Sormunen, Silja; Leskelä, Lasse; Saramäki, Jari

Distinguishing subsampled power laws from other heavy-tailed distributions

Published in:
Physical Review E

DOI:
10.1103/PhysRevE.109.054308

Published: 08/05/2024

Document Version
Publisher's PDF, also known as Version of record

Distinguishing subsampled power laws from other heavy-tailed distributions

Silja Sormunen [1,*] Lasse Leskelä [2] and Jari Saramäki [1]
[1]*Department of Computer Science, Aalto University, 00076 Espoo, Finland*
[2]*Department of Mathematics and Systems Analysis, Aalto University, 00076 Espoo, Finland*

(Received 20 July 2023; accepted 11 April 2024; published 8 May 2024)

Distinguishing power-law distributions from other heavy-tailed distributions is challenging, and this task is often further complicated by subsampling effects. In this work, we evaluate the performance of two commonly used methods for detecting power-law distributions—the maximum likelihood method of Clauset *et al.* and the extreme value method of Voitalov *et al.*—in distinguishing subsampled power laws from two other heavy-tailed distributions, the lognormal and the stretched exponential distributions. We focus on a random subsampling method commonly applied in network science and biological sciences. In this subsampling scheme, we are ultimately interested in the frequency distribution of elements with a certain number of constituent parts—for example, species with k individuals or nodes with k connections—and each part is selected to the subsample with an equal probability. We investigate how well the results obtained from low-subsampling-depth subsamples generalize to the original distribution. Our results show that the power-law exponent of the original distribution can be estimated fairly accurately from subsamples, but classifying the distribution correctly is more challenging. The maximum likelihood method falsely rejects the power-law hypothesis for a large fraction of subsamples from power-law distributions. While the extreme value method correctly recognizes subsampled power-law distributions with all tested subsampling depths, its capacity to distinguish power laws from the heavy-tailed alternatives is limited. However, these false positives tend to result not from the subsampling itself but from the estimators' inability to classify the original sample correctly. In fact, we show that the extreme value method can sometimes be expected to perform better on subsamples than on the original samples from the lognormal and the stretched exponential distributions, while the contrary is true for the main tests included in the maximum likelihood method.

DOI: 10.1103/PhysRevE.109.054308

I. INTRODUCTION

Power-law distributions have frequently been observed in both natural and artificial systems. They naturally emerge by mechanisms such as preferential attachment [1] and restarts in telecom and queueing networks [2,3], and they also serve to explain other complex phenomena such as fractional Gaussian noises [4] and small-world phenomena [5]. The apparent ubiquity of power laws has been taken to indicate a universal self-organizing mechanism at play, and power laws have acquired a reputation as a hallmark of complex systems. However, there is no universally accepted method for identifying power laws, and their ubiquity has. been argued to result from lacking statistical testing rather than their actual universality [6]. The debate has been especially heated in the field of network science, where networks with a power-law degree distribution, the so-called scale-free networks, play a prominent role. For these networks, the probability that a randomly chosen node has k connections to other nodes varies as a power of the degree k. Using the maximum likelihood methods presented in Ref. [7], Broido and Clauset concluded that—contrary to the widespread belief—scale-free networks are rare [6]. A power law was required to hold only for the largest degrees:

$$P(k) = \frac{1}{\zeta(\alpha, k'_{\min})} k^{-\alpha}, \text{ for degrees } k \geqslant k'_{\min} \geqslant 1, \qquad (1)$$

where $\alpha > 1$ is the tail exponent, $k'_{\min}$ is the smallest degree for which the power law holds, and $\zeta(\alpha, k'_{\min})$ is the Hurwitz zeta function which normalizes the degree distribution so that $\sum_{k=k'_{\min}}^{\infty} P(k) = 1$. In contrast to this, however, Voitalov *et al.* [8] found more evidence of scale-free networks using a method based on extreme value theory. In their work, the definition of power-law distribution was extended to include all regularly varying distributions, defined as distributions whose complementary cumulative distribution function approaches a power law asymptotically in the tail while deviating arbitrarily from a pure power law for smaller degrees. In line with this, Serafino *et al.* [9] found many empirical degree distributions to satisfy the scale-free hypothesis when finite-size effects were taken into account with tools from statistical physics. Note that while these articles use the terminology of network science, the methods can equally well be applied to frequency distributions in any other domain of science.

In summary, distinguishing power-law distributions from other heavy-tailed distributions has proven to be challenging, and this task is further complicated if the data are *subsampled*. Often only a part of the system can be observed, and depending on the subsampling strategy, the subsample represents the original system more or less accurately. For example, if we

*silja.sormunen@aalto.fi

sample nodes of a network with equal probability and record their degrees, the subsampling is trivial in the sense that the degree distribution stays unchanged. However, this kind of unbiased subsampling is often not possible; we might, for example, sample each node with equal probability but only be able to observe the connections between the chosen nodes, in which case the observed degree distribution does not faithfully reflect that of the whole system. In such cases, making inferences from the subsample without further consideration might lead to erroneous conclusions.

In this work, we assess how reliably two state-of-the-art methods for recognizing power-law distributions—that of Clauset *et al.* [6,7] and Voitalov *et al.* [8]—determine whether subsampled data originate from a power-law distribution. We focus on a random subsampling method where the probability of observing a node depends on its degree. In network science, this method is known as the incident subgraph sampling strategy, where each edge is included in the subsample with probability π together with the nodes that it connects. The degree distribution of the subsamples is given by

$$P_s(k) = \sum_{i \geqslant k}^{\infty} P(i) \binom{i}{k} \pi^k (1-\pi)^{i-k}, \tag{2}$$

where $P(k)$ denotes the degree distribution of the original network. This subsampling strategy is equivalent to selecting each node with a probability linear in degree. The same strategy can be applied to frequency distributions arising in other contexts. In biodiversity studies, for example, $P(k)$ might represent the fraction of species with k individuals, and each individual would subsequently be picked to the subsample with probability π. In general, this subsampling method is common in situations where observing the whole system or population is impossible due to the sheer size of the system, such as when recording the size distribution of neuronal avalanches in the brain [10], assessing relative species abundance in an ecological community [11] or investigating the diversity of cells in immunological studies [12]. In such cases, it is crucial to know to what extent the results obtained for a subsample can be generalized to the original distribution.

For power laws, generalizing the results to the original distribution is not straightforward. Stumpf *et al.* [13] have argued that the degree distribution of a scale-free network is not closed under the subsampling strategy described by Eq. (2), meaning that the degree distribution of the subsampled network and that of the whole network do not belong to the same family of probability distributions. The same applies to other heavy-tailed distributions such as the lognormal and the stretched exponential distributions [13], which are notoriously difficult to distinguish from power laws and have commonly been used as alternatives to the power-law hypothesis in the previous literature (see, for example, Refs. [14–17]). For power-law distributions, deviation from the original power law grows larger as the subsampling depth decreases or the power-law exponent increases [13]. However, the form of the distribution is mostly affected for small degrees, and the tail of the subsampled distribution still approaches the original power law asymptotically for $k \gg 1$ [18]. This finding is fruitfully exploited by Levina *et al.* [10], who propose a method for differentiating between power-law and exponential distributions by further subsampling the data—a subsample itself—and scaling the subsampled sequences in a way that collapses the scaled tails of the subsamples to the original power law.

In general, methods considering only the tail of a distribution should, in theory, continue to work on subsampled power-law distributions, which still belong to the larger class of regularly varying distributions. However, we do not currently know how incident subgraph sampling affects the separability of other heavy-tailed distributions from power-law distributions. Subsampling might distort other distributions to resemble power laws; for example, Han *et al.* [19] have previously observed that when the subsampling scheme consists of selecting a fraction p of the nodes and a fraction q of those nodes neighbors, subsamples from networks with exponential, truncated normal and Poisson degree distributions can mimic power-law-like behavior, when the resemblance to power law is assessed based on the degree of linearity between logarithms of the degree k and the number of nodes with degree k.

In this work, we evaluate how reliably the methods of Clauset *et al.* [6,7] and Voitalov *et al.* [8] succeed in distinguishing subsampled power-law distributions from two other types of heavy-tailed distributions—namely the lognormal and the stretched exponential distributions—when the above-described incident subgraph sampling strategy is applied to subsamples from simulated degree distributions. We use the term stretched exponential distribution to refer to the subclass of Weibull distributions with $\beta \in (0, 1)$ to maintain consistency with nomenclature in our core references. For convenience, we use the name maximum likelihood (ML) method to refer to the power-law hypothesis test by Broido and Clauset [6] based on the methods of Clauset *et al.* [7] and the name extreme value (EV) method to refer to the method of Voitalov *et al.* As heavy-tailed distributions are by definition heavier than any exponential distribution, we assess the methods' performance on discrete exponential distributions (i.e., geometric distributions) as well. Our analysis is restricted to these two methods because their implementation is readily available, excluding, e.g., the finite-size scaling method of Ref. [9]. It is also worth noting that this finite-size scaling method uses the tools of the ML method to determine how large a fraction of the distribution's tail is to be considered in the analysis, and therefore its performance depends partly on how well this estimation succeeds.

Our results show that the power-law exponent can be estimated fairly accurately from simulated subsamples of power-law distributions, but the classification of the distribution's type should be taken with caution. Finally, we show that subsampling affects the performance of the two methods differently: While the EV method can in some cases be expected to perform better on subsampled data than on the original distribution, the opposite applies to the main tests included in the ML method.

II. ESTIMATORS

We start by briefly presenting the ML and the EV methods; the reader already familiar with these can move straight to Sec. III. Note that we have chosen to use these names for the sake of convenience, and they do not necessarily capture the essence of the methods; the EV method, for example, incor-

porates estimators that are based on the maximum likelihood approach, while the ML method contains additional tests not based on maximum likelihood estimation.

A. Maximum likelihood method

The ML method of Refs. [6,7] for assessing whether a sample originates from a power-law distribution starts with estimating the optimal values of the start of the power law and the corresponding power-law exponent. The rationale behind not necessarily including the entire sample in the analysis is that many empirical distributions are expected to follow a power law only for large values of k [7]. In the following, we denote the true start of the power law (a property of the distribution) with $k'_{\min}$. We use $\hat{k}_{\min}$ to refer to the best estimate of $k'_{\min}$ produced by the ML method; only data points $k \geqslant \hat{k}_{\min}$ are used for testing the power-law hypothesis. Furthermore, we employ the symbol $k_{\min}$ to denote the smallest value of k included in the analysis in cases where this value is not selected by the automatic procedure of the ML method (e.g., where it is chosen manually or where one sweeps through a range of values).

To find the optimal $\hat{k}_{\min}$, each unique value of k present in the data is in turn used as $k_{\min}$, and a maximum likelihood estimate for the power-law exponent α is calculated considering only data points $k \geqslant k_{\min}$. Subsequently, the value of $k_{\min}$ minimizing the Kolmogorov-Smirnov (KS) distance between the cumulative distribution function (CDF) of the data points larger than or equal to $k_{\min}$ and the CDF of the fitted power-law model in the same region is selected as the optimal $\hat{k}_{\min}$.

Next, the statistical plausibility of the best-fitting model is assessed with a goodness-of-fit test. Denoting the sample size with n and the number of data points larger than or equal to $\hat{k}_{\min}$ with n_{tail}, a number of synthetic datasets are generated with a semiparametric bootstrap approach, where each data point is drawn from the best-fitting power-law model with probability n_{tail}/n and else from the empirical sample with $k < \hat{k}_{\min}$. A power-law model is then fitted to each of these bootstrapped samples, and the KS distance between the CDF of the original empirical distribution and its best-fitting power-law model is compared to the distribution of KS statistics between the generated synthetic datasets and their fitted models, and the test is considered to reject the power-law hypothesis if the fraction of KS statistics at least as extreme as the KS distance of the empirical distribution is smaller than a given p value. Finally, the power-law model is compared to four alternative distributions (lognormal, exponential, Weibull, and truncated power law) using a normalized log-likelihood-ratio test originally presented by Vuong *et al.* [20].

In Ref. [6], the sample is subsequently classified to one of six categories based on how strong evidence for the power-law hypothesis it is deemed to show. Here we group these into two categories. First, as in Ref. [6], a sample is said to show strong evidence for the power-law hypothesis if the following four conditions are met:

(1) The estimated exponent of the power law is between 2 and 3.

(2) The number of data points in the tail, n_{tail}, is at least 50.

(3) The goodness-of-fit test cannot reject the power-law hypothesis (p value $\geqslant 0.1$).

(4) None of the alternative distributions is favored over power law in the log-likelihood-ratio test.

In Ref. [6], a sample falls into the category "not scale-free" if neither of conditions 3 or 4 is met. Consequently, we say that a sample shows some evidence for the power-law hypothesis if it fulfills at least one of these two conditions.

B. Extreme value method

Voitalov *et al.* [8] broaden the definition of power-law distribution to encompass all regularly varying distributions. The complementary cumulative distribution function (CCDF) of a regularly varying function is given by $\bar{F}(k) = l(k)k^{-(\alpha-1)}$, where $l(k)$ is a slowly varying function defined by the property $\lim_{k\to\infty} \frac{l(tk)}{l(k)} = 1$ for any $t > 0$. All distributions with a probability density function given by $P(k) = l(k)k^{-\alpha}$ are regularly varying, but the converse is not true.

The EV method of Voitalov *et al.* is based on extreme value theory, which is concerned with the limit distribution of the sample maximum. Let $X_1, X_2, \ldots, X_n$ form a sample of independent and identically distributed random variables following a cumulative distribution function F. Then F is said to be in the maximum domain of attraction (MDA) of an extreme value distribution with tail index ξ, denoted by $F \in D_M(G_\xi)$, if there exist normalizing constants $a_n > 0$ and $b_n \in \mathbb{R}$ and a nondegenerate distribution function G such that

$$\lim_{n\to\infty} P\left[\frac{\max(X_1, X_2, \ldots, X_n) - b_n}{a_n} \leqslant x\right] = \lim_{n\to\infty} [F(a_n x + b_n)]^n = G_\xi(x) \text{ for all } x. \tag{3}$$

By the Fisher-Tippett-Gnedenko theorem (see, e.g., Ref. [21]), there exist only three families of extreme value distributions $G_\xi(x)$—the Fréchet, Gumbel and reversed Weibull families—each characterized by the tail index ξ determining the shape of the distribution. Regularly varying distributions (both continuous and their discrete counterparts) form the MDA of the Fréchet distribution, for which $\xi > 0$. The power-law exponent of any regularly varying distribution in the MDA of the Fréchet distribution can be directly inferred from the tail index,

$$\alpha = \frac{1}{\xi} + 1. \tag{4}$$

The reversed Weibull family, in turn, is characterized by a negative tail index, indicating that the distribution is bounded from above. Finally, for distributions belonging to the Gumbel MDA, the tail index equals zero. The continuous lognormal and stretched exponential distributions belong to the MDA of the Gumbel distribution. As both lognormal and stretched exponential distributions belong to the class of long-tailed distributions [22], also their discretized versions remain in the same MDA [23]. In contrast, while the continuous exponential distribution is in the Gumbel MDA, its discrete counterpart does not belong to any MDA [23].

Voitalov *et al.* [8] propose estimating the power-law exponent α with three statistically consistent estimators of the tail index—Hill, moments, and kernel—which can be automated

using a double bootstrap method for finding the optimal size of the tail considered in the estimation. The authors argue that due to the nonparametric nature of the regularly varying distribution family, it is impossible to quantify the probability that a given finite sample originates from a regularly varying distribution; however, if all three estimators return a clearly positive estimate of ξ, the observed sequence is likely to come from a regularly varying distribution. Consequently, Voitalov *et al.* classify a distribution as power law if all the considered estimators estimate the tail index to be over $1/4$. The limit is set to $1/4$ instead of zero to reduce the probability of falsely accepting the power-law hypothesis. The distribution is classified as not power law if at least one of the estimators returns a nonpositive estimate, else the network is classified to be "hardly power law." In other words, the power-law exponent α needs to be between 1 and 5 for a sequence to be classified as a power law.

The first considered estimator, the Hill estimator [24], is statistically consistent for $\xi > 0$ and converges eventually to zero for $\xi = 0$. Given an ordered sample $X_{(1)} \geqslant X_{(2)} \geqslant \cdots \geqslant X_{(n)}$, the estimator operates on the κ largest observations and gives an empirical estimate of the expected excess of the log-transformed distribution over threshold $\ln(X_{(\kappa+1)})$:

$$\hat{\xi}^{\text{Hill}}_{\kappa,n} = \frac{1}{\kappa}\sum_{i=1}^{\kappa} \ln\left[\frac{X_{(i)}}{X_{(\kappa+1)}}\right]. \tag{5}$$

As the other two estimators, the Hill estimator converges to the true tail index as $\kappa, n \to \infty$ and $\kappa/n \to 0$.

The moments estimator [25] is an extension of the Hill estimator consistent for all $\xi \in \mathbb{R}$:

$$\hat{\xi}^{\text{Moments}}_{\kappa,n} = \hat{\xi}^{\text{Hill}}_{\kappa,n} + 1 - \frac{1}{2}\left[1 - \frac{\left(\hat{\xi}^{\text{Hill}}_{\kappa,n}\right)^2}{\hat{\xi}^{\text{Hill},2}_{\kappa,n}}\right]^{-1}, \tag{6}$$

where

$$\hat{\xi}^{\text{Hill},2}_{\kappa,n} = \frac{1}{\kappa}\sum_{i=1}^{\kappa}\left[\ln\frac{X_{(i)}}{X_{(\kappa+1)}}\right]^2. \tag{7}$$

For $\xi \geqslant 0$, $\lim_{n\to\infty} \frac{(\hat{\xi}^{\text{Hill}}_{\kappa,n})^2}{\hat{\xi}^{\text{Hill},2}_{\kappa,n}} = \frac{1}{2}$ [25], meaning that the moments estimator should converge to the value of the Hill estimator for distributions belonging to the MDA of either Fréchet or Gumbel families.

The kernel estimator [26] is consistent for all $\xi \in \mathbb{R}$ as well. The number of largest observations considered is determined by the bandwidth parameter h; approximately nh observations are considered. The kernel estimator is given by

$$\hat{\xi}^{\text{Kernel}}_{h,n} = \hat{\xi}^{\text{pos}}_{h,n} - 1 + \frac{\hat{q}^{(2)}_{h,n}}{\hat{q}^{(1)}_{h,n}}, \text{ where} \tag{8}$$

$$\hat{\xi}^{\text{pos}}_{h,n} = \sum_{i=1}^{\lfloor nh\rfloor} \frac{i}{n} K_h\left(\frac{i}{n}\right)\ln\left[\frac{X_{(i)}}{X_{(i+1)}}\right] \tag{9}$$

$$= -\int_0^h uK_h(u)\, d\, \ln Q_n(1-u), \tag{10}$$

$$\hat{q}^{(1)}_{h,n} = \sum_{i=1}^{\lfloor nh\rfloor}\left(\frac{i}{n}\right)^{\gamma} K_h\left(\frac{i}{n}\right)\ln\left[\frac{X_{(i)}}{X_{(i+1)}}\right] \tag{11}$$

$$= -\int_0^h u^{\gamma} K_h(u)\, d\, \ln Q_n(1-u), \tag{12}$$

$$\hat{q}^{(2)}_{h,n} = \sum_{i=1}^{\lfloor nh\rfloor} \frac{\partial}{\partial u}[u^{\gamma+1}K_h(u)]_{u=i/n}\ln\left[\frac{X_{(i)}}{X_{(i+1)}}\right] \tag{13}$$

$$= -\int_0^h \frac{d}{du}[u^{\gamma+1}K_h(u)]\, d\, \ln Q_n(1-u), \tag{14}$$

where the kernel $K_h(u)$ is given by $\frac{15}{8h}[1-(\frac{u}{h})^2]^2$ and Q_n denotes the empirical quantile function defined as $Q_n(u) = \inf\{x : F_n(x) \geqslant u\}$, where F_n is the empirical distribution function. Following Ref. [8], we set the parameter γ equal to 0.6. Note that we use $\lfloor nh\rfloor$ as the upper limit in the above summations while the upper limit is set to $\lfloor nh\rfloor - 1$ in the code of Ref. [8]. As we illustrate in the Supplemental Material (SM) IB [27], using either $\lfloor nh\rfloor$ or $\lfloor nh\rfloor - 1$ as the upper limit yields identical results under some conditions; however, for a general h (excluding $h = 1$), the limit $\lfloor nh\rfloor$ yields correct results. We have modified the indexing in the code accordingly (see SM IB [27] for details).

III. SIMULATING SUBSAMPLING

To investigate the performance of the estimators on subsampled distributions, we generate n_0 random numbers—corresponding to the degrees of n_0 nodes—from the desired distributions, subsample these simulated degree sequences using the previously described incident subgraph sampling strategy, and apply the ML and the EV methods to the obtained subsampled degree sequences. To avoid confusion, we use the term *subsample* to refer to the samples obtained with incident subgraph sampling, while the term *sample* refers to a sample from a given distribution to which the incident subgraph sampling has not yet been applied. Correspondingly, we denote the number of data points in the original sample by n_0 and use n as general symbol for the sample size, whether of a subsample or the original sample.

A. Simulation procedure

We generate the samples with the exact search algorithm described in Ref. [7] and implemented in the Python package *powerlaw* [31]. In short, the algorithm operates by generating a random number $u \in (0, 1]$ from a uniform distribution and returning the largest integer i such that $\bar{F}(i) \geqslant u$, where $\bar{F}$ denotes the CCDF of the desired distribution. The probability density functions (PDFs) of the distributions are listed in Table I (see Fig. 1 for visualization). After generating n_0 random degrees k_j from a given distribution, we form a list where a unique identifier j is repeated k_j times for each node j, shuffle the list, and analyze the first x identifiers of the list at a time, where $x = \text{round}(\pi \sum_{j=1}^{n_0} k_j)$ and π is the subsampling probability. With this method, the degree of each node j in the subsample simply equals the number of identifiers j in the selected list. We repeat the sampling procedure twenty times for each combination of parameters. Note that if two samples consist of the same number of data points, i.e., degrees, but the other sample comes from a distribution with a heavier tail, the sum of degrees is in general higher in that sample.

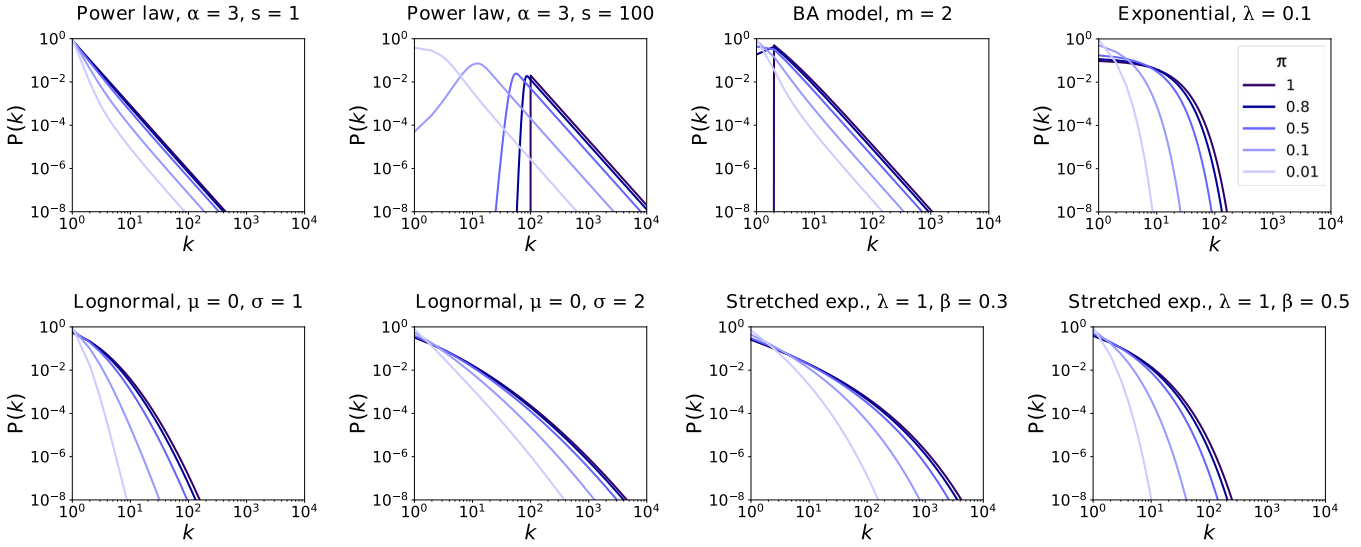


FIG. 1. Examples of probability density functions (PDFs) of the subsampled distributions obtained by applying Eq. (2) to the PDFs listed in Table I (procedure described in detail in Sec. IV A). The beginning of the distribution's support before subsampling is denoted by s. Increasing μ and σ of the lognormal distribution and decreasing λ and β of the stretched exponential distribution lead to heavier tails. Subsampling seems to increase the apparent degree of linearity on the log-log scale especially for the lognormal distributions.

Consequently, also the number of data points can be expected to be higher at any subsampling depth $\pi < 1$. Note also that the range of values of k varies greatly for different distributions, becoming more restricted for distributions with smaller subsampling depths and lighter tails (such as the exponential distributions, stretched exponential distributions with $\beta = 0.5$ and lognormal distributions with $\mu = 1$).

For the ML method, there are two available Python implementations of the original code in Matlab [7]. We use the *powerlaw* package by Alstott *et al.* [31] for the model fitting as well as the log-likelihood-ratio tests. As the goodness-of-fit test is not implemented in this package, we use a modified version of the *plpval* function from the implementation by Broido and Clauset [6], where we have altered the function so that the model fitting is again done with the *powerlaw* package (see SM IA [27] for details). We have chosen to use a combination of these implementations due to an issue we encountered in the Broido and Clauset implementation regarding calculation of the normalization constant in cases where not all integers between the minimum and maximum values of the sample are present in the data (see SM I [27]). In the Alstott *et al.* implementation, we have furthermore increased the number of allowed iterations when optimizing the parameters of alternative distributions with the *scipy.optimize.fmin* function by adding the argument *maxiter = 1000*. We do this to prevent termination of optimization before the parameters have converged (note that the variable *warnflag* warning about failed convergence is not printed by default).

For the EV method, we use the code released together with the article [8] with the modification discussed in Sec. II B. We use the default parameters of the code, including adding a random uniform value $u \in [-0.5, 0.5)$ to each discrete data point to enhance the performance of the estimators. A minor exception is the parameter *amseborder*, which reduces the likelihood of the double-bootstrap method choosing the order statistics from a region where the uniform noise dominates. We use the default value 1.0 in all (sub)samples where $k = 1$ is the smallest degree and set the value equal to the minimum degree for all other (sub)samples.

TABLE I. Probability density functions of the distributions, $P(k) = Cf(k)$. The distributions are normalized so that $\sum_{k=s}^{\infty} P(k) = 1$, where s denotes the start of the distribution's support. The parameter ranges are the following: for power law, $\alpha > 1$; for discrete exponential (i.e., geometric), $\lambda > 0$; for lognormal, $\mu \in (-\infty, \infty)$, $\sigma > 0$; and for Weibull, $\lambda > 0$, $\beta > 0$ (for the heavy-tailed subclass that we refer to as stretched exp., $\beta \in (0, 1)$). For the Barabási-Albert (BA) model, m is the number of nodes that each incoming node attaches to.

Distribution	C	$f(k)$
Power law	$\zeta(\alpha, s)^{-1}$	$k^{-\alpha}$
Exponential	$(1 - e^{-\lambda})e^{\lambda s}$	$e^{-\lambda k}$
Lognormal*	$\sqrt{\frac{2}{\pi\sigma^2}}\left[\mathrm{erfc}\left(\frac{\ln s - \mu}{\sqrt{2}\sigma}\right)\right]^{-1}$	$\frac{1}{k}\exp\left[-\frac{(\ln k - \mu)^2}{2\sigma^2}\right]$
Weibull*	$\beta\lambda e^{(\lambda s)^\beta}$	$(\lambda k)^{\beta-1}e^{-(\lambda k)^\beta}$
BA model		$\frac{2m(m+1)}{k(k+1)(k+2)}$ for $k \geqslant m$

*The discrete PDFs are approximated with the continuous PDFs listed here, as the distributions marked with a star lack an analytically defined discrete form. When generating random samples, the probabilities are obtained as $P(k) = F(k + 0.5) - F(k - 0.5)$, where $F(k)$ is the corresponding CDF, after which the probabilities are normalized by their sum.

B. Estimation accuracy of the power-law exponent

As in Ref. [8], we assess the estimation accuracy of the power-law exponent with the relative root-mean-squared error (RRMSE). RRMSE is commonly used to measure the average error in tail index estimation proportional to the real value of

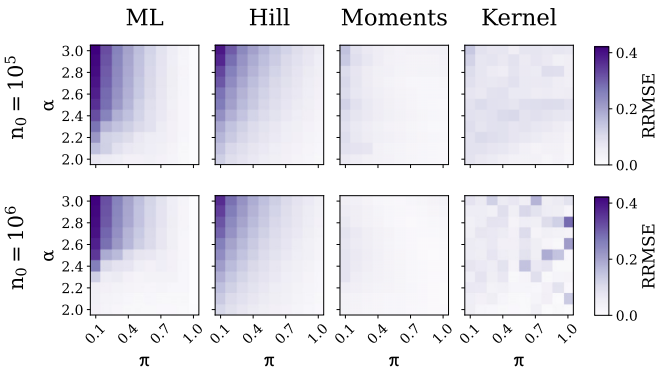


FIG. 2. Relative root-mean-squared error (RRMSE) values for the different estimates of the tail index ξ as a function of the true power-law exponent α and the subsampling depth π. The darker the color, the greater the error in the tail index estimation. The columns correspond to the different estimators, while the subfigures in the same row share the same number of data points n_0 in the original sample. The number of (sub)samples used for calculating the RRMSE value in each cell is 20.

$\xi = \frac{1}{\alpha-1}$:

$$\mathrm{RRMSE} = \frac{\sqrt{\frac{1}{r}\sum_{j=1}^{r}(\hat{\xi}_j - \xi)^2}}{\xi},$$

where r is the number of samples. RRMSE is defined in terms of ξ and not the power-law exponent α because α is defined only for distributions belonging to MDA of the Fréchet distribution, while ξ characterizes all three extreme value distributions.

We observe that when n_0, i.e., the number of simulated data points in the original sample, equals 10^5, the moments and the kernel estimates stay fairly accurate even when the sequence is subsampled to a 10th of the original size [Fig. 2 (top row)]. In contrast, both the maximum likelihood estimate of the ML method and the Hill estimate deviate further from the true power-law exponent α as the subsampling probability π decreases and α grows larger (see SM II [27] for the values of $\hat{\alpha}$ and $\hat{k}_{\min}$ as well as the median sizes and maximum values of k in the subsamples). Note that the number of data points for a certain subsampling depth differs greatly for different distributions, and the range of values k gets more and more restricted for larger values of α and lower subsampling depths. However, increasing n_0 from 10^5 to 10^6 improves the accuracy only slightly [Fig. 2 (bottom row)]. Due to the challenging shape of the heavily subsampled power-law sequences, all estimators tend to consider a too-large part of the tail; for α close to 3, the ML method consistently estimates $\hat{k}_{\min}$ to be 1, which is clearly not an optimal choice considering the pronounced downward bending shape of subsampled power-law distributions (see the PDFs in Fig. 1). Somewhat surprisingly, the kernel estimator becomes more unstable when $n_0 = 10^6$ and classifies some of the samples as belonging to the MDA of the Gumbel distribution. These misclassifications arise from a less-than-optimal estimation of the number of order statistics; almost all data points are taken into account, and, consequently, the added random noise on the small degrees has a prominent effect on the estimate.

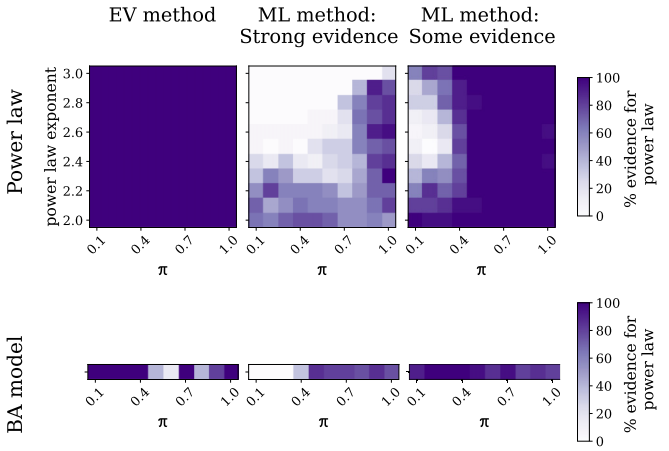


FIG. 3. Performance of the two methods with samples from power-law distributions as well as from degree distributions of networks grown by preferential attachment (BA model). The darker the color, the higher the fraction of subsamples correctly classified as power laws. The percentages are calculated over 20 samples with the number of data points before subsampling equal to 10^5.

The accuracy of the tail index estimation for small subsamples improves clearly if the original distribution has support only on higher values of k. In this case, the part of the distribution most affected by subsampling—the smallest degrees—has no probability mass to begin with and hence cannot dominate the estimation of $\hat{k}_{\min}$ or the number of order statistics. Consequently, the error of all estimators remains negligible for all distributions even for $\pi = 0.1$.

C. Classifying subsamples

1. Power-law distributions

The EV method classifies all subsamples from power-law distributions correctly for all tested subsampling depths $\pi \in [0.1, 1.0]$. As shown in the previous section, the moments and the kernel estimates stay fairly accurate even for small subsampling probabilities, and while the Hill estimator deviates from the true α with decreasing π, the estimates of α do not exceed the allowed upper limit for the tested subsampling probabilities.

The ML method works less reliably on the subsamples (Fig. 3). The fraction of subsamples exhibiting strong evidence for the power-law hypothesis diminishes with decreasing π and increasing α, primarily due to the KS minimization driving the value of $\hat{k}_{\min}$ to a too-low value, which results in an inaccurate estimate of α (see SM III [27] for performance of the individual criteria of the ML method and their different combinations). Consequently, the estimates of α often fall outside the allowed range [2,3], and the goodness-of-fit test is more likely to reject the power-law hypothesis. This tendency of the ML method to select a too-small $\hat{k}_{\min}$ is already noted in Ref. [8] and Ref. [32] with regard to some regularly varying distributions, and the problem is further aggravated by the probability mass concentrating more heavily on small degrees as a result of subsampling. The log-likelihood-ratio test, however, continues to perform better than the goodness-of-fit test on the subsamples,

resulting in a higher fraction of correctly classified subsamples when using the some evidence criteria.

In addition to pure power laws, we assess the behavior of the estimators with degree distributions of networks grown by preferential attachment. We simulate the Barabási-Albert (BA) model, where new nodes are added to the network one by one and each new node forms $m = 2$ connections to already existing nodes with probability $\frac{k_i}{\sum_j k_j}$, where k_i denotes the degree of node i and the index j ranges over all the already existing nodes of the network [33]. The degree distribution of the network asymptotically approaches a power law with $\alpha = 3$ for $k \gg 1$.

The strong evidence criteria of the ML method performs better with the subsamples from the BA model than with subsamples from a pure power law with $\alpha = 3$. Performance is better because the fraction of nodes with degree one is smaller than for a pure power law, which leads to a more accurate estimation of $\hat{k}_{\min}$. However, subsamples with low subsampling depth are again falsely deemed to not to show strong evidence for the power-law hypothesis.

The EV method, in turn, works less reliably on the subsamples from the BA model than on pure power laws (Fig. 3). This results from a tendency of the kernel estimator to occasionally estimate the subsample to belong to the MDA of the Gumbel distribution as a result of considering too-large a fraction of the distribution's tail.

2. Exponential distributions

All subsamples are classified correctly when using the strong evidence criteria of the ML method (Fig. 4). While the goodness-of-fit test and the log-likelihood ratio test do not always manage to reject the power-law hypothesis, the ML estimates of α stay consistently above the allowed upper limit, 3, which results in the lack of strong evidence.

The EV method correctly classifies all subsamples from exponential distribution with $\lambda \in [0.1, 0.5]$ as non-power-law samples for all tested subsampling depths.

3. Stretched exponential distributions

According to the ML method, none of the subsamples with $\beta \in \{0.3, 0.4, 0.5, 0.7, 0.9\}$ and $\lambda \in \{1, 2, 3\}$ show strong evidence for the power-law hypothesis (Fig. 4). A fraction of all (sub)samples shows some evidence for the power-law (PL) hypothesis, but this fraction does not seem to be notably affected by the exact value of the subsampling probability π.

The EV method consistently classifies samples with $\beta \geqslant 0.5$ as non-power-laws, but for $\beta = 0.4$ the rate of misclassification increases, and for $\beta = 0.3$ the majority of the original samples as well as the subsamples are classified as power laws. These incorrect classifications result from too-slow convergence to the limiting extreme value distribution for small values of β, a problem noted with regards to $\beta = 0.3$ already in Ref. [14].

4. Lognormal distributions

None of the lognormal sequences with $\sigma = 1$ exhibit strong evidence for the power-law hypothesis according to the ML criteria, again largely due to the estimates of α falling above the upper limit of the accepted range. These estimates decrease as σ is increased and the tail becomes heavier, and, consequently, some of the subsampled sequences with $\sigma = 2$ are falsely deemed to show strong evidence.

The EV method, in turn, erroneously classifies some of the samples with $\sigma = 1$ as power laws, but—somewhat surprisingly—the rate of incorrect classifications seems to decrease with decreasing subsampling depth. This trend is more pronounced for distributions with smaller values of μ. When σ is increased to 1.3, the vast majority of the original samples, as well as their subsamples, are falsely classified as power laws. This trend continues for $\sigma = 2$; in addition,

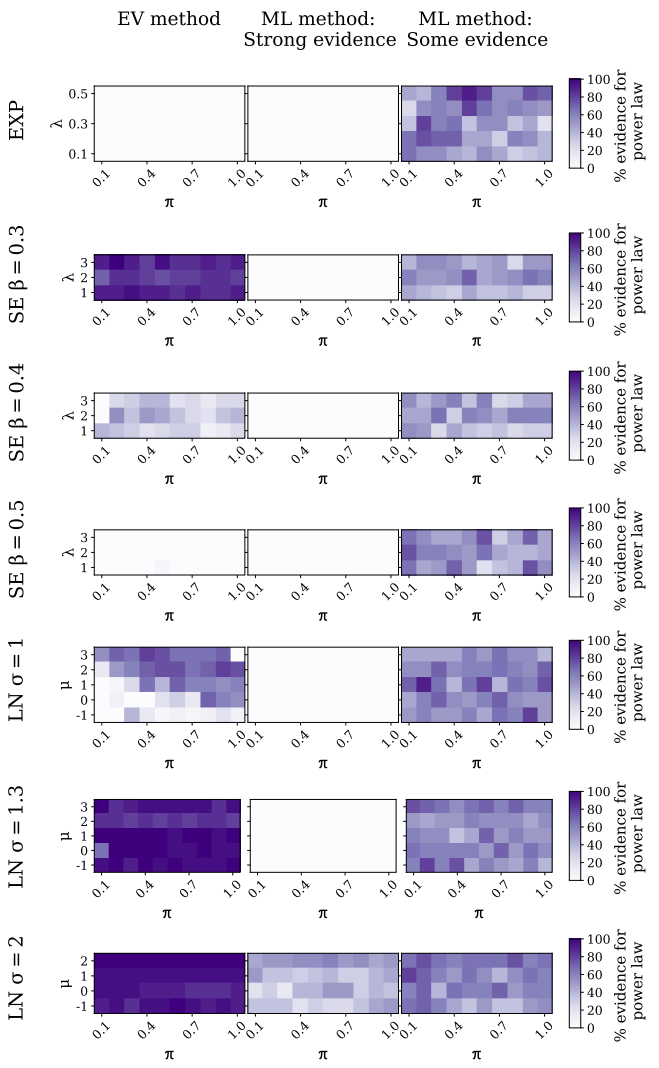


FIG. 4. Performance of the two methods with subsamples from exponential (EXP), lognormal (LN), and stretched exponential (SE) distributions. The darker the color, the higher the fraction of subsamples incorrectly classified as power laws. The number of data points in the original sample before subsampling is 10^5 for all distributions except for the lognormal distribution with $\sigma = 2$, for which $n_0 = 10^4$ due to computational limitations. For the performance of the individual EV estimators, see SM III [27]. In general, combining the classifications of the three EV estimators outperforms any individual EV estimator on its own, supporting the use of the combined results to minimize false-positive classifications.

all three EV estimates tend to lie close to each other, which usually indicates that the classification should be reliable.

As with the stretched exponential distribution, the unreliability of the EV method for lognormal distributions with larger values of σ originates from too-slow convergence to the asymptotic EV distribution. This dramatic slowing down as σ increases can be seen in the convergence rate in the von Mises condition on which the consistency analysis of the kernel method is based on (Eq. (1.1) in Ref. [26]),

$$\lim_{k_{\min}\to\infty} M(k_{\min}) = \xi, \text{ where } M(k_{\min}) = \frac{d}{dk_{\min}} \frac{1-F(k_{\min})}{F'(k_{\min})}. \tag{15}$$

For the continuous lognormal distribution with $\mu = 2$ and $\sigma = 2$, we observe that the value of $M(k_{\min})$ falls for the first time under the threshold $1/4$ (used by the EV method to separate power laws from non-power-laws) when $k_{\min} > 17\,000\,000$. In contrast, when the parameter σ is decreased to 1, the value of $M(k_{\min})$ falls under the threshold already around $k_{\min} \approx 73$.

In conclusion, it seems that the EV method and the strong evidence criteria of the ML method suffer from opposite tendencies; while the strong criteria tends to produce false negatives, the EV method is prone to falsely accepting the power-law hypothesis for some of the tested alternative heavy-tailed distributions. The some evidence criteria of the ML method suffers from a substantial rate of false positives as well. Overall, analyzing the performance of the individual criteria of the ML method shows that the log-likelihood ratio criteria performs better than the goodness-of-fit test both with regard to the rate of false positives and false negatives, and hence its overall performance is better than that of the *some evidence* criteria (see SM III [27]). Whether one prefers to use this criterion alone, the strong evidence criteria or the strong evidence criteria without the goodness-of-fit test ultimately depends on how much emphasis one places on avoiding false positives at the expense of an increased false-negative rate.

Interestingly, it would seem that the methods' ability to reliably reject the power-law hypothesis for subsamples from alternative distributions depends to a great extent on their ability to classify the original sample correctly. As an exception to this trend we observed that the accuracy of the EV method seemed to increase for some lognormal distributions as a result of subsampling. However, as the results of the simulations depend rather heavily on the sample size n, it is not clear whether this exception originates from the properties of the subsampled distributions or simply from the smaller amount of data points in the substantially downsampled data. In addition, trends visible with either larger or smaller sample sizes might go unnoticed in the simulations. To resolve these questions, we now turn to examine the behavior of the estimators on the theoretical subsampled probability distributions.

IV. PERFORMANCE ON THEORETICAL SUBSAMPLED PROBABILITY DISTRIBUTIONS OF THE HEAVY-TAILED ALTERNATIVES

In this section, we analyze the theoretical subsampled probability distributions of the heavy-tailed alternatives and assess whether correctly classifying these lognormal and stretched exponential distributions as non-power-law becomes more challenging as the subsampling depth decreases. We put all considerations of the sample size n aside and assess the behavior of the estimators on the distributions in the limit of large n (derivations in SM IV [27]), under the assumption that the subsampling depth or the choice of the smallest considered degree $k_{\min}$ does not significantly affect the rate at which an estimator converges towards its limiting value. Note that we do not consider here how the size of the tail to be considered is in practice estimated for empirical or simulated data with either the double bootstrap procedure of the EV method or the KS minimization of the ML method, since this selection depends on the sample size n. Instead, we manually vary the value of $k_{\min}$ and assess how the analytically calculated estimates change as a result.

A. Generating theoretical distributions

Given a discrete probability density function $P(k)$ (listed in Table I), the theoretical subsampled distributions are obtained by brute force using Eq. (2) with $i = 10^5$ as the upper limit in the summation. Some examples of the theoretical subsampled distributions are shown in Fig. 1. As the upper limit is finite, these numerically obtained subsampled distributions are not exact. To exclude the possibility of numerical imprecision confounding our results, we verify that increasing the upper limit to $i = 10^6$ does not visibly change the results for the largest and the smallest subsampling depth in Figs. 7 and 8 (as well as Fig. 9 in the SM [27]) for the parameter combinations producing the heaviest tails. We refrain from considering parameter combinations requiring an even larger limit, as already the limit $i = 10^6$ is computationally very heavy.

Since lognormal and stretched exponential distributions do not have an analytically defined discrete form, we approximate the original discrete distributions before subsampling by calculating the pointwise probabilities $P(k) = f(k)$ (the notation corresponds to that in Table I) from $k = 1$ up to $k = 10^5$ and normalizing the probabilities by their sum. To verify that our conclusions are not affected by the chosen discretization strategy, we repeat the analysis for one representative distribution from both lognormal and stretched exponential families with two other discretization methods; one where $P(k) = F(k+1) - F(k)$ and another where $P(k) = F(k + 0.5) - F(k - 0.5)$. While the exact numerical limit values of the estimators for a given $k_{\min}$ depend on the discretization strategy, the conclusions remain unaffected.

When analyzing the behavior of the EV estimators, we treat the probability distribution functions not as discrete functions but as step functions, where the probability remains constant from $k - 0.5$ up to $k + 0.5$ for each k. We do this to mirror the empirical analysis as closely as possible. This transformation corresponds to adding random uniform noise to each data point, as is done with the simulated data.

B. Extreme value method

As previously discussed, the successfulness of the EV method depends to a great extent on the distribution's rate

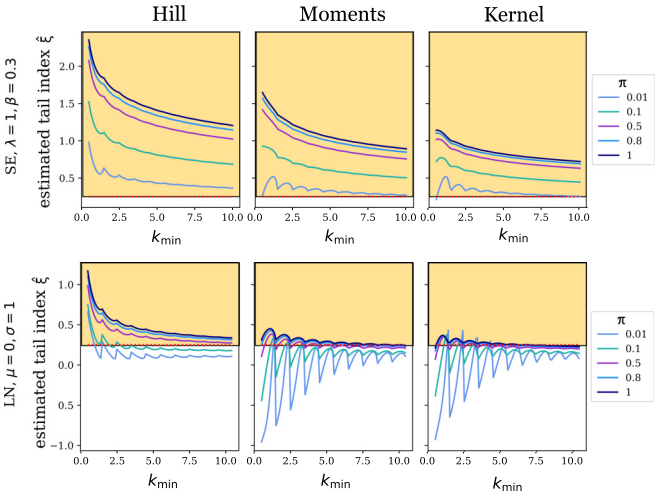


FIG. 5. Examples of the behavior of the estimated tail index $\hat{\xi}$ in the limit of large n as a function of the smallest considered degree $k_{\min}$ for a stretched exponential distribution (SE, top row) and a lognormal distribution (LN, bottom row) with different subsampling depths π. The yellow shaded area ($\hat{\xi} > 1/4$) marks the values of $\hat{\xi}$ for which the distribution is incorrectly classified as a power law.

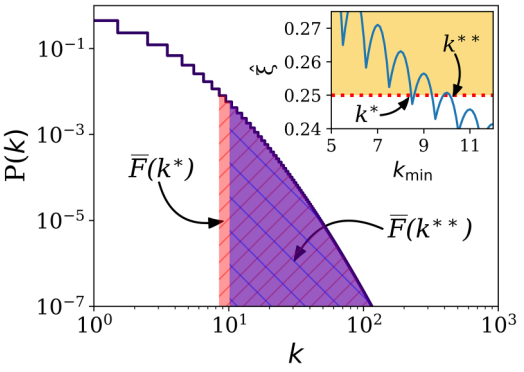


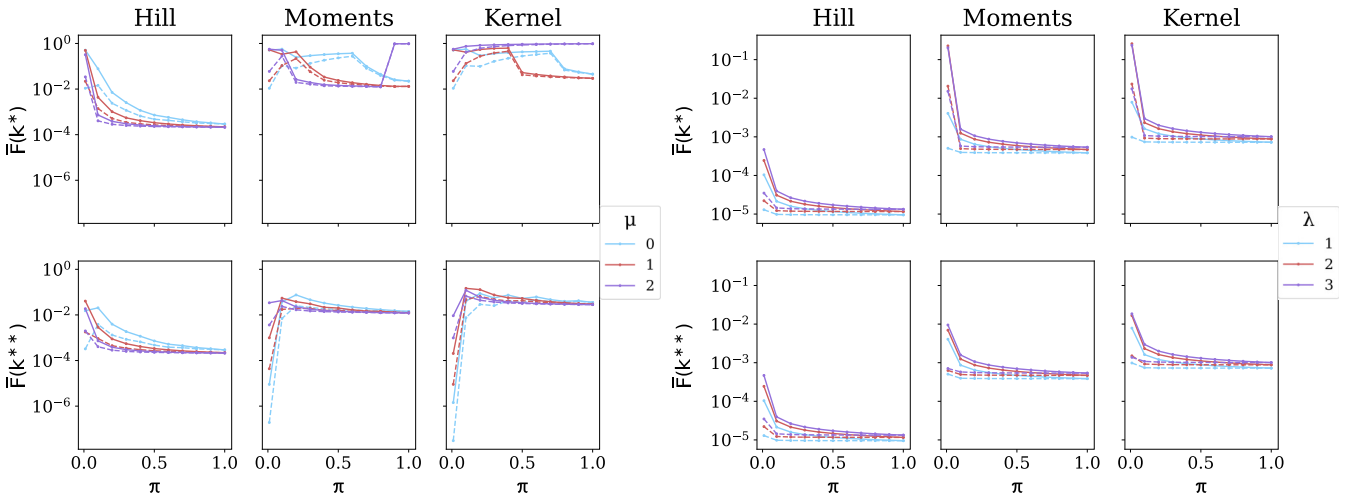
FIG. 6. Illustration of the fractions $\bar{F}(k^*)$ and $\bar{F}(k^{**})$ for the lognormal distribution with $\mu = 1$, $\sigma = 1$ and subsampling depth $\pi = 0.4$. The inset shows the analytical limit value of the moments estimator as a function of the smallest considered degree $k_{\min}$. At $k_{\min} = k^*$, this limit value falls for the first time (and at $k_{\min} = k^{**}$, permanently) under the threshold 1/4 allowing for correctly classifying the distribution as non-power-law. The shaded areas $\bar{F}(k^*)$ and $\bar{F}(k^{**})$ display the corresponding values of the theoretical subsampled probability distribution's CCDF, i.e., the fractions of the distribution's tail that can be used to correctly identify the distribution's type according to the criteria of the EV method.

of convergence to the asymptotic extreme value distribution; correct classification of the distribution's type becomes more difficult if this rate slows down considerably. In practice, this would mean that the fraction of the distribution's tail that can be used to correctly identify the maximum domain of attraction (and hence the type of the distribution) becomes too small, meaning that the sample size n would need to be unreasonably large to capture this part of the tail. To assess the effect of subsampling on the estimators' performance, we examine how the fraction of probability mass in the distribution's tail that allows for correct identification of the distribution's type changes as the subsampling depth decreases. In the following, we refer to this fraction as the usable part of the tail. In the case of lognormal and stretched exponential distributions, this fraction corresponds to the value of the subsampled distribution's CCDF at the smallest value of $k_{\min}$ (not necessarily an integer in this context) for which the tail index falls below the limit of 1/4 allowing the distribution to be correctly classified as non-power-law according to the criteria of the EV method. We denote this value of $k_{\min}$ with k^* and assess it with an accuracy of 0.01.

As we are operating with step functions, the estimates do not necessarily decrease monotonically as a function of the smallest considered degree $k_{\min}$ (see Fig. 5 for examples), and the estimate might consequently rise above the threshold 1/4 even for $k_{\min} > k^*$. Consequently, it may be more informative to assess the CCDF at the point where the estimate falls permanently under the threshold of 1/4. We denote this point with k^{**} and the corresponding value of the CCDF with $\bar{F}(k^{**})$ (see Fig. 6 for illustration). In the following, we refer to $\bar{F}(k^{**})$ as the *fully usable* fraction of the tail and to $\bar{F}(k^*)$ as the *potentially usable* fraction. For some distributions the values k^* and k^{**} coincide. In general, the moments and the kernel estimators suffer from oscillations more than the Hill estimator, and the oscillations grow larger the more heavily the probability mass concentrates on the smallest degrees. This happens when subsampling depth decreases or parameters of the distributions are changed to produce a lighter tail. When comparing the usable fraction of the tail for different subsampling depths, we are essentially asking how the expected number of nodes from the usable part would change if we were to randomly draw samples of equal size from distributions of different subsampling depths. To understand how the expected number of nodes changes in cases where the sample sizes are unequal as a result of subsampling, we assess the fraction of the tail allowing for correct identification relative to the original distribution ($\pi = 1$) as well. We do this by multiplying the obtained value of $\bar{F}(k^*)$ or $\bar{F}(k^{**})$ with the fraction of the total non-normalized probability mass of the subsampled distribution, i.e., the probability $1 - P_s(k = 0)$, where $P_s(k = 0)$ is obtained using Eq. (2).

Interestingly, the lognormal and the stretched exponential distributions tend to become easier to separate from power laws when subsampled. Identification becomes easier in the sense that the fully usable fraction $\bar{F}(k^{**})$ of the tail tends to slightly increase as the subsampling depth decreases. However, this effect is far from linear, and if the tail is not heavy enough, $\bar{F}(k^{**})$ can drop dramatically for very small values of π due to the violently oscillating pattern of the estimator [Fig. 7(a)].

For the Hill estimator also the potentially usable fraction $\bar{F}(k^*)$ tends to become larger for smaller subsampling probabilities, even to the extent that this fraction of probability mass allowing for correct classification increases when expressed relative to the original distribution before subsampling. Since the oscillations are more pronounced for the moments and the kernel estimators, the potentially usable fraction behaves in a more unpredictable manner for these estimators. However, if the tail of the distribution is heavy enough, this fraction seems to increase consistently even for these estimators [Fig. 7(b)]. Overall, these results imply that the previous



(a) Lognormal with $\sigma = 1$ (b) Stretched exponential with $\beta = 0.3$

FIG. 7. Value of the CCDF at the smallest value of $k_{\min}$ (smallest degree considered in the analysis) at which the limit value of the EV estimators falls for the first time (top row) and permanently (bottom row) below the threshold $1/4$ resulting in the distribution being correctly classified as non-power-law. Larger values of the CCDF indicate that a larger fraction of the distribution's tail allows for correctly identifying the distribution. The value of the CCDF is expressed both relative to the subsampled distribution (solid lines) and to the original distribution (dashed lines). In the top row of panel (a), the abrupt fall of $\bar{F}(k^*)$ of the moments estimator for the lognormal distribution with $\mu = 2$ at $\pi = 0.8$ originates from the fact that we start to assess the estimators performance at $k_{\min} = 1.0$ instead of $k_{\min} = 0.5$ mimicking the default parameters of the code. For $\mu = 2$, the estimates for all tested subsampling probabilities fall originally below $1/4$ at $k_{\min} = 0.5$ and subsequently rise above this threshold, but for $\pi = 1.0$ this rising happens more slowly and the moments estimate still lies below $1/4$ at $k_{\min} = 1$.

finding of lognormal subsamples being more accurately classified for lower subsampling depths may result from better discriminability of the subsampled distribution and not only on the smaller sample size.

C. Maximum likelihood method

As in the previous section, we examine how subsampling affects the ease of classification by varying the smallest considered degree $k_{\min}$ and calculating the corresponding limit values of the estimators as $n \to \infty$ for the theoretical subsampled distributions. To allow comparison between different subsampling depths, we display the results not as a function of $k_{\min}$ but as a function of the value of the CCDF $\bar{F}(k_{\min})$. In contrast to the EV method, the ML method classifies empirical data using the p values associated with the estimates, and hence we cannot directly assess any threshold for $\bar{F}(k_{\min})$ above which the power-law hypothesis would be rejected. However, we can analyze whether the criteria for the power-law hypothesis would be more likely to be met in comparison with a different subsampling depth for a given $\bar{F}(k_{\min})$.

To get insight into the behavior of the goodness-of-fit test, we calculate for each $k_{\min}$ the Kolmogorov-Smirnov distance between the subsampled distribution's CDF and the CDF of the best-fitting power law as $n \to \infty$. Our results show that the KS distance between the distribution's CDF and the CDF of the best-fitting power law for a given $\bar{F}(k_{\min})$ tends to become smaller as the subsampling depth decreases; however, this effect remains small for moderate subsampling depths and becomes less prominent when moving towards the tail of the distribution (Fig. 8, see also Fig. 9 in SM V [27] for results for distributions with other parameter combinations). In general, a smaller KS distance indicates that the goodness-of-fit test is more likely to accept the power-law hypothesis. This is because the goodness-of-fit test proceeds by generating a number of bootstrapped samples from the fitted power-law model, after which a power-law model is fitted to each sample and the KS distance is calculated for each sample with respect to its own fitted model. The larger the fraction of the bootstrapped samples with KS distance exceeding that of the original sample, the more likely the power-law model is deemed to be. Consequently, assuming that the variances of the distributions of the KS statistics are approximately equal for different subsampling depths, our results indicate that if the values of the smallest considered degree $k_{\min}$ for two subsampling depths are chosen so that $\bar{F}(k_{\min})$ is approximately the same for both, the goodness-of-fit test is less likely to correctly reject the PL hypothesis for the lower subsampling depth. While a similar KS distance can often be obtained for the smaller subsampling depth by considering a larger part of the tail, this kind of compensation is not expected to happen in practice since the optimal $\hat{k}_{\min}$ for empirical or simulated data is chosen by minimizing the KS distance, which for all tested subsampling depths except for $\pi = 0.01$ tends to decrease towards the tail of the distribution.

The log-likelihood ratios comparing the likelihood of the power-law hypothesis to that of alternative distributions behave in a similar manner. With empirical data, the power-law hypothesis is considered to get support if none of the four alternative distributions is favored over the power-law model (i.e., the log-likelihood ratios are all positive or the corresponding p values are nonsignificant). Hence, the more negative the theoretical log-likelihood ratios are, the more

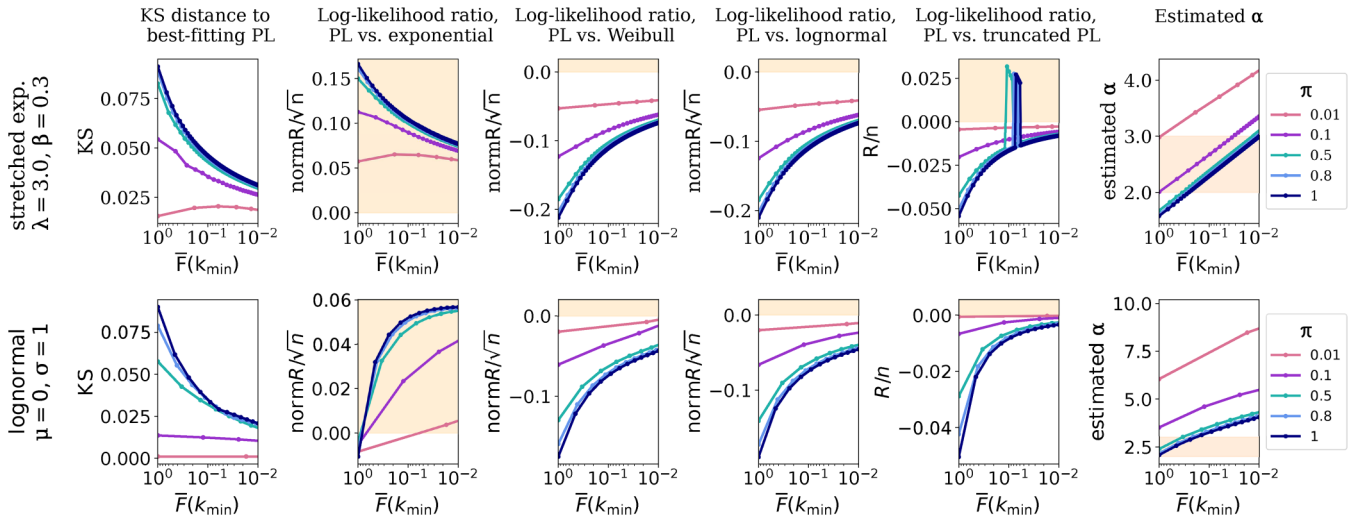


FIG. 8. Applying the estimators of the ML method to theoretical subsampled stretched exponential (upper row) and lognormal distribution (lower row) to assess how likely the power-law hypothesis can be expected to be rejected for different subsampling depths π in the limit $n \to \infty$. The analytical limit values of the estimators are calculated by varying $k_{\min}$ but we present them as a function of the value of the CCDF at $k_{\min}$, $\bar{F}(k_{\min})$, to allow better comparison between distributions. Here we define $\bar{F}(k_{\min}) = \sum_{k=k_{\min}} P(k)$. Note that the x axis has been inverted to enhance readability; when moving to the right, we are considering larger and larger values of $k_{\min}$, which correspond to smaller values of $\bar{F}(k_{\min})$. The first column shows that the KS distance between the values of the distribution's CDF and the CDF of the best-fitting power-law distribution for a given $\bar{F}(k_{\min})$ tends to become smaller as the subsampling depth π decreases, indicating that the goodness-of-fit test of the ML method can be expected to perform worse for smaller π. The next four columns show the results of the log-likelihood ratio tests where the PL hypothesis is tested against an alternative hypothesis. The best-fitting parameters of the alternative distribution are obtained with the same optimization methods as in the implementation of Alstott *et al.* [31] using as initial parameter guesses the values of the Alstott *et al.* implementation as $n \to \infty$. Depending on the alternative hypothesis, the results are expressed in terms of either the log-likelihood ratio R or its normalized version norm$R$ (details in SM IVA.3 [27]). If the ratio is positive (shaded yellow area), the incorrect PL model is favored over the alternative hypothesis, whereas a negative ratio indicates that the alternative hypothesis is favored over the PL model. The log-likelihood ratio tests should be understood as a series; if all four tests favor the power-law hypothesis over the alternative distribution or give inconclusive results (ratio close to zero, not meaningful in the limit of large n), the power-law hypothesis is considered to get support. Consequently, for the considered subsampled lognormal and stretched exponential distributions, the power-law hypothesis is more likely to be correctly rejected the more negative the likelihood ratios are. The abrupt spikes in the fifth column result from suboptimal estimation of parameters of the truncated power-law distribution. The last column shows that for a given $\bar{F}(k_{\min})$, the estimated value of the power-law exponent α is expected to converge to a larger value for smaller subsampling depths. The shaded area marks the range to which the estimated α must fall for the criteria of strong evidence for the PL hypothesis to be met.

likely the power-law hypothesis will be correctly rejected. Our results on the theoretical distributions show that when the alternative hypothesis against the power-law model is either the lognormal, the Weibull, or the truncated power-law distribution, the limit value of the likelihood ratio tends to become less negative for smaller subsampling depths for a given $\bar{F}(k_{\min})$, meaning that the alternative hypothesis is less strongly favored over the power-law model. However, the effect for moderate subsampling depths is again small and the difference between the subsampling depths decreases as the considered fraction of the tail becomes smaller. The exponential distribution is the only alternative distribution with the opposite tendency: For the tested subsampled probability distributions, the power-law hypothesis tends to be favored over the exponential model for all subsampling depths, but the ratio tends closer to zero as the subsampling depth decreases.

Finally, we observe that for a given $\bar{F}(k_{\min})$, the limit value of the estimated power-law exponent α is in general larger for smaller subsampling depths. For some tested subsampled distributions with $\pi = 0.01$, the distribution is not likely to fill the strong evidence criteria for the power-law hypothesis because the values of α lie above the allowed range for almost all values of $k_{\min}$.

Overall, our results show that the goodness-of-fit test and the log-likelihood ratio tests lose some of their power in classifying the distribution's type correctly as the subsampling depth decreases.

V. DISCUSSION

In this work, we have investigated how well the maximum likelihood method of Ref. [6,7] and the extreme value method of Ref. [8] succeed in recognizing power-law distributions when the data are heavily subsampled with the incident subsampling strategy. As subsampled power-law distributions have been shown to approach the original power law asymptotically, we hypothesized that the methods would continue to work on subsampled data as long as the sample sizes remained reasonable. With the strong evidence criteria of the ML method, however, suboptimal estimation of the beginning of the power-law tail led to a substantial false rejection rate of the power-law hypothesis for the subsamples. While the EV method correctly recognized subsampled power-law dis-

tributions, it sometimes misclassified subsamples from both lognormal and stretched exponential distributions as power laws. However, these false positives tended to result not from the subsampling itself, but from the estimators' inability to classify the original sample correctly due to the underlying distribution converging too slowly to its asymptotic extreme value distribution.

Interestingly, we observed that while especially the lognormal distribution started to visually resemble a power law as the subsampling depth decreased, subsampling seemed to enhance the performance of the EV method in correctly classifying lognormal and stretched exponential distributions. This effect was visible especially for the Hill estimator; the fraction of probability mass in the distribution's tail allowing for correctly classifying the distribution's type was in general larger for lower subsampling depths, in some cases to the extent that the expected absolute number of nodes in this part of the tail increased. The moments and the kernel estimators followed the same trend, but for many of the tested distributions, the estimators started to oscillate at very low subsampling depths ($\pi = 0.01$), which resulted in the fraction getting seemingly smaller.

Overall, our results imply that the classifications obtained with the EV method should be accepted with some caution if very heavy-tailed distributions (such as the lognormal with $\sigma = 1.3$ or the stretched exponential distribution with $\beta = 0.3$) are valid alternatives for the power-law hypothesis. As noted already in Ref. [14], a result that the data belong to the MDA of the Gumbel distribution seems to be relatively reliable, while a classification to the MDA of the Fréchet distribution (thus supporting the power-law hypothesis) is not equally informative. It has been argued, however, that while it is often important to know whether a distribution is heavy-tailed, further identifying it as a power-law distribution may not bring any considerable additional value [34]. In some cases distinguishing between a power-law and a lognormal distribution might simply not be important or, alternatively, the lognormal and the stretched exponential distributions might not be relevant candidates for the question at hand. Consequently, if the main interest lies instead in, e.g., determining whether a sample originates from an exponential distribution or from a power-law distribution, the EV method may be a suitable alternative even if the data are heavily subsampled.

While the EV method is at times too permissible, the strong evidence criteria of the ML method avoid this drawback with the cost of an increased false rejection rate. Especially the requirement of the exponent α staying in the range [2,3] results in many false rejections for power laws with α close to the limits of this range. The simulations showed that most subsamples from a pure power law did not exhibit strong evidence for the power-law hypothesis due to suboptimal estimation of the start of the power-law tail. In addition, we showed that distinguishing the alternative distributions from power laws using the goodness-of-fit and the log-likelihood ratio tests of the ML method becomes increasingly difficult for lower subsampling depths. It is important to remember, however, that these results apply directly only to incident subgraph sampling, and other subsampling methods might produce substantially different results.

In general, while the automatic determination of the fraction of the tail considered in the analysis has it benefits—including the fact that no subjective determination of the threshold is needed—one should not trust this estimate blindly (a point raised with regard to the ML method already by, e.g., Refs. [8] and [35]). At the very least, it might be useful to examine more closely the range of values of $k_{\min}$ for which certain conclusions are valid; examining how the estimates change as a function of $k_{\min}$ might in some cases even offer further insight into the distribution's type as shown in Ref. [36]. The automatic estimation is especially likely to fail if the probability mass of the distribution is heavily concentrated on the small degrees, from which the PDF decays in a seemingly convex manner on a log-log-plot. However, as noted by Stumpf *et al.* [18], this kind of convex decay on a log-log plot is not commonly observed in real-world networks, and the problem might thus be overly pronounced in the simulated subsamples.

Overall, while our results highlight the importance of analyzing the same issue with different approaches, even using the two methods in combination does not always allow one to deduce with reasonable confidence whether a subsample originates from a power-law distribution. Naturally, the situation is likely to be even more complicated when analyzing real-world networks with more noise and variation. Consequently, assessing whether other methods—such as the maximum entropy test of Bee *et al.* [37], the Wilk's test used in Ref. [14], the finite-size scaling method of Ref. [9] or the approaches presented in Refs. [38], [39], and [40]—could fruitfully complement the methods addressed in this work remains a task for future research.

ACKNOWLEDGMENTS

We acknowledge the computational resources provided by the Aalto Science-IT project.

[1] A.-L. Barabási and R. Albert, Emergence of scaling in random networks, Science **286**, 509 (1999).

[2] P. Jelenković and J. Tan, Can retransmissions of superexponential documents cause subexponential delays? in *IEEE INFOCOM 2007 - 26th IEEE International Conference on Computer Communications*, Anchorage (IEEE, Los Alamitos, CA, 2007), pp. 892–900.

[3] S. Asmussen, P. Fiorini, L. Lipsky, T. Rolski, and R. Sheahan, Asymptotic behavior of total times for jobs that must start over if a failure occurs, Math. Operat. Res. **33**, 932 (2008).

[4] I. Kaj, L. Leskelä, I. Norros, and V. Schmidt, Scaling limits for random fields with long-range dependence, Ann. Probab. **35**, 528 (2007).

[5] R. van der Hofstad, G. Hooghiemstra, and D. Znamenski, Distances in random graphs with finite mean and infinite variance degrees, Elect. J. Probab. **12**, 703 (2007).

[6] A. Broido and A. Clauset, Scale-free networks are rare, Nat. Commun. **10**, 1017 (2019).

[7] A. Clauset, C. Shalizi, and M. Newman, Power-law distributions in empirical data, SIAM Rev. **51**, 661 (2009).

[8] I. Voitalov, P. van der Hoorn, R. van der Hofstad, and D. Krioukov, Scale-free networks well done, Phys. Rev. Res. **1**, 033034 (2019).

[9] M. Serafino, G. Cimini, A. Maritan, A. Rinaldo, S. Suweis, J. Banavar, and G. Caldarelli, True scale-free networks hidden by finite size effects, Proc. Natl. Acad. Sci. USA **118**, e2013825118 (2021).

[10] A. Levina and V. Priesemann, Subsampling scaling, Nat. Commun. **8**, 15140 (2017).

[11] H. Shimadzu and R. Darnell, Attenuation of species abundance distributions by sampling, R. Soc. Open Sci. **2**, 140219 (2015).

[12] N. Heikkilä, S. Sormunen, J. Mattila, T. Härkönen, M. Knip, E.-L. Ihantola, T. Kinnunen, I. Mattila, J. Saramäki, and T. Arstila, Generation of self-reactive, shared T-cell receptor α chains in the human thymus, J. Autoimmunity **119**, 102616 (2021).

[13] M. Stumpf and C. Wiuf, Sampling properties of random graphs: The degree distribution, Phys. Rev. E **72**, 036118 (2005).

[14] Y. Malevergne, V. Pisarenko, and D. Sornette, Empirical distributions of stock returns: Between the stretched exponential and the power law? Quant. Financ. **5**, 379 (2005).

[15] Y. Malevergne, V. Pisarenko, and D. Sornette, Testing the Pareto against the lognormal distributions with the uniformly most powerful unbiased test applied to the distribution of cities, Phys. Rev. E **83**, 036111 (2011).

[16] P. Montebruno, R. Bennett, C. Lieshout, and H. Smith, A tale of two tails: Do power law and lognormal models fit firm-size distributions in the mid-Victorian era? Physica A **523**, 858 (2019).

[17] S. Foss, D. Korshunov, and S. Zachary, *An Introduction to Heavy-Tailed and Subexponential Distributions* (Springer, New York, 2013).

[18] M. Stumpf, C. Wiuf, and R. May, Subnets of scale-free networks are not scale-free: Sampling properties of networks, Proc. Natl. Acad. Sci. USA **102**, 4221 (2005).

[19] J.-D. Han, D. Dupuy, N. Bertin, M. Cusick, and M. Vidal, Effect of sampling on topology predictions of protein-protein interaction networks, Nat. Biotechnol. **23**, 839 (2005).

[20] Q. Vuong, Likelihood ratio tests for model selection and non-nested hypotheses, Econometrica **57**, 307 (1989).

[21] M. Charras-Garrido and P. Lezaud, Extreme value analysis: An introduction, J. Soc. France Stat. **154**, 66 (2013).

[22] J. Nair, A. Wierman, and B. Zwart, *The Fundamentals of Heavy-tails: Properties, Emergence, and Identification* (Cambridge University Press, Cambridge, UK, 2022).

[23] T. Shimura, Discretization of distributions in the maximum domain of attraction, Extremes **15**, 299 (2012).

[24] B. M. Hill, A simple general approach to inference about the tail of a distribution, Ann. Statist. **3**, 1163 (1975).

[25] A. L. M. Dekkers, J. H. J. Einmahl, and L. de Haan, A moment estimator for the index of an extreme-value distribution, Ann. Statist. **17**, 1833 (1989).

[26] P. Groeneboom, H. P. Lopuhaä, and P. P. de Wolf, Kernel-type estimators for the extreme value index, Ann. Statist. **31**, 1956 (2003).

[27] See Supplemental Material at http://link.aps.org/supplemental/10.1103/PhysRevE.109.054308 for derivation of the analytical results as well as for further analysis on the estimators' performance. The Supplemental Material includes Refs. [28–30].

[28] E. Parzen, Quantile functions, convergence in quantile, and extreme value distribution theory, Technical Report B-3, Texas A&M University (1980) , Accession No. ADA093000.

[29] R. Bartle and J. Joichi, The preservation ofconvergence of measurable functions under composition, Proc. Am. Math. Soc. **12**, 122 (1961).

[30] S. Bobkov and M. Ledoux, One-dimensional empirical measures, order statistics, and Kantorovich transport distances, Memoirs Am. Math. Soc. **261**, 1 (2019).

[31] J. Alstott, E. Bullmore, and D. Plenz, Powerlaw: A Python package for analysis of heavy-tailed distributions, PLoS ONE **9**, e95816 (2014).

[32] H. Drees, A. Janßen, S. Resnick, and T. Wang, On a minimum distance procedure for threshold selection in tail analysis, SIAM J. Math. Data Sci. **2**, 75 (2020).

[33] A.-L. Barabási and M. Pósfai, *Network Science* (Cambridge University Press, Cambridge, UK, 2016).

[34] M. Stumpf and M. Porter, Critical truths about power laws, Science **335**, 665 (2012).

[35] Á. Corral, F. Font, and J. Camacho, Noncharacteristic half-lives in radioactive decay, Phys. Rev. E **83**, 066103 (2011).

[36] E. K. H. Salje, A. Planes, and E. Vives, Analysis of crackling noise using the maximum-likelihood method: Power-law mixing and exponential damping, Phys. Rev. E **96**, 042122 (2017).

[37] M. Bee, M. Riccaboni, and S. Schiavo, Pareto versus lognormal: A maximum entropy test, Phys. Rev. E **84**, 026104 (2011).

[38] Y. Zhang, E. D. Kolaczyk, and B. D. Spencer, Estimating network degree distributions under sampling: An inverse problem, with applications to monitoring social media networks, Ann. Appl. Statist. **9**, 166 (2015).

[39] Á. Corral and Á. González, Power law size distributions in geoscience revisited, Earth Space Sci. **6**, 673 (2019).

[40] I. Artico, I. Smolyarenko, V. Vinciotti, and E. Wit, How rare are power-law networks really? Proc. R. Soci. A: Math. Phys. Eng. Sci. **476**, 20190742 (2020).