Nine Open problems in Complex Networks

1 Testing models against alternatives (a) what are the alternatives (b) how to establish gof likelihood (KS) and comparative pair-wise likelihood (KS and LR) (c) how to get MLE estimation for each (d) proper forms to construct and normalize data for testing

2 What are the small sample MLE and other biases and how to correct for them?

3 Convergence and complementarity in different estimation methods, some not MLE
   (Du: sample results: still needs to fit log normal, stretched exponentials, etc.)

4 What plausible missing alternatives need development? E.g., q-exponential

5 What is the q-exponential and why is it plausible?

6 Continuous q-distribution MLE solved; problems for discrete q-distributions (ntwks)

7 Alternate fitting using q-semilog (also an open problem)

8 Is there an MLE for the semilog solution generally

9 Will it work for discrete data, small samples
Bibliography (refs at end)


FIRST THREE PROBLEMS

1 Testing models against alternatives (a) what are the alternatives (b) how to establish gof likelihood (KS) and comparative pair-wise likelihood (KS and LR) (c) how to get MLE estimation for each (d) proper forms to construct and normalize data for testing

2 What are the small sample MLE and other biases and how to correct for them?

3 Convergence and complementarity in different estimation methods, some not MLE

(summarized in intro to [4] that follows)
Model Testing: Pareto, Pareto II ($q$), and other Distributions

Laurent Tambayong, Aaron Clauset, Cosma Shalizi, Douglas R. White

Draft 2 Feb 11, 2008 - DRW

Abstract

Models of point-data such as power laws and other distributions, if done correctly, allow accurate tests of fit. We test two maximum likelihood estimators for the $q$-exponential (Pareto II distribution) against other models (exponential, stretched exponential, lognormal; and nested models such as power law plus exponential cut-off). A newly derived MLE provides a single-parameter estimation for $q$ which is compared to the previous MLE solution from for the Pareto II. Fit for the $q$-exponential strongly outperforms the power law for continuous distributions (populations of cities, sales of books, forest fire sizes, blackouts) studied by Clauset, Shalizi and Newman (2007) in their review of likelihood methods for testing univariate distributions.
I. Introduction

Common fallacies of model testing include fitting the parameters of a model by minimizing the sum of squared deviations from the data (measuring $r^2$ fit of the data and “model predictions” for each data point) and that the significance test for deviation from $r^2$ fit, according to the null hypothesis, is a good measure of fit. What this ignores, among other things, is (1) formulation of the model in probabilistic terms and measuring the probability that the data observed would have been generated under the model probabilities given the fitted model parameters (hence $p \approx 1.0$ is good fit, unlike the significance test), and (2) testing alternative models using log likelihood probabilities that measure the probability the model of interest is a better fit than an alternative (again $p \approx 1.0$ is the probability that the data could have been generated from the model probabilities rather than the alternative, again unlike the significance test). Only in the social sciences do the fallacies and common usage stubbornly persist that (1) rejection of the null hypothesis by a significance test can substitute for a direct test of fit to a model, a debate called the significance-test controversy (Meehl 1967) that has never been fully laid to rest, and (2) that correlation between observed data and model “predictions” provide a valid test of fit. The latter case, however, is an error commonly made even by mathematically sophisticated physicists, in the belief that fitting a straight line to a log-log plot by $r^2$ and a significance test provides a good test of fit for a power law (or, in a semi-log plot, for an exponential distribution). Clauset et al (2007), in their Appendix A.2 show how severely these approaches are biased, even if they are in common use in both the
social and the physical sciences. Hundreds if not thousands of peer-reviewed papers have
been published testing fit, for example, between a 2-parameter power-law and the frequency
distribution of nodes in a network by their number of links (aka degree distributions).

Models of point-data such as power laws and other distributions do allow accurate tests of fit
when correctly done, as Clauset et al (2007) have laid out and shown empirically in their
review article. We follow their procedures and extend their results by testing two MLE
( estimators) for the Pareto II distribution, now more widely known as the $q$-exponential,
against other models (exponential, stretched exponential, lognormal; and nested models such
as power law plus exponential cut-off). We do so for several continuous distributions sets of
data (populations of cities, sales of books, forest fire sizes, blackouts) which Clauset et al.
(2007) use in their review to exemplify how to evaluate fit to power-law distribution models.

II. Test Procedures

As Clauset et al (2007) show, unbiased point data estimates can be obtained and tested, and
different model fits compared, in four steps:

1. Derivation and validation of appropriate maximum likelihood estimation (MLE)
   formuli for different kinds of distributions;
2. Estimating appropriate parameters with the MLE, including small sample corrections..
3. Use of the Kolmogorov-Smirnov (KS) test\(^1,2\) (Chakravarti, Laha, and J. Roy 1967) to estimate goodness-of-fit for a model with MLE parameters (using synthetic data) and estimate the lower bounds \(x_{min}\) for power-law and \(q\)-exponential behavior, or \(x_{min}\) lower-bounds on fit for other distributions that give the minimum KS value \(D\) that measures distance between distributions (Clauset et al. 2007:7-9). (Unlike the null significance test, KS probability values close to 1 are good fits, and the null hypothesis of no difference is rejected if \(p\) is small; e.g., \(p<.05\) means the distributions differ.) In these cases, the use of the KS test to measure the distance between distributions is applied to the empirical data compared to synthetic data. Unlike the commonly used \(r^2\) test the theoretical values do not form a continuous curve but variable “jiggled” lines that only approximate in their general form the theoretical curve of a model. Graphs of the synthetic data for a given model and set parameters can be quite surprising to a novice used to fitting empirical data points to a continuous and singular theoretical curve. Because of the random generation of \(x\) values for the synthetic data, no two generated “jiggled” lines are identical. They form a distribution of possible outcomes for a given theoretical distribution, a characteristic shared with “bootstrap” estimates of variance in outcomes from model probabilities. Thus, for example, “even if a data set is drawn from a perfect power-law distribution, the fit

\(^1\) http://www.itl.nist.gov/div898/handbook/prc/section2/prc212.htm
\(^2\) http://www.physics.csbsju.edu/stats/KS-test.html
between that data set and the true distribution will on average be poorer than the fit to
the best-fit [e.g., \( r^2 \)] distribution, because of statistical fluctuations” (p.11)

4. Likelihood ratio tests of competing fit as between different models. Clauset et al
(2007:13,15,19) do so only for power laws compared to fit of other distributions
(exponential, stretched exponential, lognormal, power law plus exponential cut-off).

1. Deriving an MLE

The principle of maximum likelihood parameter estimation is to find the parameter values
that make the observed data most likely under a given parameterized model. This requires
estimation that is asymptotically unbiased, consistent and efficient, i.e., providing parameter
values for large samples (or as sample size goes to infinity) that are almost always accurate
(unbiased) within error bounds that can also be accurately calculated (more specifically,
asymptotically efficient, i.e., with the lowest possible variance among all unbiased estimators,
as shown by the Cramér-Rao inequality for consistent unbiased estimators). If data drawn
with unknown parameters from a given distribution can be assumed to be independent and
identically distributed (iid), so that each element belongs to the same probability distribution
as the others and all are mutually independent) the problem simplifies because the likelihood
\( \lambda(\pi) \), for the data and model parameter set \( \pi \), can then be written as a product of univariate
probability densities \( p_\pi(x) \) for each \( x \), and can be reexpressed as a weighted sum of logs of
these densities.
1.1. Testing the MLE with synthetic data

The best way to test the accuracy of an MLE is to generate large samples of synthetic data drawn from the model with different parameter values, and see if the MLE generates these values and their expected variances. For the $q$-exponential, Clauset's QPVA.m Matlab code includes a way of generating synthetic data for the parameters from the Pareto II which, as formulated in Equation 2 of Shalizi (2007), has parameters that are algebraic equivalents of the $q$-exponential parameters.

2. Estimation with the MLE

The MLE is found by fitting the model parameter value(s) to maximize the likelihood function $\lambda(\pi)$ for the empirical data, a sum of logs weighted by these value(s). This maximum can also be found algebraically by setting to zero the derivative of the likelihood function with respect to the parameter(s).

2.1. Small sample correction for the MLE

For small samples, parameter estimates can be corrected by analysis of the tests of MLE accuracy for smaller samples (e.g., Clauset et al. 2007, Fig. 6 for power laws; Shalizi 2007,
Fig. 1, for $q$-exponentials). Error bounds estimates from asymptotic approximations should be avoided. Instead, parametric bootstrapping (Wasserman 2002:section 9.11) to obtain parameter estimates, standard errors, and confidence limits can be derived by generating many “bootstrap” samples of random numbers with the model density function $p_n(x)$.

3. Testing KS fit for a model with MLE parameters and

The likelihood framework allows tests of parameter values comparing the empirical data (using KS) to a same-sized sample of synthetic data using the model parameters estimates, returning $D$ and $p$ for each comparison. The synthetic data is generated by varying $x$ randomly (rather than in uniform intervals) and returning $y(x)$ for the function investigated. Because the intervals between rank-ordered values of $x$ are not uniform, each synthetic data set will vary, with $D$ and $p$ from the KS test varying for each comparison with the empirical data. The appropriate measure of fit is to generate point data from the model and its parameters (with the same number of points as the actual data), repeated a sufficient number of times to estimate a convergent variance around the mean of the parameter estimate, and then calculate $p$ as to where it falls within the variance estimate, with best fit at $p=1$.

Power laws, like most distributions, have cutoffs $x_{min}$ for the lower value of $x$ at which the distribution begins to lose fit to the data. The Kolmogorov-Smirnov (KS) test can be use to
estimate lower bounds $x_{min}$ for best fit (Clauset et al. 2007:7-9), again compared to synthetic data. This approach can be applied to other distributions as well.

4. Comparing different models for the same empirical data by calculating the likelihood ratio

The likelihood ratio (LR) addresses the question of which of two models is better fit to the data. Each model is has a likelihood determined by the KS test against synthetic data. Here the model with the higher ratio $R$ is the better fit, normally given as the logarithm $\ln R$ of the ratio, and with significant differences computed from the standard deviation of $\ln R$, as reviewed by (Clauset et al. 2007:13,15). By excluding nested-model comparisons (power law with exponential cut-off), non-nested LR comparisons can be normalized (2007:19).

**NEXT:**

3 Convergence and complementarity in different estimation methods, some not MLE

(Du: sample results: still needs to fit log normal, stretched exponentials, etc.)
## Fitting Results (Feb 08, 2008)

Table 1. Abbreviations of different kinds of networks

<table>
<thead>
<tr>
<th>Abbreviations</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>ISN</td>
<td>Instrumental support network</td>
</tr>
<tr>
<td>ESN</td>
<td>Emotional support network</td>
</tr>
<tr>
<td>SCN</td>
<td>Social contact network</td>
</tr>
<tr>
<td>MDN</td>
<td>Discussion networks focusing on marriage</td>
</tr>
<tr>
<td>CDN</td>
<td>Discussion networks focusing on childbearing</td>
</tr>
<tr>
<td>CoDN</td>
<td>Discussion networks focusing on contraception</td>
</tr>
<tr>
<td>ADN</td>
<td>Discussion networks focusing on ageing life</td>
</tr>
</tbody>
</table>
**Part I  Ego-Centered Networks**

<table>
<thead>
<tr>
<th>Networks</th>
<th>Before migration</th>
<th>After migration</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$a, \alpha$</td>
<td>$\beta$</td>
</tr>
<tr>
<td>ESN</td>
<td>2.94</td>
<td>1.94</td>
</tr>
<tr>
<td>SCN</td>
<td>3.44</td>
<td>2.44</td>
</tr>
<tr>
<td>ISN</td>
<td>3.10</td>
<td>2.10</td>
</tr>
</tbody>
</table>

Table 2  Fitting Results: MLE for power-law slope
Figure 1 Degree distribution of Emotional support network

Before Migration \( p = 0.58 \)  
After Migration
Before Migration $p = 0.181$

After Migration

Figure 2  Degree distribution of Social contact network
Before Migration \( p=0.895 \text{ scale}^5 \)  
After Migration \( p=0.013 \)

Figure 3 Degree distribution of instrumental support social network
### Part II  MLE for Power-Law Slope of Whole Networks

Table 2 MLE for power-law slope

<table>
<thead>
<tr>
<th>In-degree</th>
<th>Out-degree</th>
<th>HM (200)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>$a$</td>
</tr>
<tr>
<td>3.03 2.03</td>
<td>3 0.060187</td>
<td>0.203</td>
</tr>
<tr>
<td>3.28 2.28</td>
<td>3 0.053133</td>
<td>0.337</td>
</tr>
<tr>
<td>2.96 1.96</td>
<td>3 0.05262</td>
<td>0.225</td>
</tr>
<tr>
<td>2.7 1.7</td>
<td>2 0.097954</td>
<td>0</td>
</tr>
<tr>
<td>3.3 2.3</td>
<td>2 0.015401</td>
<td>0.97</td>
</tr>
<tr>
<td>2.5 1.5</td>
<td>1 0.089886</td>
<td>0</td>
</tr>
<tr>
<td>3.17 2.17</td>
<td>2 0.042422</td>
<td>0.502</td>
</tr>
</tbody>
</table>
Fig. 4. Degree distribution of HM’s instrumental support network

(a) In-degree

(b) Out-degree KS $p=0.924$
Fig. 5. Degree distribution of HM’s emotional support network
Fig. 6. Degree distribution of HM’s social contact network
Fig. 7. Degree distribution of HM’s discussion network focusing on marriage
Fig. 8. Degree distribution of HM’s discussion network focusing on childbearing
Fig. 9. Degree distribution of HM's discussion network focusing on contraception
Fig. 10. Degree distribution of HM’s discussion network focusing on ageing life
### Part III  MLE for q-exponential Model of Whole Networks – METHOD
**INCORRECT HERE AS YET – WORKING ON THIS NOW**

Fitting for q-exponential

<table>
<thead>
<tr>
<th>HM (200)</th>
<th>(x_{\text{min}})</th>
<th>(q)</th>
<th>(\kappa)</th>
<th>(\sigma)</th>
<th>(p)</th>
<th>(\text{gof})</th>
<th>(x_{\text{min}})</th>
<th>(q)</th>
<th>(\kappa)</th>
<th>(\sigma)</th>
<th>(p)</th>
<th>(\text{gof})</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.3967</td>
<td>0.79131</td>
<td>1.9949</td>
<td>0</td>
<td>0.3125</td>
<td></td>
<td>1.4384</td>
<td>0.74672</td>
<td>1.7034</td>
<td>0</td>
<td>0.27027</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1.0853</td>
<td>1.5395</td>
<td>18.043</td>
<td>0</td>
<td>0.26752</td>
<td></td>
<td>1.4541</td>
<td>0.5419</td>
<td>1.1932</td>
<td>0</td>
<td>0.30769</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1.1872</td>
<td>1.5941</td>
<td>8.5141</td>
<td>0</td>
<td>0.25153</td>
<td></td>
<td>1.3791</td>
<td>0.8567</td>
<td>2.26</td>
<td>0</td>
<td>0.26404</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1.3398</td>
<td>0.67988</td>
<td>2.0011</td>
<td>0</td>
<td>0.32394</td>
<td></td>
<td>1.4861</td>
<td>0.37002</td>
<td>0.7612</td>
<td>0</td>
<td>0.36296</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1.385</td>
<td>0.16452</td>
<td>0.42727</td>
<td>0</td>
<td>0.46281</td>
<td></td>
<td>1.4576</td>
<td>0.23426</td>
<td>0.51195</td>
<td>0</td>
<td>0.41414</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1.2588</td>
<td>0.046166</td>
<td>0.17839</td>
<td>0</td>
<td>0.64444</td>
<td></td>
<td>1.4769</td>
<td>0.17076</td>
<td>0.35807</td>
<td>0</td>
<td>0.45455</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1.3928</td>
<td>0.14457</td>
<td>0.36807</td>
<td>0</td>
<td>0.48</td>
<td></td>
<td>1.537</td>
<td>0.41939</td>
<td>0.78098</td>
<td>0</td>
<td>0.31818</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

But the fitting program is not MLE and is severely biased for discrete distributions. The probability density function is unknown as normalization does not converge.

So we review the likelihood methods that work pretty well for continuous distributions (cities, etc.) – and have known corrections for small sample biases from MLE estimates – and then hone in on an estimation method that may be unbiased for small samples although not MLE, and propose software design to find lower “cutoffs.”
4 What plausible missing alternatives need development? E.g., \textit{q-exponential}

5 What is the \textit{q-exponential} and \textit{why is it plausible}?

6 \textbf{Continuous} \textit{q}-distribution MLE solved; \textbf{problems for discrete} \textit{q}-distributions (ntwks)

7 \textbf{Alternate} fitting using \textit{q-semilog} (also an open problem)

summarized in one slide and in remainder of [4] that follows
Open systems entropy

1 The idea of entropy is that things run down in closed systems, so you need open systems to keep organized things going.

2 Tsallis 1988 generalized the entropy measure for open systems.

3 In an organized system there are lots of low energy states, tending toward randomness.

4 Then there may be some positive feedback loops from multiplicative effects.

5 These loops push some elements, links, etc., toward power-law tails of distributions.

6 Fitting power-laws PL doesn’t deal with the lower “cutoff” that starts where PL ends.

7 Tsallis’s $1/(1-q)$ fits the asymptote to PL with a lower crossover toward the exponential. The $q$-exponentials, with an extra $(1+ stretch)$, a raised power $1/(1-q)$ scale, an extra $x/\kappa$ shape ($\kappa$ the crossover parameter) to the curve $e_q^x = (1+(1-q) x/\kappa)^{1/(1-q)}$ have a very particular curvature (but are straight-line in $q$-semilog). 8 Our discovery was that there was also a lower “cutoff” for the $q$-exponential where support for organized energy and characteristic structure ends. 9 We found better Kolgomorov-Smirnov and likelihood tests gof than power laws for many phenomena [ref 4] (cities, etc.), did simulations…etc.

III. Application to the Pareto II (q-exponential): Theory

Constantino Tsallis (1988) introduced $q$-exponentials into the literature of statistical mechanics to provide a generalization of the Boltzmann-Gibbs(-Shannon?) entropy measure which he called $q$-entropy. He argued on theoretical grounds for their applicability to systems with long-range interactions. Without the application to $q$-entropy, and with a different but algebraically equivalent pair of parameters, $q$-exponentials were previously known in the statistical literature as Pareto II distributions for which MLE solutions were known (McGuire, Pearson and Wynn 1952, Silcock 1954, Harris, 1968). The algebraic transformation between the two, from the Pareto exponent (power) $\theta$ to $1/(q-1)$ is motivated partly by the fact that as $q$ approaches 1 the $q$-entropy, based on a $q$-exponential function, converges to equivalence with the standard Boltzmann-Gibbs entropy measure.

1. The $q$-exponential function and its inverse
The $q$ exponential and its inverse, the $q$ logarithm (Tsallis 2004:4-6), are functions derived from many convergent sources, including the solution to a simple dynamical equation, where $y(0)=1$, and

$$\frac{dy}{dx} = y^q \text{ (equation 1)}$$

The solution defines the $q$-exponential function

$$e_q^x \equiv [1 + (1 - q)x]^{1/(1-q)} \text{ (equation 2)}, \text{ where } (e_1^x = e^x), \text{ and } (1 + (1 - q)x > 0);$$

otherwise $e_q^x = 0$. Without loss of generality, a constant such as $\lambda = 1/\kappa$ may be inserted before $x$ and will carry over before $x$ in equations 1 and 2).

The inverse of the $q$-exponential is the $q$-logarithmic function:

$$ln_q x \equiv \frac{x^{1-q} - 1}{1 - q} \text{ (equation 3)} \text{ where } (\ln_1 x = \ln x), \text{ thus } ln_q (e_q^x) \equiv x$$

With the optional constant $\kappa$
\[ \ln_q x \kappa \equiv \frac{(x\kappa)^{1-q} - 1}{1 - q} \quad \text{(equation 3')} \]

where \( \ln_1 x \kappa = \ln x \kappa = \ln x + \ln \kappa \)

If \( e_q^x \) is normalized with \( \int_0^\infty dx \ p_{e_q^x}(x) = 1 \) to a probability distribution then

\[ p_{e_q^x}(x) \equiv (2-q)e_q^{-x}(x \geq 0) = (2-q)[1 - (1-q)x]^{1/(1-q)} \quad \text{(Tsallis equation 4)} \]

The cumulative probability distribution for the \( q \)-exponential function in (equation 1) is also a \( q \)-exponential function, with \( q_M \equiv 1/(2-q) \), so that (Tsallis 2004:8):

\[ P(X) \equiv \int_X^\infty dx \ p_{e_q^x}(x) = e_q^{-x/[\kappa]} = [1 - \frac{(1-q)}{(2-q)} x]^{(2-q)/(1-q)} \quad \text{(Tsallis equation 5)} \]

where for \( q > 2 \), \( p(x) \) is not normalizable, so that \( q \) cannot be fitted for a range at or above 2.

2. Equivalence of Pareto II and the \( q \)-exponential
Shalizi (2007) transforms the $q$-exponential function to a distribution, the Pareto II, for which the MLE is already known taking as his cumulative distribution

$$P_{q,\kappa}(X) = e^{-X/\kappa} = [1 - (1 - q_M)x/\kappa]^{1/(1-q_M)} = [1 - (1 - q/2 - q)x/\kappa]^{2-q/2} \text{ (equation 5)}$$

Substituting to obtain Pareto II parameters, $\theta$ and $\sigma$, we may take Shalizi’s equations (8) and (10) for the MLE estimation of $\hat{\theta}$, $\hat{\sigma}$ in terms equivalent to the estimates of $q, \kappa$.

$$q_M \equiv \frac{1}{2-q} = \frac{\theta + 1}{\theta} \text{ hence } q_M = \frac{\theta + 2}{\theta + 1} \neq \frac{\hat{\theta} + 1}{\hat{\theta}} \text{ as in Shalizi and thus}$$

$$\hat{q} = q_M = \frac{\hat{\theta} + 2}{\hat{\theta} + 1}$$

Shalizi’s $q$ ($q_M$), then, pertains to the $q_M$ for Tsallis equation 5.

$$\hat{\kappa} = \sigma/\theta$$
In should be the case that the cumulative $q$-exponential asymptotes in the tail to a power-law slope of $1 / (1 - q_m)$.

3. The MLE for parameters of Pareto II (and $q$-exponentials)

For maximum likelihood we need the correct probability density function. Following Shalizi (2007:1-2, and his eq. 5), under the $q$-exponential model with parameter $q$, or the Pareto II with parameter $\eta=\theta$, the probability density of $P_{\theta,\sigma}(X \geq x)$, and $\eta=\theta$ is

$$p_{\theta,\sigma}(x) = p_\eta(x) = (\theta/\sigma) \left(1 + x/\sigma\right)^{-\theta-1}$$

and the log of $p_\eta(x)$ equals the log-probability density of a sequence of independent identically-distributed samples is a log-likelihood function $\ell(\eta)$:

$$\log p_\eta(x^n) = -(\eta + 1) \sum_{i=1}^{n} \log \left(1 + x_i/\eta\right) = \ell(\eta) \text{ (equation 8)}$$

Then the derivative of the log-likelihood function with respect to $\eta=\theta$ (following Shalizi 2007:2: his equation 7) is:
\[ \frac{\delta \ell}{\delta \eta} = \frac{n}{\eta} - \sum_{i=1}^{n} \log \left(1 + \frac{x_i}{\eta}\right) \] (equation 9)

To find the MLE (maximum likelihood estimate, following Shalizi 2007:2, and his equation 8.) we take the first derivative of the log-likelihood \( \ell(\eta) \) with respect to the parameter \( \eta \) and set it equal to zero to solve for the maximum.

\[ \hat{\eta} = n \left[ \sum_{i=1}^{n} \log \left(\frac{x_i}{x_{\text{min}}}\right) \right]^{-1} \] (equation 10)

where in the simplest case, \( x_{\text{min}} = \eta \), but we may also take the value of \( x_{\text{min}} \) that maximizes fit according to the Kolmogorov-Smirnov (KS) bootstrap method.
Thus, we use Shalizi’s (2007) estimate of $q_w$ for the $q$-exponential, which he also tested for quality (Figure 1: explain)
4. The KS fit for $q$-exponential models with MLE parameters

Using Shalizi’s $q_W$ estimates, Clauset wrote a new Matlab program, QPVA.m, to complement the Matlab program package used in Clauset, Shalizi and Newman. (2007), now allowing $q$-exponential (Pareto II) fitted models to be compared to Pareto models for a given distribution. What is completely novel about his procedure is that his procedure validated the discovery by White, Tambayong and Keyžar (2007), of an $x_{min}$ for $q$-exponentials. QPVA(x, $x_{min_{test}}$) takes data x, computes the $q$-exponential behavior with a lower cutoff $x_{min_{test}}$, and the corresponding p-value for the Kolmogorov-Smirnov (KS) test. It iterates values of $x_{min_{test}}$ to obtain the value $x_{min}$ that gives the highest KS probability of fit. With $x(0)$ the lower limit of the observed data series, and $\min(x(0), x_{min}) = \min(x)$, if $\min(x) \geq 1000$, QPVA uses the continuous approximation, which is a reliable in this regime. The fitting procedure works as follows: 1) For each possible choice of $x_{min_{test}}$, we estimate the $q$-exponential parameters via the method of maximum likelihood (Shalizi’s MLE), and calculate the Kolmogorov-Smirnov goodness-of-fit statistic D. 2) We then select as our estimate of $x_{min_{test}}$, the value $x_{min}$ that gives the minimum value D over all values of $x_{min_{test}}$. (With the data we tested, $x(0) < x_{min}$ in each data series, so that the optimal $x_{min}$ represented an empirical finding pertinent to fit.)

IV. Application to the Pareto II ($q$-exponential): Results
Clauset et al. (2007) tested twenty-four empirical distributions with large sample sizes, from different domains, against the power-law model, and then compared their fit that of three plausible alternative models (log-normal, exponential and stretched exponential) and to the power law with a cut-off for the tail of the distribution. Fourteen of the empirical distributions were discrete data sets, or frequency distributions starting with low integer frequencies, and ten were continuous. Here, we consider comparisons with the $q$-exponential but only for the continuous data sets because it is not know how to do the probability normalization for the discrete case. For these data sets none showed good fit to a power law, e.g., with KS $p \geq 0.40$ and no alternative with a LR alternative significant at $p \geq 0.20$. Six had moderate fit, good in terms of KS $p \geq 0.40$ but with plausible alternatives. Six had a power-law fit but with an exponential cut-off clearly favored over a pure power-law.

Our findings are shown in Table 1. For our analysis of $q$-exponential fit, we eliminated two of the six data sets with moderate fit (wars and religion) because of irregular distributions that on visual inspection fit neither the power law nor the $q$-exponential. The size of religion distribution has a clear breakpoint between two power-law regimes, the tail for world religions having a thicker tail with greater sizes than expected for the lower power-law regime. That left four continuous data sets with moderate support for power law: bird-species sightings, customers affected by power blackouts, sales quantities for Amazon books, and populations of U.S. cities in the year 2000. Visually, each of these resembled a $q$-exponential distribution (see graphs j-m in Clauset at al. 2007). Statistically, each had a very high KS probability, as shown in Table 1, much higher in each case than KS power-law probability.
The likelihood ratio for comparison of the two distributions _____ (Laurent). In each case the tail of $q$-exponential converges with the fitted power law slope $\theta$ consistent with the theoretical expectation that the tail of the $q$-exponential asymptotes to slope $\theta=1/(1-q)$. (Laurent). The Figures 1-4 for these distributions show the fitted curves but WE ARE HAVING TROUBLE PROPERLY GRAPHING THE FIRST TWO (Laurent).

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Birds</td>
<td>0.55</td>
<td>0.99 .023</td>
<td>1.87 0.06</td>
<td>-0.882 0.38</td>
<td>-1.24 0.12</td>
<td>moderate</td>
<td></td>
</tr>
<tr>
<td>Blackouts</td>
<td>0.62</td>
<td>0.97 .033</td>
<td>1.21 0.23</td>
<td>-0.417 0.68</td>
<td>-0.382 0.38</td>
<td>moderate</td>
<td></td>
</tr>
<tr>
<td>Book sales</td>
<td>0.66</td>
<td>0.88 .018</td>
<td>2.70 0.00</td>
<td>3.885 0.00</td>
<td>-0.140 0.60</td>
<td>moderate</td>
<td></td>
</tr>
<tr>
<td>Cities</td>
<td>0.76</td>
<td>0.94 .009</td>
<td>3.65 0.00</td>
<td>0.204 0.84</td>
<td>-0.123 0.62</td>
<td>moderate</td>
<td></td>
</tr>
<tr>
<td>Surnames</td>
<td>0.20</td>
<td>0.66 .028</td>
<td>2.89 0.00</td>
<td>-0.884 0.40</td>
<td>-1.360 0.10</td>
<td>with cutoff</td>
<td></td>
</tr>
<tr>
<td>Fires</td>
<td>0.05</td>
<td>0.56 .016</td>
<td>4.00 0.00</td>
<td>-1.820 0.07</td>
<td>-5.020 0.00</td>
<td>with cutoff</td>
<td></td>
</tr>
</tbody>
</table>

*Note: higher positive log-likelihood p values in our table indicate better fit for the column as labeled

Table 1. (From Clauset et al. 2007:19 Table VI) plus $q$-entropy results.
Survival function of birds data

Cumulative probability

x
black=empirical, blue=Tsallis, red=Pareto

OUR ORIGINAL GRAPH
q=1.91  xq-min=500  sp=2.12  xp-min=6679

LAURENT’S GRAPH - WRONG

LAURENT’S GRAPH - WRONG

q=1.71  xq-min=1000  xp-min=6679

Smith et al 2003 Body mass of late Quaternary mammals
Survival function of blackouts data

\[ x_{\text{zagged}} = \text{empirical}, \ x_{\text{solid}} = \text{Tsallis}, \ x_{\text{dashed}} = \text{Pareto}, \ x = q = 1.56, \ x_{\text{q-min}} = 1000, \ x_{\text{p-min}} = 2.27, \ x_{\text{xp-min}} = 230000, \ x_{\text{p}} = 2.3 \]

OUR ORIGINAL GRAPH
Survival function of blackouts data

Cumulative probability

zagged=empirical, solid=Tsallis, dashed=Pareto,
q=1.55 \text{xq-min}=75000 \text{xp-min}=230000
Survival function of cities data

Cumulative probability

x

l=empirical, solid=Tsallis, dotted=Pareto, xq-min=5000, OUR ORIGINAL GRAPH
q=1.78 sp=2.37 xq-min=5000 xp-min=52640
2.37
Survival function of cities data

Laurent's

\( q = 1.65 \)  \( \times_{q\text{-min}} = 13908 \)  \( \times_{p\text{-min}} = 52640 \)

\[ x \]

zagged = empirical, solid = Tsallis, dotted = Pareto, \( x_{q\text{-min}} = 5000 \), \( x_{p\text{-min}} = 52640 \)

Laurent's

\[ q = 1.65 \]  \( \times_{q\text{-min}} = 13908 \)  \( \times_{p\text{-min}} = 52640 \)

0.76  0.94  0.009
VISIBLY BETTER THAN THE 5,000 CUTOFF
$q [1] 1.651811$
$kappa [1] 9659.442
$shape [1] 1.534187$
$scale [1] 14819.40$
$loglike [1] -23955.43$
$n [1] 2085$
$xmin [1] 13908$
$method [1] “mle.equation”

For the six data sets with power-law with an exponential cutoff, flares and earthquakes had irregular distributions that on visual inspection did not resemble the $q$-exponential as did fires and, to a lesser extent, surnames. Web hits and web links had a visually clear a power-law fit with an exponential cut-off favored over a pure power-law. Hence, we eliminated four of these six data sets and tested two for $q$-exponential fit: fires and surnames.

Their graphs are shown below but again, we have trouble in the first case with the graphing command.
Survival function of fires data

q = 1.72  xq-min = 2200  xp-min = 6324

LAURENT’S (WRONG)
Survival function of surnames data
V. Two-parameter $q$-semilog fits for the $q$-exponentials

Thurner, Kyriakopoulos, and Tsallis (2007:5) note that “A convenient procedure to perform a two-parameter fit is to take the $q$-logarithm of the distribution $P$, defined as”

$$Z_q(x) \equiv \ln_q P_q(x) = \frac{[P(x)]^{1-q} - 1}{1 - q} = \frac{[1 - (1 - q_M)x]^{2-q}}{1 - q}.$$ 

“This is done for a series of values of $q$. The function $Z_q(x)$ which can best be fit with a straight line determines the value of $q$, the slope being $\kappa$.”
Hence with slope $\kappa$, the two-parameter solution estimating $q$ and $\kappa$ optimizes a straight line for this theoretical expectation:
This needs a solution by numerical methods, which we have tested with Excel solver. The linear fit estimates for successive lower cutoffs are tested for KS likelihoods as iterations within a QXPVA.m that saves the solution for a cutoff with the largest likelihood p. Matlab program This is provided with QPVA.m and other materials with this article. (DRW - Will also need testing).

**VI. Conclusions: Pareto II (\(q\)-exponential) and other fits**

Clauset, Shalizi, and Newman (2007) provided an all-important review of what is needed to fit theoretical models to empirical distributions, illustrated with a focus on fitting power laws in relation to other plausible univariate distributions. They showed the fundamental importance of likelihood methods, from maximal likelihood (MLE) estimation of parameters to the use of Kolmogorov-Smirnov (KS) tests using synthetic data to generate the scatter of theoretical distributions consistent with estimated parameters for a given model. They provide a package of R programs that include estimating MLE parameters for each of four pure distributions (power law, exponential, stretched exponential, and log-normal) and one nested (power-law plus exponential cut-offs). The also provide a package of Matlab programs that generates the synthetic distributions for KS tests for each of these distributions, given

\[
 x \kappa \approx \frac{[P_q(x)]^{1-q} - 1}{1 - q} = [1 - (1 - q_M)x]^{1/(1-q_M)} > 0
\]
particular parameters, and for estimating the likelihood ratio and its variance for comparisons of goodness-of-fit for pairs of competing models. Their fits and comparisons of fit to the power law for 24 data sets shows that log-normal distributions are strong competitors in five cases (all for continuous variables: fires, HTTP, earthquakes, web hits, web links), exponentials or stretched exponentials are strong competitors in all but one of the 24, and power-law plus cut-offs for ten of the 24. These findings give extreme caution against inappropriate fitting of power laws by $r^2$ fit, as is most commonly done.

Our findings complement and reinforce those of Clauset et al., adding the $q$-exponentials to the set of models that should be compared as an alternative to power laws. We do not yet have a way of finding MLE solutions for parameter estimates of the $q$-exponentials for discrete distributions (DRW - can the one-parameter MLE be modified to suit this need?). Our contribution for continuous distributions is to show that four of the six continuous distributions classified by Clauset et al. as having moderate support for power-law fit have a much higher fit to the $q$-exponential. Of the two remaining data sets numbers by religion have a $p=0.42$ to power law with a slightly better fit for a stretched exponential (and a visual cut-off to a fat-tail), while intensity of wars oddly wobbles around a power law with $p=0.20$ KS fit over three orders of magnitude. Over all six, the $q$-exponential strongly outperforms power law fit. Further work, however, would be required for likelihood ratio tests of $q$-exponential fit versus that of the stretched exponential, exponential, or lognormal.
One of three further advantages of the $q$-exponential over the power law is that our one-parameter MLE for $\tilde{q}$ in the $q$-exponential matches that of the maximal likelihood for the estimate of slope $\alpha$ in the power law. The second is that $1/(1-q)$ asymptotes to power law slope $\alpha$ in the tail of the $q$-exponential and so subsumes many observed power-law tendencies while accounting for a fitting, as well, the more exponential form that is common in most empirical distributions that might be attributed to a power law with a lower cutoff $x_{\text{min}}$. There is, however, no theoretical interpretation for $x_{\text{min}}$ in most distributions attributed to a power law. Third, as we shown for the U.S. 2000 city-size distribution, there is a potential a priori theoretical interpretation for $x_{q\text{min}}$ although this has never been recognized to date. For example, we commonly distinguish rural settlements which lack a diverse division of labor that supports bilateral exchange from cities with a complex division of labor supporting complex forms of exchange. Cities are archaeologically distinguishable from other settlements, and their minimum size archaeologically begins at 4-5,000 people. If cities have an $x_{q\text{min}}$ of 14,000 it probably means that this is a significant threshold at which urbanity begins, perhaps only three times greater than the thresholds of urbanity 5,500 years ago. The power law and $q$-exponentials, then, in formalized form, have $x_{q\text{min}}/x_{\text{min}}$ thresholds that are estimated as a second parameter beyond their slope (scale) coefficients. But we have also shown that both the $q$-exponential have a normalized form in which a direct MLE solution is available for $q$, like that for $\alpha$ in the power law. This normalized form, however, give an additional parameter, $\kappa$, for the crossover, in normalized form between the asymptote toward power law in the tail, above $\kappa(1-q)$, and asymptote toward the horizontal axis at $P=1$ as $x \to x_{q\text{min}}$, something that is entirely lacking in the power law.
Given the methods for estimating the parameters of univariate distributions by likelihood methods, and our extension of these methods now to $q$-exponentials as proven viable alternates to power law and other distributions, two additional steps are needed. One is the calibration of small sample corrections for MLE estimates, which have been shown to be needed for power laws and $q$-exponentials. The other is to extend the package of Matlab programs the likelihood ratio tests between $q$-exponentials and univariate distributions other than the power law.

References


