Journal of Computational Science 12 (2016) 77–82

Contents lists available at ScienceDirect

Journal of Computational Science

journal homepage: www.elsevier.com/locate/jocs

Efficient computational testing of scale-free behavior in real-world systems

Guannan Zhao[a], Zhenyuan Zhao[b], Neil F. Johnson[c,*]

[a] Bayview Asset Management LLC, Coral Gables, FL 33146, USA
[b] Smartleaf Inc., Cambridge, MA 02139, USA
[c] Physics Department, University of Miami, Coral Gables, FL 33124, USA

ARTICLE INFO

Article history:
Received 1 July 2015
Received in revised form
18 November 2015
Accepted 25 November 2015
Available online 30 November 2015

Keywords:
Collective behavior
Complexity
Numerical analysis

ABSTRACT

With big data becoming available across the physical, life and social sciences, researchers are turning their attention to the question of whether universal statistical signatures emerge across systems. Power-laws are a particularly potent example, since they indicate scale-free or scale invariant behavior and are observed in physical systems near phase transitions. However, the same scale-free property that enables them to unify behaviors across multiple spatiotemporal scales, also means that usual Gaussian-based approaches cannot be used to test their presence. Here we analyze the crucial question of how to implement a power-law test efficiently, given that a key part involves multiple Monte Carlo simulations to obtain an accurate statistical *p*-value. We present such a computational scheme in detail.

1. Introduction

Researchers from all disciplines are now experiencing rapidly increasing availability of 'big-data' across multiple systems of interest, whether social, economic, medical, biological or physical in nature [1–7]. Not only is the precision of such data increasing, but also the spatial and temporal resolution, e.g. price changes at resolutions down to the millisecond from a financial market [8–10] or number of casualties with resolution down to individual intraday events from a human conflict [4,11]. One approach to analyzing such data, favored in particular within the physics community, is to look for possible universality of signals in the data since this may indicate universality in the system's underlying mechanistic dynamics – and hence ultimately, a simplification of real-world complex systems into a finite number of classes of behavior. This approach has been hugely successful in Physics with the study of critical phenomena and phase transitions, dramatically reducing the complex behaviors observed across disparate materials down to a few universality classes [12]. The new wave of big data in other disciplines now allows researchers to address whether such universality can extend to systems involving, for example, humans, and hence act as generic scientific signatures of behavior in living systems [3].

Though still a research question in progress, particular universalities have already been identified, including in human systems (see for example, Refs. [3,7,13–16]). The universal signatures in question concern a particular type of fat-tailed distribution $p(x)$ called a power-law [3,7,13–16,18,19]. As shown in Fig. 1(a) and (b), a power-law differs from the well-known Gaussian (Normal or bell-curve) distribution in that it has an algebraic tail at high x which, according to the value of the exponent α, can lead to a mean and/or standard deviation that is in principle infinite. This means that any theories of that system built around average behavior with fluctuations given by a Gaussian-fit standard deviation, will be inaccurate. The reason why it arises so much in physical systems, is that phase transitions – which are highly important collective behaviors for a system – have scaling properties that involve power-laws [12]. Suppose a distribution $p(x)$ of, say, cluster sizes near a phase transition follows the form $p(x) \sim x^{-\alpha}$ in the high x limit $(x > x_{\min})$ then re-scaling $x \to 2x = x'$ yields $p(2x) \sim (2x)^{-\alpha} \propto x^{-\alpha}$ which means that the distribution is invariant under a change of scale. This is called *scale-free* or *scale invariant* and means that theories of the system can be built which apply across all scales, as opposed to having to describe the system at every different length or time scale and then couple these different descriptions together.

Having explained the unique importance of a power-law distribution, and hence of identifying a power-law distribution in data, we turn to the focus of this paper: how can this power-law testing

[*] Corresponding author.
E-mail addresses: GuannanZhao@bayviewassetmanagement.com (G. Zhao), njohnson@physics.miami.edu (N.F. Johnson).

http://dx.doi.org/10.1016/j.jocs.2015.11.009
1877-7503/© 2015 Elsevier B.V. All rights reserved.

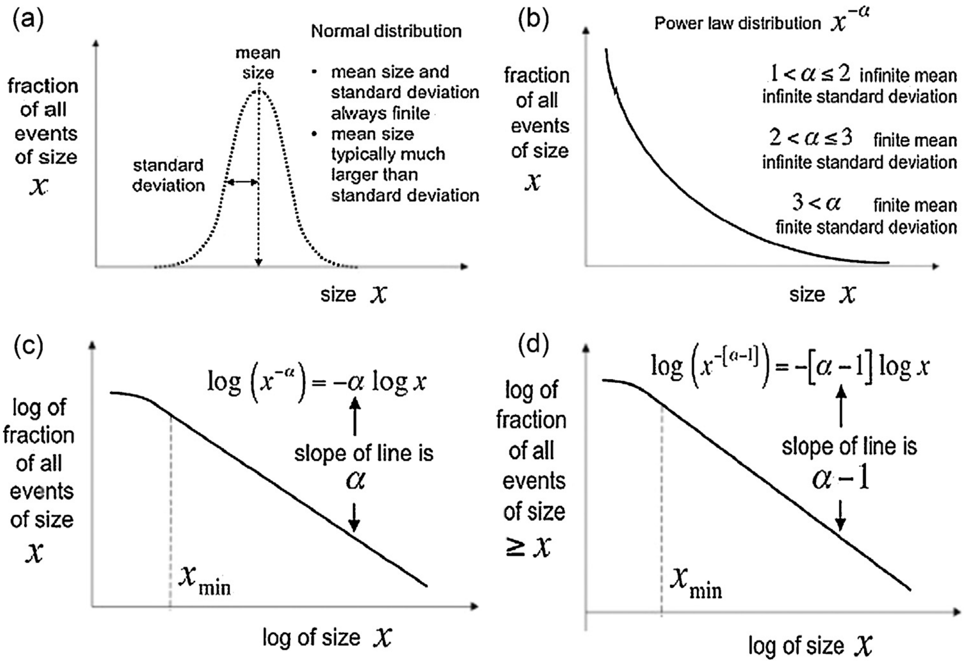
(a)
fraction of all events of size x
Normal distribution
mean size
• mean size and standard deviation always finite
• mean size typically much larger than standard deviation
standard deviation
size x
(b)
fraction of all events of size x
Power law distribution $x^{-\alpha}$
$1<\alpha\leq 2$ infinite mean infinite standard deviation
$2<\alpha\leq 3$ finite mean infinite standard deviation
$3<\alpha$ finite mean finite standard deviation
size x
(c)
log of fraction of all events of size x
$\log(x^{-\alpha}) = -\alpha \log x$
slope of line is α
$x_{\min}$
log of size x
(d)
log of fraction of all events of size $\geq x$
$\log(x^{-[\alpha-1]}) = -[\alpha-1]\log x$
slope of line is $\alpha-1$
$x_{\min}$
log of size x

Fig. 1. Schematic comparison between traditional Gaussian distribution (i.e. normal or bell-curve) in (a) and power-law in (b), for data analysis. (c) Shows the log–log plot of a power-law. (d) Shows a log–log plot of the complementary cumulative distribution which is typically used since it is a less noisy than $p(x)$.

be done in a computationally efficient way? One might think that since taking logarithms of $p(x) \sim x^{-\alpha}$ yields a linear relationship log $p(x) = -\alpha \log x + \text{constant}$, the task couldn't be easier as illustrated in Fig. 1(c) for $p(x)$, or Fig. 1(d) for the cumulative distribution which is typically preferred since it reduces noise in $p(x)$. However this would be a mistake, since such a linear regression has been shown to be a biased estimator [19].

This paper focuses on implementing the widely utilized power-law testing procedure of Ref. [19] which avoids this problem of biased estimation. We focus on presenting a particular procedure for efficient computational implementation – specifically, evaluation of the statistical p-values involving Monte-Carlo simulations. While we do not worry more about discussing real-world data, Fig. 2 illustrates the results that emerge from direct application of the procedure in this paper. The estimate of α is near $2.5 \equiv 5/2$ [4,11]. This illustrates the potentially broader universality discussed above, since 5/2 has also been reported as the distribution of stock transaction sizes, indicating that herd sizes of similar-minded traders is power-law with that exponent [9]; and also the size distribution of neuronal avalanches, given avalanche initiation by a randomly chosen neuron (i.e. $k \cdot k^{-5/2} \equiv k^{-3/2}$ [20]); and even the size distribution of pockets of superconducting coherence in fragmented materials [21]. Interestingly, the value 5/2 is also very close to the 2.3 values reported for the size distribution of gangs in Chicago, and gangs in Manchoukuo in 1935 [22].

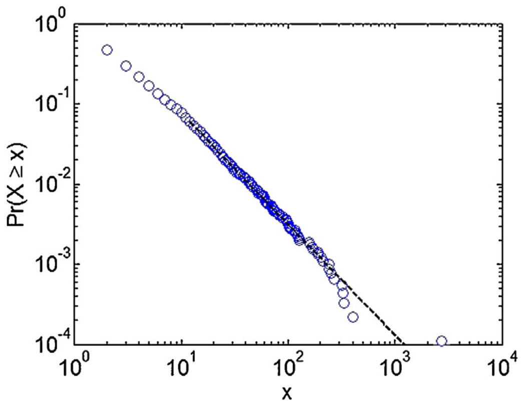

Fig. 2. Example illustrating the complementary cumulative distribution shown in Fig. 1(d), for empirical casualty data drawn from more than 9000 events in the recent war in Afghanistan [4]. Each event is a clash producing x casualties and hence having size x. It is best fit by a discrete power-law with $\alpha = 2.37$, $x_{\min} = 12$ and $p = 0.647$. Solid lines represent the best fit to the data using the methods described in the text. These methods produce an unbiased estimate, whereas a simple linear regression is biased [19].

2. Background

We start by giving the necessary preliminary information, following the notation of Ref. [19].

2.1. Power law p(x): continuous vs. discrete

Data is often discrete in nature, i.e. x is a discrete number (integer), for example the number of casualties in a given event. However for completeness, we analyze the two cases of discrete and continuous x. For the continuous case of a power-law distribution, when $\alpha > 1$, it is straightforward to show that:

$$p(x) = \frac{\alpha - 1}{x_{\min}} \left(\frac{x}{x_{\min}} \right)^{-\alpha} \tag{1}$$

In the discrete case, we have:

$$p(x) = \frac{x^{-\alpha}}{\zeta(\alpha, x_{\min})} \tag{2}$$

where

$$\zeta(\alpha, x_{\min}) = \sum_{n=0}^{\infty} (n + x_{\min})^{-\alpha} \tag{3}$$