

Underestimating extreme events in power-law behavior due to machine-dependent cutoffs

Filippo Radicchi*

Center for Complex Networks and Systems Research, School of Informatics and Computing, Indiana University, Bloomington, Indiana 47408, USA

(Received 22 April 2014; published 3 November 2014)

Power-law distributions are typical macroscopic features occurring in almost all complex systems observable in nature. As a result, researchers in quantitative analyses must often generate random synthetic variates obeying power-law distributions. The task is usually performed through standard methods that map uniform random variates into the desired probability space. Whereas all these algorithms are theoretically solid, in this paper we show that they are subject to severe machine-dependent limitations. As a result, two dramatic consequences arise: (i) the sampling in the tail of the distribution is not random but deterministic; (ii) the moments of the sample distribution, which are theoretically expected to diverge as functions of the sample sizes, converge instead to finite values. We provide quantitative indications for the range of distribution parameters that can be safely handled by standard libraries used in computational analyses. Whereas our findings indicate possible reinterpretations of numerical results obtained through flawed sampling methodologies, they also pave the way for the search for a concrete solution to this central issue shared by all quantitative sciences dealing with complexity.

DOI: [10.1103/PhysRevE.90.050801](https://doi.org/10.1103/PhysRevE.90.050801)

PACS number(s): 89.75.Da, 89.20.-a, 89.75.Hc

If the probability $P(x)$ to observe a particular value x of some quantity is inversely proportional to the λ th power of that value, i.e., $P(x) \sim x^{-\lambda}$, the quantity under observation is said to follow a power-law probability density function (PDF) or distribution. The first empirical observation of a power-law PDF dates back to the analysis by Pareto at the end of the 1800s of the distribution of income and wealth among the population of Italy [1]. Not long afterward, Zipf noted that the frequency of any word in natural languages is inversely proportional to its rank in the frequency table [2]. Historically speaking, the contributions by Pareto and Zipf represented only the beginning of a long series of empirical discoveries of power-law distributions in natural and manmade systems. Thanks to the possibility offered by modern computational technologies to retrieve, store, and analyze large-scale sets of data, the amount of such empirical evidence has drastically increased in the past 10–15 years, and it is continuously growing. Power-law PDFs have been observed by researchers in almost all scientific disciplines, including physics [3–5], biology [6], earth and planetary sciences [7,8], economics and finance [9–11], computer science [12,13], and social science [14,15], just to mention a few. For a recent review of power laws in empirical data, see [16]. Power-law distributions generally emerge in self-organized, critical, multiscale, and collective phenomena, and their ubiquity is interpreted as a consequence of the intrinsic complexity that drives the dynamics and organization of natural systems.

Computers not only play a fundamental role in the analysis of data, but they are also often used to run numerical simulations with the aim of understanding and possibly predicting the behavior of complex systems. Since power-law distributions are the emblematic features of complexity, computer simulations and analyses often rely on the construction of sequences of synthetic power-law variates, i.e., random numbers extracted from power-law distributions. Whereas the

presence and consequences of serious numerical errors have already been studied in the context of record statistics [17,18], so far none has questioned the effectiveness of the algorithms developed for the generation of power-law random variates. In this paper, we show that all methods currently adopted for this purpose are instead subjected to severe machine-dependent limitations.

For simplicity, we will focus on the case of continuous random variates obeying the power-law PDF

$$P(x) = \frac{1 - \lambda}{x_M^{1-\lambda} - x_m^{1-\lambda}} x^{-\lambda} \quad (1)$$

if $x_m \leq x \leq x_M$, and $P(x) = 0$, otherwise. $x_M \geq x_m \geq 1$ indicates the upper and lower bounds of the support of $P(x)$, whereas $\lambda > 1$ is the exponent of the power-law distribution. The k th moment of the PDF of Eq. (1) is given by

$$\langle x^k \rangle = \frac{1 - \lambda}{k + 1 - \lambda} \frac{x_M^{k+1-\lambda} - x_m^{k+1-\lambda}}{x_M^{1-\lambda} - x_m^{1-\lambda}}, \quad (2)$$

which is valid for $k + 1 \neq \lambda$. For $k + 1 = \lambda$, we have instead $\langle x^k \rangle = \frac{(1-\lambda)(\log x_M - \log x_m)}{x_M^{1-\lambda} - x_m^{1-\lambda}}$. The previous equation mathematically formalizes a fundamental property of power-law distributions: all k th moments, with $k + 1 \geq \lambda$, diverge as powers (or logarithms) of the upper bound x_M . The divergence of the moments of power-law PDFs is at the basis of many fundamental theoretical results. One requirement in the computer generation of power-law random numbers is thus to preserve this property. However, this task is only possible under very restrictive conditions. Let us illustrate in detail the origin of these numerical problems.

Suppose we want to create a sequence of computer-generated random variables extracted from the PDF of Eq. (1). Whereas there are many different ways to perform this task, here we will concentrate on the so-called inversion method [16,19]. We will come back to this point later, but we can already anticipate that the problems arising in the inversion method are common to all other methods that can possibly be used to generate nonuniform random variates.

*filiradi@indiana.edu

Our choice to focus on this method is not arbitrary, but is motivated by its popularity in numerical analyses [16]. The algorithm consists of two steps: (i) extract a random variable q from a uniform distribution defined over the interval $(0, 1)$; (ii) compute the random variate x from the desired distribution $P(x)$ by inverting the relation $\int_x^{x_M} dy P(y) = q$. While the approach is valid for any $P(x)$, it turns out to be highly applicable in all cases in which the integral on the left-hand side of the previous equation can be expressed explicitly, so that x can be written in terms of q . For power-law distributions, we can write

$$x = [x_M^{1-\lambda} - (x_M^{1-\lambda} - x_m^{1-\lambda})q]^{1/(1-\lambda)}, \quad (3)$$

and we use this relation to generate power-law distributed random variables x directly from uniformly distributed random variates q . Note that, in typical situations, $x_M \gg x_m$, so that Eq. (3) can be written as $x \simeq x_m q^{1/(1-\lambda)}$.

The inversion method is very elegant and theoretically solid, but, computationally speaking, it suffers from a severe weakness: its effectiveness relies on the goodness of the generator of random uniform variables. Generally, this fundamental role is played by pseudo-random-number generators (PRNGs). These are deterministic algorithms capable of producing high-quality uniform variates with very long periods [19,20]. However, the precision at which these numbers are created is finite. Good PRNGs available on the market today have a precision of $B_r \geq 32$ bits, meaning that the minimal distance between two random numbers generated by the PRNG is at maximum 2^{-B_r} . While this precision is completely satisfactory for the generation of uniform random variates, when applied to the transformation of Eq. (3) it translates into a series of machine-dependent thresholds at which the distribution of the samples produced by the machine drastically diverges from the theoretical distribution $P(x)$.

The first problem that arises is the dramatic worsening of the accuracy at which random variates are extracted from $P(x)$. Suppose that we want to be able to discriminate between two consecutive random numbers with an accuracy equal to a . This is possible only if the distance in the probability space between x and $x + a$ is larger than 2^{-B_r} . This requirement is first violated at $x = x^*$, with x^* defined by $P(x^*) - P(x^* + a) = 2^{-B_r}$. When $a \ll x^*$, the term of the left-hand side can be approximated by the negative derivative of $P(x)$ computed at x^* . For power-law PDFs, we thus have

$$x^* \simeq 2^{\frac{B_r}{\lambda-1}} x_m^{\frac{\lambda-1}{\lambda}} [a\lambda(\lambda-1)]^{\frac{1}{\lambda-1}}, \quad (4)$$

valid in the limit $x_M \gg x_m$. For values of $x > x^*$, power-law random variates are no longer extracted with the required accuracy a . Consecutive admissible values differ by amounts larger than a , giving rise to a unwanted discretization (see Fig. 1). The most evident effect of discretization is the presence of plateaus in the PDF of the samples. Plateaus are defined over overlapping intervals on the x axis, and sudden jumps occur among consecutive plateaus. As x increases, the discretization gets worse (i.e., plateaus become wider), until we reach the final plateau corresponding to a probability equal to 2^{-B_r} . This level is reached for $x = x^\dagger$, where x^\dagger is defined by

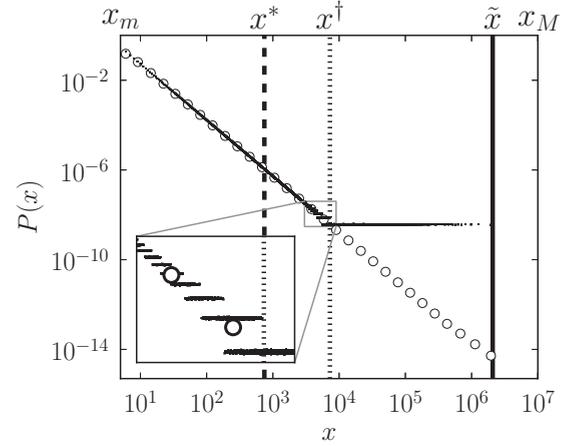


FIG. 1. Sample distribution obtained from 10^{12} random variates extracted from a power-law PDF with parameters $x_m = 5$, $x_M = 10^7$, and $\lambda = 2.5$. For the generation of the synthetic variates, we used a PRNG with $B_r = 28$ bits of precision (see the Supplemental Material in Ref. [21]). For simplicity of illustration, x values have been rounded to the closest integer value, and the sample PDF is normalized by the sum of these integer numbers. Except for the region $x < x^*$ (on the left of the dashed line), the histogram based on linear binning (black squares) reveals the presence of data discretization as discussed in the text (see the inset for a zoom of this region). The final plateau corresponds to a probability equal to 2^{-B_r} . The histogram based on logarithmic binning hinders the discretization of the data (white circles). The fact that the exponent measurable from the histogram based on logarithmic binning is equal to λ also in the region between x^\dagger and \tilde{x} (between dotted and full lines) indicates that the distance between consecutive admissible variates in the final plateau is becoming powerlike with exponent λ . No points are visible in the region $x > \tilde{x}$ (on the right of the full line).

$P(x^\dagger) = 2^{-B_r}$. For power-law PDFs, this means

$$x^\dagger \simeq 2^{\frac{B_r}{\lambda}} x_m^{\frac{\lambda-1}{\lambda}} (\lambda-1)^{\frac{1}{\lambda}}, \quad (5)$$

which is valid for $x_M \gg x_m$. The discretization of the probability and the sample space is certainly deleterious. The tail of the true distribution is not sampled at random, but in a deterministic fashion. No matter how many variables we extract, the sample PDF will never approach the true distribution, and particular values of x will never be extracted. On the other hand, the consequences of this discretization on the effective divergence of the moments of the sample distribution are not as serious as those produced by the third threshold that we are going to describe.

Proceeding in the direction of increasing x , an additional and much more severe threshold appears. Since no uniform numbers lower than 2^{-B_r} can be extracted, there is no possibility to generate random variates from the distribution $P(x)$ larger than \tilde{x} defined by $\int_{\tilde{x}}^{x_M} dy P(y) = 2^{-B_r}$. For the special case of power-law PDFs, this machine-dependent cutoff in the sample PDF can be estimated as

$$\tilde{x} \simeq x_m 2^{\frac{B_r}{\lambda-1}}, \quad (6)$$

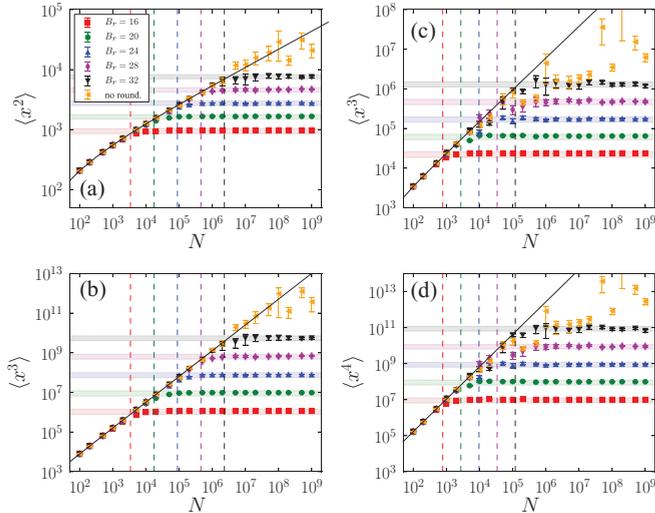


FIG. 2. (Color online) Convergence of the moments in the sample distribution of N random variates. We fix $x_m = 5$ and $x_M = N$. The power-law exponent is $\lambda = 2.7$ in panels (a) and (b), and $\lambda = 3.2$ in panels (c) and (d). In each simulation, we extracted N random variates and calculated the moments of the sample distribution. Points in the figure refer to the average value of the moments of the sample PDF over at least 100 independent realizations. Error bars, when visible, quantify the standard error associated with this measure. Black lines, corresponding to $\int_{x_m}^{x_M} dy y^k P(y)$, represent the expected value of the k th moment for the theoretical distribution. Dashed vertical lines stand for the machine-dependent cutoffs predicted by Eq. (6). Horizontal shaded areas are delimited by $\int_{x_m}^{\tilde{x}} dy y^k P(y)$ and Eq. (7). Different colors and symbol types correspond to different values of the number of bits B_r used in the generation of uniform random variates. Orange points are obtained without rounding the value of the random uniform variates (see Ref. [21] for details).

again valid for $x_M \gg x_m$. The presence of a machine-dependent cutoff implies that the k th moment of the sample distribution behaves as $\langle x^k \rangle \simeq \int_{x_m}^{\tilde{x}} dy y^k P(y) + 2^{-B_r} \tilde{x}^k$.

For power-law PDFs with $k + 1 > \lambda$, the k th moment of the sample distribution, which is supposed to diverge as x_M increases, instead converges to

$$\langle x^k \rangle \simeq x_m^k 2^{\frac{B_r(k+1-\lambda)}{\lambda-1}} \frac{k}{k+1-\lambda}. \quad (7)$$

A similar expression can be also found for $k + 1 = \lambda$.

To illustrate the practical importance of the machine-dependent limitations described so far, we present numerical evidence of problems induced by the finite precision PRNGs in two very common situations. Figure 2 illustrates the errors committed in a typical situation faced in the generation of random networks with prescribed degree distributions. These objects are generally constructed following the method developed by Molloy and Reed [22]. For a network with N nodes, the first ingredient in the Molloy-Reed recipe consists in the creation of a degree sequence composed of random integer numbers extracted from an *a priori* fixed degree distribution, which, in the case of scale-free networks, consists of a list of random variables extracted from a power-law PDF defined over the interval $[x_m, x_M]$, with x_M generally

set equal to $N - 1$. As is well known, many fundamental results for random scale-free networks rely on the fact that the second moment of the degree distribution is divergent for any $\lambda \leq 3$ [5,23,24]. Examples include the percolation threshold [25], the epidemic threshold [26], and the consensus time in the voter model [27]. As Fig. 2 shows, however, the expected divergence of the moments of the degree distribution is numerically satisfied only up to values of $N \simeq \tilde{x}$, as predicted in Eq. (6). Results obtained with different PRNGs confirm that the precise value of the cutoffs can be machine- and implementation-dependent (see Fig. 1 in the Supplemental Material [21]).

Our second example regards the statistics of the sum of power-law random variables $S = \sum_{i=1}^T x_i$. This is a quantity of interest in many practical situations. For example, the total deformation of rock materials or of the Earth's surface due to earthquakes may be modeled as the sum of seismic moment tensors [28,29], the latter obeying the Pareto distribution; cumulative economic losses or casualties due to natural catastrophes are modeled as sums of power-law distributed variables [30]; the total payoff of an insurance company is modeled as the sum of individual payoffs, each of which is distributed according to a power law [31]. Also, the sum of random power-law variates plays an important role in superdiffusive processes modeled by Lévy flights, i.e., a special class of random walks in which the step lengths obey a power-law probability distribution [32]. Numerical simulations of Lévy flights are used in several contexts: for example, in the study of movements of animals and humans [6,15] and stochastic models of physical systems [33]. Theory predicts two different behaviors for the sum of power-law random variables: for values of the power-law exponent $\lambda > 3$, the PDF $P(x)$ has finite variance, the central limit theorem applies, and the distribution $P(S)$ of the sum S approaches a normal distribution as the number of summands T grows to infinity; for $\lambda \leq 3$ instead, $P(x)$ has diverging variance, the generalized central limit theorem applies, and $P(S)$ approaches a so-called α -stable distribution as T grows [34,35]. The presence of the upper bound in the distribution $P(x)$ induces, however, catastrophic deviations from the expected behavior (see Fig. 2 in the Supplemental Material [21]). For any value of the exponent λ there exists a maximal number of summands \tilde{T} after which the distribution $P(S)$ starts to show features typical of a normal distribution: coefficient of variation, skewness, and excess kurtosis decrease to zero as the number of summands increases. This essentially means that if the number of summands is large enough, the distribution $P(S)$ starts to become peaked and symmetric around a given value, as expected for a normal distribution. On the basis of the value of the machine-dependent cutoff of Eq. (6), we should expect $P(S)$ to approach a normal distribution very slowly [36]. Our results, however, suggest that $P(S)$ enters in a regime of “normality” for small values of the number of summands. For $B_r = 32$, for example, the distribution starts to approach a normal distribution at $\tilde{T} \simeq 10^4$ when $\lambda = 1.5$, and only $\tilde{T} \simeq 10^3$ for $\lambda = 2.5$.

One may argue that the reason for the machine-dependent limitations illustrated so far resides only in the finite precision of the PRNG, so that the solution to this problem would be just

to use a PRNG that can produce a number of random bits B_r sufficiently large. Unfortunately, the solution is not as simple. When $B_r \geq 32$, as in the case of the orange points of Fig. 2, an additional machine-dependent effect may appear, namely the finite precision in the representation of numbers in our machine that affects the computation of powers. This second limitation approximately arises when we reach the so-called machine- ϵ or unit roundoff of our machine, i.e., the bits used in the so-called mantissa of the floating point number representation in the machine [37]. Under the hypothesis of having a PRNG that can produce a number of random bits $B_r > B_\epsilon$, the previous machine-dependent thresholds of Eqs. (4), (5), and (6) are still valid by substituting B_r with B_ϵ . In our implementation, we have $B_\epsilon = 52$. This is, however, a limit that can be reached only if the PRNG is truly able to generate a sufficiently large number of random bits. The results of our simulations instead indicate that strong numerical inaccuracies arise earlier.

To summarize, the current implementations of the inversion method are not suitable for the generation of synthetic random variates obeying power-law distributions. Problems arise for the violation of two basic principles at the basis of the theory of this method: (i) there exists a perfect uniform random variate generator; (ii) computers can store and manipulate real numbers. The same principles are violated for any other method developed for the computer-generation of nonuniform random variates, so that we expect to see similar problems also for other popular algorithms such as the rejection or the acceptance-complement methods [19]. In this paper, we explicitly considered the case of continuous random variates, but our results extend also to discrete variables. In addition, the same machine-dependent limitations hold for other PDFs rather than power-law distributions. For instance, exponential distributions are subjected to even stronger discretization effects. On the other hand, exponential PDFs have finite

moments, and the presence of the machine-dependent cutoff \bar{x} causes errors not as dramatic as in the case of power laws or other heavy-tailed distributions.

The practical consequences of the existence of machine-dependent distortions in the computer generation of power-law random variates are diverse. For example, the discretization of random variables in the tail may be crucial for the outcome of tests of statistical significance of empirical data. The presence of a finite cutoff is certainly more dangerous. Finite-size scaling analyses, for example, that do not account for it are all potentially subjected to uncontrollable scaling factors. Even more seriously, in predictive analyses there is the concrete risk of underestimations of extreme events. Last but not least, since the limitations are machine-dependent, neglecting their presence may cause problems of reproducibility of numerical experiments on different computers. The general and counterintuitive message of our analysis is that increasing the size of the sample does not improve the statistics, since in computer simulations fundamental properties of power-law distributions are preserved only up to relatively small sample sizes. With the current methodology, the only way to avoid all the numerical problems illustrated in this paper is to impose a forced cutoff at x^* as defined in Eq. (4), reducing, however, the support of the PDF to incredibly small ranges. Future research must thus work on finding algorithms that do not suffer from this serious problem. In addition to more radical solutions based on the use of arbitrarily long numbers of bits in the representation of both uniform random numbers and powers, we believe that effective approaches could be based on generative models of power-law distributions [13], features of dynamical systems [38], or they may rely on the scale invariance property of powers.

The author thanks C. Castellano and A. Flammini for comments on the manuscript.

-
- [1] H. Moore, *Annals Am. Acad. Polit. Social Sci.* **9**, 128 (1897).
 - [2] G. Zipf, *Human Behaviour and the Principle of Least-Effort* (Addison-Wesley, Cambridge, MA, 1949).
 - [3] H. E. Stanley, *Introduction to Phase Transitions and Critical Phenomena* (Oxford University Press, New York, 1987).
 - [4] B. Mandelbrot, *The Fractal Geometry of Nature* (Freeman, San Francisco, 1982).
 - [5] A.-L. Barabási and R. Albert, *Science* **286**, 509 (1999).
 - [6] G. M. Viswanathan, V. Afanasyev, S. V. Buldyrev, E. J. Murphy, P. A. Prince, and H. E. Stanley, *Nature (London)* **381**, 413 (1996).
 - [7] B. Gutenberg and C. F. Richter, *Frequency of Earthquakes in California*, Bulletin of the Seismological Society of America (Seismological Society of America, Pasadena, 1944).
 - [8] G. Boffetta, V. Carbone, P. Giuliani, P. Veltri, and A. Vulpiani, *Phys. Rev. Lett.* **83**, 4662 (1999).
 - [9] R. N. Mantegna and H. E. Stanley, *Nature (London)* **376**, 46 (1995).
 - [10] W. J. Reed, *Econ. Lett.* **74**, 15 (2001).
 - [11] X. Gabaix, Tech. Rep., National Bureau of Economic Research (2008), <http://pages.stern.nyu.edu/~xgabaix/papers/pl-ar.pdf>.
 - [12] M. Faloutsos, P. Faloutsos, and C. Faloutsos, *Comput. Commun. Rev.* **29**, 251 (1999).
 - [13] M. Mitzenmacher, *Internet Math.* **1**, 226 (2004).
 - [14] D. J. de Solla Price, *Science* **149**, 510 (1965).
 - [15] D. Brockmann, L. Hufnagel, and T. Geisel, *Nature (London)* **439**, 462 (2006).
 - [16] A. Clauset, C. R. Shalizi, and M. E. Newman, *SIAM Rev.* **51**, 661 (2009).
 - [17] G. Wergen, D. Volovik, S. Redner, and J. Krug, *Phys. Rev. Lett.* **109**, 164102 (2012).
 - [18] Y. Edery, A. B. Kostinski, S. N. Majumdar, and B. Berkowitz, *Phys. Rev. Lett.* **110**, 180602 (2013).
 - [19] L. Devroye, *Non-Uniform Random Variate Generation* (Springer-Verlag, Berlin, 1986).
 - [20] M. Matsumoto and T. Nishimura, *ACM Trans. Model. Comput. Simul. (TOMACS)* **8**, 3 (1998).
 - [21] See Supplemental Material at <http://link.aps.org/supplemental/10.1103/PhysRevE.90.050801> for details of the simulations and additional numerical results.
 - [22] M. Molloy and B. Reed, *Random Struct. Algorithms* **6**, 161 (1995).

- [23] R. Albert and A.-L. Barabási, *Rev. Mod. Phys.* **74**, 47 (2002).
- [24] S. N. Dorogovtsev, A. V. Goltsev, and J. F. Mendes, *Rev. Mod. Phys.* **80**, 1275 (2008).
- [25] R. Cohen, K. Erez, D. ben-Avraham, and S. Havlin, *Phys. Rev. Lett.* **85**, 4626 (2000).
- [26] R. Pastor-Satorras and A. Vespignani, *Phys. Rev. Lett.* **86**, 3200 (2001).
- [27] V. Sood and S. Redner, *Phys. Rev. Lett.* **94**, 178701 (2005).
- [28] O. Sotolongo-Costa, J. Antoranz, A. Posadas, F. Vidal, and A. Vazquez, *Geophys. Res. Lett.* **27**, 1965 (2000).
- [29] Y. Y. Kagan, *Geophys. J. Int.* **149**, 731 (2002).
- [30] J. Voit, *The Statistical Mechanics of Financial Markets* (Springer, Berlin, 2003), Vol. 2.
- [31] Y. Y. Kagan, *Commun. Statist. Stochastic Models* **13**, 775 (1997).
- [32] M. F. Shlesinger, G. M. Zaslavsky, and J. Klafter, *Nature (London)* **363**, 31 (1993).
- [33] M. F. Shlesinger, G. M. Zaslavsky, and U. Frisch (eds.), *Lévy Flights and Related Topics in Physics*, Lecture Notes in Physics, Vol. 450 (Springer, Berlin, Heidelberg, 1995).
- [34] J. Nolan, *Stable Distributions: Models for Heavy-tailed Data* (Birkhauser, Boston, 2003).
- [35] I. Zaliapin, Y. Y. Kagan, and F. P. Schoenberg, *Pure Appl. Geophys.* **162**, 1187 (2005).
- [36] R. N. Mantegna and H. E. Stanley, *Phys. Rev. Lett.* **73**, 2946 (1994).
- [37] D. Goldberg, *ACM Comput. Surv.* **23**, 5 (1991).
- [38] L. Palatella and P. Grigolini, *Physica A* **391**, 5900 (2012).