Universitat Autònoma de Barcelona

Doctoral Thesis

On the Scale Invariance of certain Complex Systems

Author:
Francesc Font Clos

Supervisor:
Dr. Álvaro Corral Cano

Tutor:
Dr. Pere Puig Casado

Programa de Doctorat en Matemàtiques
Departament de Matemàtiques
Universitat Autònoma de Barcelona

Grup de Sistemes Complexos
Centre de Recerca Matemàtica

– 2015 –

Acknowledgements

I would like to thank all those who contributed, in some or another way, to the successful completion of this Thesis. I have been lucky enough to meet a lot of interesting people during these four years, and everybody has always been so kind to me, that I can hardly find the right words to express my gratitude.

My most sincere thanks to my supervisor Álvaro Corral, without whom this Thesis would not have been possible: Álvaro, gracias por todo! Gracias por guiarme en esta aventura, por dejarme toda la libertad del mundo, por aguantarme en tu despacho tantas horas, y por animarme incondicionalmente en todo momento.

Then I'd like to mention others who also guided me into and through academia. Josep Maria Mondelo introduced me to the wonders of the command line, `C` coding, `gnuplot` and `Vim` when I was an undergraduate at UAB. I didn't know it at the time, but these have been the most useful tools I've ever learned. Thanks also to Isaac Pérez Castillo, my first mentor at King's College London, with whom I experienced for the first time the thrill of doing research; he then encouraged me to do a PhD, and I would like to thank him for that as well. Many thanks as well to Anna Deluca, with whom I first shared a desk, and who welcomed me in the nicest possible way in the Complex Systems Group at CRM. Anna, moltes gràcies! Gràcies per ensenyar-me tantes, tantes coses; gràcies per la teva infinita paciència i per mostrar-me sempre el teu suport.

I am grateful to the Centre de Recerca Matemática for providing a unique research environment, and for funding me in the first place. My deepest thanks to Nicholas R. Moloney and Gunnar Pruessner, who have always welcomed me in London, and with whom doing research is the most pleasant thing. And thanks as well to Martin Gerlach and Eduardo G. Altmann for having me, and to the rest of the Dynamical Systems and Social Dynamics group at the Max Planck Institute for Complex Systems in Dresden; and to Thorsten Rheinländer and Friedrich Hubalek at the Vienna University of Technology.

I gràcies a tota la gent de bioquímica, i altres bio-coses, per aguantar-me a l'hora de dinar –i a altres hores: Rita, Marín, Pulido, Paula, Pol, Serra... i Gisela, Jofre, Laia i Javi, gràcies per tot! I als amics de sempre, Sebas, Jordi i companyia, i a la Sílvia.

I mil gràcies a la meva família, al meu germà, i a la meva mare, i al meu pare; doncs ells m'ho han donat tot, i a ells tot els hi dec.

Contents

CHAPTER 1

Introduction

1.1 Complexity Science: a biased introduction

This section is a brief and biased introduction to some aspects of Complexity Science. It is biased because it aims at introducing only those aspects of complexity on which the rest of the Thesis sits: the concept of *universality* and the concept of *scale invariance*. Certainly much more was left out than included, but what was included is mostly self-contained, and it should suffice for the understanding of the developments of §3.

There are numerous books and review articles that aim at introducing Complexity Science, and each has its own view on the subject. Let us just cite, for instance, Bak (1996); Holland (2000); Albert and Barabási (2002); Bar-Yam (2003); Solé and Manrubia (2009); Mitchell (2009); Newman (2010) or Newman (2011).

I shall not attempt to establish when or where did Complexity Science originate –nor if it really constitutes a scientific discipline of its own– for this is surely a matter prone to endless debate. But it might be fair to consider the seminal works of Anderson (1972) and Kadanoff (1986) as marking the beginning of what is known as Complexity Science nowadays: an interdisciplinary field of study that combines ideas and methods mostly from statistical physics and critical phenomena, and applies them to study problems in almost any other field: from Biology (Gisiger, 2001) and Chemistry (Rao et al., 2010) to Economics (Boyd et al., 2003), to Geoscience (Peters et al., 2002) or even Sports Science (Balague et al., 2013). Complexity Science is concerned with the study of *complex systems*, a term used (and abused) to describe such a variety of systems, that it is difficult to give a *precise* definition. Actually, there seems to be no agreement on what the precise and formal definition of Complex System should be. But in general terms, it is usually said that Complex Systems are formed by a *large number* of elements that *interact*, giving rise to some emergent, global phenomena (which should, in principle, not be directly encoded in the interaction rules, or at least not in a very obvious way).

In essence, Complexity Science attempts to challenge the reductionist approach to scientific inquiry by claiming that "the total is more that the sum of its parts" and that, therefore, reductionism shall ultimately fail: When a problem or system is analyzed by studying its constituent units, and these units are subsequently analyzed in terms of even simpler units, and so on, then a descending hierarchy of realms of study is formed. And while the system might be somewhat understood in terms of different concepts at each different level, from the coarser description down to its most elementary units, there is no guarantee of a successful bottom-up, comprehensive "reconstruction" of the system. Reductionism only provides a way *down* the hierarchy of theories, *i.e.*, towards those supposedly more basic and elementary; Complexity aims at finding a way back home, that is, from the basic elementary units *up* to the original object of study.

The section is organized as follows: §1.1.1 introduces the concept of universality, starting from some simple mathematical examples and then moving to phase transitions in critical phenomena; and §1.1.2 covers the notion of scale invariance and some basic results concerning scale-invariant functions in one and two dimensions.

1.1.1 Universality

The famous mathematician and Field-medalist Terence Tao published in 2012 a non-technical survey on the concept of universality in complex systems (Tao, 2012). Unlike most of the literature available, Tao offers a more mathematically-oriented view on the subject, and being him one of the greatest mathematicians of our times, I cannot resist to begin this Chapter by quoting him. Tao's abstract reads:

> In this brief survey, I discuss some examples of the fascinating phenomenon of universality in complex systems, in which universal macroscopic laws of nature emerge from a variety of different microscopic dynamics. This phenomenon is widely observed empirically, but the rigorous mathematical foundation for universality is not yet satisfactory in all cases.

The survey, published in *Dædalus, the Journal of the American Academy of Arts & Sciences*, walks us from the simplest universal laws in statistics, the law of large numbers and the central limit theorem, to Zipf's law, to phase transitions, and to random matrix theory and the Riemann hypothesis. And while Tao's definition of universality is not much different from the ones that can be found elsewhere (Binney et al., 1992; Stanley, 1999; Sethna, 2006), his survey of examples and his unique perspective are truly enlightening.

Indeed, it is difficult to further define universality without discussing specific examples, precisely because universality is observed in such a broad repertoire of systems. In what follows, we will first see some examples of "universality without complexity", *i.e.* universality arising in systems composed of many *non-interacting* elements, to then turn to phase transitions in ferromagnetic materials, where interactions play a prominent rôle in the emergence of certain universal properties.

Universality without complexity

Let us come back to a definition of universality: *a universal macroscopic feature of a system is one which does not depend on its microscopic details.* But macroscopic observables are generally defined from the individual elements, *e.g.* by means of an average over the whole system or more complicated formulae. In other words, there are lots of degrees of freedom that need to be integrated out, if one wants to be left with a low-dimensional, global observable. So how can a macroscopic feature possibly be independent of the microscopic details? The key point is that it is not completely independent, it is only *almost* independent. That is, the global, universal feature depends only on a small number of parameters of its microscopic constituents, but not on the rest (*i.e.* the vast majority) of degrees of freedom.

Consider the simplest "complex" system possible: a collection of N independent, identically distributed random variables $x_1, \dots, x_N$, with $N \gg 1$, $x_i \sim X$ and X a random variable[1] with law $F(x)$. We now look for universal features of this toy model, *i.e.* features that do not depend on the precise form of $F(x)$ (the microscopic details), but rather only on a few parameters, such as the average or the standard deviation. The first two were already shown by Tao (2012), but they still deserve to be mentioned. Consider

[1] Let us assume that the first and second moments of X are finite.

first the law of large numbers. That is, the fact that the average over the system, S_N, converges to the expected value of X:

$$S_N = \frac{1}{N}\sum_{i=1}^{N} x_i; \quad \lim_{N\to\infty} S_N = \mathrm{E}[X] \tag{1.1}$$

It is obvious that the observable S_N does not depend on the precise form of the law $F(x)$: It does not matter if X is a normal, an exponential, or a Poisson random variable[2]: as long as N is large enough, all that matters is the value of $\mathrm{E}[X]$. In this sense, S_N is a universal "macroscopic" property of our system. Similarly, the second example in Tao (2012) is the central limit theorem.

We shall now give a third example of our own, but this time, we will present it "masked" in the form of an observable of our toy complex system. Suppose that the variables of our system, $x_1, \ldots, x_N$, represent the "intensity" of some elements, say luminescent cells, sitting on sites in a one-dimensional regular lattice. Further suppose the presence of a barrier separating the lattice in two regions, left (L) and right (R), with $N/2$ sites on each side of the barrier[3]. Cells are assumed to be "active" only if $x_i < h$ for some externally fixed level h. Figure 1.1 illustrates this setup. Finally, assume that the observable of

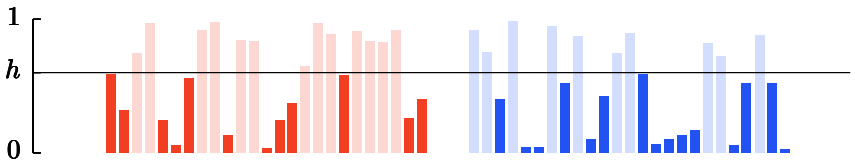

Figure 1.1 A pictorial representation of the toy model. The height of the bars represents the intensity of the elements of the system, which are placed in a 1-dimensional lattice of size $N = 50$. A barrier (not drawn) separates the left region (red) and the right region (blue). Dark colouring indicates active sites, *i.e.*, those whose intensity is below h. In this example, we chose $X \sim \mathrm{U}([0, 1])$ for simplicity.

interest of the system is the "maximum activity asymmetry" A_N, defined as the maximum difference of active sites between the two regions (divided by the system size), when h is varied along the domain of X. From the definition given above, A_N is computed as follows:

$$A_N = \frac{2}{N}\sup_h \left| \sum_{i\in L} \Theta(h - x_i) - \sum_{j\in R} \Theta(h - x_j) \right|, \tag{1.2}$$

with $\Theta(\cdot)$ the Heaviside step function, and L and R the two regions separated by the barrier.

In the toy model's world, this observable could be of interest *e.g.* because it gives a simple estimate of the degree of inhomogeneity in the system. Because we know that

[2]Or any other random variable, provided it has finite first moment.

[3]Placing the barrier in the middle, thus giving regions of equal size, is not essential for the result that will follow, but it simplifies the algebra.

$x_i \sim X$ for all i, in both sides of the barrier, it is obvious to us that there is no real asymmetry in terms of activity besides statistical fluctuations, but the point is to imagine that this is not known in the toy model's world. To complete this thought exercise, suppose that several kinds of cells are put to test in the laboratory; because each kind of cell has different luminescent patterns, each gets a different underlying parent distribution $F(x)$ in the model. Wouldn't it be unexpected, for the inhabitants of the toy model's world, to find out in the laboratory that the distribution of A_N is *independent* of the kind of luminescent cell used? What if any kind of cell ever tried gave exactly the same results? That would certainly be considered a universal feature of the system.

Well, it turns out the observable A_N is just the two-sided Kolmogorov-Smirnov statistic for samples of equal size (Gnedenko and Korolyuk, 1951), and it asymptotically converges to a limiting random variable, provided $F(x)$ is continuous. In particular,

$$\lim_{N\to\infty} \text{Prob}[A_N \leq z\sqrt{4/N}] = 1 - 2\sum_{k=1}^{\infty}(-1)^{k+1}e^{-k^2z^2} \tag{1.3}$$

which has no parameters and, besides the particular form of Eq. (1.3), can also be written as a *scaling law* (see §1.1.2),

$$\text{Prob}[A_N \leq y] \simeq \mathcal{G}\left(\sqrt{N}y\right); \qquad N \gg 1 \tag{1.4}$$

So it could happen, in the imaginary world where the toy model lives, where the absence of interactions is unknown, and where only the values of A_N are available to researchers, that the scaling law in Eq. (1.4) is considered a striking, universal law of the system, which holds for a wide range of cell-types.

Admittedly, the observable A_N was artificially planted in advance, and the whole toy model would not even be considered "complex" at all, due to the absence of interactions. However, it has almost all other typical characteristics of complex systems: a large number of elements giving rise to universal emergent macroscopic properties (A_N), independently of the microscopic properties of the system (the parent distribution $F(x)$). The three examples presented so far have in common that they are asymptotic statements, exact for $N \to \infty$, and provide very good approximations for large but finite N. And so it seems that, indeed, having a very large number of elements is enough to obtain universality and, in this sense, complex systems –formed of a large number of elements– are good candidates to display universality. But complex systems are more than just a large number of elements: usually, the elements are required to *interact* in some or another way, and it is expected that these interactions are the ones that truly give rise to some emergent, universal macroscopic feature. So our second example of universality will include interactions, and it will be, instead of purely mathematical, purely physical.

Phase transitions and critical phenomena

We briefly introduce phase transitions and critical phenomena as a paradigmatic example of universality. Indeed, the core ideas of critical phenomena and phase transitions have had, and still have, a deep impact in Complexity Science.

A simple example of a phase transition is found in ferromagnetic materials. It turns out that when ferromagnetic materials are heated up, they eventually loose their magnetic properties. This can be understood, intuitively, because magnetism is a state of order in a material; of many internal small magnets, if one wishes, pointing in the same direction. When heat is introduced in the system, it excites the particles that form the material, inducing certain disorder and disturbing the global alignment. Thus it is not surprising that the total magnetization M of a ferromagnetic material decreases when the temperature T is raised. What is more surprising, thought, is that it does so in a very peculiar way: there is a special temperature, T_c, called the *critical temperature*, above which the magnetization is always zero. If the ferromagnetic material is heated up to T_c, its magnetization will drop to zero. If it is further heated, the magnetization stays at zero. And conversely, when the material is then cooled down, as long as $T > T_c$, the magnetization will be zero, and only when T reaches T_c will the magnetization M start to raise. Because this happens in a continuous way, *i.e.*, without sudden changes of magnetization, this phase transition is called continuous or of second-order (discontinuous ones being called of first order)[4].

In addition, if the temperature is close to the critical temperature T_c, then the magnetization turns out to be a power law of the distance to the critical point,

$$M \propto |T - T_c|^{\alpha} \tag{1.5}$$

with $\propto$ denoting proportionality, and α called the *scaling exponent*. The really exciting thing, however, is that instead of finding a different exponent for each material, experimentalists find that only a few exponents keep turning up again and again, no matter which materials they put to test. In addition, when the temperature and the magnetization are rescaled in certain ways, some *scaling functions* arise, see §1.1.2, and these scaling functions are also shared by many materials see Figure 1.2. The whole picture is much more complicated, see Binney et al. (1992) for a proper introduction to the subject, but this is the basic idea: first, ferromagnetic materials can be classified in a few universality classes on the basis of their scaling exponents[5] and scaling functions; and second, these same exponents are also found in other phase transitions, such as the liquid-gas transition.

To finish, let us revise how our notion of universality fits with phase transitions in ferromagnetic materials: we have a system (say, a macroscopic sample of some ferromagnetic material), which is composed of a large number of elements (the particles that form the material). And this elements interact (via their local magnetization), giving rise to a macroscopic feature of the system (the global magnetization), whose behavior close to the critical point, characterized by scaling exponents and scaling functions, is independent of the microscopic details of the system (the specific composition of the material). Thus phase transitions in ferromagnetic materials are an example of universality in a purely physical system.

[4]Quite confusingly for mathematicians, second-order phase transitions have a discontinuity in the first derivative of the magnetization as a function of temperature. But everything becomes more clear when one learns that the terms *first* and *second* order actually refer to discontinuities in the first and second derivatives of the free energy.

[5]Actually, α is not the only scaling exponent in a phase transition, nor is Eq. (1.5) the only power

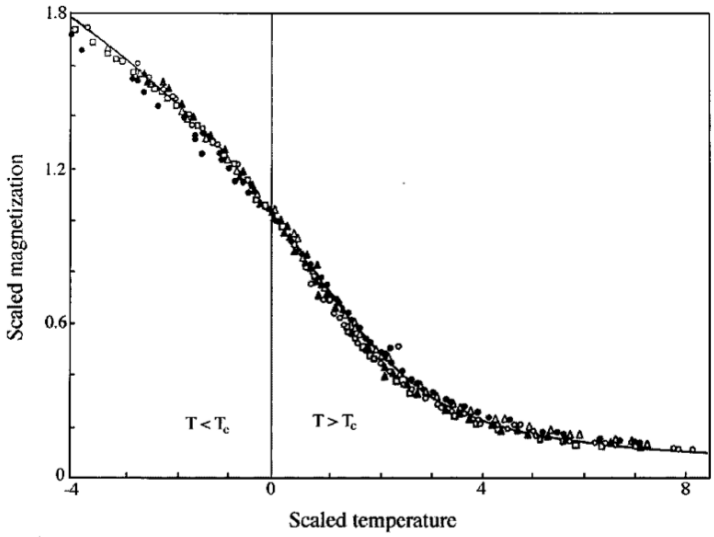

Figure 1.2 The (scaled) magnetization against the (scaled) temperature for five different materials close to their critical temperatures. The zero in the x axis corresponds to $T = T_c$. The symbols correspond to experimental measurements, while the solid line is calculated via a theoretical model. Reproduced from Stanley (1999), where more details can be found.

1.1.2 Scale Invariance

In plain words, scale-invariant objects are those that do not change when we zoom in or zoom out. Such "zoom" can be in spacial coordinates –as with a microscope or telescope–; but sometimes also in time, frequency or other coordinates. The object in turn can also take many forms: it can be a purely mathematical object, like Cantor's set; or a physical concept, like pink noise; or something found in nature, like coastlines or romanesco broccoli. In fact, it is often not the object itself that is scale-invariant, but to be more precise, a property or description of that object. For instance, in the example of pink noise, the scale invariance lies in its power spectral density, which takes the form of a power law; and for romanesco broccoli, it is its shape what is, indeed, scale invariant.

A direct consequence of scale invariance is the so-called "lack of characteristic scale": if the object does not change when we zoom in or out, we cannot know the size of what we are seeing –unless somebody tells us beforehand. And so in the hypothetical case of perfect scale invariance, there is no notion of "typical size", just like the there is no notion of "typical position" in systems with translational invariance. In any case, it turns out that in nature there are a great deal of examples with (approximately) scale-invariant properties, and given that we tend to describe these properties with mathematical functions, we shall first study scale invariance in mathematical functions.

---

law: others are defined, for instance, in terms of the correlation length, the susceptibility, the specific heat, etc. but their exponents are related via scaling relations.

Scale-invariant functions

A function $f : \mathbb{R}^n \to \mathbb{R}$ is called scale-invariant if it is invariant under a scale transformation. A scale transformation is defined as a functional $\mathcal{T} : f \to \mathcal{T}[f]$ that acts on a function $f(x_1, \ldots, x_n)$ as follows:

$$\mathcal{T}[f](x_1, \ldots, x_n) \coloneqq \frac{1}{\lambda_0} f(\lambda_1 x_1, \ldots, \lambda_n x_n), \tag{1.6}$$

with $\lambda_0, \lambda_1, \ldots, \lambda_n \in \mathbb{R}^+$. The condition of scale invariance $\mathcal{T}[f] = f$ then reads:

$$\lambda_0 f(x_1, \ldots, x_n) = f(\lambda_1 x_1, \ldots, \lambda_n x_n) \tag{1.7}$$

While Equation (1.7) can be understood in an abstract setting, *i.e.*, simply as a condition imposed on a function f, it is good to keep in mind some physical interpretation: Suppose that the variables x_i are observables of a physical system, and think of f as a property of the system. For instance, f could be a density function, thus giving the probability of finding the system in a given state:

$$\text{Prob}[X_1 = x_1, \ldots, X_n = x_n] \equiv f(x_1, \ldots, x_n) \tag{1.8}$$

Or f could be expressing a relation between $x_1, \ldots, x_n$ and yet another observable, x_0,

$$x_0 = f(x_1, \ldots, x_n). \tag{1.9}$$

But in both cases, we think of f as property of the system. With this setup in mind, the scale transformation $\mathcal{T}$ is simply a dilation (or contraction) of the observables x_i,

$$x_i \to x_i/\lambda_i, \tag{1.10}$$

including x_0 in the case of Eq. (1.9) and including the probability density function in the case of Eq. (1.8). The meaning of the scale-invariance condition becomes clear now: it implies that the property f extends to all scales of the system, because if we dilate or contract our observables x_i by *arbitrary* factors λ_i[6], the property continues to hold.

This is by no means a conventional feature: usually, properties of physical systems have a range of validity, in the sense that they hold in certain range of the observables, but they fail outside it. Such a range might be defined in a blurry way, *i.e.* sometimes one cannot establish precise values for the boundaries of the range of validity, but nonetheless, the range exists, in the sense that there is always a large enough or small enough value of x_i where f fails. In contrast, *scale invariant properties* hold at all scales, that is, their range of validity is unbounded (at least in theory), which makes them particularly interesting: By extending to all length scales, scale-invariant properties provide a bridge between different realms of physics, thus connecting the microscopic world with the macroscopic one. It is obvious that this greatly limits the functional forms f can take:

[6] We will see however that the dilation factors λ_i must be related between them in particular ways.

the scale invariance condition, $\mathcal{T}[f] = f$, is not fulfilled by most functions. For instance, if we take $f(x) = e^{-x^2}$, then scale invariance would imply that

$$\begin{aligned} \mathcal{T}[f](x) &= f(x), \\ \frac{1}{\lambda_0} e^{-\lambda^2 x^2} &= e^{-x^2}, \\ e^{(1-\lambda^2)x^2} &= \lambda_0, \end{aligned} \tag{1.11}$$

which cannot hold for all x and all λ simultaneously. Hence $f(x) = e^{-x^2}$ is *not* a scale-invariant function.

A natural question to ask then is which are the scale-invariant functions, because any scale-invariant property of a physical system must take their form. We will first respond this question for one-dimensional functions, for which the solution is somewhat more explicit, and then look at two-dimensional functions.

Scale-invariant functions in one dimension

We will now determine the set of functions $f : \mathbb{R} \to \mathbb{R}$ that fulfill the scale-invariance property, Eq. (1.7), which in one dimension reads:

$$\lambda_0 f(x) = f(\lambda x); \qquad \lambda_0, \lambda \in \mathbb{R}^+; x \in \mathbb{R}. \tag{1.12}$$

It is important to interpret this equation properly: it must hold for all λ and all x simultaneously, but λ_0 is allowed to depend on λ, *i.e.* $\lambda_0 = \lambda_0(\lambda)$. Taking the $x-$derivative of Eq. (1.12), and dividing by Eq. (1.12) itself, we find:

$$\frac{\lambda_0 f'(x)}{\lambda_0 f(x)} = \frac{f'(\lambda x)\lambda}{f(\lambda x)}, \tag{1.13}$$

and multiplying both sides by x,

$$\frac{f'(x)}{f(x)} x = \frac{f'(\lambda x)}{f(\lambda x)} \lambda x, \tag{1.14}$$

which implies that the function $xf'(x)/f(x)$ must be constant, because the equation above must hold for all $\lambda > 0$. The solution then follows easily,

$$\frac{f'(x)}{f(x)} x = b \in \mathbb{R} \implies f(x) = ax^b. \tag{1.15}$$

Thus, scale-invariant functions in one dimension are power laws. This proof can be found in Corral (2008), and equivalent ones in Christensen and Moloney (2005); Takayasu (1989). Notice that substituting Eq. (1.15) back into Eq. (1.12) fixes the dependence $\lambda_0 = \lambda_0(\lambda)$,

$$\lambda_0 a x^b = a(\lambda x)^b \implies \lambda_0 = \lambda^b. \tag{1.16}$$

This is important, because it relates the exponent of the power law, b, with the scaling factors λ_i. The exponent b is usually known as the *scaling exponent* and, in plain words, it tells us under which rescaling does scale invariance hold: if we rescale the x-axis by a factor of λ, then we must rescale the y-axis by a factor of λ^b. Under this transformation, $f(x) = ax^b$ is invariant.

Scale-invariant functions in two dimensions

We will now see that, in two dimensions, power laws are *not* the only scale-invariant functions. Instead, we will reach a more general form, usually known as a *scaling law*. The distinction between scaling laws and power laws is important, and will play a key rôle in the developments of §2.2.

Let us consider functions $f : \mathbb{R}^2 \to \mathbb{R}$ under a scale transformation $\mathcal{T}$,

$$\mathcal{T}[f](x_1, x_2) = \frac{1}{\lambda_0} f(\lambda_1 x_1, \lambda_2 x_2). \tag{1.17}$$

Let us write $\lambda_0 \equiv \lambda$, and $\lambda_1 \equiv \lambda^{b_1}, \lambda_2 \equiv \lambda^{b_2}$, with $b_1, b_2 \in \mathbb{R}$ fixed. That its, we are looking at transformations of the form

$$x_1 \to x_1/\lambda^{b_1}, \tag{1.18}$$
$$x_2 \to x_2/\lambda^{b_2}, \tag{1.19}$$

and the scale-invariance condition reads,

$$f(x_1, x_2) = \frac{1}{\lambda} f(\lambda^{b_1} x_1, \lambda^{b_2} x_2). \tag{1.20}$$

It is worth mentioning that Eq. (1.20) is also the definition of a *generalized homogeneous function* of two variables. Following Christensen and Moloney (2005), we will now show that all functions fulfilling Eq. (1.20) are of the form

$$f(x_1, x_2) = |x_1|^{1/b_1} f\left(\pm 1, \frac{x_2}{|x_1|^{b_2/b_1}}\right). \tag{1.21}$$

To see that Eq. (1.20) implies Eq. (1.21), one takes $\lambda = |x_1|^{-1/b_1}$. The converse is seen by verifying that $|x_1|^{1/b_1} f\left(\pm 1, x_2/|x_1|^{b_2/b_1}\right)$ fulfills the scale-invariant property, Eq. (1.20). With this, we have characterized the set of two-dimensional scale-invariant functions, as defined by Eq. (1.20). Because $f(\pm 1, \cdot)$ is actually a function of one argument, it customary to relabel it as $\mathcal{G}_\pm(\cdot)$, so that the general form of a function $f : \mathbb{R}^2 \to \mathbb{R}$ invariant under a scale transformation is:

$$f(x_1, x_2) = |x_1|^{1/b_1} \mathcal{G}_\pm\left(\frac{x_2}{|x_1|^{b_2/b_1}}\right). \tag{1.22}$$

The function $\mathcal{G}_\pm(\cdot)$, which is an arbitrary function of one argument, is usually called the *scaling function*, b_1, b_2 are the *scaling exponents*, and Equation (1.22), altogether, is called a *scaling law*.

Data collapses

Summarizing, so far we have shown that any one-dimensional function fulfilling scale-invariance must be a power law, and that any two-dimensional function fulfilling scale-invariance must be a scaling law. In the one-dimensional case, checking scale-invariance

with graphical methods[7] is straightforward, as one simply checks whether or not $f(x)$ is a power law of x: Plotting $f(x)$ against x in a double logarithmic scale should render a straight line of slope $+b$. For this, knowledge of the scaling exponent b is not required a priori.

In the two-dimensional case, however, things are slightly more complicated: because the scaling function $\mathcal{G}$ can take any form; plotting $f(x_1, x_2)$ as a surface in 3-dimensional space is unlikely to be helpful. Instead, a popular method consists in plotting $f(x_1, x_2)|x_1|^{-1/b_1}$ against $x_2/|x_1|^{b_2/b_1}$. Obviously, this requires knowledge of the scaling exponents a priori. If Eq. (1.22) holds, this procedure yields a single, unique curve which corresponds to the scaling function $\mathcal{G}$. Such a procedure is known as a *data collapse* because it creates the effect of different curves to collapse into a single one, if the correct scaling exponents are chosen.

Let us illustrate the process of data collapse with a simple example. Suppose that $f(x_1, x_2) = e^{-x_1/x_2}\frac{1}{x_1}$. To convince ourselves that this is truly a scale-invariant function, we notice that

$$f(x_1/\lambda, x_2/\lambda) = e^{-(x_1/\lambda)/(x_2/\lambda)}\frac{\lambda}{x_1} = \lambda f(x_1, x_2), \tag{1.23}$$

which in turn means that the scaling exponents are $b_1 = b_2 = -1$. Figure 1.3 (top) shows a 3-dimensional plot of $f(x_1, x_2)$ for $(x_1, x_2) \in [1, 3] \times [1, 3] \subset \mathbb{R}^2$. As said above, it is difficult to asses the scale-invariance of $f(x_1, x_2)$ solely on the basis of this plot. The next step consists in plotting $f(x_1, x_2)$ as a function of x_1, for different values of x_2. In Figure 1.3 (bottom), we chose $x_2 \in \{1, 2, 3\}$, and obtain three different curves (bottom left). But if we rescale our variables appropriately, the three curves collapse into a single one (bottom right): We are left with the scaling function, and scale-invariance has been "verified" (at least with what graphical methods can provide).

Obviously, the only way to *strictly* prove scale-invariance is by checking the mathematical definition, *i.e.*, verifying that indeed $\mathcal{T}[f] = f$, but this is only an option if the function $f(x_1, x_2)$ is known a priori, which is rarely the case. Instead, one typically has access to a dataset, say

$$\mathcal{S} = \{(x_i, y_i, z_i) : i = 1, \ldots, N\}, \tag{1.24}$$

and conjectures, for instance, that x, y, and z are related via a scale-invariant function, $z = f(x, y)$. It is in this circumstances where the notion of data-collapse becomes useful: by trying to "guess" the scaling exponents b_1, b_2, or by using the exponents predicted from a theoretical method, one can attempt a data collapse of the data.

[7] A related but different matter, which we do not treat here, concerns the case where $f(x)$ represents a density function and one wants to fit the exponent b with maximum-likelihood methods and validate the results with *e.g.* the Kolmogorov-Smirnov test. This has been a subject of certain debate. See Clauset et al. (2009); Deluca and Corral (2013).

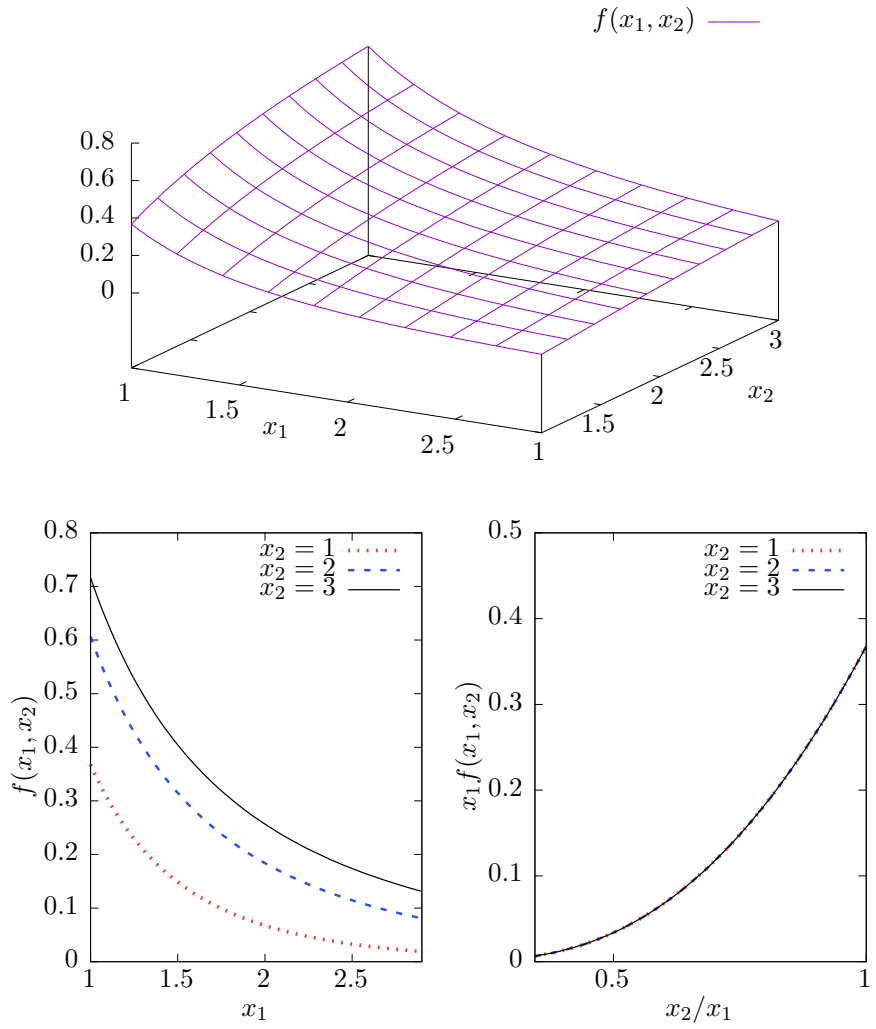

Figure 1.3 Top: A 3-dimensional plot of the function $f(x_1, x_2) = e^{-x_1/x_2}/x_1$. **Bottom:** (Left) the function $f(x_1, x_2)$ versus x_1, for different values of x_2. (Right) A data collapse of $f(x_1, x_2)$, revealing the scaling function $\mathcal{G}(y) = e^{-1/y}$.

1.2 Linguistic laws

This section is concerned with linguistic laws. First, a short discussion on how linguistic laws differ from physical laws is laid out via a simple example. §1.2.1 discusses the two main perspectives under which linguistic laws have been investigated: the perspective of Quantitative Linguistics (QL); and the perspective of Complexity Science. Finally, Zipf's law (§1.2.2) and Heaps' law (§1.2.3) are introduced.

To give an illustrative example of what a linguistics law is, consider the following question: if a given document, say an article in the encyclopedia, has a *total* length of L words, how many *different* words does it have? The answer to this question is known as Heaps' law, see §1.2.3 for details, but for now, let us approach it in a more informal and didactic way. A rather blunt answer to it would be that it depends on the article: for instance, the article for *Linguistics* in the English version of the Wikipedia has a total length of 6 353 words, of which only 1 501 are different, but the article for *Mathematics* has a length of 10 184 words, totaling 1 436 different ones. That is very accurate, but not very interesting, because it applies only to those two particular articles. But if we could claim that when an article in the Wikipedia has a total length of L words, then it has *approximately* $L^{4/5}$ different ones, that would certainly be more interesting. At the cost of less accuracy, that second answer gains in broadness, because it applies –presumably– to *any* article in the Wikipedia. Of course, we would then try to quantify in a more precise way what does *approximately* exactly mean in the sentence above, and we would try to determine the limits of its applicability (*e.g.* does it hold only for articles in English, or is it valid for any language?).

The gist of the example, however, is that linguistic laws are statements based on experimental findings. Some of them, such as Zipf's law (see §1.2.2), have been reported in numerous corpora, spanning tens of languages, and are therefore considered experimental laws. It seems clear, however, that they must belong to a different category than *e.g.* the laws of classical Physics. Altmann and Gerlach (2015) debate this question, stating that *"a creative and persistent daemon, trained in the techniques of constrained writing, can generate understandable and arbitrary long texts which deliberately violate any single law [...]"*, among other insightful remarks. Thus linguistic laws can be violated, while *e.g.* the laws of classical Physics cannot[8]. This has consequences at many different levels: from the epistemological level –can falsifiable theories, in a strict Popperian sense, be constructed from linguistic laws?– to the methodological one –how should we interpret deviations from the "predictions" of linguistic laws? are in this sense p-values meaningful?– which, admittedly, are not covered in Thesis. I would like however to make an informal remark, focusing for a second on the bright side of things: Putting aside daemons[9], exercises of constrained writing, and so on, it seems to me that linguistic laws such as Zipf's law (§1.2.2) are fulfilled with a surprising degree of accuracy, *specially* given the fact that they can, indeed, be violated.

[8] At least, not for macroscopic object at speeds not comparable to the speed of light.
[9] Here, and in the preceding quote: making reference to Maxwell's daemon

1.2.1 From Quantitative Linguistics to Complexity Science

According to Köhler et al. (2005), Quantitative Linguistics shares its subject, aims and issues with general Linguistics, but differs from it in the methods, which include *"all of the mathematical tools, i.e., especially quantitative methods"* (Köhler et al., 2005, pp. viii). Hence we might say – even if that sounds slightly tautological– that Quantitative Linguistics is the study of language with quantitative methods. Such quantitative methods range from mere counting processes to sophisticated mathematical modeling, and the field expands into almost all of the classic branches of linguistics, including phonology, morphology, syntax and semantics. Hence from the QL perspective, the linguistic laws studied in this Thesis, Zipf's law and Heaps' law, are only two out of a multitude. But they are particularly interesting due to a series of parallelisms, analogies or perhaps simple resemblances with findings in other fields outside linguistics, which fall under the umbrella of Complexity Science.

In a nutshell, the main interest of "complexologists" in some linguistic laws lies in the fact that these can been seen as *scaling laws*, in the traditional meaning this has in *e.g.* the theory of critical phenomena (CP) and as introduced in §1.1.1. We will see that Zipf's and Heaps' law take the form of scaling laws, which is considered a sign of complexity. In addition, some of the basic facts of CP, such as the presence of long-range correlations when the system is close to the critical point, have been also observed in natural languages (Ebeling and Pöschel, 1994; Montemurro and Pury, 2002; Altmann et al., 2012). All of this builds up a view in which natural language is just another system, conjectured to be in a critical state. And while it is remarkable how accurate this parallelism is in the case of natural language –in comparison with other systems–, it is not clear what lessons can be learned from such analogies. If we were to further stretch them, then we might be tempted to say that "language is in a critical state, at the edge of order and disorder", or even that "language is a self-organizing system, posed at the critical point without the need of external fine tuning". Perhaps these sort of claims are appealing to some because of the aforementioned analogy with critical phenomena or even with the theory of Self-Organized Criticality (SOC) (Bak, 1996); but their meaning beyond the analogy itself is certainly unclear. Nevertheless, there are a series of technical *prerequisites* for the whole analogy to even make sense, and it is on this prerequisites that we shall now concentrate. In particular, in relation to the notion of *scaling exponent* and *universality classes* (see §1.1):

Firstly, in critical phenomena the notion of universality class is expected to be robust: In particular, scaling exponents should not be influenced by external scales, *i.e.*, their value should not depend on the system size. Hence, if an analogy as outlined above is to be drawn, it is necessary that the exponent γ of Zipf's law (see §1.2.2) does not dependent of the system size, *i.e.* the length of the document L. Otherwise, the very concept of scaling exponent would not even be properly defined, in the case of natural languages, and the chances of further relating language and criticality would vanish, at least via this approach. In (Font-Clos et al., 2013, §3.1), we discuss this matter at depth, in relation to some prior claims of Bernhardsson et al. (2009)[10], and conclude that Zipf's exponent γ is

[10] Admittedly, some controversy followed: see the comment by Yan and Minnhagen (2014) and the subsequent reply in Font-Clos and Corral (2014).

independent of the document length[11] L.

Secondly, if one accepts that proper scaling exponents can be defined (for natural languages), the question of whether or not one can also establish the existence of universality classes follows naturally. This matter has not been investigated in this Thesis, but the following remark is worthwhile. There are many candidates for what the universality classes would be. For instance, the most ambitious result one could dream of would be that *all* corpora ever analyzed displayed exactly the same scaling exponents, for say, at least, Zipf's law. But the evidence suggests otherwise: Figure 1.4 shows the distribution of the exponent γ across 37 078 documents from the Project Gutenberg database, of which around 80% are tagged as being written in English. And although this only constitutes

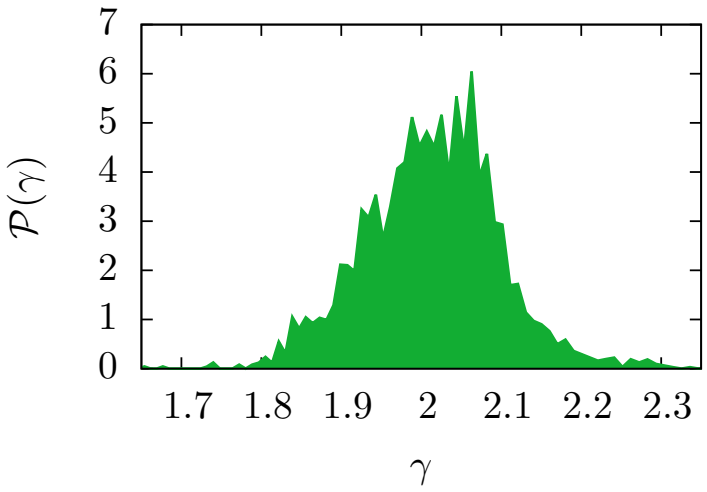

Figure 1.4 Distribution of Zipf's exponent γ across the Project Gutenberg database. Fits where done as explained in Deluca and Corral (2013). A total of 37 078 documents where included in the analysis.

a preliminary analysis, it indicates that there are important variations in the value of γ, even within the same language. A much more conservative claim would be to equate universality classes to authors, so that all works of a given author would display the same exponent. And while in this case more research would certainly be needed to establish a result, it could only ever be a very weak result: in critical phenomena, one typically finds *a few* universality classes under which all models can be categorized irrespectively of their microscopic details. This is the gist of universality. But finding *thousands* of universality classes, as would be the case if they were to be equated with authors, would be of limited interest, and probably would not qualify as universality at all. It might happen that in-between one unique universality class –which the evidence defies– and thousands of universality classes –which would be uninteresting– there is, perhaps, hope for a middle ground.

[11] Note that this refers to the length of a document as it is being read, that is, to the value of Zipf's exponent as subsets of increasing size of a fixed docum are considered

In summary, linguistic laws are fascinating phenomena that provide a bridge between Linguistics and Complexity Science. These disciplines have different long-term goals: Linguistics aims at the general study of language in all of its facets, and hence tends to put linguistic laws in relation to its whole body of knowledge; Complexity Science, on the other hand, envisages to relate linguistic laws with scaling laws, and with other systems that also display them. And while both fields are actively and separately pursuing research on the topic, the dialogue between linguists and complexologists is certainly in its infancy.

1.2.2 Zipf's law

Zipf's law (Zipf, 1949), named after George Kingsley Zipf, is without doubt the most famous and well-studied experimental law in Quantitative Linguistics. It has been reported numerous times in a variety of corpora and languages (Zipf, 1949; Zanette and Montemurro, 2005; Zanette, 2012; Corral et al., 2015), animal communication systems (McCowan et al., 1999; Hanser et al., 2004), and beyond (Czirók et al., 1995; Serrà et al., 2012). And yet it can be announced in a few lines, and anyone with a computer and some scripting skills can put it to test in a few minutes.

In its broader sense, Zipf's law has been found to hold –with varying degrees of rigor– in a wide range of systems. Indeed, gathering an exhaustive list of such examples is in itself a daunting task; here we shall rather give a non-exhaustive, but hopefully representative, list of empirical examples of Zipf's law beyond linguistics: In the field of economics, the income distribution (Malevergne et al., 2011) is the most notable example, but see also (Axtell, 2001; Clauset et al., 2009). In biology, let us cite the abundance of proteins in a cell (Furusawa and Kaneko, 2003), the presence of insects in plants (Pueyo and Jovani, 2006), or the distribution of organism's mass in ecosystems (Camacho and Solé, 2001). Finally, in the social sciences, the examples of visitors or links in web pages (Adamic and Huberman, 2002) or telephone calls to users (Newman, 2005) are typical examples.

In light of the quantity and diversity of systems that exhibit Zipf's law, it is fair to, at least, raise the following questions: What do all these systems have in common? Is there a "unifying principle", perhaps just a heuristic argument, that can be used to explain the origins of Zipf's law? Obviously, this is an open question, but at a theoretical level, certainly a great number of mechanisms have been proposed (Ferrer i Cancho and Solé, 2003; Miller, 1957; Li, 1992; Simon, 1955; Zipf, 1949; Corominas-Murtra et al., 2011; Ferrer i Cancho, 2005; Bak, 1996), and Zipf's law has been "unzipped" (Adamic, 2011; Baek et al., 2011), "explained" (Gabaix, 1999), or "understood" (Corominas-Murtra et al., 2015) far too many times, to the extend that no mechanism is considered good enough by anyone. But this is by no means a negative sign: it is rather a sign of the exceptional interest that the complex-systems community has in Zipf's law. The curious functioning of the academic world and its increasing pressure to publish in high-impact factor journals might be accountable of some exaggerated claims, but nobody truly believes to understand the origins and mechanisms leading to Zipf's law. A very good review of generative mechanisms, although now somewhat outdated, is (Mitzenmacher, 2004). See also (Newman, 2005; Saichev et al., 2009; Zanette, 2012).

Due to the broad repertoire of systems exhibiting Zipf's law, mechanisms depending too much on the details of a given system are of limited interest for the general understanding of the law. But explanations that try to accommodate a broader set of systems tend to be, by necessity, based on very abstract, undefined concepts, such as "entities", "groups", "elements", etc., and the law usually arises after imposing principles or rules of similar vagueness, all of which renders the mechanism of dubious applicability. Obviously, this is a "paradox" not specific of Zipf's law [12], but the wide range of systems where Zipf's law has been observed, and hence that any generative model should try to accommodate, makes the paradox particularly problematic. But in the end, each mechanism provides its share of insight, be it for the gradual understanding of Zipf's law, or for a specific example it is based on. In this sense, theoretical research on Zipf's law is still, and will be in the forthcoming years, of great interest for the complex-systems community.

Definition

Although Zipf's law seems to pervade a variety of systems, it is customary and perhaps more instructive to announce it in the original setting of natural languages. In its simplest form, the law can be stated as follows: given a corpus of natural language, the frequency of a word is inversely proportional to its rank, *i.e.*,

$$n \propto \frac{1}{r}, \tag{1.25}$$

where $\propto$ denotes proportionality, n the frequency of a word in the corpus and r its rank, *i.e.*, the position it occupies in the list of sorted frequencies [13,14]. In other terms, Zipf's law says that the most common word appears twice as much as the second most common word, and three times as much as the third most common word, etc. A slightly more refined version of the law introduces a parameter $\beta \simeq 1$,

$$n(r) \propto \frac{1}{r^{\beta}}. \tag{1.26}$$

Obviously, more complicated versions of the law are possible (Baayen, 2001; Li et al., 2010), but Eq. (1.26) is what is generally known as Zipf's law. We shall however refer to Eq. (1.26) as the **rank-count** representation of Zipf's law for reasons that will become clear soon. There is an alternative representation of the law in terms of frequencies of frequencies. This might sound confusing at first, but it is actually very simple: let $N(n)$ be the number of different words that have frequency n in a corpus. Then Zipf's law can also be stated as follows:

$$N(n) \propto \frac{1}{n^{\gamma}} \tag{1.27}$$

[12] This applies more generally to scientific modeling: very concrete, detailed models give rise to conclusions that can difficultly be extrapolated to other systems, while very broad, abstract models can hardly be applied to specific examples.

[13] That is, the most common word has rank 1, the second most common word has rank 2, and so on.

[14] In case of several words having exactly the same frequency n, ties are solved either at random or alphabetically. While the later seems slightly more arbitrary, it has the advantage of yielding reproducible results.

with $\gamma \simeq 2$. To this second representation of the law, Eq. (1.27), we shall refer as the **frequency-count** representation. Note that $N(n)$ is proportional to the probability mass function[15] of n, if n is regarded as a random variable. That is, suppose that we choose a word at random from the list of all available different words. Then, with probability (proportional to) $N(n)$, the chosen word will have frequency n, *i.e.* it will appear n times in the corpus. The frequency-count representation of Zipf's law, Eq. (1.27) is saying that there are many words with very low frequency, and very few words with high frequency.

These two representations of Zipf's law, Eqs.(1.26) and (1.27), are not equivalent in the most strict sense, but before discussing the relation between them, and under which conditions they are equivalent, let us show a few examples of Zipf's law. Figure 1.5, reproduced from the original book of George Kingsley Zipf, *Human behavior and the principle of least effort* (Zipf, 1949) shows the rank-count representation of Zipf's law for the book *Ulysses*, which G. K. Zipf called "rank-frequency distribution of words". In contrast, Figure 1.6 shows the frequency-count representation, *i.e.* the (normalized)

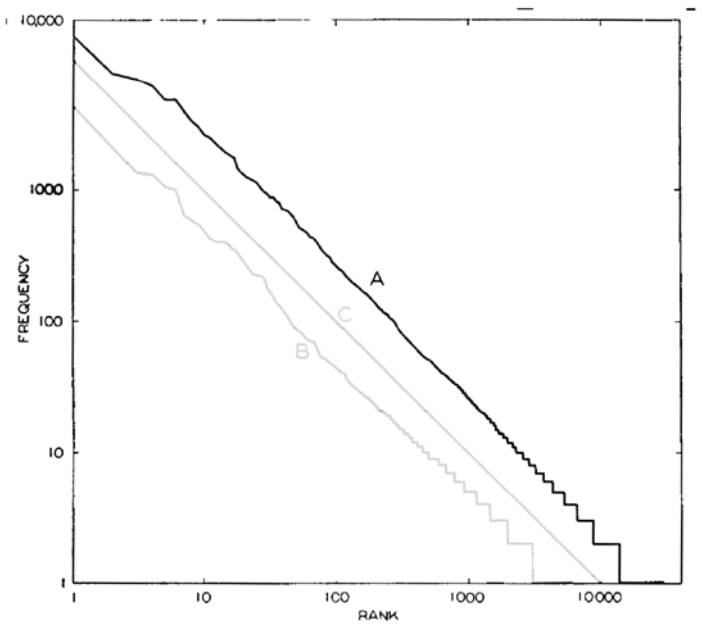

Fig. 2–1. The rank-frequency distribution of words. (A) The James Joyce data; (B) the Eldridge data; (C) ideal curve with slope of negative unity.

Figure 1.5 The rank-count representation of Zipf's law for the book *Ulysses*. Reproduced from (Zipf, 1949) (curves B and C have been shaded for clarity).

probability mass function of the frequencies of words in the books *Moby Dick* and *Ulysses*. Fitting a lower-truncated power law yields, following the method of Corral et al. (2012), a value for Zipf's exponent of $\hat{\gamma} = 1.95$ for the former, and $\hat{\gamma} = 1.98$ for the later, and

[15] Or the empirical probability mass function. This distinction is usually overlooked in the literature, and generally the *sample* and the *population* are not properly distinguished. See however (Mandelbrot, 1961).

$n_{\min} = 6$ in both cases. Thus Zipf's law in the frequency-count representation does not hold on the entire regime of frequencies for *Moby Dick* nor for *Ulysses*, but it holds in the regime $n \geq n_{\min} = 6$ in both cases.

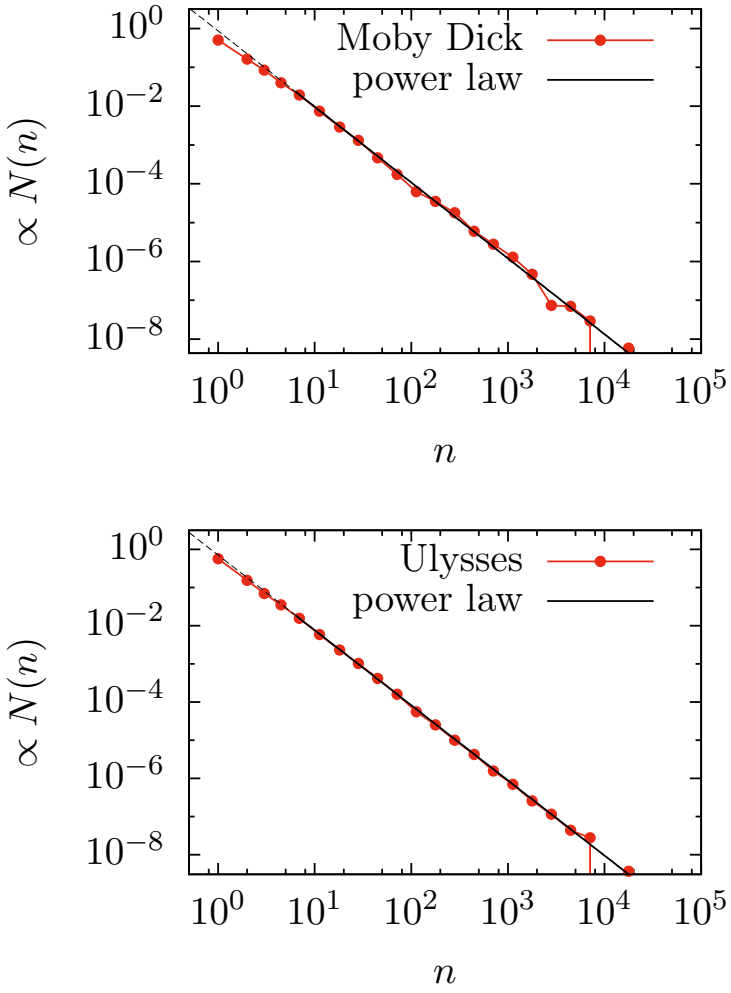

Figure 1.6 Zipf's law in the frequency-count representation for the books *Moby Dick* (top) and *Ulysses* (bottom). Red symbols correspond to the probability mass function of the frequencies n. The solid black line corresponds to fitting a lower-truncated power law, following (Corral et al., 2012), which yields fitted values of $\hat{\gamma} = 1.95$ (top) and $\hat{\gamma} = 1.98$ (bottom) and $n_{\min} = 6$ (both). The dashed thin black line extends the power law outside the fitted range, as a guide to the eye.

Comparing Figure 1.6 (bottom) with Figure 1.5 serves to exemplify the relation between exponents of the two representations $1/\beta = \gamma - 1$, see Equation (1.32), although the fitted value $\hat{\gamma} = 1.98 \simeq 2$ for the frequency-count representation cannot be compared in a fair way to the visually estimated one, $\beta \simeq 1$, of the rank-count representation.

Relation between the two representations

This section starts with a classic derivation relating the rank-count and the frequency-count representations of Zipf's law, Eqs. (1.26) and (1.27). While the derivation is fairly elementary, it also shows that the two representations are only asymptotically equivalent, for $n \to \infty$. Thus, they can only be considered *approximately* equivalent in the regime of large frequencies, $n \gg 1$, and they are definitely *not* equivalent for small frequencies, $n \sim \mathcal{O}(1)$.

The key point in the derivation is to realize that the rank r of a word with frequency n can be expressed as follows[16]:

$$r(n) = \sum_{n' \geq n} N(n') \tag{1.28}$$

That is, the rank r of a word with frequency n is the total number of words with a frequency n' greater than or equal to n. Once this is clear, the result follows easily. First, invert Eq. (1.26),

$$n(r) \propto \frac{1}{r^\beta} \implies r(n) \propto \frac{1}{n^{1/\beta}}, \tag{1.29}$$

and then insert Eqs. (1.29) and (1.27) into (1.28),

$$\frac{1}{n^{1/\beta}} \propto \sum_{n' \geq n} \frac{1}{(n')^\gamma}. \tag{1.30}$$

Assuming $n \gg 1$, we can approximate the sum by an integral,

$$\frac{1}{n^{1/\beta}} \propto \int_{n'=n}^{\infty} dn' \frac{1}{(n')^\gamma} \propto \frac{1}{n^{\gamma-1}}, \tag{1.31}$$

and the relation between exponents becomes apparent,

$$\frac{1}{\beta} = \gamma - 1. \tag{1.32}$$

It is difficult to establish who was the first to publish this (elementary) result, but similar or equivalent derivations can be found in (Mandelbrot, 1961; Baayen, 2001; Adamic and Huberman, 2002; Kornai, 2002; Zanette, 2012; Ferrer i Cancho and Hernández-Fernández, 2008; Font-Clos et al., 2013). Is is also worth mentioning that G. K. Zipf himself was aware of both representations of the law, and of the relation between their exponents, at least for the case $\beta = 1, \gamma = 2$, see (Zipf, 1949, Chapt. 2, Sec. IV). In any case, it is important to bear in mind what was assumed throughout the derivation: mainly, the scaling $n \gg 1$ was used to *approximate* a sum by an integral. The result is exact only in the asymptotic limit of $n \to \infty$, but of course, this is never attained in reality, as all

[16] Note that Eq. (1.28) only recovers the maximum rank for a given frequency n. While for small n there are many words with the same frequency, and hence strictly speaking $r(n)$ is a multi-valued function, in the regime of $n \gg 1$ (which will be assumed in the next steps) this is rarely the case, *i.e.* $r(n)$ is single valued in that regime, and no problems arise in this respect.

corpora are of finite length. Hence discreteness effects, if one wishes, must be incorporated in the analysis from the very beginning.

To sum up, we have just shown that the rank-count and the frequency count representations of Zipf's law are approximately equivalent in the regime of very large frequencies, hence small ranks, and that their exponents are related by Eq. (1.32) in this regime. The coexistence of these two representations of Zipf's law, and their approximate equivalence, have caused a somewhat surprising amount of confusion. If the rank-count and the frequency-count representations of Zipf's law are assumed to be equivalent too far away from the large n regime, then obviously incorrect results can be derived –or to put it mildly, one gets to very bad approximations. Indeed, the non-equivalence of the two representations is key to the developments of §3.2, which deals with the relation between Zipf's law and the vocabulary growth curve (see §1.2.3).

Once it has been established that the two representations are not exactly equivalent, it is natural to ask which provides a better description of a given corpus. This matter is outside the scope of this thesis, but (Moreno et al., 2015) covers it with great detail, analyzing a (non-aggregated) database of thousands of books in the public domain. Large *aggregated* corpora were previously analyzed by Ferrer i Cancho and Solé (2001) and more recently by Petersen et al. (2012) and Gerlach and Altmann (2013), all reporting the presence of two separate scaling regimes, with different scaling exponents[17].

Comparison of the two representations

A different matter worth mentioning is the pros and cons of working with rank-count or frequency-count representations. The discussion that follows is not restricted to the power-law forms of Eqs. (1.26) and (1.27); in the sense that it would equally apply to other, more refined functional forms. We shall then speak about "rank-count relation" and "frequency-count relation" in this broader sense, making reference to some unspecified functions $n(r), N(n)$ although it is helpful to keep in mind the paradigmatic case of a non-truncated power law. The discussion can be divided in the following two main points:

1. *Regarding the initial exploratory analysis of a dataset*

   Admittedly, rank-count relations offer some practical advantages, which might explain their popularity. For visual inspection purposes, the rank-count relation comes in handy: once the frequencies have been obtained, one only needs to sort them in decreasing order, and the ranks can be readily assigned. This gives a set of paired data-points, say $\{(r_1, n_1), (r_2, n_2), \dots\}$, which are then typically plotted in double logarithmic scale together with the function $n(r)$. In the case of Zipf's law as stated in Eq. (1.26), this should render a straight line of slope $-\beta \simeq -1$, which is easily distinguished at naked eye. While this is obviously not enough to establish the validity of the law, it certainly helps in the exploratory stage of analysis. In contrast, working with a frequency-count relation would require more

[17] Notice that in *aggregated* corpora frequencies of words across many documents are added up. Recently, Williams et al. (2015) have investigated the effects the aggregation process, finding strong correlations between the crossover that separates the two regimes and the average size of the aggregated documents.

work in this stage: one would need to first count the frequencies n of the words, and then the frequencies of the frequencies, $N(n)$. The resulting data-points, say $\{(n_1, N(n_1)), (n_2, N(n_2)), \dots\}$ would typically be very sparse for high frequencies, so some sort of binning would be needed –and the binning procedure in itself can be a delicate matter, specially for discrete variables, see for instance Pruessner (2009); Christensen and Moloney (2005). In this sense, it is to some extent understandable that rank-count relations have their share of popularity in the community.

2. *Regarding parameter estimation and statistical validation*

   Frequency-count relations are statements about the probability mass function of n, which can be regarded as a random variable, so that *e.g.* standard maximum likelihood methods to fit the parameters and goodness-of-fit tests to validate them can be used without further difficulties[18]. In contrast, rank-count relations are in principle a functional relationship between ranks and frequencies. Notice how the rank of a word has the peculiarity of depending on the frequencies of *all* other words in a corpus, because it is defined as the position it occupies in the *sorted* list of frequencies. It has been argued (Altmann and Gerlach, 2015) that the ranks can be regarded as a "hidden" random variable[19], and that this allows the use of maximum likelihood and hypothesis testing methods. Altmann and Gerlach (2015) acknowledge the problems of this approach (mainly, the fact that one gets an overestimation in the p-values), but they judge them to have no significant effect on the final results.

In summary, the rank-count relation offers a practical advantage during the initial exploratory analysis of a dataset, while the frequency-count relation is more appropriate for proper statistical fitting.

1.2.3 Heaps' law and the vocabulary growth law

Heaps' law (Heaps, 1978), named after Harold Stanley Heaps, is usually announced as follows:

$$V \propto L^{\alpha}, \tag{1.33}$$

with V the number of *different* words in a corpus, L the *total* number of words or corpus size, and $\alpha \in (0, 1)$ the scaling exponent. Together with the Menzerath-Altmann law (Altmann, 1980), Heaps' law is probably the second most famous law in Quantitative Linguistics (the top position, of course, is occupied by Zipf's law, see §1.2.2).

Heaps' law is known at least since 1954, when H. Guiraud announced it with $\alpha = 0.5$ in (Guiraud, 1954). A few years later, G. Herdan (Herdan, 1960) "rediscovered" it, but it was not until 1978 that Heaps published his book (Heaps, 1978). The law came to be

[18] There is however some controversy associated to the fitting of power-law distributions, see (Corral et al., 2012; Deluca and Corral, 2013; Corral et al., 2011) in relation to (Clauset et al., 2009).

[19] In this approach, each word would have a "universal" rank r_*, corresponding to the position it occupies in the sorted list of "universal" frequencies, *i.e.* some idealized, true frequencies of words, and the observed ranks r would be only an approximation, $r \simeq r_*$.

known after Heaps rather than Herdan[20] or Guiraud, but as Stigler (1980) would claim a few years later, this is a common process in science[21,22].

Similarly as with Zipf's law, researchers have found that Heaps' law holds –or at least *approximately* holds– in a variety of systems. In the case of Heaps' law, however, the evidence is not always fully convincing. Indeed, the very interpretation of Eq. (1.33) is problematic. In principle, there are two ways of understanding (1.33), and whether or not they are equivalent is a delicate issue. §3.2 analyses the relation of Zipf's law and Heaps' law, or rather, the vocabulary growth curve, see below. Here we shall only expose the two interpretations of the law, and briefly discuss their equivalence.

The first interpretation of the law assumes the existence of a collection of say $\mathcal{N}$ documents, *i.e.* of a number of *disjoint* instances of natural language. Each document $i = 1 \dots \mathcal{N}$ has a given number of different words or vocabulary V_i, and a total number of words or length L_i. In this setting, Equation (1.33) is understood as a functional relationship that the set of paired values $\{(L_1, V_1), (L_2, V_2), \dots, (L_\mathcal{N}, V_\mathcal{N})\}$ fulfills,

$$V_i \propto L_i^\alpha, \qquad i = 1, 2, \dots, \mathcal{N}. \tag{1.34}$$

We shall refer to this interpretation of Eq. (1.33) simply as **Heaps' law**.

In contrast, the second interpretation of the law applies to a single, unique document. It understands that Eq. (1.33) describes the growth of vocabulary along a given document, i.e. that L in the right-hand side means "the first L words of the document", and that V in the left-hand side means "the number of different words found in the first L words of the document". In plain terms, as a book is read, the vocabulary grows, and Eq. (1.35) describes this growth. To avoid misapprehensions, it helps to rewrite the vocabulary growth curve with lower-case letters,

$$v(\ell) \propto \ell^\alpha, \qquad \ell = 1, 2, \dots, L \tag{1.35}$$

To this second interpretation of the law, we shall refer as the **vocabulary growth law**.

It is fair to say that, in principle, the vocabulary growth law and Heaps' law are not equivalent statements. It could happen, of course, that given a collection of documents, both laws are fulfilled [23], but this is not granted in advance. Without further assumptions, all we can say is the following: in a collection of documents where the vocabulary growth law is fulfilled, with α and the proportionality constant fixed along all documents, Heaps' law will also be fulfilled. This is easy to see, by simply assuming that the vocabulary growth curve holds for a collection of documents,

$$v_i(\ell_i) \propto {\ell_i}^\alpha, \qquad \ell_i = 1 \dots L_i \; ; \; i = 1 \dots \mathcal{N} \tag{1.36}$$

and then simply taking $\ell_i = L_i \; \forall i$,

$$v_i(L_i) \propto {L_i}^\alpha, \qquad i = 1 \dots \mathcal{N} \tag{1.37}$$

[20] Although the law is sometimes referred to as Herdan's law in the field of Linguistics

[21] According to Stigler's law, *"No scientific discovery is named after its original discoverer"*. Stigler provided an explanation for this fact in (Stigler, 1980), building upon ideas of R. K. Merton, the father of the sociology of science.

[22] Zipf' law is no exception to this fact, see Petruszewycz (1973) for a historical review of Zipf's law.

[23] In the sense that the vocabulary growth curve would hold at the level of all individual documents, and at the same time Heaps' law would hold at the collection level.

Defining $V_i = v(\ell_i)$ in Eq. (1.37) renders it equal to Eq. (1.33), Heaps' law. It is clear from this (elementary) proof that the conditions $\alpha_i \equiv \alpha$, and that of the proportionality constant, are necessary. The reciprocal, however, does not hold in general. Indeed, it is possible to construct synthetic datasets fulfilling Eq. (1.34) but not Eq. (1.35), see Figure 1.7. Therefore, Heaps' law and the vocabulary growth law are not equivalent

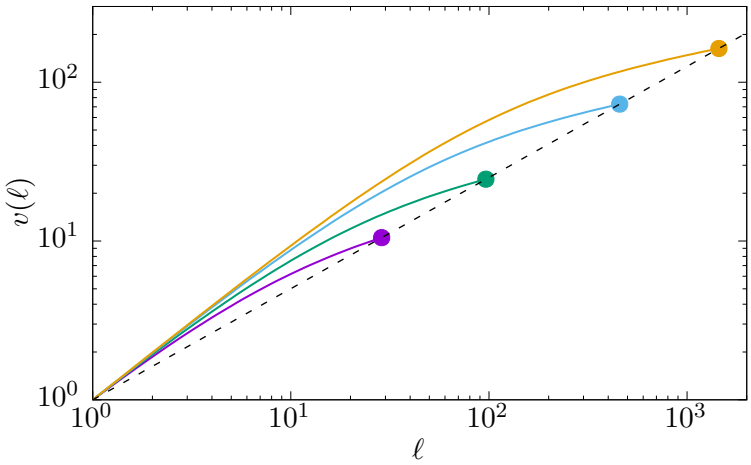

Figure 1.7 A pictorial representation of a collection of documents whose vocabulary growth curves $v_i(\ell)$, marked with solid coloured lines, are not power laws, but that fulfill Heaps' law, marked with symbols, in the sense of Eq. (1.34)

statements. Unfortunately, the distinction between both interpretations of the law is usually unclear in the literature. But if this distinction is taken into account, certain results concerning the relation between Zipf's law and Heaps' law[24] can be revisited. This is undertaken in §3.2, where we derive the exact form of the vocabulary growth law for systems where the frequency-count representation of Zipf's law holds exactly, and conclude that in this case, the vocabulary growth law is not a power law, but rather a more complicated expression in terms of the poly-logarithm function. Our results are confirmed with real corpora from the Project Gutenberg database.

[24] In the classic, ambiguous sense.

1.3 Bursty phenomena and thresholds

This section starts by introducing the "black-box" approach (§1.3.1), where the internal dynamics of a system are purposely disregarded, in favor of the study of a *signal* that emerges from the system. In the case of *bursty phenomena* (§1.3.2), this leads to the definition of *events* that turn out to display scale-invariant properties. The need to define the events via a *threshold* is exposed, and the consequences of this for the scaling exponents are briefly discussed (§1.3.4).

1.3.1 The "black-box" approach

Consider a signal that changes with time, say $a(t) > 0$. We use the label a to denote that $a(t)$ is "monitoring" the *activity* level or intensity of some underlying process, which is inaccessible to us and considered a "black-box". This is not to mean that we have zero information about the process behind the signal; rather, such a "black-box" approach is a *choice* that we do, and there might be several reasons for that. For example, it might happen that the physics behind the process are not quite well understood, or that there is some controversy in the literature regarding the mechanisms driving the process. Or it might be the case that, even if the process is thought to be well understood, the microscopic, internal mechanism cannot be observed, or observing them is too costly. In any case, these are just factors that might motivate our choice; but in the end, the approach is as legitimate as any other.

Let us take the case of financial markets as an example. Financial markets can be seen as complex systems where a large number of individual agents interact by buying and selling stock options, derivatives or other financial products. On top of this, there are plenty of external factors influencing the system in non-trivial ways: regional regulations, national political circumstances, international relations, war-related events, catastrophic natural events, lobbying, etc. Neither the interactions between agents nor the external factors are understood well enough to allow for a clear modeling strategy, which would in any case be quite involved. In addition, even in our times of big data, gathering and processing data at the level of the microscopic interactions would probably be out of reach. However, as a result of both the microscopic interactions between agents and the external factors, certain "global" quantities emerge: for example, the market price of a given commodity. This is a well-defined "observable", to which we have direct access, and from which we can obtain abundant and precise data. Thus in this example the financial market would be taken as a "black-box" process that generates a signal, the price $a(t)$ of a given commodity over time.

The key idea, however, is that by studying the properties of the signal, one can gain some insight on the nature of the underlying process. In the framework of Complexity Science, scale invariance (§1.1.2) and universality (§1.1.1) are properties considered of special interest; and thus if the signal $a(t)$ displays these properties, then this tends to be interpreted as a "mark" or "indication" of an underlying complex behavior. Clearly, this approach cannot by itself uncover the dynamics or any details of the underlying process, because no assumptions are made a priori about it. It is in this sense a more

conservative approach, which can be used in a more ample repertoire of situations, but whose conclusive power is more limited.

1.3.2 Bursty phenomena

We now focus our attention on certain phenomena that we dub, for lack of a better term, bursty phenomena. These are characterized by a tendency to occur in "bursts" of activity, combining periods of inactivity or very weak activity with periods of high-intensity activity, and the fact that such periods can last from very short times to extremely long ones, with all intermediate scales of duration and intensity being possible. Perhaps a good example is the case of rainfall: In plain words, sometimes it rains only for 5 minutes, sometimes it rains for an hour, and sometimes it rains for the whole day. We also know, from our every-day experience, that the intensity of rain seems to vary as well: it can go from very light rain where we might not even need an umbrella, to extremely intense rain that causes floods and, in some extreme cases, has devastating consequences. Besides the example of rainfall (Andrade et al., 1998; Peters et al., 2002; Peters and Christensen, 2002, 2006; Peters et al., 2010; Deluca and Corral, 2014), other examples of bursty phenomena in geosciences are earthquakes (Sornette and Sornette, 1989; Davidsen and Kwiatek, 2013; Kagan, 2010; Lippiello et al., 2012), and hurricanes (Corral, 2010); and –at least at a qualitative level– also solar flares (Baiesi et al., 2006; Boffetta et al., 1999; Paczuski et al., 2005), volcanic eruptions (Grasso and Bachélery, 1995), rock avalanches (Turcotte and Malamud, 2004) and forest fires (Malamud et al., 2005; Corral et al., 2008).

To put this qualitative description in a sound quantitative framework, let us say that the signal $a(t)$ corresponds to the rain rate in a given site, in units of *e.g.* millimeters per hour, mm/h. A typical plot of the rain rate over a long period of time is show in Figure 1.8. Notice that the signal $a(t)$, the rain rate in this example, is defined for all times, even if it takes the value 0 when it does not rain. Indeed, the bursty nature of rain makes it natural to think in terms of rain *events* rather than on a continuous, ongoing rain-rate time series: loosely speaking, a rain event starts "when it starts raining", and ends "when it ends raining". So it is all up to what is considered rain and what is not, which is actually a very delicate matter[25]. In any case, let us assume that we agree on how to define the start and the end of the rain events. Then it is easy to transform a rain rate time-series $a(t)$ into a series of "rain events", which we shall index by $i = 1, 2, \dots$, and which are characterized by a series of start-times and end-times, say $\{t_1, t_2, \dots\}$ and $\{t'_1, t'_2, \dots\}$ respectively, so that the i-th rain event starts at time t_i and ends at time t'_i. Then typically two observables are defined for each rain event: the event duration τ_i and the event size s_i,

$$\tau_i \equiv t'_i - t_i; \qquad s_i \equiv \int_{t_i}^{t'_i} a(t)\, dt \tag{1.38}$$

where the integration is generally substituted by a discrete sum if one deals with real datasets. The observable τ corresponds to the duration, in units of time, of one rain event, while s corresponds to the rain depth, *i.e.*, the total volume of rain per unit of

[25]For instance, it might be difficult to distinguish very weak rain from moisture

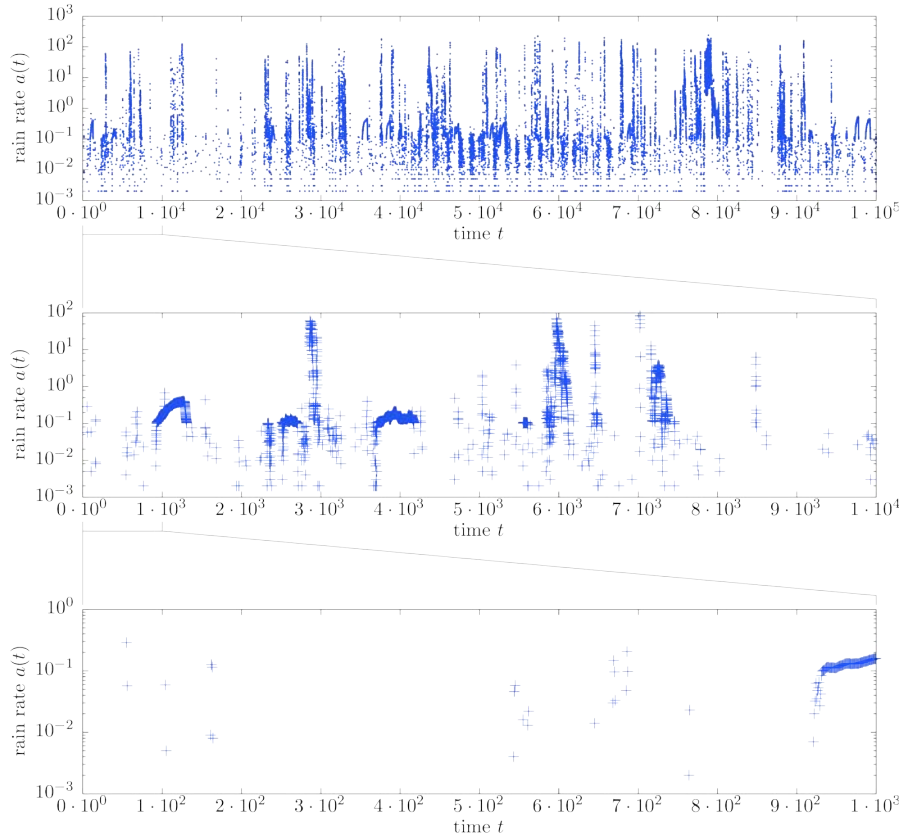

Figure 1.8 A real example of a rain-rate time series, shown at three different scales. The x-axis, time, is measured in minutes, while the rain rate in the y-axis is in units of mm/h. The dataset has a resolution of 1 minute. Missing points correspond to a recorded value of 0.

area. Now, assuming that we have collected rain data over a long period of time[26], and that we have an agreement in how to define the start and end of the rain events, we can construct a very large collection of rain events and, therefore, of associated durations $\{\tau_i\}$ and sizes $\{s_i\}$. If s and τ are regarded as random variables, then $\{\tau_i\}$ and $\{s_i\}$ constitute samples of s and τ, and we can estimate the probability distribution of s and τ,

$$\mathcal{P}(s)ds \simeq \text{Prob}[s \leq s_i < s + ds] \tag{1.39}$$

$$\mathcal{P}(\tau)d\tau \simeq \text{Prob}[\tau \leq \tau_i < \tau + d\tau]. \tag{1.40}$$

In the examples of bursty phenomena mentioned above, it turns out that $\mathcal{P}(s)$ and $\mathcal{P}(\tau)$ display scale-invariant properties, in the form of scaling laws as defined in §1.1.2. The fact that *e.g.* rainfall seems to display a scale-invariant distribution of event sizes, durations, or dry spells[27], see Figure 1.9, is considered interesting in the general complex systems

[26] There are datasets that span a period of over 10 years, with a 1 minute resolution of rain rate

[27] This observable, which we have not defined, measures the time *between* subsequent events.

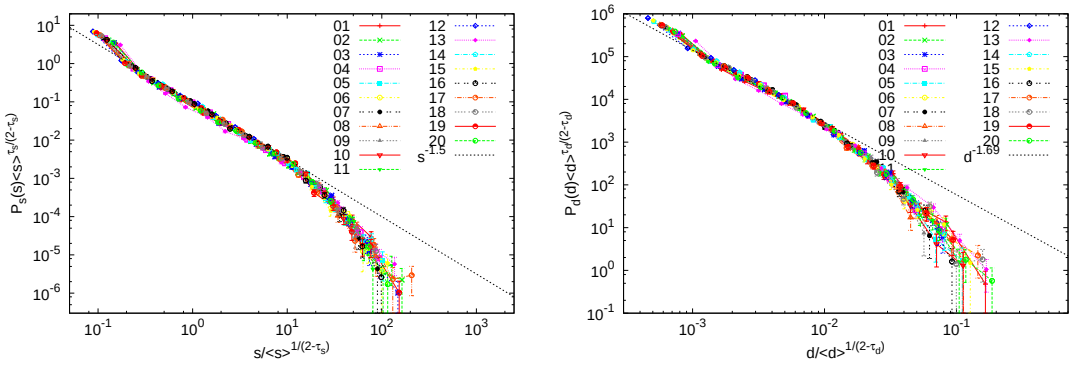

Figure 1.9 Data collapse of probability density functions of rain event sizes (left) and durations (right). Colors correspond to data from different stations in Catalonia (NE Spain), from the database maintained by the Agència Catalana de l'Aigua. Reproduced from Deluca and Corral (2014).

framework, but it is of particular interest for advocates of the theory of Self-Organized Criticality (SOC), see (Bak, 1996; Jensen, 1998; Pruessner, 2012) and many others. In short, SOC tries to give a good explanation of how and why the "critical point", a concept borrowed from the theory of critical phenomena, see §1.1.1, would be reached in other systems where some some analogies to critical phenomena have been drawn[28].

In this Thesis, however, we will *not* analyze real-world dataset of the aforementioned phenomena, nor will we enter the discussion of the case of rainfall or any other phenomena being a real-world example of SOC, or any other theory, on the basis of the possible scale invariance of some observables. Our interest will be in the *thresholding procedure*, which we introduce in what follows, and which we will analyze at a purely theoretical level. The findings of (Font-Clos et al., 2015, §3.3) provide a "warning", if one wishes, of the unexpected consequences that thresholding might have; in particular, the possibility that spurious exponents are measured due to the introduction of the threshold. This is discussed at depth in (Font-Clos et al., 2015, §3.3), but let us emphasize now how it relates with what we have explained so far: if *e.g.* rainfall is conjectured to be in some sense analogous to a critical phenomenon, for instance in the SOC-sense, then the scaling exponents are of capital importance, specially if one envisages to establish the existence universality classes as well. But we will see the threshold can, on the one hand, disturb the value of measured exponents in non-negligible ways; but it also is, on the other hand, an *unavoidable* step in the analysis of many real-world datasets.

[28]The issue is the following: while with ferromagnetic materials the critical temperature T_c is obviously not reached spontaneously, *i.e.*, we must heat the material to its critical temperature; it seems that with other systems the "critical point" is reached spontaneously, that is, the systems somehow poses itself into a "critical state". Obviously, this needs a sound explanation as to how it happens, and the theory of SOC gives one possibility for that.

1.3.3 Zero-defined events

So far we kept the example of rainfall explicit, but obviously the idea can be generalized to any signal $a(t)$. In short, a signal $a(t)$ can be transformed into a series of *events* as long as there is a convention on what marks the start and the end of events. In the case of bursty phenomena, which we are specially interested in, it seems natural to define events in the simplest possible way: the periods of inactivity, $a(t) = 0$, separate the events, and the periods of activity $a(t) > 0$, constitute the events. An illustrative depiction of this idealized situation is shown in Figure 1.10 (top). We will argue in what follows that

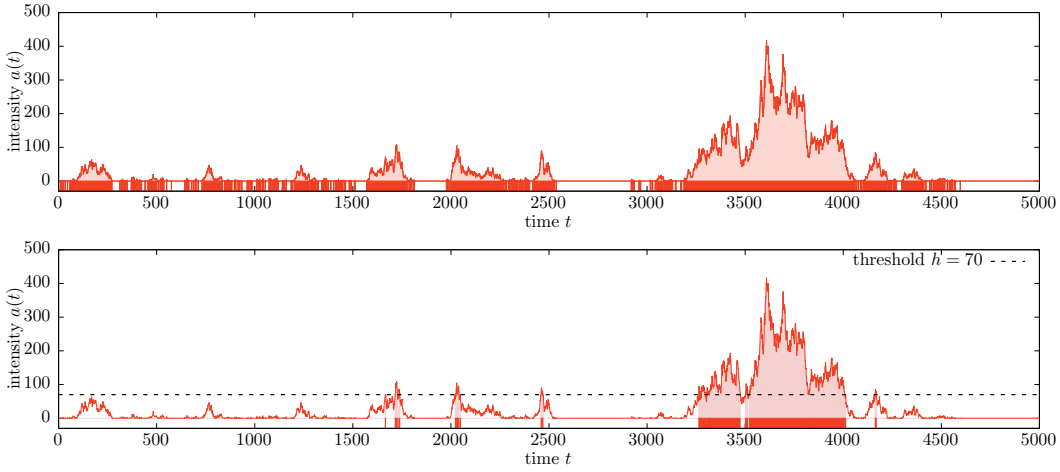

Figure 1.10 Representation of the signal generated by a bursty phenomenon, displaying large variability in its intensity $a(t)$ and in the patterns of activity/inactivity. **Top**: The original, raw signal. In this idealized example, the signal takes non-negative values, $a(t) \geq 0$, including $a(t) = 0$, and hence it defines in a natural way the periods of activity/inactivity, marked with alternating red/white strips below the $a(t) = 0$ line. **Bottom**: Alternatively, events can be defined via a threshold, set to $h = 70$ in this example.

such a procedure, although perfectly reasonable in theory, can be quite problematic when one deals with real-world measurements of certain phenomena. The technical reasons that render this "naïve" approach ill-advised vary from case to case, and are by necessity related to the nature of each phenomenon, the technical characteristics of the devices used to measure them, etc., We shall however briefly explain the most common reasons in a general manner. In short, using the value of intensity or activity *zero* to define events involves distinguishing $a(t) = 0$ from $a(t) \neq 0$, and this is problematic in many senses:

1. In a strict sense, the notion $a(t) \to 0$ involves crossing all length scales: from a macroscopic level to the mesoscale to the molecular and atomic scale; to the extend that the magnitude measured, a, might not be clearly defined at some point.

2. Even in the regime where $a(t)$ is well-defined in a physical sense, the *meaning* that we give to it cannot be distinguished from what is being measured. For example, in the case of rainfall, devices can measure the volume of water collected, but cannot

distinguish that from humidity and moisture, which contribute as well to the total volume of water, and which in principle we do not want to consider as rain.

3. The measuring devices have always a finite resolution, so that zero intensity can only be distinguished from non-zero intensity, in practice, up to certain degree.

1.3.4 Threshold-defined events

To solve the problems exposed above, an alternative way of defining events is the use of a *threshold*. The idea is simply to set a finite value $h > 0$, which we call the threshold, and below which the signal is regarded as zero. Formally, we can say that thresholding a signal $a(t)$ at threshold level $h > 0$ consists in the transformation:

$$a(t) \to \tilde{a}_h(t) := \Theta(a(t) - h)a(t), \tag{1.41}$$

where $\Theta(\cdot)$ denotes the step or Heaviside function. The thresholding procedure solves the problem of distinguishing zero from non-zero intensities by setting the signal to zero when it is below certain value h, and allows for an unambiguous definition of events. It obviously introduces new problems, because the threshold level h is in some sense arbitrary, and the new, thresholded signal, $\hat{a}_h(t)$, and the events thereof defined, depend on the value h. But before entering this discussion, which is what motivated the research exposed in (Font-Clos et al., 2015, §3.3), let us insist on the *necessity* of using thresholds in the analysis of *real-world* datasets.

First, catalogues of historical data often come with an implicit threshold, which cannot be "eliminated". This is sometimes related to the technical limits of the measuring devices, but also due to the necessity (in the past mostly) of limiting the amount of data stored. Second, the value of implicit thresholds is not always known due to incomplete (or missing) technical documentation. And finally, in large datasets that have been created by combining several datasets, or in studies that aim at comparing different datasets, each might have a different threshold. In all these cases, it might be desirable to further threshold the data, *i.e.*, to introduce a new threshold that, although by necessity higher than the implicit threshold(s), at least will be *known*, and will be homogeneous along the different datasets. Finally, let us mention the cases of solar flares and financial markets, where the signal $a(t)$ actually never ceases, *i.e.*, $a(t) > 0 \, \forall t$. In such a situation, one cannot possibly define events without introducing a threshold.

In summary, thresholding is a procedure that allows for an unambiguous definition of events and that, in many cases, cannot be avoided. But the introduction of the threshold has the consequence that, in principle, all measured quantities, including those of interest for us like the distribution of event sizes and durations, become threshold-dependent,

$$\mathcal{P}(s) \to \mathcal{P}_h(s); \qquad \mathcal{P}(\tau) \to \mathcal{P}_h(\tau). \tag{1.42}$$

In particular, this includes the values of scaling exponents, in the cases where s, τ or other observables display scale invariance. Notice that the values of the scaling exponents must be fitted from finite-length datasets, so the picture becomes quite involved: to the not-so-simple issue of fitting power-law distributions, see again Clauset et al. (2009);

Deluca and Corral (2013), we must now add the issue of the dependence of the exponent with the value of the threshold.

Interestingly, a simple heuristic argument can be constructed to conclude that scaling exponents should *not* depend on the value of threshold h, and it goes as follows: the scaling exponent is actually a statement about the asymptotically large events, and given that the value of h must be finite, those large events will be almost unaffected by the threshold. The key point behind this argument is that for *any* finite value of h, there are always large enough events whose value (size, or duration) does not change significantly due to the introduction of the threshold at level h. The drawback of this argument, which otherwise seems to stand perfectly, is that it implicitly assumes infinite amount of data: it admits that the threshold might modify the distribution of interest, say $\mathcal{P}_h(\tau)$, up to certain finite scales of τ, but that, for fixed h, there are always large enough events whose duration τ is effectively unaffected by h. And because these are the events that determine the value of the *asymptotic* scaling exponent, the argument concludes that the exponent should not depend on the threshold.

This view is challenged in detail in (Font-Clos et al., 2015, §3.3) by means of a simple stochastic process (the birth-death process). It is shown that the asymptotic exponents, when measured in a finite dataset, can be greatly disturbed by the introduction of the threshold. Interestingly, in the case of the birth-death process, a new scaling region, with a new, well-defined exponent, appears in the "intermediate" region of event durations. Further analysis and conclusions related to this publication are placed in §2.4.

CHAPTER 2

Conclusion

The main results of this Thesis are discussed in relation to the notion of scale invariance, and within the wider framework of Complexity Science. The chapter starts with a summary of the results of the three publications composing this Thesis. It is then shown that certain results can be recast into scaling laws of two variables, and a table summarizing the associated scaling exponents and scaling functions is provided. Finally, some further conclusions are outlined.

2.1 Summary of results

The first publication of this Thesis, (Font-Clos et al., 2013, §3.1), studies the dependence of Zipf's exponent γ with the system size, *i.e.*, the document length L. This is of particular interest in relation to the concept of universality in critical phenomena, see §1.1.1, where universality classes are established on the basis of scaling exponents and scaling functions. The main result of (Font-Clos et al., 2013, §3.1) can be stated as follows:

The exponent of Zipf's law γ in the frequency-count representation is independent of the document length L. This is a consequence of the fact that the word-frequency distribution $D_L(n)$ has a *shape* that is independent of the system size L and the vocabulary V_L, and it is only the *scale* of $D_L(n)$ that depends on L and V_L as follows:

$$D_L(n) = \frac{g(n/L)}{LV_L}, \tag{2.1}$$

with $g(\cdot)$ an arbitrary function of one variable (presumably containing the power law tail with exponent γ).

The rest of results are summarized as follows:

1. Equation (2.1) is seen to yield to excellent *data collapses*, see §1.1.2, both for non-lemmatized and for lemmatized texts.

2. In the case of lemmatized texts, a double power law approximates well the analyzed corpus. In this case the scaling function $g(\cdot)$ takes the form

$$g(x) \propto \frac{1}{x(a + x^{\gamma-1})}, \tag{2.2}$$

3. A rough approximation of the vocabulary growth law, see §1.2.3, can be obtained from Equation (2.1):

$$V_L = \int_{1/L}^{\infty} g(x)dx \tag{2.3}$$

The second publication of this Thesis (Font-Clos and Corral, 2015, §3.2), studies the vocabulary growth curve of Zipf's systems. Although the case of language is a natural example, the scope of the paper is more general, and includes any system where certain tokens can be grouped into types, and where Zipf's law is known to hold in the frequency-count representation (which is assumed to be a pure power law with exponent γ). The main result of (Font-Clos and Corral, 2015, §3.2) is the following:

The vocabulary growth curve $v(\ell)$ fulfills a universal data collapse that depends only on Zipf's exponent γ, after proper rescaling. Its exact form is predicted theoretically and yields excellent agreement with real data. In particular, in a system of size ℓ, total size L and total vocabulary V, the vocabulary growth curve $v(\ell)$ is given by

$$v(\ell) \simeq V\left(1 - \frac{\mathrm{Li}_\gamma(1-\ell/L)}{\zeta(\gamma)}\right). \tag{2.4}$$

This theoretical prediction is derived by assuming that Zipf's law in the frequency-count representation holds strictly, as well as a random placement of tokens in the system. In addition:

1. The theoretical prediction that Eq. (2.4) constitutes yields to excellent agreement with real data drawn from the Project Gutenberg database. This is remarkable, because of the presence of long-range correlations in real documents, contrary to the randomness assumption in the theoretical derivations.

2. A careful analysis of inter-occurrence distance distribution reveals that the first instance of a word behaves as in a random system, while subsequent occurrences display signs of clustering or long-range correlations.

3. It is hence understood why a prediction based on the randomness hypothesis yields to such a good agreement when tested against a dataset drawn from a real, non-random system.

The third and last publication of this Thesis (Font-Clos et al., 2015, §3.3), studies the *possible* effects of applying thresholds to time-series of bursty phenomena. In particular, it provides a counter-example to the claim that the threshold cannot change the scaling exponent of some relevant observables of the events, which are defined via the threshold itself. The main result of (Font-Clos et al., 2015, §3.3) is the following:

Thresholding the birth-death process introduces a spurious scaling region in the distribution of event durations. This scaling region has a scaling exponent of $-3/2$, while the scaling exponent of the original birth-death process is well-known to be -2. In addition, the distribution of durations of a thresholded birth-death process, with threshold level h, fulfills the following scaling law:

$$\mathcal{P}^{g_s}(g_s; h) \simeq 2 g_s^{-2} \mathcal{G}_>(g_s/h); \qquad \text{for } g_s \gg 1/h, \tag{2.5}$$

where g_s is the duration of the process, h the threshold level, and $\mathcal{G}_>$ the scaling function.

The rest of results are summarized as follows:

1. The Laplace transform of $\mathcal{P}^{g_s}(g_s; h)$ can be analytically computed for $g_s \gg 1$,

$$\hat{\mathcal{P}}(u; h) = \int_0^\infty dg_s e^{-g_s u} \mathcal{P}^{g_s}(g_s; h) = \frac{\sqrt{u(h+1)} K_1\left(2\sqrt{2}\sqrt{u(h+1)}\right)}{\sqrt{uh} K_1\left(2\sqrt{2}\sqrt{uh}\right)}, \tag{2.6}$$

   with $K_1(\cdot)$ the modified Bessel function of the first kind. This allows to numerically evaluate the scaling function $\mathcal{G}_>(\cdot)$.

2. An alternative scaling law, valid for short durations, can be formulated as follows:

$$\mathcal{P}^{g_s}(g_s; h) \simeq \frac{g_s^{-3/2}}{\sqrt{2\pi h}} \mathcal{G}_<(g_s h); \qquad \text{for } g_s \ll 8\pi h. \tag{2.7}$$

3. The original scaling region of the process, with scaling exponent -2, is recovered only for $g_s \gg 8\pi h$. Thus the crossover between the two scaling regions scales linearly with the threshold.

4. If the sample size is not large enough, even sophisticated fitting methods fail to capture the "true" asymptotic exponent of -2. Instead, fitted values close to the "spurious" exponent $-3/2$ are obtained.

5. The origin of the $-3/2$ spurious scaling is understood to lie in the random walk embedded in the process. For high enough values of the threshold h, relative changes in position are small, and the additive nature of the random walk "supersedes" the multiplicative nature of the original birth-death process. It is conjectured that the above discussion applies more generally to other multiplicative processes.

2.2 Results as scaling laws: a unifying picture

It is shown that some of the main results of this Thesis might be expressed in the form of scaling laws. Thus, the scale-invariance of certain properties of the systems under study is established.

As demonstrated in §1.1.2, a scale-invariant function of two variables $f(x_1, x_2)$ must take the form

$$f(x_1, x_2) = |x_1|^{1/b_1} \mathcal{G}_{\pm}\left(\frac{x_2}{|x_1|^{b_2/b_1}}\right), \tag{2.8}$$

which is known as a *scaling law*. The function $\mathcal{G}(\cdot)$ is the *scaling function*, and $b_1, b_2 \in \mathbb{R}$ are the *scaling exponents*. The triplet $\{\mathcal{G}; b_1, b_2\}$ hence completely determines the scaling law, and the function $f(x_1, x_2)$ is invariant under the *scale transformation*

$$\begin{aligned} x_1 &\to x_1/\lambda^{b_1} \\ x_2 &\to x_2/\lambda^{b_2} \\ f(x_1, x_2) &\to f(x_1, x_2)/\lambda \end{aligned} \tag{2.9}$$

for any $\lambda \in \mathbb{R}^+$. We will now show that the main results of §3.2 and §3.3, Equations (2.4) and (2.5) respectively, can be expressed as scaling laws. The main result of §3.1, Equation (2.1), does *not* fit into the definition of a scaling law if no additional assumptions are taken. To proceed in a clear and consistent manner, let us adopt the following convention: we will keep the original notation of the publications for the *variables*, but we will adopt the notation of Eq. (2.8) for the *functions*. Thus x_1, x_2 in Eq. (2.8) will be substituted by the corresponding variables of each publication, but the scale invariant function f and the scaling function $\mathcal{G}_{\pm}$ will not.

The vocabulary growth of Zipf's systems as a scaling law

The main result of (Font-Clos and Corral, 2015, §3.2),

$$v(\ell) \simeq V\left(1 - \frac{\mathrm{Li}_\gamma(1 - \ell/L)}{\zeta(\gamma)}\right), \tag{2.10}$$

gives the (average) vocabulary growth curve $v(\ell)$ in a system where Zipf's law in the frequency-count representation holds strictly, *i.e.*,

$$N(n) \propto \frac{1}{n^\gamma}; \qquad n = 1, 2, \ldots \tag{2.11}$$

The system comprises V types, with frequencies $\{n_1, \ldots, n_V\}$ drawn from Eq. (2.11), which yield a total of L tokens,

$$L = \sum_{i=1}^{V} n_i. \tag{2.12}$$

In this sense, L is regarded as the sum of V random variables, and is hence also a random variable. We now discuss the asymptotic scaling of L with V, which was not treated in

(Font-Clos and Corral, 2015, §3.2). This will allow us to rewrite Eq. (2.10) as a scaling law. If n, as a random variable, has finite mean, *i.e.*, if $\gamma > 2$, then it is trivial to see that L scales linearly with V. To be precise,

$$\lim_{V\to\infty} \frac{L}{V} = \mathrm{E}[n] < \infty, \tag{2.13}$$

which is basically the law of large numbers. We will simply write $L \sim V$ to denote such an asymptotic scaling. However in the case of $1 < \gamma < 2$ the mean diverges, and the above reasoning cannot be applied. Intuitively, one needs to rescale L with some power of V, say V^μ with $\mu > 1$, because otherwise L grows "too fast", and the sum does not converge. If the right power is chosen, then L/V^μ converges to a non-degenerate distribution. The Generalized Central Limit Theorem, see Bouchaud and Georges (1990), says that the right power is $\mu = 1/(\gamma - 1)$. More precisely, if $1 < \gamma < 2$, then

$$\lim_{V\to\infty} \frac{L}{V^{\frac{1}{\gamma-1}}} \to \mathcal{S}(\gamma), \tag{2.14}$$

where $\mathcal{S}(\gamma)$ is a random variable[1] that does *not* depend on V. We will write $L \sim V^{\frac{1}{\gamma-1}}$, or, equivalently, $V \sim L^{\gamma-1}$, to denote this asymptotic scaling relation. In summary, we have just shown that, for large V and hence large L,

$$V \sim \begin{cases} L^{\gamma-1} & \text{for } 1 < \gamma < 2 \\ L & \text{for } \gamma > 2 \end{cases} \tag{2.15}$$

Notice that Eq. (2.15) is a statement[2] about Heaps' law, in the sense of Eq. (1.34) in §1.2.3 and hence it does not contradict our original result Eq. (2.10), which refers to the vocabulary growth law in the sense of Eq. (1.35). We can finally go back to the main result of (Font-Clos and Corral, 2015, §3.2), substituting (2.15) into (2.10), to obtain

$$v(\ell, L) \propto \begin{cases} L^{\gamma-1}\left(1 - \frac{\mathrm{Li}_\gamma(1-\ell/L)}{\zeta(\gamma)}\right) & \text{for } 1 < \gamma < 2 \\ L\left(1 - \frac{\mathrm{Li}_\gamma(1-\ell/L)}{\zeta(\gamma)}\right) & \text{for } \gamma > 2 \end{cases} \tag{2.16}$$

Notice that we have made explicit the dependence of v with L, because strictly speaking v is a function of ℓ and L, $v \equiv v(\ell, L)$. It is now clear that (2.16) is a scaling law. Defining $\mathcal{G}_1(y) \equiv a_\gamma(1 - \mathrm{Li}_\gamma(1-y)/\zeta(\gamma))$, with a_γ a proportionality constant that depends on γ, $v(\ell, L) \equiv f_1^<(\ell, L)$ for $1 < \gamma < 2$, and $v(\ell, L) \equiv f_1^>(\ell, L)$ for $\gamma > 2$, we get to

$$f_1^<(\ell, L) \simeq L^{\gamma-1}\mathcal{G}_1(\ell/L); \qquad 1 < \gamma < 2 \tag{2.17}$$

$$f_1^>(\ell, L) \simeq L\ \mathcal{G}_1(\ell/L); \qquad 2 < \gamma \tag{2.18}$$

Therefore, the vocabulary growth curve is a scale-invariant property of Zipf's systems, under the assumption of a perfect power-law distribution of frequencies, independence between frequencies of different types, and random ordering in the system.

[1] In particular, $\mathcal{S}(\gamma)$ is known as a stable distribution, and it has a power-law *tail* with exponent γ.

[2] Under the following assumptions: a strict power law for Zipf's law in the frequency-count representation, and $\{n_i\}_{i=1}^V$ a set of i.i.d random variables.

The distribution of durations of a thresholded birth-death process as a scaling law

We now turn to the main data collapse in (Font-Clos et al., 2015, §3.3),

$$\mathcal{P}\left(g,h\right) \simeq 2g^{-2}\mathcal{G}_{>}\left(g/h\right); \qquad \text{for } g \gg 1/h. \tag{2.19}$$

where we have defined $g \equiv g_s$ and $\mathcal{P}(g,h) \equiv \mathcal{P}^{g_s}(g_s;h)$ to ease the notation. Remind that g corresponds to the *duration* of events in a birth-death process, and that the events are defined via a threshold at level h. Clearly, Eq. (2.19) is already in the form of a scaling law, but we might as well define $\mathcal{G}_2(y) \equiv 2y^{-2}\mathcal{G}_{>}(y)$ and $f_2(g,h) \equiv \mathcal{P}(g,h)$, to get to an equivalent formulation:

$$f_2(g,h) \simeq h^{-2}\mathcal{G}_2\left(g/h\right); \qquad \text{for } g \gg 1/h. \tag{2.20}$$

The alternative data collapse,

$$\mathcal{P}\left(g,h\right) \simeq \frac{g^{-3/2}}{\sqrt{2\pi h}}\mathcal{G}_{<}\left(gh\right); \qquad \text{for } g \ll 8\pi h, \tag{2.21}$$

can also be recast into another scaling law, by considering $\mathcal{G}_3 \equiv (2\pi)^{-1/2}y^{-3/2}\mathcal{G}_{<}(y)$, so that

$$f_3(g,h) \simeq h\,\mathcal{G}_3\left(gh\right); \qquad \text{for } g \ll 8\pi h. \tag{2.22}$$

Notice that these manipulation had the sole objective of bringing the results of §3.2 and §3.3 into the same form, so that scaling exponents, scaling functions and the rôle that the different variables play can be compared, but the original scaling laws presented in (Font-Clos et al., 2015, §3.3) are more appropriate to discuss the effects of thresholding the birth-death process. The asymptotic properties of the scaling functions $\mathcal{G}_2(\cdot)$, $\mathcal{G}_2(\cdot)$ are given in Table 2.1.

In summary, it has been shown how certain results of this Thesis can be expressed as scaling laws. Table 2.1 summarizes the scaling laws obtained, Equations (2.17,2.18,2.20) and (2.22), including the scaling functions and the scaling exponents.

2.3 Table of scaling laws

Reference	Original Eq.	Scaling law	Variables		Exponents		Scaling function
		$f(x_1,x_2) = x_2^{1/b_2}\mathcal{G}\left(x_1\ x_2^{-b_1/b_2}\right)$	x_1	x_2	b_1	b_2	$\mathcal{G}(y)$
Eq. (9) in **P[2]**	$v(\ell) = V\left(1-\frac{\mathrm{Li}_\gamma(1-\ell/L)}{\zeta(\gamma)}\right)$	$f_2^<(\ell,L) = L^{\gamma-1}\mathcal{G}_2(\ell/L)$	ℓ	L	$\frac{1}{\gamma-1}$	$\frac{1}{\gamma-1}$	$1-\frac{\mathrm{Li}_\gamma(1-y)}{\zeta(\gamma)}$
		$f_2^>(\ell,L) = L\ \mathcal{G}_2(\ell/L)$			1	1	
Eq. (18) in **P[3]**	$\mathcal{P}^{g_s}(g_s;h) \simeq 2g_s^{-2}\mathcal{G}_>(g_s/h)$	$f_3(g,h) = h^{-2}\mathcal{G}_3(g/h)$	g	h	$-1/2$	$-1/2$	$(*)$
Eq. (19) in **P[3]**	$\mathcal{P}^{g_s}(g_s;h) \simeq \frac{g_s^{-3/2}}{\sqrt{2\pi h}}\mathcal{G}_<(g_s h)$	$f_4(g,h) = h\,\mathcal{G}_4(g\cdot h)$			-1	1	$(\dagger)$

Table 2.1 The main results of this Thesis, in the form of scaling laws of two variables. To obtain the scaling laws in the third column from the original equations in the second column, the relation $V \sim L^{\min\{1,\gamma-1\}}$ was used, see §2.2

Regarding **Scaling function**:
$(*)$: Can be computed by numerically inverting the Laplace transform given in Eq. (2.6) (Equation (D.4) of **P[3]**). Its asymptotic behavior can be deduced from Eq. (18) in **P[3]**: $\mathcal{G}_2(y) \simeq 2y^{-2}$ for large y and $\mathcal{G}_2(y) \simeq (2\pi)^{-1/2}y^{-3/2}$ for small y.
$(\dagger)$: Its exact form is unknown, but the asymptotic behavior can also be deduced from Eq.(19) in **P[3]**: $\mathcal{G}_3(y) \simeq (2\pi)^{-1/2}y^{-3/2}$ for large y and $\mathcal{G}_3(y) \simeq 1/2$ for small y.

Regarding **Reference**:
P[2]: Font-Clos and Corral (2015) and §3.2
P[3]: Font-Clos et al. (2015) and §3.3

2.4 Further conclusions

This section is concerned with conclusions beyond the ones of §3.1, §3.2 and §3.3. The publications that form this Thesis have each their specific context (natural language, general Zipf's systems and bursty phenomena) and their conclusions are hence framed accordingly. It is hence only left, but also, mandatory, to give a more global, encompassing view of the results of this Thesis, in light of the unifying picture just presented.

The object of this Thesis was the study of the scale invariance of certain complex systems. It has been shown that, under certain hypothesis, the vocabulary growth curve $v(\ell, L)$ of Zipf's systems and the distribution of durations $\mathcal{P}(g, h)$ of a thresholded birth-death process take the form of a scaling law, and are hence scale invariant. The distribution of frequencies $D(n, L) \equiv D_L(n)$ as defined in the first publication of this Thesis, however, cannot be cast into a scaling law without further assumptions, and hence cannot be said to be scale-invariant (in a strict sense).

The notion of scale invariance tends to be associated with power-law distributions or power-law functional relations, and this is probably because, *in one dimension*, all scale-invariant functions are power laws. However, as explained in the introduction, when more variables are taken into account, then scale-invariant functions take the more general form of scaling laws. Notice how much richer are scaling laws compared to power laws: the scaling function $\mathcal{G}(\cdot)$ can take *any* form.

The distribution of durations of a thresholded birth-death process $\mathcal{P}(g, h)$ provides a very good example: if one only looks at g, then it looks like scale invariance is "broken": $\mathcal{P}(g, h)$ is not a power law of g, the threshold has introduced a characteristic scale (given by $8\pi h$) and scale invariance is not fulfilled, in this sense. But if the threshold h is incorporated into the system as a variable, then scale invariance is "recovered": $\mathcal{P}(g, h)$ is a scaling law, Eq. (2.20), and the scaling function takes as argument the combination g/h. Indeed, one might say that, if the durations are measured "in units of the threshold", then the characteristic scale introduced by the threshold disappears, and scale-invariance is preserved. But notice that, the scaling function has *another* power-law (left) tail, with a different exponent. All together, we learn that systems displaying several scaling regimes, each with a different exponent (*i.e.*, "double power laws", or beyond), might still be susceptible of being scale-invariant, if the right variables are taken into account.

As for the vocabulary growth curve, the scaling function is visually similar to a power law, but it displays certain convexity in log-log space. In the past, this (slight) convexity has been either purposely ignored, or attributed to other factors, or simply considered an approximate result, all in an attempt to claim –no matter what the evidence showed– that the vocabulary grows as a power law of the text length. But the interest, from a Complexity Science point of view, should be (among other things) on the scale invariance of the systems under study, not their "power-lawness". After all, power laws are not a priori better than other functions –it is their relation with scale invariance that makes them more interesting. Why then this "bias" towards finding power laws when the evidence suggests otherwise

and when, actually, a scaling law could better accommodate the data? This seems rather unjustified, particularly in cases such as the vocabulary growth law, where under some assumptions the scaling law can be theoretically derived, and the scaling function exactly calculated.

In summary, scale invariance cannot be ruled out just on the basis of a deviation from a power law in the observations: the analysis might be missing an important variable that, when taken into account, allows for the results to fall under the umbrella of a scaling law.

CHAPTER 3

Publications

3.1 A scaling law beyond Zipf's law and its relation to Heaps' law

Francesc Font-Clos, Gemma Boleda and Álvaro Corral
New Journal of Physics **15** (2013) 093033

New Journal of Physics
The open access journal for physics

A scaling law beyond Zipf's law and its relation to Heaps' law

Francesc Font-Clos[1,2,4], **Gemma Boleda**[3] **and Álvaro Corral**[1]

[1] Centre de Recerca Matemàtica, Edifici C, Campus Bellaterra, E-08193 Bellaterra, Barcelona, Spain
[2] Department de Matemàtiques, Universitat Autònoma de Barcelona, Edifici C, E-08193 Bellaterra, Barcelona, Spain
[3] Department of Linguistics, The University of Texas at Austin, 1 University Station B5100, Austin, TX, USA
E-mail: fontclos@crm.cat

New Journal of Physics **15** (2013) 093033 (16pp)
Received 4 March 2013
Published 23 September 2013
Online at http://www.njp.org/
doi:10.1088/1367-2630/15/9/093033

Abstract. The dependence on text length of the statistical properties of word occurrences has long been considered a severe limitation on the usefulness of quantitative linguistics. We propose a simple scaling form for the distribution of absolute word frequencies that brings to light the robustness of this distribution as text grows. In this way, the shape of the distribution is always the same, and it is only a scale parameter that increases (linearly) with text length. By analyzing very long novels we show that this behavior holds both for raw, unlemmatized texts and for lemmatized texts. In the latter case, the distribution of frequencies is well approximated by a double power law, maintaining the Zipf's exponent value $\gamma \simeq 2$ for large frequencies but yielding a smaller exponent in the low-frequency regime. The growth of the distribution with text length allows us to estimate the size of the vocabulary at each step and to propose a generic alternative to Heaps' law, which turns out to be intimately connected to the distribution of frequencies, thanks to its scaling behavior.

[4] Author to whom any correspondence should be addressed.

Contents

1. Introduction

Zipf's law is perhaps one of the best pieces of evidence about the existence of universal physical-like laws in cognitive science and the social sciences. Classic examples where it applies include the population of cities, company income and the frequency of words in texts or speech [1]. In the latter case, the law is obtained directly by counting the number of repetitions, i.e. the absolute frequency n, of all words in a long enough text, and assigning increasing ranks, $r = 1, 2, \ldots$, to decreasing frequencies. When a power-law relation

$$n \propto \frac{1}{r^{\beta}}$$

holds for a large enough range, with the exponent β more or less close to 1, Zipf's law is considered to be fulfilled (with $\propto$ denoting proportionality). An equivalent formulation of the law is obtained in terms of the probability distribution of the frequency n, such that it plays the role of a random variable, for which a power-law distribution

$$D(n) \propto \frac{1}{n^{\gamma}}$$

should hold, with $\gamma = 1 + 1/\beta$ (taking values close to 2) and $D(n)$ as the probability mass function of n (or the probability density of n, in a continuous approximation) [2–6]. Note that this formulation implies performing double statistics (i.e. doing statistics twice), first counting words to get frequencies and then counting repetition of frequencies to get the distribution of frequencies.

The criteria for the validity of Zipf's law are arguably rather vague (long enough text, large enough range, exponent β more or less close to 1). Generally, a long enough text means a book, a large range can be a bit more than an order of magnitude and the proximity of the exponent β to 1 translates into an interval (0.7,1.2), or even beyond that [6–8]. Moreover, no rigorous methods have been usually required for the fitting of the power-law distribution. Linear regression in double-logarithmic scale is the most common method, either for $n(r)$ or for $D(n)$, despite the fact that it is well known that this procedure suffers from severe drawbacks and can lead to flawed results [9, 10]. Nevertheless, once these limitations are assumed, the fulfillment of Zipf's law in linguistics is astonishing, being valid no matter the author, style or language [1, 6, 7]. So, the law is universal, at least in a qualitative sense.

At a theoretical level, many different competing explanations of Zipf's law have been proposed [6], such as random (monkey) typing [11, 12], preferential repetitions or proportional growth [13–15], the principle of least effort [1, 16–18], and, beyond linguistics, Boltzmann-type approaches [19] or even avalanche dynamics in a critical system [20]; most of these options have generated considerable controversy [21–23]. In any case, the power-law behavior is the hallmark of scale invariance, i.e. the impossibility to define a characteristic scale, either for frequencies or for ranks. Although power laws are sometimes also referred to as scaling laws, we will make a more precise distinction here. In short, a scaling law is any function invariant under a scale transformation (which is a linear dilation or contraction of the axes). In one dimension the only scaling law is the power law, but this is not true with more than one variable [24]. Note that in text statistics, other variables to consider in addition to frequency are the text length L (the total number of words, or tokens) and the size of the vocabulary V_L (i.e. the number of different words, or types).

Somehow related to Zipf's law is Heaps' law (also called Herdan's law [25, 26]), which states that the vocabulary V_L grows as a function of the text length L as a power law

$$V_L \propto L^{\alpha}$$

with the exponent α smaller than one. However, even simple log–log plots of V_L versus L do not show a convincing linear behavior [27] and therefore, the evidence for this law is somewhat weak (for a notable exception see [5]). Nevertheless, a number of works have derived the relationship $\beta = 1/\alpha$ between Zipf's and Heaps' exponents [2, 5, 28], at least in the infinite-system limit [29, 30], using different assumptions.

Despite the relevance of Zipf's law, and its possible relations with criticality, few systematic studies about the dependence of the law on system size (i.e. text length) have been carried out. It was Zipf himself [1, pp. 144] who first observed a variation in the exponent β when the system size was varied. In particular, 'small' samples would give $\beta < 1$, while 'big' ones yielded $\beta > 1$. However, that was attributed to 'undersampling' and 'oversampling', as Zipf believed that there was an optimum system size under which all words occurred in proportion to their theoretical frequencies, i.e. those given by the exponent $\beta = 1$. This increase of β with L has been confirmed later, see [25, 31], leading to the conclusion that the practical usefulness of Zipf's law is rather limited [25].

More recently, using rather large collections of books from single authors, Bernhardsson *et al* [32] find a decrease of the exponents γ and α with text length, in correspondence with the increase in β found by Zipf and others. They propose a size-dependent word-frequency distribution based on three main assumptions:

(i) The vocabulary scales with text length as $V_L \propto L^{\alpha(L)}$, where the exponent $\alpha(L)$ itself depends on the text length. Note however that this is not an assumption in itself, just notation, and it is also equivalent to writing the average frequency $\langle n \rangle = L/V_L$ as $\langle n(L) \rangle \propto L^{1-\alpha(L)}$.

(ii) The maximum frequency is proportional to the text length, i.e. $n_{\max} = n(r = 1) \propto L$.

(iii) The functional form of the word frequency distribution $D_L(n)$ is that of a power law with an exponential tail, with both the scale parameter $c(L)$ and the power-law exponent $\gamma(L)$

depending on the text length L. That is

$$D_L(n) = A\frac{e^{-n/c(L)}}{n^{\gamma(L)}}$$

with $1 < \gamma(L) < 2$.

Taking $c(L) = c_0 L$ guarantees that $n_{\max} \propto L$; moreover, the form of $D_L(n)$ implies that, asymptotically, $\langle n(L)\rangle \propto L^{2-\gamma(L)}$ [24], which comparing to assumption (i) leads to

$$\alpha(L) = \gamma(L) - 1,$$

so, $0 < \alpha(L) < 1$. This relationship between α and γ is in agreement with previous results if L is fixed [2, 29, 30]. It was claimed in [32] that $\alpha(L)$ decreases from 1 to 0 for increasing L and therefore $\gamma(L)$ decreases from 2 to 1. The resulting functional form

$$D_L(n) = A\frac{e^{-n/(c_0 L)}}{n^{1+\alpha(L)}}$$

is in fact the same functional form appearing in many critical phenomena, where the power-law term is limited by a characteristic value of the variable, $c_0 L$, arising from a deviation from criticality or from finite-size effects [24, 33–35]. Note that this implies that the tail of the frequency distribution is not a power law but an exponential one, and therefore the frequency of most common words is not power-law distributed. This is in contrast with recent studies that have clearly established that the tail of $D_L(n)$ is well modeled by a power law [9, 36]. However, what is most uncommon about this functional form is the fact that it has a 'critical' exponent that depends on system size. The values of exponents should not be influenced by external scales. So, here we look for an alternative picture that is more in agreement with typical scaling phenomena.

Our proposal is that, although the word-frequency distribution $D_L(n)$ changes with system size L, the *shape* of the distribution is independent of L and V_L, and only the *scale* of $D_L(n)$ changes with these variables. This implies that the shape parameters of $D_L(n)$ (in particular, any exponent) do not change with L; only one scale parameter changes with L, increasing linearly. This is explained in section 2, while section 3 one is devoted to the validation of our scaling form in real texts, using both plain words and their corresponding lemma forms; in the latter case an alternative to Zipf's law can be proposed, consisting of a double power-law distribution (which is a distribution with two power-law regimes that have different exponents). Our findings for words and lemmas suggest that the previous observation that the exponent in Zipf's law depends on text length [25, 31, 32], might be an artifact of the increasing weight of a second regime in the distribution of frequencies beyond a certain text length. Section 4 investigates the implications of our scaling approach for Heaps' law. Although the scaling ansatz we propose has a counterpart in the rank-frequency representation, we prefer to illustrate it in terms of the distribution of frequencies, as this approach has been deemed more appropriate from a statistical point of view [36].

2. The scaling form of the word-frequency distribution

Let us come back to the rank-frequency relation, in which the absolute frequency n of each type is a function of its rank r. Defining the relative frequency as $x \equiv n/L$ and inverting the

relationship, we can write

$$r = G_L(x).$$

Note that here we are not assuming a power-law relationship between r and x, just a generic function G_L, which may depend on the text length L. Instead of the three assumptions introduced by Bernhardsson *et al* we just need one assumption, which is the independence of the function G_L with respect to L; so

$$r = G(n/L). \tag{1}$$

This turns out to be a scaling law, with $G(x)$ a scaling function. It means that if in the first 10 000 tokens of a book there are five types with relative frequency larger than or equal to 2%, that is, $G(0.02) = 5$, then this will still be true for the first 20 000 tokens, and for the first 100 000 and for the whole book. These types need not necessarily be the same ones, although in some cases they might be. In fact, instead of assuming as in [32] that the frequency of the most used type scales linearly with L, what we assume is just that this is true for all types, at least on average. Notice that this is not a straightforward assumption, as, for instance [5], considers instead that n is just a (particular) function of r/V_L.

Now let us introduce the survivor function or complementary cumulative distribution function $S_L(n)$ of the absolute frequency, defined in a text of length L as $S_L(n) = \text{Prob}[\text{frequency} \geqslant n]$. Note that, estimating from empirical data, $S_L(n)$ turns out to be essentially the rank, but divided by the total number of ranks, V_L, i.e. $S_L(n) = r/V_L$. Therefore, using our ansatz for r we get

$$S_L(n) = \frac{G(n/L)}{V_L}.$$

Within a continuous approximation the probability mass function of n, $D_L(n) = \text{Prob}[\text{frequency} = n]$, can be obtained from the derivative of $S_L(n)$:

$$D_L(n) = -\frac{\partial S_L(n)}{\partial n} = \frac{g(n/L)}{LV_L}, \tag{2}$$

where g is minus the derivative of G, i.e. $g(x) = -G'(x)$. If one does not trust the continuous approximation, one can write $D_L(n) = S_L(n) - S_L(n+1)$ and perform a Taylor expansion, for which the result is the same, but with $g(x) \simeq -G'(x)$. In this way, we obtain simple forms for $S_L(n)$ and $D_L(n)$, which are analogous to standard scaling laws, except for the fact that we have not specified how V_L changes with L. If Heaps' law holds, $V_L \propto L^{\alpha}$, we recover a standard scaling law, $D_L(n) = g(n/L)/L^{1+\alpha}$, which fulfills invariance under a scaling transformation, or, equivalently, fulfills the definition of a generalized homogeneous function [24, 37]

$$D_{\lambda_L L}(\lambda_n n) = \lambda_D D_L(n),$$

where λ_L, λ_n and λ_D are the scale factors, related in this case through

$$\lambda_n = \lambda_L \equiv \lambda$$

and

$$\lambda_D = \frac{1}{\lambda^{1+\alpha}}.$$

Table 1. Total text length and vocabulary before (L_{tot}, V_{tot}) and after ($L^{(\text{l})}_{\text{tot}}$, $V^{(\text{l})}_{\text{tot}}$) the lemmatization process, for all the books considered (including also their author, language and publication year). The text length for lemmas is shorter than for words because for a number of word tokens their corresponding lemma type could not be determined, and they were ignored.

Title	Author	Language	Year	L_{tot}	V_{tot}	$L^{(\text{l})}_{\text{tot}}$	$V^{(\text{l})}_{\text{tot}}$
Artamène	Scudéry siblings	French	1649	2 078 437	25 161	1 737 556	5008
Clarissa	Samuel Richardson	English	1748	971 294	20 490	940 967	9041
Don Quijote	Miguel de Cervantes	Spanish	1605–1615	390 436	21 180	378 664	7432
La Regenta	L Alas 'Clarín'	Spanish	1884	316 358	21 870	309 861	9900
Le Vicomte de Bragelonne	A Dumas (father)	French	1847	693 947	25 775	676 252	10 744
Moby-Dick	Herman Melville	English	1851	215 522	18 516	204 094	9141
Ulysses	James Joyce	English	1918	268 144	29 448	242 367	12 469

However, in general (if Heaps' law does not hold), the distribution $D_L(n)$ still is invariant under a scale transformation but with a different relation for λ_D, which is

$$\lambda_D = \frac{V_L}{\lambda V_{\lambda L}}.$$

So, $D_L(n)$ is not a generalized homogeneous function, but presents an even more general form. In any case, the validity of the proposed scaling law, equation (1), can be checked by performing a very simple rescaled plot, displaying $LV_L D_L(n)$ versus n/L. A resulting data collapse support the independence of the scaling function with respect to L. This is undertaken in section 3.

3. Data analysis results

To test the validity of our predictions, summarized in equation (2), we analyze a corpus of literary texts, comprised by seven large books in English, Spanish and French (among them, some of the longest novels ever written, in order to have as much statistics of homogeneous texts as possible). In addition to the statistics of the words in the texts, we consider the statistics of lemmas (roughly speaking, the stem forms of the word; for instance, *dog* for *dogs*). In the lemmatized version of each text, each word is substituted by its corresponding lemma, and the statistics are collected in the same way as they are collected for word forms. Appendix A provides detailed information on the lemmatization procedure, and table 1 summarizes the most relevant characteristics of the analyzed books.

First, we plot the distributions of word frequencies, $D_L(n)$ versus n, for each book, considering either the whole book or the first L/L_{tot} fraction, where L_{tot} is the real, complete text length (i.e. if $L = L_{\text{tot}}/2$ we consider just the first half of the book, no average is performed over parts of size L). For a fixed book, we observe that different L leads to distributions with small but clear differences, see figure 1. The pattern described by Bernhardsson *et al* (equivalent to Zipf's findings for the change of the exponent β) seems to hold, as the absolute value of the slope in log–log scale (i.e. the apparent power-law exponent γ) decreases with increasing text length.

However, a scaling analysis reveals an alternative picture. As suggested by equation (2), plotting $LV_L D_L(n)$ against n/L for different values of L yields a collapse of all the curves onto

IOP Institute of Physics DEUTSCHE PHYSIKALISCHE GESELLSCHAFT

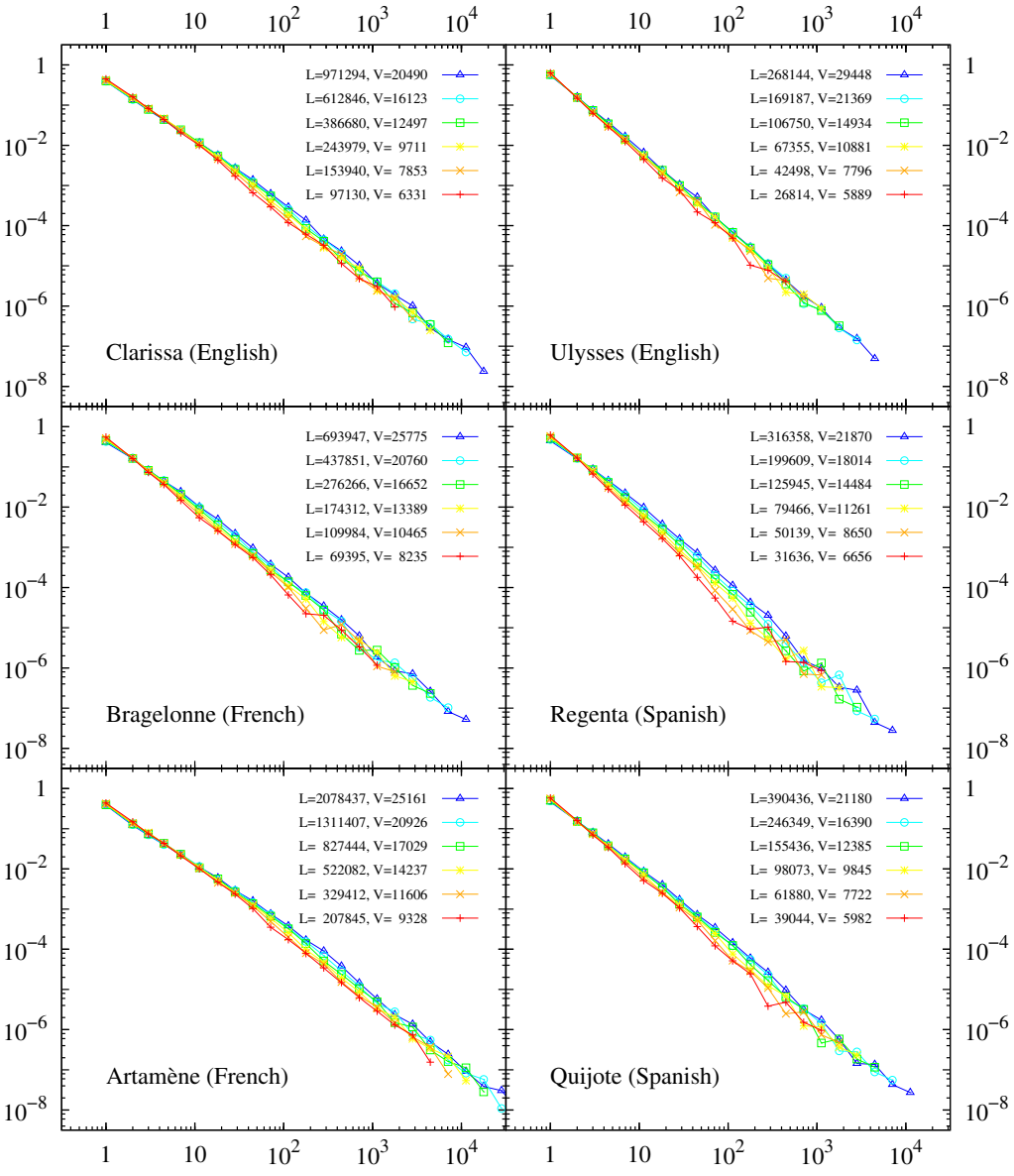

Figure 1. Density of word frequencies $D_L(n)$ (y-axis) against absolute frequency n (x-axis), for six different books, taking text length $L = L_{\text{tot}}/10$, $L_{\text{tot}}/10^{4/5}$, $L_{\text{tot}}/10^{3/5}$, ..., L_{tot}. The slope seems to decrease with text length.

a unique L-independent function for each book, which represents the scaling function $g(x)$. Figure 2 shows this for the same books and parts of the books as in figure 1. The data collapse can be considered excellent, except for the smallest frequencies. For the largest L the collapse is valid up to $n \simeq 3$ if we exclude *La Regenta*, which only collapses for about $n \geqslant 6$. So, our scaling hypothesis is validated, independently of the particular shape that $g(x)$ takes. Note that $g(x)$ is independent of L but not the book, i.e. each book has its own $g(x)$, different from the rest. In any case, we observe a slightly convex shape in log–log space, which leads to the rejection of the power-law hypothesis for the whole range of frequencies. Nevertheless, the data does not show any clear parametric functional form. A double power law, a stretched exponential, a Weibull

IOP Institute of Physics DEUTSCHE PHYSIKALISCHE GESELLSCHAFT

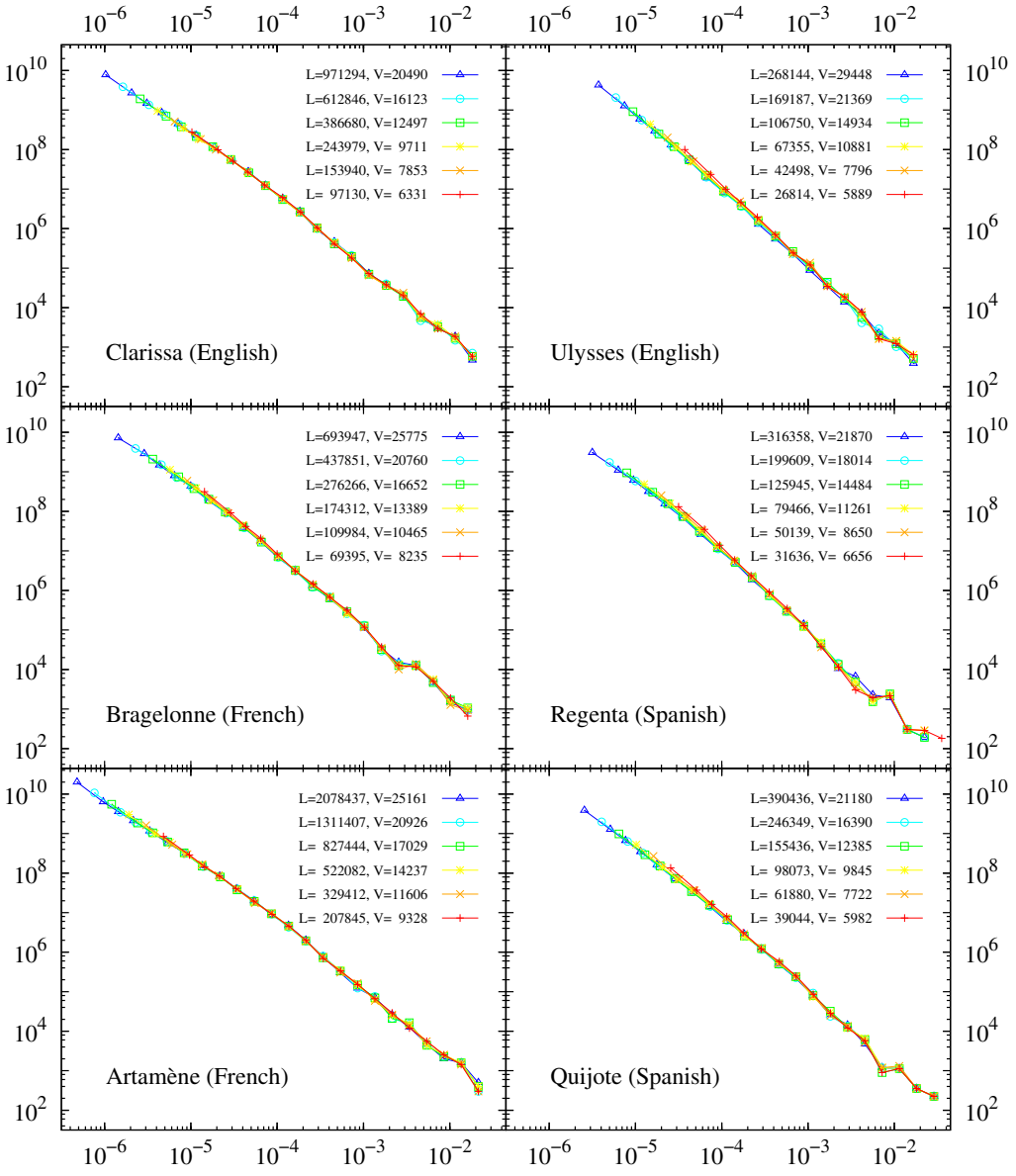

Figure 2. Rescaled densities $LV_L D_L(n)$ (y-axis) against relative frequency n/L (x-axis), for the same books and fractions of text as in figure 1. The rescaled densities collapse onto a single function, independently of the value of L, validating our proposed scaling form for $D_L(n)$ (equation (2)) and making it clear that the decrease of the log–log slope with L is not a consequence of a genuine change in the scaling properties of the distribution.

or a lognormal tail could be fit to the distributions. This is not incompatible with the fact that the large n tail can be well fit by a power law (the Zipf's law), for more than two orders of magnitude [36].

Things turn out to be somewhat different after the lemmatization process. The scaling ansatz is still clearly valid for the frequency distributions, see figure 3, but with a different kind of scaling function $g(x)$, with a more defined characteristic shape, due to a more pronounced

IOP Institute of Physics DEUTSCHE PHYSIKALISCHE GESELLSCHAFT

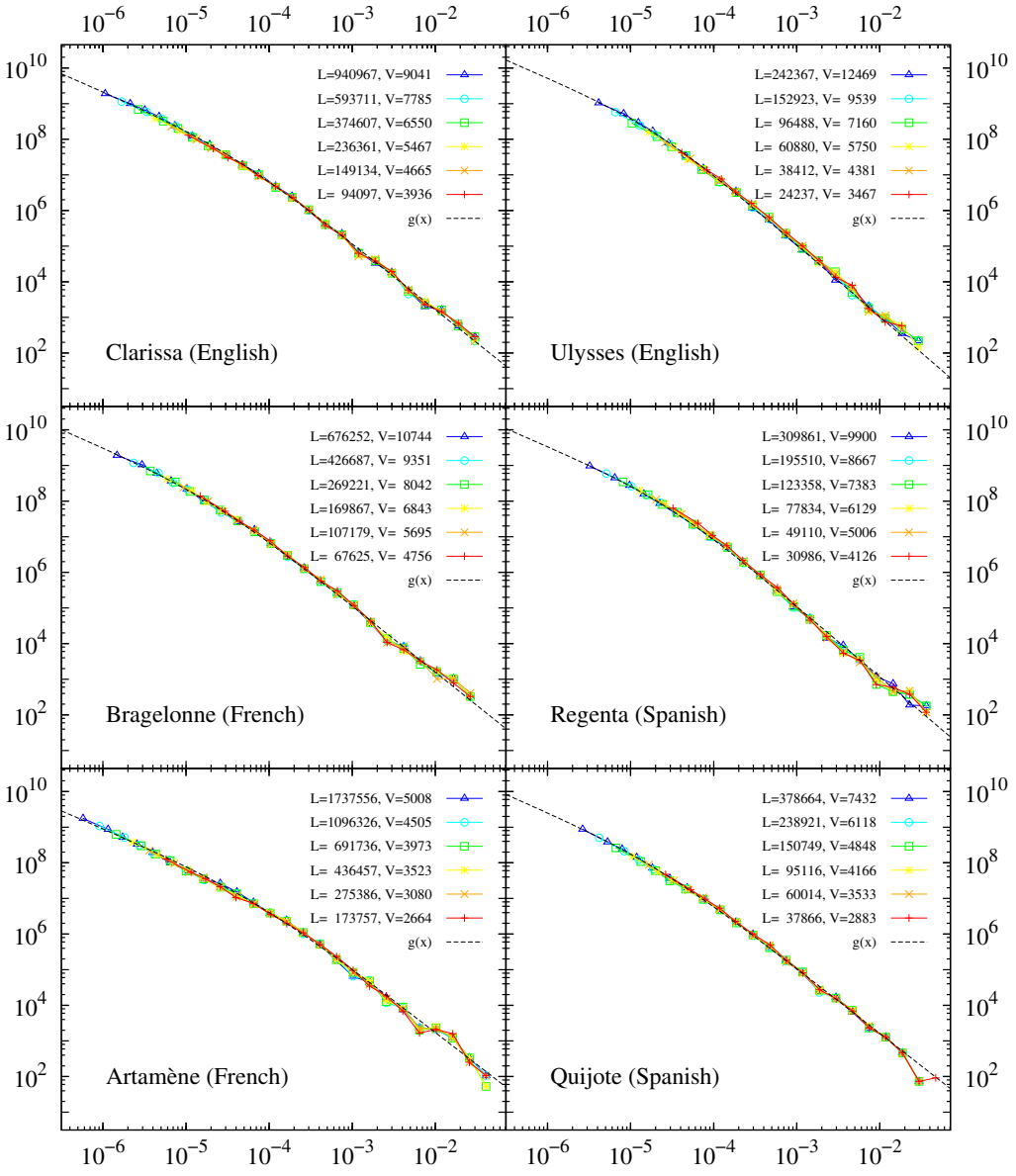

Figure 3. Same rescaled distributions as in previous figure ($LV_L D_L(n)$ versus n/L), but for the frequencies of lemmas. The data collapse guarantees the fulfillment of the scaling law also in this case. The fit resulting from the double power-law distribution, equation (3), is also included.

log–log curvature or convexity. In fact, close examination of the data leads us to conclude that the lemmatization process enhances the goodness of the scaling approximation, specially in the low-frequency zone. It could be reasoned that, as lemmatized texts have a significantly reduced vocabulary compared to the original ones, but the total length remains essentially the same, they are somehow equivalent to much longer texts, if one considers the length-to-vocabulary ratio. Although this matter needs to be further investigated, it supports the idea that our main hypothesis, the scale-invariance of the distribution of frequencies, holds more strongly for longer texts.

Table 2. Values of the parameters n_a, γ and a for the lemmatized versions (indicated with the superscript (l)) of the seven complete books. The fits are performed numerically through MLE, while the standard deviations come from Monte Carlo simulations, see appendix B.

Title	$n_a \pm \sigma_{n_a}$	$\gamma \pm \sigma_\gamma$	$a \pm \sigma_a$
Artamène$^{(l)}$	129.7 ± 12.6	1.807 ± 0.026	$(4.65 \pm 0.91) \times 10^{-4}$
Clarissa$^{(l)}$	32.70 ± 2.17	1.864 ± 0.021	$(1.40 \pm 0.24) \times 10^{-4}$
Don Quijote$^{(l)}$	7.91 ± 0.75	1.827 ± 0.020	$(1.35 \pm 0.22) \times 10^{-4}$
La Regenta$^{(l)}$	9.45 ± 0.66	1.983 ± 0.021	$(3.68 \pm 0.62) \times 10^{-5}$
Bragelonne$^{(l)}$	14.56 ± 1.23	1.866 ± 0.018	$(9.10 \pm 1.37) \times 10^{-5}$
Moby-Dick$^{(l)}$	8.21 ± 0.53	2.050 ± 0.024	$(2.42 \pm 0.47) \times 10^{-5}$
Ulysses$^{(l)}$	5.38 ± 0.31	2.020 ± 0.017	$(1.79 \pm 0.28) \times 10^{-5}$

Due to the clear curvature of $g(x)$ in the lemmatized case, we go one step further and propose a concrete function to fit these data, namely

$$g(x) = \frac{k}{x(a + x^{\gamma-1})}. \tag{3}$$

This function has two free parameters, a and γ (with $\gamma > 1$ and $a > 0$), and behaves as a double power law, that is, for large x, $g(x) \sim x^{-\gamma}$ (we still have Zipf's law), while for small x, $g(x) \sim x^{-1}$. The transition point between both power-law tails is determined by a (more precisely, by $a^{\frac{1}{\gamma-1}}$), and k is fixed by normalization. But an important issue is that it is not $g(x)$ which is normalized to one but $D_L(n)$. We select a power-law with exponent one for small x for three reasons: firstly, in order to explore an alternative to the power law in the V_L versus L relation (which is not clearly supported by the data, see next section); secondly, to allow for a better comparison of our results and those of [32]; thirdly, to keep the number of parameters minimum. Thus, we do not look for the most accurate fit but for the simplest description of the data.

Then, defining $n_a = a^{\frac{1}{\gamma-1}} L$, the corresponding word-frequency density (or, more properly, lemma-frequency density or type-frequency density) turns out to be

$$D_L(n) \propto \frac{1}{n\left(1 + (n/n_a)^{\gamma-1}\right)} \tag{4}$$

with n_a the scale parameter (recall that the scale parameter of $g(x)$ was $a^{\frac{1}{\gamma-1}}$).

The data collapse in figure 3 and the good fit imply that the Zipf-like exponent γ does not depend on L, but the transition point between both power laws, n_a, obviously does. Hence, as L grows the transition to the $\sim n^{-\gamma}$ regime occurs at higher absolute frequencies, given by n_a, but fixed relative frequencies, given by $a^{\frac{1}{\gamma-1}}$. In table 2 we report the fitted parameters for all seven books, obtained by maximum likelihood estimation (MLE) of the frequencies of the whole books, as well as Monte Carlo estimates of their uncertainties. We have confirmed the stability of γ fitting only a power-law tail from a fixed common relative frequency, for different values of L [36].

Regarding the low-frequency exponent, one could find a better fit if the exponent was not fixed to be one; however, our data does not allow this value to be well constrained. A more important point is the influence of lemmatization errors in the characteristics of the low-frequency regime. Although the tools we use are rather accurate, rare words are likely to be assigned a wrong lemma. This limitation is intrinsic to current computational tools and has to be considered as a part of the lemmatization process. Nevertheless, the fact that the behavior at low frequencies is robust in front of a large variation in the percentage of lemmatization errors implies that our result is a genuine consequence of the lemmatization. See appendix A for more details.

Although double power laws have been previously fit to rank-frequency plots for unlemmatized multi-author corpora [27, 38, 39], the resulting exponents for large ranks (low frequencies) are different than the ones obtained for our lemmatized single-author texts. Note that [27] also proposed that the crossover between both power laws happened for a constant number of types, around 7900, independently of corpus size. This corresponds indeed to $r = 7900$ and therefore, from equation (1), to a fixed relative frequency. This is certainly in agreement with our results, supporting the hypothesis that rank-frequency plots and frequency distributions are stable in terms of relative frequency.

4. An asymptotic approximation of Heaps' law

Coming back to our scaling ansatz, equation (2), the normalization of $D_L(n)$ will allow us to establish a relationship between the word-frequency distribution and the growth of the vocabulary with text length. In the continuous approximation

$$1 = \int_1^\infty D_L(n)\,\mathrm{d}n = \frac{1}{V_L}\int_1^\infty g(n/L)\frac{\mathrm{d}n}{L} = \frac{1}{V_L}\int_{1/L}^\infty g(x)\,\mathrm{d}x = \frac{1}{V_L}G\left(\frac{1}{L}\right),$$

where we have used the previous relation $g(x) = -G'(x)$, and have additionally imposed $G(\infty) \equiv 0$, for which it is necessary that $g(x)$ decays faster than a power law with exponent one. So,

$$V_L = G\left(\frac{1}{L}\right). \tag{5}$$

This just means, compared to equation (1), that the number of types with relative frequency greater or equal than $1/L$ is the vocabulary size V_L, as this is the largest rank for a text of length L. It is important to notice the difference between saying that $G_L(1/L) = V_L$, which is a trivial statement, and stating that $G(1/L) = V_L$, which provides a link between Zipf's and Heaps' law, or, more generally, between the distribution of frequencies and the vocabulary growth, by approximating the latter by the former. The quality of such an approximation will depend, of course, on the goodness of the scale-invariance approximation. In the usual case of a power-law distribution of frequencies extending to the lowest values, $g(x) \propto 1/x^\gamma$, with $\gamma > 1$, then $G(x) \propto 1/x^{\gamma-1}$, which turns into Heaps' law, $V_L \propto L^\alpha$, with $\alpha = \gamma - 1$, in agreement with previous research [2, 5, 29, 30, 32].

However, this power-law growth of V_L with L is not what is observed in texts, in general. Due to the accurate fit that we can achieve for lemmatized texts, we can explicitly derive an

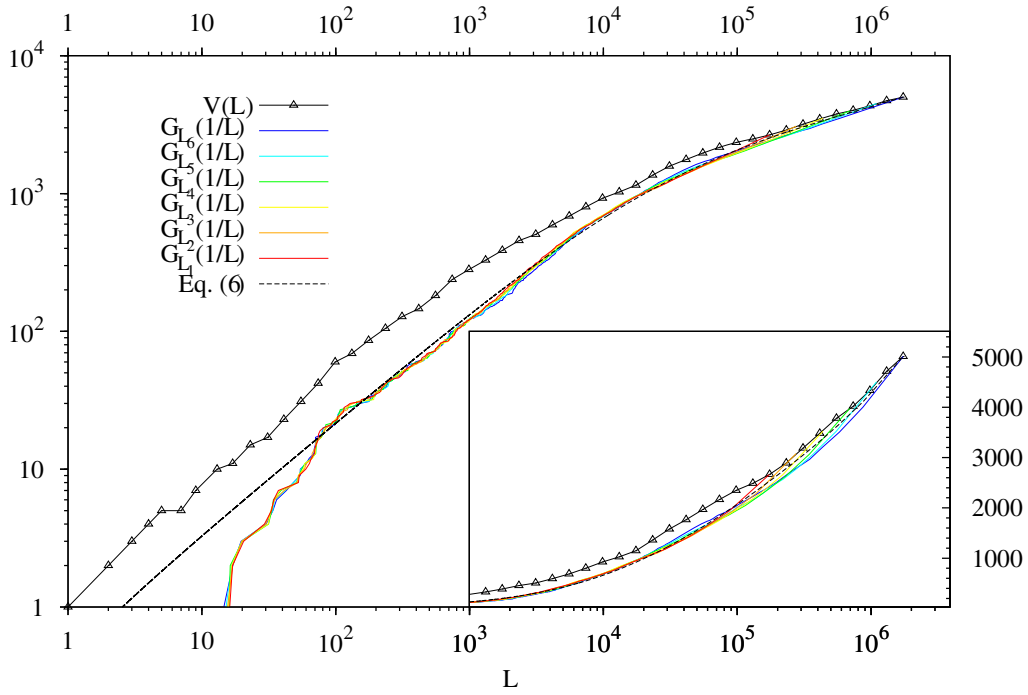

Figure 4. The actual curve V_L (solid black with triangles) for the lemmatized version of the book *Artamène*, together with the curves $V_L = G(1/L)$ obtained by using the empirical inverse of the rank-frequency plot, $r = G(n/L)$, with $L_i = L_{\text{tot}}/10^{(6-i)/5}$ (colors), and the analytical expression (7) with parameters determined from the fit of $D_{L_{\text{tot}}}(n)$, equation (6) (dashed black).

asymptotic expression for V_L given our proposal for $g(x)$. As we have just shown, $g(x)$ is not normalized to one, rather, $\int_{1/L}^{\infty} g(x)\,\mathrm{d}x = V_L$. Hence, substituting $g(x)$ from equation (2) and integrating

$$V_L = \int_{1/L}^{\infty} \frac{k}{x(a+x^{\gamma-1})}\mathrm{d}x = \frac{k}{a}\int_{1/L}^{\infty} \frac{ax^{-\gamma}}{ax^{1-\gamma}+1}\mathrm{d}x =$$
$$= \frac{k}{a(1-\gamma)}\ln(ax^{1-\gamma}+1)\Big|_{1/L}^{\infty} = \frac{k}{a(\gamma-1)}\ln(aL^{\gamma-1}+1). \qquad (6)$$

In this case V_L is not a power law, and behaves asymptotically as $\propto \ln L$. This is a direct consequence of our choice for the exponent 1 in the left-tail of $g(x)$. Indeed, it seems clear that the vocabulary growth curve greatly deviates from a straight line in log–log space, for it displays a prominent convexity, see figure 4 as an example. Nevertheless, the result from equation (6) is not a good fit either, due to a wrong proportionality constant. This is caused by the continuous approximation in equation (6).

For an accurate calculation of V_L we must treat our variables as discrete and compute discrete sums rather than integrals. In the exact, discrete treatment of $D_L(n)$, equation (6) must

be rewritten as

$$V_L = G\left(\frac{1}{L}\right) = G\left(\frac{L_{\rm tot}/L}{L_{\rm tot}}\right) = \sum_{n \geqslant L_{\rm tot}/L} \frac{g(n/L_{\rm tot})}{L_{\rm tot}}$$
$$= \frac{1}{L_{\rm tot}} \sum_{n \geqslant L_{\rm tot}/L} \frac{k}{\left(\frac{n}{L_{\rm tot}}\right)\left(a + \left(\frac{n}{L_{\rm tot}}\right)^{\gamma-1}\right)}, \quad (7)$$

where we have used the fact that $S_{L_{\rm tot}}(n') = \sum_{n \geqslant n'} D_{L_{\rm tot}}(n)$, with $n' = L_{\rm tot}/L$ (notice that in the discrete case, $g(x) \neq -G'(x)$). This is consistent with the fact that, indeed, the maximum likelihood parameters γ and a have been computed assuming a discrete probability function (see appendix B), and so has the normalization constant. We would like to stress that no fit is performed in figure 4, that is, the constant k in $g(x)$ is directly derived from the normalizing constant of $D_L(n)$, and depends only on γ and a.

5. Conclusions

In summary, we have shown that, contrary to claims in previous research [25, 31, 32], Zipf's law in linguistics is extraordinarily stable under changes in the size of the analyzed text. A scaling function $g(x)$ provides a constant shape for the distribution of frequencies of each text, $D_L(n)$, no matter its length L, which only enters into the distribution as a scale parameter and determines the size of the vocabulary V_L. The apparent size-dependent exponent found previously seems to be an artifact of the slight convexity of $g(x)$ in a log–log plot, which is more clearly observed for very small values of x, accessible only for the largest text lengths. Moreover, we find that in the case of lemmatized texts the distribution can be well described by a double power law, with a large-frequency exponent γ that does not depend on L, and a transition point n_a that scales linearly with L. The small-frequency exponent is different than the ones reported in [27, 38] for non-lemmatized corpora. Further, the stability of the shape of the frequency distribution allows one to predict the growth of vocabulary size with text length, resulting in a generalization of the popular Heaps' law.

The robustness of Zipf-like parameters under changes in system size opens the way to more practical applications of word statistics. In particular, we provide a consistent way to compare statistical properties of texts with different lengths [40]. Another interesting issue would be the application of the same scaling methods to other fields in which Zipf's law has been proposed to hold, as economics and demography, for instance.

Acknowledgments

We appreciate a collaboration with R Ferrer-i-Cancho, who also put AC in contact with GB. Financial support is acknowledged from grants FIS2009-09508 from the Ministerio de Ciencia y Tecnología, FIS2012-31324 from the Ministerio de Economía y Competitividad and 2009-SGR-164 from Generalitat de Catalunya, which also supported FF-C through grant 2012FI_B 00422 and GB through AGAUR grant 2010BP-A00070.

Table A.1. Coverage of the vocabulary by the dictionary in each language, both at the type and at the token level. Remember that we distinguish between a word *type* (corresponding to its orthographic form) and its *tokens* (actual occurrences in text).

Title	Types (%)	Tokens (%)
Clarissa	68.0	96.9
Moby-Dick	70.8	94.7
Ulysses	58.6	90.4
Don Quijote	81.3	97.0
La Regenta	89.5	97.9
Artamène	43.6	83.6
Bragelonne	89.8	97.5
Seitsemän v.	89.8	95.4
Kevät ja t.	96.2	98.3
Vanhempieni r.	96.5	98.5
Average	78.4	95.0

Appendix A. Lemmatization

To analyze the distribution of frequencies of lemmas, the texts needed to be lemmatized. To manually lemmatize the words would have exceeded the possibilities of this project, so we proceeded to automatic processing with standard computational tools: *FreeLing*[5] for Spanish and English and *TreeTagger* [41] for French. The tools carry out the following steps:

1. *Tokenization*. Segmentation of the texts into sentences and sentences into words (tokens).
2. *Morphological analysis*. Assignment of one or more lemmas and morphological information (tag) to each token. For instance, *found* in English can correspond to the past tense of the verb *find* or to the base form of the verb *found*. At this stage, both are assigned whenever the word form *found* is encountered.
3. *Morphological disambiguation*. An automatic tagger assigns the single most probable lemma and tag to each word form, depending on the context. For instance, in *I found the keys* the tagger would assign the lemma *find* to the word *found*, while in *He promised to found a hospital*, the lemma *found* would be preferred.

All these steps are automatic, such that errors are introduced at each step. However, the accuracy of the tools is quite high (e.g. around 95–97% at the token level for morphological disambiguation), so a quantitative analysis based on the results of the automatic process can be carried out. Also note that step 2 is based on a pre-existing dictionary (of words, not of lemmas, also called a lexicon): only the words that are in the dictionary are assigned a reliable set of morphological tags and lemmas. Although most of the tools used heuristically assign tag and/or lemma information to words that are not in the dictionary, we only count tokens of lemmas for which the corresponding word types are found in the dictionary, so as to minimize the amount of error introduced by the automatic processing. This comes at the expense of losing some data. However, the dictionaries have quite a good coverage of the vocabulary, particularly at the token level, but also at the type level (see table A.1). The exceptions are *Ulysses*, because

[5] FreeLing (http://nlp.lsi.upc.edu/freeling).

of the stream of consciousness prose, which uses many non-standard word forms, and *Artamène*, because 17th century French contains many word forms that a dictionary of modern French does not include.

Appendix B. Maximum likelihood fitting

The fitted values of table 2 have been obtained by MLE. This well-known procedure consists firstly in computing the log-likelihood function $\mathcal{L}$, which in our case reads

$$\mathcal{L} = \frac{1}{V_L}\sum_{i=1}^{V_L} \ln\, D_L(n_i) = \ln\, K - \frac{1}{V_L}\sum_{i=1}^{V_L} \ln\left(n_i(b+n_i^{\gamma-1})\right)$$

with n_i the V_L values of the frequency and the normalization constant K in the discrete case equal to

$$K = \left[\sum_{n=1}^{n_{\max}} \frac{1}{n(b+n^{\gamma-1})}\right]^{-1}.$$

Note that we have reparameterized the distribution compared to the main text, introducing $b = n_a^{\gamma-1} = aL^{\gamma-1}$. Then, $\mathcal{L}$ is maximized with respect to the parameters γ and b; this has been done numerically using the simplex method [42]. The error terms σ_γ and σ_b, representing the standard deviation of each estimator, are computed from Monte Carlo simulations. From the resulting maximum-likelihood parameters γ^* and b^*, synthetic data samples are simulated, and the MLE parameters of these samples are calculated in the same way; their fluctuations yield σ_γ and σ_b. We stress that no continuous approximation has been made, that is, the simulated data follows the discrete probability function $D_L(n)$ (this is done using the rejection method, see [36, 43] for details for a similar case). In a summarized recipe, the procedure simply is:

1. numerically compute the MLE parameters, γ^* and b^*;
2. draw M datasets, each of size V_L, from the discrete probability function $D_L(n; \gamma^*, b^*)$;
3. for each dataset $m = 1, \ldots, M$, compute the MLE parameters γ^m, b^m;
4. compute the standard deviations σ_γ and σ_b of the sets $\{\gamma^m\}_{m=1}^M$ and $\{b^m\}_{m=1}^M$;

The standard deviations of n_a and a are computed in the same way using their relationship to b and γ.

References

[1] Zipf G K 1949 *Human Behavior and the Principle of Least Effort* (Reading, MA: Addison-Wesley)
[2] Mandelbrot B B 1961 *Structures of Language and its Mathematical Aspects* ed R Jacobsen (New York: American Mathematical Society) pp 214–7
[3] Ferrer i Cancho R and Hernández-Fernández A 2008 Power laws and the golden number *Problems of General, Germanic and Slavic Linguistics* ed G Altmann, I Zadorozhna and Y Matskulyak (Chernivtsi: Books–XXI) pp 518–23
[4] Adamic L A and Huberman B A 2002 *Glottometrics* **3** 143
[5] Kornai A 2002 *Glottometrics* **4** 61
[6] Zanette D 2012 *Statistical Patterns in Written Language*
[7] Zanette D and Montemurro M 2005 *J. Quantum Linguist.* **12** 29

[8] Ferrer i Cancho R 2005 *Eur. Phys. J.* B **44** 249
[9] Clauset A, Shalizi C R and Newman M E J 2009 *SIAM Rev.* **51** 661
[10] Corral A, Font F and Camacho J 2011 *Phys. Rev.* E **83** 066103
[11] Miller G A 1957 *Am. J. Psychol.* **70** 311
[12] Li W 1992 *IEEE Trans. Inform. Theory* **38** 1842
[13] Simon H A 1955 *Biometrica* **42** 425
[14] Newman M E J 2005 *Contemp Phys.* **46** 323
[15] Saichev A, Malevergne Y and Sornette D 2010 *Theory of Zipf's Law and Beyond* (*Lecture Notes in Economics and Mathematical Systems* vol 632) (Berlin: Springer)
[16] Ferrer i Cancho R and Solé R V 2003 *Proc. Natl Acad. Sci. USA* **100** 788
[17] Corominas-Murtra B, Fortuny J and Solé R V 2011 *Phys. Rev.* E **83** 036115
[18] Ferrer i Cancho R 2005 *Eur. Phys. J.* B **47** 449
[19] Düring B, Matthes D and Toscani G 2009 *Riv. Mat. Univ. Parma* **1** 199
[20] Bak P 1996 *How Nature Works: The Science of Self-Organized Criticality* (New York: Copernicus)
[21] Mitzenmacher M 2004 *Internet Math.* **1** 226
[22] Ferrer i Cancho R and Elvevåg B 2010 *PLoS ONE* **5** e9411
[23] Dickman R, Moloney N R and Altmann E G 2012 *J. Stat. Mech.* **2012** P12022
[24] Christensen K and Moloney N R 2005 *Complexity and Criticality* (London: Imperial College Press)
[25] Baayen H 2001 *Word Frequency Distributions* (Dordrecht: Kluwer)
[26] Herdan G 1964 *Quantitative Linguistics* (London: Butterworths)
[27] Gerlach M and Altmann E G 2013 *Phys. Rev.* X **3** 021006
[28] van Leijenhorst D and van der Weide T 2005 *Inform. Sci.* **170** 263
[29] Serrano M A, Flammini A and Menczer F 2009 *PLoS ONE* **4** e5372
[30] Lü L, Zhang Z-K and Zhou T 2010 *PLoS ONE* **5** e14139
[31] Powers D M W 1998 *NeMLaP3/CoNLL '98*: *Proc. of the Joint Conf. on New Methods in Language Processing and Computational Natural Language Learning* (Stroudsburg, PA: Association for Computational Linguistics) pp 151–60
[32] Bernhardsson S, da Rocha L E C and Minnhagen P 2009 *New J. Phys.* **11** 123015
[33] Stauffer D and Aharony A 1994 *Introduction To Percolation Theory* 2nd edn (Boca Raton, FL: CRC)
[34] Zapperi S, Lauritsen K B and Stanley H E 1995 *Phys. Rev. Lett.* **75** 4071
[35] Corral A and Font-Clos F 2013 *Self-Organized Critical Phenomena* ed M Aschwanden (Berlin: Open Academic Press) pp 183–228
[36] Corral A, Boleda G and Ferrer-i-Cancho R 2013 in preparation
[37] Hankey A and Stanley H E 1972 *Phys. Rev.* B **6** 3515
[38] Ferrer i Cancho R and Solé R V 2001 *J. Quantum Linguist.* **8** 165
[39] Petersen A M, Tenenbaum J N, Havlin S, Stanley H E and Perc M 2012 *Sci. Rep.* **2** 943
[40] Baixeries J, Elvevåg B and Ferrer-i R 2013 Cancho *PLoS ONE* **8** e53227
[41] Schmid H 1994 *Proc. Int. Conf. on New Methods in Language Processing* vol 12 (Manchester: Citeseer) pp 44–9
[42] Press W H, Teukolsky S A, Vetterling W T and Flannery B P 1992 *Numerical Recipes in FORTRAN* 2nd edn (Cambridge: Cambridge University Press)
[43] Devroye L 1986 *Non-Uniform Random Variate Generation* (New York: Springer)

3.2 Log-log Convexity of Type-Token Growth in Zipf's Systems

Francesc Font-Clos, Gunnar Pruessner, Nicholas R. Moloney and Anna Deluca
Physical Review Letters (In press, 2015)

Log-log Convexity of Type-Token Growth in Zipf's Systems

[Accepted for publication in Physical Review Letters: 05 May 2015]

Francesc Font-Clos[1, 2] and Álvaro Corral[1, 2]

[1] *Centre de Recerca Matemàtica, Edifici C, Campus Bellaterra, E-08193 Barcelona, Spain.*
[2] *Departament de Matemàtiques, Facultat de Ciències, Universitat Autònoma de Barcelona, E-08193 Barcelona, Spain*

It is traditionally assumed that Zipf's law implies the power-law growth of the number of different elements with the total number of elements in a system - the so-called Heaps' law. We show that a careful definition of Zipf's law leads to the violation of Heaps' law in random systems, with growth curves that have a convex shape in log-log scale. These curves fulfil universal data collapses that only depend on the value of the Zipf's exponent. We observe that real books behave very much in the same way as random systems, despite the presence of burstiness in word occurrence. We advance an explanation for this unexpected correspondence.

PACS numbers:

A great number of systems in social science, economy, cognitive science, biology, and technology have been proposed to follow Zipf's law [1–6]. All of them have in common that they are composed by some "elementary" units, which we will call tokens, and that these tokens can be grouped into larger, concrete or abstract entities, called types. For instance, if the system is the population of a country, the tokens are its citizens, which can be grouped into different concrete types given by the cities where they live [7]. If the system is a text, each appearance of a word is a token, associated to the abstract type given by the word itself [8]. Zipf's law deals with how tokens are distributed into types, and can be formulated in two different ways, which are generally considered as equivalent [1, 3, 8, 9].

The first one is obtained when the number of tokens associated to each type are counted and the types are ranked in decreasing order of counts. We call this the rank-count representation. If a (decreasing) power law holds between the number of tokens of each type and the rank of the type, with an exponent close to one, this indicates the fulfilment of Zipf's law. An alternative version of the law arises when a second statistics is performed, considering the number of types that have the same number of counts; as the counts play the role of the random variable what one gets is the distribution of counts. If a (decreasing) power law is obtained, with an exponent around two, one gets a different formulation of Zipf's law, in principle.

However, in general, the fulfilment of Zipf's law has not been tested with rigorous statistical methods [2, 10]; rather, researchers have become satisfied with just qualitative resemblances between empirical data and power laws. In part, this can be justified by the difficulties of obtaining clear statistics from the rank-count representation, in particular for high ranks (that is, for rare types), and also by poor methods of estimation of probability distributions [2]. Despite the lack of unambiguous empirical support, from the theoretical point of view the search for explanations of Zipf's law has been extensive, but without a clearly accepted preferred mechanism [1, 11–14].

The presence of temporal order is an important feature in many Zipf-like systems, but this is not captured in Zipf's law. Indeed, this law only provides a static picture of the system (as the law is not altered under re-ordering of the data). In contrast, a suitable statistic that can unveil some of the dynamics is the type-token growth curve, which counts the total number of types, v, as a function of the total number of tokens, ℓ, as a system evolves, i.e., as citizens are born or a text is being read. Note that ℓ is a measure of system size (as system grows) and v is a measure of richness or diversity of types (with the symbol v borrowed from linguistics, where it stands for the size of the vocabulary).

It has long been assumed that Zipf's law implies also a power law for the type-token growth curve, i.e.,

$$v(\ell) \propto \ell^{\,\alpha}, \tag{1}$$

with exponent α smaller than one, and this is referred to as Heaps' law in general or Herdan's law in quantitative linguistics [15–17]. Indeed, Mandelbrot [18] and the authors of Ref. [19] obtain Heaps' law when tokens are drawn independently from a Zipf's system. Baeza-Yates and Navarro [16] argue that, if both Zipf's law and Heaps' law are fulfilled, their exponents are connected. A similar demonstration, using a different scaling of the variables, is found in Ref. [20], and with some finite-size corrections in Ref. [9]. Other authors have been able to derive Heaps' law from Zipf's law using a master equation [21] or pure scaling arguments [22]. Alternatives to Heaps' formula are listed in Ref. [23], but without a theoretical justification.

However, even simple visual inspection of the log-log plot of empirical type-token growth curves shows that

Heaps' law is not even a rough approximation of the reality. On the contrary, a clear convexity (as seen from above) is apparent in most of the plots (see, for instance, some of the figures in [9, 24, 25]). This has been attributed to the fact that the asymptotic regime is not reached or to the effects of the exhaustion of the number of different types [26]. Nevertheless, the effect persists in very large systems, composed by many millions of tokens, and where the finiteness of the number of available types is questionable [22].

In the few reported cases where there seems to be a true power-law relation between number of tokens and number of types, as in Ref. [20], this turns out to come from a related but distinct statistics. Instead of considering the type-token growth curve in a single, growing system ($v(\ell)$ for $\ell = 1 \dots L$), one can look for the total type-token relationship in a collection or ensemble of $\mathcal{N}$ systems (V_j versus L_j, for $j = 1 \dots \mathcal{N}$, with $V_j = v(L_j)$), see also Refs. [21, 27–29]. We are, in contrast, interested in the type-token relation of a single growing system.

The fact that Heaps' law is so clearly violated for the type-token growth, given that this law follows directly from Zipf's law, casts doubts on the very validity of the latter law. But one may notice that, although the two versions of Zipf's law mentioned above are usually considered as equivalent, they are only asymptotically equivalent for high values of the count of tokens (i.e., for low ranks) [15, 17, 18]. However, the type-token growth curve emerges mainly from the statistics of the rarest types (i.e., the types with $n = 1$, for each value of ℓ), as it is only when a type appears for the first time that it contributes to the growth curve [22], and these are precisely the types for which the usual description in terms of the rank-count representation becomes problematic. So, the election of which is the form of Zipf's law that one considers to hold true becomes crucial for the derivation of the type-token growth curve and the fulfilment of Heaps' law or not.

Although most previous research has focused in Zipf's law in the rank-count representation, *i.e.*, the first version mentioned above, we argue that it is the second version of the law, that of the distribution of counts, the one that becomes relevant to describe the real type-token growth curve, at least in the case of written texts. Indeed, let us notice that the previously mentioned derivations of Heaps' law were all based on the rank-count representation [9, 16, 18–22]; therefore, the violation of Heaps' law for real systems invalidates the (exact) fulfilment of Zipf's law for the rank-count representation.

In contrast, when the viewpoint of Zipf's law for the distribution of counts is adopted, we prove that Heaps' law cannot be sustained for random systems and we derive an alternative law, which leads to "universal-like" shapes of the rescaled type-token growth curves, with the only dependence on the value of the Zipf's exponent. Quite unexpectedly, our prediction for random uncorrelated systems holds very well also for real texts. We are able to explain this effect despite the significant clustering or burstiness of word occurrences [30, 31], due to the singular role that the first appearance of a type plays in the type-token growth curve, in contrast to subsequent appearances.

Let us consider a Zipf's system of total size L, and a particular type with overall number of counts n; this means that the complete system contains n tokens of that type (and then L is the sum of counts of all types, $L = \sum_i n_i$). In fact, Zipf's law tells us that there can be many types with the same counts n, and we denote this number as $N_L(n)$. Quantitatively, in terms of the distribution of counts, Zipf's law reads

$$N_L(n) \propto \frac{1}{n^\gamma}, \tag{2}$$

for $n = 1, 2, \dots$ with the exponent γ close to 2. Note that $N_L(n)$ is identical, except for normalisation, to the probability mass function of the number of counts. For a part of the system of size ℓ, with $\ell \leq L$, the number of types with k counts will be $N_\ell(k)$. The dependence of this quantity with the global $N_L(n)$ will be computed for a random system, which is understood as a sequence of tokens where these are taken at random from some underlying distribution. The $N_L(n)$ words with number of counts n in the whole system will lead, on average, to $N_L(n)h_{k,n}$ types with counts k in the subset, with $k \leq n$ and $h_{k,n}$ given by the hypergeometric distribution,

$$h_{k,n} = \frac{\binom{n}{k}\binom{L-n}{\ell-k}}{\binom{L}{\ell}}. \tag{3}$$

This is the probability to get k instances of a certain type when drawing, without replacement, ℓ tokens from a total population of L tokens of which there are n tokens of the desired type. The dependence of $h_{k,n}$ on ℓ and L is not explicit, to simplify the notation. The average number of types with k counts in the subset of size ℓ will result from the sum of $N_L(n)h_{k,n}$ for all $n \geq k$, i.e.,

$$N_\ell(k) = \sum_{n \geq k} N_L(n) h_{k,n}. \tag{4}$$

We will use this relationship between $N_\ell(k)$ and $N_L(n)$ to derive the type-token growth curve. For a subset of size ℓ we will have that, out of the total V types, $v(\ell)$ will be present whereas $N_\ell(0)$ will not have appeared (and so, their number of counts will be $k = 0$); therefore, $v(\ell) = V - N_\ell(0)$, and substituting Eq. (4) for $k = 0$ and using that $N_L(0) = 0$, then,

$$v(\ell) = V - \sum_{n \geq 1} N_L(n) h_{0,n}. \tag{5}$$

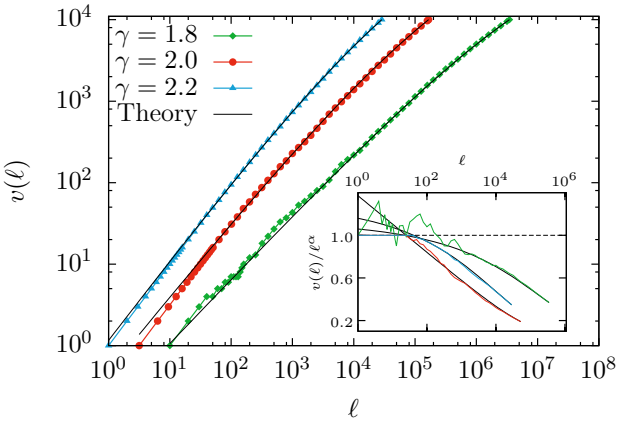

FIG. 1: **Main:** Type-token growth curve $v(\ell)$ for three random systems with number of counts drawn from a discrete power-law distribution $N_L(n) \propto n^{-\gamma}$, and $\gamma = 1.8$ (green diamonds), 2.0 (red circles) and 2.2 (blue triangles). The black lines correspond to our theoretical predictions, Eq. (9) for $\gamma \geq 2$ and Eq. (10) for $\gamma < 2$ (plotted with the help of the GSL libraries). No average over the reshuffling procedure is performed. Curves are consecutively shifted by a factor of $\sqrt{10}$ in the x-axis. **Inset:** The ratio $v(\ell)/\ell^{\alpha}$ is displayed, with $\alpha = \min\{1, \gamma - 1\}$, showing that an approximation of the form $v(\ell) \propto \ell^{\alpha}$ is too crude.

This formula relates the type-token growth curve with the distribution of counts in a random system, where it is exact, if we interpret $v(\ell)$ as an average over the random ensemble.

We now show that a power-law distribution of type counts does not lead to a power law in the type-token growth curve, in other words, Zipf's law for the distribution of counts does not lead to Heaps' law, in the case of a random system. Assuming that $n \ll L$, the "zero-success" probability $h_{0,n}$ can be approximated as follows (see SI for details),

$$h_{0,n} = \frac{\binom{L-n}{\ell}}{\binom{L}{\ell}} =\simeq \left(1 - \frac{\ell}{L}\right)^n , \tag{6}$$

which in practice holds for all types; in fact, the smallest number of counts, for which the approximation is better, give the largest contribution to Eq. (5), due to the power-law form of $N_L(n)$. This is given, taking into account a normalisation constant A, by

$$N_L(n) = V \frac{A}{n^{\gamma}}, \tag{7}$$

for $n = 1, 2, \ldots$ (and zero otherwise), with $\sum_{n \geq 1} N_L(n) = V$. Let us substitute the previous expressions for $h_{0,n}$ and $N_L(n)$ into Eq. (5), then

$$v(\ell) \simeq V \left(1 - A \sum_{n \geq 1} \frac{(1 - \ell/L)^n}{n^{\gamma}}\right). \tag{8}$$

Although there exists a maximum number of counts $n_{\max}$ beyond which $N_L(n) = 0$, as a first approximation the sum can be safely extended up to infinity, and hence we reach the following expression:

$$v(\ell) \simeq V \left(1 - \frac{\mathrm{Li}_{\gamma}(1 - \ell/L)}{\zeta(\gamma)}\right), \tag{9}$$

where we have made use of the polylogarithm function, $\mathrm{Li}_{\gamma}(z) = \sum_{n=1}^{\infty} z^n / n^{\gamma}$, defined for $|z| < 1$, and of the fact that the normalisation of Zipf's law is given by $A = 1/\zeta(\gamma)$, with $\zeta(\gamma)$ the Riemann zeta function, $\zeta(\gamma) = \mathrm{Li}_{\gamma}(1)$. Notice that, for random systems with fixed γ, Eq. (9) yields a "universal" scaling relationship between the number of types $v(\ell)$, if expressed in units of the total number of types V, and the text position ℓ expressed in units of the total size L.
In fact, Eq. (9) can lead to an overestimation of $v(\ell)$ due to finite-size effects, but this is rarely noticeable in practice. If one wants a more precise version of Eq. (9), then, going back to Eq. (8) and limiting the sum up to $n_{\max}$ gives, after some algebra (see SI for details),

$$v(\ell) = V \left(1 - \frac{\mathrm{Li}_{\gamma}(q) - q^{n_{\max}+1} \Phi(q, \gamma, n_{\max} + 1)}{\zeta(\gamma) - \Phi(1, \gamma, n_{\max} + 1)}\right), \tag{10}$$

with $q = 1 - \ell/L$, and $\Phi(z, \gamma, a) = \sum_{n=0}^{\infty} \frac{z^n}{(a+n)^{\gamma}}$, $|z| < 1; a \neq 0, -1, \ldots$ the Lerch transcendent. Obviously, Eq. (10) gives better results at the cost of using an additional parameter, $n_{\max}$. As a rule of thumb, it appears to be worth the cost in cases where $\gamma < 2$, $\ell \ll L$ and L is not too large. In most practical cases Eq. (9) gives an excellent approximation; nevertheless, we include its more refined version, Eq. (10), for the sake of completeness.
In order to test these predictions, we simulate a random Zipf's system as follows: Let us draw $V = 10^4$ random numbers $n_1, n_2, \ldots n_V$, from the discrete probability distribution $N_L(n)/V = n^{-\gamma}/\zeta(\gamma)$, with $\gamma = 1.8, 2.0$ and 2.2. Each of these V values of n represents a type, with a number of counts given by the value of n. For each type $i = 1, \ldots, V$, we create then n_i copies (tokens) of its associated type, and make a list with all of them,

$$\underbrace{1, \ldots, 1}_{n_1}, \underbrace{2, \ldots, 2}_{n_2}, \ldots, \underbrace{V, \ldots, V}_{n_V}. \tag{11}$$

Then, the list is reshuffled in order to create a random system, of size $L = n_1 + n_2 + \cdots + n_V$. Figure 1 shows the resulting type-token growth together with the approximation given either by Eq. (9), which only depends on γ, or by Eq. (10), which depends on γ and $n_{\max}$. The agreement is nearly perfect, except for very small ℓ.

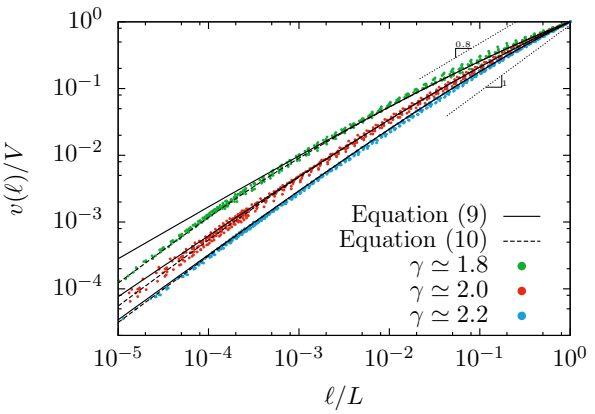

FIG. 2: The rescaled vocabulary-growth curve of 28 books from the PG database with exponents $\gamma = \{1.8, 2.0, 2.2\} \pm 0.01$ fitted for $n \geq 1$ or $n \geq 2$. The values of L and V range from $27,873$ to $146,845$ and from $5,639$ to $30,912$ respectively. As it is apparent, all books with the same exponent collapse into a single curve, which Eqs. (9) and (10) accurately capture. For the case of Eq. (10), we have used a value of $n_{\max}/L = 0.05$. The dotted straight lines (shifted for clarity) indicate the behaviour predicted by Heaps' law.

So far we have shown that Eqs. (9) and (10) capture very accurately the type-token growth curve for synthetic systems that have a perfect power-law distribution of counts but are completely random. Real systems, however, can have richer structures beyond the distribution of counts [30–32] and so one wonders if our derivations can provide acceptable predictions for them. In the following, we show that this is indeed the case when the system considered is that of natural language, and provide a qualitative explanation of this remarkable fact.

We analyse books from the *Project Gutenberg* (PG) database [33], selecting those whose distribution of frequencies $N_L(n)$ is statistically compatible with a pure, discrete power-law distribution. We fit the γ exponent with rigorous methods, see Refs. [34, 35]. In analogy with the previous section, we plot in Fig. 2 $v(\ell)/V$ versus ℓ/L for a total of 28 books for which $\gamma = 1.8, 2.0$, or 2.2. Books with the same Zipf's exponent collapse between them and into the corresponding theoretical curves, Eqs. (9) and (10). This is rather noticeable, as it points to the idea that the vocabulary-growth curve is unaffected by clustering, correlations, or by syntactic or discursive constraints. In other words, the vocabulary-growth curve of a real book fulfilling Zipf's law as given by Eq. (2) is not a power law but can be predicted using only its associated Zipf's exponent.

In order to understand why a prediction that heavily depends on the randomness hypothesis works so well for real books, we analyse the inter-occurrence-distance distribution of words. Given a word (type) with frequency n, we define its k-th inter-occurrence distance τ_k as the number of words (tokens) between its $k-1$-th and k-th appearances, plus one; i.e.,

$$\tau_k = \ell_k - \ell_{k-1} \tag{12}$$

(with ℓ_k the position of its k-th appearance and $k \leq n$). For the case of $k = 1$, we compute the number of words from the beginning of the text up to the first appearance, i.e., $\tau_1 = \ell_1$. If real books were completely random, then τ_k would be roughly exponentially distributed, and the rescaled distances

$$\hat{\tau}_k = \frac{\tau_k}{\langle \tau_k \rangle} \tag{13}$$

would be, for any value of n, exponentially distributed with parameter 1. Deviations from an exponential distribution for inter-occurrence distances in real books are well-known when all $k > 1$ are considered together, and constitute the so-called clustering or burstiness effect: instances of a given word tend to appear grouped together in the book, forming clusters and hence both very short and very long inter-occurrence distances are much more common than what an exponential distribution predicts [30, 31].

In contrast to previous works [30, 31], our analysis introduces the distinction between $k > 1$ and $k = 1$. Note that for what concerns the vocabulary-growth curve, all that matters is $k = 1$, as it is only the first appearance of each word that adds to the vocabulary. Figure 3 shows the (estimated) probability mass function $\mathcal{P}(\hat{\tau}_k)$ of the rescaled inter-occurrence distance for the book *Moby Dick* as an example (top), and for the one hundred longest books in the PG database (bottom). As it is apparent, for $k > 1$, the distributions of distances are not exponentially distributed, and we recover a trace of the clustering effect; however the case $k = 1$ displays a clearly different shape, much more close to an exponential distribution. This explains, at a qualitative level, why our derivations, based on a randomness assumption, continue to work in the case of real books that display clustering effects.

In conclusion, we have shown that Eqs. (9) and (10), which are not power laws but contain the polylogarithm function and the Lerch transcendent, provide a continuum of universality classes for type-token growth, depending only on the value of Zipf's exponent for the distribution of counts. We have verified our results both on synthetic random systems and on real books, showing that despite correlations or clustering effects, they remain valid as long as Zipf's law is fulfilled for the distribution of counts. Our results open the door to investigations in other contexts beyond linguistics, where the validity of Heaps' law could be questioned in a similar manner.

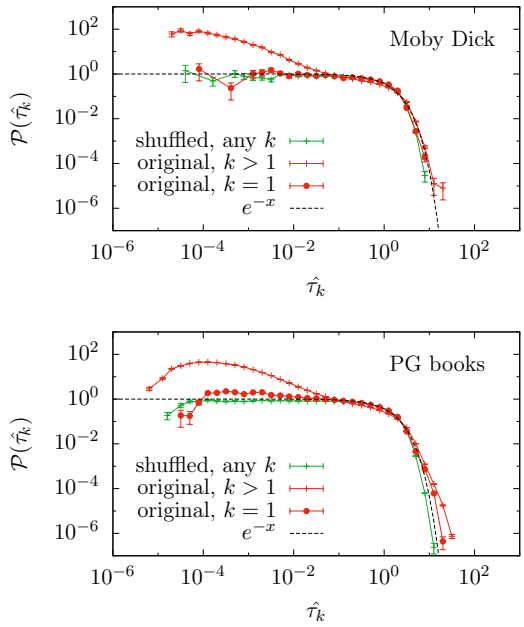

FIG. 3: Distribution of the rescaled inter-occurrence distances $\hat{\tau}_k$, see Eq. (13). The scale parameter $\langle \tau_k \rangle$ was computed from the data of each book (types with $n = 1$ or with $N(n) = 1$ were not included in the analysis). The original books (red) display clear deviations from an exponential distribution for $k > 1$, but not for $k = 1$. Shuffled versions of the books (green) do not show, as expected, any clustering effect, and hence their rescaled inter-occurrence distances are roughly exponentially distributed. **Top**: The book *Moby Dick*, by Herman Melville, as an illustrative example. **Bottom**: The one hundred longest books in the PG database.

Acknowledgements. The authors appreciate comments from E. Beltran-Saez and M. Gerlach. A. C. has enjoyed a long-term collaboration with G. Boleda and R. Ferrer-i-Cancho. Research projects in which this work is included are FIS2012-31324, from Spanish MINECO, and 2014SGR-1307 and 2012FI-B-00422, from AGAUR.

[1] M. E. J. Newman. Power laws, Pareto distributions and Zipf's law. *Cont. Phys.*, 46:323 –351, 2005.
[2] A. Clauset, C. R. Shalizi, and M. E. J. Newman. Power-law distributions in empirical data. *SIAM Rev.*, 51:661–703, 2009.
[3] L. A. Adamic and B. A. Huberman. Zipf's law and the Internet. *Glottometrics*, 3:143–150, 2002.
[4] C. Furusawa and K. Kaneko. Zipf's law in gene expression. *Phys. Rev. Lett.*, 90:088102, 2003.
[5] R. L. Axtell. Zipf distribution of U.S. firm sizes. *Science*, 293:1818–1820, 2001.
[6] J. Serrà, A. Corral, M. Boguñá, M. Haro, and J. Ll. Arcos. Measuring the evolution of contemporary western popular music. *Sci. Rep.*, 2:521, 2012.
[7] Y. Malevergne, V. Pisarenko, and D. Sornette. Testing the Pareto against the lognormal distributions with the uniformly most powerful unbiased test applied to the distribution of cities. *Phys. Rev. E*, 83:036111, 2011.
[8] D. Zanette. *Statistical Patterns in Written Language.* 2012.
[9] L. Lü, Z.-K. Zhang, and T. Zhou. Zipf's law leads to Heaps' law: Analyzing their relation in finite-size systems. *PLoS ONE*, 5(12):e14139, 12 2010.
[10] A. Corral, G. Boleda, and R. Ferrer-i-Cancho. Zipf's law for word frequencies: word forms versus lemmas in long texts. *PLoS ONE*, (submitted), 2015.
[11] M. Mitzenmacher. A brief history of generative models for power law and lognormal distributions. *Internet Math.*, 1 (2):226–251, 2004.
[12] R. Ferrer i Cancho and R. V. Solé. Least effort and the origins of scaling in human language. *Proc. Natl. Acad. Sci. U.S.A.*, 100:788–791, 2003.
[13] A. Saichev, Y. Malevergne, and D. Sornette. *Theory of Zipf's Law and of General Power Law Distributions with Gibrat's Law of Proportional Growth.* Lecture Notes in Economics and Mathematical Systems. Springer Verlag, Berlin, 2009.
[14] B. Corominas-Murtra, R. Hanel, and S. Thurner. Understanding scaling through history-dependent processes with collapsing sample space. *Proc. Natl. Acad. Sci. USA*, in press, 2015.
[15] H. S. Heaps. *Information retrieval: computational and theoretical aspects.* Academic Press, 1978.
[16] R. Baeza-Yates and G. Navarro. Block addressing indices for approximate text retrieval. *J. Am. Soc. Inform. Sci.*, 51(1):69–82, 2000.
[17] H. Baayen. *Word Frequency Distributions.* Kluwer, Dordrecht, 2001.
[18] B. Mandelbrot. On the theory of word frequencies and on related Markovian models of discourse. In R. Jakobson, editor, *Structure of Language and its Mathematical Aspects*, pages 190–219. American Mathematical Society, Providence, RI, 1961.
[19] D.C. van Leijenhorst and Th.P. van der Weide. A formal derivation of Heaps' law. *Inform. Sciences*, 170:263 – 272, 2005.
[20] A. Kornai. How many words are there? *Glottom.*, 2:61–86, 2002.
[21] M. A. Serrano, A. Flammini, and F. Menczer. Modeling statistical properties of written text. *PLoS ONE*, 4(4):e5372, 2009.
[22] F. Font-Clos, G. Boleda, and A. Corral. A scaling law beyond Zipf's law and its relation with Heaps' law. *New J. Phys.*, 15:093033, 2013.
[23] G. Wimmer and G. Altmann. On vocabulary richness. *J. Quant. Linguist.*, 6:1–9, 1999.
[24] Y. Sano, H. Takayasu, and M. Takayasu. Zipf's law and Heaps' law can predict the size of potential words. *Prog. Theor. Phys. Supp.*, 194:202–209, 2012.
[25] S. Bernhardsson, S. K. Baek, and P. Minnhagen. A paradoxical property of the monkey book. *J. Stat. Mech.*, 2011(07):P07013, 2011.
[26] L. Lü, Z.-K. Zhang, and T. Zhou. Deviation of Zipf's and Heaps' Laws in human languages with limited dictionary sizes. *Sci. Rep.*, 3:1–7, 2013.
[27] A. M. Petersen, J. N. Tenenbaum, S. Havlin, H. E.

Stanley, and M. Perc. Languages cool as they expand: Allometric scaling and the decreasing need for new words. *Sci. Rep.*, 2:943, 2012.

[28] M. Gerlach and E. G. Altmann. Stochastic model for the vocabulary growth in natural languages. *Phys. Rev. X*, 3:021006, 2013.

[29] M. Gerlach and E. G. Altmann. Scaling laws and fluctuations in the statistics of word frequencies. *New J. Phys.*, 16(11):113010, 2014.

[30] A. Corral, R. Ferrer-i-Cancho, and A. Díaz-Guilera. Universal complex structures in written language. *http://arxiv.org*, 0901.2924, 2009.

[31] E. G. Altmann, J. B. Pierrehumbert, and A. E. Motter. Beyond word frequency: Bursts, lulls, and scaling in the temporal distributions of words. *ArXiv*, 0901.2349v1, 2009.

[32] C. D. Manning and H. Schütze. *Foundations of Statistical Natural Language Processing*. MIT Press, Cambridge, Massachusetts, 1999.

[33] http://www.gutenberg.org/.

[34] A. Deluca and A. Corral. Fitting and goodness-of-fit test of non-truncated and truncated power-law distributions. *Acta Geophys.*, 61:1351–1394, 2013.

[35] A. Corral, A. Deluca, and R. Ferrer-i-Cancho. A practical recipe to fit discrete power-law distributions. *ArXiv*, 1209:1270, 2012.

Supplementary Information: Log-log Convexity of Type-Token Growth in Zipf's Systems

Francesc Font-Clos[1, 2] and Álvaro Corral[1, 2]

[1] *Centre de Recerca Matemàtica, Edifici C, Campus Bellaterra, E-08193 Barcelona, Spain.*
[2] *Departament de Matemàtiques, Facultat de Ciències, Universitat Autònoma de Barcelona, E-08193 Barcelona, Spain*

This document contains Supplementary Information (SI) for the article entitled "Log-log Convexity of Type-Token Growth in Zipf's Systems". It consists of two sections that give details of derivations and algebraic manipulations used to reach equations (6) and (10) in the main text.

DERIVATION OF THE APPROXIMATION FOR THE "ZERO-SUCCESS" PROBABILITY $h_{0,n}$

Equation (6) in the main text gives an approximation for the zero-success probability $h_{0,n}$.

$$h_{0,n} = \frac{\binom{L-n}{\ell}}{\binom{L}{\ell}} \simeq \left(1 - \frac{\ell}{L}\right)^n, \tag{S1}$$

While the derivation is fairly elementary, some caution must be taken in order to get the desired result. Note that other approximations are possible, but they are not useful for our purposes.
The first equal sign in Eq. (S1) is just the definition of $h_{0,n}$, i.e. the probability mass function of a hypergeometric random variable that takes value $k = 0$ (number of successes), with parameters ℓ (number of draws), L (total population) and n (number of successes in the population). To get to the desired result, we write the binomial coefficients in terms of factorials, cancel out one $\ell!$ term, and regroup the rest of terms as follows:

$$\frac{\binom{L-n}{\ell}}{\binom{L}{\ell}} = \left[\frac{(L-n)!}{(L-n-\ell)!\,\ell!}\right] \Big/ \left[\frac{L!}{(L-\ell)!\,\ell!}\right] \tag{S2}$$

$$= \frac{(L-n)!}{L!} \times \frac{(L-\ell)!}{(L-\ell-n)!}. \tag{S3}$$

Notice that each fraction in equation (S3) has n terms, so that

$$\frac{\binom{L-n}{\ell}}{\binom{L}{\ell}} = \frac{(L-\ell)(L-\ell-1)\ldots(L-\ell-n+1)}{L(L-1)\ldots(L-n+1),} \tag{S4}$$

$$= \prod_{j=0}^{n-1} \left(\frac{L-\ell-j}{L-j}\right). \tag{S5}$$

The first term, $j = 0$, is equal to $(1 - \ell/L)$. If n is small compared to L, then $L - j \simeq L$ for all $j = 0, \ldots, n-1$, and each of the n terms in the above product can be approximated by the first one. The result then follows easily, as

$$h_{0,n} = \frac{\binom{L-n}{\ell}}{\binom{L}{\ell}} = \prod_{j=0}^{n-1} \left(\frac{L-\ell-j}{L-j}\right) \simeq \prod_{j=0}^{n-1} \left(\frac{L-\ell}{L}\right) = \left(1 - \frac{\ell}{L}\right)^n. \tag{S6}$$

It is important to bear in mind that the only assumption used was $L - n \simeq L$, or equivalently, $n \ll L$, but nothing was assumed regarding the ratio ℓ/L, and thus the approximation should work equally well for any ℓ, once L and n are fixed. In practice, the most common type, with $n_{\max}$ counts, will give an upper bound to the error that this approximation introduces. In the case of written books, $n_{\max}/L \simeq 0.05$ is a typical value for many languages, so that even in the worst case, the approximation is already quite good. In addition, most types have low counts (this is in essence Zipf's law), and hence the sum of equation (8) in the main text will be dominated by terms where the approximation is very good.

DERIVATION OF THE TYPE-TOKEN GROWTH CURVE $v(\ell)$ WHEN $n_{\max}$ IS TAKEN INTO ACCOUNT

Equation (10) in the main text offers a more refined version of the type-token growth curve, equation (8), by limiting the sum up to a given maximum value of the number of counts $n_{\max}$. Let us denote $q = 1 - \ell/L$ and introduce the Lerch transcendent $\Phi(z, \gamma, a)$, defined as

$$\Phi(z, \gamma, a) = \sum_{n=0}^{\infty} \frac{z^n}{(a+n)^\gamma}, \quad |z| < 1; a \neq 0, -1, \dots \ . \tag{S7}$$

We first perform the sum $\sum_{n=1}^{n_{\max}} q^n/n^\gamma$ as follows:

$$\sum_{n=1}^{n_{\max}} \frac{q^n}{n^\gamma} = \sum_{n=1}^{\infty} \frac{q^n}{n^\gamma} - \sum_{n=n_{\max}+1}^{\infty} \frac{q^n}{n^\gamma} \tag{S8}$$

$$= \mathrm{Li}_\gamma(q) - q^{n_{\max}+1} \sum_{n=n_{\max}+1}^{\infty} \frac{q^{n-n_{\max}-1}}{n^\gamma}, \tag{S9}$$

and defining $m = n - n_{\max} - 1$,

$$\sum_{n=1}^{n_{\max}} \frac{q^n}{n^\gamma} = \mathrm{Li}_\gamma(q) - q^{n_{\max}+1} \sum_{m=0}^{\infty} \frac{q^m}{(m + n_{\max} + 1)^\gamma} \tag{S10}$$

$$= \mathrm{Li}_\gamma(q) - q^{n_{\max}+1} \Phi(q, \gamma, n_{\max} + 1). \tag{S11}$$

The normalization constant can be similarly computed,

$$A^{-1} = \sum_{n=1}^{n_{\max}} \frac{1}{n^\gamma} = \sum_{n=1}^{\infty} \frac{1}{n^\gamma} - \sum_{n=n_{\max}+1}^{\infty} \frac{1}{n^\gamma} \tag{S12}$$

$$= \zeta(\gamma) - \Phi(1, \gamma, n_{\max} + 1), \tag{S13}$$

and the result in the main text follows:

$$v(\ell) = V \left(1 - A \sum_{n=1}^{n_{\max}} \frac{(1 - \ell/L)^n}{n^\gamma} \right) \tag{S14}$$

$$= V \left(1 - \frac{\mathrm{Li}_\gamma(q) - q^{n_{\max}+1} \Phi(q, \gamma, n_{\max} + 1)}{\zeta(\gamma) - \Phi(1, \gamma, n_{\max} + 1)} \right). \tag{S15}$$

As discussed in the main text, this improved version of the type-token growth curve makes use of an additional parameter, $n_{\max}$, and so it is not surprising that it gives better results. However, in practice only for $\gamma < 2$ is this usually noticeable. This is related to the tail of the distribution of counts, as for fixed $n_{\max}, \ell/L$ the overall weight given to the types with $n > n_{\max}$ [1] increases if γ is decreased. In any case, it is worth noting that (i) the additional "parameter" $n_{\max}$ is usually known in practical cases, so that it does not need to be fitted, and (ii) the polylogarithm function, Riemman's zeta function and the Lerch transcendent are functions with well-studied properties and usually can be found in most numerical packages.

---

[1] Obviously, these types are not present in the system, but the simple version of the type-token curve, Eq. (9) in the main text, includes them in the sum.

3.3 The perils of thresholding

Francesc Font-Clos, Gunnar Pruessner, Nicholas R. Moloney and Anna Deluca
New Journal of Physics **17** (2015) 043066

IOP Publishing *New J. Phys.* **17** (2015) 043066 doi:10.1088/1367-2630/17/4/043066

New Journal of Physics
The open access journal at the forefront of physics

Deutsche Physikalische Gesellschaft DPG
IOP Institute of Physics

Published in partnership with: Deutsche Physikalische Gesellschaft and the Institute of Physics

OPEN ACCESS

RECEIVED
30 November 2014

REVISED
18 March 2015

ACCEPTED FOR PUBLICATION
23 March 2015

PUBLISHED
30 April 2015

PAPER

The perils of thresholding

Francesc Font-Clos[1,2], Gunnar Pruessner[3], Nicholas R Moloney[4] and Anna Deluca[5]

[1] Centre de Recerca Matemàtica, Edifici C, Campus Bellaterra, E-08193 Bellaterra, Barcelona, Spain
[2] Department de Matemàtiques, Universitat Autònoma de Barcelona, Edifici C, E-08193 Bellaterra, Barcelona, Spain
[3] Department of Mathematics, Imperial College London, 180 Queen's Gate, London SW7 2BZ, UK
[4] London Mathematical Laboratory, 14 Buckingham Street, London WC2N 6DF, UK
[5] Max Planck Institute for the Physics of Complex Systems, Nöthnitzer Straße 38, D-01187 Dresden, Germany

E-mail: fontclos@crm.cat

Keywords: thresholding, double scaling, birth–death process

Abstract

The thresholding of time series of activity or intensity is frequently used to define and differentiate events. This is either implicit, for example due to resolution limits, or explicit, in order to filter certain small scale physics from the supposed true asymptotic events. Thresholding the birth–death process, however, introduces a scaling region into the event size distribution, which is characterized by an exponent that is unrelated to the actual asymptote and is rather an artefact of thresholding. As a result, numerical fits of simulation data produce a range of exponents, with the true asymptote visible only in the tail of the distribution. This tail is increasingly difficult to sample as the threshold is increased. In the present case, the exponents and the spurious nature of the scaling region can be determined analytically, thus demonstrating the way in which thresholding conceals the true asymptote. The analysis also suggests a procedure for detecting the influence of the threshold by means of a data collapse involving the threshold-imposed scale.

1. Introduction

Thresholding is a procedure applied to (experimental) data either deliberately, or effectively because of device limitations. The threshold may define the onset of an event and/or an effective zero, such that below the threshold the signal is regarded as 0. An example of thresholding is shown in figure 1. Experimental data often comes with a detection threshold that cannot be avoided, either because the device is insensitive below a certain signal level, or because the signal cannot be distinguished from noise. The quality of a measurement process is often quantified by the noise to signal ratio, with the implication that high levels of noise lead to poor (resolution of the) data. Often, the rationale behind thresholding is to weed out small events which are assumed irrelevant on large scales, thereby retaining only the asymptotically big events which are expected to reveal (possibly universal) large-scale physics.

Most, if not all, of physics is due to some basic interactions that occur on a 'microscopic length scale', say the interaction between water droplets or the van der Waals forces between individual water molecules. These length scales separate different realms of physics, such as between micro-fluidics and molecular physics or between molecular physics and atomic physics. However, these are *not* examples of the thresholds we are concerned with in the following. Rather, we are interested in an often arbitrary microscopic length scale well above the scale of the microscopic physics that governs the phenomenon we are studying, such as the spatiotemporal resolution of a radar observing precipitation (which is much coarser than the scale set by microfluidics), or the resolution of the magnetometer observing solar flares (which is much coarser than the scale set by atomic physics and plasma magnetohydrodynamics).

Such thresholds often come down to the device limitations of the measuring apparatus, the storage facilities connected to it, or the bandwidth available to transmit the data. For example, the earthquake catalogue of Southern California is only complete above magnitude 3, even though the detection-threshold is around magnitude 2 [1]. One fundamental problem is the noise-to-signal ratio mentioned above. Even if devices were to improve to the level where the effect of noise can be disregarded, thresholding may still be an integral part of the

IOP Publishing New J. Phys. 17 (2015) 043066 F Font-Clos et al

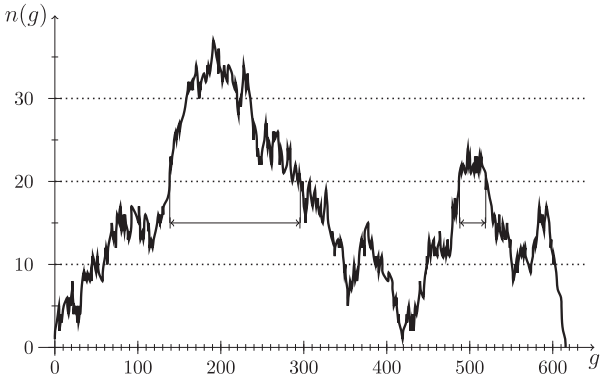

Figure 1. Example of thresholding of a time series. An event begins when the signal exceeds the threshold (dotted lines, $h = 10,\ 20,\ 30$) and ends as soon as the signal falls below the threshold. Increasing levels of the threshold lead (non-monotonically) to different numbers of events and, provided the signal eventually ends, monotonically smaller total event durations. The main focus of this paper is on the statistics of the individual event durations, as exemplified by the two intervals for the intermediate threshold.

measurement. For example, the distinction between rainfall and individual drops requires a separation of microscale and macroscale which can be highly inhomogeneous [2]. Solar flares, meanwhile, are defined to start when the solar activity exceeds the threshold and end when it drops below, but the underlying solar activity never actually ceases [3].

Thresholding has also played an important rôle in theoretical models, such as the Bak–Sneppen model [4] of self-organized criticality [5], where the scaling of the event-size distribution is a function of the threshold [6] whose precise value was the subject of much debate [7, 8]. Finite size effects compete with the threshold-imposed scale, which has been used in some models to exploit correlations and predict extreme events [9].

Often, thresholding is tacitly assumed to be 'harmless' for the (asymptotic) observables of interest and beneficial for the numerical analysis. We will argue in the following that this assumption may be unfounded: the very act of thresholding can distort the data and the observables derived from it. To demonstrate this, we will present an example of the effect of thresholding by determining the apparent scaling exponents of a simple stochastic process, the birth–death process (BDP). We will show that thresholding obscures the asymptotic scaling region by introducing an additional prior scaling region, solely as an artefact. Owing to the simplicity of the process, we can calculate the exponents, leading order amplitudes and the crossover behaviour analytically, in excellent agreement with simulations. In doing so, we highlight the importance of sample size since, for small samples (such as might be accessible experimentally), only the 'spurious' threshold-induced scaling region that governs the process at small scales may be accessible. Finally, we discuss the consequences of our findings for experimental data analysis, where detailed knowledge of the underlying process may not be available, usually the mechanism behind the process of interest is unclear, and hence such a detailed analysis is not feasible. But by attempting a data collapse onto a scaling ansatz that includes the threshold-induced scale, we indicate how the effects of thresholding can be revealed.

The outline of the paper is as follows: in section 2 we introduce the model and the thresholding applied to it. To illustrate the problems that occur when thresholding real data, we analyse in detail some numerical data. The artefact discovered in this analysis finds explanation in the theory present in section 3. We discuss these findings and suggest ways to detect the problem in the final section.

2. Model

In order to quantify numerically and analytically the effect of thresholding, we study the BDP [10] with Poissonian reproduction and extinction rates that are proportional to the population size. More concretely, we consider the population size $n(g)$ at (generational) time $g \geqslant 0$. Each individual in the population reproduces and dies with the same rate of 1/2 (in total unity, so that there are $n(g)$ birth or death events or 'updates' per time unit on average); in the former case (birth) the population size increases by 1, in the latter (death) it decreases by 1. The state $n(g) = 0$ is absorbing [11]. Because the instantaneous rate with which the population $n(g)$ evolves is $n(g)$ itself, the exponential distributions from which the random waiting times between events are drawn are themselves parameterized by a random variable, $n(g)$.

Because birth and death rates balance each other, the process is said to be at its critical point [12], which has the peculiar feature that the expectation of the population is constant in time, $\langle n(g) \rangle = n(g_0)$, where $\langle \cdot \rangle$

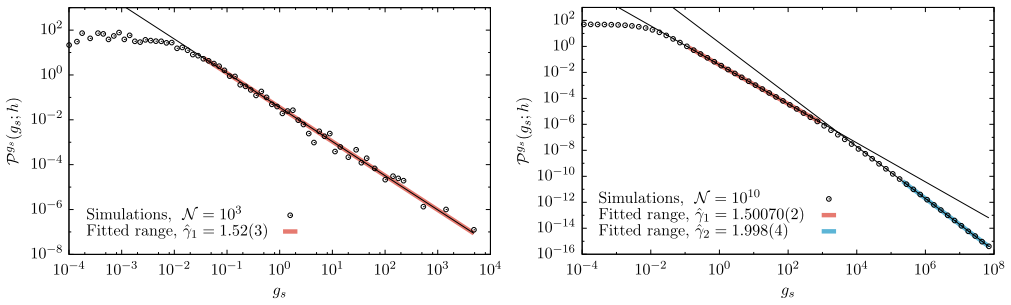

Figure 2. (a): the PDF $\mathcal{P}^{g_s}(g_s;\ h)$ of the survival time g_s of a thresholded BDP, with a threshold of $h = 100$, estimated from Monte Carlo simulations using a limited sample size of $\mathcal{N} = 10^3$. Fitting a power law yields an exponent of $\hat{\gamma}_1 = 1.52(3)$ over the range $[0.031,\ 1.259 \times 10^5]$, with a p-value of 0.71. (b): same as above, but using a sample size of $\mathcal{N} = 10^{10}$. In this case, two power laws can be fitted in two different regimes: below $g_X = 8\pi h$, we find $\hat{\gamma}_1 = 1.50070(2)$ in the (fixed) range $[10^{-2},\ 10^3]$, while above g_X, the fit leads to $\hat{\gamma}_2 = 1.998(4)$ over the range $[1.99 \times 10^5,\ 3.16 \times 10^8]$, with a p-value of 0.55. Monte Carlo simulations are shown as symbols, while the small (large) regime power-law fit is plotted with full black lines, and the fitted range marked with red (blue) shading.

denotes the expectation and $n(g_0)$ is the initial condition, set to unity in the following. This constant expectation is maintained by increasingly fewer surviving realizations, as each realization of the process terminates almost surely. We therefore define the survival time as the time $g_s - g_0$ such that $n(g) > 0$ for all $g_0 \leqslant g < g_s$ and $n(g) = 0$ for all $g \geqslant g_s$. For simplicity, we may shift times to $g_0 = 0$, so that g_s itself is the survival time. It is a continuous random variable, whose probability density function (PDF) is well known to have a power law tail in large times, $\mathcal{P}^{g_s}(g_s) \propto g_s^{-2}$ [12, as in the branching process].

In the following, we will introduce a threshold, which mimics the suppression of some measurements either intentionally or because of device limitations. For the BDP this means that the population size (or, say, 'activity') below a certain, prescribed level, h, is treated as 0 when determining survival times. In the spirit of [3, also solar flares, 13], the threshold allows us to distinguish events, which, loosely speaking, start and end whenever $n(g)$ passes through h.

Explicitly, events start at g_0 when $\lim_{\epsilon \to 0^+} n(g_0 - \epsilon) = h$ and $n(g_0) = h + 1$. They end at g_s when $n(g_s) = h$, with the condition $n(g) > h$ for all $g_0 \leqslant g < g_s$. This is illustrated in figures 1 and 4. No thresholding takes place (i.e. the usual BD process is recovered) for $h = 0$, in which case the initial condition is $n(g_0) = 1$ and termination takes place at g_s when $n(g_s) = 0$. For $h > 0$ one may think of $n(g)$ as an 'ongoing' time series which never ceases and which may occasionally 'cross' h from below (starting the clock), returning to h some time later (stopping the clock). In a numerical simulation one would start $n(g)$ from $n(g_0) = h + 1$ at $g_0 = 0$ and wait for $n(g)$ to arrive at $n(g) = h$ from above. The algorithm may be summarized as

```
for i = 1...𝒩 do
  n ← h+1
  g_i ← 0
  while n > h do
    g_i ← g_i + ξ(n)
    n ← n+b
  end while
end for
```

where $\xi(n)$ is an exponential random variable with rate n, and b stands for a random variable that takes the values $\{-1,\ 1\}$ with probability 1/2. In our implementation of the algorithm, all random variables are handled with the GNU Scientific Library [14].

2.1. Numerics and data analysis

Monte-Carlo runs of the model reveal something unexpected: The exponent of the PDF of the thresholded BDP appears to change from $\mathcal{P}^{g_s}(g_s) \propto g_s^{-2}$ at $h = 0$ to $\mathcal{P}^{g_s}(g_s) \propto g_s^{-3/2}$ at $h = 100$ or, in fact, any reasonably large $h \gtrsim 10$. Figure 2 shows $\mathcal{P}^{g_s}(g_s)$ for the case of $h = 100$ and two different sample sizes, $\mathcal{N}_1 = 10^3$ and $\mathcal{N}_2 = 10^{10}$, corresponding to 'scarce data' and 'abundant data', respectively. In the former case, the exponent of the PDF is estimated to be $\hat{\gamma}_1 = 1.52(3) \approx 3/2$; in the latter, the PDF splits into two scaling regimes, with exponents $\hat{\gamma}_1 = 1.50070(2) \approx 3/2$ and $\hat{\gamma}_2 = 1.998(4) \approx 2$. This phenomenon can be investigated systematically for different sample sizes $\mathcal{N}$ and thresholds h.

We use the fitting procedure introduced in [15], which is designed not only to estimate the exponent, but to determine the range in which a power law holds in an objective way. It is based on maximum likelihood

IOP Publishing New J. Phys. 17 (2015) 043066 F Font-Clos et al

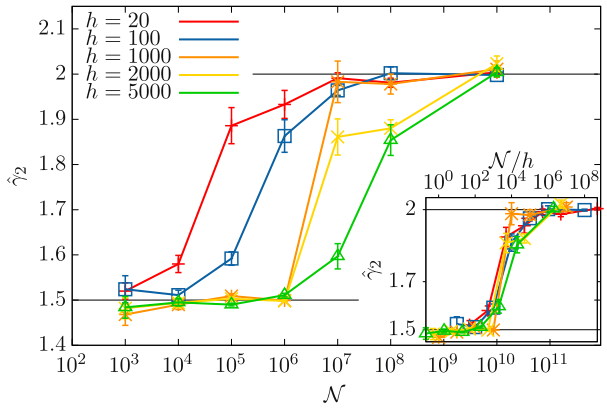

Figure 3. Estimated large scale exponent $\hat{\gamma}_2$ for different thresholds h and sample size $\mathcal{N}$. The error bars correspond to one standard deviation and are inversely proportional to the number of data within the fitted range. *Inset*: estimated large scale exponent $\hat{\gamma}_2$ as a function of a 'rescaled sample size' $\mathcal{N}/h$.

methods, the Kolmogorov–Smirnov (KS) test and Monte Carlo simulations of the distributions, see appendix A for details. In figure 3 we show the evolution of the estimated large scale exponent, $\hat{\gamma}_2$, for different $\mathcal{N}$ and for different h. The fits are made by assuming that there is a true power law in a finite range $[a, b]$. For values of the exponent between 1.5 and 2 larger error bars are observed. For these cases, less data is fitted but the fitting range is always at least two orders of magnitude wide.

It is clear from figure 3 that $\mathcal{N}$ has to be very large in order to see the true limiting exponent. Even the smallest h investigated, $h = 20$, needs a sample size of at least $\mathcal{N} = 10^7$, while for $h = 5000$ the correct exponent is not found with less than about $\mathcal{N} = 10^{10}$. It is natural to ask how large the applied thresholds are compared to the average signal amplitude A or maximum M. Focusing on the case shown in figure 2(a), where $h = 100$ and $\mathcal{N} = 10^3$, we find that $h \simeq 0.07\langle A\rangle \simeq 0.02\langle M\rangle$, so that in this sense, the thresholds can be regarded as 'small'.

The mere introduction of a threshold therefore changes the PDF of events sizes significantly. It introduces a new, large scaling regime, with an exponent that is misleadingly different from that characterizing large scale asymptotics. In fact, for small sample sizes ($\mathcal{N}_1 = 10^3$, see figure 2(a)), the only visible regime is that induced by thresholding (in our example, $\gamma_1 = 3/2$), while the second exponent ($\gamma_2 = 2$), which, as will be demonstrated below, governs the large scale asymptotics, remains hidden unless much larger sample sizes are used (figure 2(b)).

In the inset of figure 3 we plot the fitted values $\hat{\gamma}_2$ as a function of the rescaled sample size $\mathcal{N}/h$. The data collapse is remarkable: the sample size required to recover the exponent $\hat{\gamma}_2$ grows linearly with the threshold h. This is in agreement with the scaling of the crossover that separates the two scaling regimes, $g_\times \propto h$, see section 3.2.1.

Although the algorithm is easy to implement, finding the two scaling regimes numerically can be challenging. There are a number of caveats:

(1) The crossover point $g_\times$ between the two scaling regimes scales linearly with the threshold, $g_\times = 8\pi h$ (see section 3.2.1), effectively shifting the whole g_s^{-2} asymptotic regime to larger and thus less likely values of g_s. To maintain the same *number* of events above $g_\times \propto h$, one needs $\mathcal{N}\int_{g_\times}^{\infty} \mathrm{d}g_s g_s^{-2} = \text{const}$, i.e. $\mathcal{N} \propto h$.

(2) Because the expected running time of the algorithm diverges, one has to set an upper cutoff on the maximum generational timescale, say $g_s < G$. If the computational complexity for each update is constant, an individual realization, starting from $n(0) = h + 1$ and running up to $n(g_s) = h$ with $g_s < G$, has complexity $\mathcal{O}(g_s^2)$ in large g_s where g_s^2 is the scaling of the expected survival time of the mapped random walker introduced below. The expected complexity of realizations that terminate before G (with rate $\sim 1/g_s^2$) is therefore linear in G, $\int_1^G \mathrm{d}g_s g_s^{-2} g_s^2 = G - 1$. With the random walker mapping it is easy to see that the expected population size $n(g)$ of realizations that terminate after G (and therefore have to be discarded as g_s exceeds G) is of the order $n(g_s) \sim G$ for $g_s = G$. These realizations, which appear with frequency $\propto 1/G$, have complexity $\mathcal{O}(G^2)$, i.e. the complexity of realizations of the BDP is $\mathcal{O}(G)$ both for those counted into the final tally and those dismissed because they exceed G. There is no point probing beyond G if $\mathcal{N}$ is too small to produce a reasonable large sample on a logarithmic scale, $\mathcal{N}\int_G^{2G} \mathrm{d}g_s g_s^{-2} = \text{const}$, so that $\mathcal{N} \sim G$

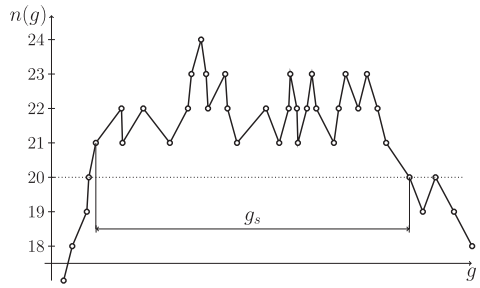

Figure 4. Magnification of the right interval in figure 1. The clock starts when $n(g)$ exceeds the threshold and stops when $n(g)$ returns to the threshold.

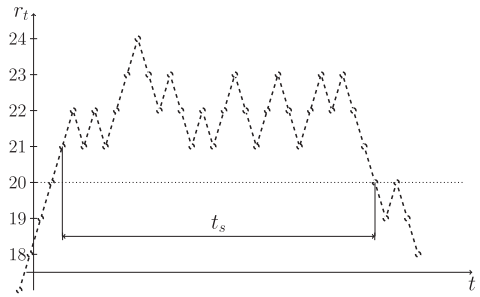

Figure 5. The same data as in figure 4 but on the mapped time scale of the random walker, which evolves in equally spaced, discrete steps. The survival time is necessarily odd, $t_s = 2T - 1$, $T \in \mathbb{N}$ ($t_s = 29$ in this example).

and thus the overall complexity of a sample of size $\mathcal{N}$ is $\mathcal{O}(\mathcal{N}^2)$ and thus $\mathcal{O}(h^2)$ for $G \sim g_X \sim h$ and $\mathcal{N} \propto h$ from above.

That is, larger h necessitates larger $\mathcal{N}$, leading to *quadratically* longer CPU time. In addition, parallelization of the algorithm helps only up to a point, as the (few) biggest events require as much CPU time as all the smaller events taken together. The combination of all these factors has the unfortunate consequence that, for large enough values of h, observing the $\mathcal{P}^{g_s}(g_s) \propto g_s^{-2}$ regime is simply out of reach (even for moderate values of h, such as $h = 100$, to show the crossover as clearly as in figure 2, a sample size as large as $\mathcal{N} = 9 \times 10^9$ was necessary, which required about 1810 h of CPU time).

3. Results

While it is straightforward to set up a recurrence relation for the generating function if the threshold is $h = 0$, the same is not true for $h > 0$. This is because the former setup ($h = 0$) does not require an explicit implementation of the absorbing wall since the process terminates naturally when $n(g) = 0$ (there is no individual left that can reproduce or die). If, however, $h > 0$, the absorbing wall has to be treated explicitly and that is difficult when the evolution of the process (the effective diffusion constant) is a function of its state, i.e. the noise is multiplicative. In particular, a mirror charge trick cannot be applied.

However, the process can be mapped to a simple random walk by 'a change of clocks', a method detailed in [16]. For the present model, we observe that $n(g)$ performs a fair random walk r_t by a suitable mapping of the generational timescale g to that of the random walker, $r_t(g) = n(g)$ with $t(g) \in \mathbb{N}$. In fact, because of the Poissonian nature of the BD process, birth and death almost surely never occur simultaneously and a suitable, unique $t(g)$ is found by $t(0) = 0$ and

$$\lim_{\epsilon \to 0^+} t(g + \epsilon) - t(g - \epsilon) = \lim_{\epsilon \to 0^+} |n(g + \epsilon) - n(g - \epsilon)| \tag{1}$$

i.e. $t(g)$ increases whenever $n(g)$ changes and is therefore an increasing function of g. With this map, r_t is a simple random walk along an absorbing wall at h, see figure 5. The challenge is to derive the statistics of the survival times g_s on the time scale of the BD process from the survival times t_s on the time scale of the random walk.

IOP Publishing New J. Phys. 17 (2015) 043066 F Font-Clos et al

In the following, we first approximate some important properties of the survival times in a handwaving manner before presenting a mathematically sound derivation in section 3.2.

3.1. Approximation

The expected waiting time[6] between two events in the BDP is $1/n$, if n is the current population size, with $n = n_x + h$ such that n_x is the excess of n above h. As discussed in detail in section 3.2, n_x is a time-dependent random variable, and so taking the ensemble average of the waiting time is a difficult task. But on the more convenient time scale t, the excess n_x performs a random walk and it is in that ensemble, with that time scale, where we attempt to find the expectation

$$\overline{g_s\left(t_s;\, h\right)} = \sum_{t=0}^{t_s-1} \left\langle \frac{1}{n_x(t) + h} \right\rangle_{\mathcal{R}\left(t_s\right)}, \tag{2}$$

which is the expected survival time of a thresholded BD process given a certain return (or survival) time t_s of the random walker. In this expression $n_x(t)$ is a time-dependent random variable and the ensemble average $\langle \cdot \rangle_{\mathcal{R}(t_s)}$ is taken over all random walker trajectories $\mathcal{R}(t_s)$ with return time t_s. To ease notation, we will include the argument of $\mathcal{R}(t_s)$ only where necessary. Replacing the random variable g_s by its mean $\overline{g_s(t_s;\, h)}$, the PDFs for t_s and g_s are approximately related via,

$$\mathcal{P}^{g_s}(g_s)\frac{\mathrm{d}}{\mathrm{d}t_s}\overline{g_s\left(t_s;\, h\right)} \approx \mathcal{P}^{t_s}\left(t_s\right). \tag{3}$$

This map will be made rigorous in section 3.2, avoiding the use of $\overline{g_s(t_s;\, h)}$ in lieu of the random variable.

In a more brutal approach, one may approximate the time dependent excess $n_x(t)$ in equation (2) by its expectation conditional to a certain survival time t_s,

$$\begin{aligned}\left\langle \frac{1}{h + n_x(t)} \right\rangle_{\mathcal{R}} &= \frac{1}{h + \langle n_x(t)\rangle_{\mathcal{R}}} \left\langle \frac{1}{1 + \dfrac{n_x(t) - \langle n_x(t)\rangle_{\mathcal{R}}}{h + \langle n_x(t)\rangle_{\mathcal{R}}}} \right\rangle \\ &= \frac{1}{h + \langle n_x(t)\rangle_{\mathcal{R}}} + (\text{higher order terms})\end{aligned} \tag{4}$$

so that the expected survival time $g_s(t_s)$ given a certain return time t_s is approximately $t_s/(h + \langle n_x(t)\rangle_{\mathcal{R}})$.

The quantity $\langle n_x(t)\rangle_{\mathcal{R}}$ is the expected excursion of a random walker, which is well-known to be

$$\langle n_x(t)\rangle_{\mathcal{R}} \approx \sqrt{\frac{\pi}{8}}\, t_s^{1/2} \tag{5}$$

in the continuum limit (with diffusion constant 1/2) (e.g. [17, 18]). Thus

$$\overline{g_s\left(t_s;\, h\right)} \approx \frac{t_s}{h + \sqrt{\pi t_s/8}}. \tag{6}$$

At small times, $h \gg \sqrt{\pi t_s/8}$, the relation between g_s and t_s is essentially linear, $g_s \approx t_s/h$, whereas for large times, $h \ll \sqrt{\pi t_s/8}$, the asymptote is $g_s \approx \sqrt{8t_s/\pi}$. Writing the right-hand side of equation (6) in the form $\sqrt{8t_s/\pi}\,\frac{1}{1+\sqrt{8h^2/(\pi t_s)}}$ allows us to extract the scaling of the crossover time. The argument of the square root is of order unity when $t_\times = 8h^2/\pi$, for which $g_s(t_\times,\, h) \approx 4h/\pi$. Moreover, one can read off the scaling form

$$\overline{g_s\left(t_s;\, h\right)} \approx t_s^{1/2}\mathcal{G}\left(t_s/h^2\right), \tag{7}$$

with $\mathcal{G}(x) = \sqrt{8/\pi}/(1 + \sqrt{8/(\pi x)})$ and asymptotes $\mathcal{G}(x) \approx \sqrt{x}$ for small x and $\lim_{x\to\infty}\mathcal{G}(x) = \sqrt{8/\pi}$.

The PDF of the survival time

$$\mathcal{P}^{t_s}\left(t_s\right) = \frac{1}{\sqrt{4\pi D t_s}}\frac{a}{D t_s}\exp\left(-\frac{a^2}{4Dt_s}\right) \tag{8}$$

of a random walker along an absorbing wall is well-known to be a power law $\propto t_s^{-3/2}$ for times t_s large compared to the time scale set by the initial condition, i.e. the distance a of the random walker from the absorbing wall at time $t = 0$. The precise value of a is effectively determined by the details the continuum approximation, here $a = 1$, $D = 1/2$, and so we require $1 \ll 2t_s$.

[6] In a numerical simulation this would be the time increment.

To derive the PDF of the BD process, note that equation (6) has the unique inverse $t_s(g_s) = \frac{\pi g_s^2}{16}\mathcal{T}(\frac{16h}{\pi g_s})$, where $\mathcal{T}(y) = 1 + y + \sqrt{1+2y}$. Evaluating the crossover time by setting $y = 1$ yields $g_X = 16h/\pi$. The PDF of the survival time of the BD process finally reads

$$\mathcal{P}^{g_s}\left(g_s;\, h\right) \sim \left(\frac{\pi}{16}\mathcal{T}(y)\right)^{-1/2} g_s^{-2}\left(2 - \frac{y\mathcal{T}'(y)}{\mathcal{T}(y)}\right), \tag{9}$$

where $y = \dfrac{16h}{\pi g_s}$. For small y, the last bracket converges to 2, so $\mathcal{P}^{g_s}(g_s;\, h) \sim 2\sqrt{8/\pi}\, g_s^{-2}$ for large g_s. For large y, the last bracket converges to 1, so $\mathcal{P}^{g_s}(g_s;\, h) \sim (1/\sqrt{h})\, g_s^{-3/2}$ for small g_s.

This procedure recovers the results in section 3.2: for $g_s \ll 16h/\pi$ the PDF of the survival times in the BD process goes like $g_s^{-3/2}$, and for $g_s \gg 16h/\pi$ like g_s^{-2}, independent of h. Equation (9) also gives a prescription for a collapse, since $\mathcal{P}^{g_s}(g_s;\, h)\, g_s^2$ plotted versus g_s/h should, for sufficiently large g_s, reproduce the same curve, as confirmed in figures 7 and 8.

Applying a threshold introduces a new scale, $16\,h/\pi$, below which the PDF displays a clearly discernible power law, $g_s^{-3/2}$, corresponding to the return time of a random walker. The 'true' g_s^{-2} power law behaviour (the large g_s asymptote) is visible only well above the threshold-induced crossover.

3.2. Detailed analysis

In the previous section we made a number of assumptions, in particular the approximation of replacing the random variable by its expectation, and the approximation in equation (4), which both require further justification.

In the present section we proceed more systematically. In particular, we will be concerned with the statistics of the BD survival time $g_s(\mathcal{R})$ *given* a particular *trajectory* $\mathcal{R} = \{r_0, r_1, \ldots, r_{t_s}\}$ of the random walk, where $t_s = 2T - 1$, necessarily odd, $T \in \mathbb{N}$, see figures 5 and B1 . We will then relax the constraint of the trajectory and study the whole ensemble Ω of random walks terminating at a particular time $2T - 1$, denoting as $g_s(\Omega(T))$ a survival time drawn from the distribution of all survival times of a BD process with a mapping to a random walker that terminates at $2T - 1$ or, for simplicity, just $g_s(\Omega)$. This will allow us to determine the existence of a limiting distribution for $g_s(\Omega)/\sqrt{T}$ and to make a quantitative statement about its mean and variance. We will *not* make any assumptions about the details of that limiting distribution; in order to determine the asymptotes of $\mathcal{P}^{g_s}(g_s;\, h)$ we need only know that the limit exists.

For a given trajectory $\mathcal{R}$ of the random walk, the resulting generational survival time $g_s(\mathcal{R})$ may be written as

$$g_s(\mathcal{R}) = \sum_{t=0}^{2T-2} \xi_t\left(r_t + h\right), \tag{10}$$

where $\xi_t(\alpha)$ is a random variable drawn at time t from an exponential distribution with rate α, i.e. drawn from $\alpha e^{-\alpha\xi}$, and r_t is the position of the random walk at time t, with initial condition $r_0 = 1$ and terminating at $2T - 1$ with $r_{2T-1} = 0$ (see figure B1).

The mean and standard deviation of ξ_t are $1/(r_t + h)$, necessarily finite, so that by the central limit theorem the limiting distribution of $g_s(\mathcal{R})/\sqrt{T}$ *given* a trajectory $\mathcal{R}$ is Gaussian (for $T \gg 1$). This ensures that $g_s(\Omega)/\sqrt{T}$ has a limiting distribution (see appendix C).

It is straightforward to calculate the mean and standard deviation of $g_s(\mathcal{R})$ for a particular trajectory $\mathcal{R}$ that terminates after $2T - 1$ steps. Slightly more challenging is the mean $\mu(\Omega)$ and variance $\sigma^2(\Omega)$ of $g_s(\Omega)$ for the entire ensemble Ω of such trajectories. The details of this calculation are relegated to appendix B. Here, we state only the key results. For the mean of the survival time, we find

$$\mu(\Omega) \simeq 2\sqrt{\pi T} + 2h\psi\left(\frac{h}{\sqrt{T}}\right) \tag{11}$$

(see equation (B.22)) with $\psi(x) = e^{-x^2}(\mathrm{Ei}(x) - \pi\mathcal{E}(\imath x)/\imath)$ and asymptotes

$$\mu(\Omega) \simeq \begin{cases} 2\sqrt{\pi T} & \text{for } T \gg h^2 \\ 2T/h & \text{for } T \ll h^2 \end{cases} \tag{12}$$

see equation (B.24). The variance is

$$\sigma^2(\Omega) \simeq T\ \mathcal{I}(x) - \mu(\Omega)^2 + \mathcal{K}(x) \tag{13}$$

(see equation (B.27)) with integrals $\mathcal{I}(x)$ and $\mathcal{K}(x)$ defined in equation (B.28*a*) and with asymptotes

$$\sigma^2(\Omega) \simeq \begin{cases} 4\pi T \dfrac{\pi - 3}{3} & \text{for } T \gg h^2, \\ 2T/h^2 & \text{for } T \ll h^2, \end{cases} \tag{14}$$

see equation (B.32). All these results are derived in the limit $T \gg 1$ in which the mapped random walker takes more than just a few steps, corresponding to a continuum approximation. However, as shown in the following, the results remain valid even for T close to one.

To assess the quality of the continuum approximation and the validity of the asymptotes, we extracted the mean $\mu(\Omega(T))$ and variance $\sigma^2(\Omega(T))$ of the survival time $g_s(\Omega(T))$ from simulated BDPs starting with a population size $n(0) = h + 1$ and returning to $n(g_s) = h$ after $2T - 1$ updates (births or deaths), i.e. the process was conditioned to a particular value of T. In particular, we set the threshold at $h = 100$, and simulated a sample of 10^5 constrained BDPs for values $T = 2^k$, $k = 0 \ldots 20$. The results are shown in figure 6 and confirm the validity of the large $T \gg 1$ approximation in equations (11) and (13), as well as the asymptotes (12) and (14). Remarkably, as previously stated, equations (11) and (13) are seen to be valid even when the condition $T \gg 1$ does not reasonably hold.

3.2.1. Distribution of g_s

For large T, the generational survival time g_s *given* a survival time $2T - 1$ of the mapped random walk has PDF

$$\mathcal{P}^{g_s}\left(g_s;\, h;\, T\right) \simeq \frac{1}{\sqrt{\sigma^2(\Omega(T))}} \Phi\left(\frac{g_s - \mu(\Omega(T))}{\sqrt{\sigma^2(\Omega(T))}}\right), \tag{15}$$

where $\Phi(x)$ denotes the limiting distribution of the rescaled survival time $(g_s - \mu(\Omega(T)))/\sqrt{\sigma^2(\Omega(T))}$, and the mean $\mu(\Omega(T))$ and variance $\sigma^2(\Omega(T))$ are given by equations (11) and (13). We demonstrate that Φ exists and find its precise (non-Gaussian) form in appendix C for completeness, but we will not use this result in what follows: to extract the asymptotic exponents and first order amplitudes, see below, knowledge of the mean $\mu(\Omega)$ and variance $\sigma^2(\Omega)$ is sufficient.

As the ensembles $\Omega(T)$ are disjoint for different T, the overall distribution $\mathcal{P}^{g_s}(g_s;\, h)$ of survival generational times is therefore given by the sum of the constrained distribution $\mathcal{P}^{g_s}(g_s;\, h;\, T)$ weighted by the probability of the mapped random walk to terminate after $2T - 1$ steps. In the limit of large T, as assumed throughout, that weight is $T^{-3/2}/(2\sqrt{\pi})$ [19]. Therefore

$$\mathcal{P}^{g_s}\left(g_s;\, h\right) = \sum_{T=1}^{\infty} \frac{T^{-3/2}}{2\sqrt{\pi}} \frac{1}{\sqrt{\sigma^2(\Omega(T))}} \Phi\left(\frac{g_s - \mu(\Omega(T))}{\sqrt{\sigma^2(\Omega(T))}}\right). \tag{16}$$

To extract asymptotic behaviour for $T \ll h^2$ and $T \gg h^2$ we make a crude saddle point, or 'pinching' approximation, by assuming that $\Phi(x)$ essentially vanishes for $|x| > 1/2$ and is unity otherwise. This fixes the random walker time T via $g_s - \mu(\Omega(T)) = 0$, while the number of terms in the summation is restricted to satisfy $|g_s - \mu(\Omega(T))| \leqslant \sqrt{\sigma^2(\Omega(T))}$. After some algebra we find

$$\mathcal{P}^{g_s}\left(g_s;\, h\right) = \begin{cases} \dfrac{h+1}{2} & \text{for } g_s \ll 1/h, \\ \dfrac{g^{-3/2}}{\sqrt{2\pi h}} & \text{for } 1/h \ll g_s \ll 8\pi h, \\ 2g^{-2} & \text{for } g_s \gg 8\pi h. \end{cases} \tag{17}$$

The qualitative scaling of these two asymptotes was anticipated after equation (9). The crossover time $g_X = 8\pi h$, shown in figures 7 and 8, can be determined by assuming continuity of $\mathcal{P}^{g_s}(g_s;\, h)$ and thus imposing $\frac{1}{\sqrt{2h\pi}} g_X^{-3/2} = 2g_X^{-2}$. Figure 7 shows $\mathcal{P}^{g_s}(g_s;\, h) g_s^2$ versus g_s/h for varying h, comparing Monte Carlo simulations for varying h with the numerical evaluation of equation (16) for $h = 100$, thus confirming the validity of the data collapse proposed in equation (9). In particular, the shape of the transition between the two asymptotic regimes, predicted to take place near $g_X/h = 8\pi$, is recovered from equation (16) with great accuracy. As an alternative to the numerical evaluation of equation (16), we introduce in appendix D a complementary approach that provides the Laplace transform of $\mathcal{P}^{g_s}(g_s;\, h)$, see equation (D.4). Unfortunately, inverting the Laplace transform analytically does not seem feasible, but numerical inversion provides a perhaps simpler means of evaluating $\mathcal{P}^{g_s}(g_s;\, h)$ in practice.

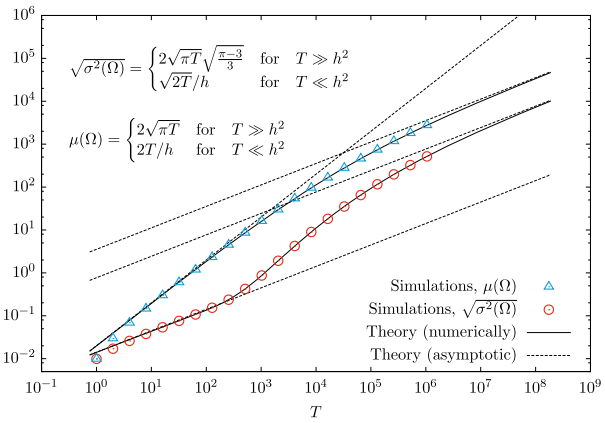

Figure 6. Numerical comparison of the approximations equations (11) and (13) (shown as full lines), their asymptotes equations (12) and (14) (dashed) and the numerical estimates based on a sample of 10^5 realizations per datapoint in a Monte-Carlo simulation of a birth–death process constrained to $2T - 1$ updates, with $h = 100$ and $T = 2^k$, $k = 0 \ldots 20$.

In addition to the two asymptotic regimes discussed so far, one notices that figure 8 displays yet another 'regime' (left-most, green shading), which corresponds to extremely short survival times. This regime is almost exclusively due to the walker dying on the first move via the transition $n(0) = h + 1$ to $n(g_s) = h$. In this case, the sum in equation (10) only has one term, and hence the PDF of g_s can be approximated as $\mathcal{P}^{g_s}(g_s; h) = \frac{1}{2}(h+1)\mathrm{e}^{-(h+1)g_s} \sim \frac{h+1}{2}$, where the factor 1/2 corresponds to the probability of $T = 1$, and the limit of small g_s has been taken. Thus, for very short times $g_s \ll 1/h$, the PDF of g_s is essentially 'flat'. In order to estimate the transition point to this third regime, we impose again continuity of the solution, so that $(h+1)/2 = g_{\mathrm{XX}}^{-3/2}/\sqrt{2\pi h}$ and hence (dropping the constants) $g_{\mathrm{XX}} = 1/h$, as shown in equation (17) as well as figures 7 and 8.

Given the *three* regimes shown in figure 7, $\mathcal{P}^{g_s}(g_s; h)$ can be collapsed either by ignoring the very short scale, (see equation (9))

$$\mathcal{P}^{g_s}\left(g_s; h\right) \simeq 2g_s^{-2}\mathcal{G}_{>}\left(g_s/h\right) \qquad \text{for} \qquad g \gg 1/h \tag{18}$$

with $\mathcal{G}_{>}(x) = 1$ for large x and $\mathcal{G}_{>}(x) = \sqrt{x/(8\pi)}$ in small x, or according to

$$\mathcal{P}^{g_s}\left(g_s; h\right) \simeq \frac{g_s^{-3/2}}{\sqrt{2\pi h}}\mathcal{G}_{<}\left(g_s h\right) \qquad \text{for} \qquad g \ll 8\pi h \tag{19}$$

with $\mathcal{G}_{<}(x) = 1$ for large x and $\mathcal{G}_{<}(x) = x^{3/2}\sqrt{\pi/2}$ for small x. Power-law scaling (crossover) functions offer a number of challenges, as they affect the 'apparent' scaling exponent [20]. Also, there is no hard cutoff in the present case, i.e. moments $\langle g_s^m \rangle = \int \mathrm{d}g_s \mathcal{P}^{g_s}(g_s; h) g_s^m$ do not exist for $m \geqslant 2$.

4. Summary and discussion

The main goal of the present paper has been to understand how thresholding influences data analysis. In particular, how thresholding can change the scaling of observables and how one might detect this.

To this end, we worked through the consequences of thresholding in the BDP, which is known to have a power-law PDF of survival times with exponent $\gamma = 2$. We have shown, both analytically and via simulations, that the survival times g_s for the thresholded process include a new scaling regime with exponent $\gamma = 3/2$ in the range $1/h \ll g_s \ll 8\pi h$ (see figure 8), where h is the intensity level of the threshold.

We would like to emphasize how difficult it is to observe the asymptotic $\gamma = 2$ exponent, even for such an idealized toy model. For large values of the threshold, $h = 5000$, sample sizes as large as 10^{10} are needed in order to populate the histogram for large survival times. Real-world measurements are unlikely to meet the demand for such vast amounts of data. An illustration of what might then occur for realistic amounts of data that are subject to threshold is given by figure 2, where only the threshold-induced scaling regime associated with exponent $-3/2$ is visible.

Intriguingly, a qualitatively similar scaling phenomenology is observed in renormalized renewal processes with diverging mean interval sizes [21]. The random deletion of points (that, together with a rescaling of time,

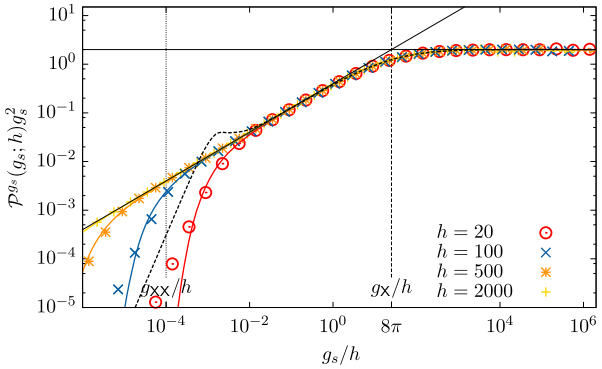

Figure 7. Collapse of the PDFs for different thresholds h for large $g_s \gg 1/h$, plotting $\mathcal{P}^{g_s}(g_s; h)g_s^2$ against g_s/h, according to equations (9) and (18), capturing equation (17). Symbols correspond to simulations at different threshold levels $h = 20,\ 200,\ 500$ and 2000. The black full lines indicate the asymptotes according to equation (17), the dashed lines show the crossovers at $g_s/h = 8\pi$ and $g_s/h = 1/h^2$ for $h = 100$. Finally, the black thick dashed line corresponds to the analytical solution computed from equation (16) for $h = 100$, while coloured full lines were computed by numerically inverting the Laplace transform given in equation (D.4), see appendix D. Another collapse is possible according to equation (19).

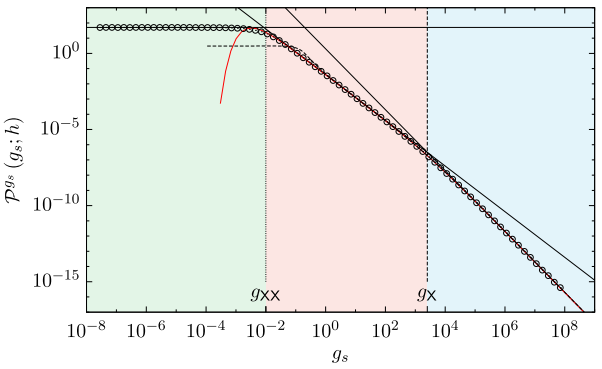

Figure 8. The PDF of survival times $\mathcal{P}^{g_s}(g_s; h)$ for $h = 100$. Three scaling regimes partitioned by g_X (thin dashed line) and g_{XX} (thin dotted line) exist: for very short times $g_s \ll 1/h$ (green shading), the exponential waiting time to the first (death) event dominates, so that $\mathcal{P}^{g_s}(g_s; h) \sim (h+1)/2$. For 'intermediate' times (red shading) $1/h \ll g_s \ll 8\pi h$, the effect of the threshold dominates, and hence $\mathcal{P}^{g_s}(g_s; h) \sim g_s^{-3/2}/\sqrt{2\pi h}$. For long times (blue shading) $g_s \gg 8\pi h$, $\mathcal{P}^{g_s}(g_s; h) \sim 2g_s^{-2}$, independently of h. Monte-Carlo simulation results are shown as symbols, asymptotes of $\mathcal{P}^{g_s}(g_s; h)$, equation (17), as solid lines, and the analytical solution $\mathcal{P}^{g_s}(g_s; h)$, computed via equation (16) as a black thick dashed line, and via numerical inversion of the Laplace transform, equation (D.4), as a red solid line.

constitutes the renormalization procedure) is analogous to the raising of a threshold. It can be shown that the non-trivial fixed point distribution of intervals is bi-power law. The asymptotic scaling regime has the same exponent as that of the original interval sizes. But, in addition, a prior scaling regime emerges with a different exponent, and the crossover separating the two regimes moves out with increasing threshold.

A fundamental difference between theoretical models and the analysis of real-world processes is that in the former, asymptotic exponents are defined in the limit of large events, with everything else dismissed as irrelevant, whereas real world phenomena are usually concerned with finite event sizes. In our example, the effect of the threshold dominates over the 'true' process dynamics in the range $1/h \ll g_s \ll 8\pi h$, and grows with increasing h before eventually taking over the whole region of physical interest.

Of course, real data may not come from an underlying BDP. But we believe that the specific lessons of the BDP apply more generally to processes with multiplicative noise, i.e. a noise whose amplitude changes with the dynamical variable (the degree of freedom). Let us cite two specific examples from the literature to illustrate our point: in [22], Laurson *et al* apply thresholds to Brownian excursion, but since noise is *additive* in Brownian motion, the asymptotic exponent of $-3/2$ is recovered at any threshold level. On the other hand, Larremore *et al* [23] apply thresholds to networks of excitable nodes and critical branching processes, i.e. to processes with *multiplicative* noise, and report strong effects of the threshold on the asymptotic exponents.

Indeed, in a process with multiplicative noise, at large thresholds small changes of the dynamical variable are negligible and an effectively additive process is obtained (the plain random walker in our example). Only for large values of the dynamical variable is the original process recovered. These large values are rare, in particular when another cutoff (such as, effectively, the sample size) limits the effective observation time ($2T - 1$ above). In the worst case, thresholding may therefore bury the asymptotics which would only be recovered for *much* longer observation times. However, if the threshold can easily be changed, its effect can be studied systematically by attempting a data collapse onto the scaling ansatz $\mathcal{P}^{g_s}(g_s; h) = g_s^{-\gamma}\mathcal{G}(g_s/h^D)$, equations (9) and (18), with exponents γ and D to be determined, as performed in figure 7 with $\gamma = 2$ and $D = 1$. The threshold plays an analogous role to the system size in finite-size scaling (albeit for intermediate scales). In the present case, the exponents in the collapse, together with the asymptote of the scaling function, identify two processes at work, namely the BDP as well as the random walk.

Acknowledgments

FFC would like to acknowledge support from projects 2012FI_B 00422 and 2014SGR-1307, from AGAUR; and FIS2012-31324, from the Spanish MINECO.

Appendix A. Power law fitting procedure

We use a fitting procedure valid for both truncated and non-truncated power-law distributions [15, 24]. It is based on maximum likelihood estimation of the exponent, the KS goodness-of-fit test, and Monte Carlo simulations.

A continuous random variable x is power-law distributed if its probability density is given by

$$P(x) = \frac{\gamma - 1}{a^{1-\gamma} - 1/b^{\gamma-1}}\left(\frac{1}{x}\right)^{\gamma}, \tag{A.1}$$

where $a > 0$ and b are the lower and upper ends of the range, respectively. If b is finite, the distribution is truncated, while if $b \to \infty$, the distribution is non-truncated. In the latter case, $\gamma > 1$ is required for a normalizable distribution.

The key to fitting power-law distributions properly to real-world data is to have an objective criterion for deciding when the power law starts (and, in the truncated case, when it ends); this is the fitting range. Given a sample $X_1, X_2, \ldots, X_n$, we would like to estimate the parameter γ and determine the interval $[a, b]$ where the power-law holds. In order to obtain a reliable estimate of the exponent γ, we use maximum likelihood estimation, with a and b fixed. The log-likelihood reads

$$\ell(\gamma) = \ln\frac{\gamma - 1}{1 - r^{\gamma-1}} - \gamma \ln\frac{g}{a} - \ln a, \quad (\gamma \neq 1), \tag{A.2}$$

where $r = a/b$ and g is the geometric mean. The value $\hat{\gamma}$ which maximizes $\ell(\gamma)$ is the maximum likelihood estimator of the exponent.

Having estimated γ, we quantify the goodness-of-fit via a KS test [25]. The KS statistic is the absolute value between the theoretical and empirical cumulative distributions, where the empirical cumulative distribution is given by the fraction of X_i smaller than x, within the interval $[a, b]$.

Using the $\hat{\gamma}$ obtained from the data, we generate surrogate power law samples via Monte Carlo in order to assign a p-value to the KS statistic. Under the null hypothesis, the p-value is the probability that the KS statistic takes a value larger than that obtained empirically. Next, we apply the same procedure for all possible ranges $[a, b]$ and retain those fits (i.e., the triplets $\{a, b, \hat{\gamma}\}$) with p-values greater than p_c. In this analysis we have taken $p_c = 0.5$, which is quite conservative. Under the null hypothesis, the p-value is uniformly distributed such that half of the correct models would be rejected.

Finally, we select one fitting range among all the listed triplets. For non-truncated power laws ($b = \infty$), we select the largest interval, i.e., the smallest a. For truncated power laws, one can either select the interval that maximizes the number of data points contained within, or the size of the log-range b/a, see [15] for a discussion. In this analysis, we have maximized the log-range, which tends to select power laws nearer to the tail of the distribution.

Appendix B. Mean and variance of the survival time

This appendix contains the details of the calculations leading to the approximation (in large T), equations (11) and (13), as well as their asymptotes equations (12) and (14), for the mean $\mu(\Omega)$ and the variance $\sigma^2(\Omega)$

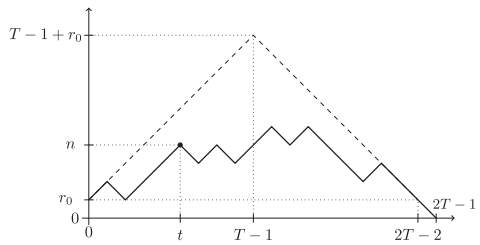

Figure B1. Sample path of a random walk along an absorbing wall at 0. The walker starts at $t=0$ from r_0 and terminates at $2T-1$ by reaching the wall $r_{2T-1}=0$, i.e. $r_{2T-2}=1$. By construction, it cannot escape from the region demarcated by the dashed line. When counting distinct paths, the number of paths terminating at $r_{2T-1}=0$ equals the number of paths passing through $r_{2T-2}=1$.

respectively, averaged over the ensemble $\Omega(T)$, or Ω for short, of the mapped random walks with the constraint that they terminate at $2T-1$, see figure B1.

In the following, we will use the notation ξ_t for $\xi_t(r_t+h)$, but it is important to note that any two $\xi_t(r_t+h)$ are independent, even though the consecutive r_t are not. The random variable $g_s(\mathcal{R})$ in equation (10) is thus a sum of *independent* random variables ξ_t, whose mean and variance at consecutive t, however, are correlated due to r_t being a trajectory of a random walk. Because $h+r_t>0$ for $t<2T-1$, the limiting distribution of $(g_s(\mathcal{R})-\mu(\mathcal{R}))/\sqrt{\sigma^2(\mathcal{R})}$ as $T\to\infty$ is a Gaussian with unit variance. Mean $\mu(\mathcal{R})$ and variance $\sigma^2(\mathcal{R})$ are defined as

$$\mu(\mathcal{R})=\left\langle g_s(\mathcal{R})\right\rangle_{\mathcal{R}}=\sum_{t=0}^{2T-2}\langle\xi_t\rangle_{\mathcal{R}}, \tag{B.1a}$$

$$\begin{aligned}\sigma^2(\mathcal{R})&=\left\langle\left(g_s(\mathcal{R})\right)^2\right\rangle_{\mathcal{R}}-\left\langle g_s(\mathcal{R})\right\rangle_{\mathcal{R}}^2\\&=\sum_{t,t'=0}^{2T-2}\langle\xi_t\xi_{t'}\rangle_{\mathcal{R}}-\langle\xi_t\rangle_{\mathcal{R}}\langle\xi_{t'}\rangle_{\mathcal{R}}\end{aligned} \tag{B.1b}$$

and are functions of the trajectory $\mathcal{R}$ with $\langle\,\cdot\,\rangle_{\mathcal{R}}$ taking the expectation across the ensemble of ξ for given, fixed $\mathcal{R}$, i.e. $\langle\xi_t\rangle_{\mathcal{R}}=1/(r_t+h)$ and $\langle\xi_t^2\rangle_{\mathcal{R}}-\langle\xi_t\rangle_{\mathcal{R}}^2=1/(r_t+h)^2$. Because $\langle\xi_t\xi_{t'}\rangle_{\mathcal{R}}=\langle\xi_t\rangle_{\mathcal{R}}\langle\xi_{t'}\rangle_{\mathcal{R}}$ for $t\neq t'$ the mean and the variance are in fact just

$$\mu(\mathcal{R})=\sum_{t=0}^{2T-2}\frac{1}{r_t+h}, \tag{B.2a}$$

$$\sigma^2(\mathcal{R})=\sum_{t=0}^{2T-2}\frac{1}{\left(r_t+h\right)^2}. \tag{B.2b}$$

If $\rho_n(\mathcal{R})$ *counts* the number of times r_t attains a certain level

$$\rho_n(\mathcal{R})=\sum_{t=0}^{2T-2}\delta_{n,r_t} \tag{B.3}$$

then $\sum_{t=0}^{2T-2}f(r_t)=\sum_{t=0}^{2T-2}\sum_{n=0}^{\infty}\delta_{n,r_t}f(n)=\sum_{n=0}^{\infty}\rho_n(\mathcal{R})f(n)$, so

$$\mu(\mathcal{R})=\sum_{n=r_0}^{T-1+r_0}\frac{\rho_n(\mathcal{R})}{n+h}, \tag{B.4a}$$

$$\sigma^2(\mathcal{R})=\sum_{n=r_0}^{T-1+r_0}\frac{\rho_n(\mathcal{R})}{(n+h)^2}. \tag{B.4b}$$

where we used the fact that within time $2T-2$ our random walker cannot stray further away from r_0 than $T-1+r_0$, as illustrated in figure B1.

In the same vein, we can now proceed to find mean and variance of g_s over the entire ensemble $\Omega=\Omega(T)$ of trajectories $\mathcal{R}$ that terminate at $2T-1$. In the following $\langle\,\cdot\,\rangle_{\Omega}$ denotes the ensemble average over all trajectories $\mathcal{R}\in\Omega$, each appearing with the same probability

$$\left\langle f\left(\xi_t\right)\right\rangle_{\Omega} = \frac{1}{|\Omega|}\sum_{\mathcal{R}}\left\langle f\left(\xi_t\right)\right\rangle_{\mathcal{R}}, \tag{B.5}$$

where $f(\xi_t)$ is an arbitrary function of the random variable ξ_t. We therefore have

$$\begin{aligned}\mu(\Omega) &= \left\langle \sum_{t=0}^{2T-2}\xi_t \right\rangle_{\Omega} = \frac{1}{|\Omega|}\sum_{\mathcal{R}}\sum_{t=0}^{2T-2}\frac{1}{r_t+h} \\ &= \frac{1}{|\Omega|}\sum_{\mathcal{R}}\sum_{n=r_0}^{T-1+r_0}\frac{\rho_n(\mathcal{R})}{n+h} = \sum_{n=r_0}^{T-1+r_0}\frac{\left\langle\rho_n(\mathcal{R})\right\rangle_{\Omega}}{n+h},\end{aligned} \tag{B.6}$$

where $\langle\rho_n(\mathcal{R})\rangle_{\Omega}$ is in fact the expected number of times a random walker terminating at $2T-1$ attains level n.

The variance turns out to require a bit more work. The second moment

$$\left\langle g_s(\mathcal{R})^2\right\rangle_{\Omega} = \left\langle\left(\sum_{t=0}^{2T-2}\xi_t\right)^2\right\rangle_{\Omega} = \frac{1}{|\Omega|}\sum_{\mathcal{R}}\sum_{t,t'=0}^{2T-2}\langle\xi_t\xi_{t'}\rangle_{\mathcal{R}} \tag{B.7}$$

simplifies significantly when $t \neq t'$ in which case the lack of correlations means that the expectation factorizes $\langle\xi_t\xi_{t'}\rangle_{\mathcal{R}} = \langle\xi_t\rangle_{\mathcal{R}}\langle\xi_{t'}\rangle_{\mathcal{R}}$, so that we can write

$$\sum_{t,t'=0}^{2T-2}\langle\xi_t\xi_{t'}\rangle_{\mathcal{R}} = \sum_{t,t'=0}^{2T-2}\langle\xi_t\rangle_{\mathcal{R}}\langle\xi_{t'}\rangle_{\mathcal{R}} + \sum_{t=0}^{2T-2}\left(\langle\xi_t^2\rangle_{\mathcal{R}} - \langle\xi_t\rangle_{\mathcal{R}}^2\right). \tag{B.8}$$

Obviously $\sum_{t,t'=0}^{2T-2}\langle\xi_t\rangle_{\mathcal{R}}\langle\xi_{t'}\rangle_{\mathcal{R}} = \left(\sum_{t=0}^{2T-2}\langle\xi_t\rangle_{\mathcal{R}}\right)^2$, but that is not a useful simplification for the time being.

The square of the first moment, equation (B.6), is best written as

$$\left\langle g_s(\mathcal{R})\right\rangle_{\Omega}^2 = \frac{1}{|\Omega|^2}\sum_{\mathcal{R},\mathcal{R}'}\sum_{t,t'=0}^{2T-2}\langle\xi_t\rangle_{\mathcal{R}}\langle\xi_{t'}\rangle_{\mathcal{R}'} \tag{B.9}$$

so that

$$\begin{aligned}\sigma^2(\Omega) &= \left\langle g_s(\mathcal{R})^2\right\rangle_{\Omega} - \left\langle g_s(\mathcal{R})\right\rangle_{\Omega}^2 \\ &= \frac{1}{|\Omega|}\sum_{\mathcal{R}}\sum_{t,t'=0}^{2T-2}\langle\xi_t\rangle_{\mathcal{R}}\langle\xi_{t'}\rangle_{\mathcal{R}} + \frac{1}{|\Omega|}\sum_{\mathcal{R}}\sum_{t=0}^{2T-2}\left(\langle\xi_t^2\rangle_{\mathcal{R}} - \langle\xi_t\rangle_{\mathcal{R}}^2\right) \\ &\quad - \frac{1}{|\Omega|^2}\sum_{\mathcal{R},\mathcal{R}'}\sum_{t,t'=0}^{2T-2}\langle\xi_t\rangle_{\mathcal{R}}\langle\xi_{t'}\rangle_{\mathcal{R}'}.\end{aligned} \tag{B.10}$$

The first and the last pair of sums can be written as

$$\frac{1}{|\Omega|^2}\sum_{\mathcal{R},\mathcal{R}'}\sum_{t,t'=0}^{2T-2}\langle\xi_t\rangle_{\mathcal{R}}\left(\langle\xi_{t'}\rangle_{\mathcal{R}} - \langle\xi_{t'}\rangle_{\mathcal{R}'}\right) \tag{B.11}$$

using $\sum_{\mathcal{R}}(1/|\Omega|) = 1$, so that

$$\begin{aligned}\sigma^2(\Omega) &= \frac{1}{|\Omega|^2}\sum_{\mathcal{R},\mathcal{R}'}\sum_{t,t'=0}^{2T-2}\langle\xi_t\rangle_{\mathcal{R}}\left(\langle\xi_{t'}\rangle_{\mathcal{R}} - \langle\xi_{t'}\rangle_{\mathcal{R}'}\right) \\ &\quad + \frac{1}{|\Omega|}\sum_{\mathcal{R}}\sum_{t=0}^{2T-2}\left(\langle\xi_t^2\rangle_{\mathcal{R}} - \langle\xi_t\rangle_{\mathcal{R}}^2\right).\end{aligned} \tag{B.12}$$

In the first sum, the two terms can be separated into those in t' and one in t. Using the same notation as above, equation (B.3) we have

$$\sum_{t'=0}^{2T-2}\left(\langle\xi_{t'}\rangle_{\mathcal{R}} - \langle\xi_{t'}\rangle_{\mathcal{R}'}\right) = \sum_{n'=r_0}^{T-1+r_0}\frac{\rho_{n'}(\mathcal{R}) - \rho_{n'}\left(\mathcal{R}'\right)}{n'+h} \tag{B.13}$$

and $\sum_{t=0}^{2T-2}\langle\xi_t\rangle_{\mathcal{R}} = \sum_{n=r_0}^{T-1+r_0}\frac{\rho_n(\mathcal{R})}{n+h}$.

The second sum recovers the earlier result in equation (B.4*b*), as $\langle \xi_t^2 \rangle_{\mathcal{R}} = \frac{2}{(r_t+h)^2}$ and $\langle \xi_t \rangle_{\mathcal{R}} = \frac{1}{r_t+h}$, so that

$$\sum_{t=0}^{2T-2} \left(\langle \xi_t^2 \rangle_{\mathcal{R}} - \langle \xi_t \rangle_{\mathcal{R}}^2 \right) = \sum_{n=r_0}^{T-1+r_0} \frac{\rho_n(\mathcal{R})}{(n+h)^2} \tag{B.14}$$

and therefore

$$\begin{aligned}
\sigma^2(\Omega) &= \frac{1}{|\Omega|^2} \sum_{\mathcal{R},\mathcal{R}'} \sum_{n,n'=r_0}^{T-1+r_0} \frac{\rho_n(\mathcal{R})}{n+h} \frac{\rho_{n'}(\mathcal{R}) - \rho_{n'}(\mathcal{R}')}{n'+h} \\
&\quad + \frac{1}{|\Omega|} \sum_{\mathcal{R}} \sum_{n=r_0}^{T-1+r_0} \frac{\rho_n(\mathcal{R})}{(n+h)^2} \\
&= \frac{1}{|\Omega|} \sum_{\mathcal{R}} \sum_{n,n'=r_0}^{T-1+r_0} \frac{\rho_n(\mathcal{R})\rho_{n'}(\mathcal{R})}{(n+h)(n'+h)} - \left(\frac{1}{|\Omega|} \sum_{\mathcal{R}} \sum_{n=r_0}^{T-1+r_0} \frac{\rho_n(\mathcal{R})}{n+h} \right)^2 \\
&\quad + \frac{1}{|\Omega|} \sum_{\mathcal{R}} \sum_{n=r_0}^{T-1+r_0} \frac{\rho_n(\mathcal{R})}{(n+h)^2} \\
&= \sum_{n,n'=r_0}^{T-1+r_0} \frac{\langle \rho_n(\mathcal{R})\rho_{n'}(\mathcal{R}) \rangle_\Omega}{(n+h)(n'+h)} - \left(\sum_{n=r_0}^{T-1+r_0} \frac{\langle \rho_n(\mathcal{R}) \rangle_\Omega}{n+h} \right)^2 + \sum_{n=r_0}^{T-1+r_0} \frac{\langle \rho_n(\mathcal{R}) \rangle_\Omega}{(n+h)^2}.
\end{aligned} \tag{B.15}$$

We now have the mean $\mu(\Omega)$, equation (B.6), and the variance $\sigma^2(\Omega)$, equation (B.15), in terms of $\langle \rho_n(\mathcal{R}) \rangle_\Omega$ and $\langle \rho_n(\mathcal{R})\rho_{n'}(\mathcal{R}) \rangle_\Omega$. In the following, we will determine these two quantities and then return to the original task of finding a closed-form expression for $\mu(\Omega)$ and $\sigma^2(\Omega)$.

B.1. $\langle \rho_n(\mathcal{R}) \rangle_\Omega$ and $\langle \rho_n(\mathcal{R})\rho_{n'}(\mathcal{R}) \rangle_\Omega$

Of the two expectations, $\langle \rho_n(\mathcal{R}) \rangle_\Omega$ is obviously the easier one to determine. In fact, $\sum_n \rho_n(\mathcal{R}) = 2T-1$ implies $\sum_{n'} \langle \rho_n(\mathcal{R})\rho_{n'}(\mathcal{R}) \rangle = (2T-1)\langle \rho_n(\mathcal{R}) \rangle$, i.e. $\langle \rho_n(\mathcal{R}) \rangle_\Omega$ is a 'marginal' of $\langle \rho_n(\mathcal{R})\rho_{n'}(\mathcal{R}) \rangle_\Omega$.

To determine $\langle \rho_n(\mathcal{R}) \rangle_\Omega$, we use the method of images (or mirror charges). The number of positive paths $(r_i > 0)$ from $(t=0,\ r_0)$ to (t,n) are $\binom{t}{\frac{n-r_0+t}{2}} - \binom{t}{\frac{n+r_0+t}{2}}$ for $n+r_0+t$ even and $n>0$. By construction, the number of paths passing through $n=0$ is exactly 0, thereby implementing the boundary condition. The set of paths (to be considered in the following) which terminate at time $2T-1$ by reaching $r_{2T-1}=0$ is, up to the final step, identical to the set of paths passing through $(2T-2,\ 1)$, i.e. $r_{2T-1}=0$. The number of positive paths (see figure B1) originating from $(0,\ r_0=1)$ and terminating at $(t=2T-1,\ r_{2T-1}=0)$ therefore equals the number of positive paths from $(0,\ r_0=1)$ to $(t=2T-2,\ n=1)$, so that $|\Omega| = \binom{2T-2}{T-1} - \binom{2T-2}{T} = \frac{1}{T}\binom{2T-2}{T-1}$, which are the Catalan numbers [26, 27]. For $r_0=1$ we also have

$$\binom{t}{\frac{n-1+t}{2}} - \binom{t}{\frac{n+1+t}{2}} = \frac{n}{t+1} \binom{t+1}{\frac{n+1+t}{2}} \tag{B.16}$$

again for $n+r_0+t$ even. This is the number of positive paths from $(0,\ 1)$ to (t,n) and by symmetry also the number of paths from $(2T-2-t,\ n)$ to $(2T-2,\ 1)$, given that the walk is unbiased (see figure B1). If $\langle \rho_n(t;\mathcal{R}) \rangle_\Omega$ is the expected fraction of paths passing through (t,n) (illustrated in figure B1), we therefore have

$$\left\langle \rho_n\left(t;\mathcal{R}\right) \right\rangle_\Omega = \underbrace{\frac{T}{\binom{2T-2}{T-1}}}_{1/|\Omega|} \underbrace{\tfrac{n}{t+1}\binom{t+1}{\frac{n+1+t}{2}}}_{\text{from(0,1) to (t, n)}} \underbrace{\tfrac{n}{2T-1-t}\binom{2T-1-t}{\frac{n+2T-1-t}{2}}}_{\text{from (t, n) to (2T−2, 1)}} \tag{B.17}$$

which is normalized by construction, i.e. $\sum_n \langle \rho_n(t;\mathcal{R}) \rangle_\Omega = 1$. The first binomial factor in the denominator is due to the normalization, whereas of the last two, the first is due to paths from $(0,\ 1)$ to (t,n) and the second due to paths from (t,n) to $(2T-2-t,\ 1)$. In the following we are interested in the fraction of times a random

walker reaches a certain level during its lifetime, $\langle \rho_n(\mathcal{R})\rangle_\Omega = \sum_t \langle \rho_n(t;\mathcal{R})\rangle_\Omega$. Using $\binom{a}{b} \simeq 2^a (a\pi/2)^{-1/2} \exp\left(-\frac{2}{a}\left(b-\frac{a}{2}\right)^2\right)$ we find

$$\left\langle \rho_n\left(t;\mathcal{R}\right)\right\rangle_\Omega \simeq \frac{8\,T^{3/2}}{\sqrt{\pi}} \frac{n^2}{\tilde{t}^{3/2}\left(2T-\tilde{t}\right)^{3/2}} \exp\left(-\frac{n^2}{2\tilde{t}} - \frac{n^2}{2\left(2T-\tilde{t}\right)}\right), \tag{B.18}$$

where we have used $T \gg 1$ and $\tilde{t} = t+1$. Simplifying further gives

$$\left\langle \rho_n(\mathcal{R})\right\rangle_\Omega = \sum_{\tilde{t}=n}^{2T-n} \left\langle \rho_n\left(t;\mathcal{R}\right)\right\rangle_\Omega \simeq 8\nu^2\sqrt{\frac{T}{\pi}} \sum_{\tilde{t}=n}^{2T-n} \frac{\exp\left(-\frac{\nu^2}{\tau(2-\tau)}\right)}{T\left(\tau(2-\tau)^{3/2}\right)} \tag{B.19}$$

with the sum running over the $\tilde{t}$ with the correct parity and $\tau = \tilde{t}/T$ and $\nu = n/\sqrt{T}$. In the limit of large $T \gg 1$ we find [28]

$$\lim_{T\to\infty} \frac{\left\langle \rho_n(\mathcal{R})\right\rangle_\Omega}{\sqrt{T}} = \frac{4\nu^2}{\sqrt{\pi}} \int_0^2 \mathrm{d}\tau \frac{\exp\left(-\frac{\nu^2}{\tau(2-\tau)}\right)}{(\tau(2-\tau))^{3/2}} = 4\nu \mathrm{e}^{-\nu^2}, \tag{B.20}$$

where the parity has been accounted for by dividing by 2. In the last step, the integral was performed by some substitutions, as $\tau(2-\tau)$ is symmetric about 1. It follows that in the limit of large $T \gg 1$

$$\left\langle \rho_n(\mathcal{R})\right\rangle_\Omega \simeq 4n \exp\left(-\frac{n^2}{T}\right). \tag{B.21}$$

Using that expression in equation (B.6) gives equation (11), namely

$$\begin{aligned} \frac{\mu(\Omega)}{\sqrt{T}} &\simeq 4 \sum_{n=r_0}^{T-1+r_0} \frac{1}{\sqrt{T}} \frac{\nu}{\nu + \frac{h}{\sqrt{T}}} \mathrm{e}^{-\nu^2} \\ &\simeq \int_0^{\sqrt{T}} \mathrm{d}\nu \frac{4\nu}{\nu+\frac{h}{\sqrt{T}}} \mathrm{e}^{-\nu^2} \simeq \int_0^\infty \mathrm{d}\nu \frac{4\nu}{\nu+\frac{h}{\sqrt{T}}} \mathrm{e}^{-\nu^2} = 2\sqrt{\pi} + 2\frac{h}{\sqrt{T}} \psi\left(\frac{h}{\sqrt{T}}\right) \end{aligned} \tag{B.22}$$

with [29, equation 27.6.3]

$$\psi(x) = -\mathrm{e}^{-x^2}\left(2\sqrt{\pi}\int_0^x \mathrm{d}s \mathrm{e}^{s^2} + \fint_{-x^2}^\infty \mathrm{d}y \frac{\mathrm{e}^{-y}}{y}\right), \tag{B.23}$$

where we have used $r_0 = 1$. The second integral is known as the exponential integral function $\fint_{-x}^\infty \mathrm{d}y \frac{\mathrm{e}^{-y}}{y} = -\mathrm{Ei}(x)$ and the first as (a multiple of) the imaginary error function $2\sqrt{\pi}\int_0^x \mathrm{d}s \mathrm{e}^{s^2} = \pi\mathcal{E}(\imath x)/\imath$. In the limit of large arguments x, the function $\psi(x)$ is $-\sqrt{\pi}/x + 1/x^2 - \sqrt{\pi}/(2x^3) + 1/x^4 + \mathcal{O}(x^{-5})$, in the limit of small arguments by $\gamma + 2\ln(x)$, where γ is the Euler-Mascheroni constant. We conclude that

$$\mu(\Omega) \simeq \begin{cases} 2\sqrt{\pi T} + 2h(\gamma + 2\ln(h/\sqrt{T})) & \text{for } T \gg h^2 \\ 2T/h - \sqrt{\pi}T^{3/2}/h^2 + 2T^2/h^3 & \text{for } T \ll h^2 \end{cases} \tag{B.24}$$

(see equation (12)) provided T is large compared to 1, which is the key assumption of the approximations used above. It is worth stressing this distinction: T has to be large compared to 1 in order to make the various continuum approximations (effectively continuous in time, so sums turn into integrals and continuous in state, so binomials can be approximated by Gaussians), but no restrictions were made regarding the ratio T/h^2.

The correlation function $\langle \rho_n(\mathcal{R})\rho_{n'}(\mathcal{R})\rangle_\Omega$ can be determined using the same methods, starting with equation (B.17):

$$
\begin{aligned}
&\left\langle \rho_n\left(t;\mathcal{R}\right)\rho_{n'}\left(t';\mathcal{R}\right)\right\rangle_{\Omega}\\
&=\sum_{t}\sum_{t'<t}\underbrace{\frac{T}{\binom{2T-2}{T-1}}}_{1/|\Omega|}\underbrace{\frac{n}{t'+1}\binom{t'+1}{\frac{n+t'+1}{2}}}_{\text{from }(0,1)\text{ to }(t',n)}\\
&\times\underbrace{\left[\binom{t-t'}{\frac{t-t'+n-n'}{2}}-\binom{t-t'}{\frac{t-t'+n+n'}{2}}\right]}_{\text{from }(t',n)\text{ to }(t,n')}\\
&\times\underbrace{\frac{n'}{2T-1-t}\binom{2T-1-t}{\frac{n'+2T-1-t}{2}}}_{\text{from}(t,n')\text{ to}(2T-2,1)}\\
&+\sum_{t}\sum_{t'\geqslant t}\underbrace{\frac{T}{\binom{2T-2}{T-1}}}_{1/|\Omega|}\underbrace{\frac{n}{t+1}\binom{t+1}{\frac{n+t+1}{2}}}_{\text{from }(0,1)\text{ to }(t,n)}\\
&\times\underbrace{\left[\binom{t'-t}{\frac{t'-t+n-n'}{2}}-\binom{t'-t}{\frac{t'-t+n+n'}{2}}\right]}_{\text{from }(t,n)\text{ to }(t',n')}\\
&\times\underbrace{\frac{n'}{2T-1-t}\binom{2T-1-t}{\frac{n'+2T-1-t'}{2}}}_{\text{from}(t',n')\text{ to}(2T-2,1)}.
\end{aligned}
\tag{B.25}
$$

Because both t and t' are dummy variables, one might be tempted to write the entire expression as twice the first double sum, which is indeed correct as long as $n \neq n'$. In that case, the case $t' = t$ does not contribute because the 'middle chunk' (from (t, n) to (t', n')) vanishes. However, if $n = n'$ that middle chunk is unity and therefore needs to be included separately. This precaution turns out to be unnecessary once the binomials are approximated by Gaussians and the sums by integrals.

The resulting convolutions are technically tedious, but can be determined in closed form on the basis of Laplace transforms and tables [29, equations 29.3.82 and 29.3.84], resulting finally in

$$
\left\langle \rho_n(\mathcal{R})\rho_{n'}(\mathcal{R})\right\rangle_{\Omega} \simeq 8\,T\left(\mathrm{e}^{-n^2/T}-\mathrm{e}^{-(n+n')^2/T}\right) \tag{B.26}
$$

to leading order in T.

We proceed to determine equation (B.15) using equations (B.21) and (B.26) in the limit of large T. Again, we interpret the sums as Riemann sums, to be approximated by integrals, resulting in equation (13),

$$
\sigma^2(\Omega) \simeq T\ \mathcal{I}(x) - \mu(\Omega)^2 + \mathcal{K}(x) \tag{B.27}
$$

with $x = h/\sqrt{T}$ and

$$
\mathcal{K}(x) = \int_0^{\infty} \mathrm{d}n \frac{4n\mathrm{e}^{-n^2}}{(n+x)^2} = -4 + 4x\sqrt{\pi} + 2(2x^2-1)\psi(x), \tag{B.28a}
$$

$$
\mathcal{I}(x) = 16\int_0^{\infty}\mathrm{d}n\int_0^{n}\mathrm{d}n'\frac{\mathrm{e}^{-n^2}-\mathrm{e}^{-(n+n')^2}}{(n+x)(n'+x)} \tag{B.28b}
$$

(for the definition of $\psi(x)$ see equation (B.23)). Unfortunately, we were not able to reduce $\mathcal{I}(x)$ further.

Because of the structure of equation (B.27), where $T\ \mathcal{I}(x)r - \mu(\Omega)^2$ scale linearly in T at fixed $x = h/\sqrt{T}$, whereas $\mathcal{K}(x)$ remains constant, a statement about the leading order behaviour in T is no longer equivalent to a

statement about the leading order behaviour in $1/x^2$. This is complicated further by the assumption made throughout that T is large. The limits we are interested in, are in fact $T \gg h^2$ with $T \gg 1$ and $1 \ll T \ll h^2$. In the following, we need to distinguish not only large x from small x, but also different orders of T.

It is straightforward to determine the asymptote of $\mathcal{I}(x)$ in large x, where the denominator of the integrand is dominated by x^2 while the numerator vanishes at least as fast as e^{-n^2}, because $\mathrm{e}^{-n^2} - \mathrm{e}^{-(n+n')^2} = \mathrm{e}^{-n^2}(1 - \mathrm{e}^{-2nn'-n'^2})$ and $0 \leqslant (1 - \mathrm{e}^{-2nn'-n'^2}) < 1$, so [28]

$$\begin{aligned}\mathcal{I}(x) = \frac{16}{x^2}\int_0^\infty \mathrm{d}n \int_0^n \mathrm{d}n' \Bigg[& \mathrm{e}^{-n^2}(1 - \mathrm{e}^{-2nn'-n'^2}) \\ & \times \left(1 - \frac{n}{x} + \frac{n^2}{x^2} + \cdots\right)\left(1 - \frac{n'}{x} + \frac{n'^2}{x^2} + \cdots\right)\Bigg] \\ = \frac{4}{x^2} - \frac{4\sqrt{\pi}}{x^3} + \frac{34}{3x^4} + \mathcal{O}\left(x^{-5}\right).\end{aligned} \tag{B.29}$$

Similarly, or using the expansion of $\psi(x)$ introduced above, we find $\mathcal{K}(x) = 2/x^2 + \mathcal{O}(x^{-3})$. Since $\mu(\Omega) = T(2/x - \sqrt{\pi}/x^2 + 2/x^3 + \cdots)$, the first two terms in the expansion of $\mathcal{I}(x)$ for large x cancel, and we arrive at

$$\begin{aligned}\sigma^2(\Omega) &= \frac{2}{x^2} + \mathcal{O}\left(x^3\right) + T\left(\frac{34}{3x^4} - \frac{8+\pi}{x^4} + \mathcal{O}\left(x^5\right)\right) \\ &= \frac{2T}{h^2} + \frac{10 - 3\pi}{h^4}T^3 + \cdots\end{aligned} \tag{B.30}$$

for $T \ll h^2$, containing the rather unusual looking ('barely positive', one might say) difference $10 - 3\pi$. The second term in equation (B.30) is clearly subleading in large x and no ambiguity arises in that limit, not even if $T \gg 1$.

The limit $h/\sqrt{T} = x \to 0$, on the other hand, $\mathcal{I}(x)$ is

$$\mathcal{I}(x) = \frac{4}{3}\pi^2 + \mathcal{O}(x) \tag{B.31}$$

using [29, equation 27.7.6] so that $T\mathcal{I}(x) - \mu(\Omega)^2 = T(4\pi^2/3 - 4\pi + \mathcal{O}(x))$, whereas $\mathcal{K}(x) = -4\ln(x) - 4 - 2\gamma$ *diverges* in small x. Although this latter term therefore dominates in small x, the former, $T\mathcal{I}(x) - \mu(\Omega)^2$, does for large $T \gg h^2$ at finite, fixed h.

We are now in the position to determine the relevant asymptotes of $\sigma^2(\Omega)$, as stated in (14),

$$\sigma^2(\Omega) \simeq \begin{cases} 4\pi T\dfrac{\pi - 3}{3} & \text{for } T \gg h^2, \\ \dfrac{2T}{h^2} & \text{for } T \ll h^2. \end{cases} \tag{B.32}$$

Appendix C. Limiting distribution of $g_s(\Omega)/\sqrt{T}$

In this second appendix, we explicitly find the limiting distribution of $g_s(\Omega)/\sqrt{T}$. We begin by noting that, for $T \gg 1$, $g_s(\Omega)$ can be approximated as $g_s(\Omega) \simeq \int_0^{2T} \mathrm{d}t \frac{1}{x(t)+h}$, where $x(t)$ performs a Brownian excursion of length $2T$. While for large but finite T this is clearly an approximation (e.g. the exponential random variables have been replaced by their mean), in the limit of $T \to \infty$ the approximation becomes exact. In particular, the 'noise' due to the variance of the exponential random variables scales like $\log T$, see equation (B.28*a*), and thus vanishes after rescaling with respect to $\sqrt{T}$. In addition, owing to the scaling properties of Brownian motion,

$$\lim_{T\to\infty} g_s(\Omega)/\sqrt{T} = \lim_{T\to\infty}\int_0^2 \mathrm{d}t \frac{1}{x(t) + h/\sqrt{T}} = \int_0^2 \mathrm{d}t \frac{1}{x(t)}, \tag{C.1}$$

where $x(t)$ is a Brownian excursion of length 2. Functionals of this kind have recently been discussed in detail in [30]. To find the distribution of this quantity, we first define $y(t) = \int_0^t \mathrm{d}t' 1/x(t')$, and the propagator

$Z(x, y, x_0, y_0, t)$, i.e. the probability for a Brownian particle to go from (x_0, y_0) to (x, y) in time t, without touching the line $x = 0$. Using standard techniques [31], the associated Fokker–Plank equation for the propagator takes the form

$$\left[\partial_t + \frac{1}{x}\partial_y - \frac{1}{2}\partial_{xx}\right] Z(x, y, x_0, y_0, t) = 0, \tag{C.2}$$

with initial condition

$$Z\left(x, y, x_0, y_0, 0\right) = \delta\left(x - x_0\right)\delta\left(y - y_0\right), \tag{C.3}$$

and boundary condition

$$Z\left(0, y, x_0, y_0, t\right) = 0. \tag{C.4}$$

Taking the Laplace transform with respect to t yields

$$\left[s + \frac{1}{x}\partial_y - \frac{1}{2}\partial_{xx}\right] \hat{Z}\ (x, y, x_0, s) = \delta(x - x_0)\delta(y), \tag{C.5}$$

$$\hat{Z}\left(0, y, x_0, s\right) = 0. \tag{C.6}$$

We first solve the associated homogeneous equation, from which we will be able to construct the solution to the inhomogeneous problem. After substituting the ansatz $\hat{Z}_{\mathrm{hom}}(x, y, s) = \Psi(x, s)\rho(y, s)$, the equation separates into

$$-1/2\partial_{xx}\Psi(x, s) + (s - \lambda/x)\Psi(x, s) = 0, \tag{C.7}$$

$$-\partial_y\rho\left(y, s\right) + \lambda\rho\left(y, s\right) = 0, \tag{C.8}$$

where λ is an arbitrary real constant. Equation (C.7) is an eigenvalue problem for $\Psi(x, s)$ with respect to the weight $1/x$. The solutions that vanish at infinity take the form $\Psi_\lambda(x, s) \propto \mathrm{e}^{-\sqrt{2s}x}U\left(-\lambda/\sqrt{2s}, 0, 2\sqrt{2s}x\right)$, but only for $\lambda_k = \sqrt{2s}k$, $k = \{1, 2, ...\}$ do they vanish at $x = 0$. The correctly normalized eigenfunctions that satisfy boundary conditions are therefore

$$\Psi_k(x, s) = \frac{\mathrm{e}^{-\sqrt{2s}x}U\left(-k, 0, 2\sqrt{2s}x\right)}{\sqrt{k!\,(k-1)!}}. \tag{C.9}$$

These functions are an orthonormal set with respect to the weight $1/x$, $\int_0^\infty \mathrm{d}x\Psi_j(x, s)\Psi_k(x, s)\frac{1}{x} = \delta_{j,k}$, and the corresponding closure relation reads $\sum_{k=1}^\infty \Psi_k(x, s)\Psi_k(x', s)\frac{1}{x} = \delta(x - x')$. One can use this to construct the solution of the original equation. In particular

$$\hat{Z}(x, y, x_0, s) = \Theta(y)\sum_{k=1}^{\infty}\Psi_k(x, s)\Psi_k(x_0, s)\mathrm{e}^{-\sqrt{2s}ky}. \tag{C.10}$$

We now return to the original problem of finding the probability of a Brownian excursion with functional $\int_0^t 1/x(t')\mathrm{d}t' = y(t)$. We make use of the device $x_0 = x = \epsilon$, and let $\epsilon \to 0$ only after normalization. In short

$$\lim_{T\to\infty}\mathrm{Prob}\left(g_s(\Omega)/\sqrt{T} = y\right) = \lim_{\epsilon\to 0}\frac{Z(\epsilon, \epsilon, y, t)}{Z_\epsilon}\bigg|_{t=2}, \tag{C.11}$$

where $Z_\epsilon = \frac{1}{\sqrt{2\pi t}}(1 - \mathrm{e}^{-2\epsilon^2/t})$ is the well-known normalizing constant (see e.g. [18]). From (C.10) and expanding for small $x = x_0 = \epsilon$ term by term, we find

$$\frac{\hat{Z}\left(\epsilon, \epsilon, y, s\right)}{Z_\epsilon} \simeq \sqrt{2\pi t}\sum_{k=1}^{\infty}\frac{\Psi_k(\epsilon, s)^2}{\left(1 - \mathrm{e}^{-2\epsilon^2/t}\right)}\mathrm{e}^{-\sqrt{2s}ky}. \tag{C.12}$$

Using the fact that $\Psi_k(\epsilon, s)^2 \simeq 8sk\epsilon^2$ for small ϵ, we finally arrive at

$$\begin{aligned}\lim_{\epsilon\to 0}\frac{\hat{Z}\left(\epsilon, \epsilon, y, s\right)}{Z_\epsilon} &\simeq \lim_{\epsilon\to 0}\sqrt{2\pi t}\sum_{k=1}^{\infty}\frac{8sk\epsilon^2}{2\epsilon^2/t}\mathrm{e}^{-\sqrt{2s}ky} \\ &= 4s\sqrt{2\pi}\,t^{3/2}\frac{\mathrm{e}^{\sqrt{2s}y}}{\left(\mathrm{e}^{\sqrt{2s}y} - 1\right)^2} = \sqrt{2\pi}\,t^{3/2}\frac{s}{\sinh^2\left(\sqrt{s/2}\,y\right)}.\end{aligned} \tag{C.13}$$

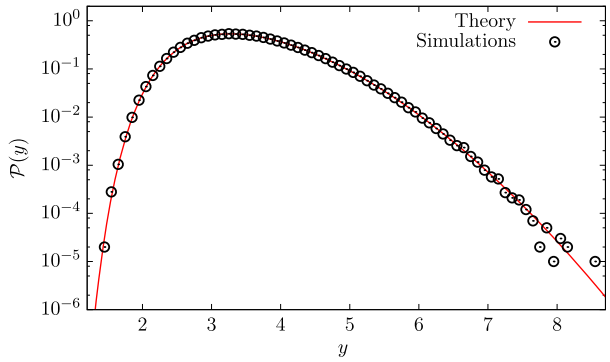

Figure C1. The distribution of $y = \int_0^2 \mathrm{d}t\ 1/x\,(t)$, where $x(t)$ is a Brownian excursion of length 2. The red full line is the analytical result and black symbols correspond to simulations.

Inverting terms involving s yields

$$\lim_{T\to\infty} \mathrm{Prob}\left(g_s(\Omega)/\sqrt{T} = y\right) = \left[\frac{2\sqrt{2\pi}\,t^{3/2}\,\pi^2}{y^6}\sum_{k=1}^{\infty}(2k)^2 \mathrm{e}^{-(2\pi k)^2 t/\left(2y^2\right)}\left((2\pi k)^2 t - 3y^2\right)\right]_{t=2} \tag{C.14}$$

$$= \left[\frac{2y}{t^2}\sum_{k=1}^{\infty}\mathrm{e}^{-(ky)^2/(2t)}k^2\left(k^2y^2 - 3t\right)\right]_{t=2}. \tag{C.15}$$

The first equation is obtained by collecting residues from double poles, and is useful for a small y expansion. The second equation is obtained by expanding (C.13) and inverting term by term, and is useful for a large y expansion. Both expressions converge rapidly and, evaluating at $t = 2$, are in excellent agreement with simulations, see figure C1.

Appendix D. Laplace transform of $\mathcal{P}^{g_s}(g_s,\ h)$

In this final appendix, we take yet another route in the calculation of $\mathcal{P}^{g_s}(g_s,\ h)$ by finding its Laplace transform. The key point in this approach is to approximate the embedded random walk of the process by standard Brownian motion. Therefore, we expect our approximation to hold as long as $T \gg 1$. The approach is very similar in spirit to that of appendix C, but both Appendices are self-contained and can be read independently.

Let $x(t)$ denote the trajectory of a Brownian particle starting at $x\,(0) = x_0$, and t_f its first passage time to 0. Then we argue that, in the Brownian motion picture, the original observable of interest of the process g_s corresponds to the quantity $\mathcal{G}_h$,

$$\mathcal{G}_h = \int_0^{t_f} \mathrm{d}t U_h(x\,(t)), \tag{D.1}$$

with $U_h(x) = 1/(x + h)$. Effectively, the underlying exponential random variables $\xi\,(x\,(t))$ are replaced by their average. Such an approximation, which can be seen as a self-averaging property of the process, is well-justified because (i) the Brownian particle visits any state infinitely many times, and (ii) the exponential distribution has finite moments of any order. We are hence left with computing the distribution of the integral of a function $U_h(x)$ along a Brownian trajectory starting at $x\,(0) = x_0$ and ending at $x\,(t_f) = 0$. As usual, the problem is most conveniently solved by taking the Laplace transform of $\mathcal{G}_h$ (see the excellent review by Majumdar, [18]). In particular, the Laplace transform of $\mathcal{P}\,(\mathcal{G}_h)$, which we denote by $\hat{\mathcal{P}}\,(u;\ h,\ x_0)$, fulfills the following differential equation:

$$\frac{1}{2}\frac{\partial^2}{\partial x_0^2}\hat{\mathcal{P}}\left(u;\ h,\ x_0\right) - u\ U_h(x_0)\ \hat{\mathcal{P}}\left(u;\ h,\ x_0\right) = 0 \tag{D.2}$$

with boundary conditions $\lim_{x_0\to\infty}\hat{\mathcal{P}}\,(u;\ h,\ x_0) = 0$ and $\lim_{x_0\to 0}\hat{\mathcal{P}}\,(u;\ h,\ x_0) = 1$. Note that this is a differential equation with respect to the initial position x_0. The general solution to this differential equation is given by

$$\sqrt{2}\,C_1\sqrt{u\left(h+x_0\right)}\,I_1\left(2\sqrt{2}\sqrt{u\left(h+x_0\right)}\right) - \sqrt{2}\,C_2\sqrt{u\left(h+x_0\right)}\,K_1\left(2\sqrt{2}\sqrt{u\left(h+x_0\right)}\right), \tag{D.3}$$

where $I_1(x)$ and $K_1(x)$ are modified Bessel functions of the first and second kind respectively, and C_1 and C_2 are constants to be determined via the boundary conditions. Because $I_1(x_0)$ diverges for $x_0 \to \infty$, C_1 must be zero, and C_2 is then fixed via the other boundary condition. Finally, by setting $x_0 = 1$ we reach a remarkably simple expression for the Laplace transform of $\mathcal{P}^{g_s}(g_s,\, h)$,

$$\hat{\mathcal{P}}(u;\, h) = \frac{\sqrt{u(h+1)}\;\; K_1\left(2\sqrt{2}\sqrt{u(h+1)}\right)}{\sqrt{uh}\;\; K_1\left(2\sqrt{2}\sqrt{uh}\right)}. \tag{D.4}$$

This result is not only of interest in itself, but also provides a convenient way of evaluating $\mathcal{P}^{g_s}(g_s,\, h)$ by numerically inverting equation (D.4) (see figure 7 in the main text). We can also recover the asymptotic exponents γ_1, γ_2 of $\mathcal{P}^{g_s}(g_s,\, h)$ directly from its Laplace transform, equation (D.4). To see this, we consider the first and second derivatives of $\hat{\mathcal{P}}(u;\, h)$,

$$-\partial_u\hat{\mathcal{P}}(u;\, h) \sim \sqrt{2/(hu)} \text{ for } 1 \ll h, \tag{D.5}$$

$$\partial_{uu}\hat{\mathcal{P}}(u;\, h) \sim \frac{2}{u} \text{ for } u \ll 1. \tag{D.6}$$

The first equation assumes large h, while the second does not; this allows us to recover the two scaling regions mentioned in the main text. Then it is easy to check that an application of a Tauberian theorem [32, p 192] leads to equation (17) in the main text, recovering not only the asymptotic exponents γ_1, γ_2, but also their associated first order amplitudes.

References

[1] Schorlemmer D and Woessner J 2008 Probability of detecting an earthquake *Bull. Seismol. Soc. Am.* **98** 2103–17
[2] Lovejoy S, Lilley M, Desaulniers-Soucy N and Schertzer D 2003 Large particle number limit in rain *Phys. Rev.* E **68** 025301
[3] Paczuski M, Boettcher S and Baiesi M 2005 Interoccurrence times in the Bak–Tang–Wiesenfeld sandpile model: a comparison with the observed statistics of solar flares *Phys. Rev. Lett.* **95** 181102
[4] Bak P and Sneppen K 1993 Punctuated equilibrium and criticality in a simple model of evolution *Phys. Rev. Lett.* **71** 4083–6
[5] Pruessner G 2012 *Self-Organized Criticality* (Cambridge: Cambridge University Press)
[6] Paczuski M, Maslov S and Bak P 1996 Avalanche dynamics in evolution, growth, and depinning models *Phys. Rev.* E **53** 414–43
[7] Sneppen K 1995 Minimal SOC: intermittency in growth and evolution *Scale Invariance, Interfaces, and Non-Equilibrium Dynamics* ed A McKane, M Droz, J Vannimenus and D Wolf (New York: Plenum) pp 295–302
Sneppen K 1994 *NATO Advanced Study Institute on Scale Invariance, Interfaces, and Non-Equilibrium Dynamics (Cambridge, UK, 20–30 June 1994)* pp 20–30
[8] Grassberger P 1995 The Bak-Sneppen model for punctuated evolution *Phys. Lett.* A **200** 277–82
[9] Garber A, Hallerberg S and Kantz H 2009 Predicting extreme avalanches in self-organized critical sandpiles *Phys. Rev.* E **80** 026124
[10] Gardiner C W 1997 *Handbook of Stochastic Methods* 2nd edn (Berlin: Springer)
[11] Hinrichsen H 2000 Non-equilibrium critical phenomena and phase transitions into absorbing states *Adv. Phys.* **49** 815–958
[12] Harris T E 1963 *The Theory of Branching Processes* (Berlin: Springer)
[13] Peters O, Hertlein C and Christensen K 2002 A complexity view of rainfall *Phys. Rev. Lett.* **88** 018701
[14] Galassi M, Davies J, Theiler J, Gough B, Jungman G, Alken P, Booth M and Rossi F 2009 *GNU Scientific Library Reference Manual* Network Theory Ltd. 3rd edn (v1.12) (www.network-theory.co.uk/gsl/manual/) accessed 18 August 2009
[15] Deluca A and Corral A 2013 Fitting and goodness-of-fit test of non-truncated and truncated power −law distributions *Acta Geophys.* **61** 1351–94
[16] Rubin K J, Pruessner G and Pavliotis G A 2014 Mapping multiplicative to additive noise *J. Phys. A: Math. Theor.* **47** 195001
[17] Majumdar S N and Comtet A 2004 Exact maximal height distribution of fluctuating interfaces *Phys. Rev. Lett.* **92** 225501
[18] Majumdar S N and Comtet A 2005 Airy distribution function: from the area under a brownian excursion to the maximal height of fluctuating interfaces *J. Stat. Phys.* **119** 777–826
[19] Mohanty G 1979 *Lattice Path Counting and Applications* (New York: Academic)
[20] Christensen K, Farid N, Pruessner G and Stapleton M 2008 On the scaling of probability density functions with apparent power-law exponents less than unity *Eur. Phys. J.* B **62** 331–6
[21] Corral A 2009 Point-occurrence self-similarity in crackling-noise systems and in other complex systems *J. Stat. Mech.* **2009** P01022
[22] Laurson L, Illa X and Alava M J 2009 The effect of thresholding on temporal avalanche statistics *J. Stat. Mech.* **2009** P01019
[23] Larremore D B, Shew W L, Ott E, Sorrentino F and Restrepo J G 2014 Inhibition causes ceaseless dynamics in networks of excitable nodes *Phys. Rev. Lett.* **112** 138103
[24] Peters O, Deluca A, Corral A, Neelin J D and Holloway C E 2010 Universality of rain event size distributions *J. Stat. Mech.* **11** P11030
[25] Press W H, Teukolsky S A, Vetterling W T and Flannery B P 2002 *Numerical Recipes in Fortran* 3rd edn (Cambridge: Cambridge University Press)
[26] Knuth D E 1997 Fundamental algorithms *The Art of Computer Programming* vol 1, 3rd edn (Reading, MA: Addison-Wesley)
[27] Stanley R P 1999 *Enumerative Combinatorics (Cambridge Studies in Advanced Mathematics* vol 2) (Cambridge: Cambridge University Press)
[28] Wolfram Research Inc. 2011 *Mathematica* (Champaign, IL: Wolfram Research Inc.) Version 8.0.1.0

[29] Abramowitz M and Stegun I A (ed) 1970 *Handbook of Mathematical Functions* (New York: Dover)
[30] Perret A, Comtet A, Majumdar S N and Schehr G 2015 On certain functionals of the maximum of Brownian motion and their applications (arXiv:1502.01218)
[31] Chaichian M and Demichev A 2001 *Path Integrals in Physics: Stochastic Processes and Quantum Mechanics (Institute of Physics Series in Mathematical and Computational Physics* vol 1) (London: Taylor and Francis)
[32] Widder D V 1946 *The Laplace Transform* (Princeton, NJ: Princeton University Press)

Bibliography

Adamic, L. (2011). Complex systems: Unzipping Zipf's law. *Nature*, 474(7350):164–165. (page 16).

Adamic, L. A. and Huberman, B. A. (2002). Zipf's law and the Internet. 3:143–150. (pages 16, 20).

Albert, R. and Barabási, A. (2002). Statistical mechanics of complex networks. *Rev. Mod. Phys.*, 74:47–97. (page 2).

Altmann, E. G., Cristadoro, G., and Degli Esposti, M. (2012). On the origin of long-range correlations in texts. *Proc. Natl. Acad. Sci. USA*, 109(29):11582–11587. (page 14).

Altmann, E. G. and Gerlach, M. (2015). Statistical laws in linguistics. *arXiv*, (1502.03296v1). (pages 13, 22).

Altmann, G. (1980). Prolegomena to Menzerath's law. *Glottometrika 2*, 2:1–10. (page 22).

Anderson, P. W. (1972). More is different. *Science*, 177:393–396. (page 2).

Andrade, R. F. S., Schellnhuber, H. J., and Claussen, M. (1998). Analysis of rainfall records: possible relation to self-organized criticality. *Physica A*, 254:557–568. (page 26).

Axtell, R. L. (2001). Zipf distribution of U.S. firm sizes. *Science*, 293:1818–1820. (page 16).

Baayen, H. (2001). *Word Frequency Distributions*. Kluwer, Dordrecht. (pages 17, 20).

Baek, S. K., Bernhardsson, S., and Minnhagen, P. (2011). Zipf's law unzipped. *New J. Phys.*, 13(4):043004. (page 16).

Baiesi, M., Paczuski, M., and Stella, A. L. (2006). Intensity thresholds and the statistics of the temporal occurrence of solar flares. *Phys. Rev. Lett.*, 96:051103. (page 26).

Bak, P. (1996). *How Nature Works: The Science of Self-Organized Criticality*. Copernicus, New York. (pages 2, 14, 16, 28).

Balague, N., Torrents, C., R., H., Davids, K., and Araújo, D. (2013). Overview of complex systems in sport. *J Syst Sci Complex*, 26(1):4–13. (page 2).

Bar-Yam, Y. (2003). *Dynamics Of Complex Systems*. Westview Press. (page 2).

Bernhardsson, S., da Rocha, L. E. C., and Minnhagen, P. (2009). The meta book and size-dependent properties of written language. *New J. Phys.*, 11(12):123015. (page 14).

Binney, J. J., Dowrick, N. J., Fisher, A. J., and Newman, M. E. J. (1992). *The Theory of Critical Phenomena*. Oxford University Press, Oxford. (pages 3, 6).

Boffetta, G., Carbone, V., Giuliani, P., Veltri, P., and Vulpiani, A. (1999). Power laws in solar flares: Self-organized criticality or turbulence? *Phys. Rev. Lett.*, 83:4662–4665. (page 26).

Bouchaud, J.-P. and Georges, A. (1990). Anomalous diffusion in disordered media: statistical mechanisms, models and physical applications. *Phys. Rep.*, 195:127–293. (page 38).

Boyd, R., Gintis, H., Bowles, S., and Richerson, P. J. (2003). The evolution of altruistic punishment. *Proc. Natl. Acad. Sci. USA*, 100(6):3531–3535. (page 2).

Camacho, J. and Solé, R. V. (2001). Scaling in ecological size spectra. *Europhys. Lett.*, 55:554. (page 16).

Christensen, K. and Moloney, N. R. (2005). *Complexity and Criticality*. Imperial College Press, London. (pages 9, 10, 22).

Clauset, A., Shalizi, C. R., and Newman, M. E. J. (2009). Power-law distributions in empirical data. *SIAM Rev.*, 51:661–703. (pages 11, 16, 22, 30).

Corominas-Murtra, B., Fortuny, J., and Solé, R. V. (2011). Emergence of Zipf's Law in the Evolution of Communication. *Phys. Rev. E*, 83:036115. (page 16).

Corominas-Murtra, B., Hanel, R., and Thurner, S. (2015). Understanding scaling through history-dependent processes with collapsing sample space. *Proc. Natl. Acad. Sci. USA*. (page 16).

Corral, A. (2008). Scaling and universality in the dynamics of seismic occurrence and beyond. In Carpinteri, A. and Lacidogna, G., editors, *Acoustic Emission and Critical Phenomena*, pages 225–244. Taylor and Francis, London. (page 9).

Corral, A. (2010). Tropical cyclones as a critical phenomenon. In Elsner, J. B., Hodges, R. E., Malmstadt, J. C., and Scheitlin, K. N., editors, *Hurricanes and Climate Change: Volume 2*, pages 81–99. Springer, Heidelberg. (page 26).

Corral, A., Boleda, G., and Ferrer-i-Cancho, R. (2015). Zipf's law for word frequencies: Word forms versus lemmas in long texts. *PLoS ONE*, (In press). (page 16).

Corral, A., Deluca, A., and Ferrer-i-Cancho, R. (2012). A practical recipe to fit discrete power-law distributions. *arXiv*, (1209.1270). (pages 18, 19, 22).

Corral, A., Font, F., and Camacho, J. (2011). Non-characteristic half-lives in radioactive decay. *Phys. Rev. E*, 83:066103. (page 22).

Corral, A., Telesca, L., and Lasaponara, R. (2008). Scaling and correlations in the dynamics of forest-fire occurrence. *Phys. Rev. E*, 77:016101. (page 26).

Czirók, A., Mantegna, R. N., Havlin, S., and Stanley, H. E. (1995). Correlations in binary sequences and a generalized Zipf analysis. *Phys. Rev. E*, 52(1):446. (page 16).

Davidsen, J. and Kwiatek, G. (2013). Earthquake interevent time distribution for induced micro-, nano-, and picoseismicity. *Phys. Rev. Lett.*, 110:068501. (page 26).

Deluca, A. and Corral, A. (2013). Fitting and goodness-of-fit test of non-truncated and truncated power-law distributions. *Acta Geophys.*, 61:1351–1394. (pages 11, 15, 22, 31).

Deluca, A. and Corral, A. (2014). Scale invariant events and dry spells for medium-resolution local rain data. *Nonlinear Proc. Geophys.*, 21:555–567. (pages 26, 28).

Ebeling, W. and Pöschel, T. (1994). Entropy and long-range correlations in literary English. *Europhys. Lett.*, 26:241–246. (page 14).

Ferrer i Cancho, R. (2005). Zipf's law from a communicative phase transition. *Eur. Phys. J. B*, 47:449–457. (page 16).

Ferrer i Cancho, R. and Hernández-Fernández, A. (2008). Power laws and the golden number. In Altmann, G., Zadorozhna, I., and Matskulyak, Y., editors, *Problems of General, Germanic and Slavic Linguistics*, number Books - XII, pages 518–523. Chernivtsi. (page 20).

Ferrer i Cancho, R. and Solé, R. V. (2001). Two regimes in the frequency of words and the origin of complex lexicons: Zipf's law revisited. *J. Quant. Linguist.*, 8(3):165–173. (page 21).

Ferrer i Cancho, R. and Solé, R. V. (2003). Least effort and the origins of scaling in human language. *Proc. Natl. Acad. Sci. USA*, 100:788–791. (page 16).

Font-Clos, F., Boleda, G., and Corral, A. (2013). A scaling law beyond Zipf's law and its relation to Heaps' law. *New J. Phys.*, 15(9):093033. (pages 14, 20, 34).

Font-Clos, F. and Corral, A. (2014). Reply to "Comment on 'A Scaling law beyond Zipf's law and its relation to Heaps' law'". *arXiv*, (1405.0207). (page 14).

Font-Clos, F. and Corral, A. (2015). Log-Log Convexity of Type-Token Growth in Zipf's Systems. *Phys. Rev. Lett.*, (In press). (pages 35, 37, 38, 40).

Font-Clos, F., Pruessner, G., Moloney, N. R., and Deluca, A. (2015). The perils of thresholding. *New J. Phys.*, 17(4):043066. (pages 28, 30, 31, 36, 39, 40).

Furusawa, C. and Kaneko, K. (2003). Zipf's law in gene expression. *Phys. Rev. Lett.*, 90:088102. (page 16).

Gabaix, X. (1999). Zipf s law for cities: an explanation. *Quart. J. Econ.*, 114:739–767. (page 16).

Gerlach, M. and Altmann, E. G. (2013). Stochastic model for the vocabulary growth in natural languages. *Phys. Rev. X*, 3:021006. (page 21).

Gisiger, T. (2001). Scale invariance in biology: coincidence or footprint of a universal mechanism? *Biol. Rev.*, 76:161–209. (page 2).

Gnedenko, B. V. and Korolyuk, V. (1951). On the maximum discrepancy between two empirical distributions. *Dokl. Akad. Nauk SSSR*, 80:525–528. (page 5).

Grasso, J. R. and Bachélery, P. (1995). Hierarchical organization as a diagnostic approach to volcano mechanics: Validation on Piton de la Fournaise. *Geophys. Res. Lett.*, 22(21):2897–2900. (page 26).

Guiraud, H. (1954). *Les charactères statistiques du vocabulaire*. Presses Universitaires de France. (page 22).

Hanser, S. F., Doyle, L. R., McCowan, B., and Jenkins, J. M. (2004). Information Theory Applied to Animal Communication Systems and Its Possible Application to SETI. In Norris, R. and Stootman, F., editors, *Bioastronomy 2002: Life Among the Stars*, volume 213 of *IAU Symposium*, page 514. (page 16).

Heaps, H. S. (1978). *Information retrieval: computational and theoretical aspects*. Academic Press. (page 22).

Herdan, G. (1960). *Type-Token Mathematics*. Mouton de Gruyter, Berlin. (page 22).

Holland, J. (2000). *Emergence: From Chaos to Order*. Popular science / Oxford University Press. Oxford University Press. (page 2).

Jensen, H. J. (1998). *Self-Organized Criticality*. Cambridge University Press, Cambridge. (page 28).

Kadanoff, L. P. (1986). Fractals: Where's the Physics? *Phys. Today*, 39(2):6. (page 2).

Kagan, Y. Y. (2010). Earthquake size distribution: Power-law with exponent $\beta \equiv 1/2$? *Tectonophys.*, 490:103–114. (page 26).

Köhler, R., Altman, G., and Piotrowski, R. G., editors (2005). *Quantitative Linguistik. Ein internationales Handbuch. Quantitative Linguistics. An international Handbook.* de Gruyter, Berlin. (page 14).

Kornai, A. (2002). How many words are there? 2:61–86. (page 20).

Li, W. (1992). Random texts exhibit zipf's-law-like word frequency distribution. *IEEE Trans. Inf. Theory*, 38(6):1842–1845. (page 16).

Li, W., Miramontes, P., and Cocho, G. (2010). Fitting ranked linguistic data with two-parameter functions. *Entropy*, 12(7):1743–1764. (page 17).

Lippiello, E., Corral, A., Bottiglieri, M., Godano, C., and de Arcangelis, L. (2012). Scaling behavior of the earthquake intertime distribution: Influence of large shocks and time scales in the Omori law. *Phys. Rev. E*, 86:066119. (page 26).

Malamud, B. D., Millington, J. D. A., and Perry, G. L. W. (2005). Characterizing wildfire regimes in the United States. *Proc. Natl. Acad. Sci. USA*, 102:4694–4699. (page 26).

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. *Phys. Rev. E*, 83:036111. (page 16).

Mandelbrot, B. (1961). On the theory of word frequencies and on related Markovian models of discourse. In Jakobson, R., editor, *Structure of Language and its Mathematical Aspects*, pages 190–219. American Mathematical Society, Providence, RI. (pages 18, 20).

McCowan, B., Hanser, S. F., and Doyle, L. R. (1999). Quantitative tools for comparing animal communication systems: information theory applied to bottlenose dolphin whistle repertoires. *Anim. Behav.*, 57(2):409–419. (page 16).

Miller, G. A. (1957). Some effects of intermittent silence. *Am. J. Psychol.*, 70(2):311–314. (page 16).

Mitchell, M. (2009). *Complexity: A Guided Tour*. Oxford University Press. (page 2).

Mitzenmacher, M. (2004). A brief history of generative models for power law and lognormal distributions. *Internet Math.*, 1 (2):226–251. (page 16).

Montemurro, M. and Pury, P. A. (2002). Long-range fractal correlations in literary corpora. *Fractals*, 10:451–461. (page 14).

Moreno, I., Font-Clos, F., and Corral, A. (2015). In preparation. (page 21).

Newman, M. E. J. (2005). Power laws, Pareto distributions and Zipf's law. *Cont. Phys.*, 46:323 –351. (page 16).

Newman, M. E. J. (2010). *Networks: an Introduction*. Oxford University Press. (page 2).

Newman, M. E. J. (2011). Complex systems: A survey. *Am. J. Phys.*, 79:800–810. (page 2).

Paczuski, M., Boettcher, S., and Baiesi, M. (2005). Interoccurrence times in the Bak-Tang-Wiesenfeld sandpile model: A comparison with the observed statistics of solar flares. *Phys. Rev. Lett.*, 95:181102. (page 26).

Peters, O. and Christensen, K. (2002). Rain: Relaxations in the sky. *Phys. Rev. E*, 66:036120. (page 26).

Peters, O. and Christensen, K. (2006). Rain viewed as relaxational events. *J. Hidrol.*, 328:46–55. (page 26).

Peters, O., Deluca, A., Corral, A., Neelin, J. D., and Holloway, C. E. (2010). Universality of rain event size distributions. *J. Stat. Mech.*, P11030. (page 26).

Peters, O., Hertlein, C., and Christensen, K. (2002). A complexity view of rainfall. *Phys. Rev. Lett.*, 88:018701. (pages 2, 26).

Petersen, A. M., Tenenbaum, J. N., Havlin, S., Stanley, H. E., and Perc, M. (2012). Languages cool as they expand: Allometric scaling and the decreasing need for new words. *Sci. Rep.*, 2. (page 21).

Petruszewycz, M. (1973). L'histoire de la loi d'estoup-zipf : documents. *Math. Sci. Humaines*, 44:41–56. (page 23).

Pruessner, G. (2009). Probability densities in complex systems, measuring. In Meyers, R. A., editor, *Encyclopedia of Complexity and Systems Science*, pages 6990–7009. Springer New York. (page 22).

Pruessner, G. (2012). *Self-Organised Criticality: Theory, Models and Characterisation*. Cambridge University Press, Cambridge. (page 28).

Pueyo, S. and Jovani, R. (2006). Comment on "A keystone mutualism drives pattern in a power function". *Science*, 313:1739c–1740c. (page 16).

Rao, F., Garrett-Roe, S., and Hamm, P. (2010). Structural inhomogeneity of water by complex network analysis. *J. Phys. Chem. B*, 114(47):15598–15604. (page 2).

Saichev, A., Malevergne, Y., and Sornette, D. (2009). *Theory of Zipf's Law and of General Power Law Distributions with Gibrat's Law of Proportional Growth*. Lecture Notes in Economics and Mathematical Systems. Springer Verlag, Berlin. (page 16).

Serrà, J., Corral, A., Boguñá, M., Haro, M., and Arcos, J. L. (2012). Measuring the evolution of contemporary western popular music. *Sci. Rep.*, 2:521. (page 16).

Sethna, J. P. (2006). *Statistical Mechanics: Entropy, Order Parameters, and Complexity*. Oxford University Press, New York. (page 3).

Simon, H. A. (1955). On a class of skew distribution functions. *Biomet.*, 42:425–440. (page 16).

Solé, R. V. and Manrubia, S. C. (2009). *Orden y Caos En Sistemas Complejos. Aplicaciones*. Edicions UPC SL. (page 2).

Sornette, A. and Sornette, D. (1989). Self-organized criticality and earthquakes. *Europhys. Lett.*, 9:197–202. (page 26).

Stanley, H. E. (1999). Scaling, universality, and renormalization: Three pillars of modern critical phenomena. *Rev. Mod. Phys.*, 71:S358–S366. (pages 3, 7).

Stigler, S. (1980). *Science and Social Structure: A Festschrift for Robert K. Merton.* Number 2. New York Academy of Science Translations. (page 23).

Takayasu, H. (1989). *Fractals in the Physical Sciences.* Manchester University Press, Manchester. (page 9).

Tao, T. (2012). E pluribus unum: From Complexity, Universality. *Dædalus*, 141(3):23–24. (pages 3, 4).

Turcotte, D. L. and Malamud, B. D. (2004). Landslides, forest fires, and earthquakes: examples of self-organized critical behavior. *Physica A*, 340:580–589. (page 26).

Williams, J. R., Bagrow, J. P., Danforth, C. M., and Dodds, P. S. (2015). Text mixing shapes the anatomy of rank-frequency distributions. *Phys. Rev. E*, 91:052811. (page 21).

Yan, X.-Y. and Minnhagen, P. (2014). Comment on 'A scaling law beyond Zipf's law and its relation to Heaps' law'. (1303.0705). (page 14).

Zanette, D. (2012). Statistical patterns in written language. *arXiv*, (1412.3336). (pages 16, 20).

Zanette, D. and Montemurro, M. (2005). Dynamics of text generation with realistic Zipf's distribution. *J. Quant. Linguist.*, 12(1):29–40. (page 16).

Zipf, G. K. (1949). *Human Behavior and the Principle of Least Effort.* Addison-Wesley. (pages 16, 18, 20).