected to a node j that also does not belong to the giant component, and this occurs with probability up_l. Therefore $u = (1 - p_l + up_l)^{N-1}$, because j can represent $N - 1$ different nodes. From Equation (4.5) it can be proved (Erdös and Rényi, 1959; Barabási and Pósfai, 2016) that for $\langle k \rangle > 1$ a giant percolating cluster exists and its size scales as the network size N.

At this point we can work out the probability p_{FC} that we pick up a random link not belonging to the giant cluster:

$$p_{FC} = \frac{u(Nu-1)}{N-1}, \tag{4.6}$$

where the right hand side is the ratio of the average number of links not belonging to the giant cluster $p_l Nu/2(Nu-1)$ and the average number of links of the network $p_l N/2(N-1)$.

On this network structure, an avalanche of ruptures is described by the following algorithm. Notice that while the probability of connection p_l remains constant, the probability of breaking $p(t)$ can depend on time, and therefore is able to capture time relaxing processes.

- We pick at random a link of the network and we break it (removing it from the network), this initiates the avalanche;
- we look for the links coming out from both extremities of the broken link;
- we break all of them with a probability $p(t)$, where at the first time step $t = 0$;

- we consider the links coming out from the ones broken at the previous time step from both extremities and break them with probability $p(t+1)$.

We repeat the last point and increase the time step by one unit, until the avalanche stops when there are no more links that break, either because the probability of breaking is too low or because of the damage caused to the network, which will have decreased the abundance of unbroken links. The time at which the avalanche stops, t^*, defines the duration of the avalanche and the total size of the avalanche Z is the total number of broken links. Here, we focus on the distribution of Z, which is, up to some conversion factors, equivalent to the energy released during the avalanche. Overall the stochastic process has 2 sources of randomness: the random structure of the network and the avalanche process itself which is a stochastic process.

Rupture avalanches propagate on networks that remain static during the rupture process, since actin polymerization and active cross-linkers act on time scales much longer than the time scale of the experiment, which is of the order of 1 s. Nevertheless this assumption is relaxed in section 4.5.1, to account for the fact that cross-linkers can be restored on very short time scales even by physical contact.

Statistically speaking an avalanche is a binomial process of probability $p(t)$ of success, with a random number of trials (the number of available links). If the number of available links were always drawn from the same Poisson distribution, with parameter $\lambda = 2Np_l$ (the factor 2 is because we consider both extremities of a broken link), we could conclude that the distribution of broken links at time t is also a Poisson distribution with a new parameter $\lambda_1 = 2p(t)Np_l$. The choice of the Poisson distribution is motivated by the degree distribution, which, as we said, is approximately Poissonian in the $N \gg \langle k \rangle$ limit keeping $\langle k \rangle$ fixed, and the parameter $\lambda = 2Np_l$ corresponds to the average number of links available at the first time step. This can be shown by applying the law of total probability to the probability of having k broken links:

$$p_b(k) = \sum_{n=k}^{\infty} \mathcal{B}(n,p)\,\mathcal{P}(\lambda)_n = \frac{(2pNp_l)^k}{k!} e^{-2pNp_l} = \mathcal{P}(\lambda_1 = 2pNp_l, k). \tag{4.7}$$

The probability of having k broken links is then given by the product of the Poisson distribution, which represents the probability of having n links available for breaking, times the binomial distribution, which represents the probability of having k successes out of n trial. To include all possible disjoint events we have to sum over $n > k$, up to $N-1$, which in the considered limit tends to infinity.

The problem with this reasoning is that the number of available links is not drawn from the same distribution as far as the avalanche moves forward. Indeed the number of available links depends on the particular path of the avalanche, since the broken links disappear from the network. Therefore as far as the avalanche progresses, the number of available links may not even follow a Poisson distribution, making the analytic expression of the number of broken links impossible. The same is true if we consider the restoration of crosslinks, as in section 4.5.1, because the avalanche process introduces an effective and uncontrolled time dependence of p on t, by changing the actual number of available links. This effect is the main reason why we do not find perfect log-normals in our model, as in

our first 1D mean field model (Polizzi et al., 2018) or related models without a network structure (see Chapter 3). Actually, this was checked on a random multiplicative process with a multiplicative factor drawn, at each time step, from the same distribution as in Equation (4.7) for different parameters λ_1 proving that, as expected from our results in Chapter 3, there is a region where the avalanche size distribution is exactly log-normal. Besides, we have studied the effect of loops (the path between two nodes may be not unique) by running the avalanches on a tree graph, thus a graph without loops, and considering avalanches moving in only one direction, thus eliminating uncontrolled effects on $p(t)$. In this way, in the large N limit the avalanche size statistics seems indeed to converge to a log-normal. However these results are not shown here, because there were some unclear finite size effects.

We considered $\langle k \rangle > 1$ in all our simulations, in order to have a giant component in the network, as explained previously. We also checked the effect of picking the first link initiating the avalanche out of the giant component. Indeed, from Equations (4.5) and (4.6) we can compute p_{FC}. For $\langle k \rangle = 2$ for instance $p_{FC} = 0.04$ is already irrelevant for our results, and for $\langle k \rangle = 4$, $p_{FC} \simeq 4 \cdot 10^{-4}$, is completely negligible. This was checked numerically in all the following simulations and the results do not change by running the avalanches only on the giant component. We should also mention that avalanches containing only the first randomly chosen link are not taken into account for the statistics, this making even more negligible the effect stated above.

It is conceptually interesting to note that the proposed avalanche model is actually equivalent to an epidemic model running with probability $p(t)$ on the line graph of the original network, *i.e.* on the graph obtained by transposing links into nodes and nodes into links.

Simulations of avalanches were done with self-written codes using open source software R version 4.0.2 and Python 3.6. Figures and data analysis were done with Python 3.6 using basic packages and a customized version of the module *powerlaw* (Alstott et al., 2014) with a corrected computation of the cumulative distribution function and added graphical features. For all the following simulations it was checked that changing the size of the network N (but keeping the same values of $\langle k \rangle$), the phase diagram did not change significantly, showing always the same types of distribution for $N = 1000, 5000, 20000, 50000$. It was also checked that running the avalanche algorithm on the same network for each stochastic realization (of course with all links restored), was equivalent, since the first initiating link is chosen randomly, to run the avalanche on a different network for each realisation. The first option is computationally faster and was then used for detailed results.

4.4 Avalanche statistics with constant probability of breaking

First, we implemented the avalanche model with a probability of breaking $p(t) = p$ constant with time. We detected 3 regimes with different distributions of the avalanche size. A phase diagram is shown in Figure 4.3**B**. The three phases are *pop* (blue), *mixed*

(purple), *explosive* (red). The *pop* regime presents avalanche size distributions which decrease faster than a power-law, containing then small avalanches, with a low number of broken links of less than 100. On the other hand the *explosive* regime is composed of large avalanches, scaling as the system size and then causing global damage to the network. In between these two regimes the *mixed* regime is composed of avalanche distributions made of a mixture of the two previous regimes.

A likelihood-ratio test, based on the ratio of the likelihood of the model fitted by two different distributions, can be done to compare possible outcome distributions (Severini, 2001). In Figure 4.3**D**, we show the maximum of likelihood (ML) fit for the log-normal distribution and the truncated power-law distribution, with density function $kx^{-\alpha}e^{-\lambda x}$ (a power law with an exponential cut-off with λ decay rate and k normalization constant). The truncated power-law distribution is chosen to take into account the finite size of the system (see for instance Ch. 4 of (Barabási and Pósfai, 2016)).

When applying the likelihood-ratio test to compare the two distributions, it turns out that the log-normal is always a better fit than the truncated power-law. For the example of Fig. 4.3**D**, the p-value is $1 \cdot 10^{-22}$, rejecting the null hypothesis that the likelihood ratio is equal to 1, which would mean that none of the selected distributions is preferable. This result is coherent with those obtained on 1-D models in Chapter 3, where for low multiplicative factors, the size distribution was approximately log-normal, with the difference that here the process is discrete (not allowing sizes < 1). Therefore we have only the right tail of the distribution, while in the 1-D models, the multiplicative process was continuous. We can thus conclude that in the *pop* regime the distribution is statistically compatible with a log-normal tail.

The boundary between the *pop* and *mixed* regimes is computed by an algorithm detecting for which parameters $\langle k \rangle$ and p the avalanche size distribution is better fitted by a truncated power-law over the whole domain. We chose this criterion because power-law distributions are typical of critical points and at the same time we wanted to include the finite size cut-off. Moreover, when comparing the tails of the distributions (by using the algorithm described in (Clauset et al., 2009) and implemented in (Alstott et al., 2014)), at the boundary the most likely distribution is the truncated power-law, with always a very significant p-value $\ll 10^{-4}$. The null hypothesis is again of a likelihood ratio equal to 1 (compared to all log-normal, exponential and power-law). In Figure 4.3**C** is shown a typical distribution of the avalanche size at the boundary between *pop* and *mixed*. For comparison we also show the maximum of likelihood fit with a power-law.

In the *mixed* regime the size distribution consists of two separated parts: a small size part similar to the *pop* regime distribution and a large size part corresponding to a very narrow bump, whose avalanches scale as the system size. In Figure 4.3**E** we show the full Z distribution in a log-log plot, in the inset is shown in detail the distribution of the large avalanche size narrow bump, represented by only one dot in the large field view. In this regime, moving toward the *explosive* regime causes a shift to the right of the large avalanche size bump and a decrease of importance of the small avalanche size part. Notice that the large avalanche size bump is not disappearing in the large N limit: if we take a distribution in the *mixed* regime for a given p and k and we increase the size of the network N (we did it for $N = 2000, 5000, 10000, 20000$) the proportion of avalanches

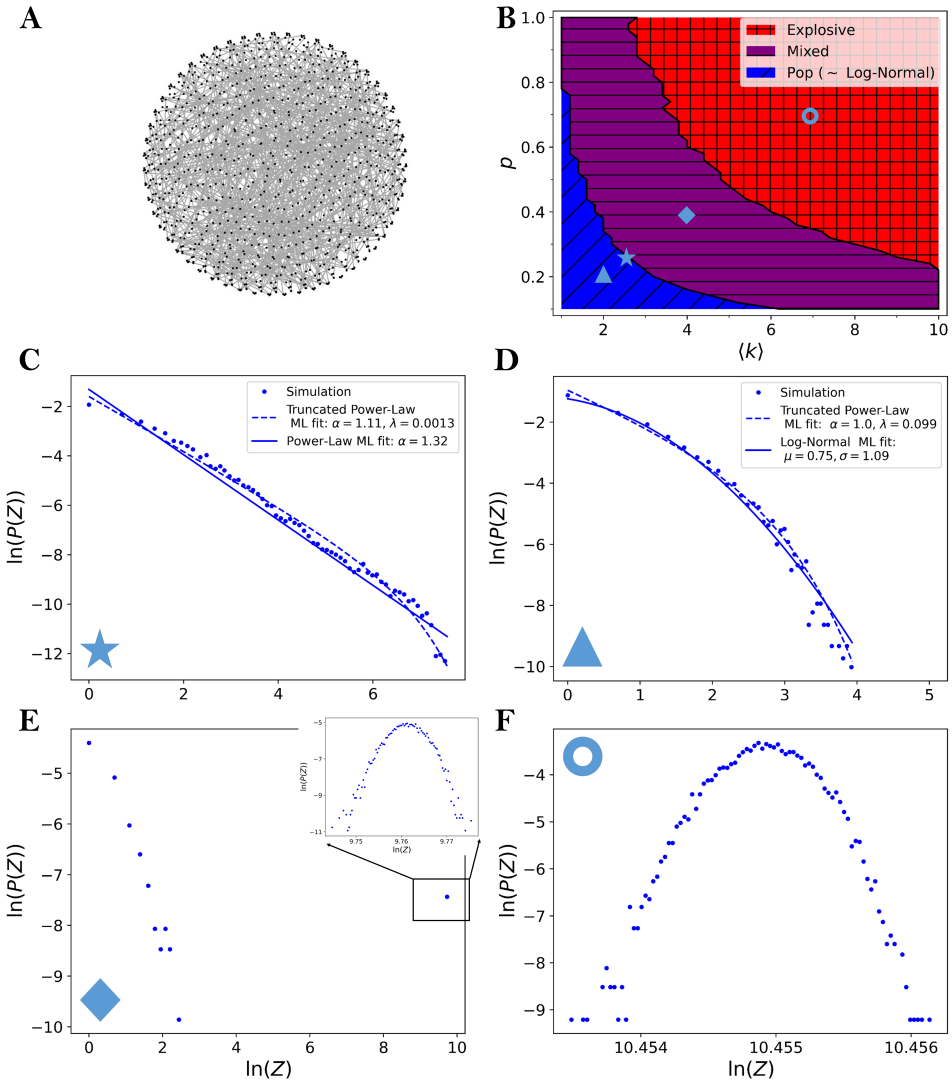

Figure 4.3 – **A.** Picture of the random network model, embedded in a sphere. Notice that this is just a choice of representation, the shape of the network is not fixed: our network model is only a set of connections, it can be embedded in any metric space. Dots are the nodes of the network and lines are the links. For better visibility the network has here only $N = 1000$ nodes and a probability of connection $p_l = 0.003$. **B.** Phase diagram of the possible resulting avalanche size distributions with respect to the model parameters $\langle k \rangle$ (the average number of connections of the network) and p (the probability of breaking a link). The three identified regimes are: *pop* (blue), which has an avalanche distribution statistically compatible with a log-normal tail; *mixed* (purple) where a mixture of the two different behaviors, *i.e.* big large size avalanches and pop avalanches, are observed at the same time; *explosive* (red) where almost only big avalanches spanning all the network size are observed. For details about the detection of the transitions between regimes see text. The number of nodes of the network is $N = 10000$ and the statistics is done out of $n = 10000$ repetitions. Symbols indicate the points taken to show some distribution examples in the remaining panels. Star has coordinates $(2.6, 0.26)$, triangle $(2, 0.2)$, diamond $(4, 0.4)$ and doughnut $(7, 0.7)$. **C.** Avalanche size distribution exactly at the transition between *pop* and *mixed*. At this transition the distribution is a power-law for a range of almost 3 decades, more precisely the distribution is a truncated power-law, given the finite size of the system. Dots show the simulation distribution and the dashed (resp. full) line shows the maximum of likelihood fit with a truncated power-law (resp. a power-law) distribution. **D.** Avalanche size distribution in the *pop* regime. Dots show the simulation distribution and the full (resp. dashed) line shows the maximum of likelihood fit with a log-normal (resp. truncated power-law) distribution. **E.** Typical distribution in the *mixed* regime, the inset shows a zoom on the distribution of large size avalanches not visible in the main panel. **F.** Typical distribution in the explosive regime.

in the bump stays the same, even though the bump shifts to the right, because those avalanche sizes are proportional to the system size.

Finally, the *explosive* regime, a typical distribution of which is shown in Figure 4.3**F**, is detected by the criterion of less than 1% of small size avalanches. In this regime the distribution of avalanche sizes is a very narrow bump having a size comparable to the number of links of the system. The integrity of the whole network is therefore compromised and only a few sparse links are intact.

Note that even though it could be tempting to say that the bump in *mixed* and *explosive* regimes is a log-normal distribution (remember that in a log-log representation a log-normal becomes a parabola), this cannot be assessed, because the distribution is very narrow and therefore a log-normal cannot be distinguished from a normal distribution (see Section 1.3.2). In none of the three regimes observed so far we can recognize avalanche size distributions similar to those observed experimentally, *i.e.* approximately log-normals and of finite size. Nonetheless, we observe that the distributions of the *pop* regime (showing a log-normal tail) are strikingly similar to those observed for avalanches of firing rates in freely behaving rats in (Ribeiro et al., 2010). In that work the distributions are also claimed to follow a log-normal tail, but they justified its appearance with the undersampling of the data, supposed instead to be at a critical state (and therefore power-law distributed). Now we can give an other possible reason of the observed distributions, which can be given either by a small connectivity of the network or by an extremely fast absorption of the avalanches.

We shall stress here the fact that our network model does not have a metric, hence no concept of distance, it is only a network of interactions. It does not matter if the links are close in space (see for a possible representation of the network Fig. 4.3**A**), but only if they «communicate». However, it is possible to embed the random network model in any metric space and in any shape, considering for example nodes that are connected to each other closer than others, without impacting our results: as far as the interactions follow this structure the shape of the network does not matter. With this interpretation the small *pop* avalanches in the *mixed* regime can be thought as a boundary effect: they represent avalanches where the starting randomly chosen link is in a less than average connected region.

4.5 Avalanche statistics by introducing visco-elasticity

We now introduce the exponential relaxation coming from fractional visco-elasticity (see Section 4.2.1). We let then the probability of breaking be dependent on time in the following way:

$$p(t) = p_0 e^{-(t/\tau)^\alpha}. \tag{4.8}$$

The Γ coefficient in Eq. (4.1), is here embedded in the constant τ. This relation imposes some temporal correlation during the avalanche, added to the interaction structure.

Conceptually, Eq. (4.8) assumes that the local fracture mechanics follows the same mechanical response as the global network given by the shear-relaxation modulus (see Section 4.2.1). It is important to note that the limit $\tau \gg 1$ is equivalent to setting $\alpha = 0$, since in both cases $p(t)$ loses the dependence on time and we recover the results of section 4.4. We can thus interpret the model in section 4.4 as a purely elastic response.

In Figure 4.4 we show the results after introducing a time dependent $p(t)$. The phase diagram also shows 3 different regimes. The *pop* and the *explosive* regimes have the same statistics as in Section 4.4, and an example of them is shown in Figure 4.4**C** and **D**.

The important difference introduced by a fractional visco-elastic relaxation is in the *mixed* regime, which shows now the emergence of an approximately log-normal finite size bump, and it is thus called the *log-normal bump* regime, in magenta. Indeed the very narrow bump observed in the non-visco-elastic model broadens here due to the time relaxing probability and it is not at a size comparable to the size of the system, therefore not causing a global destruction of the network. Avalanches in this regions do not scale as the systems size. Observing Figure 4.4**B** we can see that there are still some small avalanches of *pop* type in the distribution. These may still be an effect of picking low reticulated zones as starting point of the avalanche, but as previously said, even running the avalanches only on the giant cluster they are still present. Notice that while moving toward the red region the log-normal bumps shrinks and moves to larger sizes becoming then similar to the distribution in Figure 4.3**E**.

The boundaries between the regimes are detected with the same algorithms as described in Section 4.4. Notice that here we can let the probability p_0 become larger than 1, since the stretched exponential decay will always make $p(t)$ converge toward 0. Simulations in Figure 4.4 are shown for $\tau = 4$, and the influence of varying τ has also been studied. Increasing τ produces a shift of the distribution to the right, since the avalanche tends to last longer. However for having a log-normal bump in the size distribution, τ must not be too large, because otherwise the distribution would tend to that one shown in Figure 4.3, with a larger proportion of small avalanches similar to the *pop* regime and a very narrow large avalanche size bump. That is coherent with the fact that for $\tau \to \infty$, $p(t) \to p_0$ finding again thus a constant probability of breaking as in Section 4.4. On the other hand, decreasing τ would make the relaxation very fast and shift the distribution toward the *pop* region. If decreasing τ is compensated by an increase of $\langle k \rangle$ the weight of the small size avalanches in Fig. 4.4**B** is decreased, making the log-normal bump even more pronounced (see Fig. 4.5), and broadening the magenta region. As a rule of thumb, $\tau < 5$ (in arbitrary units) is a good compromise to find an approximately log-normal bump. The value of α, coherently to rheological experiments on living cells, is set to 0.3 in all the simulations.

In conclusion, the best receipt to have log-normal avalanches, without at the same time having small *pop* avalanches, is then to decrease τ in such a way that if increasing p_0 or $\langle k \rangle$ the avalanche stops fast enough to escape from the red regime (and then a very narrow large avalanche size bump) possibly together with a few small *pop* avalanches, but still p_0 or $\langle k \rangle$ has to be large enough to quit the blue region. In other words we get log-normal avalanches if the avalanche starts explosive and relaxes fast, ending gently. We showed this in Figure 4.5, where the same simulations as before were run, but with $\tau = 1$. Comparing Figures 4.4 and 4.5 we can observe that the magenta region is broadened, and

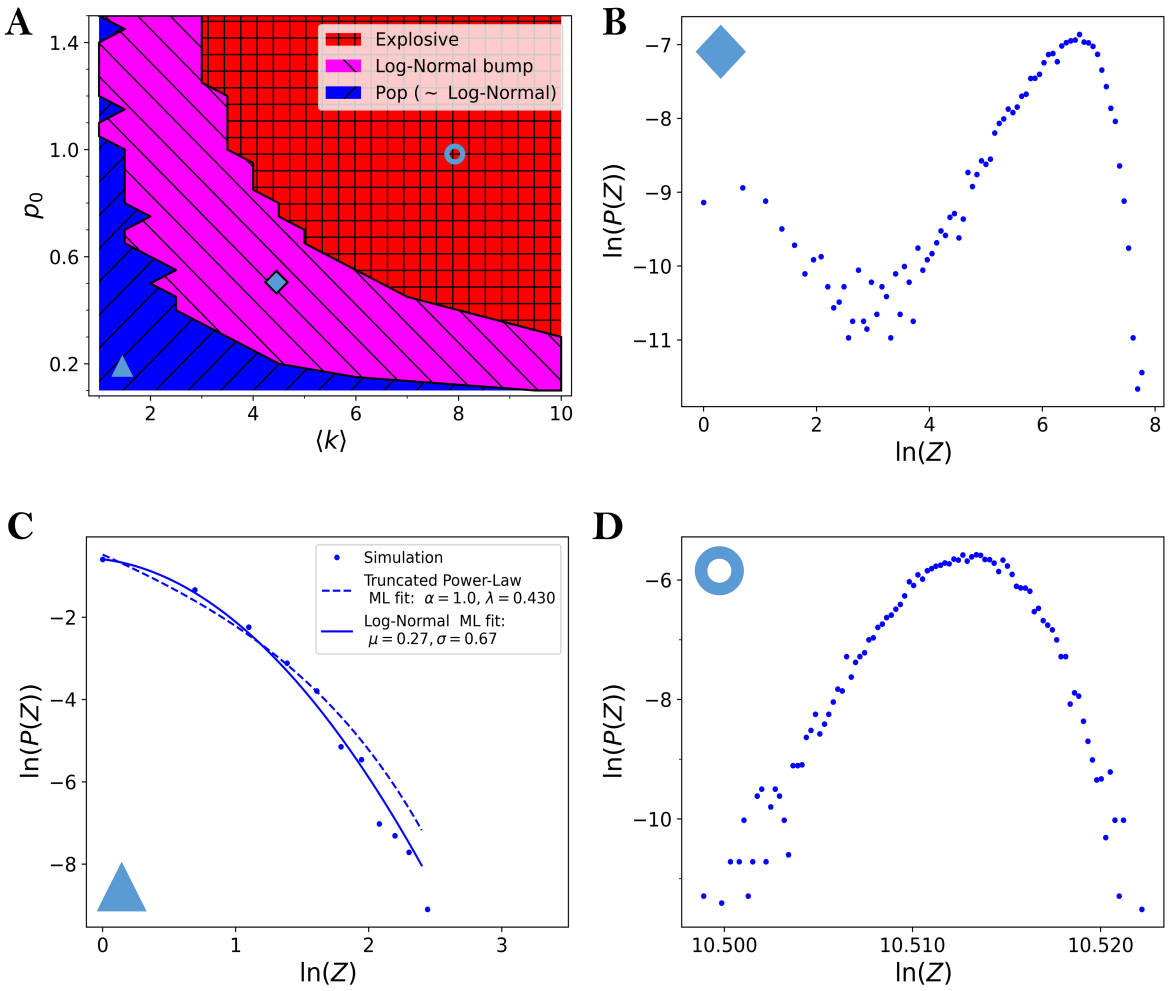

Figure 4.4 – **A.** Phase diagram of the resulting avalanche size distributions with respect to the model parameters $\langle k \rangle$ and p_0, the constant pre-factor of the breaking probability (see Eq. (4.8)). The time constant is $\tau = 4$ and the rheological exponent is $\alpha = 0.3$. Symbols indicate the points taken to show some distribution examples in the remaining panels. The *pop* (blue) and *explosive* (red) regimes are the same as in Figure 4.3, while the *mixed* regime changes, becoming composed of a broader bump, approximately log-normally distributed (in magenta), together with some small avalanches reminiscent of the *pop* regime. This regime is called here *log-normal bump* (magenta). **B.** Example distribution of the *log-normal bump* regime with coordinates at the diamond point $(4.5, 0.55)$. **C.** Example distribution of the *pop* regime, with coordinates at the triangle point $(1.5, 0.2)$. Dots show the simulation distribution and the full (resp. dashed) line shows the maximum of likelihood fit with a log-normal (resp. truncated power-law) distribution. **D** Example distribution of the *explosive* regime, with coordinates at the doughnut point $(8, 1)$.

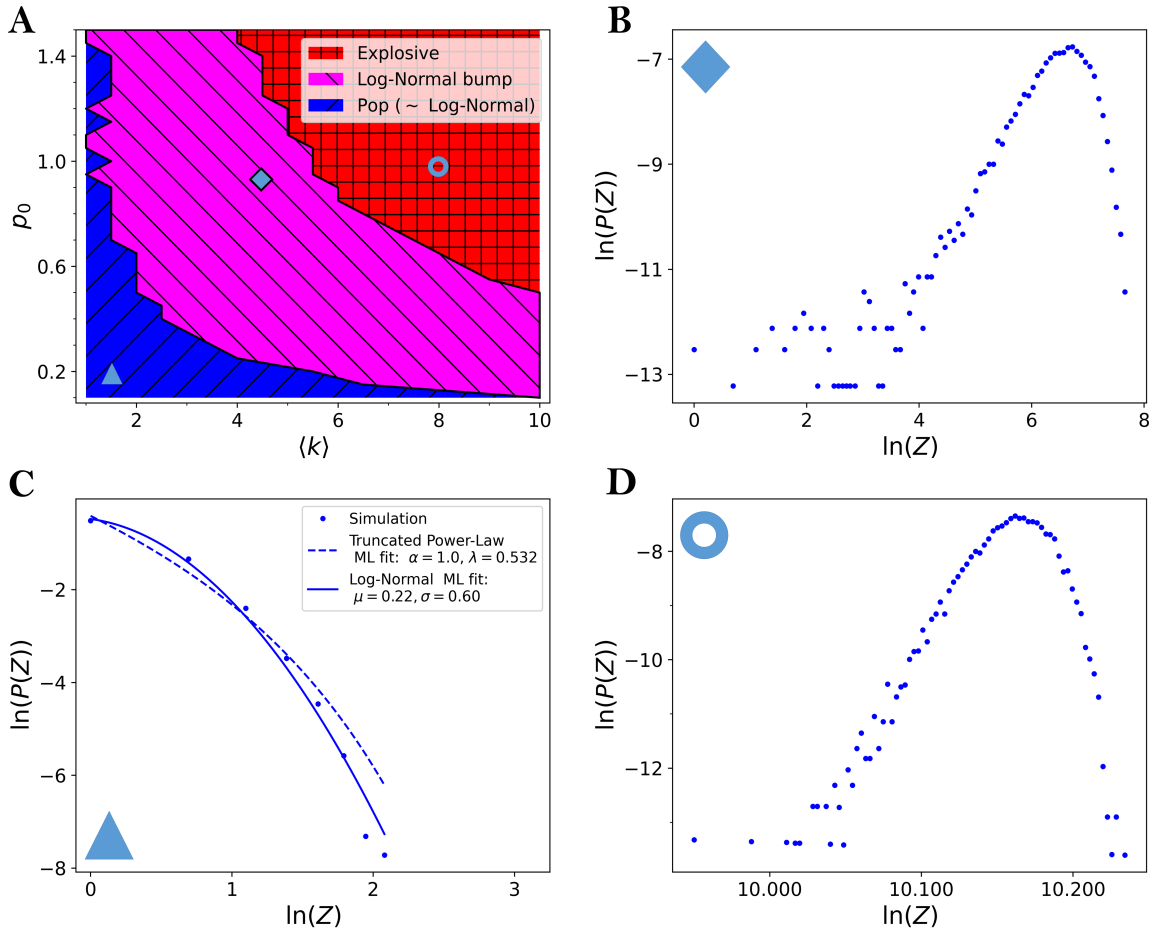

Figure 4.5 – A. Phase diagram of the resulting avalanche size distributions with respect to the model parameters $\langle k \rangle$ and p_0, the constant pre-factor of the breaking probability (see Eq. (4.8)). The time constant is here $\tau = 1$ and the rheological exponent is again fixed at $\alpha = 0.3$. Symbols indicate the points taken to show some distribution examples in the remaining panels. The *pop* (blue) and *explosive* (red) regimes are the same as in Figure 4.3, while the *log-normal bump* (magenta) regime is composed of approximately log-normal avalanche sizes. Note that here the regime is much wider than in Figure 4.4 because of the different choice of τ. **B.** Example distribution of the *log-normal bump* regime with coordinates at the diamond point $(4.5, 0.95)$. **C.** Example distribution of the *pop* regime, with coordinates at the triangle point $(1.5, 0.2)$. Dots show the simulation distribution and the full (resp. dashed) line shows the maximum of likelihood fit with a log-normal (resp. truncated power-law) distribution. **D.** Example distribution of the *explosive* regime, with coordinates at the doughnut point $(8, 1)$.

in the new area, which from red became magenta, distributions with a very low proportion of *pop* avalanches emerge.

We note that these new distributions are not exactly log-normal, but only approximately. It is for example possible that they could be as well approximated by other skewed distributions, as the Gamma distribution, but for sure not by power-laws. It is interesting that these skewed distributions of the *log-normal bump* region are originated by both the introduction of time correlation (in the form of visco-elastic relaxation) and the non-linear dynamics of the avalanche propagation. Similar arguments (but instead of time correlations it was for phase correlations) were used in (Ivanov et al., 1996, 1998) to explain

the appearance of skewed (non-power-law) distributions for the temporal variability of the heart rate, suggesting that an avalanche process of this type may also be latent in that case.

4.5.1 Avalanche statistics with local restoring of cross-linkers

In order to test the effects of local cross-linkers restoring on the final avalanche size distribution, we consider the most extreme case, where all broken links are restored after only 1 time step. Therefore the broken links disappear from the network only for the first time step after their rupture. This could be interpreted as an extremely fast network restoring. In this scenario, the probability of breaking must relax with time, otherwise the avalanche would never stop as soon as $\lambda_1 > 1$ (see Eq. (4.7)). We thus take the same law of breaking as in Eq. (4.8), with $\tau = 4$.

The phase diagram and the resulting avalanche size distributions, when propagating avalanches with local restoring of cross-linkers, are almost the same as in Section 4.5 (compare Figs. 4.4 and 4.6). We have chosen exactly the same points in the phase diagram as in Fig. 4.4 for ease of comparison. The *pop* region is very similar to the one in Fig. 4.4. The two other distributions have the same shape as before, but they are shifted to larger values, because the number of links available for the avalanches is more abundant than in Fig. 4.4. This is a direct consequence of the link repair mechanism introduced for this simulation. We can thus conclude that our results about the *log-normal bump* regime are robust and do not seem to depend on the details of the particular avalanche algorithm, provided a visco-elastic relaxation is included in the model. However these details can be important to tune the average and the width of the resulting distributions.

4.6 Avalanche propagation along the weaknesses of the network

In this short section we consider an alternative rule of propagation, in which the avalanche propagates with the following algorithm:

- we pick at random a link from the network and we break it (removing it from the network), as previously;
- we attribute a random (uniformly chosen) number to each of the neighbour links;
- we break the link who has the smallest attributed number.

The avalanche then stops only when there are no more links attached to the broken one.

This algorithm simulates a propagation by attributing different strengths to the cross-linkers, and can be interpreted thus as an avalanche propagation along the weaknesses of the material.

Notice that here the only control parameter is the average number of connections k. In Figure 4.7 we show two typical distributions for the avalanche size with this new rule.

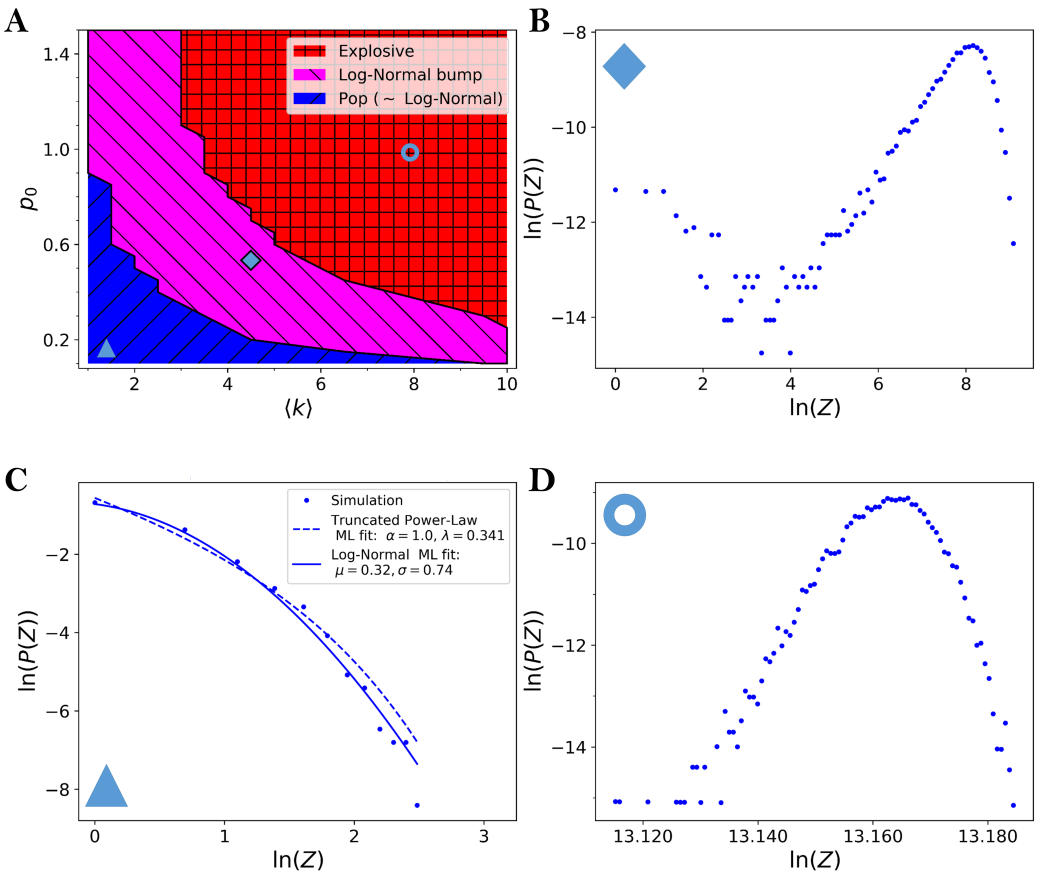

Figure 4.6 – **A.** Phase diagram of the resulting avalanche size distributions with respect to the model parameters $\langle k \rangle$ and p_0 (see Eq. (4.8)), in the case of link restoring. The time constant $\tau = 4$ and the rheological exponent is $\alpha = 0.3$. Symbols indicate the points taken to show some distribution examples in the remaining panels. The *pop* (blue) and *explosive* (red) regimes are the same as in Figure 4.3, while the *log-normal bump* regime is composed of approximately log-normally distributed avalanche sizes (in magenta), together with some small avalanches reminiscent of the *pop* regime. Compared to Figure reffig:3 the phase diagram does not show significant differences. **B.** Example distribution of the *log-normal bump* regime with coordinates at the diamond point $(4.5, 0.55)$. **C.** Example distribution of the *pop* regime, with coordinates at the triangle point $(1.5, 0.2)$. Dots show the simulation distribution and the full (resp. dashed) line shows the maximum of likelihood fit with a log-normal (resp. truncated power-law) distribution. **D.** Example distribution of the *explosive* regime, with coordinates at the doughnut point $(8, 1)$.

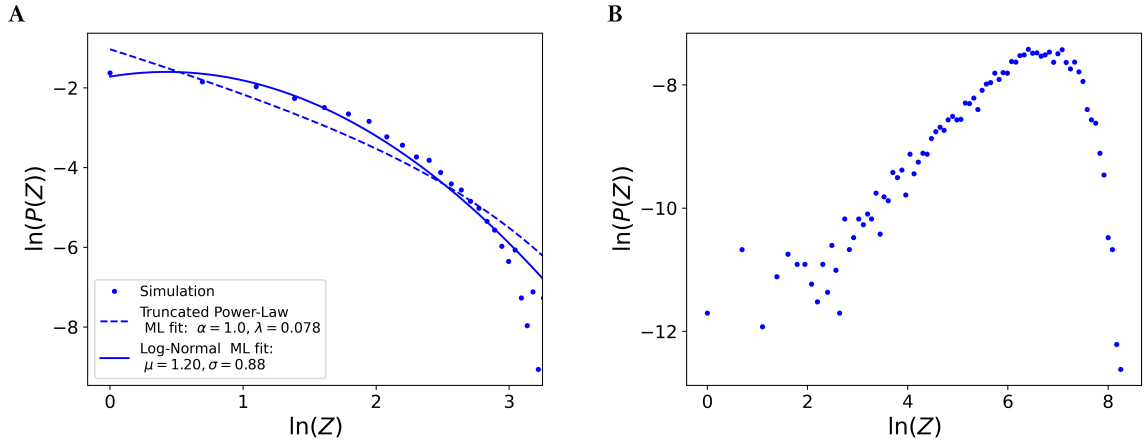

Figure 4.7 – A. Avalanche size distribution obtained for $k = 1$, with the propagation rule described in Section 4.6. Dots show the simulation distribution and the full (resp. dashed) line shows the maximum of likelihood fit with a log-normal (resp. truncated power-law) distribution. **B.** Avalanche size distribution obtained for $k = 4$, with the propagation rule described in Section 4.6.

We can see that even with this completely different breaking rule, the distributions look very similar to the previously discussed ones. We observe a *pop* distribution in Fig. 4.7**A** (for $k = 1$), showing small avalanches with an approximately log-normal tail, and an approximately log-normal distribution in Fig. 4.7**B** (for $k = 4$). The latter is actually much wider than the ones obtained with the previous propagation rule in Sections 4.5 and 4.5.1, but still the shape looks conserved, and the same 3 regimes appear to be there. However, we did not do an extensive study as before in this case. These findings suggests that the observed distributions are a general feature of avalanches on networks, no matter what particular avalanche propagation rule is chosen.

4.7 Discussion and future directions

Our phenomenological model shows that it is possible to model approximately log-normal avalanches, as experimentally observed. We showed this on a network structure which does not have a natural metric, but it is primarily a structure of interactions. Therefore interactions are not necessarily due to physical contacts, but can for example be given by vibrational modes, or other forms of interaction. However, this characteristic should not be thought of as a limitation, but rather as a feature that makes the model more general. Indeed, the network structure can always be embedded in some metric space if desired, *e.g.* on a 3-D sphere, and the same results would remain valid.

Interestingly, in order to have approximately log-normal avalanches, an elastic-instantaneous breaking mechanism is not sufficient. With this first rule we only get small *pop* avalanches, *explosive* avalanches leading to a global destruction of the network, or a mixture of both regimes. Along the critical line, where large avalanches proportional to the network size start being present, the distribution looks like an exponentially truncated power-law, typical of critical systems in physics. A key ingredient for having log-normal type avalanches is the introduction of a breaking rule taking into account the

visco-elastic memory of cells (or other glassy materials), thus introducing time correlations. This behaviour is shown to be robust even with different avalanche propagation rules, as for example by introducing the local restoration of cross-linkers. We should also mention that even with a completely different breaking rule the qualitative avalanche behaviour is the same. We did preliminary studies on avalanches with a propagation rule consisting in breaking at each time step only the weakest link (by setting randomly attributed «strength» to the links), supposing that the avalanche propagates along the weaknesses of the material, and we could again recognize the same 3 regimes. We were then able to obtain avalanche size distributions closer to a log-normal by imposing faster relaxation times as in Figure 4.5. In this way the *log-normal bump* region becomes larger, thus showing slightly new distributions with a less important proportion of small *pop* avalanches, while the other resulting distributions stay the same.

Our model does not give quantitative information, because we did not make the link with physical units. Accounting for this, however, may be possible by introducing an appropriate scaling factor, to make the size of the avalanche a physical measurable quantity (as the released energy). At the same time the correspondence of the lasting time of the avalanche with the experimental one can be done. This would attribute a unit to τ, hence leading to an estimation of the physical relaxation time needed to have log-normal avalanches. Altogether, these results suggest that, at least for cells, the local fracture mechanism resembles that of the global mechanical response of the cytoskeleton network. The global shear-relaxation modulus can thus be interpreted as a sequence of local avalanches of ruptures.

We conclude by saying that, in contrast to simpler 1-D models without a network structure (see Chapter 2 and 3), here the resulting distributions are not exactly log-normal. It is for example possible that they could as well be approximated by skewed distributions other than log-normals, such as the Gamma distribution. In any case, whichever distribution is preferred, we get an alternative to power-laws, but still obtaining skewed, stationary avalanche size distributions, non-proportional to the size of the system. Notably, these skewed distributions in the *log-normal bump* region are originated by both the introduction of time correlations (in the form of visco-elastic relaxation) and non-linear dynamics in the avalanche propagation. These are the same ingredients (but instead of time correlations it was for phase correlations) identified to cause the appearance of skewed distributions for the variations in the heart-rate signal in (Ivanov et al., 1996, 1998), behind which an avalanche process may also be latent. This paves the way to possible applications to physiological data, where non-linear units are organized in a networked communicating structure, such as brain seizures, heartbeats or chemical signalling inside cells.

Chapter 5

Conclusion and perspectives

This work was initially motivated by the observation of avalanches of local fractures in living cells, when poked by sharp cantilever tips. The analysis of these data, reported in Chapter 2 of this manuscript highlights important mechanical properties of living cells interwoven in the complex network structure of their cytoskeleton. We used the wavelet transform mathematical microscope to characterise the length and the energy of these fracture events, identified from force-indentation curves, recorded with an atomic force microscope, operated in constant speed indenting mode. The distribution of the energy released along the avalanche of fractures revealed 2 different regimes:

- A *ductile* cascade regime, containing cascades involving a small number of cross-linkers unbindings and small energies ($\bar{E} \sim 200\,\mathrm{k_B T}$). It corresponds to a liquid-like behaviour, since it acts at a scale comparable to dynamical actin repolymerisation, thus preserving the structural integrity of the perinuclear cytoskeleton.

- A *brittle* cascade regime, containing much more dramatic failures involving approximately 7 times more unbindings and released energies ($\bar{E} \sim 1300\,\mathrm{k_B T}$). It corresponds to a solid-like behaviour, since it leads to a plastic deformation and thus to an irreversible disruption of the perinuclear cytoskeleton.

These 2 regimes were observed in both healthy and leukaemic haematopoietic cell lines, but what significantly separates cancer cells from healthy ones, is the proportion of brittle avalanches with respect to the total number of avalanches (34% for leukaemic cells *vs* 26% for healthy cells). This is a signature of a modification of the cytoskeleton architecture of these cancer cells, which show a greater mechanical fragility and a strengthening of their network reticulation.

What was really surprising from a physicist point of view was the statistics of the rupture events: all the physical quantities were not following a power-law distribution, typical of the theory of self-organised criticality and observed for avalanches of fractures in solid and amorphous materials (Song et al., 2013; Chuang et al., 2013; Salje et al., 2017), but were instead log-normally distributed. More precisely the distribution of the released energy was found to follow two overlapping log-normal distributions, one for the ductile regime and one for the brittle regime.

Motivated the experimental evidence of this atypical avalanche statistics we developed a first minimal model explaining the propagation mechanics of cascades of ruptures. This first model was based on a generalisation of a branching process, with a reproduction rate relaxing exponentially with time on average. Therefore the spreading of an avalanche was described as a front consisting of cross-linker detachments that can, at each time step, either trigger subsequent activity (new avalanches) or die out. Note that every singular event in the experimental curves can either be interpreted as a single avalanche or as the sum of several interacting avalanches, because the model includes the possibility to start with several simultaneous cross-linker unbindings. The final released energy of the avalanche was then the sum of all the energies released at each time step. It is commonly recognised that the sum of independent random variables results in a normally distributed random variable, however introducing correlations between the summed variables, through the temporal relaxation, we demonstrated that this distribution can be changed. Indeed, with our model we reproduced the same log-normal statistics as in experiments, and we highlighted the fundamental role of correlations in this process. The mechanistic description provided by our minimal rupture cascade model was applied here to leukaemic and healthy cells, but the same statistics was also observed in different cell lines studied in our group, such as myoblasts or myotubes (Streppa, 2017), and could be a characteristic of other soft glassy materials (Radhakrishnan et al., 2017). Besides confirming that brittle failures are more frequent in cancer leukaemic cells than in healthy ones, our model predicts that the rate of released energy (released energy per time step) increases abruptly (and even more abruptly for brittle avalanches) at the beginning of the avalanche and then relaxes exponentially. In retrospect, with our minimal model we could set the basis for modelling log-normal avalanches.

At this point the question whether this log-normal statistics is typical of living systems, or is a more general characteristic of the cross-linked network structure of glassy materials arose. Concerning the sum of random variables generated by a multiplicative process, which incidentally makes every individual variable approximately log-normally distributed, another more theoretical-statistical issue emerged. What exactly makes the distribution of the sum log-normal, instead of normally or differently distributed, and also what is relevant for causing the shift between the more common power-law distribution and the log-normal? Would it be possible to relate the avalanche size distribution to some intrinsic mechanisms of the multiplicative process, and in which way?

The latter question was addressed in a particular, but still quite general way in Chapter 3. In that chapter we analysed the avalanche size distribution for a branching process evolving with a continuous random reproduction rate (also called multiplicative factor, growth factor or branching ratio). The avalanche size distribution was studied with respect to the first (resp. the second) central moment, $\bar{a}$ (resp. $\tilde{a}$), of its distribution. Detailed results were obtained for a uniformly distributed reproduction rate, but the phenomenology of the results is the same as long as the distribution has a compact support (for example excluding heavy tailed distributions). More precisely all results are expected the same if the branching ratio distribution belongs at least to the symmetric exponentially bounded class of distribution (but the symmetry condition can be released if the distribution has compact support). In fact, the same behaviour described hereafter

was observed in simulations considering a normally distributed reproduction rate.

We showed that the sum Z of z_i values resulting from a multiplicative cascade (branching) process can give rise to different types of distributions and very rich behaviours, which depend on both parameters of the reproduction factor distribution. The global «phase plane» was then divided in 3 regions.

In a first region, in which both the average and the variance of Z converge, we found finite scale distributions with all finite moments. These distributions can be very close to log-normals depending on how fast the average of Z decreases with the number of iteration steps and therefore on how many individual variables z_i are important for the sum (after a certain N^* all the summed z_i become negligible). For very low values of N^* the distribution moves away from a log-normal showing two bumps separating normal size avalanches and very small size ($Z < 1$) avalanches, with comparable probabilities. In this region the resulting distributions were significantly shrunk when removing correlations, but their shape was basically preserved. The effect of correlation was then mainly to importantly broaden the distributions.

In the second region we observed power-law tailed distributions, which look very close to Lévy $\alpha-$stable distributions with a Paretian tail. The exponent of the tails varies from 3 to 1, by increasing $\tilde{a}$ and $\bar{a}$. For $\tilde{a} = 1$ the exponent is equal to 3 and does not depend on the value of $\bar{a}$, on the other hand, for $\bar{a} = 1$ the exponent is equal to 2 and does not depend on the value of $\tilde{a}$.

The third region showed a very robust statistical behaviour for Z, whose distribution was found to converge to an exact log-normal distribution. The velocity of convergence turned out to depend on the choice of the model parameters (the bigger $\bar{a}$ and $\tilde{a}$, the faster the convergence), thus the log-normal limit can be reached in experimentally reasonable time lengths N of the process, provided the system is large enough. Notably, as opposed to the two previous regions the distribution here is not stationary, since both parameters of the log-normal depend on N. However with a standardisation of the variable Z this dependence disappears and the asymptotic distribution is then well-defined. Notice that the critical lines separating the 3 regimes were computed analytically for a uniformly distributed reproduction rate.

We proved that correlations were important for all outcoming distributions, but were even crucial in the last two regions and could not be neglected. Indeed in the second and third region, by removing them we obtained strongly different distributions which respect the Central Limit Theorem. This means that correlations make impossible the use of the Central Limit Theorem, even considering the Lyapunov version (for not too different distributions of the addends) or a version including weakly correlated random variables. Note that by adding correlations we observed that the sum of log-normal random variables (the z_i of this model) gives to a log-normal random variable, and hence the log-normal distribution can be as well a signature of scale invariance of a system. In other words, a log-normally distributed extensive physical quantity at small scales can be also log-normally distributed at larger scales, if the subparts composing the system are correlated.

The randomness of the growth factor actually resumes all the aleatory processes that can rise in real systems and can be thought as an *effective* random reproduction rate. For

example, for epidemics it would contain the randomness of the contacts of an individual and the randomness of transmitting the infection. Remarkably, for the avalanche size to be bounded, not only the average reproduction rate $\bar{a}$ has to be lower than one, but also its second moment, $\tilde{a}$. Again with the example of the epidemics, this two conditions would be necessary for avoiding the epidemic to spread massively.

Notably, if the accessible physical quantity is Z, by plotting its distribution we can have some information about the often inaccessible parameters of the underlying avalanche, like the growth factor $\bar{a}$ and $\tilde{a}$. Moreover a value of the tail exponent of the Z distribution is uniquely related to a line in the parameter plane, whose coordinates can be analytically computed. Therefore if we have power-law distributed experimental avalanche data, by measuring with a suitable statistical technique the exponent of the tail, the parameters of the corresponding branching ratio distribution can be determined.

Another important result was that our model can provide an explanation of some of the power-law distributions arising in environmental and human systems. In the same way it could give another perspective in the interpretation of many observed fat tailed distributions in neural driven physiological systems, such as heart and brain (Ivanov et al., 1996, 1998; Lo et al., 2002). Most of their exponents (see for a list (Newman, 2005)) are in the range of the possible outcome exponents of the model, and many of them can actually be interpreted as the result of a sum of an underlying multiplicative process. In addition, this connects the power-law measured exponent to the growth factor of the multiplicative process, giving an immediate hint for changes in the exponent. An interesting line of investigation related to this model would be allowing the growth factor a to have negative values (which can make sense in some cases, like stock pricing), in order to interpret distributions with two different power-law regimes, as observed for stock prices. The same type of distribution was also proposed for USA family names and web hits distributions in (Newman, 2005), and could reproduce well also the abundance of species of birds distribution shown in the same reference. More generally it would be interesting to investigate in more detail, also from an analytical point of view, the results for different growth factor distributions and for example relate them to results in geophysics where the branching ratio follows a power-law distribution (Saichev et al., 2005).

We point out that this avalanche process can also be used as a guide for building a network whose degree distribution is the same as one of the possible avalanche size distributions uncovered in this study. Indeed, starting from a network of $\mathfrak{N}$ nodes, to each node we could simulate an avalanche process as the one described here (actually a discretised version of it, the number of connections being an integer number). Hence, by setting the suitable parameters needed for the desired network degree distribution, we could attribute to each node a number of connections equal to the result of the corresponding avalanche process. Ultimately, the growth of a network can itself be seen as an avalanche process, thinking for example of the preferential attachment rule (Albert et al., 1999). This model can henceforth be proposed as an interpretation and a generative model for the occurrence of not scale-free networks (Broido and Clauset, 2019).

Our study reported in Chapter 3 suggests that the reason why we have seen close to log-normal distributions for the size of the avalanches for the model in Chapter 2 and

(Polizzi et al., 2018), is related to the short duration of the avalanches. Actually the observed distributions were not exact log-normals and therefore could not be interpreted as outcome of the third region, even though deviations from log-normality could also be due to other reasons. Nevertheless this was consistent with the large width of the observed distributions and can be interpreted with an effective average reproduction rate lower than 1.

Later, in Chapter 4, we increased the dimensionality and the complexity of our avalanche modelling by introducing a network structure. In this way, we could address more precisely the question of what is the impact of the network structure on the avalanche size distribution. It turned out that we could give some arguments to exclude that the cytoskeleton organisation follows a scale free network (with a power-law degree distribution). Therefore we chose a Gilbert graph (which was for us equivalent to an Erdős-Rényi graph), to model the architecture of the cytoskeleton network. The most important property of this graph is to have a finite scale (close to Poisson) degree distribution. We identified the network nodes as the actin filaments and its links as the cross-linkers of the cytoskeleton network. On this structure we performed avalanche simulations with different propagation rules.

First, by using a breaking rule with a constant probability of breaking p, we modelled the avalanche as an instantaneous-elastic rupture process. The avalanche size statistics was found to follow three different types of distribution, none of them being log-normal. We obtained i) a *pop* regime, composed of small avalanches, whose size follows a distribution decaying faster than a power-law. ii) A *mixed* regime with avalanche sizes distributed following a distribution composed of both *pop* avalanches and large avalanches with a size scaling as the system size. Along the critical line between the *pop* and the *mixed* regime the avalanche size distribution follows a power-law with an exponential cut-off taking into account the finite size of the system. iii) An *explosive* regime composed only of avalanches scaling as the system size, and thus causing a global disruption of the network. Their distribution was very narrow and indistinguishable between a normal or a log-normal distribution.

Subsequently, we introduced in the model a rule taking into account the fractional visco-elastic relaxation observed experimentally in living cells and thus considering memories in the deformation of the system. This was the key ingredient for having approximately log-normal avalanches, which appeared instead of the *mixed* regime. The time correlations introduced in this way led to a kind of synchronisation of the avalanches, whose size has here become finite. At the same time the proportion of small *pop* avalanches in the region was reduced when decreasing the visco-elastic relaxation time. The same three regions (including the *log-normal bump* region) where found again when introducing in the model the local restoration of cross-linkers, while always keeping a visco-elastic relaxation of the breaking probability. The avalanche behaviour was therefore shown to be robust and not depending on the details of the particular avalanche propagating rule.

Our phenomenological model was based on a network structure not provided of a natural metric: this network was defined primarily as a structure of interactions. Beyond considering this characteristic as a limitation, it confers to the model a wider generality and richness of applications. Indeed, the network structure can always be embedded in a

metric space, such as a 3-D sphere, if desired, and the same results would remain valid.

Since it was a phenomenological model we did not try to extract from it in Chapter 4 any quantitative information. Nevertheless, that may be possible by introducing appropriate scaling factors to make the avalanche size a physical measurable quantity (as the released energy). At the same time the correspondence of the lasting time of the avalanche with the experimental one can be done. This would attribute a unit to τ, allowing for an estimation of the physical relaxation time (at the local level) needed to have log-normal avalanches. Altogether these results suggest that at least for cells the local fracture mechanism behaves in the same way as the global mechanical response of the cytoskeleton network. The global shear-relaxation modulus can thus be interpreted as a sequence of local avalanches of ruptures. Consequently, the observation of the log-normal distribution at different scales (locally for avalanche sizes distribution and globally for the shear-relaxation modulus) can be considered as a hint for scale invariance in the response mechanism of the polymer network. Surprisingly the scale invariance did not result in power-law distributions and this needs to be better understood. In any case note that our scale invariance is in the response mechanics and not in the avalanche size. Other log-normal distributions appearing in living systems (Loewenstein et al., 2011; Buzsáki and Mizuseki, 2014) could be related to this scale invariance and therefore to these log-normal type local avalanches (showing up as noise in the force indentation curves).

Conceptually, we note that in contrast to simpler (mean-field) 1-D models without a network structure as in Chapter 2 and 3, we did not find any region in the parameter space where the avalanche size distribution was exactly log-normal. This points out the role of the network structure which introduces, in an uncontrolled way, an additional effective time dependence on the probability of breaking $p(t)$, and then on the reproduction factor distribution. This effect is mainly due to the avalanche propagation on the network, since it changes the actual number of available links along its way, but also loops in the network (the path between two nodes may be not unique) could have an effect. The difference on the outcome distribution is therefore entirely due to the network structure and would disappear in a mean field 1-D model, like the one in Chapter 3

In reality, in Chapter 4 the avalanche size distribution in the *log-normal bump* region could as well be approximated by a skewed distribution other than the log-normal, such as the Gamma distribution. Either log-normal or Gamma distribution, this avalanche model, by finding non power-law, but still stationary and skewed, avalanches, constitutes an alternative to self-organized criticality models published in the literature. Notably, we demonstrated that a skewed distribution for the avalanche size can be originated from two main ingredients: the introduction of time correlations (in the form of visco-elastic relaxation) and the non-linear dynamics of the avalanche propagation. These are the same ingredients (but instead of time correlations it was there for phase correlations) which were identified to cause the appearance of skewed distributions for the variations in the heart rate signal in (Ivanov et al., 1996, 1998), behind which an avalanche process may also be latent.

As a more global discussion, it could have been interesting to model the observed avalanches of fractures in cells (but not only) by making our models closer to the mechanical system. For example each link of the network may have some specific mechanical (elastic

or visco-elastic) properties and/or a precise metrics can be introduced in the network, allowing for the study of the nodes displacements. This can be a future perspective, however it was not our main goal for several reasons. First, it would have implied some strong assumptions about the network structure and its organisation in a 3-D space in a living cell, while so far precise data about this are not available in the literature. Besides, it would have caused an increase in complexity which might have not helped in the identification of the most important and general elements leading to the observed behaviour in cells, or other soft glassy materials. Actually, the advantage of this work is its generality and thus it can be applied to completely different systems in nature for which a log-normal statistics is observed. In addition, it was more challenging for us to deal with the problem of the determination of the network structure of interactions, starting only from the avalanche size experimental observation, instead of introducing a rigid space structure for which there is no evidence.

Relating our discussion to the Introduction of this thesis (see Chapter 1) we can now give a more general view to the experimental observations in the context of complex systems. The complex system is the cytoskeleton network of the living cell and the environment is the AFM measuring instrument. At this scale the log-normal distribution appears as an emergent property not related to the individual characteristics of the components (here the cross-linkers). If we had not considered the collective behaviour the log-normal distribution may not have arisen. In particular the importance of correlations has been highlighted as a signature of cooperative effects and it led to consider the log-normal distribution as a signature of scale invariance in living systems.

In that *Middle Way*, standing between the physics of particles and the cosmological theories, there is the realm of randomness and incertitude, the realm of life.

Appendix A

Supplemental material of Chapter 2

Table A.1 – Distributions of energy released during ductile and brittle rupture events in normal and CML cells: simulations *versus* experiments. Characteristics ($\overline{N}$, σ_N, $\mu = \overline{log_{10}(E)}$, $\sigma = \sigma_{log_{10}E}$, $\overline{E}$, $\tilde{E}$) of the numerical cascades simulated with equation (2.24) for parameters values a_0, $\hat{a}$, $\Delta E_0 = 12\,\mathrm{k_B T}$, $\Delta E^* = 0\,\mathrm{k_B T}$ *versus* experimental data (Figures 2.7 **E** and **F**)

	Normal Cells	**CML Cells**
Ductile		
Model Parameters	$a_0 = 1.3$, $\hat{a} = 14$, $\Delta a = 0.44$	$a_0 = 1.3$, $\hat{a} = 15$, $\Delta a = 0.52$
Simulations	$\overline{N} = 14$, $\sigma_N = 5$ $\mu_1 = 2.11$, $\sigma_1 = 0.35$ $\overline{E}_1 = 178\,\mathrm{k}_B\mathrm{T}$, $\tilde{E}_1 = 129\,\mathrm{k}_B\mathrm{T}$	$\overline{N} = 12$, $\sigma_N = 6$ $\mu_1 = 2.10$, $\sigma_1 = 0.41$ $\overline{E}_1 = 197\,\mathrm{k}_B T$, $\tilde{E}_1 = 126\,\mathrm{k}_B\mathrm{T}$
Experiments	$\mu_1 = 2.14$, $\sigma_1 = 0.34$ $\overline{E}_1 = 188\,\mathrm{k}_B\mathrm{T}$, $\tilde{E}_1 = 138\,\mathrm{k}_B\mathrm{T}$	$\mu_1 = 2.15$, $\sigma_1 = 0.40$ $\overline{E}_1 = 216\,\mathrm{k}_B\mathrm{T}$, $\tilde{E}_1 = 141\,\mathrm{k}_B\mathrm{T}$
Brittle		
Model Parameters	$a_0 = 1.3$, $\hat{a} = 45$, $\Delta a = 0.36$	$a_0 = 1.3$, $\hat{a} = 54$, $\Delta a = 0.42$
Simulations	$\overline{N} = 39$, $\sigma_N = 12$ $\mu_2 = 2.84$, $\sigma_2 = 0.42$ $\overline{E}_2 = 1104\,\mathrm{k}_B\mathrm{T}$, $\tilde{E}_2 = 692\,\mathrm{k}_B\mathrm{T}$	$\overline{N} = 36$, $\sigma_N = 13$ $\mu_2 = 2.89$, $\sigma_2 = 0.53$ $\overline{E}_2 = 1635\,\mathrm{k}_B\mathrm{T}$, $\tilde{E}_2 = 776\,\mathrm{k}_B\mathrm{T}$
Experiments	$\mu_2 = 2.87$, $\sigma_2 = 0.41$ $\overline{E}_2 = 1158\,\mathrm{k}_B\mathrm{T}$, $\tilde{E}_2 = 741\,\mathrm{k}_B\mathrm{T}$	$\mu_2 = 2.89$, $\sigma_2 = 0.51$ $\overline{E}_2 = 1547\,\mathrm{k}_B\mathrm{T}$, $\tilde{E}_2 = 776\,\mathrm{k}_B\mathrm{T}$

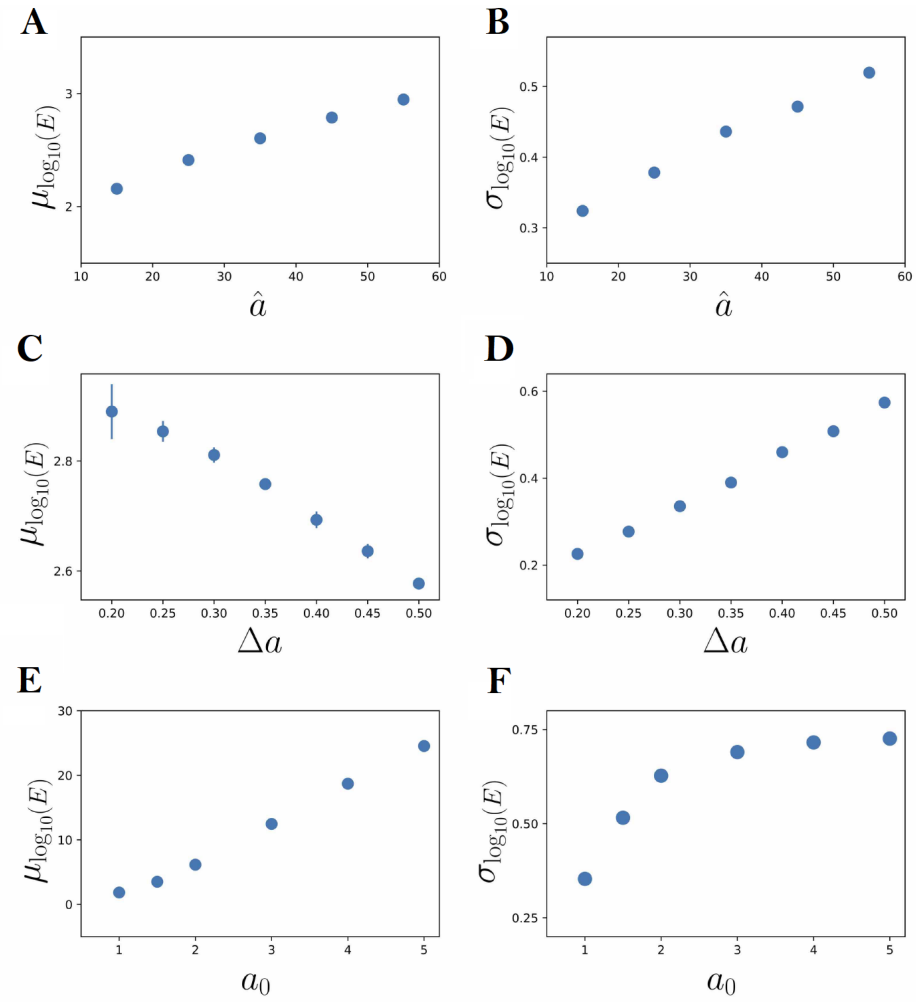

Figure A.1 – Dependence of the average number of steps N and its standard deviation on the model parameters parameters $\tilde{a}$, Δa and a_0. Fixed parameter values are: $\Delta E_0 = 12\,k_BT$, $\Delta E^* = 0$. **A.** $\overline{\log_{10}(E)}$ *vs* $\hat{a}$. **B.** $\sigma_{\log_{10}(E)}$ *vs* $\hat{a}$, for fixed $a_0 = 1.3$, $\Delta a = 0.4$. **C.** $\overline{\log_{10}(E)}$ *vs* Δa. **D.** $\sigma_{\log_{10}(E)}$ *vs* Δa, for fixed $a_0 = 1.3$, $\hat{a} = 40$. **E.** $\overline{\log_{10}(E)}$ *vs* a_0. **B.** $\sigma_{\log_{10}(E)}$ *vs* a_0, for fixed $\hat{a} = 40$, $\Delta a = 0.4$.

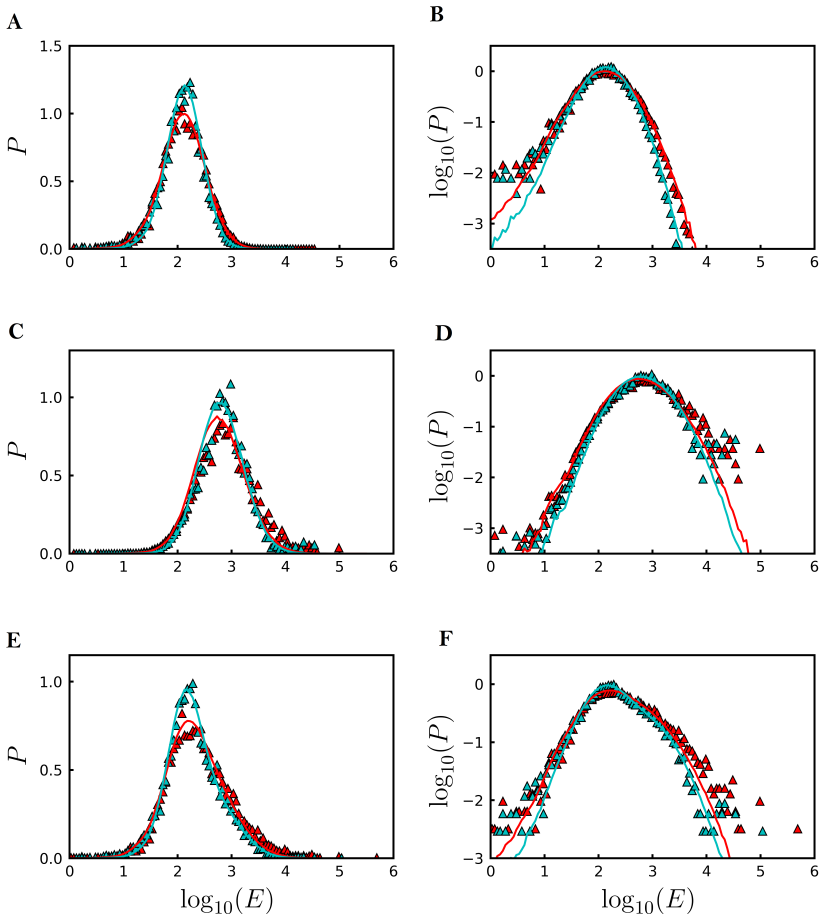

Figure A.2 – Computed model simulations of the p.d.fs $P(E)$ of energy (k_BT) released during rupture events in normal (blue) and CML (red) cells. Ductile rupture events: **A.** semi-log representation ; **B.** log-log representation. Brittle rupture events: **C.** semi-log representation ; **D.** log-log representation. All rupture events: **E.** semi-log representation ; **F.** log-log representation. The triangles (△) represent the experimental data (Figures 2.7 **E** and **F**); ductile and brittle rupture event log-normal p.d.fs were disentangled using a classical two-component Gaussian mixture model (see Eq. (2.22)). The curves represent the prediction of the released energy cacade model defined by equation (2.24); the corresponding sets of model parameters are given in Table A.1.

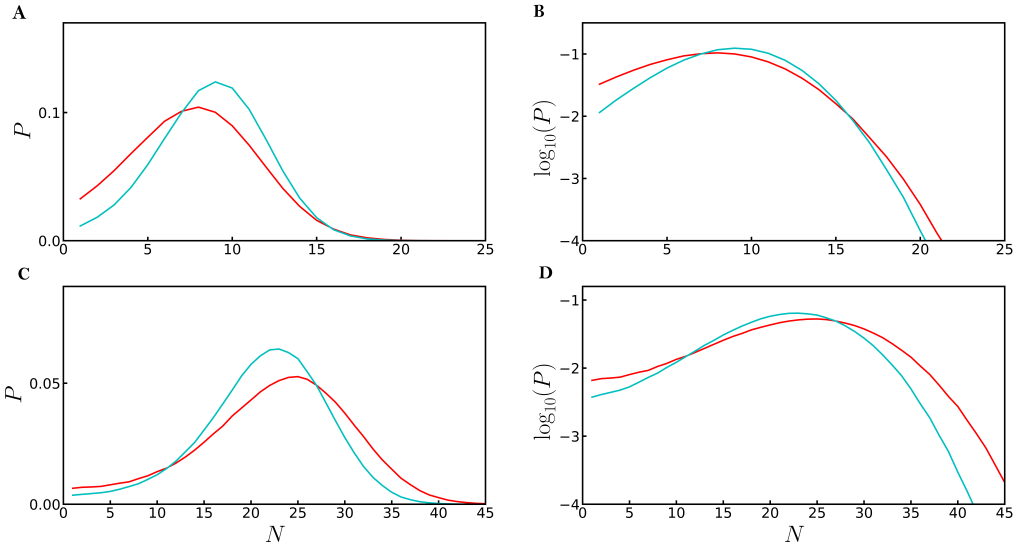

Figure A.3 – Distributions of the number of steps N of the log-normal rupture cascade model. P.d.f. of N computed from 1.2×10^6 realisations of the multiplicative process defined by equation (2.24), for parameters values given in Table 2.1. **A.** Ductile rupture events in normal (blue) and CML (red) cells. **B.** semi-log representation. **C.** Brittle rupture events in normal (blue) and CML (red) cells. **D.** semi-log representation.

Appendix B

Table of parameters for simulations of Chapter 3

	b	c	$\bar{a}$	$\tilde{a}$	N	n
Finite Scale						
	0	1	0.5	0.33	1000	10^4
	0.2	0.9	0.55	0.34	500	10^5
	0.5	1.2	0.85	0.76	500	10^5
	0.5	1.2	0.85	0.76	500	10^6
	0.57	1.23	0.9	0.85	500	10^5
Power Law	0.67	1.3	0.985	1	500	10^6
	0.68	1.3	0.99	1.01	500	10^6
	0.7	1.3	1	1.03	500	10^6
	0.72	1.3	1.01	1.05	500	10^6
Log-Normal	0.9	1.3	1.1	1.22	500	10^6

Bibliography

Abidine, Y., Laurent, V., Michel, R., Duperray, A. and Verdier, C. (2013), 'Microrheology of complex systems and living cells using afm', *Computer Methods in Biomechanics and Biomedical Engineering* **16**(suppl), 15–16.

Aitchison, J. and Brown, J. A. C. (1957), *The Log-normal Distribution*, Cambridge University Press, Cambridge, England, U.K.

Albert, R., Jeong, H. and Barabási, A.-L. (1999), 'Diameter of the world-wide web', *Nature* **401**(6749), 130–131.

Alford, P. W., Humphrey, J. D. and Taber, L. A. (2008), 'Growth and remodeling in a thick-walled artery model: effects of spatial variations in wall constituents', *Biomechanics and Modeling in Mechanobiology* **7**(4), 245.

Alstott, J., Bullmore, E. and Plenz, D. (2014), 'Powerlaw: A python package for analysis of heavy-tailed distributions', *PLoS ONE* **9**(1).

Amar, M. B. and Bianca, C. (2016), 'Towards a unified approach in the modeling of fibrosis: A review with research perspectives', *Physics of life reviews* **17**, 61–85.

Ambrosi, D., Ateshian, G. A., Arruda, E. M., Cowin, S., Dumais, J., Goriely, A., Holzapfel, G. A., Humphrey, J. D., Kemkemer, R., Kuhl, E. et al. (2011), 'Perspectives on biological growth and remodeling', *Journal of the Mechanics and Physics of Solids* **59**(4), 863–883.

Anderson, P. W. (1972), 'More is different', *Science* **177**(4047), 393–396.

Arneodo, A., Audit, B., Decoster, N., Muzy, J.-F. and Vaillant, C. (2002), Wavelet based multifractal formalism: applications to dna sequences, satellite images of the cloud structure, and stock market data, *in* 'The science of Disasters', Springer, Berlin, Heudelberg, DE, pp. 26–102.

Arneodo, A., Audit, B., Kestener, P. and Roux, S. (2008), 'Wavelet-based multifractal analysis', *Scholarpedia* **3**(3), 4103.

Arneodo, A., Bacry, E., Graves, P. and Muzy, J.-F. (1995), 'Characterizing long-range correlations in dna sequences from wavelet analysis', *Physical Review Letters* **74**(16), 3293.

Arneodo, A., Bacry, E. and Muzy, J. F. (1995), 'The thermodynamics of fractals revisited with wavelets', *Physica A: Statistical Mechanics and its Applications* **213**(1-2), 232–275.

Arneodo, A., Bacry, E. and Muzy, J.-F. (1998), 'Random cascades on wavelet dyadic trees', *Journal of Mathematical Physics* **39**(8), 4142–4164.

Arneodo, A., Vaillant, C., Audit, B., Argoul, F., d'Aubenton Carafa, Y. and Thermes, C. (2011), 'Multi-scale coding of genomic information: From dna sequence to genome structure and function', *Physics Reports* **498**(2-3), 45–188.

Azeloglu, E. U. and Costa, K. D. (2011), Atomic force microscopy in mechanobiology: Measuring microelastic heterogeneity of living cells, *in* 'Methods in Molecular Biology', Vol. 736, Humana Press Inc., pp. 303–329.

Bacry, E., Muzy, J. F. and Arnéodo, A. (1993), 'Singularity spectrum of fractal signals from wavelet analysis: Exact results', *Journal of Statistical Physics* **70**(3-4), 635–674.

Badre, D. and Wagner, A. D. (2006), 'Computational and neurobiological mechanisms underlying cognitive flexibility', *Proceedings of the National Academy of Sciences* **103**(18), 7186–7191.

Bak, P. (2013), *How nature works: the science of self-organized criticality*, Springer Science & Business Media, New York, NY, USA.

Barabási, A.-L. and Albert, R. (1999), 'Emergence of scaling in random networks', *Science* **286**(5439), 509–512.

Barabási, A.-L., Albert, R. and Jeong, H. (1999), 'Mean-field theory for scale-free random networks', *Physica A: Statistical Mechanics and its Applications* **272**(1-2), 173–187.

Barabási, A.-L. and Pósfai, M. (2016), *Network Science*, Cambridge University Press, Cambridge, England, U.K.

Barbour, B., Brunel, N., Hakim, V. and Nadal, J.-P. (2007), 'What can we learn from synaptic weight distributions?', *TRENDS in Neurosciences* **30**(12), 622–629.

Baró, J. and Davidsen, J. (2018), 'Universal avalanche statistics and triggering close to failure in a mean-field model of rheological fracture', *Physical Review E* **97**(3), 033002.

Beaulieu, N. C. (2012), 'An extended limit theorem for correlated lognormal sums', *IEEE Transactions on Communications* **60**, 23–26.

Beggs, J. M. and Plenz, D. (2003), 'Neuronal Avalanches in Neocortical Circuits', *Journal of Neuroscience* **23**(35), 11167–11177.

Berryman, A. A. (1992), 'The orgins and evolution of predator-prey theory', *Ecology* **73**(5), 1530–1535.

Black, F. and Scholes, M. (1973), 'The pricing of options and corporate liabilities', *Journal of Political Economy* **81**(3), 637–654.

Blanchoin, L., Boujemaa-Paterski, R., Sykes, C. and Plastino, J. (2014), 'Actin dynamics, architecture, and mechanics in cell motility', *Physiological Reviews* **94**(1), 235–263.

Bonakdar, N., Gerum, R., Kuhn, M., Spörrer, M., Lippert, A., Schneider, W., Aifantis, K. E. and Fabry, B. (2016), 'Mechanical plasticity of cells', *Nature Materials* **15**(10), 1090–1094.

Bonfanti, A., Kaplan, J. L., Charras, G. and Kabla, A. J. (2020), 'Fractional viscoelastic models for power-law materials', *Soft Matter* **16**, 6002–6020.

Bouchard, J.-P. and Potters, M. (2003), *Theory of financial risk and derivative pricing: From statistical physics to risk management*, Cambridge Univ. Press, New York, NY, USA.

Broedersz, C. P. and MacKintosh, F. C. (2014), 'Modeling semiflexible polymer networks', *Reviews of Modern Physics* **86**(3), 995.

Broedersz, C., Storm, C. and MacKintosh, F. (2008), 'Nonlinear elasticity of composite networks of stiff biopolymers with flexible linkers', *Physical Review Letters* **101**(11), 118103.

Broido, A. D. and Clauset, A. (2019), 'Scale-free networks are rare', *Nature Communications* **10**(1), 1–10.

Brunner, C., Niendorf, A. and Käs, J. A. (2009), 'Passive and active single-cell biomechanics: a new perspective in cancer diagnosis', *Soft Matter* **5**(11), 2171–2178.

Buzsáki, G. and Mizuseki, K. (2014), 'The log-dynamic brain: how skewed distributions affect network operations', *Nature Reviews Neuroscience* **15**(4), 264–278.

Cai, W., Chen, L., Ghanbarnejad, F. and Grassberger, P. (2015), 'Avalanche outbreaks emerging in cooperative contagions', *Nature Physics* **11**(11), 936–940.

Cappella, B. and Dietler, G. (1999), 'Force-distance curves by atomic force microscopy', *Surface Science Reports* **34**(1-3), 1–3.

Carmichael, B., Babahosseini, H., Mahmoodi, S. and Agah, M. (2015), 'The fractional viscoelastic response of human breast tissue cells', *Physical Biology* **12**(4), 046001.

Carreras, B. A., Lynch, V. E., Dobson, I. and Newman, D. E. (2002), 'Critical points and transitions in an electric power transmission model for cascading failure blackouts', *Chaos: An Interdisciplinary Journal of Nonlinear Science* **12**(4), 985–994.

Chauvière, A., Preziosi, L. and Verdier, C. (2010), *Cell mechanics: from single scale-based models to multiscale modeling*, CRC Press, Boca Raton, FL, USA.

Cheng, K., Kurzrock, R., Qiu, X., Estrov, Z., Ku, S., Dulski, K. M., Wang, J. Y. and Talpaz, M. (2002), 'Reduced focal adhesion kinase and paxillin phosphorylation in BCR-ABL-transfected cells', *Cancer* **95**(2), 440–450.

Cheng, Y.-T. and Cheng, C.-M. (2004), 'Scaling, dimensional analysis, and indentation measurements', *Materials Science and Engineering: R: Reports* **44**(4-5), 91–149.

Choi, S. W. (2016), 'Life is lognormal! What to do when your data does not follow a normal distribution', *Anaesthesia* **71**(11), 1363–1366.

Chuang, C.-P., Yuan, T., Dmowski, W., Wang, G.-Y., Freels, M., Liaw, P. K., Li, R. and Zhang, T. (2013), 'Fatigue-induced damage in zr-based bulk metallic glasses', *Scientific Reports* **3**, 2578.

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

Cross, R., Jackson, A., Citi, S., Kendrick-Jones, J. and Bagshaw, C. (1988), 'Active site trapping of nucleotide by smooth and non-muscle myosins', *Journal of Molecular Biology* **203**(1), 173–181.

Da Costa, V., Henry, Y., Bardou, F., Romeo, M. and Ounadjela, K. (2000), 'Experimental evidence and consequences of rare events in quantum tunneling', *The European Physical Journal B-Condensed Matter and Complex Systems* **13**(2), 297–303.

Daubechies, I. (1992), *Ten Lectures on Wavelets*, Society for Industrial and Applied Mathematics.

Davison, A. C. and Hinkley, D. V. (1997), *Bootstrap methods and their application*, Cambridge university press, New York, NY, USA, chapter 2.

Digiuni, S., Berne-Dedieu, A., Martinez-Torres, C., Szecsi, J., Bendahmane, M., Arneodo, A. and Argoul, F. (2015), 'Single cell wall nonlinear mechanics revealed by a multiscale analysis of AFM force-indentation curves', *Biophysical Journal* **108**(9), 2235–2248.

Dos Remedios, C., Chhabra, D., Kekic, M., Dedova, I., Tsubakihara, M., Berry, D. and Nosworthy, N. (2003), 'Actin binding proteins: regulation of cytoskeletal microfilaments', *Physiological Reviews* **83**(2), 433–473.

Dufresne, D. (2008), Sums of lognormals, *in* 'Actuarial Research Conference', pp. 1–6.

Duhem, P. (2016), *La théorie physique. Son objet, sa structure*, ENS éditions.

Dvir, H., Elbaz, I., Havlin, S., Appelbaum, L., Ivanov, P. C. and Bartsch, R. P. (2018), 'Neuronal noise as an origin of sleep arousals and its role in sudden infant death syndrome', *Science Advances* **4**(4), eaar6277.

Eghiaian, F., Rigato, A. and Scheuring, S. (2015), 'Structural, mechanical, and dynamical variability of the actin cortex in living cells', *Biophysical Journal* **108**(6), 1330–1340.

Ehrlicher, A. J., Krishnan, R., Guo, M., Bidan, C. M., Weitz, D. A. and Pollak, M. R. (2015), 'Alpha-actinin binding kinetics modulate cellular dynamics and force generation', *Proceedings of the National Academy of Sciences* **112**(21), 6619–6624.

Erdös, P. and Rényi, A. (1959), 'On random graphs i', *Publicationes Mathematicae (Debrecen)* **6**, 290–297.

Esue, O., Tseng, Y. and Wirtz, D. (2009), 'α-actinin and filamin cooperatively enhance the stiffness of actin filament networks', *PLoS ONE* **4**(2), e4411.

Ewing, J. (1928), 'Neoplastic disease', *A Treatment on Tumors* .

Fabry, B., Maksym, G. N., Butler, J. P., Glogauer, M., Navajas, D. and Fredberg, J. J. (2001), 'Scaling the microrheology of living cells', *Physical Review Letters* **87**(14), 148102.

Feller, W. (1971), *An Introduction to Probability Theory and Its Applications*, Vol. 2, Wiley, New York, NY, USA.

Fenton, L. (1960), 'The sum of log-normal probability distributions in scatter transmission systems', *IRE Transactions on Communications Systems* **8**(1), 57–67.

Ferrer, J. M., Lee, H., Chen, J., Pelz, B., Nakamura, F., Kamm, R. D. and Lang, M. J. (2008), 'Measuring molecular rupture forces between single actin filaments and actin-binding proteins', *Proceedings of the National Academy of Sciences of the United States of America* **105**(27), 9221–9226.

Feynman, R., Leighton, R. and Sands, M. (1966), *The Feynman Lectures on Physics*, Vol. I, Addison-Wesley, Reading, CA, USA, chapter 46.

Fisher, R. A. (1930), *The genetical theory of natural selection*, Oxford University Press, Oxford, UK.

Fletcher, D. A. and Mullins, R. D. (2010), 'Cell mechanics and the cytoskeleton', *Nature* **463**(7280), 485–492.

Frank, J. and Massey, J. (1951), 'The Kolmogorov-Smirnov test for goodness of fit', *Journal of the American Statistical Association* **46**(253), 68–78.

Fredberg, J. J., Inouye, D., Miller, B., Nathan, M., Jafari, S., HELIOUI RABOUDI, S., Butler, J. P. and Shore, S. A. (1997), 'Airway smooth muscle, tidal stretches, and dynamically determined contractile states', *American Journal of Respiratory and Critical Care Medicine* **156**(6), 1752–1759.

Frisch, U. and Kolmogorov, A. N. (1995), *Turbulence: the legacy of AN Kolmogorov*, Cambridge university press, Cambridge, NY, USA.

Gardel, M. L., Kasza, K. E., Brangwynne, C. P., Liu, J. and Weitz, D. A. (2008), 'Mechanical response of cytoskeletal networks', *Methods in Cell Biology* **89**, 487–519.

Gerasimova-Chechkina, E., Streppa, L., Schaeffer, L., Devin, A., Argoul, P., Arneodo, A. and Argoul, F. (2018), 'Fractional rheology of muscle precursor cells', *Journal of Rheology* **62**(6), 1347–1362.

Gerstein, G. L. and Mandelbrot, B. (1964), 'Random walk models for the spike activity of a single neuron', *Biophysical Journal* **4**(1), 41–68.

Gibrat, R. (1931), *Les inégalits économiques*, Librairie du Recueil Sirey, Paris, France.

Gilbert, E. N. (1959), 'Random graphs', *The Annals of Mathematical Statistics* **30**(4), 1141–1144.

Gnedenko, B., Kolmogorov, A., Gnedenko, B. and Kolmogorov, A. (1954), 'Limit distributions for sums of independent', *Am. J. Math* **105**.

Good, B. H., McDonald, M. J., Barrick, J. E., Lenski, R. E. and Desai, M. M. (2017), 'The dynamics of molecular evolution over 60,000 generations', *Nature* **551**(7678), 45–50.

Gorenflo, R., Loutchko, J. and Luchko, Y. (2002), Computation of the mittag-leffler function eα, β (z) and its derivative, *in* 'Fract. Calc. Appl. Anal', Citeseer.

Gralka, M. and Kroy, K. (2015), 'Inelastic mechanics: a unifying principle in biomechanics', *Biochimica et Biophysica Acta (BBA)-Molecular Cell Research* **1853**(11), 3025–3037.

Grassberger, P. (1983), 'On the critical behavior of the general epidemic process and dynamical percolation', *Mathematical Biosciences* **63**(2), 157–172.

Grossmann, A. and Morlet, J. (1984), 'Decomposition of Hardy Functions into Square Integrable Wavelets of Constant Shape', *SIAM Journal on Mathematical Analysis* **15**(4), 723–736.

Gutenberg, B. and Richter, C. F. (1942), 'Earthquake magnitude, intensity, energy, and acceleration', *Bulletin of the Seismological society of America* **32**(3), 163–191.

Gutenberg, B. and Richter, C. F. (1956), 'Earthquake magnitude, intensity, energy, and acceleration: (second paper)', *Bulletin of the Seismological Society of America* **46**(2), 105–145.

Haase, K. and Pelling, A. E. (2015), 'Investigating cell mechanics with atomic force microscopy', *Journal of The Royal Society Interface* **12**(104), 20140970.

Hamdan, M. (1971), 'The logarithm of the sum of two correlated log-normal variates', *Journal of the American Statistical Association* **66**(333), 105–106.

Harris, T. E. (1963), *The Theory of Branching Processes, Dover Phoneix Editions*, Springer - Verlag, Berlin, Germany.

Hartmann, N. (2012), *New ways of ontology*, Transaction Publishers, pp. 9–10.

Hasan, P. M., Sulaiman, N. A., Soleymani, F. and Akgül, A. (2019), 'The existence and uniqueness of solution for linear system of mixed volterra-fredholm integral equations in banach space', *AIMS Mathematics* **5**(1), 226–235.

Hertz, H. (1882), 'Über die berührung fester elastischer körper und über die härte [the contact of solid elastic bodies and their harnesses]', *Verhandlungen des Vereins zur Beförderung des Gewerbfleißes* pp. 449–463.

Higgins, M., Proksch, R., Sader, J. E., Polcik, M., Mc Endoo, S., Cleveland, J. and Jarvis, S. (2006), 'Noninvasive determination of optical lever sensitivity in atomic force microscopy', *Review of Scientific Instruments* **77**(1), 013701.

Hildebrandt, J. (1969), 'Comparison of mathematical models for cat lung and viscoelastic balloon derived by laplace transform methods from pressurevolume data', *The Bulletin of Mathematical Biophysics* **31**(4), 651–667.

Hoffman, B. D. and Crocker, J. C. (2009), 'Cell mechanics: dissecting the physical responses of cells to force', *Annual Review of Biomedical Engineering* **11**, 259–288.

Hoffman, B. D., Massiera, G., Van Citters, K. M. and Crocker, J. C. (2006), 'The consensus mechanics of cultured mammalian cells', *Proceedings of the National Academy of Sciences* **103**(27), 10259–10264.

Huber, F., Boire, A., Lopez, M. P. and Koenderink, G. (2015), 'Cytoskeletal crosstalk: when three different personalities team up', *Current Opinion in Cell Biology* **32**, 39–47.

Huber, F., Schnauß, J., Rönicke, S., Rauch, P., Müller, K., Fütterer, C. and Käs, J. (2013), 'Emergent complexity of the cytoskeleton: from single filaments to tissue', *Advances in Physics* **62**(1), 1–112.

Hutter, J. L. (2005), 'Comment on tilt of atomic force microscope cantilevers: effect on spring constant and adhesion measurements', *Langmuir* **21**(6), 2630–2632.

Huvet, M., Nicolay, S., Touchon, M., Audit, B., d'Aubenton Carafa, Y., Arneodo, A. and Thermes, C. (2007), 'Human gene organization driven by the coordination of replication and transcription', *Genome Research* **17**(9), 1278–1285.

Ising, E. (1925), 'Beitrag zur theorie des ferromagnetismus', *Zeitschrift für Physik* **31**(1), 253–258.

Isope, P. and Barbour, B. (2002), 'Properties of unitary granule cell → purkinje cell synapses in adult rat cerebellar slices', *Journal of Neuroscience* **22**(22), 9668–9678.

Ivanov, P. C., Rosenblum, M. G., Peng, C.-K., Mietus, J., Havlin, S., Stanley, H. E. and Goldberger, A. L. (1996), 'Scaling behaviour of heartbeat intervals obtained by wavelet-based time-series analysis', *Nature* **383**(6598), 323–327.

Ivanov, P. C., Rosenblum, M., Peng, C.-K., Mietus, J., Havlin, S., Stanley, H. and Goldberger, A. (1998), 'Scaling and universality in heart rate variability distributions', *Physica A: Statistical Mechanics and its Applications* **249**(1-4), 587–593.

James, W. (1890), 'The principles of psychology', *Holt and Company* .

Jia, F. and Amar, M. B. (2020), 'Scaling laws and snap-through events in indentation of perforated membranes', *Journal of the Mechanics and Physics of Solids* **135**, 103797.

Jin, X., Zhang, Y., Wang, F., Li, L., Yao, D., Su, Y. and Wei, Z. (2009), 'Departure headways at signalized intersections: A log-normal distribution model approach', *Transportation Research Part C: Emerging Technologies* **17**(3), 318–327.

Johnson, N. L., Kotz, S. and Balakrishnan, N. (1994), *Continuous univariate distributions*, Vol. 1, Wiley, New York, NY, USA, chapter 14.

Jones, D. (2010), *Neoplastic hematopathology: experimental and clinical approaches*, Springer Science & Business Media, New York, NY, USA.

Juelicher, F., Kruse, K., Prost, J. and Joanny, J.-F. (2007), 'Active behavior of the cytoskeleton', *Physics Reports* **449**(1-3), 3–28.

Jülicher, F., Ajdari, A. and Prost, J. (1997), 'Modeling molecular motors', *Reviews of Modern Physics* **69**(4), 1269–1281.

Kai, F., Laklai, H. and Weaver, V. M. (2016), 'Force matters: biomechanical regulation of cell invasion and migration in disease', *Trends in Cell Biology* **26**(7), 486–497.

Kelly, M. K., Vaudo, R. P., Phanse, V. M., Görgens, L., Ambacher, O. and Stutzmann, M. (1999), 'Large free-standing gan substrates by hydride vapor phase epitaxy and laser-induced liftoff', *Japanese Journal of Applied Physics* **38**(3A), L217.

Kesten, H. (1973), 'Random difference equations and Renewal theory for products of random matrices', *Acta Mathematica* **131**(1), 207–248.

Khalilgharibi, N., Fouchard, J., Asadipour, N., Barrientos, R., Duda, M., Bonfanti, A., Yonis, A., Harris, A., Mosaffa, P., Fujita, Y. et al. (2019), 'Stress relaxation in epithelial monolayers is controlled by the actomyosin cortex', *Nature Physics* **15**(8), 839–847.

Klaus, A., Yu, S. and Plenz, D. (2011), 'Statistical analyses support power law distributions found in neuronal avalanches', *PLoS ONE* **6**(5), e19779.

Kollmannsberger, P. and Fabry, B. (2011), 'Linear and nonlinear rheology of living cells', *Annual Review of Materials Research* **41**, 75–97.

Kollmannsberger, P., Mierke, C. T. and Fabry, B. (2011), 'Nonlinear viscoelasticity of adherent cells is controlled by cytoskeletal tension', *Soft Matter* **7**(7), 3127–3132.

Laperrousaz, B., Berguiga, L., Nicolini, F., Martinez-Torres, C., Arneodo, A., Satta, V. M. and Argoul, F. (2016), 'Revealing stiffening and brittling of chronic myelogenous leukemia hematopoietic primary cells through their temporal response to shear stress', *Physical Biology* **13**(3), 03LT01.

Laperrousaz, B., Drillon, G., Berguiga, L., Nicolini, F., Audit, B., Satta, V. M., Arneodo, A. and Argoul, F. (2016), From elasticity to inelasticity in cancer cell mechanics: A loss of scale-invariance, *in* 'AIP Conference Proceedings', Vol. 1760, AIP Publishing LLC, p. 020040.

Laperrousaz, B., Jeanpierre, S., Sagorny, K., Voeltzel, T., Ramas, S., Kaniewski, B., Ffrench, M., Salesse, S., Nicolini, F. E. and Maguer-Satta, V. (2013), 'Primitive CML cell expansion relies on abnormal levels of BMPs provided by the niche and on BMPRIb overexpression.', *Blood* **122**(23), 3767–3777.

Legrand, J., Grais, R. F., Boelle, P.-Y., Valleron, A.-J. and Flahault, A. (2007), 'Understanding the dynamics of ebola epidemics', *Epidemiology & Infection* **135**(4), 610–621.

Li, R. and Gundersen, G. G. (2008), 'Beyond polymer polarity: how the cytoskeleton builds a polarized cell', *Nature Reviews Molecular Cell Biology* **9**(11), 860–873.

Licata, I. (2008), Emergence and computation at the edge of classical and quantum systems, *in* 'Physics of emergence and organization', World Scientific, pp. 1–25.

Licata, I. (2011), *Complessità: un'introduzione semplice*, Duepunti.

Lieleg, O., Claessens, M. M. and Bausch, A. R. (2010), 'Structure and dynamics of cross-linked actin networks', *Soft Matter* **6**(2), 218–225.

Limpert, E., Stahel, W. A. and Abbt, M. (2001), 'Log-normal Distributions across the Sciences : Keys and Clues', **51**(5), 341–352.

Lo, C.-C., Amaral, L. N., Havlin, S., Ivanov, P. C., Penzel, T., Peter, J.-H. and Stanley, H. E. (2002), 'Dynamics of sleep-wake transitions during sleep', *EPL (Europhysics Letters)* **57**(5), 625.

Lo, C.-C., Chou, T., Penzel, T., Scammell, T. E., Strecker, R. E., Stanley, H. E. and Ivanov, P. C. (2004), 'Common scale-invariant patterns of sleep–wake transitions across mammalian species', *Proceedings of the National Academy of Sciences* **101**(50), 17545–17548.

Loewenstein, Y., Kuras, A. and Rumpel, S. (2011), 'Multiplicative dynamics underlie the emergence of the log-normal distribution of spine sizes in the neocortex in vivo', *Journal of Neuroscience* **31**(26), 9481–9488.

Lorenz, E. N. (1963), 'Deterministic nonperiodic flow„ vol. 20, no', *Journal of the Atmospheric Sciences* **20**(2), 130–141.

Mahaffy, R. E., Park, S., Gerde, E., Käs, J. and Shih, C. K. (2004), 'Quantitative Analysis of the Viscoelastic Properties of Thin Regions of Fibroblasts Using Atomic Force Microscopy', *Biophysical Journal* **86**(3), 1777–1793.

Malevergne, Y., Pisarenko, V. and Sornette, D. (2011), 'Testing the Pareto against the lognormal distributions with the uniformly most powerful unbiased test applied to the distribution of cities', *Physical Review E - Statistical, Nonlinear, and Soft Matter Physics* **83**(3).

Mallat, S. (1999), *A wavelet tour of signal processing*, Elsevier, New York, NY, USA.

Mandelbrot, B. (1960), 'The pareto-levy law and the distribution of income', *International Economic Review* **1**(2), 79–106.

Mandelbrot, B. B. (1982), *The Fractal Geometry of Nature*, WH Freeman, New York, USA, pp. 394–397.

McCulloch, W. S. and Pitts, W. (1943), 'A logical calculus of the ideas immanent in nervous activity', *The Bulletin of Mathematical Biophysics* **5**(4), 115–133.

McWhirter, J. and Wang, J. (1993), 'An actin-binding function contributes to transformation by the Bcr-Abl oncoprotein of Philadelphia chromosome-positive human leukemias.', *The EMBO Journal* **12**(4), 1533–1546.

Meyer, G. and Amer, N. M. (1988), 'Erratum: Novel optical approach to atomic force microscopy [appl. phys. lett. 5 3, 1045 (1988)]', *Applied Physics Letters* **53**(24), 2400–2402.

Meyers (Eds), R. A. (2009), *Encyclopedia of complexity and systems science*, Springer, New York, NY, USA.

Milevsky, M. A. and Posner, S. E. (1998), 'Asian options, the sum of lognormals, and the reciprocal gamma distribution', *Journal of Financial and Quantitative Analysis* **33**(3), 409–422.

Misiti, M., Misiti, Y., Oppenheim, G. and Poggi (Eds.), J.-M. (2007), *Wavelets and their Applications*, Vol. 330, ISTE, London, UK, pp. 363–379.

Mitzenmacher, M. (2004), 'A brief history of generative models for power law and lognormal distributions', *Internet Mathematics* **1**(2), 226–251.

Moran, P. A. P. (1958), Random processes in genetics, *in* 'Mathematical proceedings of the cambridge philosophical society', Vol. 54, Cambridge University Press, pp. 60–71.

Morris, V. J., Kirby, A. R., Gunning, A. P. et al. (1999), *Atomic force microscopy for biologists*, Vol. 57, World Scientific.

Mould, R. F., Lahanas, M., Asselain, B., Brewster, D., Burgers, S. A., Damhuis, R. A., De Rycke, Y., Gennaro, V. and Szeszenia-Dabrowska, N. (2004), 'Methodology for lognormal modelling of malignant pleural mesothelioma survival time distributions: a study of 5580 case histories from europe and usa', *Physics in Medicine & Biology* **49**(17), 3991.

Mould, R. F., Lederman, M., Tai, P. and Wong, J. K. (2002), 'Methodology to predict long-term cancer survival from short-term data using Tobacco Cancer Risk and Absolute Cancer Cure models', *Physics in Medicine and Biology* **47**(22), 3893–3924.

Mouri, H. (2013), 'Log-normal distribution from a process that is not multiplicative but is additive', *Physical Review E* **88**(4), 042124.

Müller, M. and Wyart, M. (2015), 'Marginal stability in structural, spin, and electron glasses', *Annu. Rev. Condens. Matter Phys.* **6**(1), 177–200.

Muzy, J. F., Bacry, E. and Arneodo, A. (1991), 'Wavelets and multifractal formalism for singular signals: Application to turbulence data', *Physical Review Letters* **67**(25), 3515–3518.

Muzy, J. F., Bacry, E. and Arneodo, A. (1994), 'The multifractal formalism revisited with wavelets', *International Journal of Bifurcation and Chaos* **04**(02), 245–302.

Naus, J. (1969), 'The distribution of the logarithm of the sum of two log-normal variates', *Journal of the American Statistical Association* **64**(326), 655–659.

Newman, M. E. (2002), 'Spread of epidemic disease on networks', *Physical Review E* **66**(1), 016128.

Newman, M. E. (2005), 'Power laws, Pareto distributions and Zipf's law', *Contemporary Physics* **46**(5), 323–351.

Newman, M. E. J., Strogatz, S. H. and Watts, D. J. (2001), 'Random graphs with arbitrary degree distributions and their applications', *Physical Review E* **64**, 1–17.

Olmeda, F. and Amar, M. B. (2019), 'Clonal pattern dynamics in tumor: the concept of cancer stem cells', *Scientific Reports* **9**(1), 1–18.

Orr, A. W., Helmke, B. P., Blackman, B. R. and Schwartz, M. A. (2006), 'Mechanisms of mechanotransduction', *Developmental Cell* **10**(1), 11–20.

Parry, B. R., Surovtsev, I. V., Cabeen, M. T., O'Hern, C. S., Dufresne, E. R. and Jacobs-Wagner, C. (2014), 'The bacterial cytoplasm has glass-like properties and is fluidized by metabolic activity', *Cell* **156**(1-2), 183–194.

Pastor-Satorras, R. and Vespignani, A. (2001), 'Epidemic spreading in scale-free networks', *Physical Review Letters* **86**(14), 3200.

Peinado, H., Zhang, H., Matei, I. R., Costa-Silva, B., Hoshino, A., Rodrigues, G., Psaila, B., Kaplan, R. N., Bromberg, J. F., Kang, Y. et al. (2017), 'Pre-metastatic niches: organ-specific homes for metastases', *Nature Reviews Cancer* **17**(5), 302–317.

Penson, K. and Górska, K. (2010), 'Exact and explicit probability densities for one-sided lévy stable distributions', *Physical Review Letters* **105**(21), 210604.

Perez-Reche, F. J. (2017), Modeling avalanches in martensites, *in* 'Avalanches in Functional Materials and Geophysics', Springer International Publishing, Dordrecht, Netherlands, chapter 6, pp. 99–136.

Perez-Reche, F. J., Triguero, C., Zanzotto, G. and Truskinovsky, L. (2016), 'Origin of scale-free intermittency in structural first-order phase transitions', *Physical Review B* **94**(14), 144102.

Pérez-Reche, F. J. and Vives, E. (2003), 'Finite-size scaling analysis of the avalanches in the three-dimensional gaussian random-field ising model with metastable dynamics', *Physical Review B* **67**(13), 134421.

Pessa, E. (2008), Phase transitions in biological matter, *in* 'Physics of emergence and organization', World Scientific, pp. 165–228.

Poincaré, H. (1890), 'Sur le problème des trois corps et les équations de la dynamique', *Acta Mathematica* **13**(1), A3–A270.

Polizzi, G. (2013), 'Aspetti filosofici dell'opera di Poincaré', *Lettera Matematica Pristem* **2013**(84-85), 66–79.

Polizzi, S., Laperrousaz, B., Perez-Reche, F., Nicolini, F., Satta, V., Arneodo, A. and Argoul, F. (2018), 'A minimal rupture cascade model for living cell plasticity', *New Journal of Physics* **20**(5), 053057.

Purcell, E. M. (1977), 'Life at low reynolds number', *American Journal of Physics* **45**(1), 3–11.

Radhakrishnan, R., Divoux, T., Manneville, S. and Fielding, S. M. (2017), 'Understanding rheological hysteresis in soft glassy materials', *Soft Matter* **13**(9), 1834–1852.

Radmacher, M. (2002), '4.-measuring the elastic properties of living cells by the atomic force microscope', *Methods in Cell Biology* **68**(1), 67–90.

Redner, S. (1990), 'Random multiplicative processes: An elementary tutorial', *American Journal of Physics* **58**(3), 267–273.

Ribeiro, T. L., Copelli, M., Caixeta, F., Belchior, H., Chialvo, D. R., Nicolelis, M. A. and Ribeiro, S. (2010), 'Spike avalanches exhibit universal dynamics across the sleep-wake cycle', *PloS one* **5**(11), e14129.

Rigato, A., Miyagi, A., Scheuring, S. and Rico, F. (2017), 'High-frequency microrheology reveals cytoskeleton dynamics in living cells', *Nature Physics* **13**(8), 771–775.

Ritort, F. and Sollich, P. (2003), 'Glassy dynamics of kinetically constrained models', *Advances in Physics* **52**(4), 219–342.

Romeo, M., Da Costa, V. and Bardou, F. (2003), 'Broad distribution effects in sums of lognormal random variables', *The European Physical Journal B-Condensed Matter and Complex Systems* **32**(4), 513–525.

Rumpel, S., Hatt, H. and Gottmann, K. (1998), 'Silent synapses in the developing rat visual cortex: evidence for postsynaptic expression of synaptic plasticity', *Journal of Neuroscience* **18**(21), 8863–8874.

Sabhapandit, S., Shukla, P. and Dhar, D. (2000), 'Distribution of avalanche sizes in the hysteretic response of the random-field ising model on a bethe lattice at zero temperature', *Journal of Statistical Physics* **98**(1-2), 103–129.

Sader, J. E. (1998), 'Frequency response of cantilever beams immersed in viscous fluids with applications to the atomic force microscope', *Journal of Applied Physics* **84**(1), 64–76.

Sader, J. E., Chon, J. W. and Mulvaney, P. (1999), 'Calibration of rectangular atomic force microscope cantilevers', *Review of Scientific Instruments* **70**(10), 3967–3969.

Sader, J. E., Lu, J. and Mulvaney, P. (2014), 'Effect of cantilever geometry on the optical lever sensitivities and thermal noise method of the atomic force microscope', *Review of Scientific Instruments* **85**(11), 113702.

Saichev, A., Helmstetter, A. and Sornette, D. (2005), 'Power-law distributions of offspring and generation numbers in branching models of earthquake triggering', *Pure and Applied Geophysics* **162**(6-7), 1113–1134.

Salje, E. K., Saxena, A. and Planes, (Eds), A. (2017), *Avalanches in Functional Materials and Geophysics*, Springer Int. Publ., Cham, Switzerland.

Sander, L., Warren, C., Sokolov, I., Simon, C. and Koopman, J. (2002), 'Percolation on heterogeneous networks as a model for epidemics', *Mathematical Biosciences* **180**(1-2), 293–305.

Savona, M. and Talpaz, M. (2008), 'Getting to the stem of chronic myeloid leukaemia', *Nature Reviews Cancer* **8**(5), 341–350.

Sayer, R., Friedlander, M. and Redman, S. (1990), 'The time course and amplitude of epsps evoked at synapses between pairs of ca3/ca1 neurons in the hippocampal slice', *Journal of Neuroscience* **10**(3), 826–836.

Serfling, R. J. (2002), *Approximation theorems of mathematical statistics*, 1st edn, John Wiley & Sons, New York, NY, USA, p. 27.

Sethna, J. P., Dahmen, K., Kartha, S., Krumhansl, J. A., Roberts, B. W. and Shore, J. D. (1993), 'Hysteresis and hierarchies: Dynamics of disorder-driven first-order phase transformations', *Physical Review Letters* **70**(21), 3347.

Severini, T. (2001), *Likelihood Methods in Statistics*, Oxford Univ. Press, New York, NY, USA, pp. 105–137.

Sharon, E. and Fineberg, J. (1996), 'Microbranching instability and the dynamic fracture of brittle materials', *Physical Review B* **54**(10), 7128.

Sheth, R. K. (1996), 'Galton-watson branching processes and the growth of gravitational clustering', *Monthly Notices of the Royal Astronomical Society* **281**(4), 1277–1289.

Sirghi, L., Ponti, J., Broggi, F. and Rossi, F. (2008), 'Probing elasticity and adhesion of live cells by atomic force microscopy indentation', *European Biophysics Journal* **37**(6), 935–945.

Sneddon, I. N. (1965), 'The relation between load and penetration in the axisymmetric boussinesq problem for a punch of arbitrary profile', *International Journal of Engineering Science* **3**(1), 47–57.

Sollich, P. (1998), 'Rheological constitutive equation for a model of soft glassy materials', *Physical Review E* **58**(1), 738.

Song, S., Sjöström, P. J., Reigl, M., Nelson, S. and Chklovskii, D. B. (2005), 'Highly nonrandom features of synaptic connectivity in local cortical circuits', *PLoS Biol* **3**(3), e68.

Song, Y., Chen, X., Dabade, V., Shield, T. W. and James, R. D. (2013), 'Enhanced reversibility and unusual microstructure of a phase-transforming material', *Nature* **502**(7469), 85–88.

Sornette, D. (2006), *Critical phenomena in natural sciences: chaos, fractals, selforganization and disorder: concepts and tools*, Springer Science & Business Media, New York, NY, USA.

Stehmann, C. and De Waard, M. A. (1996), 'Sensitivity of populations ofbotrytis cinerea to triazoles, benomyl and vinclozolin', *European Journal of Plant Pathology* **102**(2), 171–180.

Streppa, L. (2017), Characterizing mechanical properties of living C2C12 myoblasts with single cell indentation experiments: application to Duchenne muscular dystrophy, PhD thesis, Lyon.

Streppa, L., Ratti, F., Goillot, E., Devin, A., Schaeffer, L., Arneodo, A. and Argoul, F. (2018), 'Prestressed cells are prone to cytoskeleton failures under localized shear strain: an experimental demonstration on muscle precursor cells', *Scientific Reports* **8**(1), 1–16.

Sutton, J. (1997), 'Gibrat's legacy', *Journal of Economic Literature* **35**(1), 40–59.

Trepat, X., Grabulosa, M., Buscemi, L., Rico, F., Farré, R. and Navajas, D. (2005), 'Thrombin and histamine induce stiffening of alveolar epithelial cells', *Journal of Applied Physiology* **98**(4), 1567–1574.

Tseng, Y., Kole, T. P., Lee, J. S., Fedorov, E., Almo, S. C., Schafer, B. W. and Wirtz, D. (2005), 'How actin crosslinking and bundling proteins cooperate to generate an enhanced cell mechanical response', *Biochemical and Biophysical Research Communications* **334**(1), 183–192.

Varshney, L. R., Sjöström, P. J. and Chklovskii, D. B. (2006), 'Optimal information storage in noisy synapses under resource constraints', *Neuron* **52**(3), 409–423.

Wagner, B., Tharmann, R., Haase, I., Fischer, M. and Bausch, A. R. (2006), 'Cytoskeletal polymer networks: the molecular structure of cross-linkers determines macroscopic properties', *Proceedings of the National Academy of Sciences* **103**(38), 13974–13978.

Winkelman, J. D., Suarez, C., Hocky, G. M., Harker, A. J., Morganthaler, A. N., Christensen, J. R., Voth, G. A., Bartles, J. R. and Kovar, D. R. (2016), 'Fascin-and α-actinin-bundled networks contain intrinsic structural features that drive protein sorting', *Current Biology* **26**(20), 2697–2706.

Woehlke, G., Ruby, A. K., Hart, C. L., Ly, B., Hom-Booher, N. and Vale, R. D. (1997), 'Microtubule interaction site of the kinesin motor', *Cell* **90**(2), 207–216.

Wolff, L., Fernandez, P. and Kroy, K. (2010), 'Inelastic mechanics of sticky biopolymer networks', *New Journal of Physics* **12**(5), 053024.

World Health Organization (2020), 'Coronavirus disease 2019 (COVID-19) situation report–51', https://www.who.int/docs/default-source/coronaviruse/situationreports/20200311-sitrep-51-covid-19.pdf?sfvrsn=1ba62e57_10. Accessed 2020-08-07.

Wright, S. (1931), 'Evolution in mendelian populations', *Genetics* **16**(2), 97–159.

Wu, J., Mehta, N. B. and Zhang, J. (2005), Flexible lognormal sum approximation method, *in* 'GLOBECOM'05. IEEE Global Telecommunications Conference, 2005.', Vol. 6, IEEE, pp. 3413–3417.

Titre : Émergence de distributions log-normales dans des processus d'avalanche, validation de modèles stochastiques 1D et sur réseaux aléatoires, avec une application à la caractérisation de la plasticité des cellules cancéreuses

Résumé : Plusieurs matériaux vitreux ont des comportements caractéristiques suite à fractures induites par des contraintes. Ces fractures se présentent comme des processus d'avalanche dont la statistique dans la plus part des cas suit une loi de puissance, rappel de comportements collectifs et critiques auto-organisés. Des avalanches de fractures sont observées aussi dans des systèmes vivants, qui peuvent être vus dans certains cas comme un réseau vitreux, avec une structure figée soutenue par le cytosquelette (CSK). Des expériences ont montré que les cellules répondent à des contraintes extérieures par cascades d'événements aléatoires de ruptures, en suggérant qu'elles se comportent comme des réseaux aléatoires quasi-rigides de filaments interconnectés. Étonnement, la distribution de la taille de ces cascades, ne suit pas une distribution en loi de puissance typique de phénomènes critiques, mais au contraire s'avère être log-normale. Dans le but de donner une interprétation de ce comportement particulier nous proposons d'abord un modèle stochastique minimal (1D). Ce modèle donne une interprétation de l'énergie relâchée dans les cascades de ruptures, au regard d'une somme (étant l'énergie additive) d'un processus multiplicatif de cascade avec une relaxation temporelle. Nous identifions 2 types d'événements de ruptures : des fractures friables susceptibles de représenter des ruptures irréversibles dans un CSK rigide et très connecté, et des fractures ductiles résultant des décrochements dynamiques des cross-linkers pendant la déformation plastique sans perte d'intégrité du CSK. Entre autre notre modèle montre que les fractures friables sont relativement plus importantes dans les cellules leucémiques, en témoignant leur plus grande fragilité et leur différente architecture du CSK, plus rigide et réticulée. Ce modèle minimal motive la question plus générale de quelles sont les distributions résultantes pour la somme de variables corrélées provenant d'un processus multiplicatif. En conséquence nous analysons la distribution de la somme d'un processus de branchement généralisé évoluant avec un facteur de croissance aléatoire continu. Le processus dépend de 2 paramètres : les 2 premiers moments centrés de la distribution du facteur de croissance. Nous créons un diagramme de phase en montrant 3 régions différentes : une région où la distribution finale a tous les moments finis et qui est approximativement log-normale. 2) Une région où la distribution asymptotique est une loi de puissance, avec un exposant inclus dans l'intervalle [1 ;3], dont la valeur est déterminée par les paramètres du modèle. 3) Enfin dans la dernière région une distribution exactement log-normale, mais non-stationnaire. Dans tous les cas, les corrélations se révèlent fondamentales. Nous proposons ensuite un modèle de réseau aléatoire Erdös-Rényi pour modéliser le CSK, en identifiant le nœuds en tant que filaments d'actine et les liens en tant que cross-linkers. Sur cette structure nous simulons la propagation d'avalanches de ruptures. Nos simulations montrent que l'on peut reproduire une statistique log-normale avec deux concepts simples : un réseau aléatoire sans échelle d'espace caractéristique et une règle de rupture capturant la visco-élasticité des cellules. Ce travail ouvre la voie pour des applications futures à plusieurs phénomènes dans les systèmes vivants qui contiennent de larges populations d'éléments individuels, non-linéaires (cerveau, cœur, épidémies), où des statistiques log-normales similaires ont été observées.

Mots clés : Modelisation multi-échelle, statistique d'avalanche, processus stochastiques, réseaux, micro-environnement cellulaire, cancer, distribution log-normale

Abstract: Many glassy and amorphous materials, like martensites, show characteristic behaviours during constraint induced fractures. These fractures are avalanche processes whose statistics is known to follow in most cases a power-law distribution, reminding of collective behaviour and self-organised criticality. Avalanches of fractures are observed as well in living systems which, if we do not consider active remodelling, can be seen as a glassy network, with a frozen structure sustained by the cytoskeleton (CSK). Experiments revealed that cells respond to external constraints by a cascade of random and abrupt ruptures of their CSK, suggesting that they behave as a quasi-rigid random network of intertwined filaments. Surprisingly, the distribution of the size a of these rupture events do not follow the power-law statistics typical of critical phenomena. In fact, the avalanche size turns out to be log-normal, suggesting that the mechanics of living systems in catastrophic events would not fit into self-organised critical systems (power-laws). In order to give an interpretation of this peculiar behaviour we first propose a minimal (1D) stochastic model. This model gives an interpretation of the energy released along the rupture events, in terms of the sum (being energy additive) of a multiplicative cascade process relaxing with time. We distinguish 2 types of rupture events, brittle failures likely corresponding to irreversible ruptures in a stiff and highly cross-linked CSK and ductile failures resulting from dynamic cross-linker unbindings during plastic deformation without loss of CSK integrity. We also show that brittle failures are relatively more prominent in leukemic than in healthy cells, suggesting their greater fragility and their different CSK architecture, stiffer and more reticulated. This minimal model motivates the more general question of what are the resulting distributions of a sum of correlated random variables coming from a multiplicative process. Therefore, we analyse the distribution of the sum of a generalised branching process evolving with a continuous random reproduction (growth) rate. The process depends only on 2 parameters: the first 2 central moments of the reproduction rate distribution. We then create a phase diagram showing 3 different regions: 1) a region where the final distribution has all central moments finite and is approximately log-normal. 2) A region where the asymptotic distribution is a power-law, with a decay exponent belonging to the interval [1;3], whose value is determined by the model parameters. 3) Finally, we found an exact log-normal size, non-stationary, distribution region. In all cases correlations are fundamental. Increasing the level of complexity for avalanche modelling, we propose then a random Erdös-Rényi network to model a cell CSK, identifying the network nodes as the actin filaments, and its links as actin cross-linkers. On this structure we simulate avalanches of ruptures. Our simulations show that we can reproduce the log-normal statistics with two simple ingredients: a random network without characteristic length scale, and a breaking rule capturing the observed visco-elasticity of living cells. This work paves the way for future applications to many phenomena in living systems that include large populations of individual, non-linear, elements (brain, heart, epidemics) where similar log-normal statistics have also been observed.

Keywords: Multi-scale analysis, avalanche statistics, stochastic processes, networks, log-normal distribution, cancer, cell micro-environment

Laboratoire Ondes et Matière d'Aquitaine (LOMA) - UMR 5798 CNRS/Univ. de Bordeaux, 351, Cours de la Libération, Bât.A4N, 33400 Talence