Is Least-Squares Inaccurate in Fitting Power-Law Distributions? The Criticism is Complete Nonsense

Xiaoshi Zhong*
School of Computer Science
Beijing Institute of Technology, China
xszhong@bit.edu.cn

Muyin Wang
School of Computer Science
Beijing Institute of Technology, China
mynwang@bit.edu.cn

Hongkun Zhang
School of Computer Science
Beijing Institute of Technology, China
hkzhang@bit.edu.cn

ABSTRACT

Ordinary least-squares estimation is proved to be the best linear **unbiased** estimator according to the Gauss-Markov theorem. In the last two decades, however, some researchers criticized that least-squares was substantially inaccurate in fitting power-law distributions; such criticism has caused a strong bias in research community. In this paper, we conduct extensive experiments to rebut that such criticism is complete nonsense. Specifically, we sample different sizes of discrete and continuous data from power-law models, showing that even though the long-tailed noises are sampled from power-law models, they cannot be treated as power-law data. We define the correct way to bin continuous power-law data into data points and propose an average strategy for least-squares to fit power-law distributions. Experiments on both simulated and real-world data show that our proposed method fits power-law data perfectly. We uncover a fundamental flaw in the popular method proposed by Clauset et al. [12]: it tends to discard the majority of power-law data and fit the long-tailed noises. Experiments also show that the reverse cumulative distribution function is a bad idea to plot power-law data in practice because it usually hides the true probability distribution of data. We hope that our research can clean up the bias about least-squares fitting power-law distributions.

Source code can be found at https://github.com/xszhong/LSavg.

CCS CONCEPTS

• **Information systems** → *Network science*; Scale-free networks.

KEYWORDS

Power-law distributions, least-squares estimation (LSE), average strategy, long-tailed noises

ACM Reference Format:
Xiaoshi Zhong, Muyin Wang, and Hongkun Zhang. 2022. Is Least-Squares Inaccurate in Fitting Power-Law Distributions? The Criticism is Complete Nonsense. In *Proceedings of the ACM Web Conference 2022 (WWW'22), April 25–29, 2022, Virtual Event, Lyon, France.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3485447.3511995

*X. Zhong is the corresponding author. M. Wang and H. Zhang contribute equally.

WWW'22, April 25–29, 2022, Virtual Event, Lyon, France

ACM ISBN 978-1-4503-9096-5/22/04…$15.00
https://doi.org/10.1145/3485447.3511995

1 INTRODUCTION

Power-law distributions have been reported to appear in a variety of natural and societal systems [1, 2, 5, 7, 8, 15, 16, 20, 21, 23, 25, 29–34, 39, 41, 43, 54, 56, 58, 60]. It is important to correctly estimate power-law models so as to understand the characteristics of these systems. Ordinary least-squares estimation (LSE) is proved to be the best linear **unbiased** estimator with minimum variance among the class of linear unbiased estimators according to the Gauss-Markov theorem [14, 28, 42] (the Gauss-Markov theorem can be easily found in a textbook about regression analysis). In the last two decades, however, some researchers conducted **inappropriate** experiments to criticize LSE for being substantially inaccurate in fitting power-law distributions to data [6, 11, 12, 24, 36, 55]. Such criticism has caused a strong bias in research community about using least-squares to estimate parameters of power-law models. Recently, researchers who are concerned with power-law distributions tend to start their research with statements like these: "This naive form of linear regression generates significant errors under relatively common conditions and gives no warning of its mistakes" [52], "intrinsic limitations of the least square (LS) fits to logarithmically scaled data" [27], and similar others [4, 9, 18, 19, 44, 57]. What is even worse is that some reviewers may start their reviews on a submission with comments like this: "This approach (authors note: namely LSE) to fitting power-law distributions to data is entirely unreliable, and its results cannot be trusted." In this paper, we conduct extensive experiments to rebut such criticism, and experimental results demonstrate that such criticism is complete nonsense and that the popular CSN2009 method proposed by Clauset et al. [12] is actually misleading and fundamentally flawed.

Specifically, we sample both discrete and continuous data from power-law models (see Eq. (1)) with scaling exponent $\alpha = 2.5$ and minimal value $x_{min} = 1$, similar to the criticizers [6, 11, 12, 24, 36, 55]; but unlike those criticizers who only sample data in a specific size (e.g., 10^4 or 10^6), we sample data in different sizes from 10^3 to 10^8 so as to fully illustrate the characteristics of these sampled data. The statistics of both discrete and continuous sampled data shows that not all the data sampled from a power-law model follow a power-law distribution and that the long tails of finite samples are sampling noises and cannot be treated as power-law data. Experiments also show that it is the long-tailed noises causing the inaccuracy of LSE in fitting power-law distributions. Those criticizers mistakenly treat the data problem as the model problem.

We propose an average strategy for ordinary LSE termed LS_{avg} to estimate parameters of power-law models (see Section 3.1). The idea behind the average strategy is that if a set of data points perfectly follows a power-law distribution, then using any subset of two or more data points can find the optimum slope of the line on

doubly logarithmic system, and all these slopes and their arithmetic average are equal. For a discrete sample, it is denoted by a set of data points and we directly apply LS_{avg} to these data points after filtering out the long-tailed noises. For a continuous sample, we firstly bin the sample into a set of data points by our proposed binning method as described in Section 3.2 and then apply LS_{avg} to these data points after filtering out the long-tailed noises. Unlike most previous research that use the centers of the bins, we use the geometric averages of bins to precisely represent the binned data. Experimental results demonstrate that our LS_{avg} achieves unbiased estimation of power-law exponents on sampled discrete data and almost unbiased estimation on sampled continuous data.

To examine the goodness of power-law fit, we propose to use the maximal statistic of the two-sample Kolmogorov-Smirnov (KS) test [48, 49] among a large group of samples that are drawn from the estimated power-law model as the threshold to determine whether or not to accept the power-law hypothesis (see Section 3.3). Our proposed method overcomes the problem of the traditional KS test in mistakenly rejecting the hypothesis that two samples drawn from the same power-law model follow the same distribution.

We apply our proposed estimation method LS_{avg} and hypothesis-testing method to real-world datasets and compare our methods with the widely known CSN2009 [12]. Experiments demonstrate that our methods successfully fit and identify power-law distributions to empirical data. By contrast, CSN2009 tends to discard the majority of data and fit the long-tailed noises, and the hypothesis-testing method proposed by Clauset et al. [12] mistakenly treats the long-tailed noises as being well-fitted by power-law distributions. Experiments also demonstrate that the reverse cumulative distribution function (rCDF) widely used to plot power-law data is a bad idea in practice because it usually hides the true probability density function (PDF) of data.

To summarize, we make in this paper the following contributions.

- To the best of our knowledge, this paper is the first work to conduct extensive experiments to rebut the criticism about LSE in fitting power-law distributions. Experimental results show that LSE fits power-law data perfectly. We hope that our research can clean up the bias in the research community about LSE fitting to power-law distributions.
- We propose an average strategy for LSE to fit power-law distributions to data, and experiments demonstrate its remarkable success in fitting power-law distributions to both simulated power-law data and real-world data.
- We define the correct way to bin continuous power-law data into data points for LSE to fit power-law distributions, which has never been correctly reported in previous research.
- We find that the traditional KS test fails to examine whether two power-law samples are drawn from the same distribution. To resolve the problem, we propose to set a threshold derived from a large group of generated power-law samples to examine the power-law hypothesis. Experiments on real-world data demonstrate the effectiveness of our proposed method.
- We uncover a fundamental flaw in the widely known method CSN2009 proposed by Clauset et al. [12]: it tends to discard the majority of power-law data and fit the long-tailed noises. Such flaw invalidates the reliability of all the research based on CSN2009 and all those works need to be re-investigated.

2 RELATED WORKS

Since there is no existing research rebutting the criticism about LSE in fitting power-law distributions [6, 11, 12, 24, 36, 55], we analyze previous research about fitting power-law distributions.

2.1 Parameter Estimation of Power-Law Model

2.1.1 Least-Squares Estimation. There are mainly two kinds of binning methods used for LSE [17, 45, 51]: linear binning and logarithmic binning. Linear binning bins observed data into fixed-width histograms, and then plot the frequencies against the centers of the bins [21, 22, 30, 34, 40]. By contrast, logarithmic binning uses logarithmic bins that have fixed width on doubly logarithmic system. Comparing with linear binning, logarithmic binning can reduce the number of low-frequency bins [8, 26, 47].

Our analysis and those criticizers show that linear binning with bin center representation on the whole data (including long-tailed noises) results in substantially biased estimation. While logarithmic binning obtains more accurate estimation, it cannot rule out the long-tailed noises. In addition, without using the correct representation of bins, logarithmic binning cannot accurately estimate the constant parameter of a power-law model.

2.1.2 Maximum Likelihood Estimation. The maximum likelihood estimation (MLE) has been the most frequently reported method to fit power-law distributions since the criticizers conducted inappropriate experiments to criticize LSE [6, 11, 12, 24, 36, 55]. MLE mainly estimates the parameters that maximize the likelihood of the model given the observed data [12, 36, 59]. The CSN2009 method [12] seemed to become a "standard" for fitting power-law distributions.

Our analysis uncovers a fundamental flaw in CSN2009 [12]: it tends to discard the majority of data (even though they are sampled from a power-law model) and fit the long-tailed noises. Such flaw invalidates the reliability of all the research based on CSN2009.

2.1.3 Cumulative Distribution Function. Some researchers estimate parameters of power-law distributions by estimating their CDF ($P_{cdf}(x) = \int_{-\infty}^{x} p(t)dt$) or rCDF ($P_{rcdf}(x) = \int_{x}^{\infty} p(t)dt$) [6, 8, 12, 36, 40, 46, 53]. Newman [36] advocates to use rCDF to plot power-law distributions on doubly logarithmic system, and such rCDF plot has become popular in visualizing power-law data.

Our experiments show that rCDF is a bad idea to plot power-law data in practice, because it usually hides the true PDF of data.

2.2 Goodness-of-Fit Test

The KS test [48, 49] is widely used for power-law hypothesis testing [3, 12, 24, 27, 35, 52, 53]. The Chi-squared test [13, 38] is also used for power-law hypothesis testing [6, 11]. Clauset et al. [12] propose to sample a large number of synthetic data from estimated power-law model to calculate p-value for hypothesis test.

Our experiments show that the p-value of traditional KS test fails to test power-law hypothesis in practice (see Section 3.3 and Table 1), and that the hypothesis-testing method proposed by Clauset et al. [12] fails to reject the hypothesis that the long-tailed noises follow a power-law distribution (see Section 5, A.3, and Table 10). We propose to set a KS statistic threshold derived from a large group of samples drawn from the estimated power-law model to test power-law hypothesis. Experiments on real-world data demonstrate that our proposed method is effective in power-law hypothesis testing.

3 METHODOLOGY

For a set of data points, if they follow a power-law distribution, then they can be characterized by Eq. (1):

$$p(x) = K \cdot x^{-\alpha} \tag{1}$$

where α is the scaling exponent and K is the constant. Like many other methods, we mainly discuss the situation where $x \geq 1$, $\alpha > 1$, and $K > 0$. Mathematically, Eq. (1) possesses Property 1.

Property 1. *Strictly decreasing property of function $f(x)$ on domain $\mathbb{D}$: $\forall x_i, \forall x_j \in \mathbb{D}$, if $x_i < x_j$, then $f(x_i) > f(x_j)$.*

If a data point satisfies Property 2, then we say that the data point possesses the strictly decreasing property.

Property 2. *Strictly decreasing property of data point $(x_i, f(x_i))$ in the set of data points $\{(x, f(x))\}$ on domain $\mathbb{D}$: $\forall x \in \mathbb{D}$, if $x < x_i$, then $f(x) > f(x_i)$ and if $x_i < x$, then $f(x_i) > f(x)$.*

3.1 Least-Squares Estimation with Average Strategy for Power-Law Distributions

Eq. (1) is equivalent to Eq. (2):

$$s = \alpha \cdot t + c \tag{2}$$

where $s = -\log p(x)$, $t = \log x$, and $c = -\log K$.

Given data points $\{(t_i, s_i)\}$, $1 \leq i \leq N$, that follow a line as Eq. (2), using LSE to find the optimum value of α is to solve Eq. (3):

$$\hat{\alpha} = \arg\min_{\alpha} \sum_{i=1}^{N} (s_i - (\alpha \cdot t_i + c))^2 \tag{3}$$

Setting the gradient of the loss function to zero, we get $\hat{\alpha}$ by Eq. (4):

$$\hat{\alpha} = \frac{\sum_{i=1}^{N} (t_i - \bar{t})(s_i - \bar{s})}{\sum_{i=1}^{N} (t_i - \bar{t})^2} \tag{4}$$

where $\bar{t} = \frac{1}{N} \sum_{i=1}^{N} t_i$ and $\bar{s} = \frac{1}{N} \sum_{i=1}^{N} s_i$.

If the set of data points $\{(t_i, s_i)\}$, $1 \leq i \leq N$, can be perfectly fitted by Eq. (2), then using any subset of two or more data points can find the optimum $\hat{\alpha}$, and all these optimum $\hat{\alpha}$ and their arithmetic average are equal. Let $\hat{\alpha}_j$ denote the optimum $\hat{\alpha}$ estimated by using the first j data points, where $2 \leq j \leq N$. Under perfect-fitting condition, $\hat{\alpha}_2 = \cdots = \hat{\alpha}_N = \hat{\alpha}_{avg}$, where $\hat{\alpha}_{avg}$ is defined by Eq. (5):

$$\hat{\alpha}_{avg} = \frac{1}{N-1} \sum_{j=2}^{N} \hat{\alpha}_j \tag{5}$$

While there is rare perfect fitting in practice and in most cases not all the $\hat{\alpha}_j$ are equal, using their arithmetic average can reduce the impact of those data points deviated from the line. We use LS_{norm} to denote the normal LSE defined by Eq. (4) while use LS_{avg} to denote the average-strategy LSE defined by Eq. (5). (While this paper mainly discusses $\alpha > 1$, LS_{norm} and LS_{avg} apply to $\alpha > 0$.)

3.2 Binning for Continuous Power-Law Data

For a continuous sample, if it perfectly follows a power-law distribution, then it forms a perfect straight-line on doubly logarithmic system. Figure 1 illustrates a power-law distribution spanned by the bin from t_{b_i} to $t_{b_{i+1}}$ on doubly logarithmic system. The integral of the line on the bin equals to the areas of the trapezoid $ABDC$ and of the rectangle $ABFE$; namely $\mathcal{A}_{ABDC} = \mathcal{A}_{ABFE}$, where $\mathcal{A}$ indicates the area. The point $G\ (t_{c_i}, s_{c_i})$ is the center of the line segment CD where $t_{c_i} = \frac{1}{2}(t_{b_i} + t_{b_{i+1}})$ and $s_{c_i} = s(t_{c_i}) = \frac{\mathcal{A}_{ABDC}}{|AB|}$. Suppose the marginal values of the bin on original coordinate system are

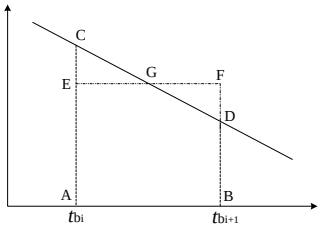


Figure 1: Illustration of a continuous power-law distribution spanned by the bin from t_{b_i} to $t_{b_{i+1}}$ on the doubly logarithmic system. G is the center of the line segment CD. The areas of the trapezoid $ABDC$ and the rectangle $ABFE$ are equal.

Table 1: Maximal $D_{n,n}$ values and their p-values of the two-sample KS test of ten groups of samples drawn from the power-law model with $\alpha = 2.5$, $x_{min} = 1$, and $n = 10^4$ and $n = 10^6$. "Size" indicates the number of samples in the group.

Size	Discrete				Continuous			
	$D_{10^4,10^4}$	p_{value}	$D_{10^6,10^6}$	p_{value}	$D_{10^4,10^4}$	p_{value}	$D_{10^6,10^6}$	p_{value}
100	0.0214	0.0205	0.0020	0.0332	0.0313	1.1e-4	0.0027	1.5e-3
200	0.0272	0.0012	0.0021	0.0227	0.0349	1.0e-5	0.0029	3.6e-4
300	0.0272	0.0012	0.0025	0.0032	0.0367	2.8e-6	0.0031	1.5e-4
400	0.0273	0.0012	0.0025	0.0032	0.0367	2.8e-6	0.0035	8.0e-6
500	0.0273	0.0012	0.0027	0.0015	0.0401	2.1e-6	0.0035	8.0e-6

x_{b_i} and $x_{b_{i+1}}$, where $t_{b_i} = \log x_{b_i}$ and $t_{b_{i+1}} = \log x_{b_{i+1}}$. Applying $t_{b_i} = \log x_{b_i}$ and $t_{b_{i+1}} = \log x_{b_{i+1}}$ to t_{c_i} we get Eq. (6).

$$t_{c_i} = \frac{1}{2}(t_{b_i} + t_{b_{i+1}}) = \log\left(x_{b_i} \cdot x_{b_{i+1}}\right)^{\frac{1}{2}} \tag{6}$$

Its value on original coordinate system is given by Eq. (7).

$$x_{c_i} = e^{t_{c_i}} = \left(x_{b_i} \cdot x_{b_{i+1}}\right)^{\frac{1}{2}} \tag{7}$$

$s_{c_i} = s(t_{c_i})$ and equals to the height of rectangle $ABFE$ as given by Eq. (8). Its value on original coordinate system is given by Eq. (9).

$$s_{c_i} = \frac{\mathcal{A}_{ABDC}}{|AB|} = \frac{\int_{t_{b_i}}^{t_{b_{i+1}}} s(t)dt}{|t_{b_{i+1}} - t_{b_i}|} = s(t_{c_i}) \tag{8}$$

$$y_{c_i} = e^{-s_{c_i}} = e^{-s(t_{c_i})} = p(x_{c_i}) \tag{9}$$

Since $\mathcal{A}_{ABDC} = \mathcal{A}_{ABFE}$ is always true, if the sample perfectly follows a power-law distribution, then Eq. (6) and Eq. (8) as well as Eq. (7) and Eq. (9) apply to arbitrary binning. The bin is represented by the point $G\ (t_{c_i}, s_{c_i})$ on doubly logarithmic system and the corresponding point $G'\ (x_{c_i}, y_{c_i})$ on original coordinate system.

3.3 Power-Law Hypothesis Test

We use the Kolmogorov-Smirnov (KS) test [48, 49] to examine whether to accept the hypothesis: the investigated data follow a power-law distribution. The KS statistic (D_n) is defined by Eq. (10):

$$D_n = \sup_x |F_n(x) - F(x)| \tag{10}$$

where $F_n(x)$ is the CDF of a sample or a set of data points, $F(x)$ is the CDF of the theoretic distribution, sup_x is the supremum of the set of distances, and $D_n \in [0, 1]$.

The two-sample KS statistic ($D_{n,m}$) is defined by Eq. (11):

$$D_{n,m} = \sup_x |F_n(x) - F_m(x)| \tag{11}$$

where $F_n(x)$ and $F_m(x)$ are the CDF of two samples.

Table 2: Statistics of different sizes of discrete data that are sampled from a power-law model with $\alpha = 2.5$ and $x_{min} = 1$. X_{5th} indicates the fifth data point; X_f indicates the last data point that all the data points of $\{X_f\}$ satisfy Property 2 while the next data point does not satisfy Property 2; X_1 indicates the first data point whose frequency is $\frac{1}{n}$; $\hat{X}_1^T$ the estimated X_1^T using the $\hat{\alpha}_f^{avg}$ described in Section 4.2; and X_{max} indicates the last data point. $Count(1)$ denotes the number of data points whose $f(x) = \frac{1}{n}$ and $Count(0)$ the number of x-values where $f(x) = 0$ for $x \leq X_{max}$. $Rate(X)$ denotes the total frequency of data points $\{X\}$. For each sample size, statistical results are reported by *mean* ± *standard deviation* based on 500 samples.

Size	X_{5th}	$Rate(X_{5th})$	X_f	$Rate(X_f)$	X_1	$Rate(X_1)$	$\hat{X}_1^T$	X_{max}	$Count(1)$	$Count(0)$
10^3	5.0±0.0	0.9619±6.2e-3	5.5±1.3	0.9640±1.5e-2	11.7±2.7	0.9877±5.2e-3	14.2±1.1	145.48±338.8	8.3±2.6	124.3±338.4
10^4	5.0±0.0	0.9617±2.0e-3	9.1±1.8	0.9824±5.7e-3	24.2±3.6	0.9958±1.1e-3	35.5±1.2	749.6±1697.6	21.1±4.1	697.0±1697.5
10^5	5.0±0.0	0.9617±6.2e-4	15.2±2.7	0.9915±2.4e-3	54.6±6.3	0.9988±2.4e-4	88.9±1.3	3486.0±10643.0	53.0±6.2	3353.9±10643.0
10^6	5.0±0.0	0.9617±1.9e-4	24.6±3.8	0.9959±9.6e-4	125.3±12.7	0.9996±6.2e-5	223.3±1.3	14645.8±20550.6	133.2±10.3	14313.1±20549.5
10^7	5.0±0.0	0.9619±5.9e-5	38.4±5.3	0.9979±4.4e-4	292.1±25.6	0.9999±1.5e-5	561.0±1.4	73029.1±150829.3	334.1±15.8	72194.5±150828.9
10^8	5.0±0.0	0.9617±2.0e-5	61.7±7.7	0.9990±2.0e-4	691.9±49.8	0.99997±4.5e-6	1409.1±1.8	515207.4±3735626.3	840.6±24.8	513107.5±3735626.0

Table 3: Statistics of binned continuous data (width=1) that are sampled from a power-law model with $\alpha = 2.5$ and $x_{min} = 1$

Size	X_{5th}	$Rate(X_{5th})$	X_f	$Rate(X_f)$	X_1	$Rate(X_1)$	$\hat{X}_1^T$	X_{max}	$Count(1)$	$Count(0)$
10^3	5.48±7.0e-4	0.9328±8.3e-3	6.5±1.5	0.9430±2.3e-2	14.8±2.9	0.9835±6.1e-3	18.6±1.4	84.4±57.5	10.1±3.0	57.6±56.3
10^4	5.48±6.4e-5	0.9320±2.5e-3	10.7±2.1	0.9716±8.7e-3	31.4±4.4	0.9943±1.4e-3	46.2±1.3	414.1±264.0	26.9±4.6	345.4±263.0
10^5	5.48±6.1e-6	0.9320±8.5e-4	17.2±3.0	0.9859±3.7e-3	69.5±7.8	0.9983±3.3e-4	115.9±1.5	2008.3±1762.5	69.0±7.9	1834.5±1761.1
10^6	5.48±6.6e-7	0.9320±2.5e-4	28.1±4.3	0.9932±1.6e-3	161.9±14.8	0.9995±7.2e-5	291.8±1.5	9194.7±8943.2	174.6±11.8	8756.2±8941.8
10^7	5.48±6.3e-8	0.9320±8.2e-5	44.7±5.8	0.9966±6.8e-4	377.2±30.5	0.9999±1.8e-5	734.0±1.9	41488.1±16322.9	440.3±18.4	40385.1±39880.8
10^8	5.48±6.1e-9	0.9320±2.6e-5	71.0±7.9	0.9983±3.0e-4	892.6±62.1	0.99996±5.1e-6	1846.3±2.6	187714.8±155880.0	1108.1±29.85	184941.6±155883.3

In Eq. (10), the null hypothesis (H_0) is that the sample is drawn from the theoretic distribution, while the alternative (H_1) is that it is not from the theoretic distribution. In Eq. (11), the null hypothesis (H_0') is that the two samples are drawn from the same distribution, while the H_1' is that they are not from the same distribution.

We find the KS statistic useful but its derived p-value fails to test power-law hypothesis in practice. Table 1 reports the maximal two-sample KS statistics and their p-values of ten groups of samples that are drawn from power-law models with $\alpha = 2.5$, $x_{min} = 1$, and $n = 10^4$ and $n = 10^6$. All the p-values are less than 0.05, and they reject the hypothesis H_0'. However, these samples are indeed drawn from the same power-law model. Such contradiction indicates that KS test's p-value fails to examine power-law hypothesis. To resolve the failure of KS test's p-value, we set the maximal two-sample KS statistic among a large group of samples that are drawn from a power-law model with n as the threshold D_n^T to determine whether or not to accept the hypothesis. D_n^T is defined by Eq. (12).

$$D_n^T = \max\{D_n^{i,j}\} \tag{12}$$

where $D_n^{i,j}$ is the KS statistic of the i-th and j-th samples.

We can get the true D_n^T only when the group size approaches infinity, which is impossible in practice. Empirically, we find to get a reasonable D_n^T when the group size reaches 300, given that the calculation of D_n^T for large samples is time-consuming.

In practice, we test power-law hypothesis by the following strategy. Firstly, estimate a power-law model from a dataset and calculate D_n between the dataset and the estimated power-law model. Secondly, draw 300 samples from the estimated power-law model with the same sample size as the dataset and calculate D_n^T among the 300 samples. Finally, compare D_n and D_n^T to examine the hypothesis H_0: if $D_n \leq D_n^T$, then **accept** H_0; if $D_n^T < D_n \leq 2 \times D_n^T$, then **moderately accept** H_0; and if $2 \times D_n^T < D_n$, **reject** H_0.

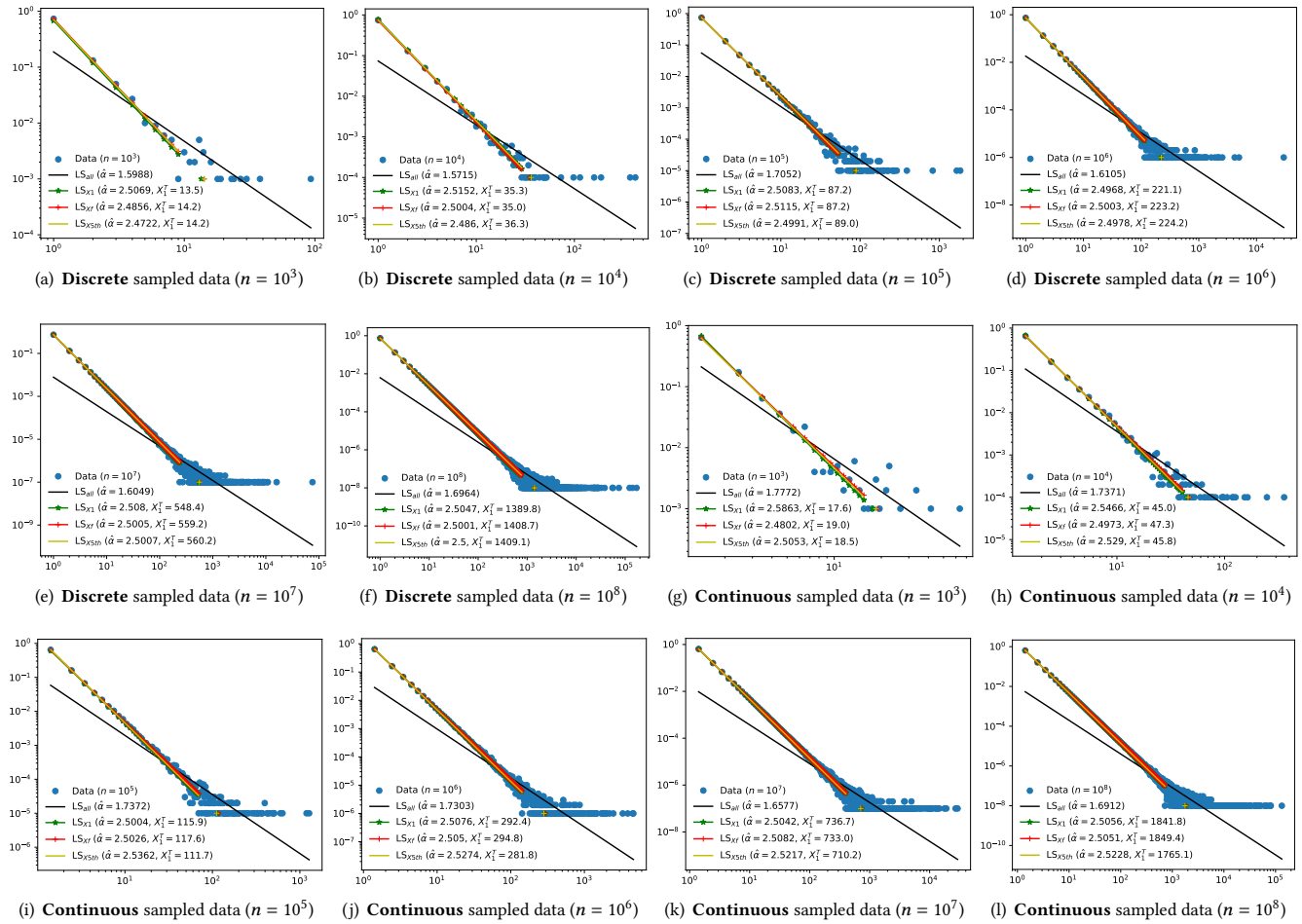
4 FITTING POWER-LAW DISTRIBUTIONS TO DATA DRAWN FROM POWER-LAW MODEL

4.1 Sampling Data from Power-Law Models and their Critical Statistics

Similar to those criticizers [6, 11, 12, 24, 36, 55], we sample data from power-law models with scaling exponent $\alpha = 2.5$ and minimal value $x_{min} = 1$, under both discrete and continuous types in different sizes[1]: 10^3, 10^4, 10^5, 10^6, 10^7, and 10^8. Each continuous sample is binned into a set of data points with $width = 1$ by the method described in Section 3.2, so that these data points are comparable to the ones of discrete samples. For each sample size, we draw 500 samples on which statistics and experimental results are based. After continuous samples are binned, each sample is denoted by a set of data points $\{(x_i, f(x_i))\}$ in ascending order by their x-values, where $f(x_i)$ is the frequency of x_i in the sample.

For each set of data points, we set four critical data points whose x-values are $X_{5th}, X_f, X_1, X_{max}$. For convenience, we use the four critical x-values to represent their corresponding data points and use them with curly brackets to represent the subset of data points from the first data point to them; for example, $\{X_{5th}\}$ represents the set of the first five data points while $\{X_1\}$ represents the set of data points from the first one to X_1. The definitions of these critical data points are as follows. X_{5th} indicates the fifth data point. X_f indicates the last data point that all the data points of $\{X_f\}$ satisfy Property 2; while its very first subsequent data point does not satisfy Property 2. $X_1 = \min\{x_i | f(x_i) = \frac{1}{n}\}$ indicates the first data point whose frequency is $\frac{1}{n}$. $X_{max} = \max\{x_i\}$ indicates the last data point. According to the definitions of X_f, X_1, and X_{max}, the relationship among them is $X_f \leq X_1 \leq X_{max}$.

[1] We use the *numpy.random.zipf* function in the NumPy package to generate discrete samples and use the *randht* module described in Clauset et al. [12] to generate continuous samples.



(a) **Discrete** sampled data ($n = 10^3$) (b) **Discrete** sampled data ($n = 10^4$) (c) **Discrete** sampled data ($n = 10^5$) (d) **Discrete** sampled data ($n = 10^6$)

(e) **Discrete** sampled data ($n = 10^7$) (f) **Discrete** sampled data ($n = 10^8$) (g) **Continuous** sampled data ($n = 10^3$) (h) **Continuous** sampled data ($n = 10^4$)

(i) **Continuous** sampled data ($n = 10^5$) (j) **Continuous** sampled data ($n = 10^6$) (k) **Continuous** sampled data ($n = 10^7$) (l) **Continuous** sampled data ($n = 10^8$)

Figure 2: LSE methods fit different sizes of discrete and binned (width=1) continuous data that are sampled from power-law models with $\alpha = 2.5$ and $x_{min} = 1$. LS_{all} indicates applying LS_{norm} on all the data to estimate α. LS_{X5th} indicates applying LS_{avg} on $\{X_{5th}\}$ to estimate α; LS_{Xf} applying LS_{avg} on $\{X_f\}$; and LS_{X1} applying LS_{avg} on $\{X_1\}$. $X_1^T = (\hat{K} \cdot n)^{\frac{1}{\hat{\alpha}}}$ is the x-value of $(X_1^T, \frac{1}{n})$.

Suppose the sampling is perfect (which means the sampled data perfectly follow a power-law distribution and possess Property 1), the frequency of the lowest-frequency data point is $\frac{1}{n}$ and let this data point be $(X_1^T, \frac{1}{n})$, then solving $p(X_1^T) = K \cdot \frac{1}{n}$, we get $X_1^T = (K \cdot n)^{\frac{1}{\alpha}}$. That means $X_1^T \propto n^{\frac{1}{\alpha}}$, and when $n \to \infty$, $X_1^T \to \infty$. If the sampling is perfect, then $X_f = X_1 = X_1^T = X_{max}$ according to their definitions. For a finite-size sample, however, the sampling is not perfect in reality and these critical values are not equal. We will see that their empirical relationship is $X_f < X_1 < X_1^T \ll X_{max}$.

Table 2 and 3 summarize the statistics of these sampled discrete and continuous data. Based on these critical data points we report some critical statistics about the data: $Rate(X_{5th})$, $Rate(X_f)$, $Rate(X_1)$, $Count(1)$, and $Count(0)$. $Rate(X) = \sum_{x_i \le X} f(x_i)$ denotes the total frequency of $\{X\}$. $Count(1) = \sum_{x_i \le X_{max}} \mathbb{I}(f(x_i) = \frac{1}{n})$ denotes the number of $\frac{1}{n}$-frequency data points, where $\mathbb{I}(\cdot)$ is the indicator function. $Count(0) = \sum_{x_i \le X_{max}} \mathbb{I}(f(x_i) = 0)$ denotes the number of 0-frequency data points within X_{max}. $\hat{X}_1^T$ is the estimated X_1^T using $\hat{\alpha}_f^{avg}$ as described in Section 4.2. Figure 2 plots example discrete and binned continuous samples (indicated by "Data" and in the blue font) on doubly logarithmic system.

Table 2 and 3 show that whatever the sample size is, X_{5th} and $Rate(X_{5th})$ of both discrete and continuous data are almost invariant: $Rate(X_{5th})$ is around 96.17% for discrete data and around 93.20% for continuous data. The high $Rate(X_{5th})$ indicates that the first five data points contain most of the data. When the magnitude of the sample size increases, X_f, X_1, and X_{max} increase dramatically. $Rate(X_f)$ and $Rate(X_1)$ are even higher than $Rate(X_{5th})$; they increase steadily from 94.30%~96.40% and 98.35%~98.77% at $n = 10^3$ to 99.83%~99.90% and 99.996%~99.997% at $n = 10^8$. This indicates that $\{X_1\}$ contains almost all the data of a sample.

At each sample size, X_{max} is significantly greater than X_1. However, the data points from X_1 to X_{max} contain less than 1.65% of data. It should be noted that (almost) all the data points counted in $Count(1)$ and $Count(0)$ appear between X_1 and X_{max}. $Count(1)$ indicates that the less than 1.65% of data are mainly composed of $\frac{1}{n}$-frequency data points. $Count(0)$ is significantly larger than $Count(1)$ and it is comparable with $|X_{max} - X_1|$. This means, for

Table 4: Estimated α of fitting different sizes of discrete data that are sampled from a power-law model with $\alpha = 2.5$ and $x_{min} = 1$. For each size, experimental results are reported by *mean* ± *standard deviation* based on 500 samples.

Size	$\hat{\alpha}_{5th}$	$\hat{\alpha}_{5th}^{avg}$	$\hat{\alpha}_f$	$\hat{\alpha}_f^{avg}$	$\hat{\alpha}_{X_1}$	$\hat{\alpha}_{X_1}^{avg}$	$\hat{\alpha}_{all}$
10^3	2.5185±0.1321	2.5100±0.0855	2.4923±0.1140	2.4972±0.0820	2.5428±0.1809	2.5102±0.0750	1.5896±0.2868
10^4	2.5013±0.0426	2.5008±0.0263	2.4962±0.0445	2.4990±0.0244	2.5560±0.0800	2.5120±0.0288	1.5953±0.1855
10^5	2.4993±0.0126	2.4996±0.0078	2.5005±0.0184	2.5001±0.0079	2.5393±0.0392	2.5096±0.0133	1.6510±0.1183
10^6	2.4999±0.0043	2.5000±0.0026	2.5003±0.0072	2.5002±0.0027	2.5319±0.0210	2.5081±0.0072	1.6609±0.0732
10^7	2.5001±0.0012	2.5000±0.0007	2.5000±0.0029	2.5000±0.0010	2.5240±0.0115	2.5065±0.0040	1.6765±0.0472
10^8	2.5000±0.0005	2.5000±0.0003	2.5001±0.0013	2.5000±0.0004	2.5200±0.0070	2.5052±0.0023	1.6795±0.0290

Table 5: Estimated α of fitting binned continuous data (width=1) that are sampled from a power-law model with $\alpha = 2.5$

Size	$\hat{\alpha}_{5th}$	$\hat{\alpha}_{5th}^{avg}$	$\hat{\alpha}_f$	$\hat{\alpha}_f^{avg}$	$\hat{\alpha}_{X_1}$	$\hat{\alpha}_{X_1}^{avg}$	$\hat{\alpha}_{all}$
10^3	2.5258±0.1282	2.5315±0.0886	2.5034±0.1114	2.5189±0.0839	2.5573±0.1735	2.5261±0.0749	1.5151±0.2734
10^4	2.5199±0.0404	2.5225±0.0285	2.5117±0.0402	2.5173±0.0226	2.5524±0.0743	2.5233±0.0271	1.5559±0.1588
10^5	2.5165±0.0128	2.5225±0.0088	2.5084±0.0160	2.5141±0.0079	2.5428±0.0351	2.5164±0.0120	1.6172±0.1083
10^6	2.5170±0.0038	2.5224±0.0027	2.5049±0.0062	2.5101±0.0026	2.5303±0.0182	2.5108±0.0064	1.6511±0.0679
10^7	2.5168±0.0012	2.5222±0.0008	2.5033±0.0025	2.5076±0.0012	2.5242±0.0101	2.5081±0.0033	1.6698±0.0394
10^8	2.5168±0.0004	2.5223±0.0003	2.5023±0.0011	2.5058±0.0006	2.5195±0.0063	2.5062±0.0020	1.6798±0.0259

Table 6: Fitting results of CSN2009 on data sampled from power-law models with $\alpha = 2.5$ and $x_{min} = 1$. The "Discrete ($\hat{x}_{min} \geq 2$)" indicates that results are based on discrete samples in which CSN2009 gets $\hat{x}_{min} \geq 2$ and "Cnt" indicates the count of such $\hat{x}_{min} \geq 2$ samples. "Continuous ($Cover < 50\%$)" indicates that results are based on continuous samples in which CSN2009 covers less than 50% of data and "Cnt" indicates the count of such $Cover < 50\%$ samples. "Discrete" and "Continuous" indicate the results are based on the whole 500 samples. "$Cover$" denotes the percentage of data covered by CSN2009.

Size	Discrete			Discrete ($\hat{x}_{min} \geq 2$)				Continuous			Continuous ($Cover < 50\%$)			
	$\hat{\alpha}$	$\hat{x}_{min}$	$Cover$	$\hat{\alpha}$	$\hat{x}_{min}$	$Cover$	Cnt	$\hat{\alpha}$	$\hat{x}_{min}$	$Cover$	$\hat{\alpha}$	$\hat{x}_{min}$	$Cover$	Cnt
10^3	2.5085±0.0645	1.0340	97.65%	2.6325±0.1505	2.0625	26.48%	16	2.5082±0.0656	1.1784	84.56%	2.5436±0.1155	2.0258	37.62%	40
10^4	2.5003±0.0195	1.0180	98.66%	2.5200±0.0686	2.0000	25.54%	9	2.5006±0.0204	1.1652	85.85%	2.5069±0.0423	1.9708	37.93%	47
10^5	2.4998±0.0062	1.0180	98.66%	2.5033±0.0141	2.0000	25.45%	9	2.5007±0.0062	1.1474	86.37%	2.5024±0.0097	1.8986	39.87%	32

$x \in [X_1, X_{max}]$, most $f(x) = 0$, some $f(x) = \frac{1}{n}$, and few $f(x) > \frac{1}{n}$. Figure 2 visualizes the distributions of data points between X_1 and X_{max}. They absolutely do not satisfy Property 1 nor Property 2. They are sampling noises and cannot be treated as power-law data. We call these data between X_1 and X_{max} as **long-tailed noises**.

When the magnitude of the sample size increases, all the X_f, X_1, $\hat{X}_1^T$, and X_{max} increase (see Table 2 and 3). When $n \to \infty$, the sample on doubly logarithmic system forms a perfect straight-line, where all the data points satisfy Property 2 and $X_f = X_1 = X_1^T = X_{max} = \infty$ and the long-tailed noises disappear.

4.2 Results of Least-Squares Estimation Fitting Data Sampled from Power-Law Models

Table 4 and 5 reports the $\hat{\alpha}$ of using LSE to fit the sampled discrete and binned continuous data. $\hat{\alpha}_{5th}$, $\hat{\alpha}_f$, $\hat{\alpha}_1$, and $\hat{\alpha}_{all}$ indicate the $\hat{\alpha}$ by LS_{norm} on $\{X_{5th}\}$, $\{X_f\}$, $\{X_1\}$, and $\{X_{max}\}$, respectively. $\hat{\alpha}_{5th}^{avg}$, $\hat{\alpha}_f^{avg}$, and $\hat{\alpha}_{X_1}^{avg}$ indicates the $\hat{\alpha}$ by LS_{avg} on corresponding data points. Figure 2 plots example fitting results of $\hat{\alpha}_f^{avg}$ (indicated by "LS_{Xf}"), $\hat{\alpha}_{X_1}^{avg}$ (by "LS_{X1}"), and $\hat{\alpha}_{all}$ (by "LS_{all}") on the sampled data.

Table 4 and 5 show that $\hat{\alpha}_{all}$ on both discrete and continuous data ranges from 1.5896 to 1.7312, which is significantly biased from the true value of 2.5. Such biased estimation is consistent with the one reported by those criticizers [6, 12, 24, 36, 55]. The fitting result of LS_{all} in Figure 2(b) is consistent with the one reported by Bauke [6] (see its Fig. 1(a)), where both settings are the same: discrete samples, $\alpha = 2.5$, $x_{min} = 1$, and $n = 10^4$. As we analyze in Section 4.1 and show in the supplementary Section A.3, however, the long-tailed noises cannot be treated as power-law data.

Look at $\hat{\alpha}_{5th}$, $\hat{\alpha}_{5th}^{avg}$, $\hat{\alpha}_f$, $\hat{\alpha}_f^{avg}$, $\hat{\alpha}_1$, and $\hat{\alpha}_{X_1}^{avg}$ that are estimated without long-tailed noises. All of them range around the true value at different sizes. Specifically, $\hat{\alpha}_{5th}$, $\hat{\alpha}_f$, and $\hat{\alpha}_1$ range from 2.4923 to 2.5573, while $\hat{\alpha}_{5th}^{avg}$, $\hat{\alpha}_f^{avg}$, and $\hat{\alpha}_{X_1}^{avg}$ from 2.4972 to 2.5315. The $\hat{\alpha}_{5th}$ at $n = 10^4$ in Table 4 is consistent with the one reported in Goldstein et al. [24] (see the "Linear 5-points" item in its Table 1). That means, when excluding long-tailed noises, $\hat{\alpha}$ changes from significantly biased to almost unbiased. It is the long-tailed noises causing significant bias in LSE. Except $\hat{\alpha}_{5th}$ and $\hat{\alpha}_{5th}^{avg}$ on continuous data, when the magnitude of sample size increases, all the six $\hat{\alpha}$ approach much closer to the true value. $\hat{\alpha}_{5th}$ and $\hat{\alpha}_{5th}^{avg}$ on continuous data slightly deviate from the true value, because the method used for continuous data generation tends to generate more data in small x-values. When the magnitude of sample size increases, all the standard deviations of $\hat{\alpha}$ decrease. When the sample size is large enough, some $\hat{\alpha}$ are unbiased in discrete data, such as $\hat{\alpha}_{5th}^{avg}$ at $n = 10^6$ and $\hat{\alpha}_f^{avg}$ at $n = 10^7$. We can expect that when $n \to \infty$, an individual sample forms a perfect straight-line on doubly logarithmic system, and according to the Gauss-Markov theorem [14, 28, 42], all the $\hat{\alpha}_{5th}$, $\hat{\alpha}_{5th}^{avg}$, $\hat{\alpha}_f$, $\hat{\alpha}_f^{avg}$, $\hat{\alpha}_1$, and $\hat{\alpha}_{X_1}^{avg}$ achieve unbiased estimation.

Table 7: Statistics of real-world datasets and fitting results of LS_{avg} and CSN2009 on these datasets. "nan" indicates that LS_{avg} rejects H_0 early for the dataset. "Coverage" indicates the percentage of data that are covered by the method.

Dataset	Size	x_{min}	Type	LS_{avg}					CSN2009			
				$\hat{\alpha}$	Coverage	D_n	D_n^T	Decision	$\hat{\alpha}$	$\hat{x}_{min}$	Coverage	p-value
Words	18,855	1	discrete	1.6616	98.48%	0.0248	0.0202	**moderate**	1.95	7	15.69%	**0.49**
Metabolic	1,641	1	discrete	nan	nan	nan	nan	reject	2.8	4	45.58%	0.00
Terrorism	9,101	1	discrete	1.6607	98.60%	0.0460	0.0343	**moderate**	2.4	12	6.01%	**0.68**
Species	509	1	discrete	1.2422	97.60%	0.0807	0.1041	**accept**	2.4	4	29.42%	**0.10**
Blackouts	211	1,000	discrete	nan	nan	nan	nan	reject	2.3	230,000	27.96%	**0.62**
Cities	19,447	1	discrete	nan	nan	nan	nan	reject	2.37	52,457	2.98%	**0.76**
Fires	203,785	1	continuous	2.7496	89.15%	0.3070	0.0045	reject	2.2	63,240	0.26%	0.05
Flares	12,773	20	discrete	nan	nan	nan	nan	reject	1.79	323	12.40%	**1.00**
Quakes	19,302	0.1	continuous	nan	nan	nan	nan	reject	7.57	3.3	34.49%	0.00
Surnames	2,753	12,436	continuous	1.9868	94.48%	0.0385	0.0509	**accept**	2.5	111,920	8.68%	**0.20**
Citations	415,229	1	discrete	0.9156	99.79%	0.3442	0.0057	reject	3.16	160	0.83%	**0.20**
Weblinks	241,428, 853	1	discrete	1.4964	99.99%	0.1119	0.00018	reject	2.336	3,684	0.01%	0.00

Comparing $\hat{\alpha}_{5th}^{avg}$ vs. $\hat{\alpha}_{5th}$, $\hat{\alpha}_{f}^{avg}$ vs. $\hat{\alpha}_{f}$, and $\hat{\alpha}_{X_1}^{avg}$ vs. $\hat{\alpha}_1$, we can see that LS_{avg} performs much better than LS_{norm}. Moreover, LS_{avg} achieves lower standard deviation than LS_{norm}. The reason is that the average strategy used in LS_{avg} reduces the impact of those data points deviated from the regression line. Example plots of $\hat{\alpha}_{5th}^{avg}$ (indicated by LS_{5th}), $\hat{\alpha}_{f}^{avg}$ (by LS_{Xf}), and $\hat{\alpha}_{X_1}^{avg}$ (by LS_{X1}) in Figure 2 demonstrate that LS_{avg} fits power-law data perfectly.

4.3 LS_{avg} vs. CSN2009

We compare LS_{avg} with CSN2009 [12]. Table 6 reports the fitting results of CSN2009 on 10^3, 10^4, and 10^5 of the same sampled data used for LS_{avg}. (Only these results of CSN2009 are available because running CSN2009 on large samples is extremely time-consuming.) When consider all the 500 samples, LS_{avg} achieves similar $\hat{\alpha}$ on discrete data and slightly deviated $\hat{\alpha}$ on continuous data in comparison with CSN2009, because the method used for continuous data generation is not perfect. However, CSN2009 gets $\hat{x}_{min} \geq 2$ in some discrete samples and less than 50% coverage in some continuous samples. The reason is that CSN2009 adopts a minimum-KS-statistic strategy to choose large lower bound (i.e. $\hat{x}_{min}$) and discards those $x < \hat{x}_{min}$ data although they contain the majority of data [12]. This leads CSN2009 to suffer from the problem of low coverage and indicates a fundamental flaw in CSN2009: it treats the majority of data that are sampled from a power-law model as being not well-fitted by a power-law distribution and discards them. Such flaw becomes extremely severe when fitting power-law distributions to empirical data (see Section 5). By contrast, LS_{avg} discards long-tailed noises and fits the majority of data. The supplementary Figure 4(a) shows an example plot of such flaw in CSN2009 in comparison with LS_{avg}.

5 FITTING POWER-LAW DISTRIBUTIONS TO REAL-WORLD DATA

We apply LS_{avg} to fit power-law distributions to twelve real-world datasets that are used in Clauset et al. [12] and compare the fitting results with the ones of CSN2009 [12]. The statistics of the twelve datasets are summarized in the first four columns of Table 7. (For details of these datasets, please refer to the Section 6, Table 3, Figure 8 and 9 in Clauset et al. [12].) The setup of applying LS_{avg} to real-world data is detailed in the supplementary Section B.

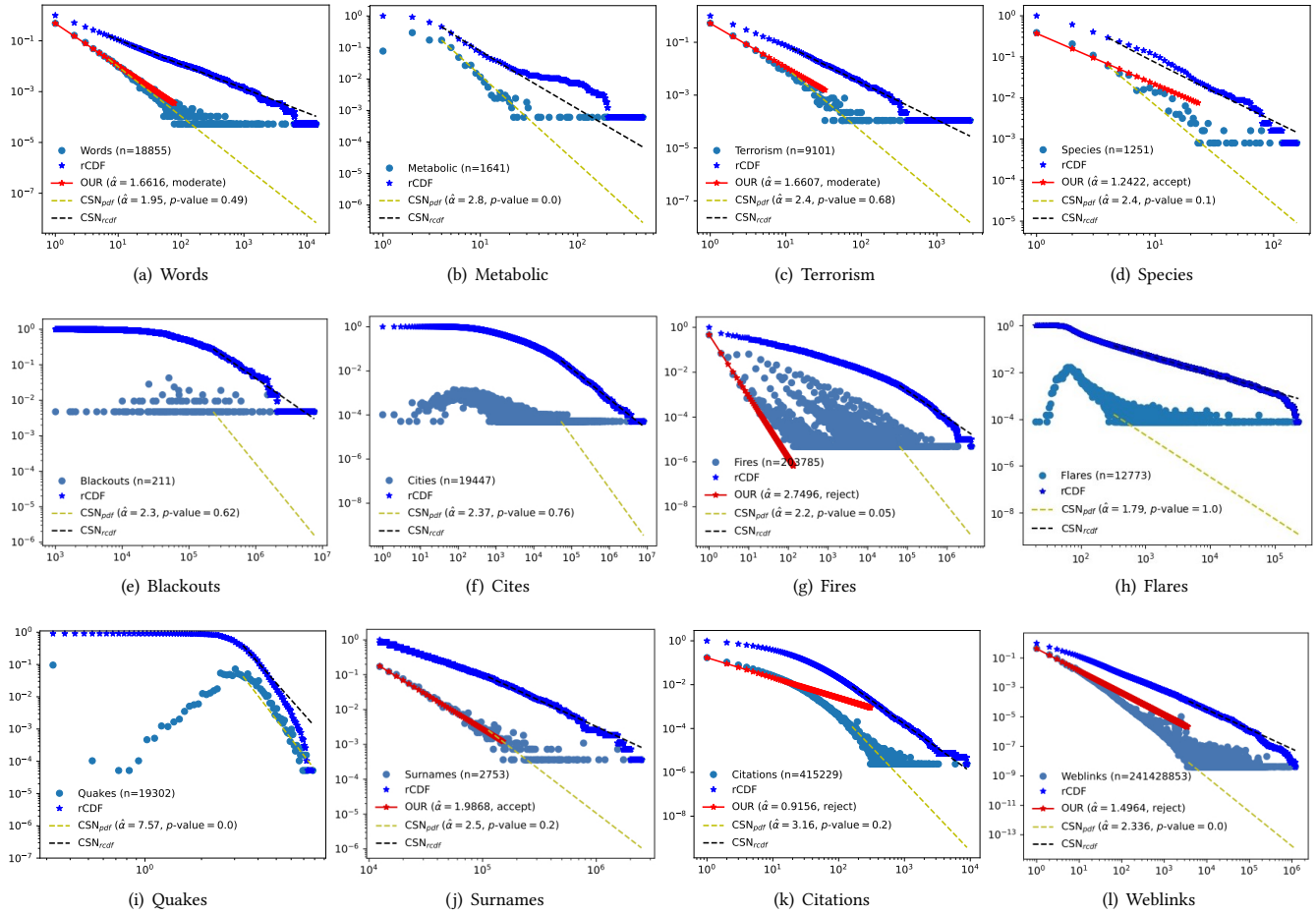
5.1 Experimental Results

The last nine columns of Table 7 report the experimental results of LS_{avg} and CSN2009 in fitting these real-world datasets. For LS_{avg}, the coverage is calculated by the total frequency of $\{X_1\}$ if the hypothesis is not rejected early; "nan" indicates that LS_{avg} rejects H_0 early for the dataset and only the decision of rejection is reported in the table. For CSN2009, the results are mainly from Clauset et al. [12] and the coverage is calculated based on the $\hat{x}_{min}$.

Figure 3 plots the results of LS_{avg} and CSN2009 fitting these real-world datasets on doubly logarithmic system. LS_{avg} displays only the PDF plot while CSN2009 displays both PDF and rCDF plots; CSN2009's rCDF plot corresponds to the "complementary cumulative distribution function (CDF)" in Clauset et al. [12].

5.2 LS_{avg} PDF vs. CSN2009 PDF

Compare LS_{avg} and CSN2009 by looking at Table 7 and PDF plots in Figure 3. Generally, LS_{avg} mainly fits the first several data points that contain the majority of data; for example, in the seven datasets that LS_{avg} does not reject H_0 early (i.e., Words, Terrorism, Species, Fires, Surnames, Citations, Weblinks), LS_{avg} covers 89.15%∼99.99% of data. By contrast, CSN2009 mainly fits the last many data points that contain only a few data; it covers only 0.01%∼45.58% of data. For the four datasets that LS_{avg} accepts or moderately accepts H_0, CSN2009 also accepts H_0 with p-values from 0.10 to 0.68. However, CSN2009 covers only 6.01%∼27.96% of these data, in comparison with the coverage of 94.48%∼98.60% by LS_{avg}. Figure 3(a), 3(c), 3(d), and 3(j) intuitively visualize such difference. For some datasets that LS_{avg} rejects H_0 early, such as Blackouts, Cites, Flares, CSN2009 treats their long tails as being following power-law distributions with high p-values from 0.62 to 1.0. Figure 3(e), 3(f), and 3(h) visualizes such ridiculous fittings by CSN2009. As we demonstrate in Section 4.1, the long-tailed noises of finite-size samples cannot be treated as power-law data. CSN2009 tends to discard the majority of data and fit the long-tailed noises and therefore produces substantially wrong fittings. The reason is as illustrated in Section 4.3 and the supplementary Section A.3 that CSN2009 adopts a minimum-KS-statistic strategy to choose the large lower bound as the beginning of power-law distributions and mistakenly teats the long-tailed noises as being following power-law distributions.



(a) Words (b) Metabolic (c) Terrorism (d) Species

(e) Blackouts (f) Cites (g) Fires (h) Flares

(i) Quakes (j) Surnames (k) Citations (l) Weblinks

Figure 3: Fitting and hypothesis testing results of LS_{avg} (denoted by "OUR") and CSN2009 (by "CSN") on real-world datasets on doubly logarithmic system. OUR displays only PDF plot. CSN$_{pdf}$ indicates CSN2009's PDF plot while CSN$_{rcdf}$ indicates CSN2009's rCDF plot. A figure without OUR plot indicates "reject," meaning that LS_{avg} rejects the hypothesis H_0 very early.

5.3 CSN2009 PDF vs. CSN2009 rCDF

Figure 3 shows that the rCDF of these data especially their long tails might seem to be fitted by CSN2009 with high p-values, however, their PDF actually do not follow power-law distributions. Such PDF vs. rCDF plots of CSN2009 are consistent with the plots of CSN2009 fitting long-tailed noises as shown in Figure 4. This indicates that using rCDF plot may hide the true probability distribution of data and lead to wrong fittings. The reason may be that only when the PDF of data follows a power-law distribution, we can use the rCDF to derive the PDF by the power-law function. We should be careful when trying to use rCDF for power-law fitting or plot.

6 CONCLUSION

In this paper, we propose an average strategy for least-squares estimation (LSE) to fit power-law distributions, define the correct way to bin continuous power-law data into data points, and propose to use the maximal statistic of the two-sample Kolmogorov-Smirnov (KS) test among a large group of power-law samples as the threshold to examine power-law hypothesis. By using these proposed methods, we conduct extensive experiments, demonstrating that the criticism about the inaccuracy of LSE in fitting power-law distributions is complete nonsense. Our experiments show that LSE fits power-law data perfectly and that it is the long-tailed noises of finite-size samples causing the inaccuracy when LSE fitting power-law distributions and such long-tailed noises cannot be treated as power-law data even though they are sampled from power-law models. Those criticizers [6, 11, 12, 24, 36, 55] mistakenly treats the data problem as the model problem. Our experiments uncover a fundamental flaw in the widely known CSN2009 method proposed by Clauset et al. [12]: it tends to discard the majority of power-law data and fit the long-tailed noises. Such flaw invalidates the reliability of all the research based on CSN2009 and all those works (e.g., [10, 19, 37, 50, 52]) need to be re-investigated. Our experiments also show that the popular reserve cumulative distribution function (rCDF) is a bad idea to plot power-law data in practice because it usually hides the true probability distribution of data. We hope that our research can clean up the bias that has been caused by those misleading research in research community about LSE in fitting power-law distributions, and that researchers should be careful not to mistakenly treat a data problem as a model problem.

REFERENCES

[1] Lada A. Adamic and Bernardo A. Huberman. 2000. The Nature of Markets in the World Wide Web. *Quarterly Journal of Electronic Commerce* 1, 1 (2000), 5–12.

[2] Reka Albert, Hawoong Jeong, and Albert-László Barabási. 1999. Diameter of the World-Wide Web. *Nature* 401 (1999), 130.

[3] I. Artico, I. Smolyarenko, V. Vinciotti, and E. C. Wit. 2020. How rare are power-law networks really?. In *Proceedings of the Royal Society A*, Vol. 476. 20190742.

[4] Eduardo M. Azevedo, Alex Deng, Jose Luis Montiel Olea, Justin Rao, and E. Glen Weyl. 2020. A/B Testing with Fat Tails. *Journal of Political Economy* 128, 12 (2020), 4614–000.

[5] Albert-László Barabási and Réka Albert. 1999. Emergence of Scaling in Random Networks. *Science* 286 (1999), 509–512.

[6] H. Bauke. 2007. Parameter estimation for power-law distributions by maximum likelihood methods. *The European Physical Journal B* 58 (2007), 167–173.

[7] Bernd Blaslus. 2020. Power-law distribution in the number of confirmed COVID-19 cases. *Chaos: An Interdisciplinary Journal of Nonlinear Science* 30, 9 (2020), 093123.

[8] Eric Bonnet, Olivier Bour, Noelle E. Odling, Philippe Davy, Ian Main, Patience Cowie, and Brian Berkowitz. 2001. Scaling of fracture systems in geological media. *Reviews of geophysics* 39, 3 (2001), 347–383.

[9] Patrick Erik Bradley and Martin Behnisch. 2019. Heavy-tailed distributions for building stock data. *Environment and Planning B: Urban Analytics and City Science* 46, 7 (2019), 1281–1296.

[10] Anna D. Broido and Aaron Clauset. 2019. Scale-free networks are rare. *Nature communications* 10, 1 (2019), 1–10.

[11] Robert Malcolm Clark, S. J. D. Cox, and Geoff M. Laslett. 1999. Generalizations of power-law distributions applicable to sampled fault-trace lengths: model choice, parameter estimation and caveats. *Geophysical Journal International* 136, 2 (1999), 357–372.

[12] Aaron Clauset, Cosma Rohilla Shalizi, and M. E. J. Newman. 2009. Power-law Distributions in Empirical Data. *SIAM Rev.* 51, 4 (2009), 661–703.

[13] William G. Cochran. 1952. The Chi-square Test of Goodness of Fit. *The Annals of Mathematical Statistics* 23, 3 (1952), 315–345.

[14] Donald Cochrane and Guy H. Orcutt. 1949. Application of least squares regression to relationships containing auto-correlated error terms. *J. Amer. Statist. Assoc.* 44, 245 (1949), 32–61.

[15] Brian Conrad and Michael Mitzenmacher. 2004. Power laws for monkeys typing randomly: the case of unequal probabilities. *IEEE Transactions on information theory* 50, 7 (2004), 1403–1414.

[16] Bernat Corominas-Murtra and Ricard V. Solé. 2010. Universality of Zipf's Law. *Physical Review E* 82, 1 (2010), 011102.

[17] Frederik Michel Dekking, Cornelis Kraaikamp, Hendrik Paul Lopuhaä, and Ludolf Erwin Meester. 2005. *A Modern Introduction to Probability and Statistics: Understanding why and how.* Springer Science & Business Media.

[18] Anna Deluca and Alvaro Corral. 2013. Fitting and Goodness-of-Fit Test of Non-Truncated and Truncated Power-Law Distributions. *Acta Geophysica* 61, 6 (2013), 1351–1394.

[19] Nicole Eikmeier and David F. Gleich. 2017. Revisiting Power-law Distributions in Spectra of Real World Networks. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 817–826.

[20] Zoltan Eisler, Imre Bartos, and Janos Kertesz. 2008. Fluctuation scaling in complex systems: Taylor's law and beyond. *Advances in Physics* 57, 1 (2008), 89–142.

[21] Brian J. Enquist, Evan P. Economo, Travis E. Huxman, Andrew P. Allen, Danielle D. Ignace, and James F. Gillooly. 2003. Scaling metabolism from organisms to ecosystems. *Nature* 423, 6940 (2003), 639–642.

[22] Brian J. Enquist and Karl J. Niklas. 2001. Invariant scaling relations across tree-dominated communities. *Nature* 410, 6829 (2001), 655–660.

[23] Xavier Gabaix. 2009. Power Laws in Economics and Finance. *Annual Review of Economics* 1, 1 (2009), 255–294.

[24] M.L. Goldstein, S.A. Morris, and G.G. Yen. 2004. Problems with fitting to the power-law distribution. *The European Physical Journal B* 41 (2004), 255–258.

[25] Beno Gutenberg and Charles F. Richter. 1944. Frequency of Earthquakes in California. *Bulletin of the Seismological Society of America* 34, 4 (1944), 185–188.

[26] Bo-Ping Han and Milan Straskraba. 1998. Size dependence of biomass spectra and population density I. The effects of size scales and size intervals. *Journal of Theoretical Biology* 191, 3 (1998), 259–265.

[27] Rudolf Hanel, Bernat Corominas-Murtra, Bo Liu, and Stefan Thurner. 2017. Fitting power-laws in empirical data with estimators that work for all exponents. *PLoS ONE* 12, 2 (2017), 1–15.

[28] Charles R. Henderson. 1975. Best Linear Unbiased Estimation and Prediction under a Selection Model. *Biometrics* 31, 2 (1975), 423–447.

[29] Hawoong Jeong, Balint Tombor, Reka Albert, Zoltan N. Oltvai, and A-L. Barabasi. 2000. The Large-Scale Organization of Metabolic Networks. *Nature* 407, 6804 (2000), 651–654.

[30] Sonia Kefi, Max Rietkerk, Concepcion L. Alados, Yolanda Pueyo, Vasilios P. Papanastasis, Ahmed ElAich, and Peter C. De Ruiter. 2007. Spatial vegetation patterns and imminent desertification in Mediterranean arid ecosystems. *Nature* 449, 7159 (2007), 213–217.

[31] Wentian Li. 2002. Zipf's Law Everywhere. *Glottometrics* 5 (2002), 14–21.

[32] Edward T. Lu and Russell J. Hamilton. 1991. Avalanches and the Distribution of Solar Flares. *The Astrophysical Journal* 380 (1991), L89–L92.

[33] R. Dean Malmgren, Daniel B. Stouffer, Adilson E. Motter, and Luis AN Amaral. 2008. A Poissonian explanation for heavy tails in e-mail communication. *Proceedings of the National Academy of Sciences* 105, 47 (2008), 18153–18158.

[34] Timothy D. Meehan. 2006. Energy Use and Animal Abundance in Litter and Soil Communities. *Ecology* 87, 7 (2006), 1650–1658.

[35] Buddhika Nettasinghe and Vikram Krishnamurthy. 2021. Maximum Likelihood Estimation of Power-law Degree Distributions via Friendship Paradox-based Sampling. *ACM Transactions on Knowledge Discovery from Data* 15, 6 (2021), 1–28.

[36] Mark EJ. Newman. 2005. Power laws, Pareto distributions and Zipf's law. *Contemporary physics* 46, 5 (2005), 323–351.

[37] Jan Overgoor, Austin R. Benson, and Johan Ugander. 2019. Choosing to Grow a Graph: Modeling Network Formation as Discrete Choise. In *Proceedings of the 2019 World Wide Web Conference*. 1409–1420.

[38] Karl Pearson. 1990. On the Criterion that a Given System of Deviations from the Probable in the Case of a Correlated System of Variables is Such that it Can be Reasonably Supposed to have Arisen from Random Sampling. *Philos. Mag.* 5, 50 (1990), 157–175.

[39] Steven T. Piantadosi. 2014. Zipf's word frequency law in natural language: A critical review and future directions. *Psychonomic bulletin & review* 21, 5 (2014), 1112–1130.

[40] G. Pickering, J. M. Bull, and D. J. Sanderson. 1995. Sampling power-law distributions. *Tectonophysics* 248 (1995), 1–20.

[41] Carla M.A. Pinto, A. Mendes Lopes, and J.A. Tenreiro Machado. 2012. A review of power laws in real life phenomena. *Commun Nonlinear Sci Number Simulat* 17 (2012), 3558–3578.

[42] Robin L. Plackett. 1949. A Historical Note on the Method of Least Squares. *Biometrika* 36, 3/4 (1949), 458–460.

[43] Derek J. De Solla Price. 1965. Networks of Scientific Papers. *Science* 149, 3683 (1965), 510–515.

[44] Salvador Pueyo and Roger Jovani. 2006. Comment on "A Keystone Mutualism Drives Pattern in a Power Function". *Science* 313, 5794 (2006), 1739–1739.

[45] John A. Rice. 2006. *Mathematical Statistics and Data Analysis* (3rd ed.). Cengage Learning.

[46] Andrea Rinaldo, Amos Maritan, Kent K. Cavender-Bares, and Sallie W. Chisholm. 2002. Cross-scale ecological dynamics and microbial size spectra in marine ecosystems. In *Proceedings of the Royal Society of London. Series B: Biological Sciences*, Vol. 269. 2051–2059.

[47] David W. Sims, David Righton, and Jonathan W. Pitchford. 2007. Minimizing errors in identifying Levy flight behaviour of organisms. *Journal of Animal Ecology* 76, 2 (2007), 222–229.

[48] Nickolay Smirnov. 1948. Table for Estimating the Goodness of Fit of Empirical Distributions. *Annals of Mathematical Statistics* 19, 2 (1948), 279–281.

[49] Michael A Stephens. 1974. EDF statistics for goodness of fit and some comparisons. *Journal of the American statistical Association* 69, 347 (1974), 730–737.

[50] Alex Stivala, Garry Robins, and Alessandro Lomi. 2020. Exponential random graph model parameter estimation for very large directed networks. *PLoS ONE* 15, 1 (2020), e0227804.

[51] Gilbert Strang. 2016. *Introduction to Linear Algebra* (5th ed.). Wellesley-Cambridge Press.

[52] Yogesh Virkar and Aaron Clauset. 2014. Power-law distributions in binned empirical data. *The Annals of Applied Statistics* 8, 1 (2014), 89–119.

[53] Ding Wang, Haibo Cheng, Ping Wang, Xinyi Huang, and Gaopeng Jian. 2017. Zipf's Law in Passwords. *IEEE Transactions on Information Forensics and Security* 12, 11 (2017), 2776–2791.

[54] Geoffrey B. West, James H. Brown, and Brian J. Enquist. 1997. A General Model for the Origin of Allometric Scaling Laws in Biology. *Science* 276, 5309 (1997), 122–126.

[55] Ethan P. White, Brian J. Enquist, and Jessica L. Green. 2008. On estimating the exponent of power-law frequency distributions. *Ecology* 89, 4 (2008), 905–912.

[56] J. C. Willis and G. Udny Yule. 1922. Some Statistics of Evolution and Geographical Distribution in Plants and Animals, and their Significance. *Nature* 109 (1922), 177–179.

[57] Chengxi Zang, Peng Cui, and Wenwu Zhu. 2018. Learning and Interpreting Complex Distributions in Empirical Data. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 2682–2691.

[58] Xiaoshi Zhong. 2020. *Time Expression and Named Entity Analysis and Recognition.* Ph.D. Dissertation. Nanyang Technological University, Singapore.

[59] Tommaso Zillio and Richard Condit. 2007. The impact of neutrality, niche, differentiation and species input on diversity and abundance distributions. *Oikos* 116 (2007), 931–940.

[60] George Zipf. 1949. *Human Behavior and the Principle of Least Effort: An Introduction to Human Ecology.* Addison-Wesley Press, Inc.

Table 8: Estimated α of fitting different sizes of continuous data (with varied-width binning and logarithmic binning) sampled from a power-law model with $\alpha = 2.5$ and $x_{min} = 1$. Experimental results are based on 500 samples for each size.

Size	Varied-width Binning						Logarithmic Binning	
	$\hat{\alpha}_{5th}$	$\hat{\alpha}_{5th}^{avg}$	$\hat{\alpha}_f$	$\hat{\alpha}_f^{avg}$	$\hat{\alpha}_{X_1}$	$\hat{\alpha}_{X_1}^{avg}$	$\hat{\alpha}_{log}$	$\hat{\alpha}_{log}^{avg}$
10^3	2.5435±0.1447	2.5240±0.0794	2.4968±0.0988	2.5099±0.0725	2.5559±0.1702	2.5234±0.0762	2.3920±0.1653	2.4901±0.0590
10^4	2.5161±0.0425	2.5121±0.0257	2.5121±0.0422	2.5131±0.0224	2.5517±0.0758	2.5213±0.0283	2.4252±0.1063	2.4940±0.0257
10^5	2.5171±0.0140	2.5130±0.0079	2.5009±0.0171	2.5127±0.0080	2.5398±0.0353	2.5163±0.0125	2.4444±0.0830	2.4966±0.0139
10^6	2.5160±0.0043	2.5125±0.0025	2.5068±0.0068	2.5100±0.0027	2.5280±0.0194	2.5107±0.0068	2.4613±0.0611	2.4976±0.0083
10^7	2.5159±0.0013	2.5124±0.0007	2.5049±0.0029	2.5083±0.0012	2.5217±0.0110	2.5082±0.0037	2.4708±0.0434	2.4986±0.0049
10^8	2.5161±0.0004	2.5125±0.0003	2.5033±0.0013	2.5067±0.0006	2.5172±0.0067	2.5060±0.0021	2.4786±0.0364	2.4991±0.0036

Table 9: Estimated α of fitting different sizes of sampled continuous data with using bin center representation for fixed-width, varied-width, and logarithmic binning. Experimental results are based on 500 samples for each size.

Size	Fixed-width Binning						Varied-width Binning						Logarithmic Binning	
	$\hat{\alpha}_{5th}$	$\hat{\alpha}_{5th}^{avg}$	$\hat{\alpha}_f$	$\hat{\alpha}_f^{avg}$	$\hat{\alpha}_{X_1}$	$\hat{\alpha}_{X_1}^{avg}$	$\hat{\alpha}_{5th}$	$\hat{\alpha}_{5th}^{avg}$	$\hat{\alpha}_f$	$\hat{\alpha}_f^{avg}$	$\hat{\alpha}_{X_1}$	$\hat{\alpha}_{X_1}^{avg}$	$\hat{\alpha}_{log}$	$\hat{\alpha}_{log}^{avg}$
10^3	2.6306	2.6733	2.5981	2.6522	2.6084	2.6170	2.6439	2.6011	2.5894	2.5852	2.6275	2.6017	2.3920	2.4901
10^4	2.6246	2.6639	2.5751	2.6225	2.5795	2.5821	2.6150	2.5883	2.5870	2.5916	2.5952	2.5852	2.4252	2.4940
10^5	2.6209	2.6639	2.5512	2.5959	2.5569	2.5522	2.6160	2.5892	2.5649	2.5843	2.5638	2.5611	2.4444	2.4966
10^6	2.6215	2.6638	2.5335	2.5723	2.5373	2.5313	2.6149	2.5886	2.5464	2.5712	2.5406	2.5393	2.4613	2.4976
10^7	2.6213	2.6636	2.5229	2.5546	2.5277	2.5196	2.6148	2.5885	2.5337	2.5591	2.5282	2.5254	2.4708	2.4986
10^8	2.6213	2.6637	2.5157	2.5410	2.5213	2.5124	2.6150	2.5886	2.5234	2.5470	2.5204	2.5158	2.4786	2.4991

A FITTING POWER-LAW DISTRIBUTIONS TO DATA DRAWN FROM POWER-LAW MODEL

A.1 Our Bin Representation Applies to any Kinds of Binning Methods

In Table 3 we report the fitting results of LS_{norm} and LS_{avg} on the sampled continuous data that are binned by our binning method with fixed width=1. In Table 8 we report the fitting results of LS_{norm} and LS_{avg} on the sampled continuous data that are binned by our binning method with varied widths and logarithmic widths. In varied-width binning, the bins in odd numbers are set by width=1 while the bins in even numbers by width=2. In logarithmic binning, the base of logarithm for bins is set by 2. Comparing Table 8 with Table 3, we can see that LS_{norm} achieves similar $\hat{\alpha}_{5th}$, $\hat{\alpha}_f$, and $\hat{\alpha}_1$ in both fixed-width binning and varied-width binning, and LS_{avg} achieves similar $\hat{\alpha}_{5th}^{avg}$, $\hat{\alpha}_f^{avg}$, and $\hat{\alpha}_1^{avg}$ in both kinds of binning. LS_{norm} with logarithmic binning achieves underestimated $\hat{\alpha}_{log}$ mainly because the sampling is not perfect; the $\hat{\alpha}_{log} = 2.4252$ on 10^4 is consistent to the one in Bauke [6] (i.e., 2.43; see its Fig. 1(c)). However, when the magnitude of sample size increases, the $\hat{\alpha}_{log}$ is steadily closer to the true value of 2.5. LS_{avg} achieves much better estimation mainly because the average strategy reduces the impact of those deviated data points. As expectation, when the magnitude of sample size increases, the standard deviations of both LS_{norm} and LS_{avg} decrease in all the fixed-width, varied-width, and logarithmic binning. This experimentally verifies that our bin representation for continuous data applies to any kinds of binning methods.

A.2 Our Bin Representation is Better than the Bin Center Representation

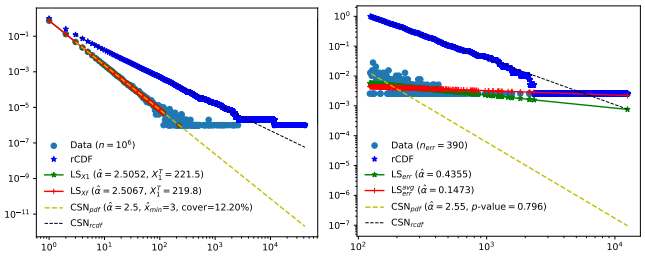
As shown in Figure 1, only the area of the rectangle $ABFE$ equals to the one of the trapezoid $ABDC$ (i.e., $\mathcal{A}_{ABFE} = \mathcal{A}_{ABDC}$). The area of any other rectangles with AB being the width does not equal to $\mathcal{A}_{ABDC}$. Therefore, using any other points to represent the bin is incorrect and will result in biased estimated model, either the exponent or in the constant or in both.

Table 9 reports the fitting results of applying LS_{norm} and LS_{avg} on the sampled continuous data that are binned by a way in which other settings are the same as our binning method except x_{c_i} is set by the center of the bin (i.e., $x_{c_i} = \frac{1}{2}(x_{b_i} + x_{b_{i+1}})$) instead of Eq. (7). Such bin center representation (i.e., represent the bin by its center) is used by most previous research [21, 22, 30, 34, 40] and some criticizers who conduct inappropriate experiments to criticize LSE in fitting power-law distributions [55].

Table 9 shows that when using bin center representation, both LS_{norm} and LS_{avg} achieve biased estimation on the exponent α in both fixed-width and varied-fixed binning. For logarithmic binning, it is expected that the bin center representation achieves the same $\hat{\alpha}$ as our bin representation because the bin center presentations and our bin representations form parallel lines with the same slope on doubly logarithmic system. We should note that while the parallel lines have the same slope, their intercepts are different, and thus the estimated constants of power-law models are different.

A.3 Not All the Data Drawn from a Power-Law Model Follow a Power-Law Distribution

In Section 4.1, statistics of the sampled data indicates that the long-tailed noises cannot be treated as power-law data. In this supplementary section, we use power-law models to fit these long-tailed noises and plot the fitting results on doubly logarithmic system. Table 10 reports the $\hat{\alpha}$ of using LS_{norm}, LS_{avg}, and CSN2009 [12] to fit these long-tailed noises of sampled discrete and continuous data. Figure 4(b) plots example results of using LS_{norm}, LS_{avg}, and CSN2009 to fit the long-tailed noises of discrete data that are drawn from a power-law model with $\alpha = 2.5$, $x_{min} = 1$, and $n = 10^6$.



(a) LS_{avg} vs. CSN2009 on a drawn power-law sample

(b) LS_{norm} vs. LS_{avg} vs. CSN2009 on the long-tailed noises of a power-law sample

Figure 4: (a) LS_{avg} and CSN2009 on a discrete power-law sample. (b) LS_{norm}, LS_{avg}, and CSN2009 on long-tailed noises. The discrete sample and long-tailed noises are drawn from a power-law model with $\alpha = 2.5$, $x_{min} = 1$, and $n = 10^6$.

Table 10: Estimated α of LS_{norm}, LS_{avg}, and CSN2009 fitting long-tailed noises of discrete and continuous data.

Size	Discrete			Continuous		
	$\hat{\alpha}_{err}$	$\hat{\alpha}_{err}^{avg}$	$\hat{\alpha}_{csn}$/p-value	$\hat{\alpha}_{err}$	$\hat{\alpha}_{err}^{avg}$	$\hat{\alpha}_{csn}$/p-value
10^3	0.1543	0.1518	3.0780/0.5472	0.2369	0.2134	2.7486/0.5974
10^4	0.2583	0.1171	2.5619/0.5633	0.3341	0.1453	2.6062/0.5530
10^5	0.3273	0.1137	2.5456/0.5569	0.3856	0.1358	2.5331/0.5467
10^6	0.3745	0.1177	-	0.4171	0.1317	-
10^7	0.4221	0.1253	-	0.4525	0.1352	-
10^8	0.4564	0.1321	-	0.4840	0.1405	-

As Table 10 shows[2], the long-tailed noises of both discrete and continuous data are fitted by LS_{norm} with $\hat{\alpha}_{err}$ ranging from 0.1543 to 0.4840 and by LS_{avg} with $\hat{\alpha}_{err}^{avg}$ from 0.1137 to 0.2134. Especially, when the sample size reaches 10^4, α_{err}^{avg} ranges from 0.1137 to 0.1453. All the $\hat{\alpha}_{err}$ and $\hat{\alpha}_{err}^{avg}$ are close to zero and far less than the true value of power-law models. It is reasonable because the long-tailed noises are mainly composed of $\frac{1}{n}$-frequency data points where the slope should be close to zero. In comparison, CSN2009 fits the long-tailed noises with $\hat{\alpha}_{csn}$ from 2.5468 to 3.0780. What is worse is that CSN2009 achieves relatively high p-values around 0.55 for its fittings; this means that CSN2009 mistakenly accepts the hypothesis with high confidence that such long-tailed noises follow a power-law distribution. The worst thing is that CSN2009 uses rCDF to plot the data.[3] Figure 4(b) shows that even though the long-tailed noises do not follow a power-law distribution (as displayed by their PDF and the statistics reported in Table 2 and 3), their rCDF seems to be fitted by a power-law distribution. This indicates that such rCDF plot hides the true PDF of data, and may hide the truth that the data points do not follow a power-law distribution.

After illustrating that not all the data that are sampled from a power-law model follow a power-law distribution, we explain why this happens. The reason is that the power-law distribution defined by Eq. (1) is not uniform but strongly skewed in favour of those small x-values and a finite-size sample cannot cover all the possible x-values. Looking at the item where $n = 10^4$ in Table 2, for example, $\{X_1\}$ includes 24.2 possible x-values and contains 99.58% of data, namely, 9958 observations; while the long-tailed noises include 725.4 possible x-values but contain only 0.42% of data, namely, 42 observations. That means, for the long-tailed noises (i.e., data between X_1 and X_{max}), the number of observations (i.e., 42) is far less than the number of possible x-values (i.e., 725.4). This is the reason why there are a large number of 0-frequency and $\frac{1}{n}$-frequency data points between X_1 and X_{max}. Using these 0-frequency and $\frac{1}{n}$-frequency data points to estimate parameters of power-law distributions will result in inaccurate estimation.

We use an analogy to explain why we cannot use the long-tailed noises to estimate parameters of power-law models. Suppose the occurrence probabilities of the six sides of a dice are not the uniform $\frac{1}{6}$ but $\frac{32}{63}$, $\frac{16}{63}$, $\frac{8}{63}$, $\frac{4}{63}$, $\frac{2}{63}$, and $\frac{1}{63}$, we can expect that when drawing infinite observations from the dice, we will obtain unbiased estimation for the occurrence probabilities of the dice. But if we draw only three observations, it is possible that the three observations contain one $\frac{32}{63}$-side, one $\frac{8}{63}$-side, and one $\frac{1}{63}$-side (or other distributions). Using the three observations to estimate the occurrence probabilities of the dice will result in substantial bias. Similarly, using the long-tailed noises of finite-size samples to estimate parameters of power-law models will result in substantial bias.

B FITTING POWER-LAW DISTRIBUTIONS TO REAL-WORLD DATA

Applying LS_{avg} to real-world data mainly contains two stages: fitting real-world data and testing power-law hypothesis.

B.1 Fitting Real-World Data

For each set of real-world data points, we apply LS_{avg} on $\{X_{5th}\}$, $\{X_f\}$, and $\{X_1\}$, and denote the $\hat{\alpha}$ by $\hat{\alpha}_{5th}^{avg}$, $\hat{\alpha}_{X_f}^{avg}$, and $\hat{\alpha}_{X_1}^{avg}$. After getting their estimated power-law models, we calculate their corresponding KS statistics on $\{X_1\}$, denoted by D_n^{5th}, $D_n^{X_f}$, and $D_n^{X_1}$, and choose the minimal one (i.e., $D_n = \min\{D_n^{5th}, D_n^{X_f}, D_n^{X_1}\}$) and its corresponding model as the final results of our method. Note that when estimating the power-law model, we use $\{X_{5th}\}$, $\{X_f\}$, and $\{X_1\}$, but when calculating KS statistic, we apply the estimated power-law model on $\{X_1\}$.

B.2 Testing Power-Law Hypothesis

If any of $\hat{\alpha}_{5th}^{avg}$, $\hat{\alpha}_{X_f}^{avg}$, and $\hat{\alpha}_{X_1}^{avg}$ does not exist, we reject the hypothesis H_0 immediately (i.e., early rejection). For the estimated power-law model $p(x) = \hat{K} \cdot x^{-\hat{\alpha}}$, we draw 300 samples from it and use Eq. (12) to calculate D_n^T and then use the strategy described in Section 3.3 to decide whether to accept, moderately accept, or reject H_0.

B.3 Real-World Datasets

We tried to collect other twelve datasets by the given links and contacting the authors of those datasets according to Aaron Clauset's instruction at https://aaronclauset.github.io/powerlaws/data.htm, but unfortunately, those authors either did not reply us or could not provide us those datasets. Clauset et al. [12] said they have no permission to make those datasets publicly available.

[2]Only the results of CSN2009 on the long-tailed noises of these 10^3, 10^4, and 10^5 samples are available because CSN2009 is extremely time-consuming on large sizes of samples.

[3]According to its definition ($P_{rcdf}(x) = \int_x^{\infty} p(t)dt$, where $0 \le p(x) \le 1$), the rCDF is a monotonically non-increasing function on its domain $\mathbb{D}$: $\forall x_i, \forall x_j \in \mathbb{D}$, if $x_i < x_j$, then $P_{rcdf}(x_i) \ge P_{rcdf}(x_j)$. The rCDF plots reported in Newman [36] and Clauset et al. [12] are incorrect because they exclude those x-values whose frequencies are 0, even though Newman [36] advocates to use rCDF to plot power-law data. The rCDF plots reported in Bauke [6] are correct.