

On Internet Traffic Dynamics and Internet Topology I

High Variability Phenomena

Walter Willinger

AT&T Labs-Research

walter@research.att.com

[This is joint work with J. Doyle and D. Alderson (Caltech)]

Topics Covered

- Motivation
- A working definition
- Some illustrative examples
- Some simple constructions
- Some key mathematical results
- Resilience to ambiguity
- Heavy tails and statistics
- A word of caution

Motivation

- Internet is full of “high variability”
 - Link bandwidth: Kbps – Gbps
 - File sizes: a few bytes – Mega/Gigabytes
 - Flows: a few packets – 100,000+ packets
 - In/out-degree (Web graph): 1 – 100,000+
 - Delay: Milliseconds – seconds and beyond
- How to deal with “high variability”
 - High variability = large, but finite variance
 - High variability = infinite variance

A Working Definition

- A distribution function $F(x)$ or random variable X is called heavy-tailed if for some $\alpha > 0$

$$P[X > x] = 1 - F(x) \approx cx^{-\alpha}, x \rightarrow \infty$$

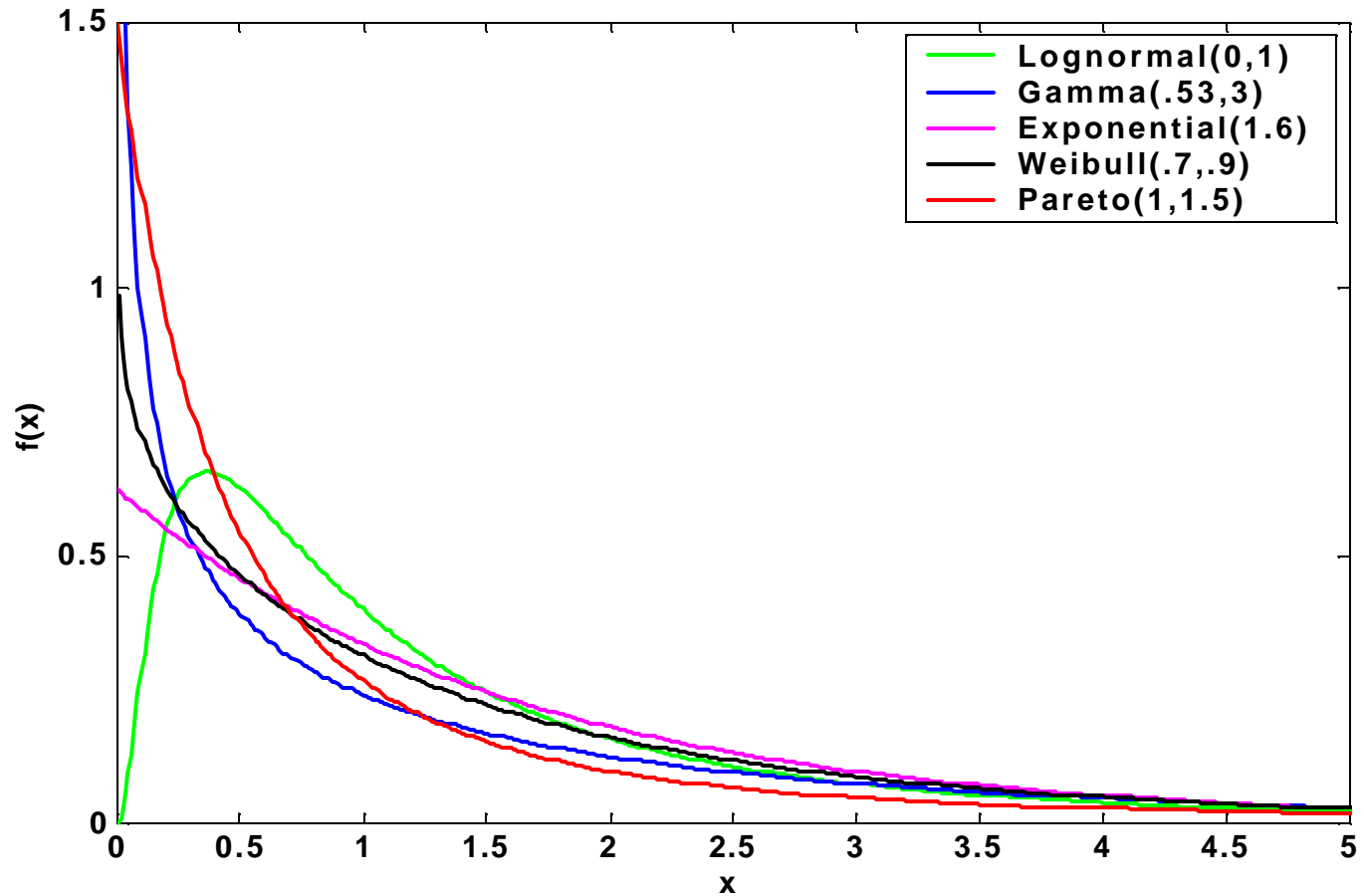
where $c > 0$ and finite

- F is also called a power law or scaling distribution
- The parameter α is called the tail index
- $1 < \alpha < 2$, F has infinite variance, but finite mean
- $0 < \alpha < 1$, the variance and mean of F are infinite
- “Mild” vs “wild” (Mandelbrot): $\alpha \geq 2$ vs $\alpha < 2$

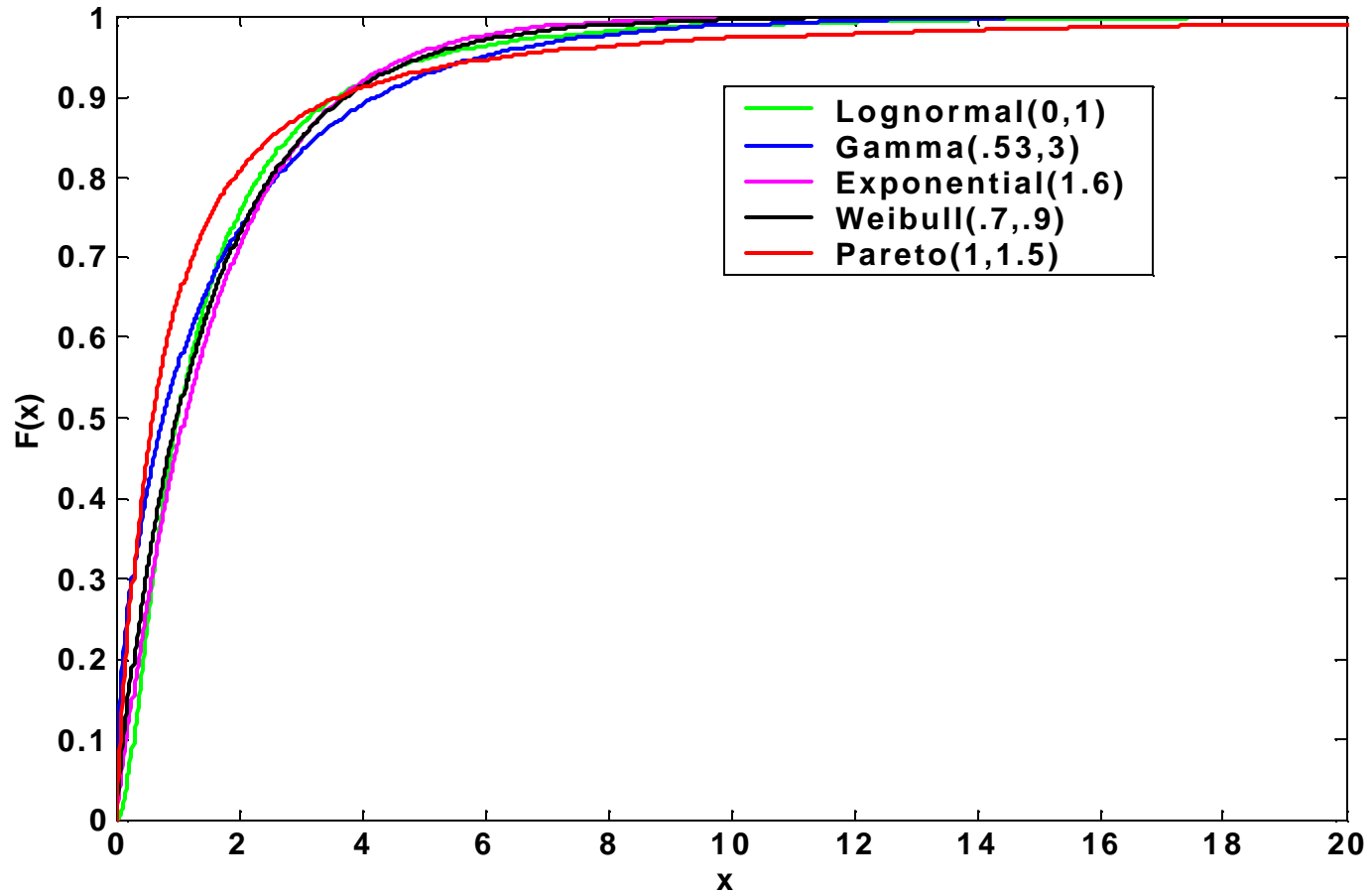
Some Illustrative Examples

- Some commonly-used plotting techniques
 - Probability density functions (pdf)
 - Cumulative distribution functions (CDF)
 - Complementary CDF (CCDF)
- Different plots emphasize different features
 - Main body of the distribution vs. tail
 - Variability vs. concentration
 - Uni- vs. multi-modal

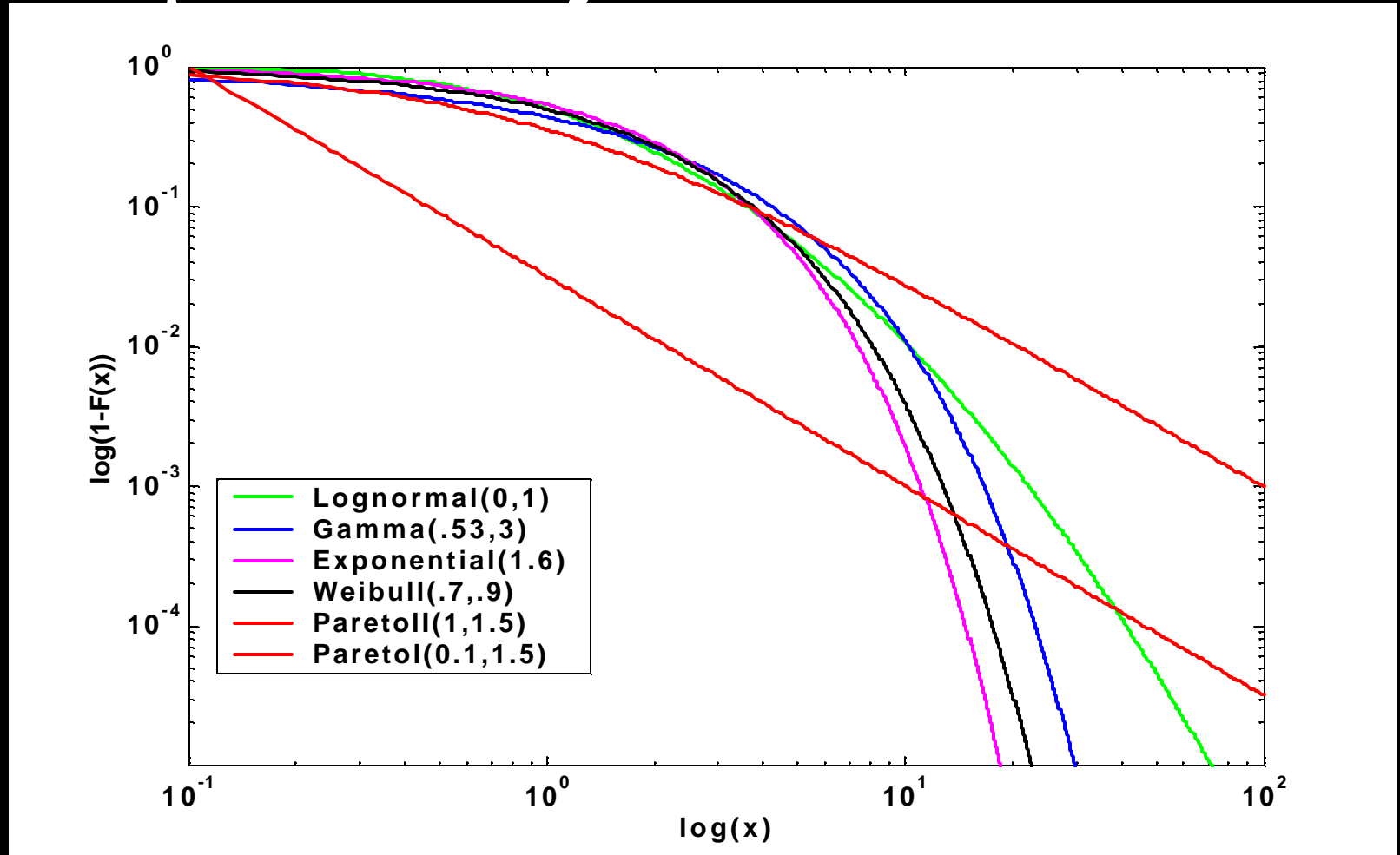
Probability density functions



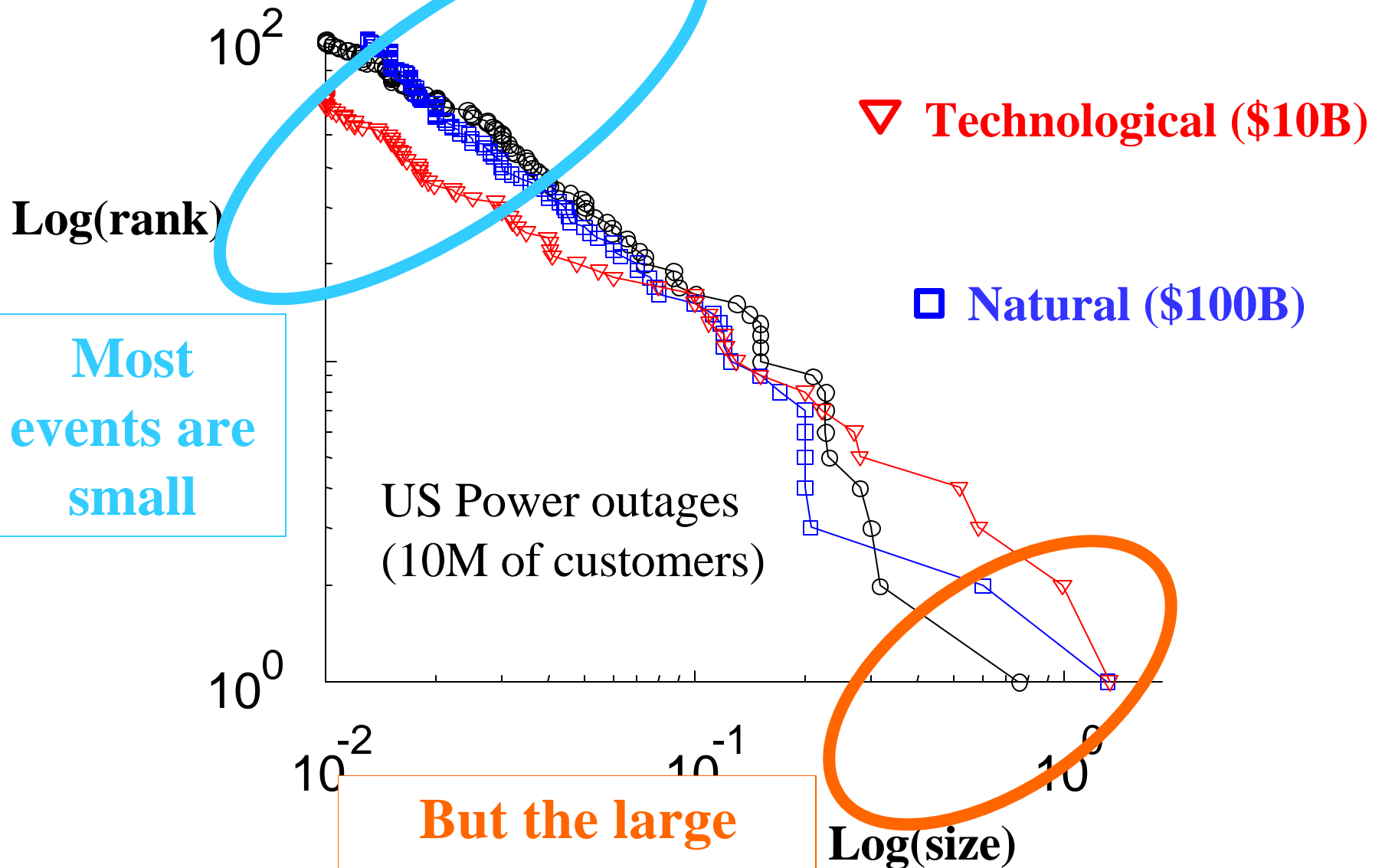
Cumulative Distribution Function



Complementary CDFs



20th Century's 100 largest disasters worldwide



Why “Heavy Tails” Matter ...

- Risk modeling (insurance)
- Load balancing (CPU, network)
- Job scheduling (Web server design)
- Combinatorial search (Restart methods)
- Complex systems studies (SOC vs. HOT)
- Towards a theory for the Internet ...

Some First Properties

- Heavy-tailed or “scaling” distribution
 - $P[X > x | X > w] = P[X > x] / P[X > w] \approx c_1 x^{-a}$
 - Compare to exponential distribution:

$$P[X > x | X > w] = \exp(-(x - w))$$

- Linearly increasing mean residual lifetime
 - $E[X - x | X > x] \approx cx$
 - Compare to exponential distribution

$$E[X - x | X > x] = \text{const}$$

Some Simple Constructions

- For U uniform in $[0,1]$, set $X=1/U$
 - X is heavy-tailed with $\alpha=1$
- For E (standard) exponential, set $X=\exp(E)$
 - X is heavy-tailed with $\alpha=1$
- The mixture of exponential distributions with parameter $1/\delta$ having a (centered) Gamma(a,b) distribution is a Pareto distribution with $\alpha=a$
- The distribution of the time between consecutive visits of a symmetric random walk to zero is heavy-tailed with $\alpha=1/2$

Key Mathematical Properties of Scaling Distributions

- Invariant under aggregation
 - Non-classical CLT and stable laws
- (Essentially) invariant under maximization
 - Domain of attraction of Frechet distribution
- (Essentially) invariant under mixture
 - Example: The largest disasters worldwide

Linear Aggregation: Classical Central Limit Theorem

- A well-known result
 - $X(1), X(2), \dots$ independent and identically distributed random variables with distribution function F (mean μ and variance 1)
 - $S(n) = X(1)+X(2)+\dots+X(n)$ n-th partial sum
 - $(S(n) - n\mu) / \sqrt{n} \rightarrow N(0,1)$, as $n \rightarrow \infty$
- More general formulations are possible
- Often-used argument for the ubiquity of the normal distribution

Linear Aggregation: Non-classical Central Limit Theorem

- A not so well-known result
 - $X(1), X(2), \dots$ independent and identically distributed with common distribution function F that is heavy-tailed with $1 < \alpha < 2$
 - $S(n) = X(1) + X(2) + \dots + X(n)$ n -th partial sum
 - $(S(n) - n\mathbf{m}) / \sqrt[n]{n} \rightarrow$ stable law, as $n \rightarrow \infty$
- The limit distribution is heavy-tailed with index α
- More general formulations are possible
- Rarely taught in most Stats/Probability courses!

Maximization:

Maximum Domain of Attraction

- A not so well-known result (extreme-value theory)
 - $X(1), X(2), \dots$ independent and identically distributed with common distribution function F that is heavy-tailed with $1 < \alpha < 2$
 - $M(n) = \max(X(1), \dots, X(n))$, n -th successive maxima
 - $M(n) / \sqrt[n]{n} \rightarrow G$, as $n \rightarrow \infty$
- G is the Fréchet distribution $\exp(-x^{-\alpha})$
- G is heavy-tailed with index α

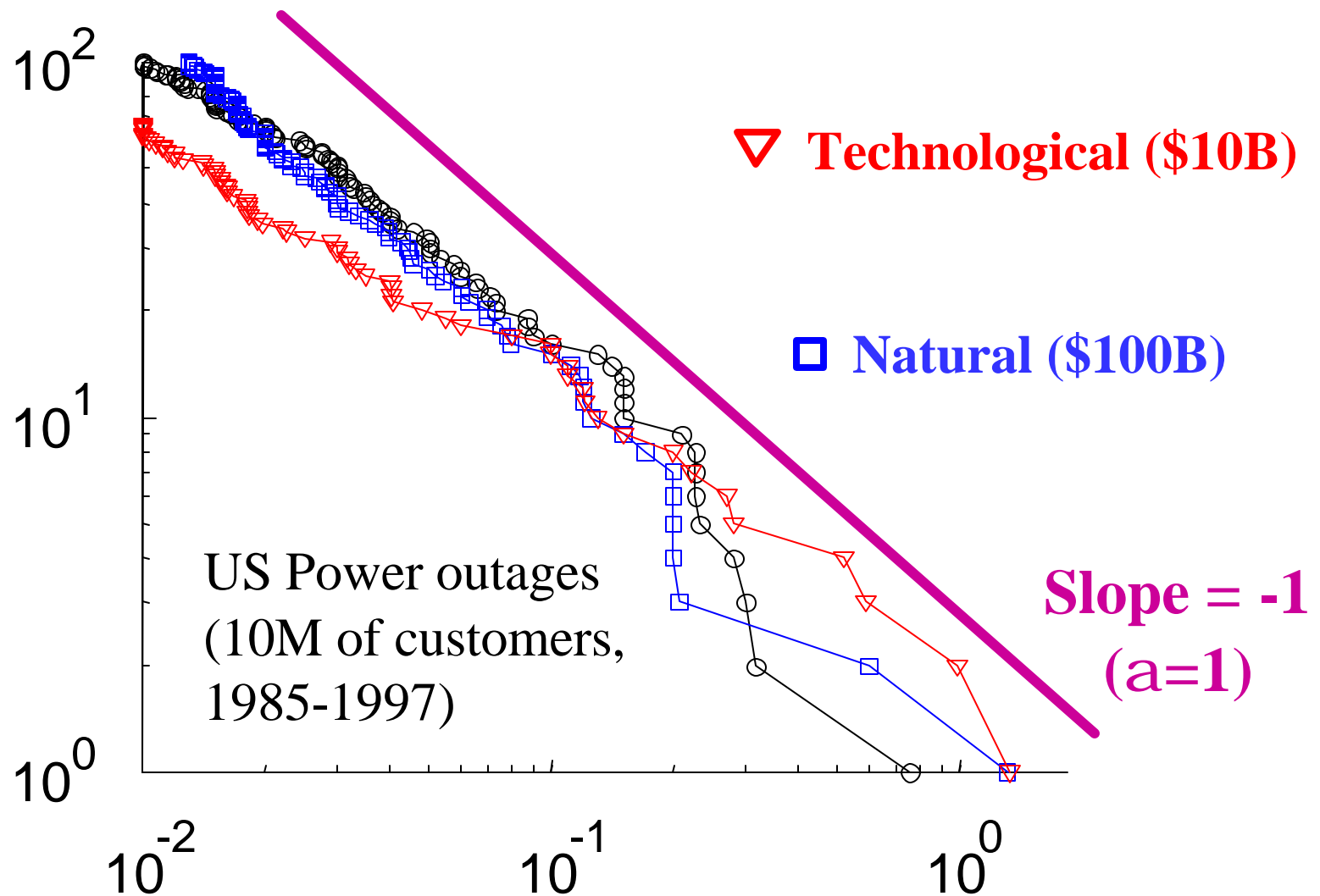
Intuition for “Mild” vs. “Wild”

- The case of “mild” distributions
 - “Evenness” – large values of $S(n)$ occur as a consequence of many of the $X(i)$'s being large
 - The contribution of each $X(i)$, even of the largest, is negligible compared to the sum
- The case of “wild” distributions
 - “Concentration” – large values of $S(n)$ or $M(n)$ occur as a consequence of a single large $X(i)$
 - The largest $X(i)$ is dominant compared to $S(n)$

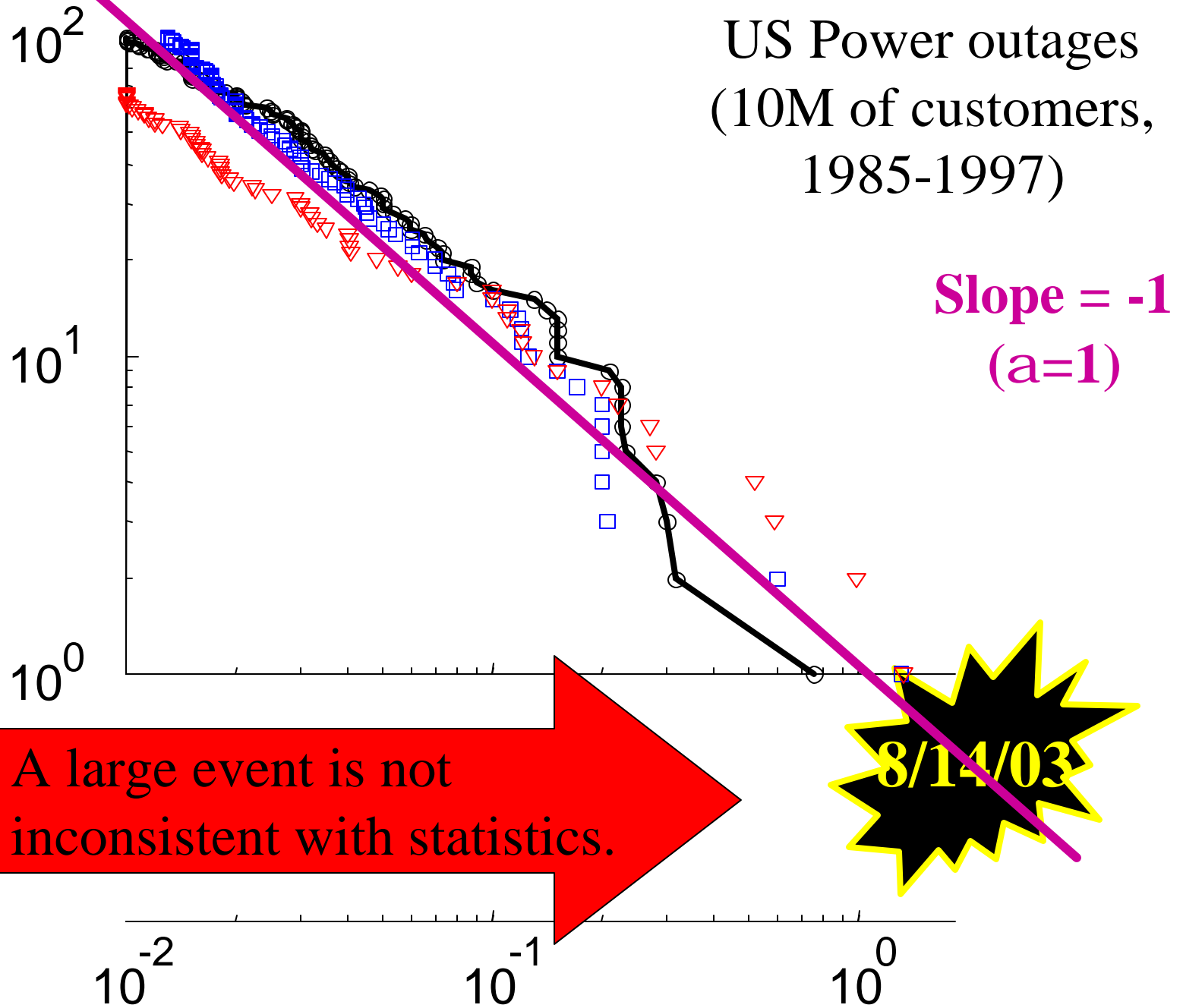
Weighted Mixture

- A little known result
 - $X(1), X(2), \dots$ independent and identically distributed with common distribution function F that is heavy-tailed with $1 < \alpha < 2$
 - $p(1), p(2), \dots, p(n)$ iid in $[0, 1]$ with $p(1) + \dots + p(n) = 1$
 - $W(n) = p(1)X(1) + \dots + p(n)X(n)$
 - $P[W(n) > x] \approx c_W x^{-a}$, for large x
- Invariant “distributions” are $F_W(u) = 1 - cu^{-a}$
- Condition on $X(i) > a > 0$

20th Century's 100 largest disasters worldwide



US Power outages
(10M of customers,
1985-1997)



A large event is not
inconsistent with statistics.

8/14/03

Resilience to Ambiguity

- Scaling distributions are robust under
 - ... aggregation, maximization, and mixture
 - ... differences in observing/reporting/accounting
 - ... varying environments, time periods
- The “value” of robustness
 - Discoveries are easier/faster
 - Properties can be established more accurately
 - Findings are not sensitive to the details of the data gathering process

On the Ubiquity of Heavy Tails

- Heavy-tailed distributions are attractors for averaging (e.g., non-classical CLT), but are the only distributions that are also (essentially) invariant under maximizing and mixing.
- Gaussians (“normal”) distributions are also attractors for averaging (e.g., classical CLT), but are not invariant under maximizing and mixing
- This makes heavy tails more ubiquitous than Gaussians, so no “special” explanations should be required ...

Heavy Tails and Statistics

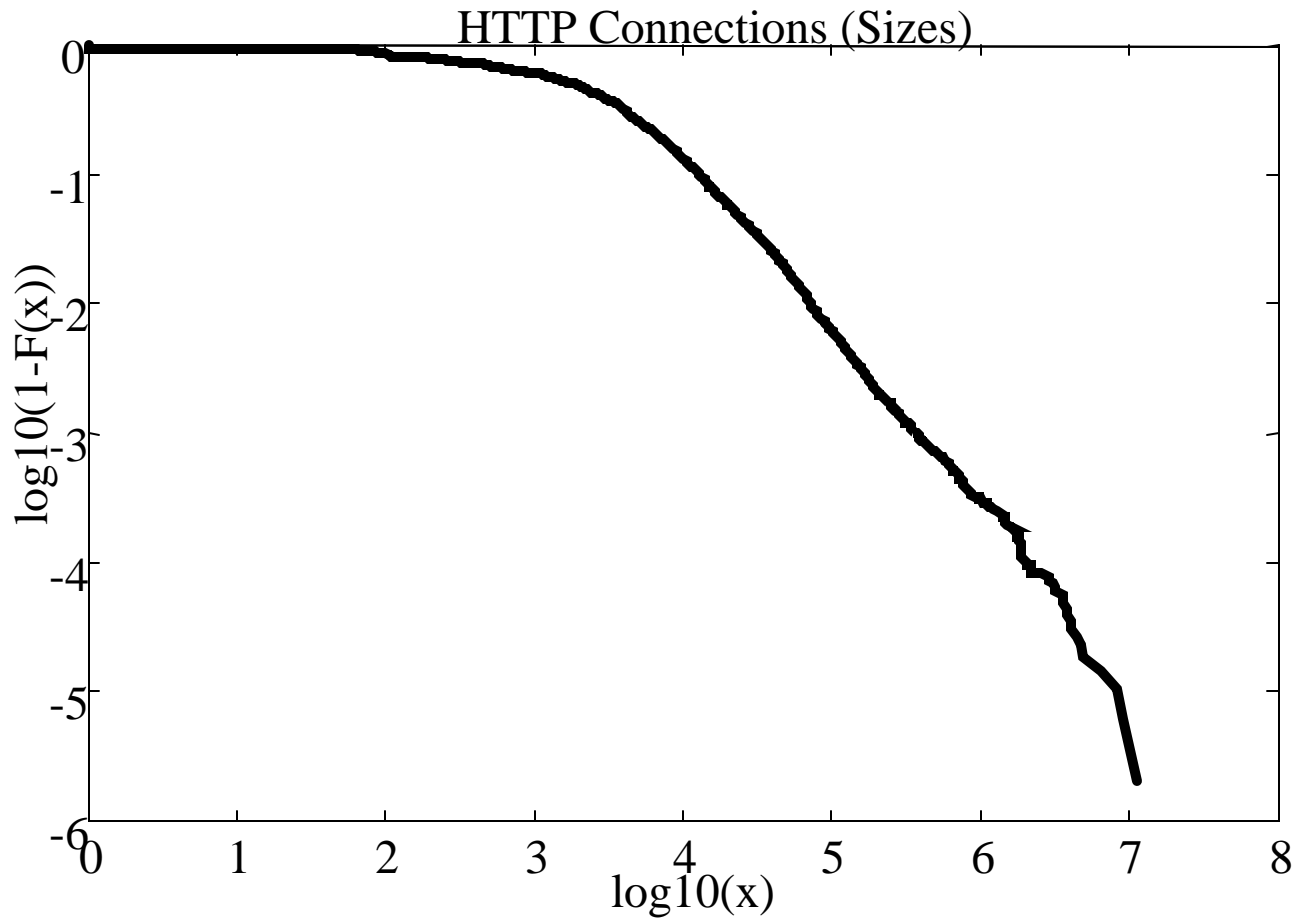
- The traditional “curve-fitting” approach
- “Curve-fitting” by example
- Beyond “curve-fitting” – “Borrowing strength”
- “Borrowing strength” by example
- What “science” in “scientific modeling”?
- Additional considerations

“Curve-fitting” approach

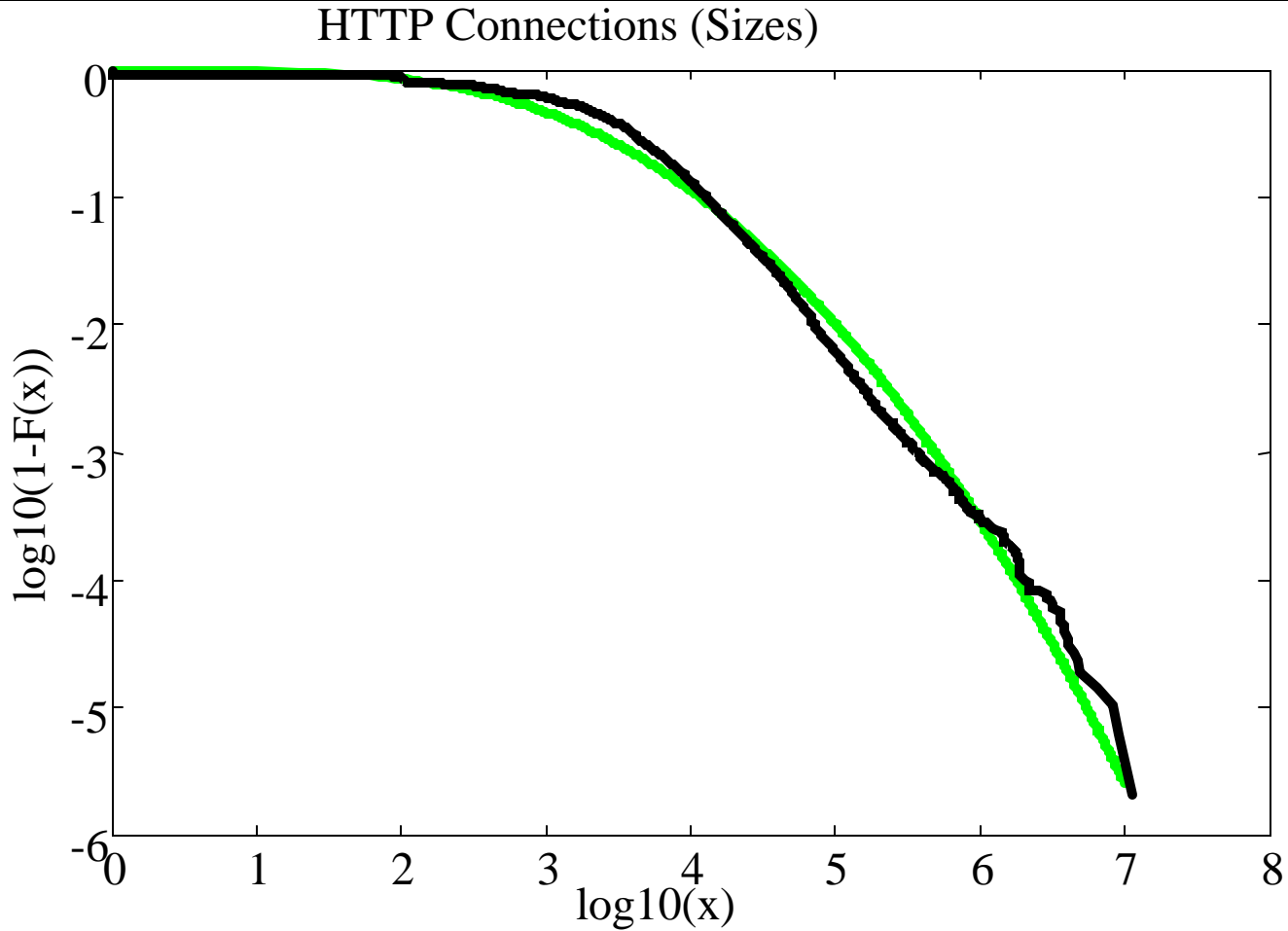
- Model selection
 - Choose parametric family of distributions
- Parameter estimation
 - Take a strictly static view of data
 - Assume moment estimates exist/converge
- Model validation
 - Select “best-fitting” model
 - Rely on some “goodness-of-fit” criteria/metrics
- “Black box-type” modeling, “data-fitting” exercise

“Curve-fitting” by example

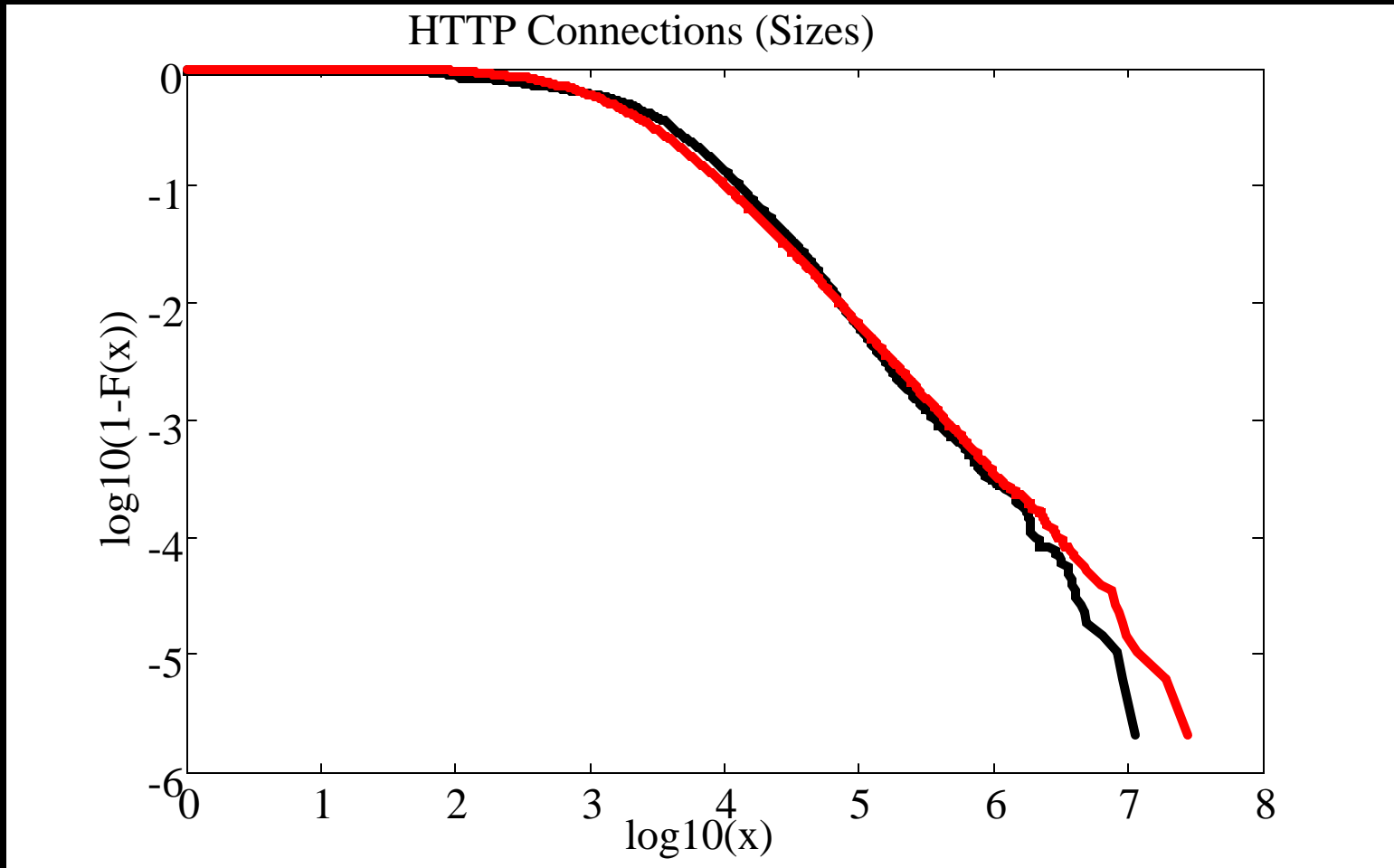
- Randomly picked data set
 - LBL’s WAN traffic (in- and outbound)
 - 1:30, June 24 – 1:30, June 25 (PDT), 1996
 - 243,092 HTTP connection sizes (bytes)
 - Courtesy of Vern Paxson (thanks!)
- Illustration of widely-accepted approach



CCDF plot on log-log scale



Fitted 2-parameter Lognormal
distribution ($\mu=6.75, \sigma=2.05$)



Fitted 2-parameter Pareto distribution
($\alpha=1.27, m=2000$)

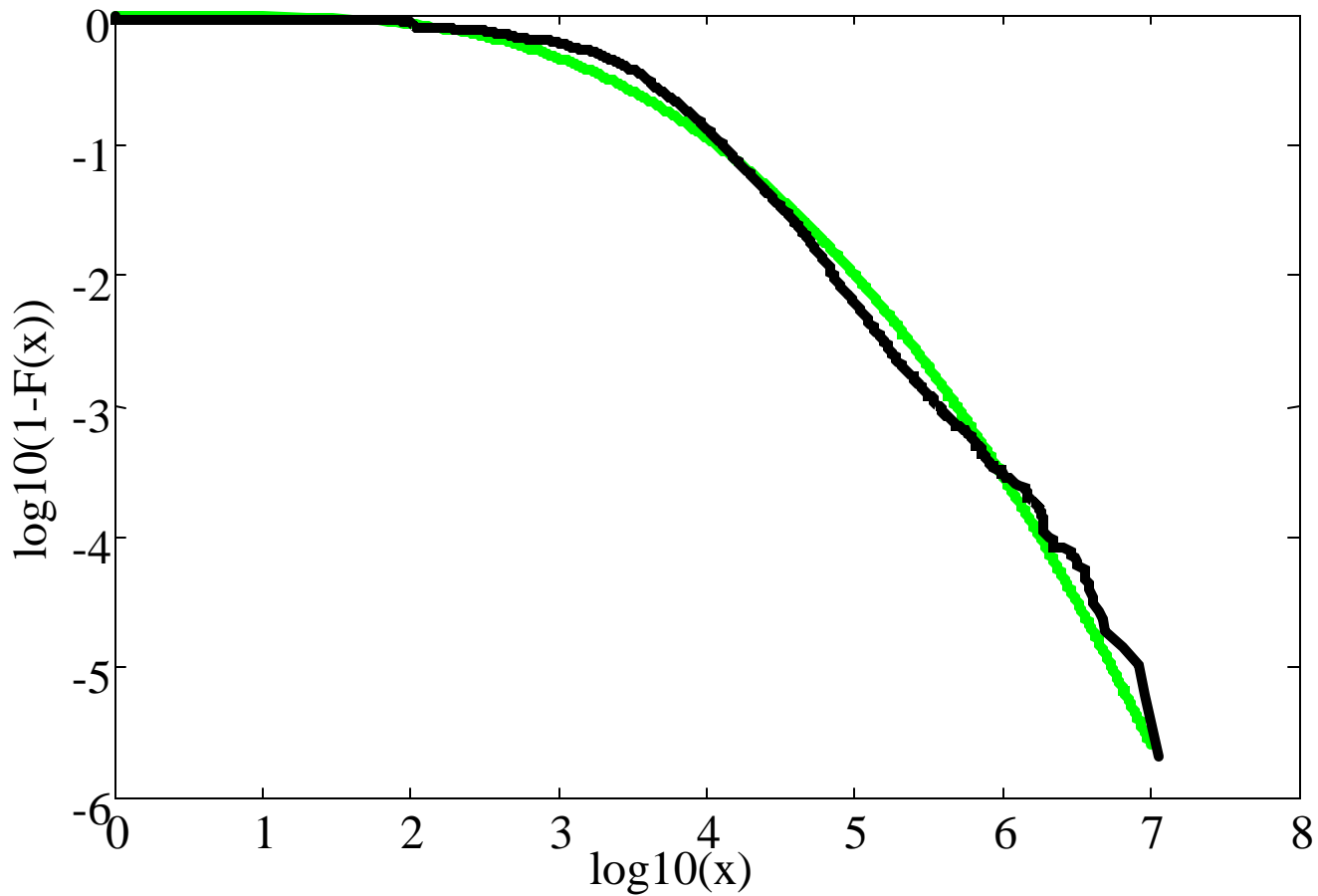
The “truth” about “curve-fitting”

- Highly predictable outcome
 - Always doable, no surprises
 - Cause for endless discussions (“Which model is better?”)
- When “more” means “better” ...
 - 2-parameter distributions (Pareto, Lognormal, ...)
 - 3-parameter distributions (Weibull, Gamma, ...)
 - 5-parameter distribution (Double-Pareto, -Lognormal, ...), etc.
- Inadequate “goodness-of-fit” criteria due to
 - Voluminous data sets
 - Data with strong dependencies (LRD)
 - Data with high variability (heavy tails)
 - » Data with non-stationary features

“Borrowing strength” approach

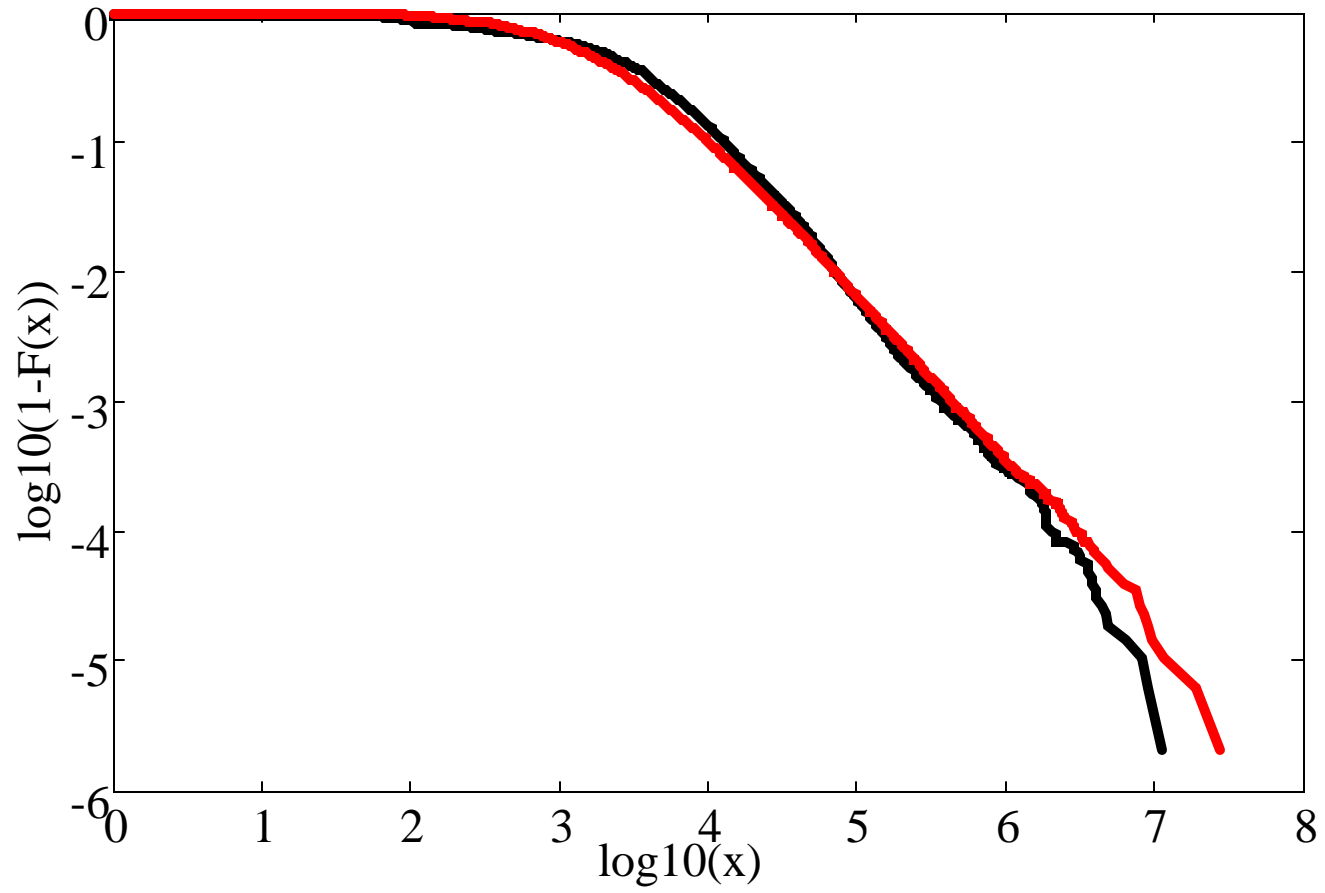
- Mandelbrot & Tukey to the rescue
 - Sequential moment plots (Mandelbrot)
 - Borrowing strength from large data (Tukey)
- “Borrowing strength” – dynamic view of data
 - Rely on traditional approach for initial (small) subset of available data
 - Consider successively larger subsets
 - Look out for inherently consistent models
 - Identify “patchwork “ of “fixes”

HTTP Connections (Sizes)



Lognormal?

HTTP Connections (Sizes)

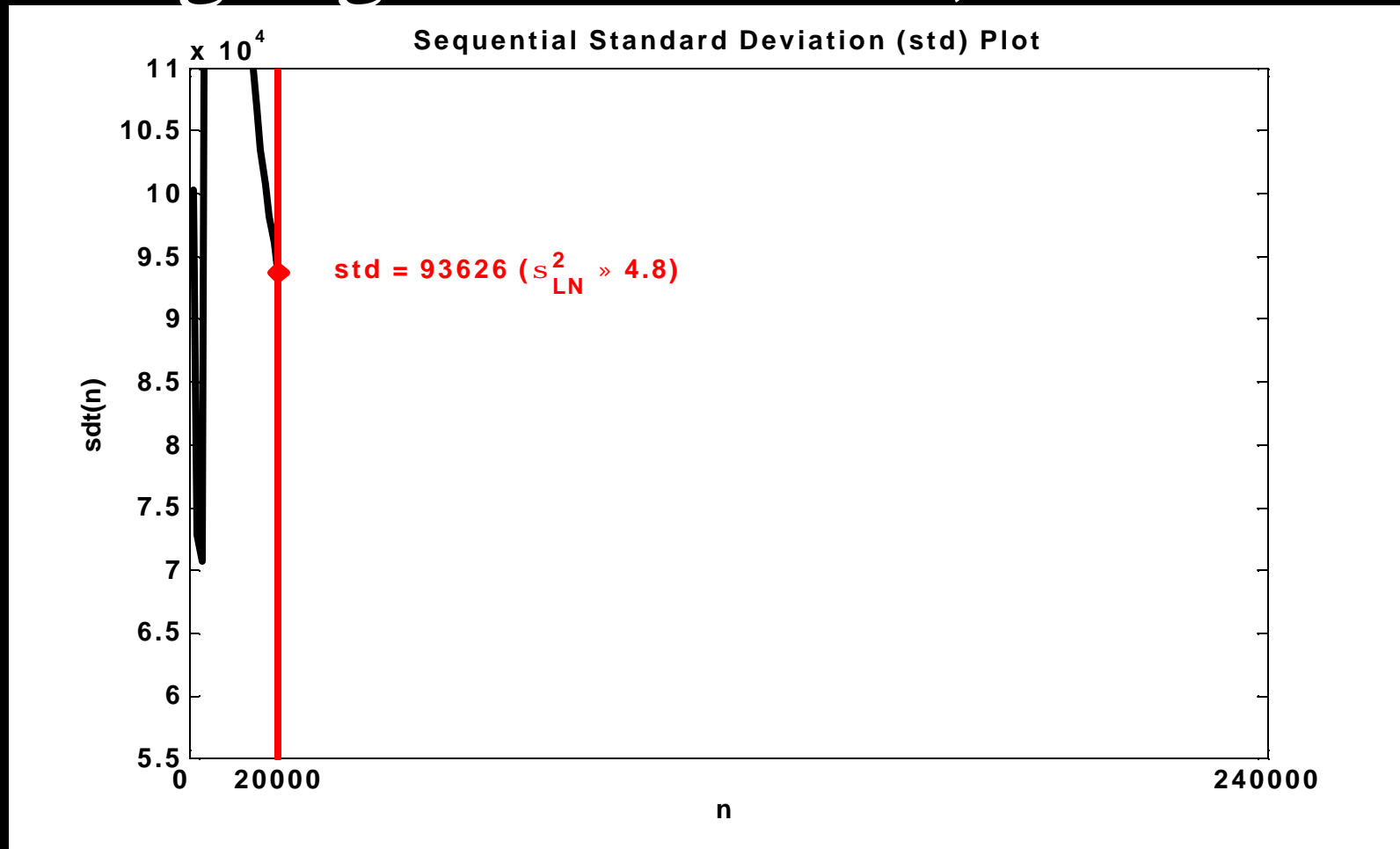


Pareto?

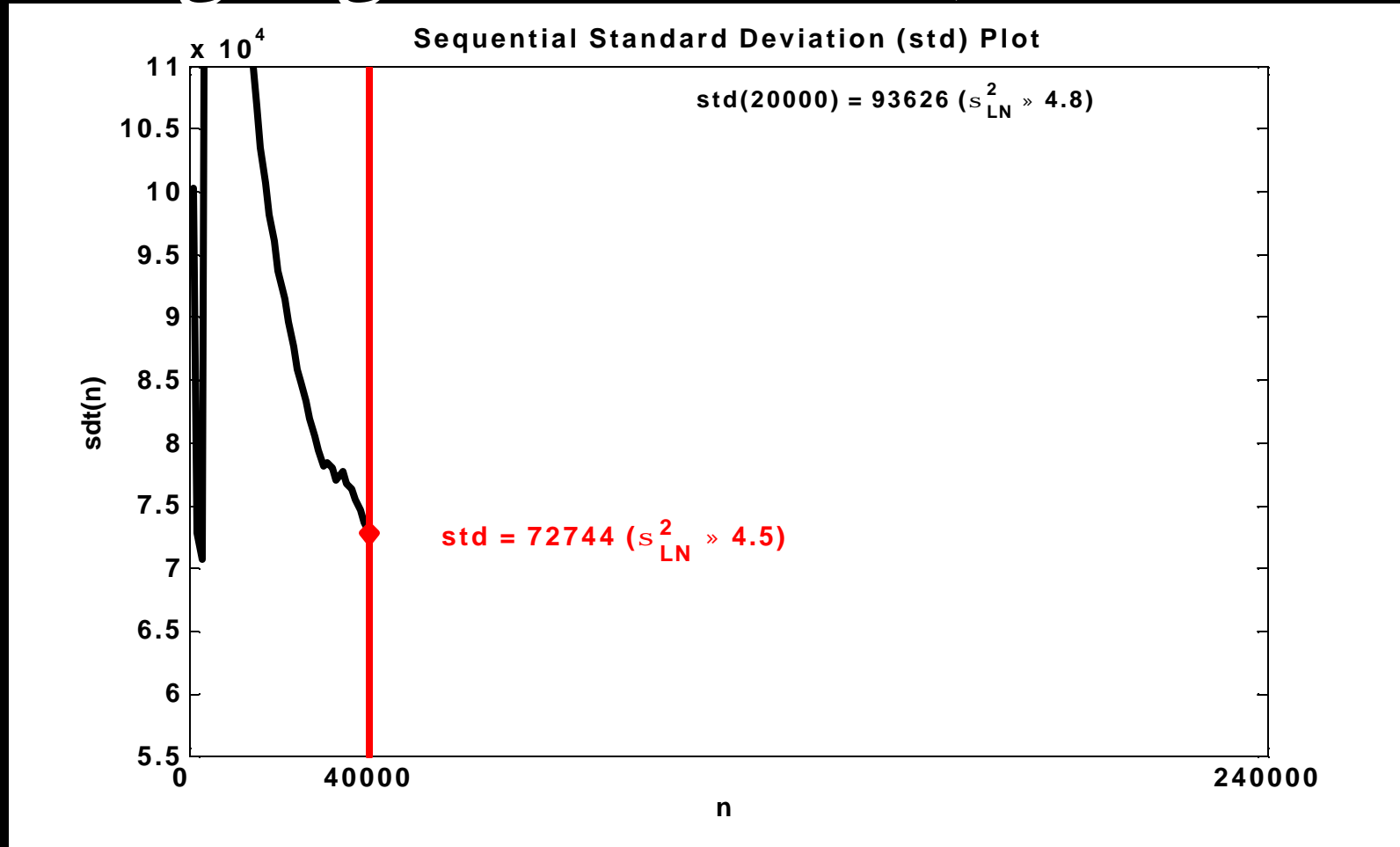
“Borrowing strength” (example 1)

- Use same data set as before
- Illustration of Mandelbrot-Tukey approach (1)
 - Sequential standard deviation plots
 - Lack of robustness
 - A case against Lognormal distributions
- More on sequential standard deviation plots
- Scaling distributions to the rescue

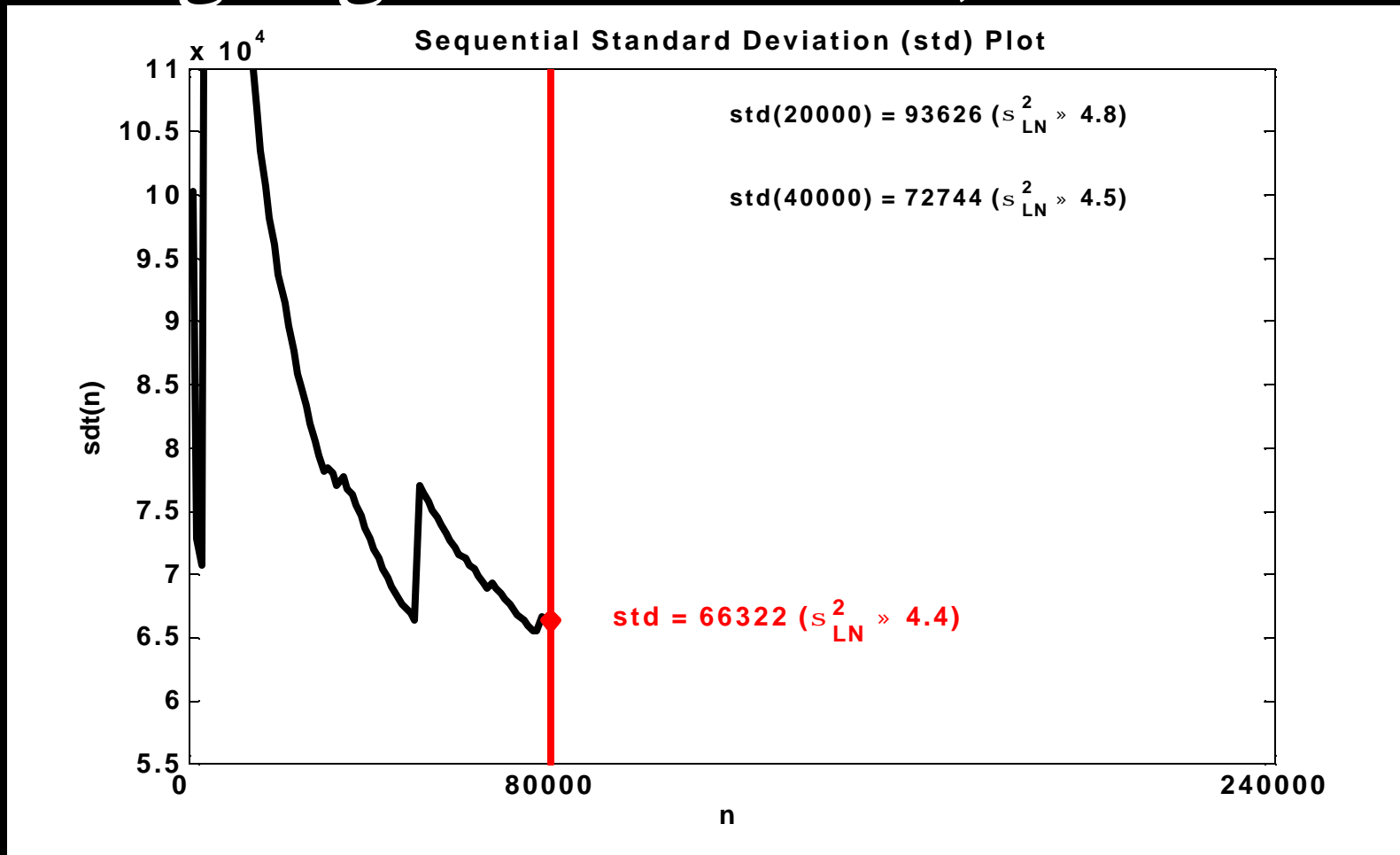
Fitting lognormal: $n=20,000$



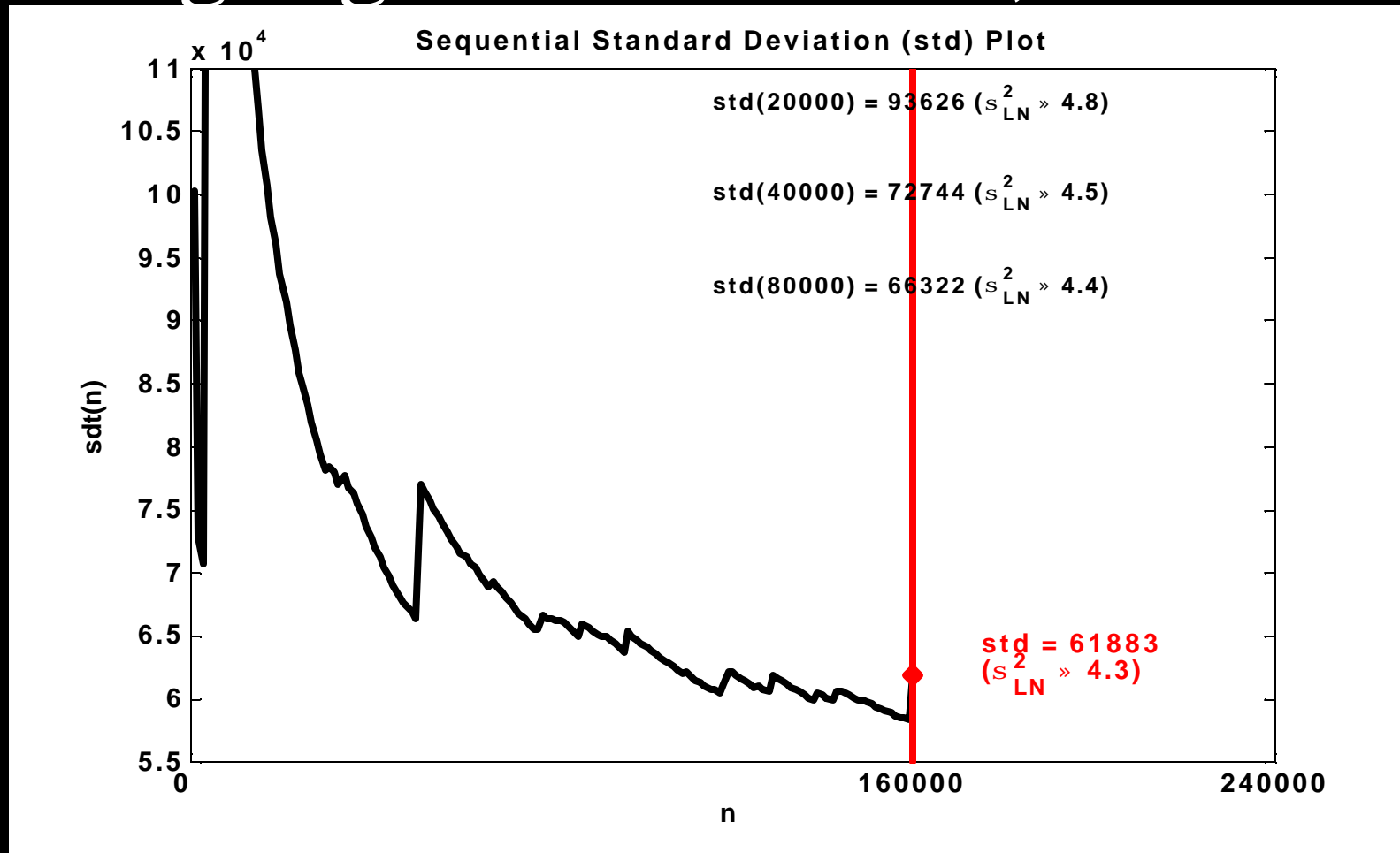
Fitting lognormal: n=40,000



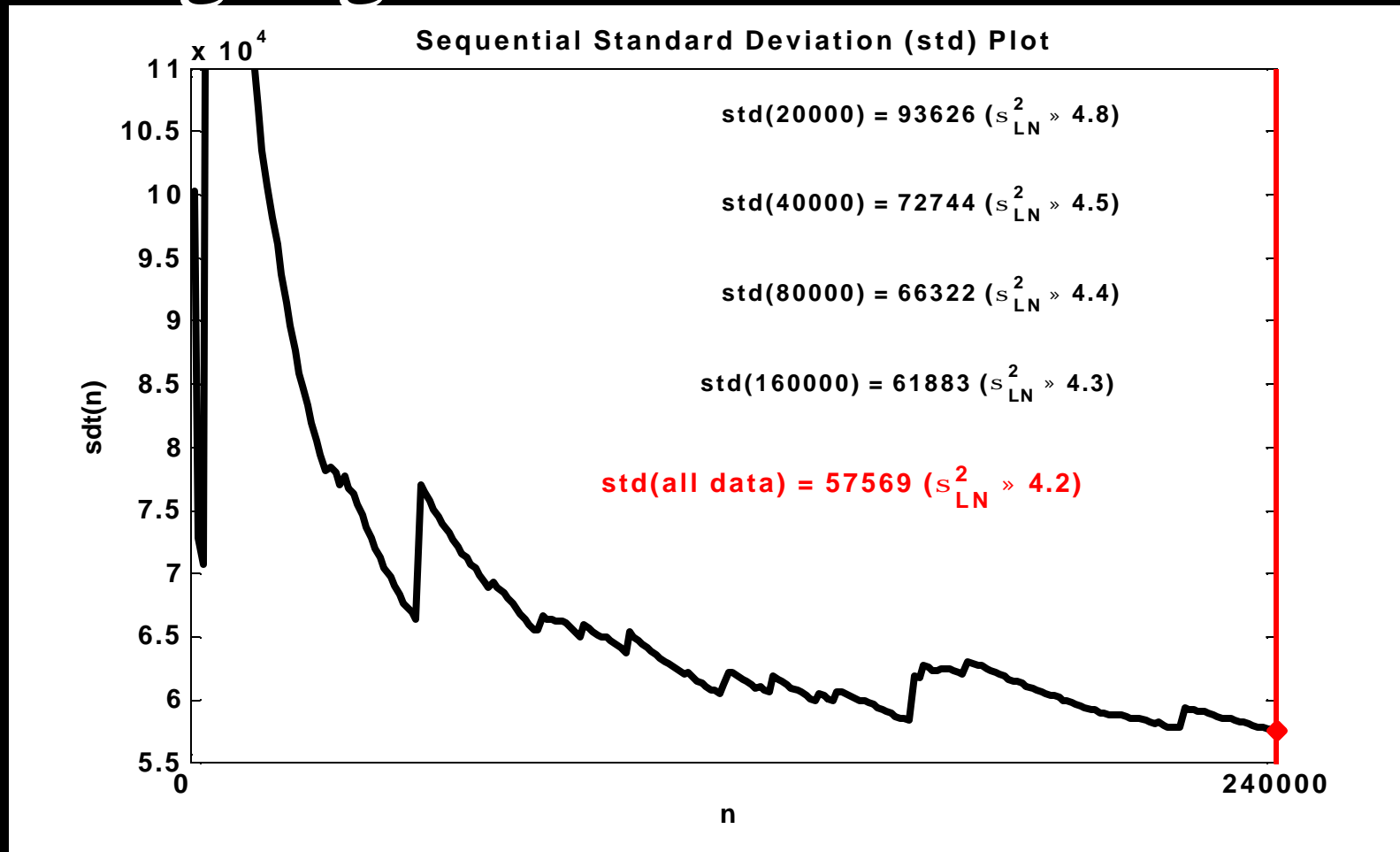
Fitting lognormal: n=80,000



Fitting lognormal: n=160,000



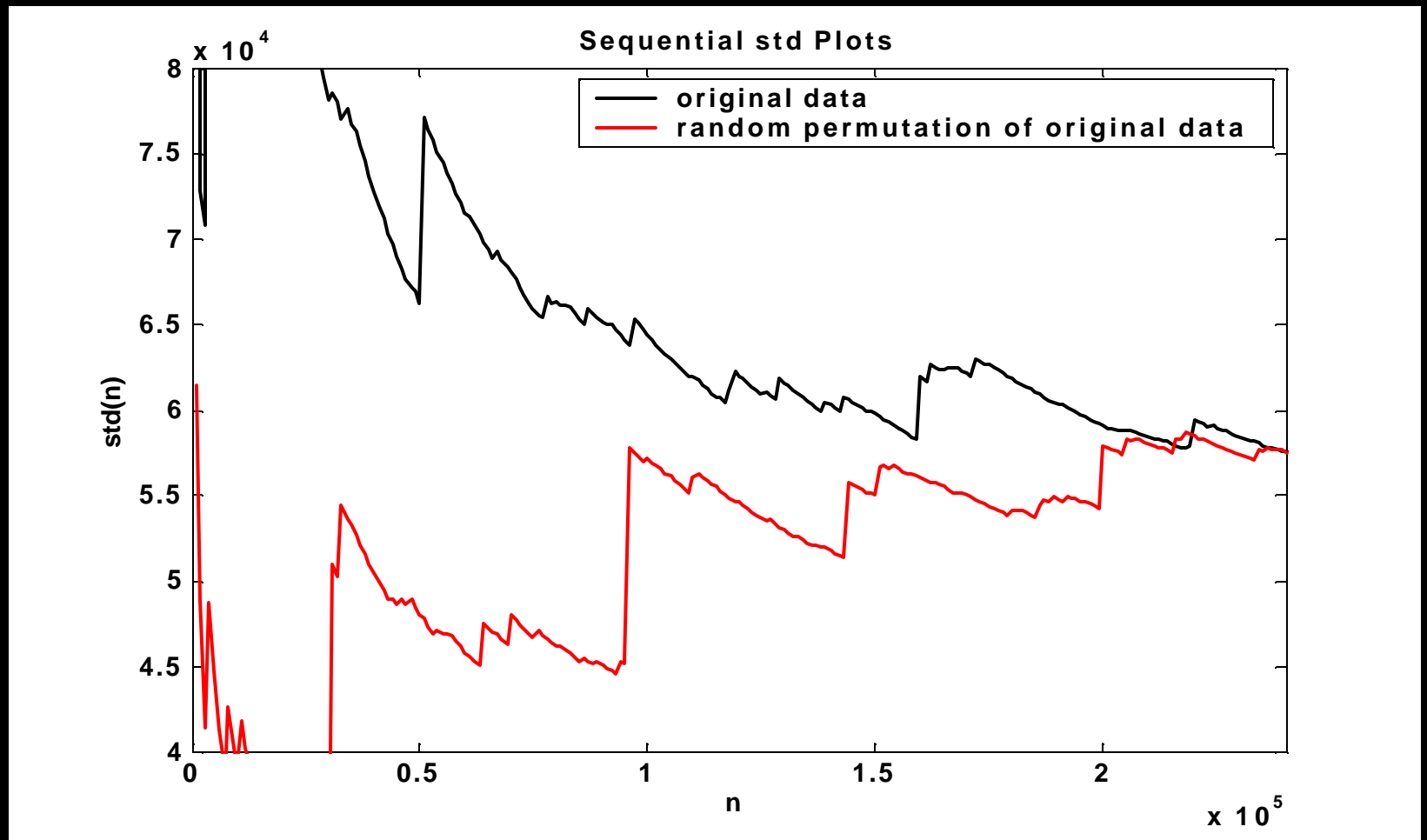
Fitting lognormal: All data



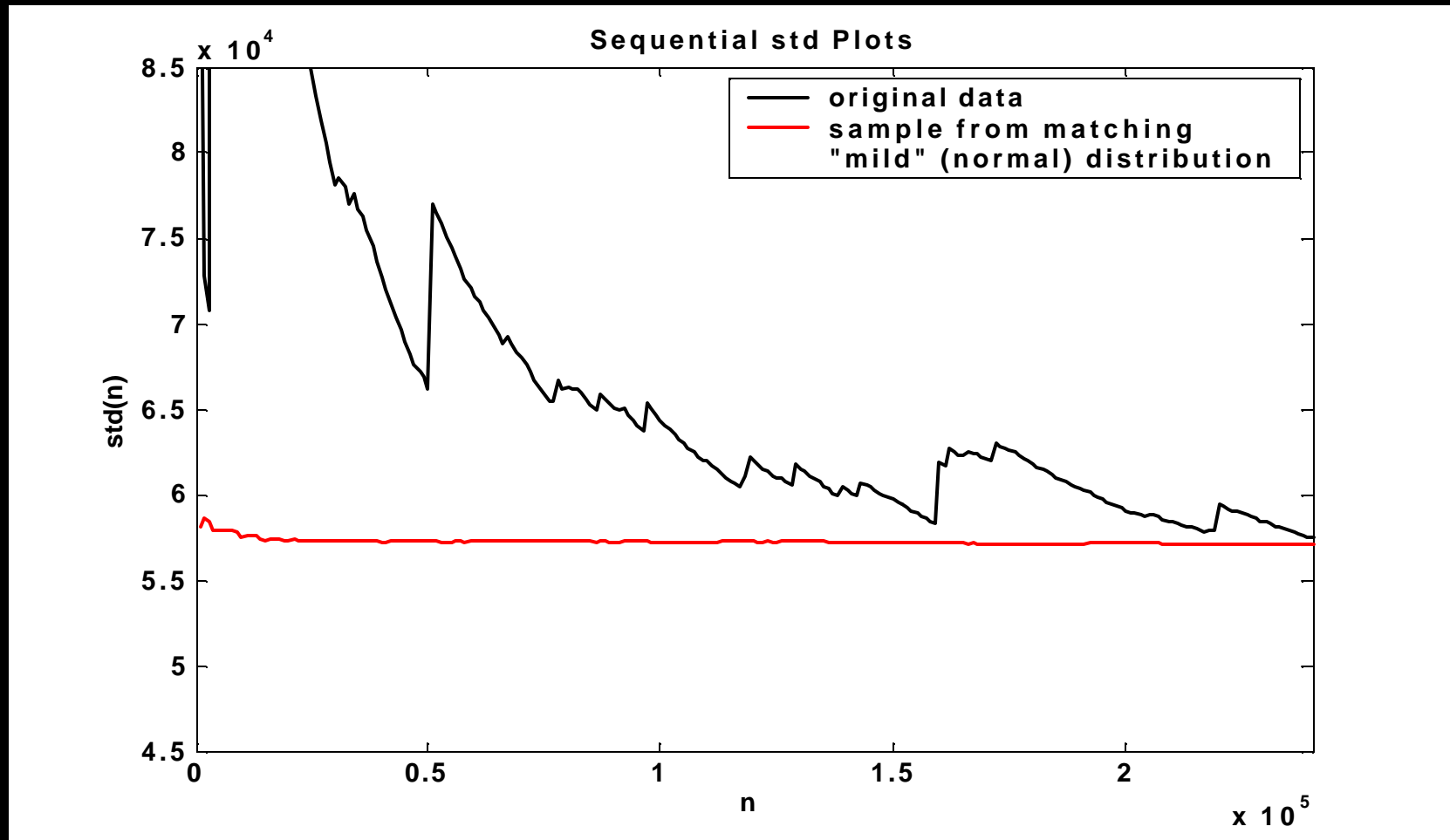
The case against lognormal

- The lognormal model assumes
 - existence/convergence of 2nd moment
 - parameter estimates are inherently consistent
- However, sequential std plot indicates
 - non-existence/divergence of 2nd moment
 - inherently inconsistent parameter estimates
- What “science” in “scientific modeling”?
 - Curve/data fitting is not “science”
 - “Patchwork” of “fixes” (Mandelbrot)

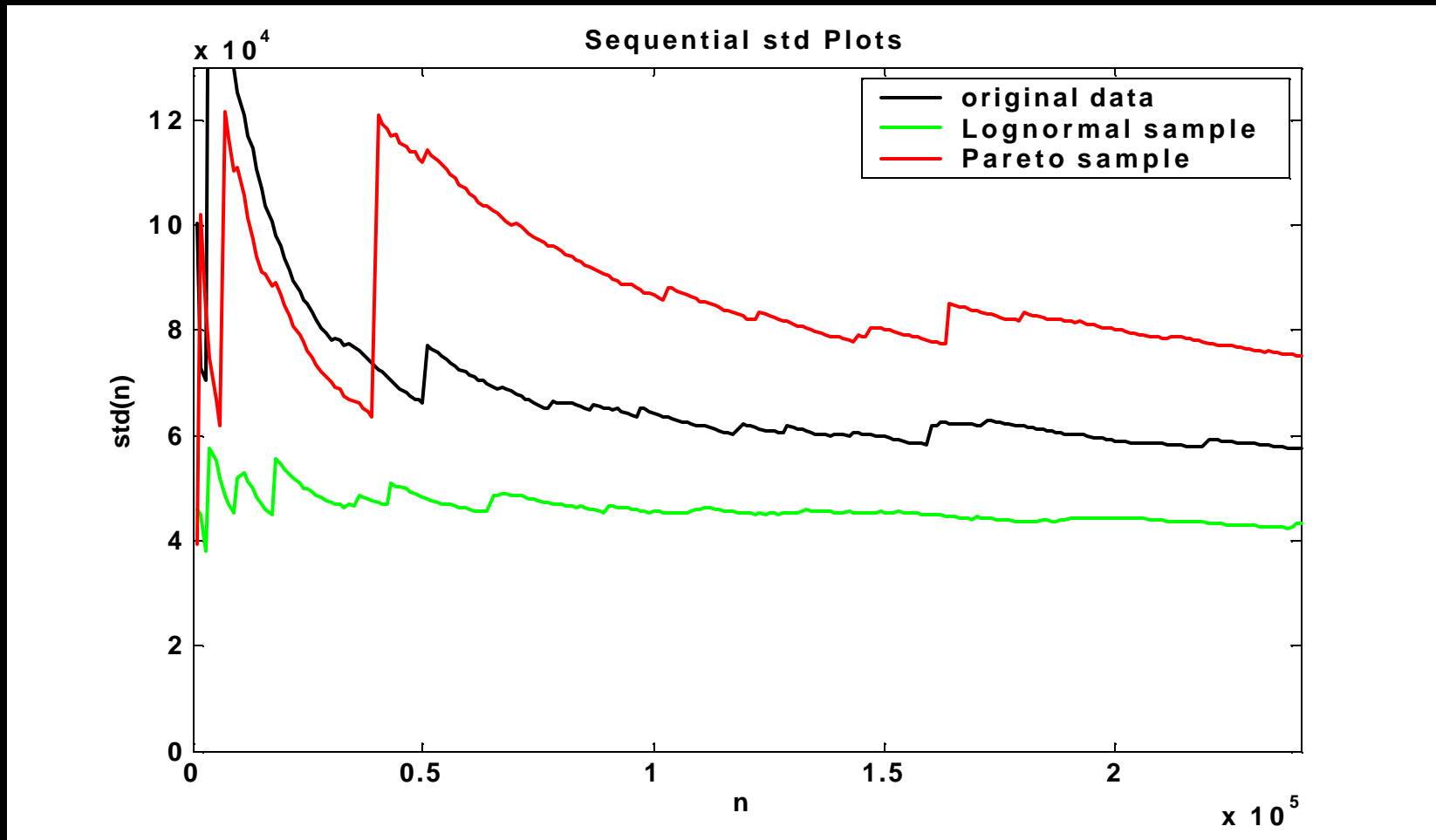
Randomizing observations



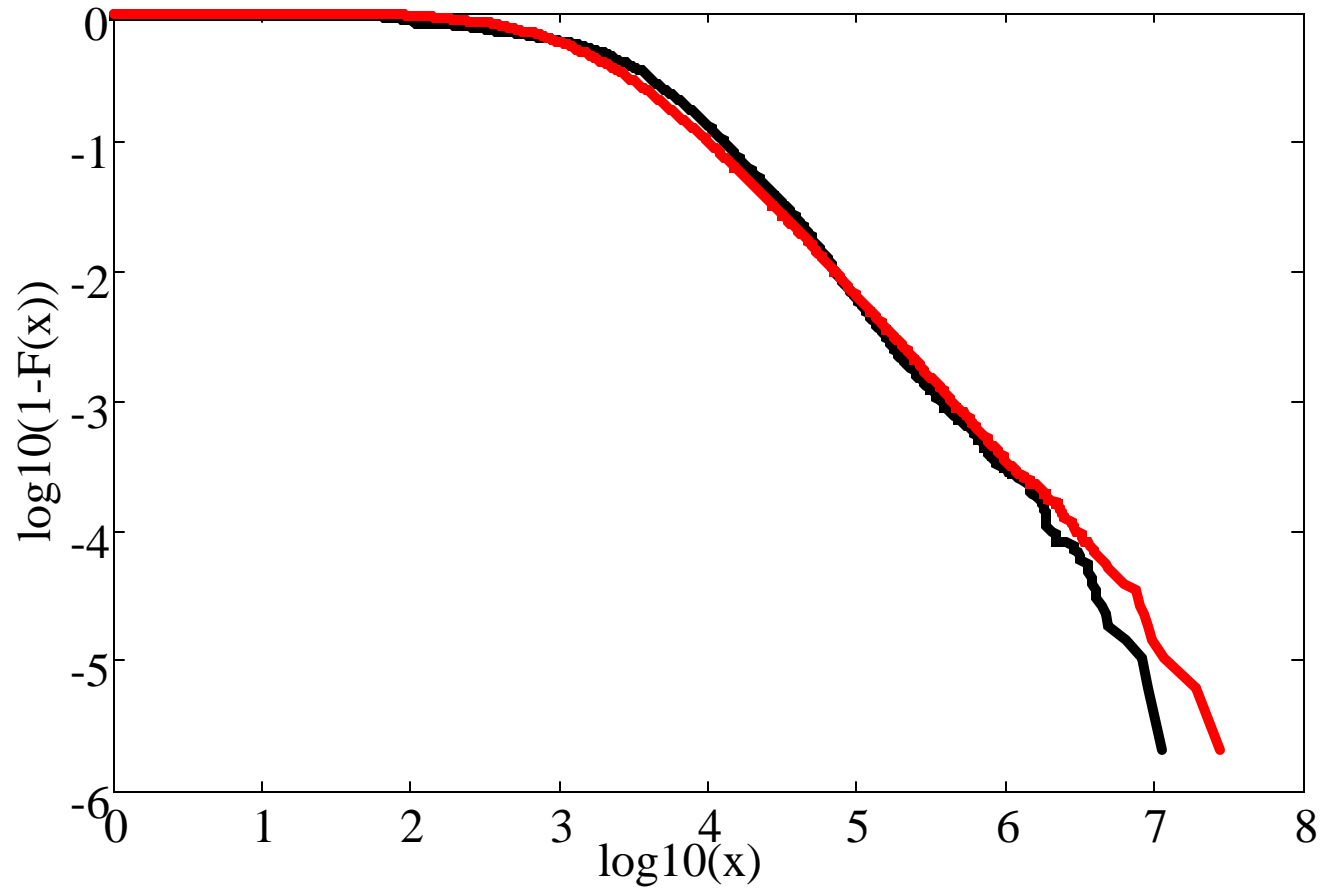
Matching “mild” distributions



Lognormal or scaling distribution



HTTP Connections (Sizes)

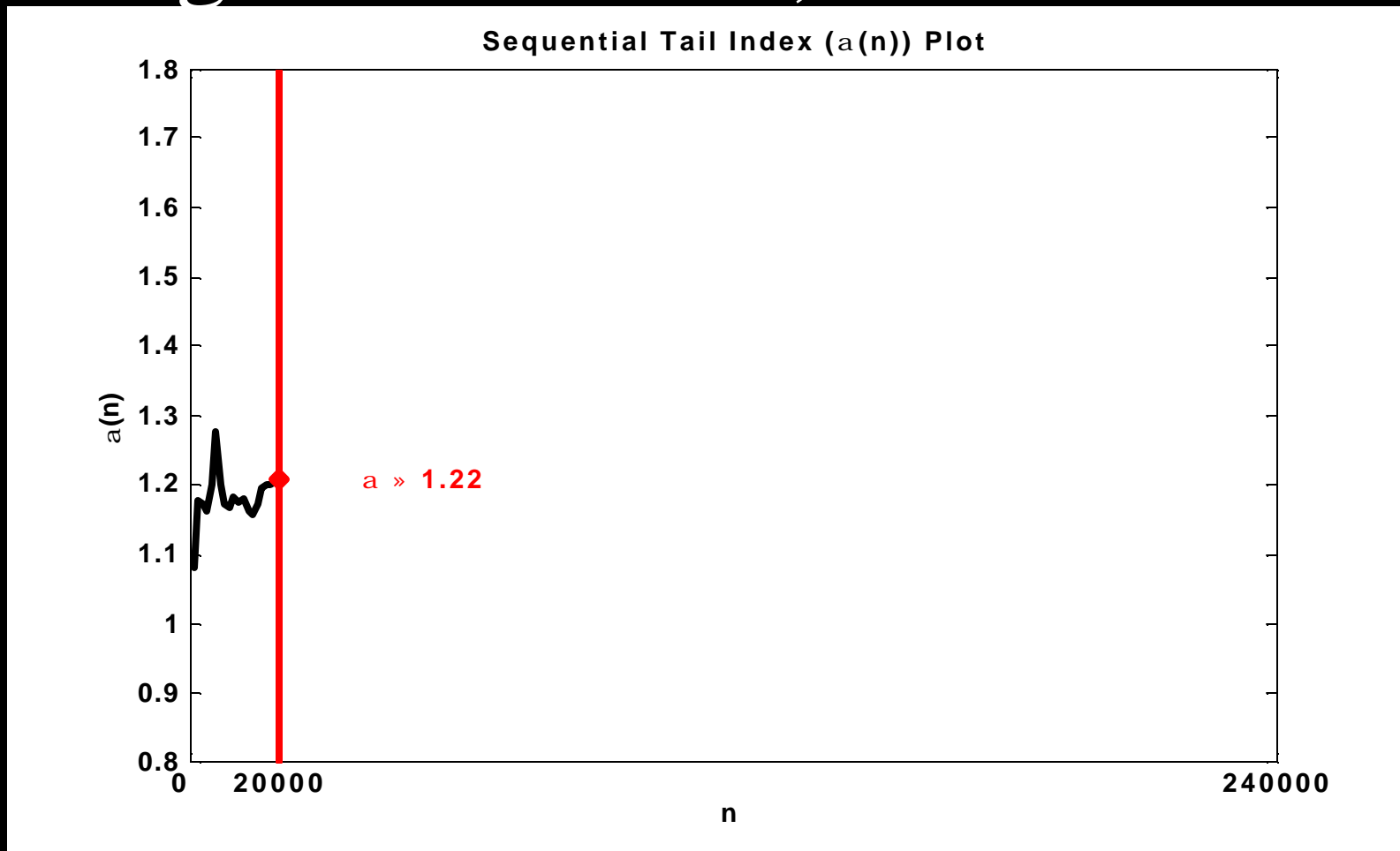


Pareto?

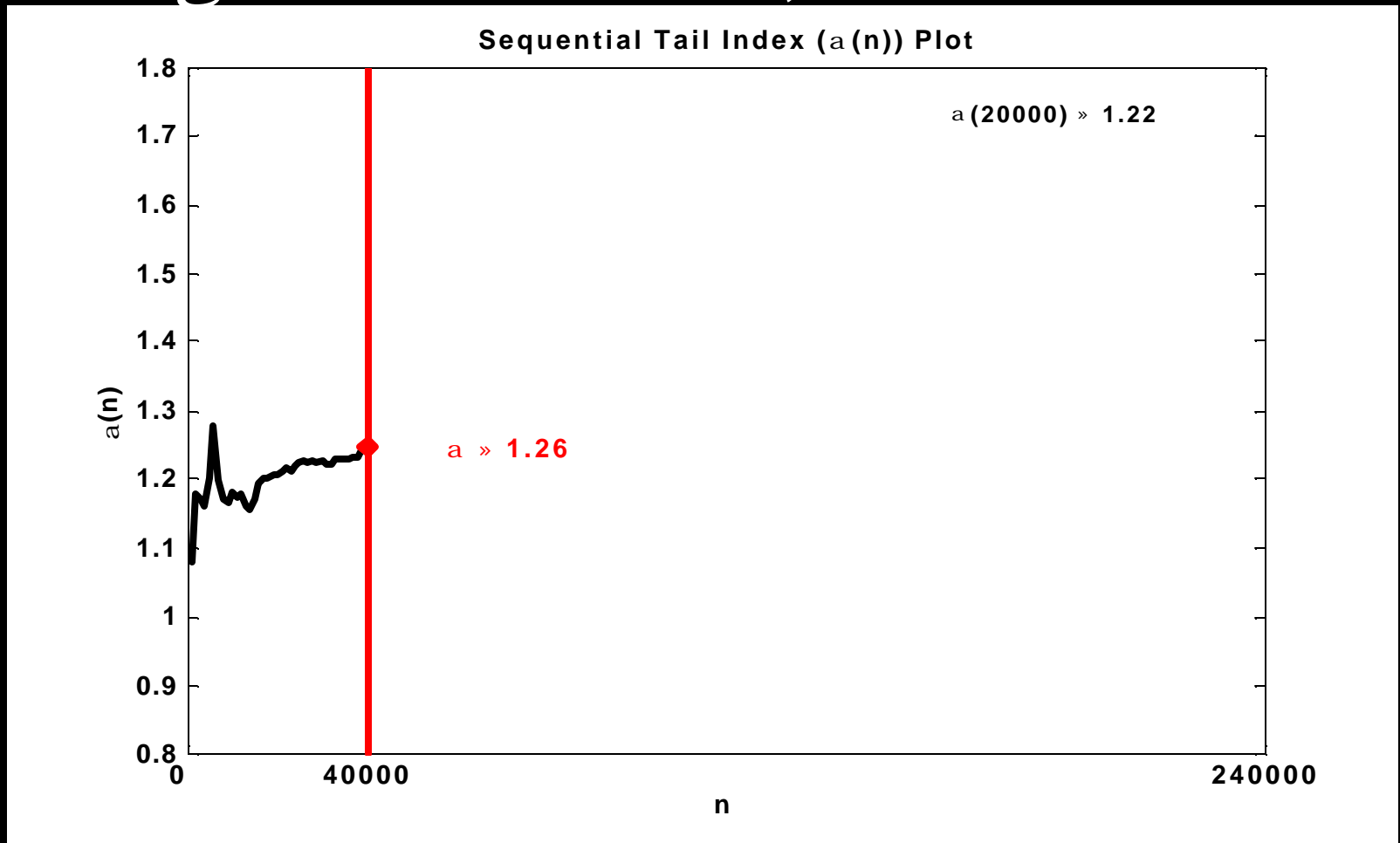
“Borrowing strength” (example 2)

- Use same data set as before
- Illustration of Mandelbrot-Tukey approach (2)
 - Sequential tail index plots
 - Strong robustness properties
 - A case for scaling distributions
- A requirement for future empirical studies

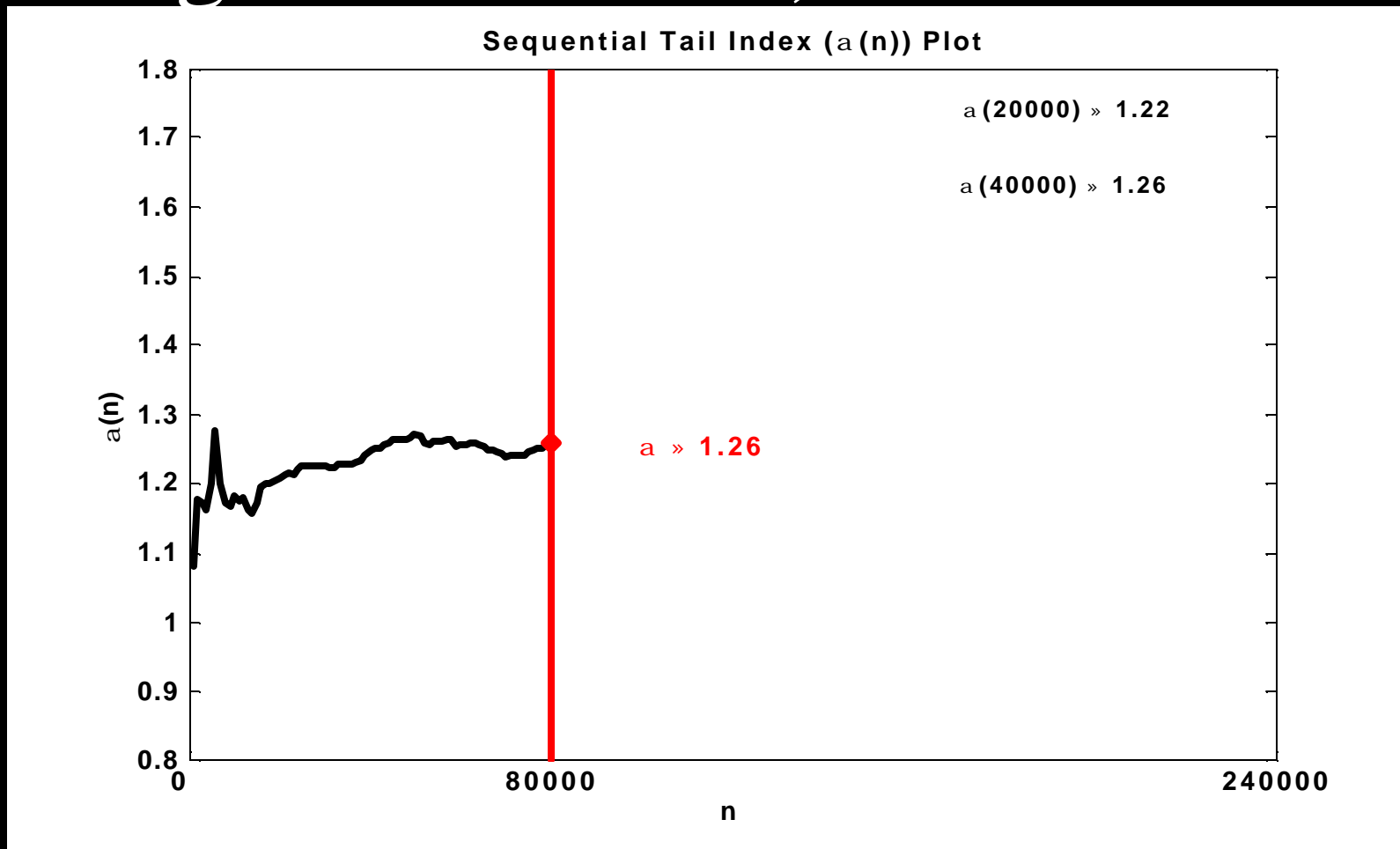
Fitting Pareto: $n=20,000$



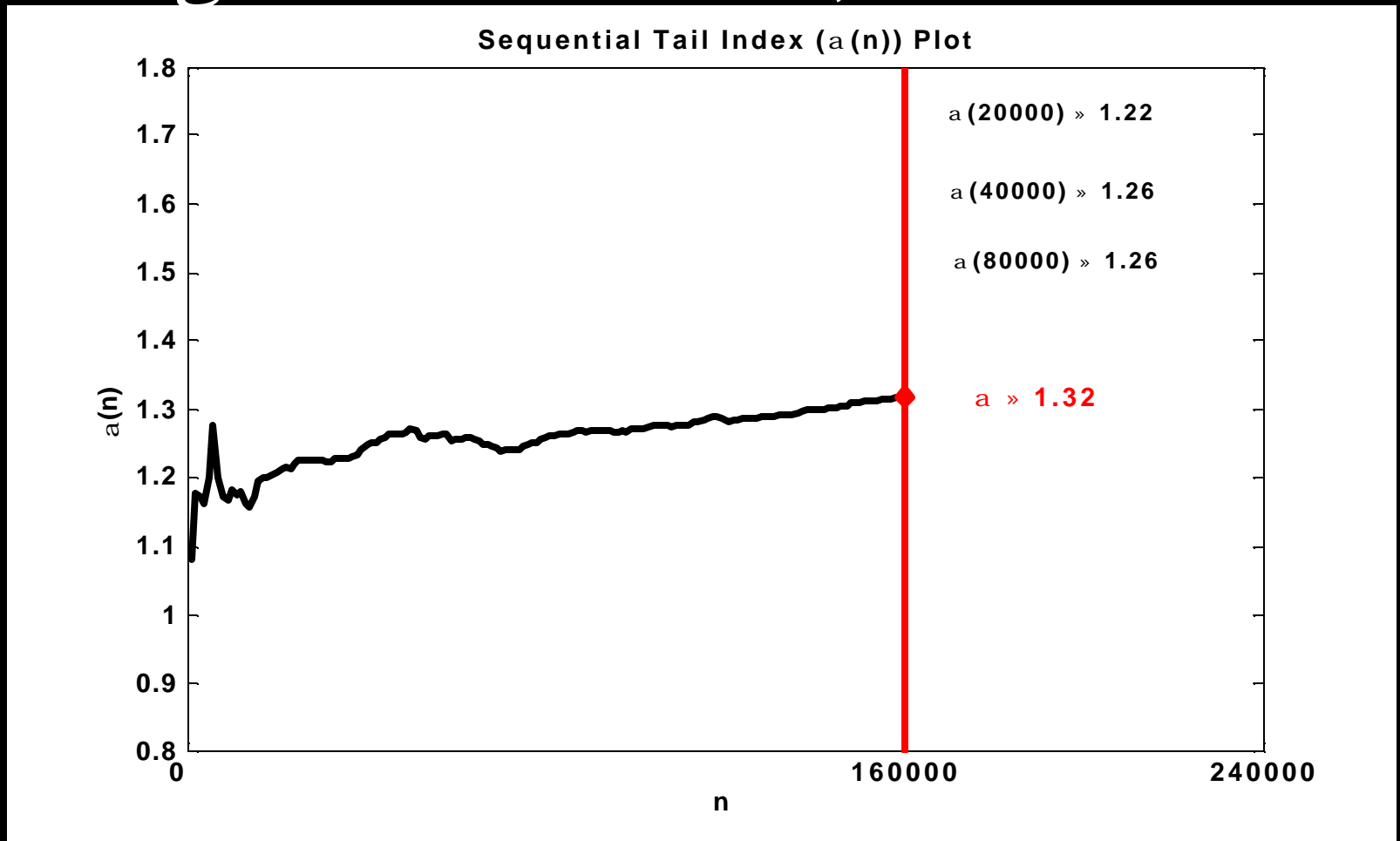
Fitting Pareto: $n=40,000$



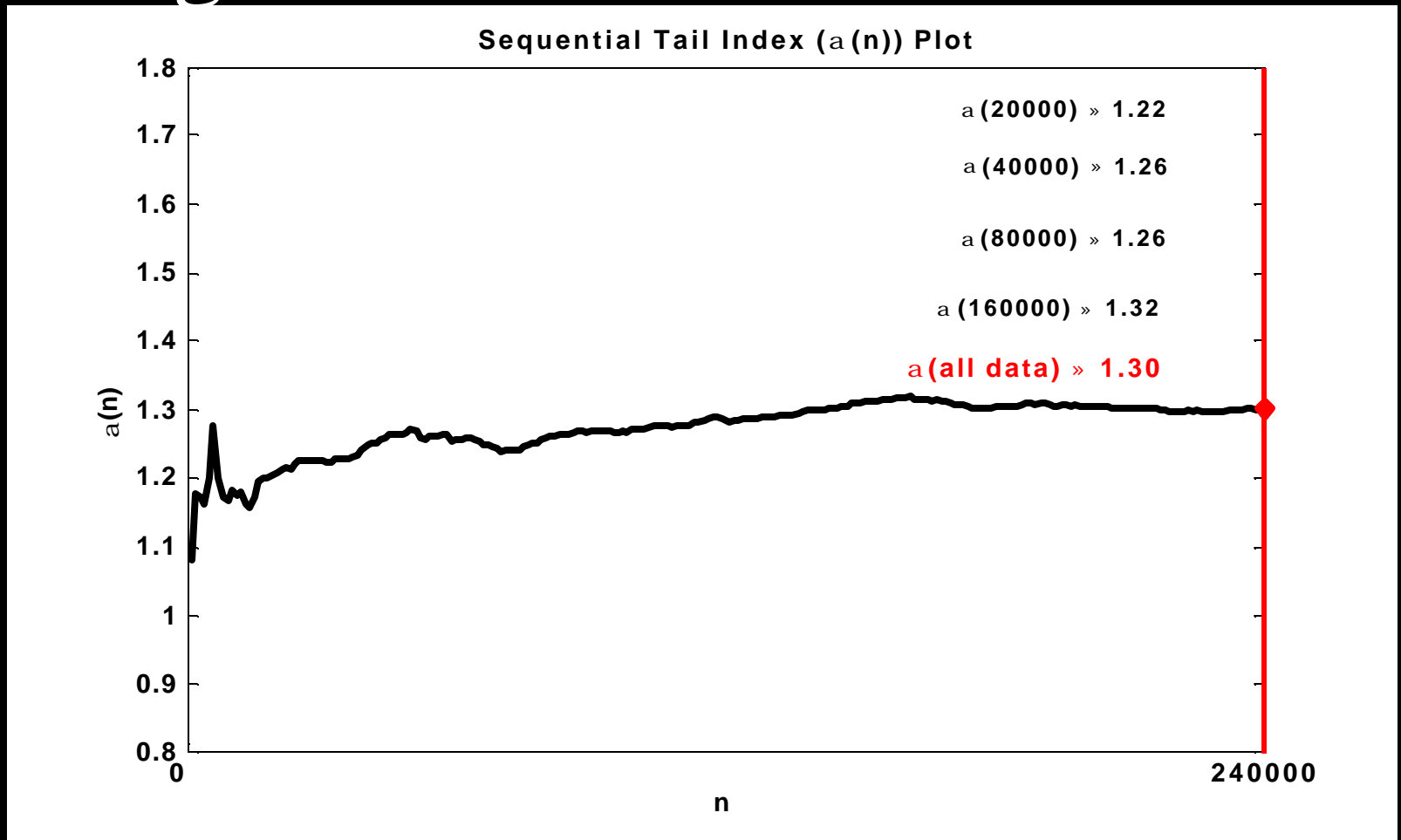
Fitting Pareto: $n=80,000$



Fitting Pareto: $n=160,000$



Fitting Pareto: All data



The case for scaling distributions

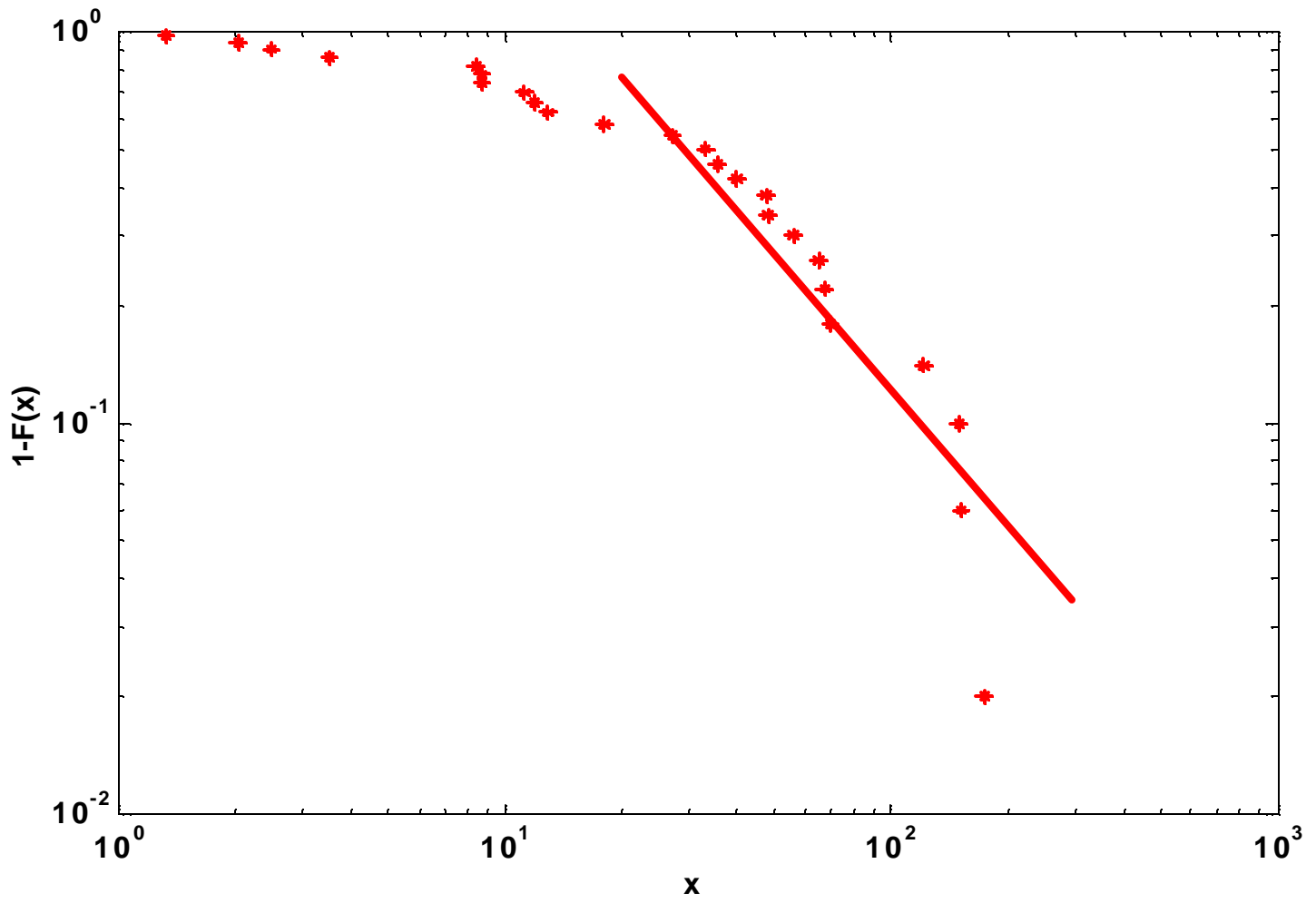
- The “creativity” of scaling distributions
 - Data: Divergent sequential moment plots
 - Mathematics: Allow for infinite moments
- Re-discover the “science” in “scientific modeling”
 - Scientifically “economical” modeling (when more data doesn’t mean more parameters)
 - Statistically “efficient” modeling (when more data mean more model accuracy/confidence)
 - Trading “goodness-of-fit” for “robustness”

Looking ahead ...

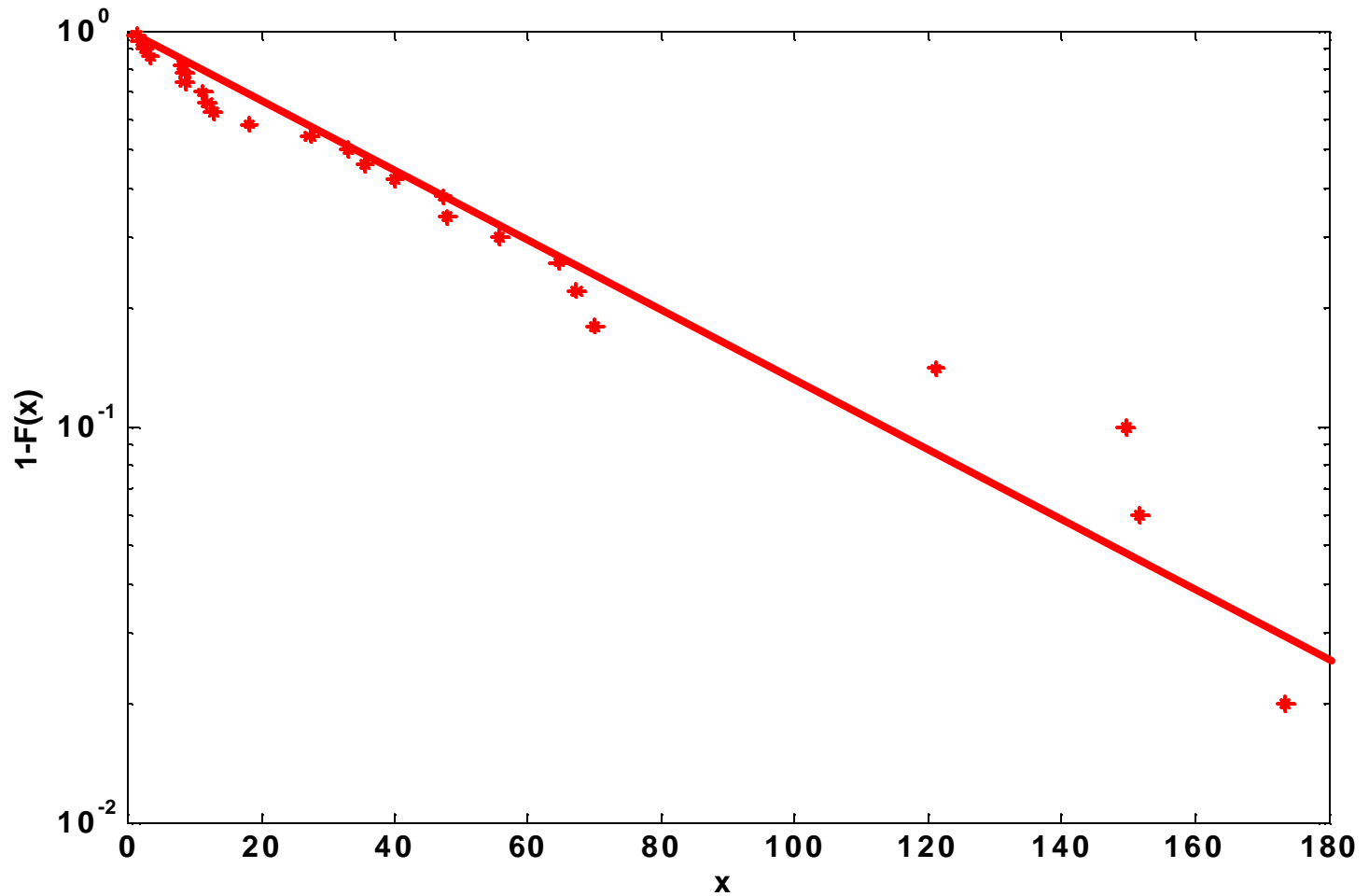
- Main objective of current empirical studies
“The observations are consistent (in the sense of “curve fitting”) with model/distribution X , but are not consistent with model/distribution Y .”
- Requirement for future empirical studies
“The observations are consistent (in the sense of “borrowing strength”) with model/distribution X , and X is not sensitive to the methods of measuring and collecting the observations.”

Some Words of Caution ...

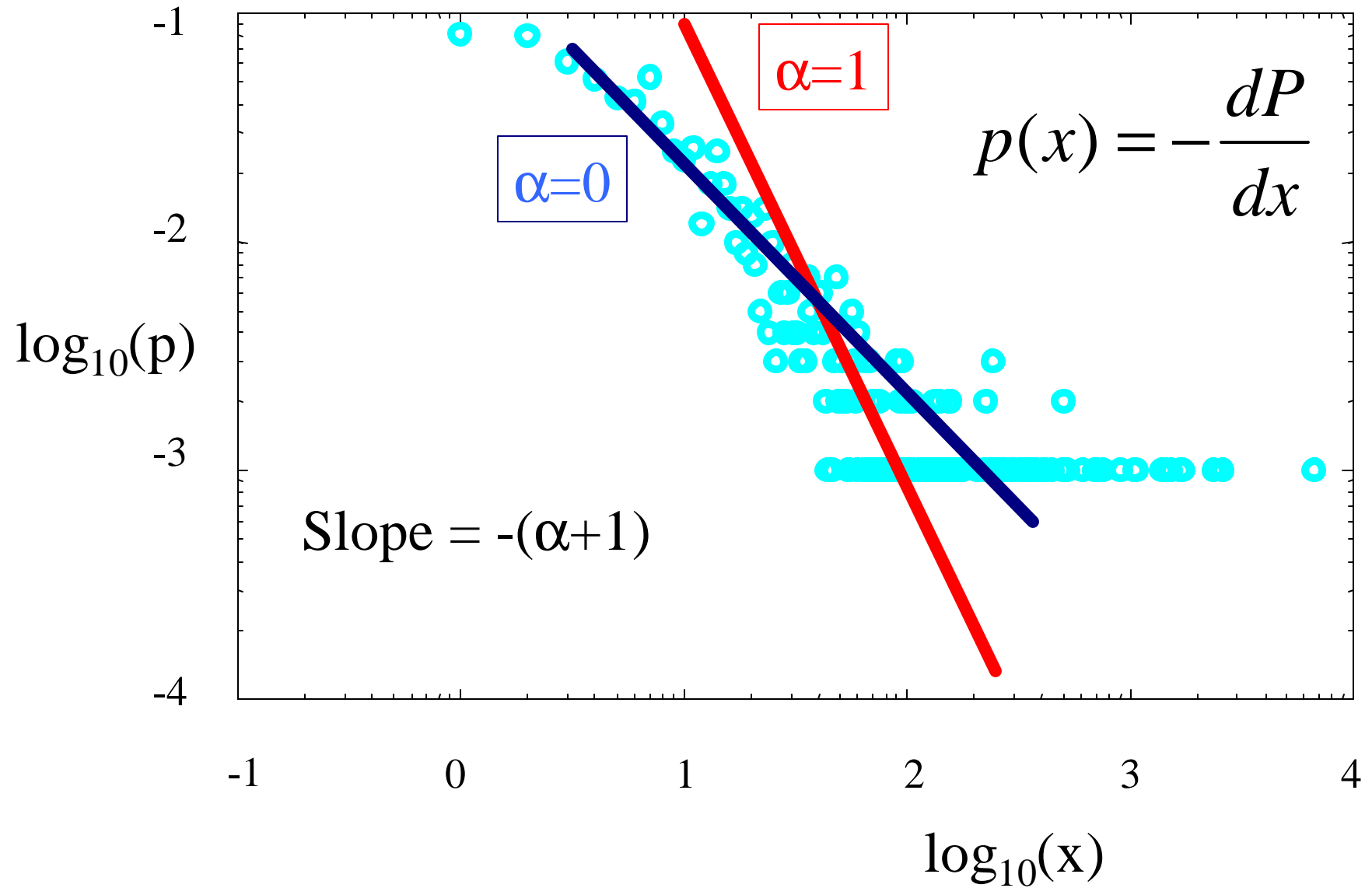
- Not every “straight-looking” log-log plot means “heavy tails”!
- Never use frequency plots to infer heavy tails – even though physicists do it all the time!



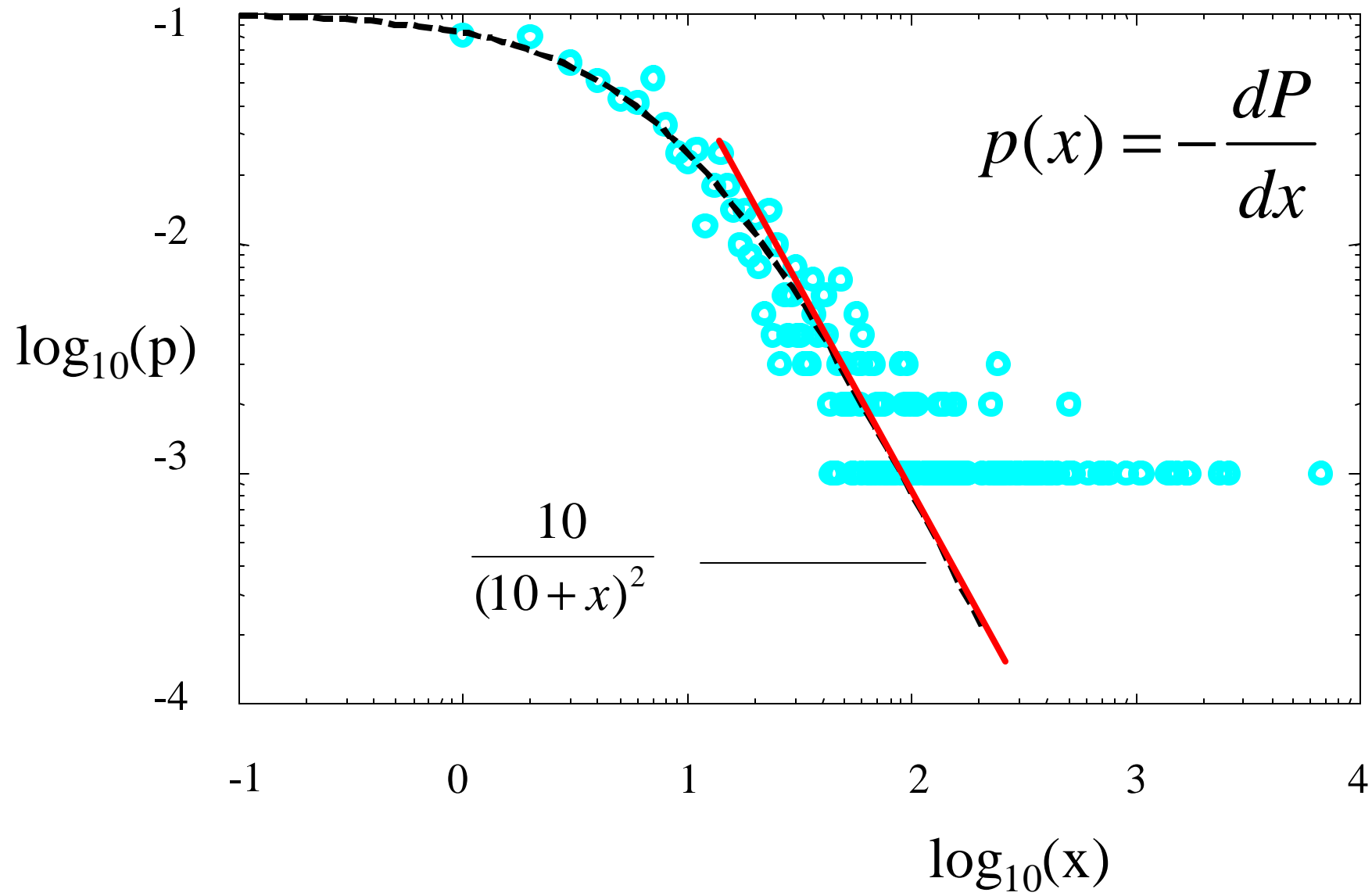
Straight-looking log-log plot?

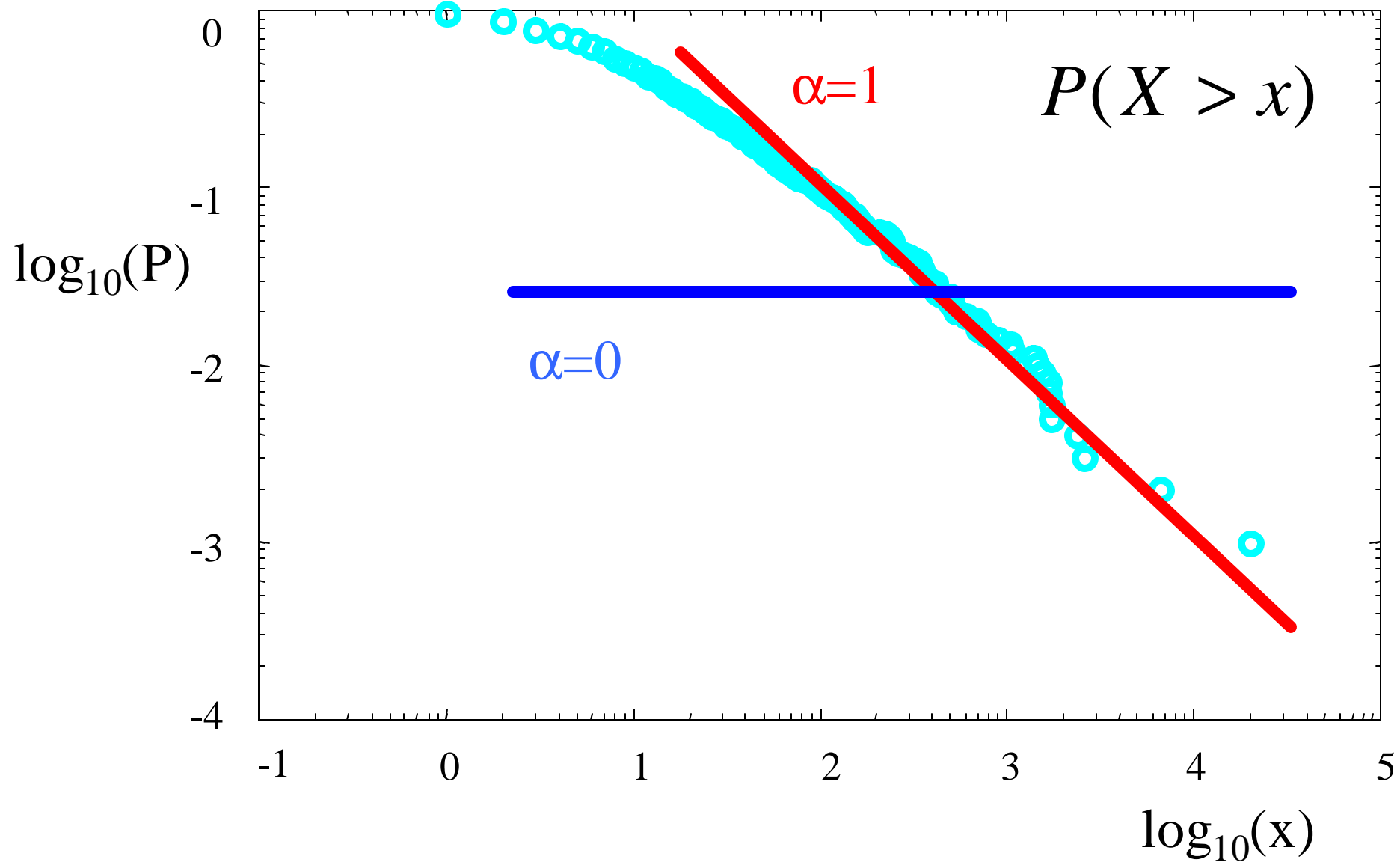


No! 25 samples from Exp(50) – linear semi-log plot!



NEVER infer α from frequency plots on log-log scale!





ALWAYS infer α from CCDF plots on log-log scale!

A Word of Wisdom ...

*In my view, even if an accumulation of quick
“fixes” were to yield an adequately fitting
“patchwork”, it would bring no understanding.*

– B.B. Mandelbrot, 1997