The perils of thresholding

2015 New J. Phys. 17 043066

(http://iopscience.iop.org/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

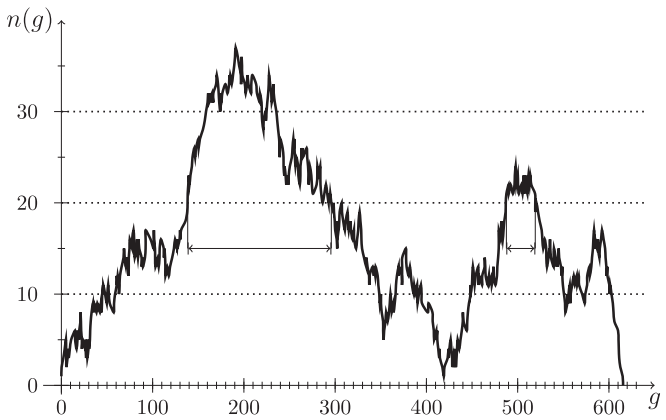

Figure 1. Example of thresholding of a time series. An event begins when the signal exceeds the threshold (dotted lines, $h = 10,\ 20,\ 30$) and ends as soon as the signal falls below the threshold. Increasing levels of the threshold lead (non-monotonically) to different numbers of events and, provided 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 $\langle \cdot \rangle$

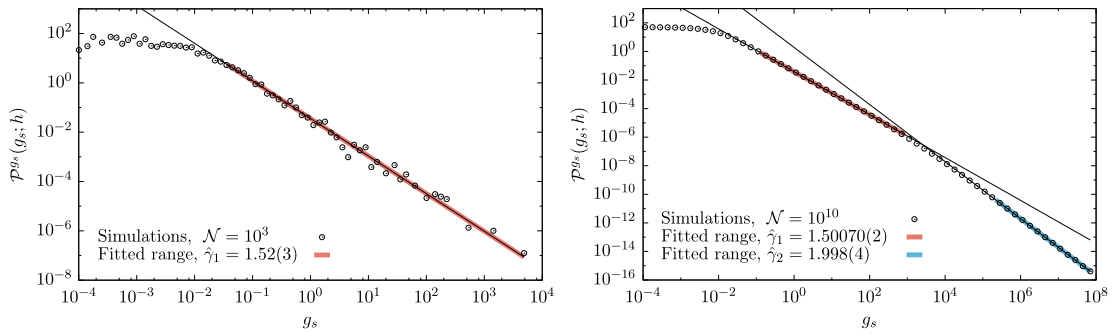

Figure 2. (a): the PDF $\mathcal{P}^{g_s}(g_s;\ h)$ of the survival time g_s of a thresholded BDP, with a threshold of $h = 100$, estimated from Monte Carlo simulations using a limited sample size of $\mathcal{N} = 10^3$. Fitting a power law yields an exponent of $\hat{\gamma}_1 = 1.52(3)$ over the range $[0.031,\ 1.259 \times 10^5]$, 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

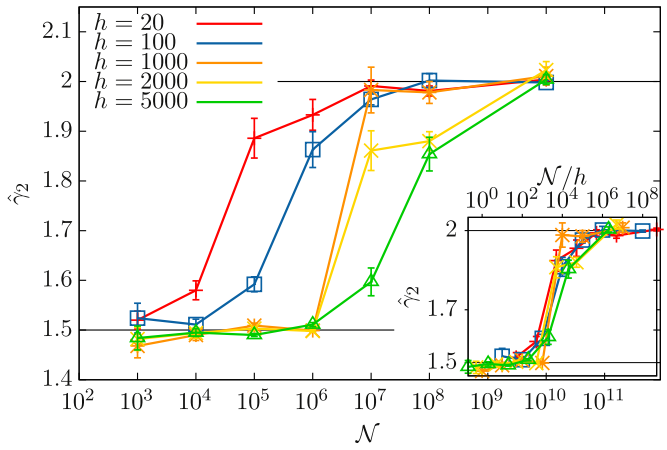

Figure 3. Estimated large scale exponent $\hat{\gamma}_2$ for different thresholds h and sample size $\mathcal{N}$. The error bars correspond to one standard deviation and are inversely proportional to the number of data within the fitted range. *Inset*: estimated large scale exponent $\hat{\gamma}_2$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

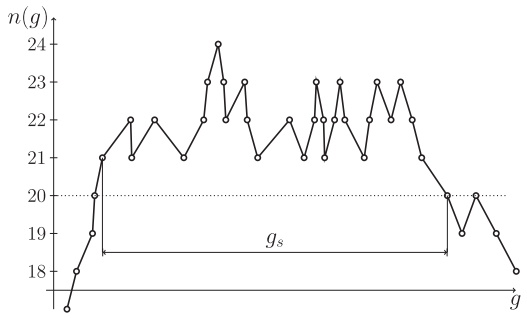

Figure 4. Magnification of the right interval in figure 1. The clock starts when $n(g)$ exceeds the threshold and stops when $n(g)$ returns to the threshold.

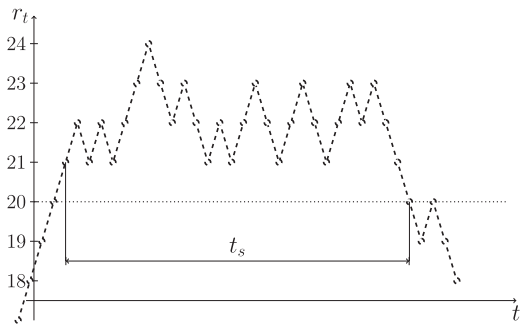

Figure 5. The same data as in figure 4 but on the mapped time scale of the random walker, which evolves in equally spaced, discrete steps. The survival time is necessarily odd, $t_s = 2T - 1$, $T \in \mathbb{N}$ ($t_s = 29$ in this example).

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 $G \sim g_\times \sim h$ 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.

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 $t_X = 8h^2/\pi$, for which $g_s(t_X,\, h) \approx 4h/\pi$. 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 $g_\times = 16h/\pi$. The PDF of the survival time of the BD process finally reads

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

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 $g_\times = 8\pi h$, shown in figures 7 and 8, can be determined by assuming continuity of $\mathcal{P}^{g_s}(g_s;\, h)$ and thus imposing $\frac{1}{\sqrt{2h\pi}} g_\times^{-3/2} = 2g_\times^{-2}$. 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 $g_\times/h = 8\pi$, 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.

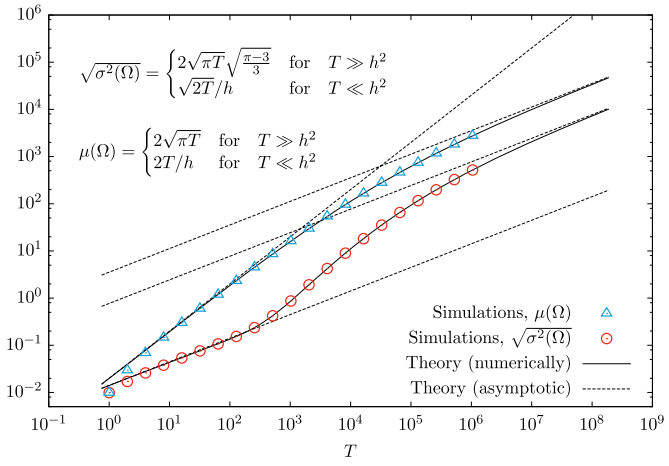

Figure 6. Numerical comparison of the approximations equations (11) and (13) (shown as full lines), their asymptotes equations (12) and (14) (dashed) and the numerical estimates based on a sample of 10^5 realizations per datapoint in a Monte-Carlo simulation of a birth–death process constrained to $2T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

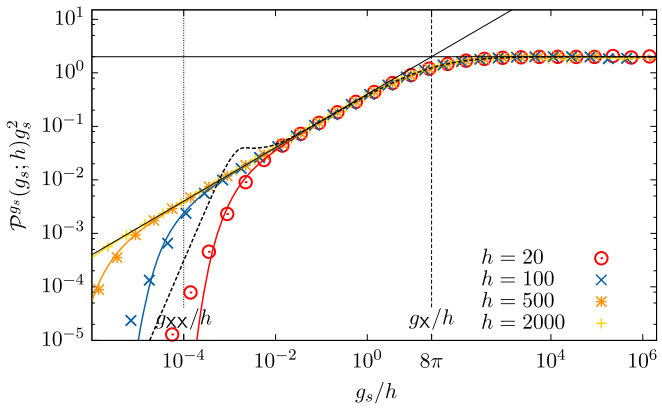

Figure 7. Collapse of the PDFs for different thresholds h for large $g_s \gg 1/h$, plotting $\mathcal{P}^{g_s}(g_s;\, h)g_s^{\,2}$ against g_s/h, according to equations (9) and (18), capturing equation (17). Symbols correspond to simulations at different threshold levels $h = 20,\ 200,\ 500$ and 2000. 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).

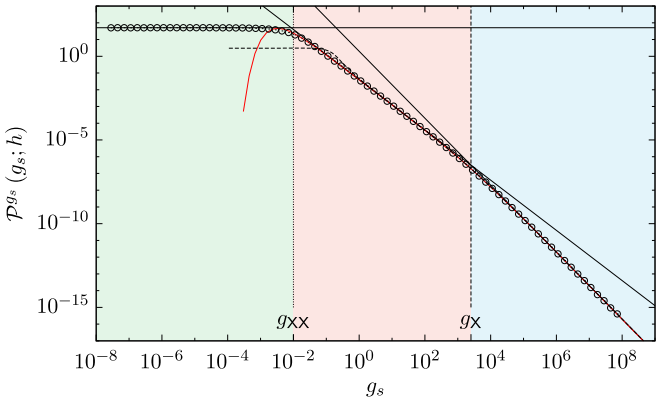

Figure 8. The PDF of survival times $\mathcal{P}^{g_s}(g_s;\, h)$ for $h = 100$. Three scaling regimes partitioned by g_X (thin dashed line) and g_{XX} (thin dotted line) exist: for very short times $g_s \ll 1/h$ (green shading), the exponential waiting time to the first (death) event dominates, so that $\mathcal{P}^{g_s}(g_s;\, h) \sim (h+1)/2$. 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 $\sigma^2(\Omega)$

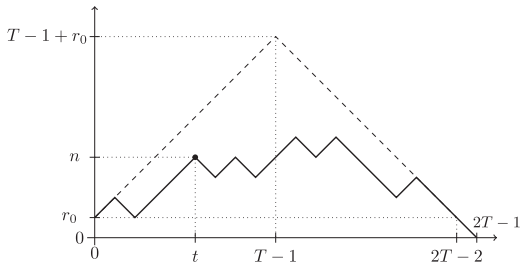

Figure B1. Sample path of a random walk along an absorbing wall at 0. The walker starts at $t=0$ from r_0 and terminates at $2T-1$ by reaching the wall $r_{2T-1}=0$, i.e. $r_{2T-2}=1$. By construction, it cannot escape from the region demarcated by the dashed line. When counting distinct paths, the number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

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

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

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

(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):

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

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]

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

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 $\hat{Z}_{\text{hom}}(x, y, s) = \Psi(x, s)\rho(y, s)$, 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

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

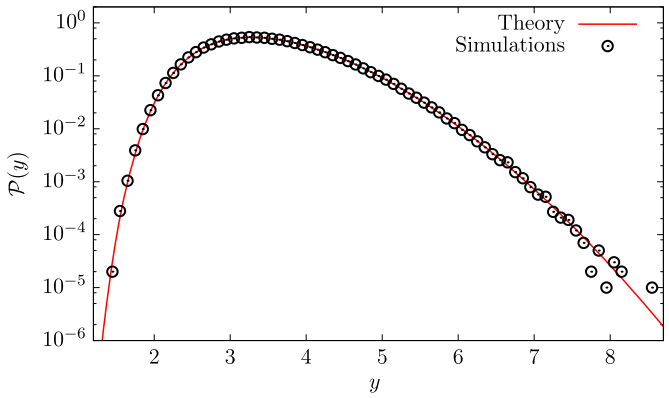

Figure C1. The distribution of $y = \int_0^2 \mathrm{d}t\ 1/x(t)$, where $x(t)$ is a Brownian excursion of length 2. The red full line is the analytical result and black symbols correspond to simulations.

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)