PLOS | ONE

RESEARCH ARTICLE

Fitting power-laws in empirical data with estimators that work for all exponents

Rudolf Hanel[1¤], Bernat Corominas-Murtra[1¤], Bo Liu[1¤], Stefan Thurner[1,2,3,4¤]*

1 Section for Science of Complex Systems, Medical University of Vienna, Spitalgasse 23, 1090 Vienna, Austria, **2** Santa Fe Institute, 1399 Hyde Park Road, Santa Fe, NM 87501, United States of America, **3** IIASA, Schlossplatz 1, 2361 Laxenburg, Austria, **4** Complexity Science Hub Vienna, Josefstädterstrasse 39, A-1090 Vienna, Austria

¤ Current address: Section for Science of Complex Systems, CeMSIIS, Medical University of Vienna, Spitalgasse 23, Bauteil 86, A-1090, Vienna, Austria
* stefan.thurner@meduniwien.ac.at

OPEN ACCESS

Citation: Hanel R, Corominas-Murtra B, Liu B, Thurner S (2017) Fitting power-laws in empirical data with estimators that work for all exponents. PLoS ONE 12(2): e0170920. doi:10.1371/journal.pone.0170920

Editor: Eduardo G. Altmann, University of Sydney, AUSTRALIA

Received: September 17, 2016

Accepted: January 12, 2017

Published: February 28, 2017

Data Availability Statement: All relevant data are within the paper and its Supporting Information files.

Funding: This work was supported in part by the Austrian Science Foundation “Fonds zur Förderung der wissenschaftlichen Forschung” (FWF: https://www.fwf.ac.at/) under grant P29252. BL is grateful for the support by the China Scholarship Council, file-number 201306230096. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Abstract

Most standard methods based on maximum likelihood (ML) estimates of power-law exponents can only be reliably used to identify exponents smaller than minus one. The argument that power laws are otherwise not normalizable, depends on the underlying sample space the data is drawn from, and is true only for sample spaces that are unbounded from above. Power-laws obtained from bounded sample spaces (as is the case for practically all data related problems) are always free of such limitations and maximum likelihood estimates can be obtained for arbitrary powers without restrictions. Here we first derive the appropriate ML estimator for arbitrary exponents of power-law distributions on bounded discrete sample spaces. We then show that an almost identical estimator also works perfectly for continuous data. We implemented this ML estimator and discuss its performance with previous attempts. We present a general recipe of how to use these estimators and present the associated computer codes.

Introduction

The omnipresence of power-laws in natural, socio-economic, technical, and living systems has triggered immense research activity to understand their origins. It has become clear in the past decades that there exist several distinct ways to generate power-laws (or asymptotic power-laws), for an overview see for example [1, 2]. In short, power-laws of the form

$$p(x) = Cx^{-\lambda} \quad , \tag{1}$$

arise in critical phenomena [3, 4], in systems displaying self-organized criticality [5], preferential attachment type of processes [6–9], multiplicative processes with constraints [10], systems described by generalized entropies [11, 12], or sample space reducing processes [13], i.e. processes that reduce the number of possible outcomes (sample space) as they unfold. Literally thousands of physical, natural, man-made, social, and cultural processes exhibit power-laws, the most famous being earthquake magnitudes [14, 15], city sizes [16, 17], foraging and distribution pattern of various animal species [18], evolutionary extinction events [19], or the frequency of word occurrences in languages, known as Zipf's law [20].

Competing interests: The authors have declared that no competing interests exist.

It is obvious that estimating power-law exponents from data is a task that sometimes should be done with high precision. For example if one wants to determine the universality class a given process belongs to, or when one estimates probabilities of extreme events. In such situations small errors in the estimation of exponents may lead to dramatically wrong predictions with potentially serious consequences.

Estimating power-law exponents from data is not an entirely trivial task. Many reported power-laws are simply not exact power-laws, but follow other distribution functions. Such partial or imperfect power-laws come in various flavours: power-laws with an exponential cut-off, asymptotic power-laws, q-exponential distributions, which approach exact power-laws only in the tail of the distribution function. Other families of distribution functions may possess one or more regions where a power-law fit may be reasonable. Strategies for identifying such regions in imperfect power-law distributions may depend on the scenario. In some scenarios, e.g. for q-exponential distribution functions, ML algorithms that are specifically tailored to fit the particular class of distribution functions, may be considered. When no parametric representation of an imperfect power-law distribution is available, other strategies may be followed. For instance, screening the data for ranges where the null-hypothesis, that the data has been sampled from an exact power-law, can not be rejected for a particular significance level. Despite the importance of developing adequate methods for distinguishing real power-laws from alternative hypotheses, we will, except for some remarks, not address this issue in depth here, since good standard literature on the topic of Bayesian *alternative hypotheses testing* exists, see for example [21, 22]. For power-laws some of these matters have been discussed also in [23]. In particular, the possibility to use the Kolmogorov-Smirnov (KS) goodness of fit test (GOF) [24, 25] for finding optimal fitting ranges. In which sense the KS GOF-test is adequate in this context and how so called p-values of the KS GOF-test needs to be interpreted is a non-trivial question. The p-value of the KS GOF-test does not directly correspond to the rate of falsely rejected power-laws. The two-sided KS GOF-test in fact rejects an unacceptably large percentage of samples drawn from exact power-laws. What the two-sided KS GOF-test in fact rejects or accepts at a certain confidence level, is that the power-law with the estimated exponent represent the same distribution function as the data, which has been used to estimate the exponent. For more details on how to calibrate critical values of the statistical parameter used in the KS GOF-test in order to control the false rejection rate of power-laws with respect to the estimator, see S3 File APPENDIX C.

Here we simply focus on estimating power-law exponents from data on a sound probabilistic basis, using a classic Bayesian parameter estimation approach, see e.g. [26, 27], that provides us with *maximum likelihood* (ML) estimators for estimating power-law exponents over the full range of reasonably accessible values. Other approaches restrict the limit of application to power-laws with exponents smaller than -1 [23, 28, 29]. Having such estimators without any a priori restriction of the exponent value is of particular interest for a large classes of situations where exponents close to $\lambda \sim 1$ appear (Zipf's law). We will argue here that whenever dealing with data we can assume discrete and bounded samples spaces (domains), which guarantees that power-laws are normalizable for arbitrary powers λ. We then show that the corresponding ML estimator can then also be used to estimate exponents from data that is sampled from continuous sample spaces, or from sample spaces that are not bounded from above. Our approach is equivalent to the one provided in [27].

Questions before fitting power-laws

In physics the theoretical understanding of a process sometimes provides us with the luxury of knowing the exact form of the distribution function that one has to fit to the data. For instance think of critical phenomena such as Ising magnets in 2 dimensions at the critical temperature, where it is understood that the susceptibility follows a power-law of the form $(T - T_c)^{-\gamma}$, with γ a critical exponent, that occasionally even can be predicted mathematically. However, often—and especially when dealing with complex systems—we do not enjoy this luxury and usually do not know the exact functions to fit to the data.

In such a case, let us imagine that you have a data set and from first inspection you think that a power-law fit could be a reasonable thing to do. It is then essential, before starting with the fitting procedures, to clarify what one knows about the process that generated this data. The following questions may help to do so.

- Do you have information about the dynamics of the process that is generating what appears to be a power-law?
- Is the data generated by a Bernoulli process (e.g. tossing dice), or not (e.g. preferential attachment)?
- Is the data available as a collection of samples (a list of measurements), or only coarse-grained in form of a histogram (binned or aggregated data).
- Is the data sampled from a discrete (e.g. text) or continuous sample space (e.g. earthquakes)?
- Does the data have a natural ordering (e.g. magnitudes of earthquakes), or not (e.g. word frequencies in texts)?

The decisions one has to take before starting to estimate power-law exponents are shown as a decision-tree in (Fig 1). If it is known that the process generating the data is not a Bernoulli process (for example if the process belongs to the family of history dependent processes such as e.g. preferential attachment), then one has the chance to use this information for deriving parameter estimators that are tailored exactly for the particular family of processes. If no such detailed information is available one can only treat the process as if it were a Bernoulli process, i.e. information about correlations between samples is ignored. If we know (or assume) that the data generation process is a Bernoulli process, the next thing to determine is whether the data is available as a collection of data points, or merely as coarse grained information in form of a histogram that collects distinct events into bins (e.g. histograms of logarithmically binned data).

If data is available in form of a data set of samples (not binned), a surprisingly general maximum likelihood (ML) estimator can be used to predict the exponent of an underlying power-law $p(x) \propto x^{-\lambda}$. This estimator that we refer to as ML*, will be derived in the main section. Its estimates for the underlying exponent λ, are denoted by λ^*. The code for the corresponding algorithm we refer to as `r_plfit`. If information is available in form of a histogram of binned data, a different estimator becomes necessary. The corresponding algorithm (`r_plhistfit`) is discussed in S1 File APPENDIX A and in the section below on discrete and continuous sample spaces. Both algorithms are available as matlab code [30]. For how to use these algorithms, see S2 File APPENDIX B.

If we have a dataset of samples (not binned), so that the `r_plfit` algorithm can be used, it still has to be clarified whether the data has a natural order or not? Numerical observables such as earthquake magnitudes are *naturally* ordered. One earthquake is always stronger or smaller than the other. If observables are non-numeric, such as word types in a text, then a natural order can not be known *a priori*. The natural order can only be inferred approximately by using so-called *rank-ordering*; or alternatively—by using the so-called *frequency distribution* of the data. Details are discussed below in the section on rank-order, frequency distributions, and natural order.

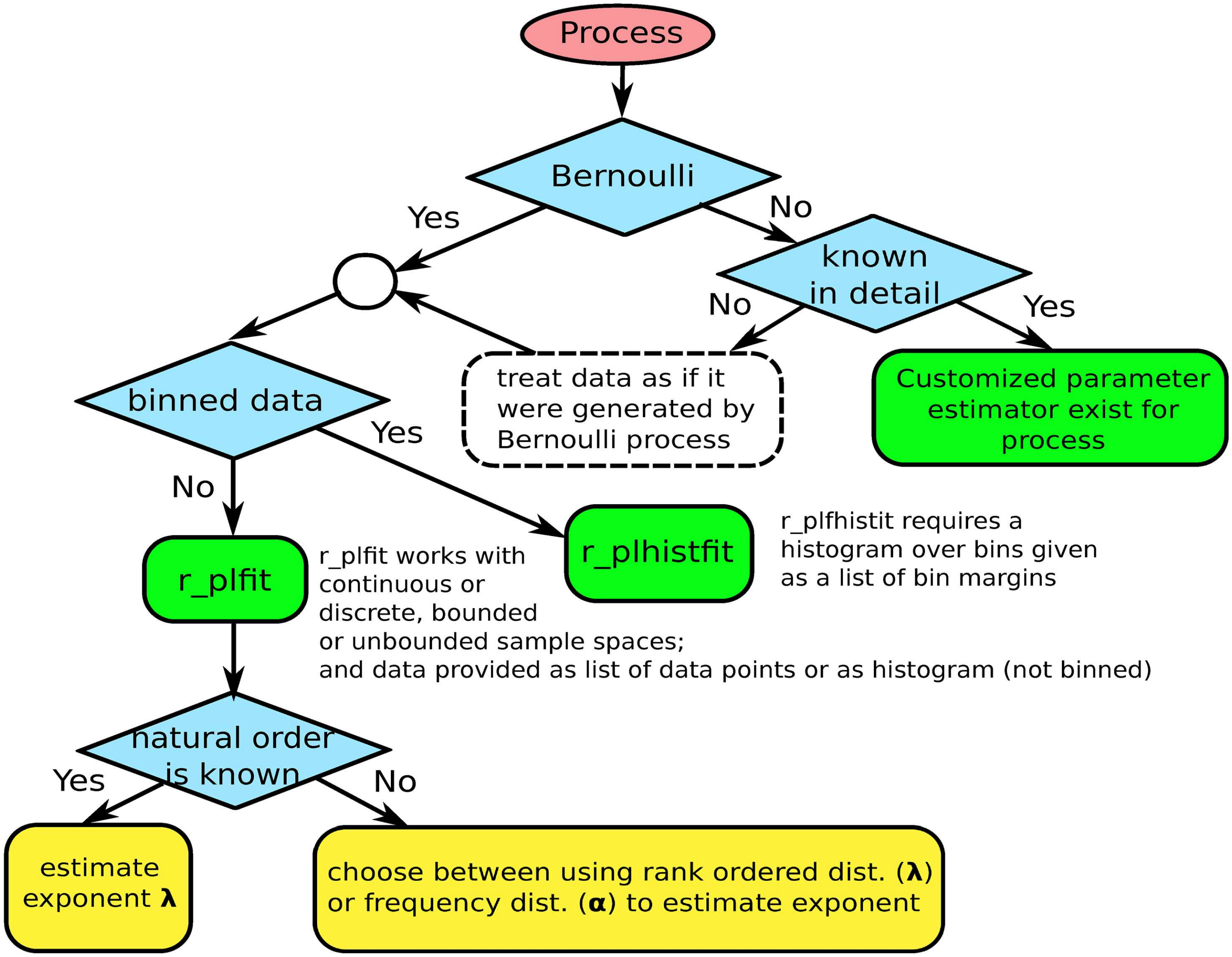
Process
Bernoulli
Yes
No
known in detail
No
Yes
treat data as if it were generated by Bernoulli process
Customized parameter estimator exist for process
binned data
Yes
No
r_plfit
r_plhistfit
r_plfit works with continuous or discrete, bounded or unbounded sample spaces; and data provided as list of data points or as histogram (not binned)
r_plfhistit requires a histogram over bins given as a list of bin margins
natural order is known
Yes
No
estimate exponent λ
choose between using rank ordered dist. (λ) or frequency dist. (α) to estimate exponent

Fig 1. Decision tree of questions that should be clarified before estimating power-law exponents from data. The tree shows under which conditions the fitting algorithms developed in this paper r_plfit and r_plhistfit can be used.

doi:10.1371/journal.pone.0170920.g001

Other issues to clarify are to see if a given sample space is continuous or discrete, and if the sample space is bounded or unbounded. These questions however, turn out to be not critical. One might immediately argue that for unbounded power-law distribution functions normalization becomes an issue for exponents $\lambda \leq 1$. However, this is only true for Bernoulli processes on *unbounded* sample spaces. Since all real-world data sets are collections of finite discrete values one never has to actually deal with normalization problems. Moreover, since most experiments are performed with apparati with finite resolution, most data can be treated as being sampled from a bounded, discrete sample space, or as binned data. For truly continuous processes the probability of two sampled values being identical is zero. Therefore, data sampled from continuous distributions can be recognized by sample values that are unique in a data set. See S1 File APPENDIX A for more details.

Statistically sound ways to fit power-laws were advocated and discussed in [23, 26–29, 31, 32]. They overcome intrinsic limitations of the *least square* (LS) fits to logarithmically scaled data, which were and are widely (and often naively) used for estimating exponents. The ML estimator that was presented in [23] we refer to as the $\mathrm{ML}_{\mathrm{CSN}}$ (for Clauset-Shalizi-Newman) estimator; its estimates for the exponent we denote by $\hat{\lambda}$. The approach that leads to $\mathrm{ML}_{\mathrm{CSN}}$ focuses on continuous data x that follows a power-law distribution from Eq (1), and that is bounded from below $x > x_{\min} > 0$ but is not bounded from above (i.e. $x_{\max} > x$ with $x_{\max} = \infty$). In [23] emphasis is put on how ML estimators can be used to infer whether an observed distribution function is likely to be a power-law or not. Also the pros and cons of using cumulative distribution functions for ML estimates are discussed, together with ways of treating discrete data as continuous data. For the continuous and unbounded case, simple explicit equations for the $\mathrm{ML}_{\mathrm{CSN}}$ estimator can be derived [23, 28]. The continuous approach however, even though it seemingly simplifies computations, introduces unnecessary self-imposed limitations with respect to the range of exponents that can be reliably estimated. $\mathrm{ML}_{\mathrm{CSN}}$ works very well for a range of exponents between −3.5 and −1.5, see Fig (3).

Here we show how to overcome these limitations—and by doing so extend the accessible range of exponents—by presenting the exact methodology for estimating λ for discrete bounded data with the estimator ML^*. While this approach appears to be more constrained than the continuous one we can show also theoretically that data from continuous and potentially unbounded sample spaces can be handled within essentially the same general ML framework as well. The key to the ML^* estimator is that it is not necessary to derive explicit equations for finding λ^*. Implicit equations in λ exist for power-law probability distributions over discrete or continuous sample spaces that are both bounded from below *and* above, see Eq (9), and also [27]. Solutions λ^* can be easily obtained numerically. An implementation of the respective algorithms can be found in [30], for a tutorial see S2 File APPENDIX B.

Rank-order, frequency distributions & natural order

There exist three distinct types of distribution functions that are of interest in the context of estimating power-law exponents:

1. The *probability distribution* $p(x)$ assigns a probability to every observable state-value x. Discrete and bounded sample spaces are characterized by W state-types $i = 1, \cdots, W$, with each type i being associated with a distinct value $x = z_i$.
2. The *relative frequencies*, $f_i = k_i/N$, where k_i is the number of times that state-type i is observed in N experiments. $k = (k_1, \cdots, k_W)$ is the *histogram* of the data. As explained below in detail, the relative frequencies can be ordered in two ways.
    - If f_i is ordered according to their descending magnitude this is called the *rank ordered* distribution.
    - If f_i is ordered according to the descending magnitude of the probability distribution $p(z_i)$, then they are *naturally ordered* relative frequencies.
3. The *frequency distribution* $\phi(n)$ counts how many state-types i fulfill the condition $k_i = n$.

In (Fig 2) we show these distribution functions. There $N = 10000$ data points are sampled from $x \in \{1, \cdots, 1000\}$, with probabilities $p(x) \propto x^{-0.7}$. The probability distribution is shown (red). The relative frequency distribution f is plotted in natural order (blue), the rank-ordered distribution is shown with the yellow line, which clearly exhibits an exponential decay towards the the tail. The inset shows the frequency distribution $\phi(n)$ of the same data. We next discuss how different sampling processes can be characterized in terms of natural order, rank-order, or frequency distributions.

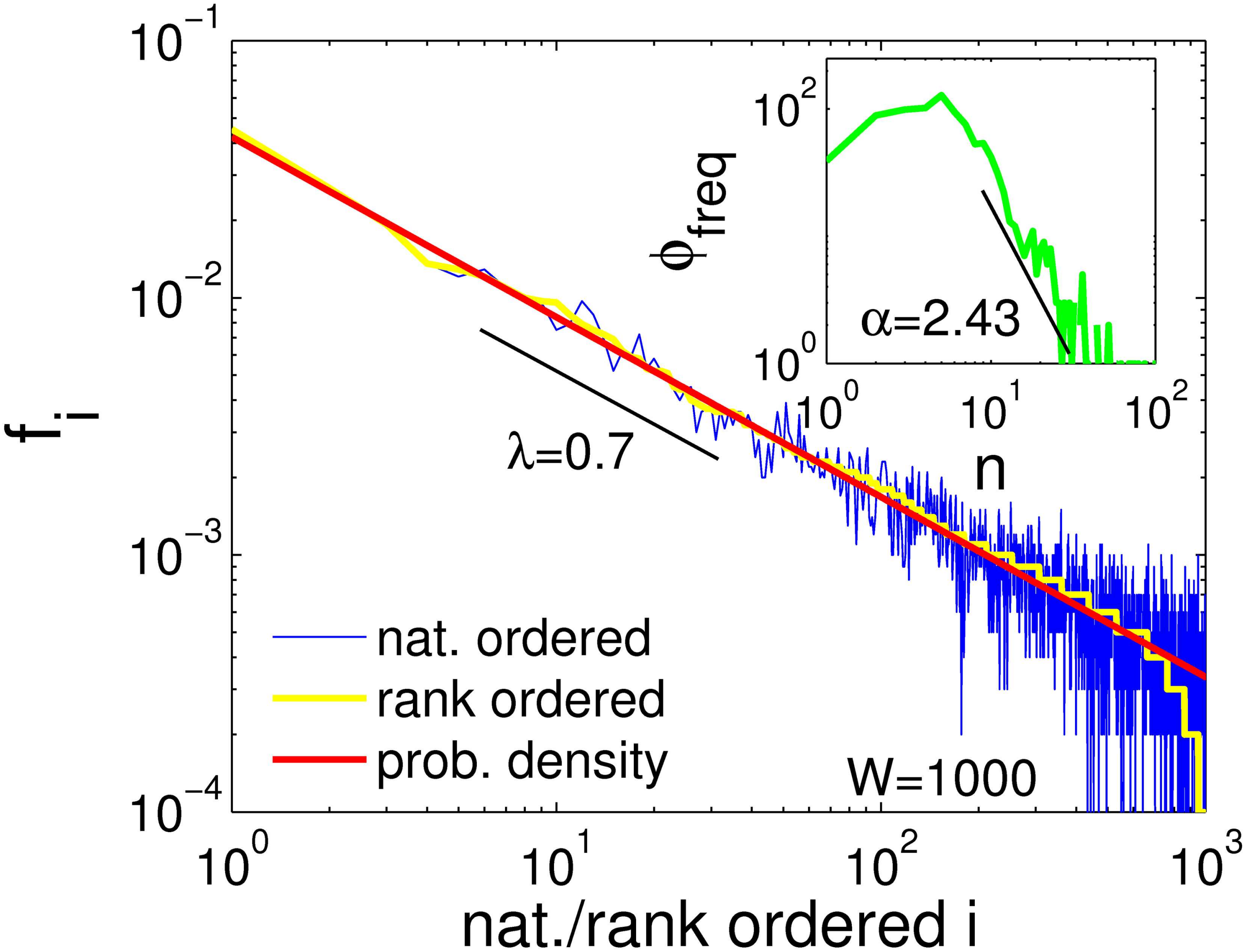

Fig 2. The four types of distribution functions. Data is sampled from a power-law distribution $p(x) \propto x^{-\lambda}$ with an exponent $\lambda = 0.7$ (red line). The relative frequencies f_i are shown for $N = 10000$ sampled data points according to their natural (prior) ordering that is associated with p (blue). The rank-ordered distribution (posterior) is shown in yellow, where states i are ordered according to their observed relative frequencies f_i. The rank-ordered distribution follows a power-law, except for the exponential decay that starts at rank ~ 500. A low frequency cut-off should be used to remove this part for estimating exponents. The inset shows the frequency distribution $\phi(n)$ that describes how many states x appear n times (green). The frequency distribution has a maximum and a power-law tail with exponent $\alpha = 1 + 1/\lambda \sim 2.43$. To estimate α, one should only consider the tail of the frequency distribution function.

doi:10.1371/journal.pone.0170920.g002

Processes with naturally ordered observables. For some sampling processes the ordering of the observed states is known. For example think of x representing the numerical values of earthquake magnitudes. Here any two observations x and x' can be ordered with respect to their numerical value, or their *natural order*. Since power-law distributions $p(x) \propto x^{-\lambda}$ are monotonic this is equivalent to ranking observations according to the probability distribution p they are sampled from: The most likely event has *natural rank* 1, the second most likely rank 2, etc. In other words, we can order state-types x in a way that over the sample space $\Omega = \{z_i | i = 1, \cdots, W\}$, $p = (p(z_1), \cdots, p(z_W))$ is a monotonic and decreasing function.

Processes with rank-ordered observables. If p is not known *a priori* because the state-types i have no numerical values z_i attached, as happens for example with words in a text, we can only count relative frequencies f_i (a normalized histogram) of states of type i, *a posteriori*, i.e. after sampling. To be clear, let $k = (k_1, \cdots, k_W)$ be the histogram of N recorded states. k_i is the number of times we observed type i, then $f_i = k_i/N$ is the relative frequency of observing states of type i. After all samples are taken, one can now order states with respect to f_i, such that the rank 1 is assigned to state i with the largest f_i, rank 2 to i' with the second largest $f_{i'}$, etc. $f = (f_1, \cdots, f_W)$ is called the *rank-ordered* distribution of the data.

The natural order imposed by p and the rank-order imposed by f are not identical for finite N. However, if data points have been sampled independently, then f converges toward p (for $N \to \infty$) and the rank-order induced by f will asymptotically approach the natural order induced by p. (This is true when data gets sample independently from a fixed distribution, but not in general. If the observable states are not bounded from above as the number of samples increases, then such processes showing emergent power-laws with exponents larger than -1 in rank, cannot be Bernoulli processes. In such cases ML estimators can be derived specifically for the particular sampling process, given the sampling process is known (compare the decision tree in (Fig 1)). If the sampling process is not known, as it often is the case, then one is left with the only option, to treat the data points as if they were sampled independently.) The highest uncertainty on estimating the order induced by p using f is associated with the least frequent observations. Therefore, when estimating exponents from rank-ordered distributions, one might consider to use a low-frequency cut-off to exclude infrequent data.

Frequency distributions. Exponents of power-laws can also be estimated from *frequency* distributions $\phi(n)$. These counts how many distinct state-types i occur exactly n times in the data. It does not depend on the natural (prior) order of states and therefore is sometimes preferred to the (posterior) rank-ordered distribution. However, complications may be encountered for naturally or rank ordered distributions as well as for frequency distributions. If we find a well defined power-like probability distribution $p \propto x^{-\lambda}$ (and asymptotically to f), then the associated frequency distribution $\phi(n)$ is not an exact power-law, but a non-monotonic distribution (with a maximum). Only the tail of the frequency distribution decays as a power-law, $\phi(n) \propto n^{-\alpha}$. Conversely, if the frequency distribution is a well defined power-law $\phi \propto n^{-\alpha}$, then the probability distribution p will not show a well defined power-law $p \propto x^{-\lambda}$ over the entire range of states x. As a consequence, one frequently needs to specify the data range for estimating the power-law exponent. The exponents λ and α are related through the well known equation [33]

$$\alpha = 1 + 1/\lambda \quad . \tag{2}$$

If the probability distribution has exponent λ, the tail of the associated frequency distribution has exponent α. Since the frequency distribution behaves like a power-law only in its tail, estimating α makes it necessary to constrain the observed data to large values of n. Note that this is equivalent to using a low-frequency cut-off. One option to do that is to derive a maximum entropy functional for $\phi(n)$ and fit the resulting (approximate) max-ent solution to the data. We do not follow this route here.

If the natural order of the data is known, one can directly use the natural ordered data in the ML estimates for the exponents. If it is not known, either the rank-ordered distribution can be used to estimate λ, or the frequency distribution to estimate α, see (Fig 1).

One might also estimate both, λ in the rank ordered distribution, and α in the frequency distribution of the data. Using Eq (2) to compare the two estimates may be used as a rough quality-check. If estimates do not reasonably coincide one should check whether the used data ranges have been appropriately chosen. If large discrepancies remain between α and $1 + 1/\lambda$ this might indicate that the observed distribution function in question is only an approximate power-law, for which Eq (2) need not hold. For a tutorial on how to use r_plfit to perform estimates see S2 File APPENDIX B.

Discrete and continuous sample spaces & normalization

Data can originate from continuous sample spaces $\Omega_c = [x_{\min}, x_{\max}]$, or discrete ones $\Omega_d = \{z_1, z_2, \cdots, z_W\}$. To each state-type $i = 1, \cdots, W$, there is assigned a state-value z_i. Whether a distribution function $p(x) = Z_\lambda^{-1} x^{-\lambda}$, with $x \in \Omega$, is normalizable or not, can only be decided once the sample space Ω has been specified. The normalization factors for continuous and discrete Ω are

$$\begin{aligned} Z_\lambda(\Omega_c) &\equiv \int_{x_{\min}}^{x_{\max}} dx\ x^{-\lambda} &= \frac{x_{\max}^{1-\lambda} - x_{\min}^{1-\lambda}}{1-\lambda} \\ Z_\lambda(\Omega_d) &\equiv \sum\nolimits_{x\in\Omega_d} x^{-\lambda} &= \sum\nolimits_{i=1}^{W} z_i^{-\lambda} \quad . \end{aligned} \tag{3}$$

For bounded sample spaces with $0 < x_{\min} \leq x \leq x_{\max} < \infty$, power-laws are always normalizable for arbitrary exponents λ, and a well defined ML estimator of λ^* exists (see below). The normalization constants in Eq (3) can be specified in r_plfit (see S2 File APPENDIX B).

Data sampled from a continuous sample space Ω_c can essentially be treated as if it were sampled from a discrete sample space Ω_d, where $x \in \Omega_d$ are given by the unique collection of distinct values in the data set. That is, the data set $x = (x_1, \cdots, x_N)$ contains N data points x_n (that have W unique values z_i, the states of type i) which we collect in the discrete sample space $\Omega = \{z_1, \cdots, z_W\}$. For truly continuous data we have $N = W$, since the probability of $x_m = x_n$ for $n \neq m$ is vanishing. As a consequence the histogram k_i, which counts the number of times z_i appears in the data, is essentially given by $k_i = 1$ for all $i = 1, \cdots, W$. This provides us with a practical criterion for when to use the normalization constant for discrete or continuous data. For details see S1 File APPENDIX A.

The equation for the ML estimator ML*, that yields the estimate λ^*, only requires the knowledge of the relative frequency distribution $f_i = k_i/N$ (in natural- or rank-order) of the observed state-types i, as we will see in Eq (9) below. Therefore r_plfit can work either with data sets x or histograms k over the unique values in the data sets. If data comes in coarse grained form, i.e. histograms, where each bin may contain a whole range of observable values x, then an estimator is required that is different from ML* [32], see also S1 File APPENDIX A. The corresponding code r_plhistfit can also be downloaded from [30].

The ML*-estimator for power-laws from discrete sample spaces

Consider a family of random processes Y that is characterized by the parameters $\theta = (\theta_1, \cdots, \theta_R)$. Let Y be defined on a discrete sample space $\Omega = \{z_1, z_2, \cdots, z_W\}$, with $0 < z_i < \infty$. The process Y samples values $x \in \Omega$ with probability,

$$p(x|\theta, \Omega) \quad . \tag{4}$$

Let us repeat the process Y in N independent experiments to obtain a data set $y = (y_1, \cdots, y_N)$. $k = (k_1, \cdots, k_W)$ is the histogram of the events recorded in y, i.e. k_i is the number of times z_i appears in y. Note that $\sum_{i=1}^{W} k_i = N$. As a consequence of independent sampling, the probability to sample exactly k is,

$$P(k|\theta, \Omega) = \binom{N}{k} \prod_{i=1}^{W} p(z_i|\theta, \Omega)^{k_i}, \tag{5}$$

where $\binom{N}{k} = N!/\prod_{i=1}^{W} k_i!$ is the multinomial factor. Bayes' formula allows us to get an estimator for the parameters θ,

$$P(\theta|k,\Omega) = P(k|\theta,\Omega)\frac{P(\theta|\Omega)}{P(k|\Omega)} \ . \tag{6}$$

Obviously, $P(k|\Omega) = \int d\theta\, P(k|\theta,\Omega)P(\theta|\Omega)$ does not depend on θ. Without further available information we must assume that the parameters θ are uniformly distributed between their upper and lower limits. As a consequence, $P(\theta|\Omega)$ also does not depend on θ within the limits of the parameter range and can be treated as a constant. (Unfortunately, what works for parameters in θ such as λ does not work for parameters such as $x_{\min}$ and $x_{\max}$. For those variables it turns out that $P(\theta|\Omega)$ can not be assumed to be constant between upper and lower bounds of the respective parameter values. Bayesian estimators for $x_{\min}$ and $x_{\max}$ require to explicitly consider a non-trivial function $P(\theta|\Omega)$. Though in principle feasible, we ignore the possibility of deriving Bayesian estimates for $x_{\min}$ and $x_{\max}$ in this paper.) From Eq (6) it follows that the value θ^* that maximizes $P(\theta|k,\Omega)$ also maximizes $P(k|\theta,\Omega)$. The most likely parameter values $\theta^* = (\theta_1, \cdots, \theta_R)^*$ are now found by maximizing the log-likelihood,

$$0 \ = \ \frac{\partial}{\partial\theta_r}\frac{1}{N}\log P(\theta|k,\Omega) = \sum\nolimits_{i=1}^{W} f_i \frac{\partial}{\partial\theta_r}\log p(z_i|\theta,\Omega) = -\frac{\partial}{\partial\theta_r} H_{\text{cross}}(f||p(z|\theta,\Omega)) \quad , \tag{7}$$

for all parameters $r = 1, \cdots, R$. Here $H_{\text{cross}}(f||p(z|\theta,\Omega)) \equiv -\sum_{i=1}^{W} f_i \log p(z_i|\lambda,\Omega)$, is the so-called *cross-entropy*. In other words, ML-estimates maximize the cross-entropy with respect to the parameters θ_r.

The ML*-algorithm for power-laws

To apply Eq (7) for ML-estimates of power-law exponents, one specifies the finite sample space $\Omega = \{z_1, z_2, \cdots, z_W\}$, and the family of probability density functions is,

$$p(x|\lambda,\Omega) = \frac{x^{-\lambda}}{Z_\lambda(\Omega)} \quad , \tag{8}$$

with $x \in \Omega$. Note that the set of parameters θ defined above now only contains λ, or $\theta = \{\lambda\}$. The normalization constant is $Z_\lambda(\Omega) = \sum_{x\in\Omega} x^{-\lambda}$. The derivative with respect to λ of the cross-entropy, $H_{\text{cross}}(f||p(z|\theta,\Omega)) = \lambda\sum_{i=1}^{W} f_i \log z_i + \log Z_\lambda(\Omega)$, has to be computed, and setting $dH_{\text{cross}}/d\lambda = 0$ yields

$$\sum_{i=1}^{W} f_i \log z_i = \left(\sum_{i=1}^{W} z_i^{-\lambda}\right)^{-1} \sum_{i=1}^{W} z_i^{-\lambda} \log z_i \quad . \tag{9}$$

An equivalent version of this equation is derived in [27] using different means. The solution to this implicit equation, $\lambda = \lambda^*$, can not be written in closed form but can be easily solved numerically. See [30] for the corresponding algorithm and S2 File APPENDIX B for a tutorial.

How to determine λ*

One possibility to find the solution $\lambda = \lambda^*$ from the implicit equation Eq (9), is to iteratively refine approximate solutions. For this, select $M + 1$ values λ from the interval $[\lambda_{\min}, \lambda_{\max}]$, where M is a finite fixed number, say $M = 100$. Those values may be chosen to be given by the expression

$$\lambda_r(m) = \underline{\lambda}_r + \frac{m}{M}\left(\overline{\lambda}_r - \underline{\lambda}_r\right) , \tag{10}$$

for $m = 0, \cdots, M$. The parameters $\underline{\lambda}_r$ and $\overline{\lambda}_r$ are defined in the following way: First define $\underline{\lambda}_1 = \lambda_{\mathrm{min}}$, and $\overline{\lambda}_1 = \lambda_{\mathrm{max}}$, where λ_{max} and λ_{min} are parameters of the algorithm. Then define $\delta\lambda_1 = \Delta\lambda/M$ with $\Delta\lambda = \lambda_{\mathrm{max}}-\lambda_{\mathrm{min}}$. If $\lambda_1(m_1^*)$ is the optimal solution of Eq (9) for some m_1^*, then we can choose $\underline{\lambda}_2 = \lambda_1(m_1^*) - \delta\lambda_1$, and $\overline{\lambda}_2 = \lambda_1(m_1^*) + \delta\lambda_1$ and $\delta\lambda_2 = 2\delta\lambda_1/M$. One then continues by iterating r times until $\delta\lambda_r < \varepsilon$, where ε is the desired accuracy of the estimate of λ^*. As a consequence, the value m_r^*, for which $|\lambda^* - \lambda_r(m_r^*)| < \varepsilon$ holds, optimally estimates λ^* in the r'th iteration with an error smaller than ε. Note that ε is the error of the ML*-estimator with respect to the exact value of the predictor λ^*, and is not the error of λ^* with respect to the (typically unknown) value of the exponent λ of the sampling distribution.

Controlling the fit region over which the power-law should be obtained therefore becomes a matter of restricting the sample space to a convenient $\Omega' \subset \Omega$. This can be used for dynamically controlling low-frequency cut-offs. These cut-offs are set to exclude states for which,

$$p(z_i|\lambda, \Omega)N < k_{\mathrm{min}} \quad , \tag{11}$$

where k_{min} is the minimal number of times that any state-type i is represented in the data set. This means that we re-estimate λ on $\Omega' \subset \Omega$ with

$$\Omega' = \{z_i \in \Omega | p(z_i|\lambda, \Omega)N \geq k_{\mathrm{min}}\} \quad . \tag{12}$$

We see in Eq (9) that iteratively adapting Ω to subsets Ω', and then re-evaluating λ, requires to solve,

$$\sum_{i\in I(\Omega')} f_i' \log(z_i) = \left(\sum_{i\in I(\Omega')} z_i^{-\lambda}\right)^{-1} \sum_{i\in I(\Omega')} z_i^{-\lambda} \log z_i \quad , \tag{13}$$

where $N' = \sum_{i \in I(\Omega')} k_i$ is the restricted sample-size and $f_i' = k_i/N'$ are the relative frequencies re-normalized for Ω'. $I(\Omega') = \{i|z_i \in \Omega'\}$ is the index-set of Ω'.

Iterating this procedure either leads to a fixed point or to a limit cycle between two low-frequency cut-offs with two slightly different estimates for λ^*. These two possibilities need to be considered in order to implement an efficient stopping criterion for the iterative search of the desired low-frequency cut-off in the data. The algorithm therefore consists of two nested iterations. The "outer iteration" searches for the low-frequency cut-off, the "inner iteration" solves the implicit equation for the power-law exponent. The matlab code for the algorithm is found in [30], see S2 File APPENDIX B for a tutorial.

Testing the new estimator with numerical experiments and known data sets

To test the proposed algorithm implementing the estimator ML*, we first perform numerical experiments and then test its performance on a number of well known data sets.

Testing with numerical experiments

For 400 different values of λ, ranging from 0 to 4, we sample $N = 10{,}000$ data points $x \in \Omega = \{1, \cdots, W\}$, with $W = 1000$ states, with probabilities $p(x|\lambda, \Omega) \propto x^{-\lambda}$. We fit the data in three ways, using (i) least square fits (LS), (ii) the CSN algorithm ML$_{\mathrm{CSN}}$ providing estimates $\hat{\lambda}$, and

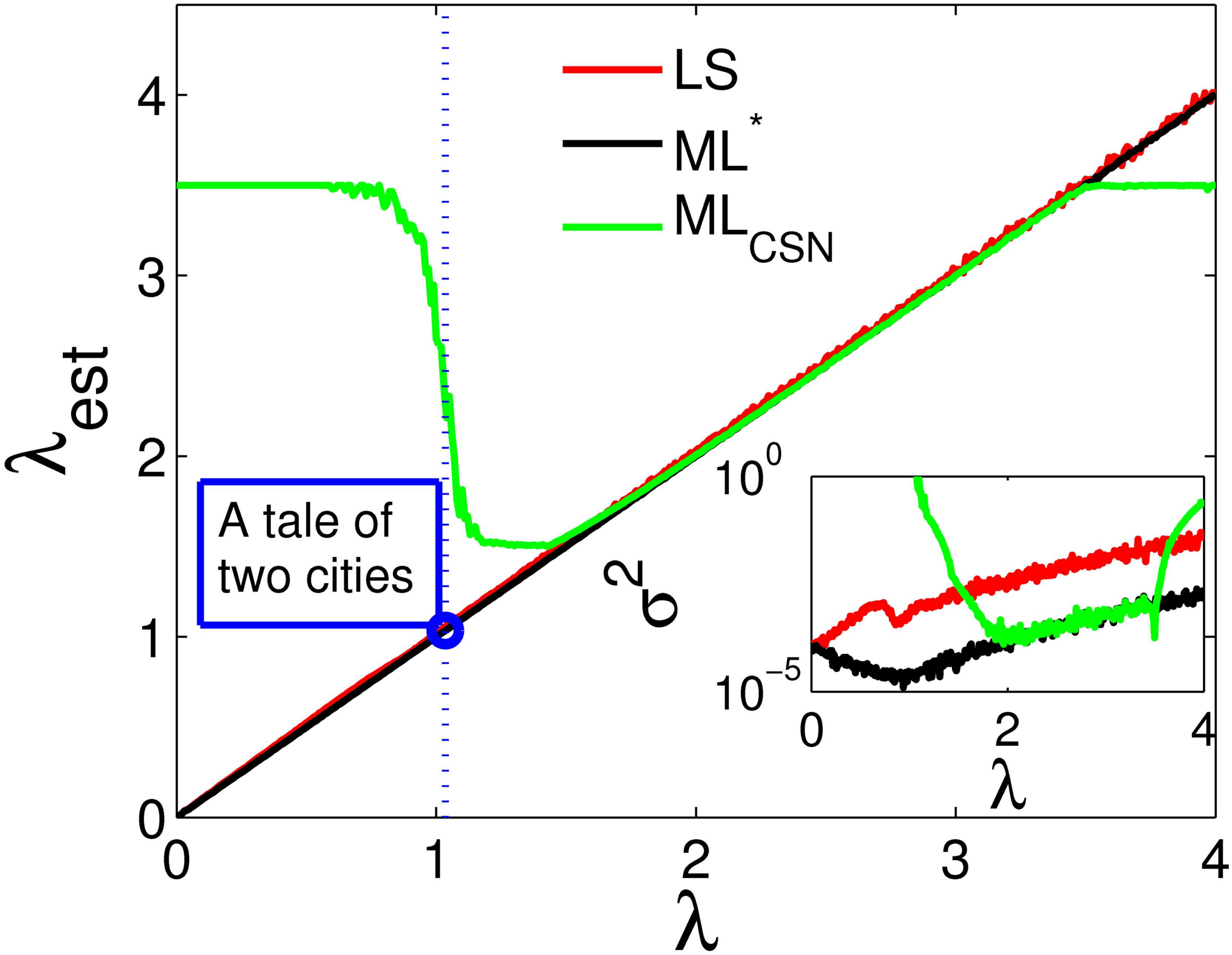

Fig 3. Comparison of the three power-law exponent estimators, LS, $\mathrm{ML_{CSN}}$, and ML*. For 400 values of λ in the range between 0 and 4, we sample $N = 10{,}000$ events from $\Omega = \{1, \cdots, 1{,}000\}$, from a power-law probability distribution $p(x|\lambda, \Omega) \propto x^{-\lambda}$. The estimated exponents λ_{est} for the estimators LS (red), the $\mathrm{ML_{CSN}}$ (green, $\lambda_{\mathrm{est}} = \hat{\lambda}$), and the new ML* (black, $\lambda_{\mathrm{est}} = \lambda^*$), are plotted against the true value of the exponent λ of the probability distribution samples are drawn from. Clearly, below $\lambda \sim 1.5$ the $\mathrm{ML_{CSN}}$ estimator no longer works reliably. $\mathrm{ML_{CSN}}$ and ML* work equally well in a range of $1.5 < \lambda < 3.5$. Outside this range ML* performs consistently better than the other methods. The inset shows the mean-square error σ^2 of the estimated exponents. The LS-estimator has a much higher σ^2 over the entire region, than the ML*-estimator. The blue dot represents the ML* estimate for the Zipf exponent of C. Dickens' "A tale of two cities". Clearly, this exponent could never reliably be obtained from the rank ordered distribution using $\mathrm{ML_{CSN}}$, whereas ML* works fine even for values of $\lambda \sim 0$.

doi:10.1371/journal.pone.0170920.g003

(iii) the implicit ML^* method providing estimates λ^*. In Fig 3 we show these estimates for the power exponents, as a function of the true *values* of λ. The LS, $\mathrm{ML_{CSN}}$, ML^* estimators are shown as the red, green, and black curves respectively. Obviously ML^* and $\mathrm{ML_{CSN}}$ work equally well for power-law exponents λ with values $1.5 < \lambda < 3.5$. In this range the three approaches coincide. However, note that in the same region the mean square error. (The mean square error is defined as $\sigma^2(\lambda) = N_{\mathrm{rep}}^{-1} \sum_{m=1}^{N_{\mathrm{rep}}} (\lambda_{\mathrm{est}}(m) - \lambda)^2$, where N_{rep} is the number of repetitions, i.e. the number of data-sets we sampled from the $p(x|\lambda, \Omega)$, $x = 1, \ldots, W$. $\lambda_{\mathrm{est}}(m)$ is the value estimated for λ from the *m*th data set. Depending on the estimator λ_{est} corresponds to

Table 1. Comparison of the estimators ML* and $\mathbf{ML_{CSN}}$ on empirical data sets that were used in [23]. These include the frequency of surnames, intensity of wars, populations of cities, earthquake intensity, numbers of religious followers, citations of scientific papers, counts of words, wealth of the Forbes 500 firms, numbers of papers authored, solar flare intensity, terrorist attack severity, numbers of links to websites, and forest fire sizes. We added the word frequencies in the novel "A tale of two cities" (C. Dickens). The second column states if α or λ were estimated. The exponents reported in [23] are found in column CSN_1, those reproduced by us applying their algorithm to data [23, 34–37] is shown in column CSN_2. The latter correspond well with the new ML* algorithm. For values λ < 1.5, CSN can not be used. We list the corresponding values for Kolmogorov-Smirnov test for the two estimators, KS_{CSN} and KS*.

	exp.	CSN_1	CSN_2	ML*	KS_{CSN}	KS*
blackouts	λ	2.3	2.27	2.25	0.061	0.031
surnames	α	2.5	2.49	2.66	0.041	0.019
int. wars	λ	1.7	1.73	1.83	0.078	0.076
city pop.	λ	2.37	2.36	2.31	0.019	0.016
quake int.	λ	1.64	1.64	1.88	0.092	0.085
relig. fol.	λ	1.8	1.79	1.61	0.091	0.095
citations	λ	3.16	3.16	3.10	0.010	0.018
words	α	1.95	1.95	1.99	0.009	0.015
wealth	λ	2.3	2.34	2.30	0.063	0.066
papers	λ	4.3	4.32	3.89	0.079	0.082
sol. flares	λ	1.79	1.79	1.81	0.009	0.021
terr. attacks	λ	2.4	2.37	2.36	0.018	0.017
websites	λ	2.336	2.12	1.72	0.025	0.056
forest fires	λ	2.2	2.16	2.46	0.036	0.034
Dickens novel	λ	-	-	1.04	-	0.017

doi:10.1371/journal.pone.0170920.t001

$\hat{\lambda}$ (ML_{CSN}), λ^*, (ML^*), or the LS estimator. We used $W = 1000$ and $N_{rep} = 25$ for any given λ.) σ^2 for the LS method is much larger than for ML^* and ML_{CSN}. Outside this range the assumptions and approximations used for ML_{CSN} start to lose their validity and both LS and ML^* estimates outperform the ML_{CSN} estimates. The inset also shows that ML^* consistently estimates λ much better than the LS estimator (two orders of magnitude better in terms of σ^2) for the entire range of λ. The blue dot in Fig 3 represents the ML^* estimate for the Zipf exponent of C. Dickens' 'A tale of two cities'. Clearly, this small exponent could never be obtained by ML_{CSN}, see also Table 1.

Testing with empirical data sets

We finally compare the new estimator ML^* on several empirical data sets that were used for demonstration in [23]. In Table 1 we collect the results. The second column states if λ or α were estimated. Column CSN_1 presents the value of the estimator ML_{CSN} as presented in [23]. Column CSN_2 contains the values of the same estimator using the data from [23] and using the algorithm provided by [31]. (The reason for the differences might be that some of the data has been updated since the publication.) The results for the ML^* estimator agrees well with those of ML_{CSN} in the range where the latter works well. To demonstrate how ML^* works perfectly outside of the comfort zone of ML_{CSN} (for $\lambda < 1.5$), we add the result of the rank distribution of word counts in the novel "A tale of two cities" (Charles Dickens, 1859), which shows an exponent of $\lambda \sim 1.035$. This exponent can be fitted directly from the data using the proposed ML^* algorithm, while ML_{CSN} can not access this range, at least not without the detour of first producing a histogram from the data and then fitting the tail of the frequency distribution. The values for the corresponding Kolmogorov-Smirnov tests (see e.g. [23]) for the two estimates, KS_{CSN} and KS^*, are similar for most cases.

Conclusions

We discuss the generic problem of estimating power-law exponents from data sets. We list a series of questions that must be clarified before estimates can be performed. We present these questions in form of a decision tree that shows how the answers to those questions lead to different strategies for estimating power-law exponents.

To follow this decision tree can be seen as a recipe for fitting power exponents from empirical data. The corresponding algorithms were presented and can be downloaded as matlab code. The two algorithms we provide are based on a very general ML estimator that maximizes an appropriately defined cross entropy. The method can be seen as a straight forward generalization of the idea developed in [23]. The two estimators (one for binned histograms and ML* for raw data sets) allow us to estimate power-law exponents in a much wider range than was previously possible. In particular, exponents lower than $\lambda < 1.5$ can now be reliably obtained.

Supporting information

S1 File. APPENDIX A: Sampling from continuous sample spaces. Provides a short discussion on ML estimates performed on data sampled from continuous sample spaces.
(PDF)

S2 File. APPENDIX B: Using r_plfit. Gives a short introduction on how to use the matlab implementations of the algorithms r_plfit and r_plhistfit.
(PDF)

S3 File. APPENDIX C: The false rejection rate of power-laws. Discusses the *p*-value of the method and how the *p*-values obtained from the KS goodness of fit test can be interpreted. It also includes a figure that exemplifies how r_plfit_calib_eval.m can be used to visualize calibration curves sampled with r_plfit_calibrate.m, in order to control the false rejection rate (*p*-value) of r_plfit.m.
(PDF)

S4 File. APPENDIX D: Code. Contains the code for all the m-files in print form. Alternatively to the download we provide, [30], the code can be accessed using copy-and-paste.
(PDF)

Acknowledgments

This work was supported in part by the Austrian Science Foundation "Fonds zur Förderung der wissenschaftlichen Forschung" (FWF: https://www.fwf.ac.at/) under grant P29252. BL is grateful for the support by the China Scholarship Council, file-number 201306230096. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Author Contributions

Conceptualization: RH.

Data curation: BL.

Formal analysis: RH BL.

Funding acquisition: ST BL.

Investigation: BL.

Methodology: RH BCM.

Project administration: RH ST.

Software: RH.

Supervision: RH ST.

Validation: RH BCM BL ST.

Visualization: RH BL.

Writing – original draft: RH ST.

References

1. Newman MEJ (2005) Power-laws, Pareto distributions and Zipf's law, Contemporary physics 46 323–51. doi: 10.1080/00107510500052444
2. Mitzenmacher M (2004) A Brief History of Generative Models for Power-Law and Lognormal Distributions, Internet Mathematics 1 226–51. doi: 10.1080/15427951.2004.10129088
3. Kadanoff LP, Gotze W, Hamblen D, Hecht R, Lewis EAS, Palciauskas VV et al. (1967) Static Phenomena Near Critical Points: Theory and Experiment, Rev Mod Phys 39 395–413. doi: 10.1103/RevModPhys.39.395
4. Sornette D (2006) Critical Phenomena in Natural Sciences, in Springer, Berlin.
5. Bak P, Tang C, Wiesenfeld K (1987) Self-Organized Criticality: An Explanation of 1/f Noise, Phys Rev Lett 59 381–84. doi: 10.1103/PhysRevLett.59.381 PMID: 10035754
6. Simon HA (1955) On a class of skew distribution functions, Biometrika 42 425–40. doi: 10.1093/biomet/42.3-4.425
7. Réka A, Barabási AL (2002) Statistical mechanics of complex networks, Rev Mod Phys 74 47–97. doi: 10.1103/RevModPhys.74.47
8. Barabási AL, Réka A (1999) Emergence of scaling in random networks, Science 286 509–12. doi: 10.1126/science.286.5439.509 PMID: 10521342
9. Yule GU (1925) A Mathematical Theory of Evolution, based on the Conclusions of Dr. J. C. Willis, F.R. S, Phil Trans Royal Soc B 213 21–87. doi: 10.1098/rstb.1925.0002
10. Takayasu H, Sato AH, Takayasu M (1997) Stable Infinite Variance Fluctuations in Randomly Amplified Langevin Systems, Phys Rev Lett 79 966–67. doi: 10.1103/PhysRevLett.79.966
11. Tsallis C (2009) Introduction to nonextensive statistical mechanics, Springer, New York.
12. Hanel R, Thurner S, Gell-Mann M (2014) How multiplicity of random processes determines entropy: derivation of the maximum entropy principle for complex systems, Proc Nat Acad Sci USA 111 6905–10. doi: 10.1073/pnas.1406071111 PMID: 24782541
13. Corominas-Murtra B, Hanel R, Thurner S (2015) Understanding scaling through history-dependent processes with collapsing sample space, Proc Nat Acad Sci USA 112 5348–53. doi: 10.1073/pnas.1420946112 PMID: 25870294
14. Gutenberg B, Richter CF (1944) Frequency of earthquakes in California, Bull Seismol Soc Amer 34 185–88.
15. Christensen K, Danon L, Scanlon T, Bak P (2002) Unified scaling law for earthquakes, Proc Nat Acad Sci USA 99 2509–13. doi: 10.1073/pnas.012581099 PMID: 11875203
16. Auerbach F (1913) Das Gesetz der Bevölkerungskonzentration, Petermanns Geographische Mitteilungen 59 74–76.
17. Gabaix X (1999) Zipf's Law for Cities: An Explanation, Quart. J. Econ. 114 739–67. doi: 10.1162/003355399556133
18. Shaffer CA (2014) Spatial foraging in free ranging bearded sakis: Traveling salesmen or Lévy walkers?, Amer J. Primatology 76 472–84. doi: 10.1002/ajp.22227 PMID: 24166852
19. Newman MEJ, Palmer RG (2003) Modeling extinction, Oxford University Press.
20. Zipf GK (1949) Human Behavior and the Principle of Least Effort, Addison-Wesley, Cambridge, Massachusetts.
21. Press SJ (2010) Subjective and Objective Bayesian Statistics: Principles, Models, and Applications, Wiley Series in Prob and Stat.

22. Berger JO (1985) Statistical decision theory and Bayesian Analysis, Springer, New York. doi: 10.1007/978-1-4757-4286-2

23. Clauset A, Shalizi CR, andNewman MEJ (2009) Power-Law Distributions in Empirical Data, SIAM Rev 51 661–703. doi: 10.1137/070710111

24. Kolmogorov A (1933) Sulla determinazione empirica di una legge di distribuzione, G Ist Ital Attuari 4 83–91.

25. Smirnov N (1948) Table for estimating the goodness of fit of empirical distributions, Ann of Math Stat 19 279–281. doi: 10.1214/aoms/1177730256

26. Fisher RA (1912) On an absolute criterion for fitting frequency curves, Messenger of Math 41 155–60.

27. Bauke H (2007) Parameter estimation for power-law distributions by maximum likelihood methods, Eur Phys J B, 58:167–173. doi: 10.1140/epjb/e2007-00219-y

28. Goldstein ML, Morris SA, Yen GG (2004) Problems with fitting to the power-law distribution, Eur Phys J B 41 255–258. doi: 10.1140/epjb/e2004-00316-5

29. Deluca A, Corral A (2013) Fitting and goodness-of-fit test of non-truncated and truncated power-law distributions, Acta Geophysica 61 1351–94 doi: 10.2478/s11600-013-0154-9

30. http://www.complex-systems.meduniwien.ac.at/SI2016/r_plfit.m
http://www.complex-systems.meduniwien.ac.at/SI2016/r_plhistfit.m
http://www.complex-systems.meduniwien.ac.at/SI2016/r_randi.m
http://www.complex-systems.meduniwien.ac.at/SI2016/r_plfit_calibrate.m
http://www.complex-systems.meduniwien.ac.at/SI2016/r_plfit_calib_eval.m Alternatively, see also S4 File Appendix D for the code.

31. http://tuvalu.santafe.edu/∼aaronc/powerlaws/

32. Virkar Y, Clauset A (2014) Power-law distributions in binned empirical data, Ann of Appl Stat 8 89–119. doi: 10.1214/13-AOAS710

33. Adamic L, Huberman BA (2002) Zipf's law and the Internet, Glottometrics 3, 143–150.

34. Broder A, Kumar R, Maghoul F, Raghavan P, Rajagopalan S, Stata R et al. (2000) Graph structure in the web, Computer networks 33 309–20. doi: 10.1016/S1389-1286(00)00083-9

35. Roberts DC, Turcotte D.L (1998) Fractality and self-organized criticality of wars, Fractals 6 351–57. doi: 10.1142/S0218348X98000407

36. Redner S (1998) How popular is your paper? An empirical study of the citation distribution, EPJ B 4 131–34. doi: 10.1007/s100510050359

37. Clauset A, Young M, Gleditsch KS (2007) On the frequency of severe terrorist events, J of Conflict Resolution 51 58–87. doi: 10.1177/0022002706296157