Making Empirical Power-Law Distributions Useful

or: How to Fit Curves & Do Statistical Validation With High-Variance Data, and an Application of these Techniques to Some Empirical "Power Laws" (with Predictable results)

Aaron Clauset
Santa Fe Institute

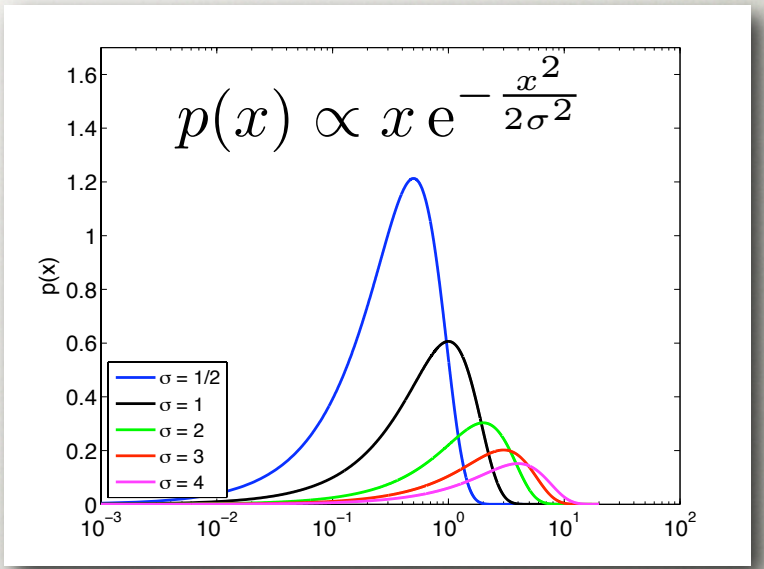
Algorithms, Inference, and Statistical Physics Workshop
4 May 2007

Joint with Cosma Shalizi and M. E. J. Newman

Statistical Distributions

Low-variance ("thin-tailed") distributions

- Maxwell-Boltzmann
- Gaussian (Normal)
- Rayleigh
- etc.

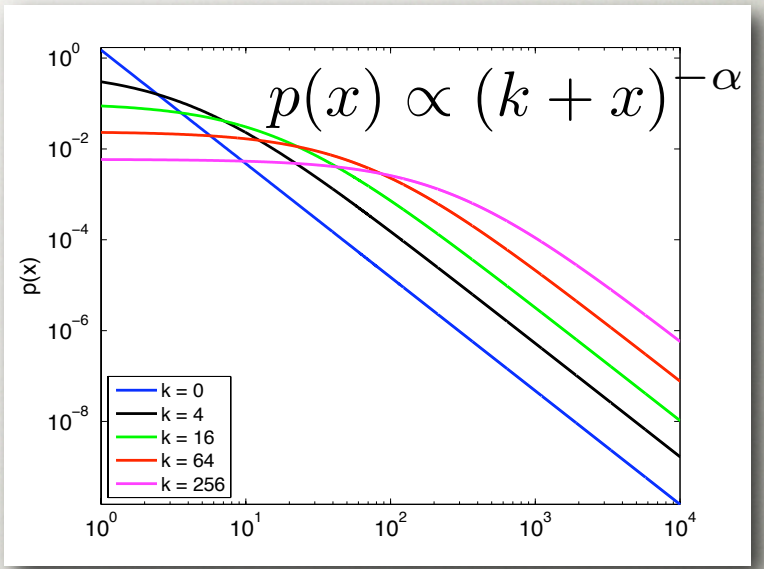


Mean is representative of almost all samples

Statistical Distributions

High-variance ("heavy-tailed") distributions

- Zipf
- Yule-Simon
- "power law"
- etc.



Mean is *not* representative of most samples

Motivation

PL distributions are mathematically **interesting.**

(imply heterogeneity & scalability)

PL distributions **appear ubiquitous in data.**

img. source: wikipedia

Some Problems

- Common methods misestimate the scaling behavior α

  "Doesn't scale the way you think it does."

- Behavior rarely validated (i.e., compared to alternative, non-PL models)

  *"Maybe it's **not** a power law."*

Some Goals

How can empirical PL distributions be useful?

- Useful for model selection

  "What kind of theory should I look for?"

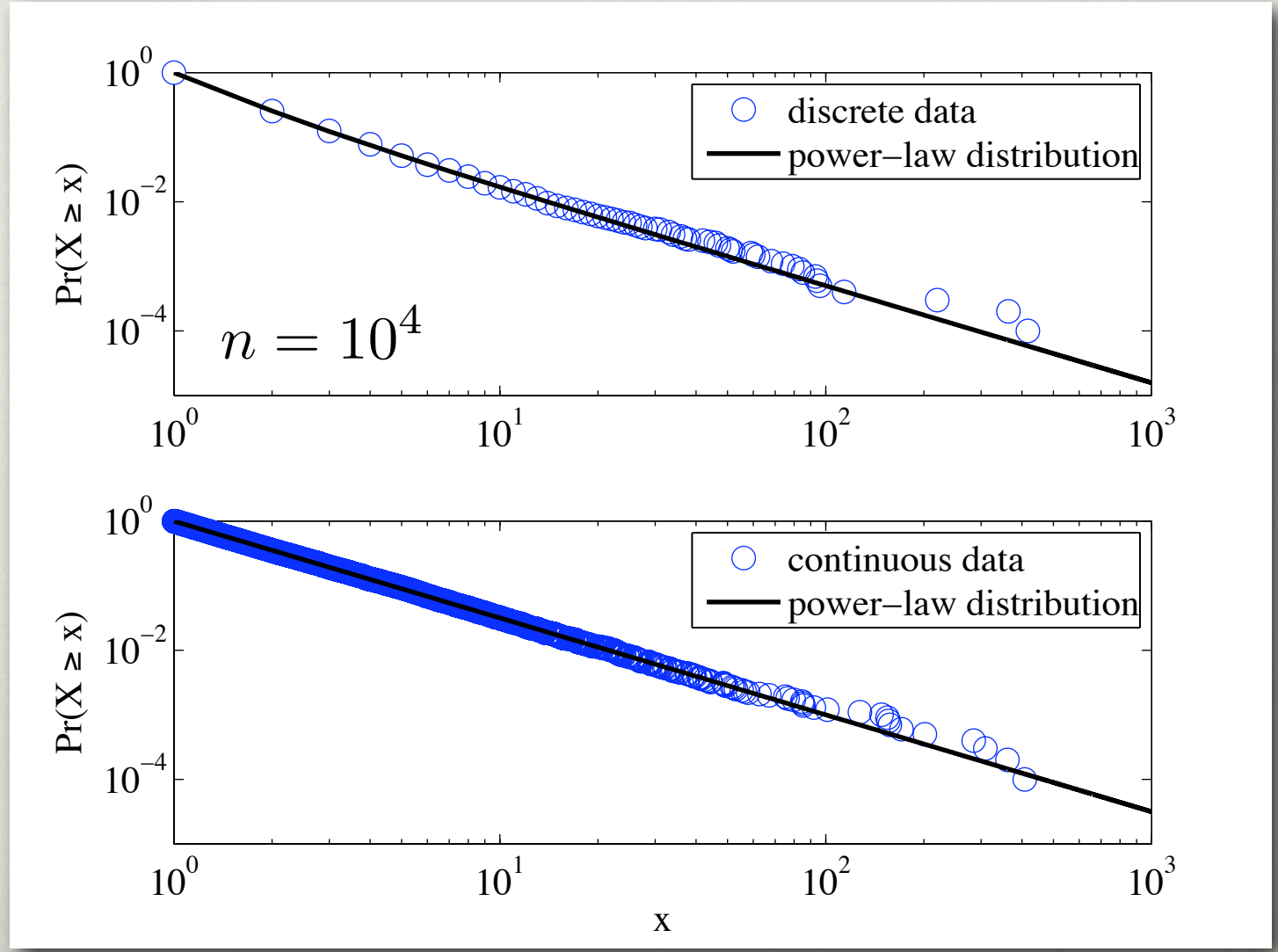
- Useful for (soft) validation

  "My theory explains my data."

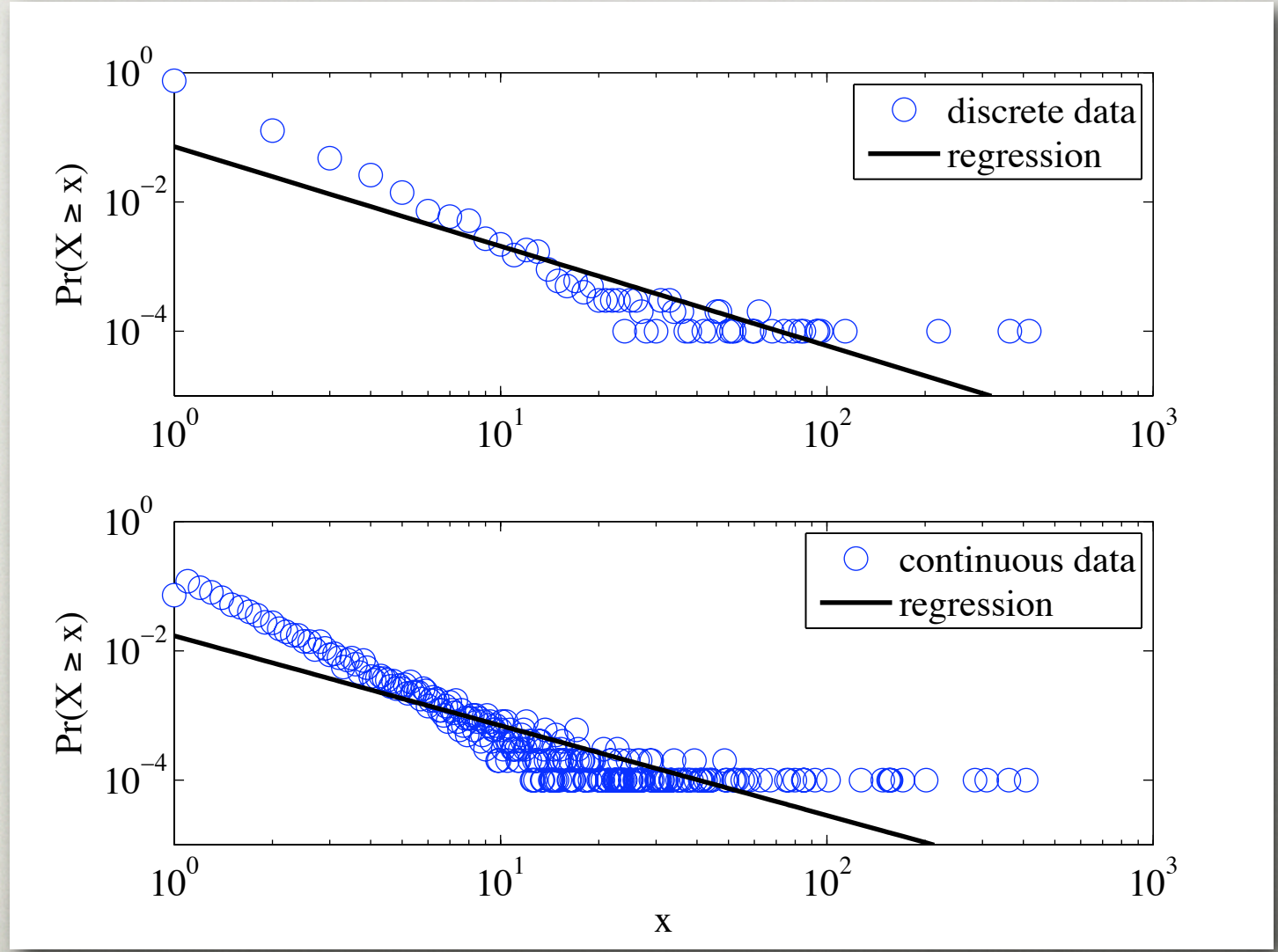
- Deviations point to interesting phenomena

  "My theory doesn't explain this part."

Estimating α



Regression on pdf



ML Estimators

Continuous power-law distribution

$$\hat{\alpha} = 1 + n \Bigg/ \left[\sum_{i=1}^{n} \ln \frac{x_i}{x_{\min}} \right] \qquad \sigma_{\hat{\alpha}} \simeq \frac{\hat{\alpha} - 1}{\sqrt{n}}$$

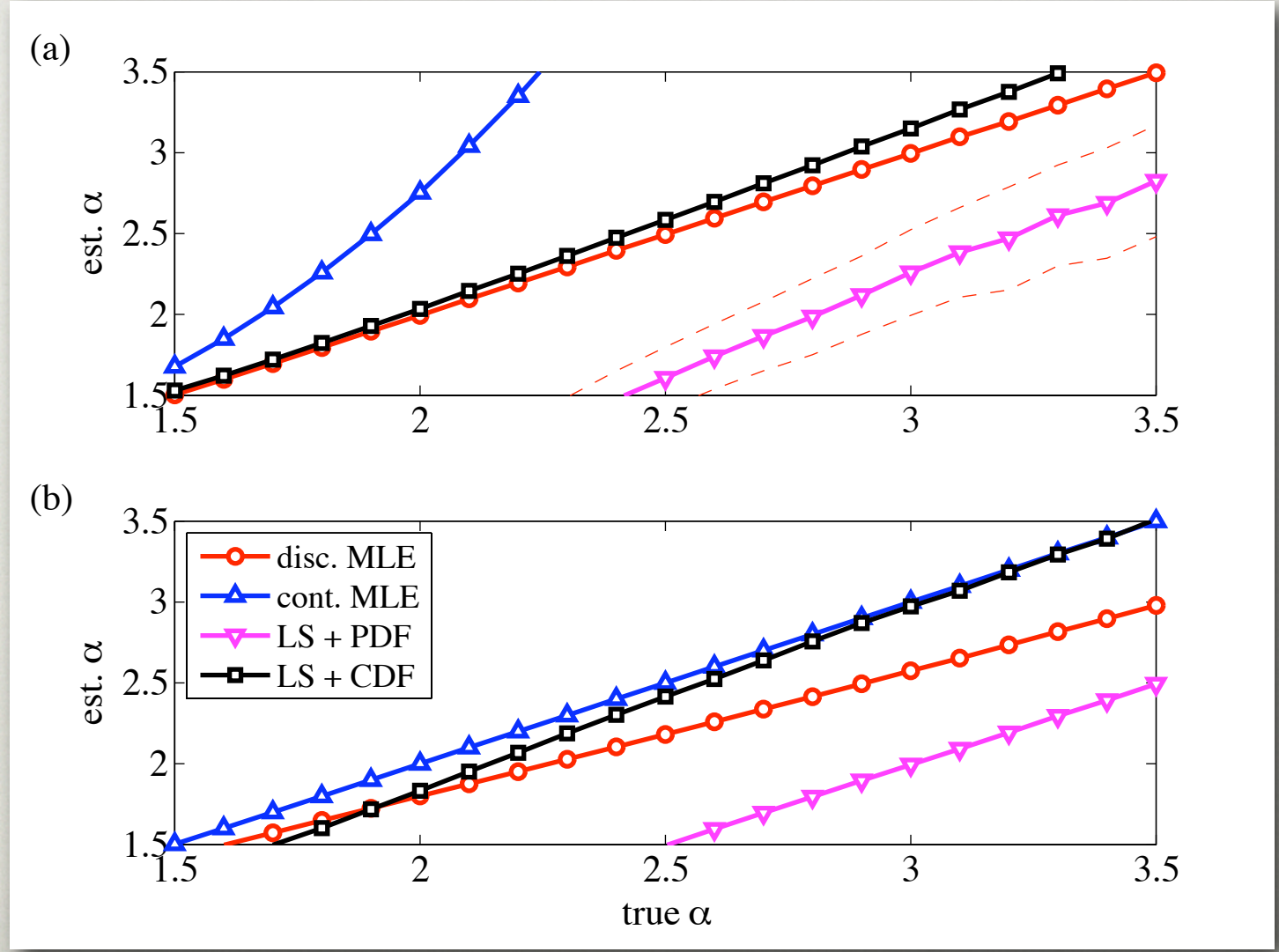
Discrete power-law distribution

$$\frac{\zeta'(\hat{\alpha}, x_{\min})}{\zeta(\hat{\alpha}, x_{\min})} = -\frac{1}{n} \sum_{i=1}^{n} \ln x_i$$

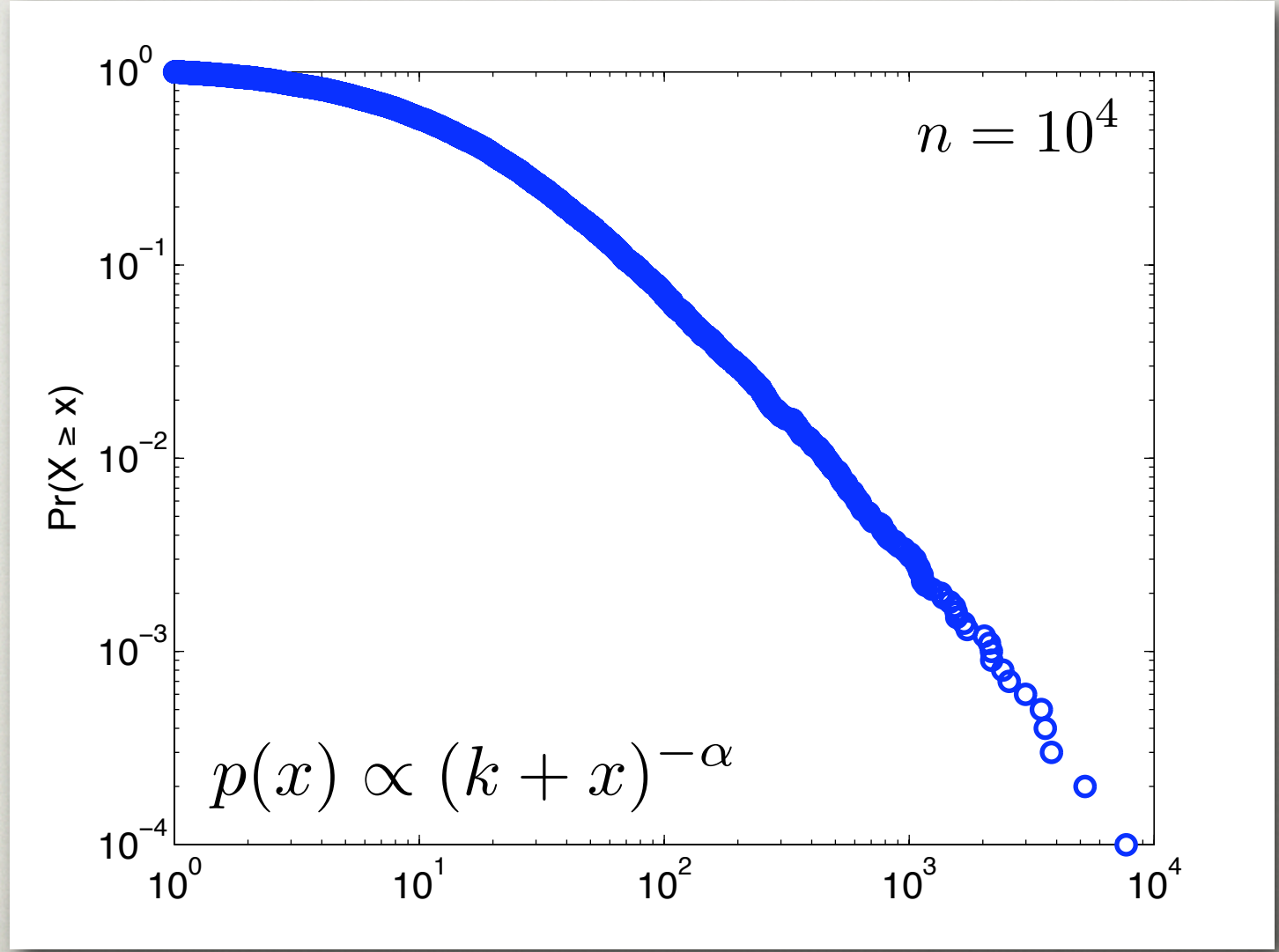
Comparison

method	notes	est. α (discrete)	est. α (continuous)
LS + PDF	const. width	1.5(1)	1.39(5)
LS + CDF	const. width	2.37(2)	2.480(4)
LS + PDF	log. width	1.5(1)	1.19(2)
LS + CDF	rank-freq	2.570(6)	2.4869(3)
cont. MLE	–	4.46(3)	**2.50(2)**
disc. MLE	–	**2.49(2)**	2.19(1)

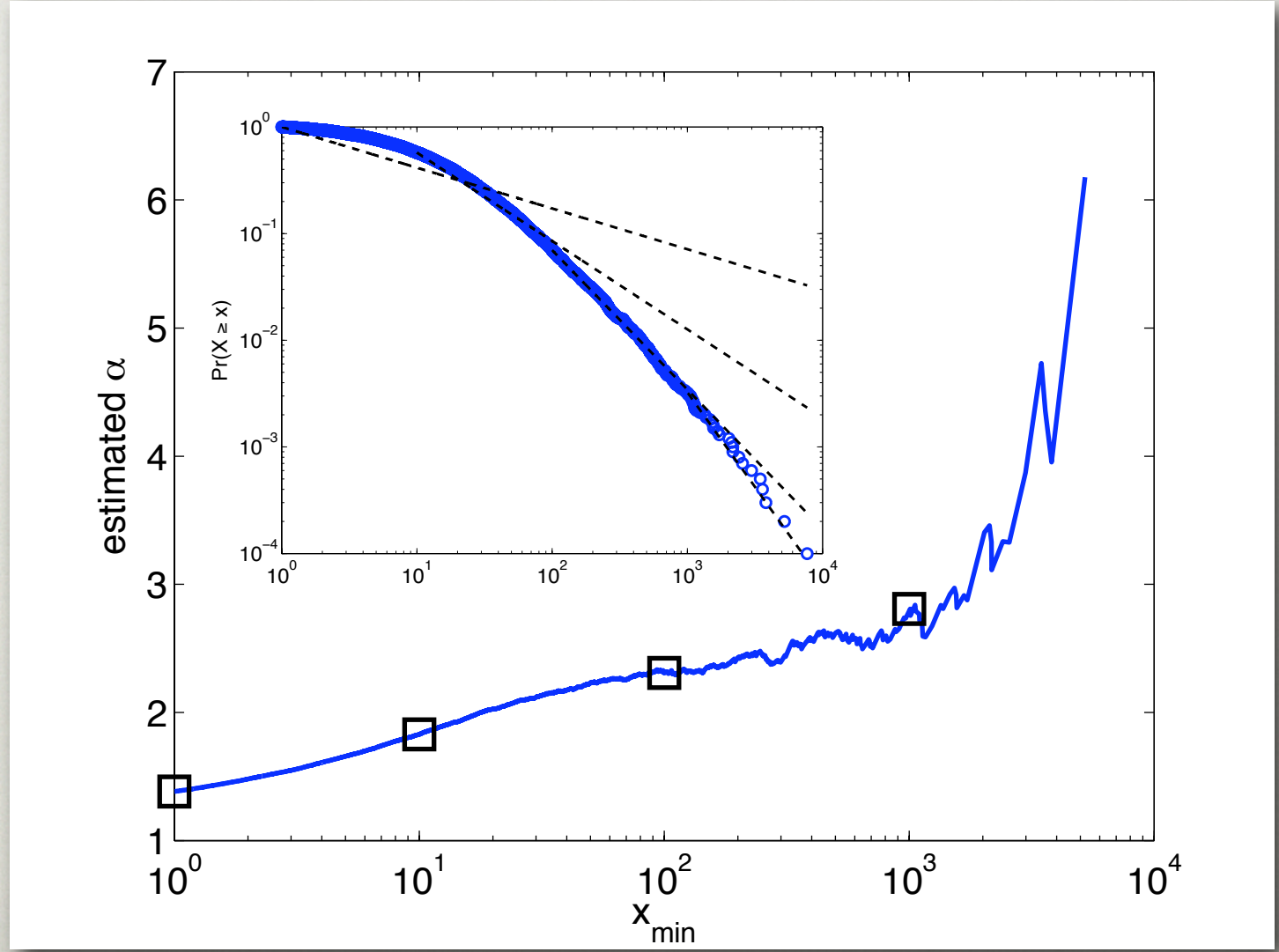
In General



Estimating $x_{\min}$



A few choices



An Objective Method

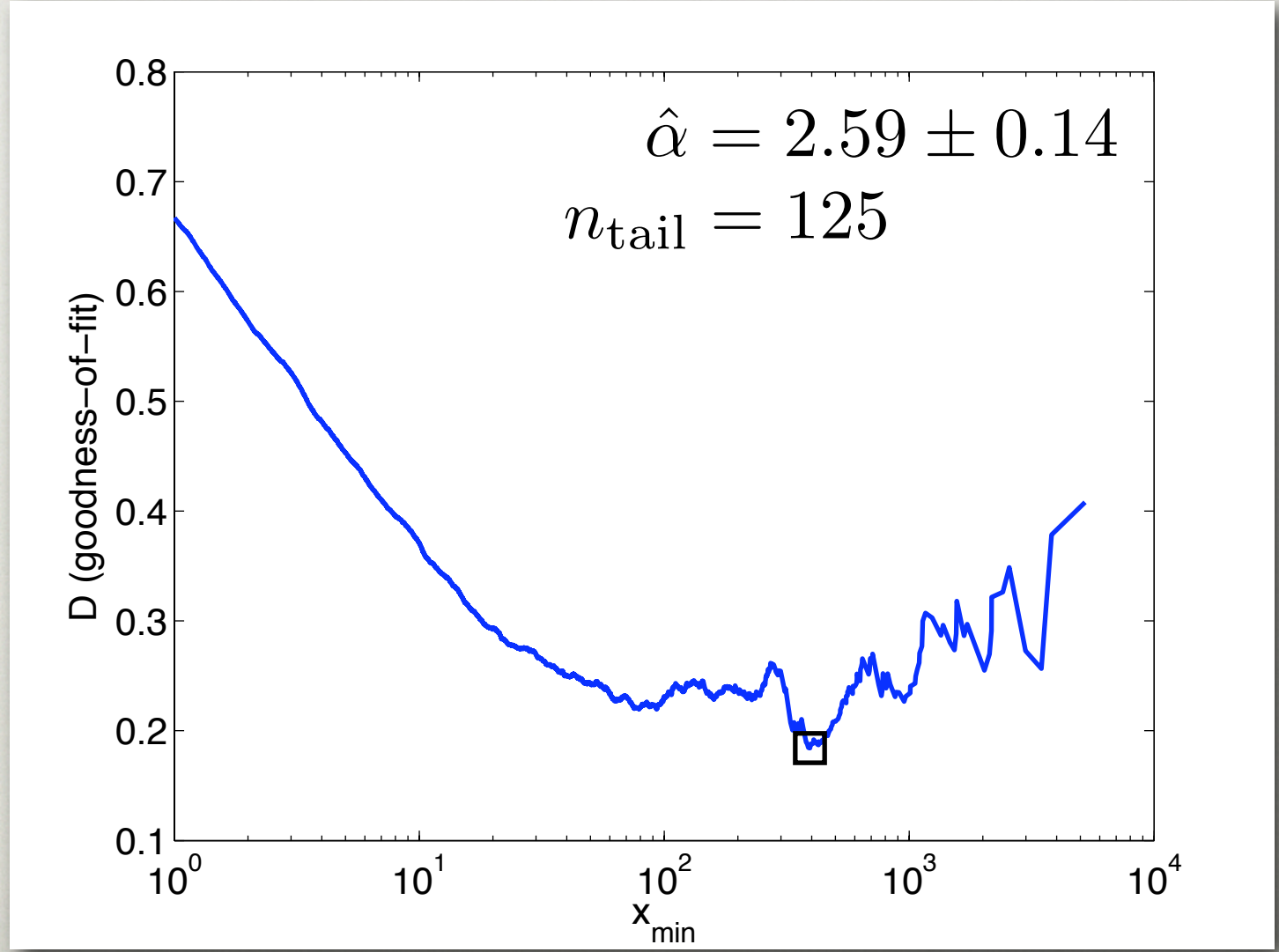
Choose power law that minimizes "distance" between model and data:

Kolmogorov-Smirnov goodness-of-fit statistic

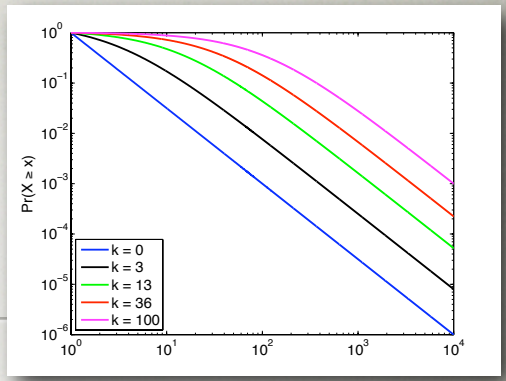
$$D(x_{\min}) = \max_{x \geq x_{\min}} |S_n(x) - P(x)|$$

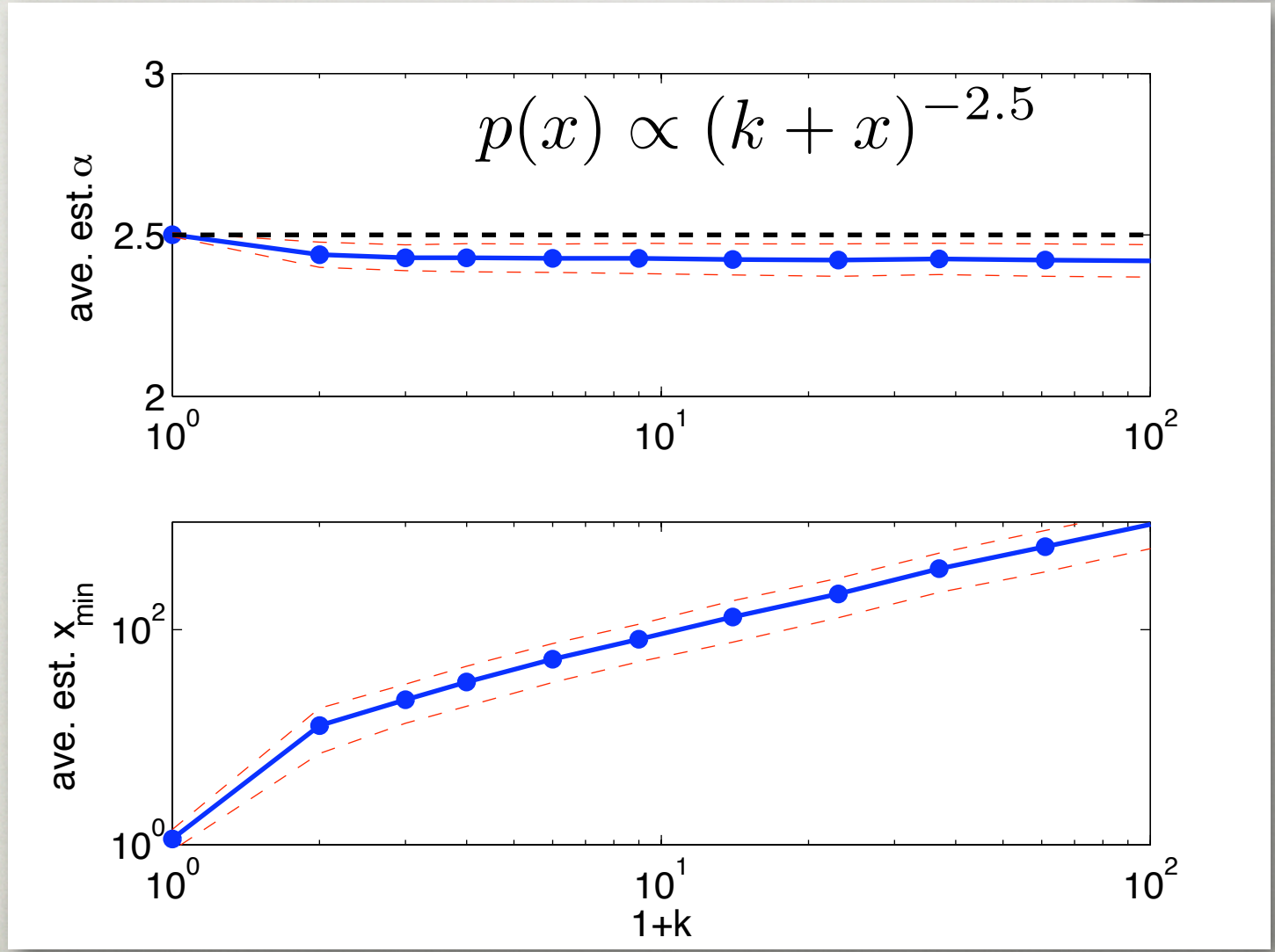
We choose $\hat{x}_{\min} = \min_y D(y)$

GoF Results


$\hat{\alpha} = 2.59 \pm 0.14$
$n_{\text{tail}} = 125$
D (goodness−of−fit)
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
10^0
10^1
10^2
10^3
10^4
x_{min}

In General

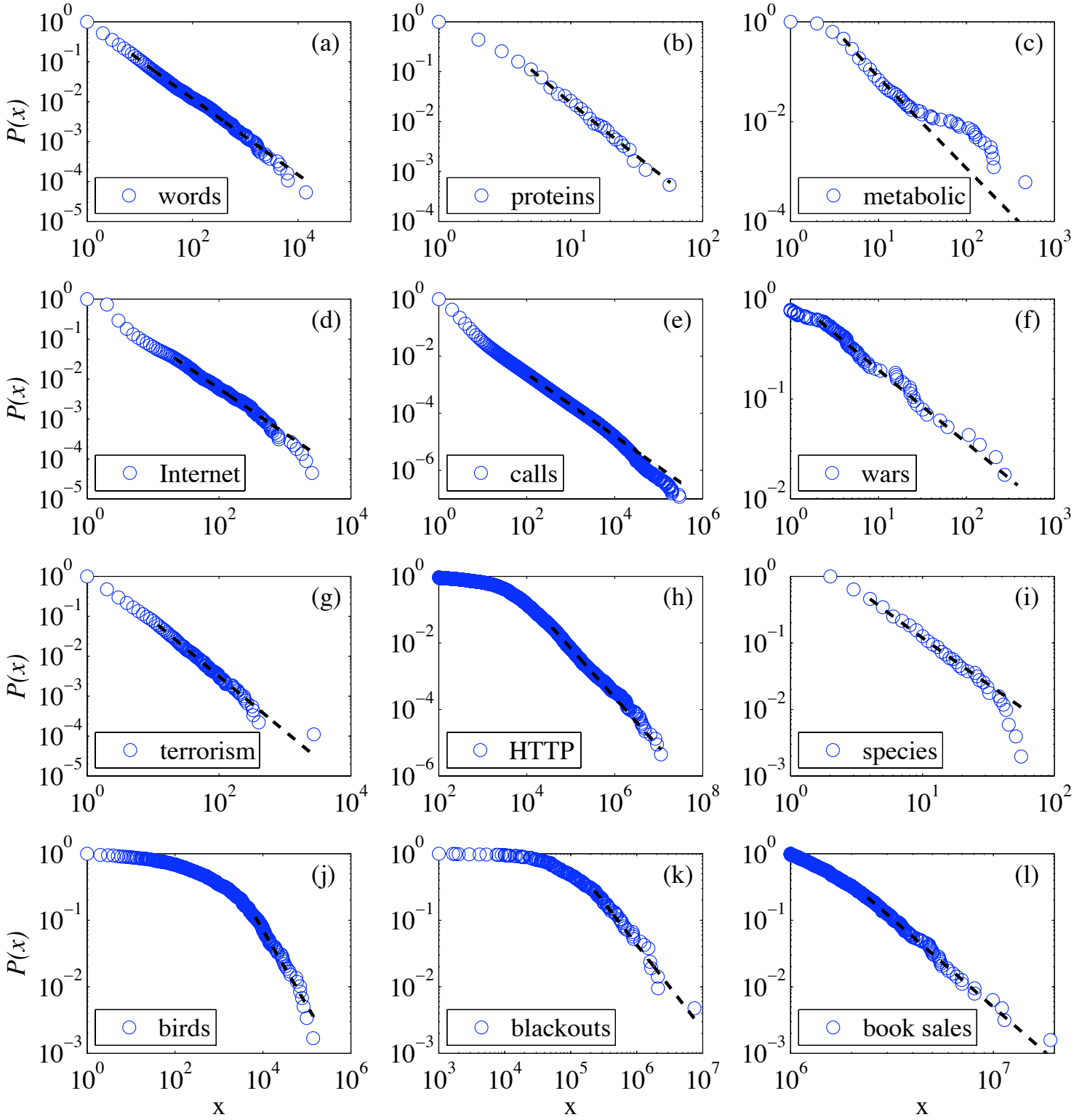

Pr(X ≥ x)
k = 0
k = 3
k = 13
k = 36
k = 100

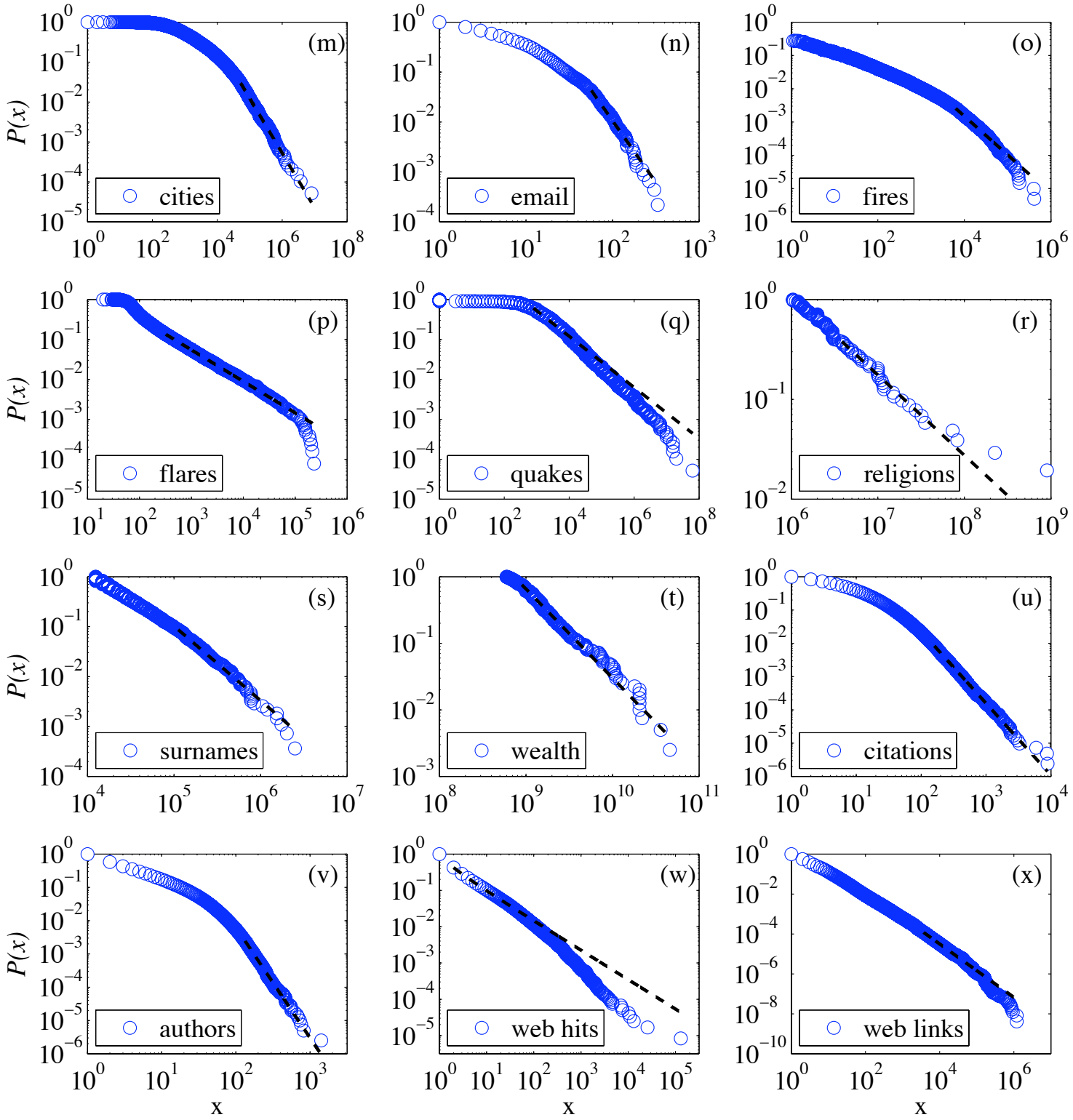

$p(x) \propto (k+x)^{-2.5}$
ave. est. α
3
2.5
2
10^0
10^1
10^2
ave. est. x_{min}
10^2
10^0
1+k

Some Real Data

count of word use
protein interaction degree
metabolic degree
Internet degree
telephone calls received
intensity of wars
terrorist attack severity
HTTP session size (kB)
species per genus
bird species sightings
blackouts
sales of books

population of cities
email address books size
forest fire size (acres)
solar flare intensity
quake intensity
religious followers
freq. of surnames
net worth (mil. USD)
citations to papers
papers authored
hits to web sites
links to web sites


(a)
words
(b)
proteins
(c)
metabolic
(d)
Internet
(e)
calls
(f)
wars
(g)
terrorism
(h)
HTTP
(i)
species
(j)
birds
(k)
blackouts
(l)
book sales
P(x)
x


(m)
cities
(n)
email
(o)
fires
(p)
flares
(q)
quakes
(r)
religions
(s)
surnames
(t)
wealth
(u)
citations
(v)
authors
(w)
web hits
(x)
web links
P(x)
x

Validation

Two approaches:

- *p*-value

  "Are deviations from PL model explained by statistical fluctuations?"

- likelihood ratio (LR)

  "Does this other model look more like the data?"

Support for Power Law

data set	power law p	log-normal LR	w. cut-off LR	support for power law
cities	**0.761**	-0.435	-0.298	reasonable
fire	0.045	-1.78	**-5.02**	w. cut-off
flares	0.997	-0.803	**-4.52**	w. cut-off
HTTP	0.000	1.59	0.000	none
quakes	0.000	-7.14	**-24.4**	w. cut-off
wealth	0.001	-0.0777	-0.198	none
web hits	0.000	0.255	0.000	none

Support for Power Law

data set	power law p	log-normal LR	w. cut-off LR	support for power law
Internet	**0.286**	-0.807	-1.97	reasonable
citations	**0.204**	-0.141	-0.007	reasonable
metabolic	0.000	-1.05	0.000	none
species	0.103	-1.63	**-3.80**	w. cut-off
terrorism	**0.684**	-0.278	-0.077	reasonable
words	**0.487**	0.395	-0.899	good

Conclusions

- Only maximum likelihood accurately estimates α
- Can now accurately (and objectively) estimate $x_{\min}$
- Validating power-law distributions possible
- Some "power laws" should be "revisited"
- Deviations from PL can suggest new hypotheses

Thanks (comments and data): L. Adamic, A. Boyer, A. Broder, A. Downey, J.D. Farmer, P. Holme, M. Huss, J. Karlin, J. Ladau, M. Mitzenmacher, C. Moore, S. Redner, S. Stoev, M. Wheatland, J. Wiener, W. Willinger and M. Young

Fin