Why is Internet traffic self-similar (or is it)?

Allen B. Downey

Olin College of Engineering

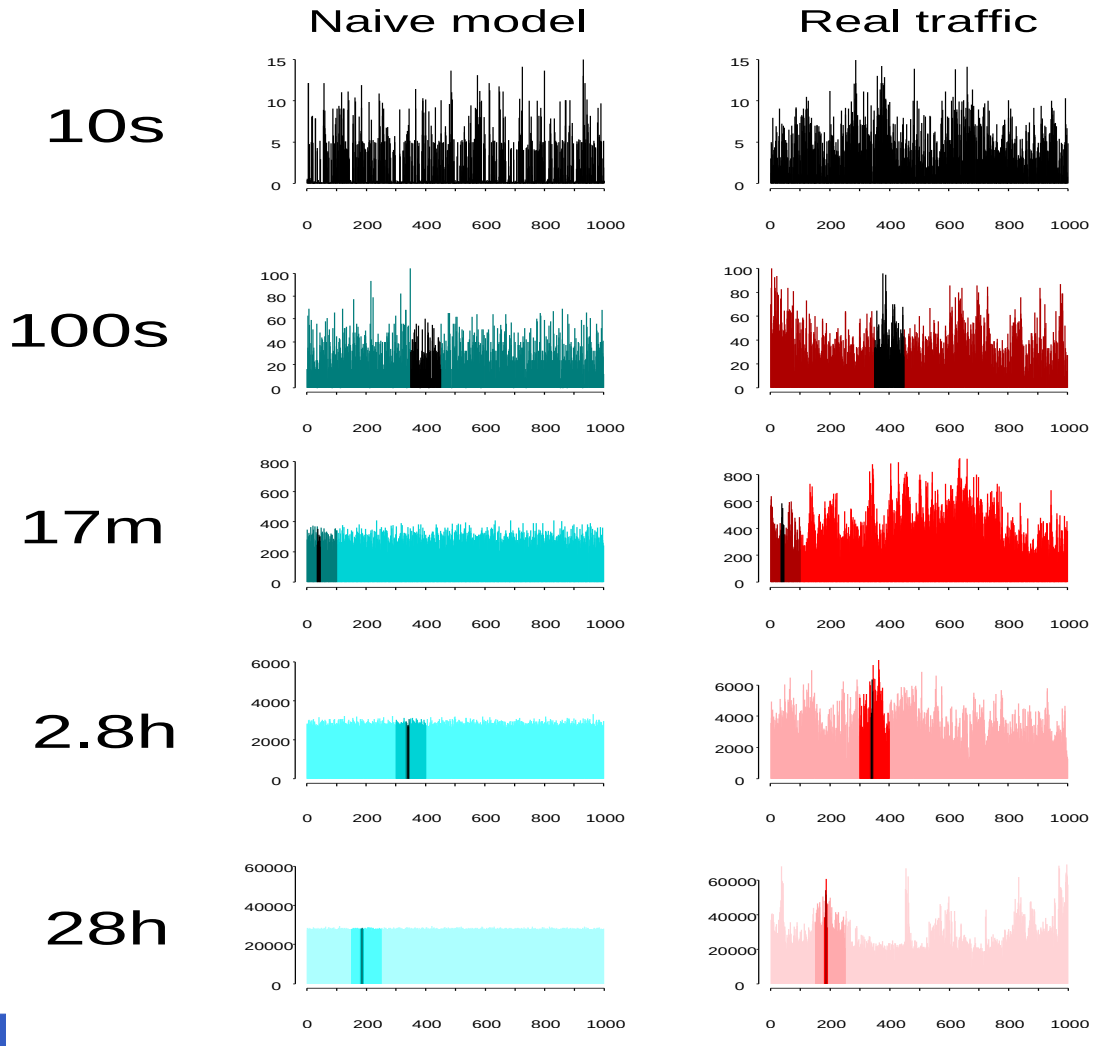
Needham MA

Outline

- Self-similar network traffic:
  - Observations.
  - Explanatory models.
- Long-tailed distributions:
  - Observations.
  - Explanatory models.
  - Implications for self-similarity.

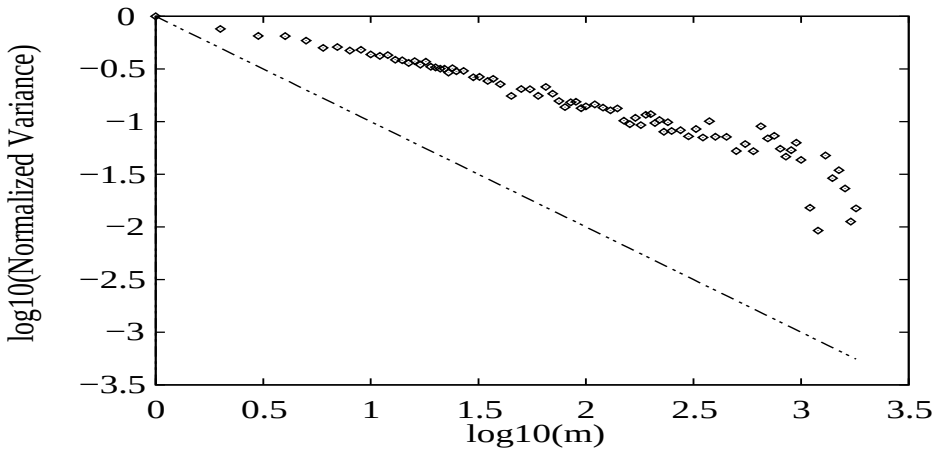
Warning: survey with overrepresentation of me.

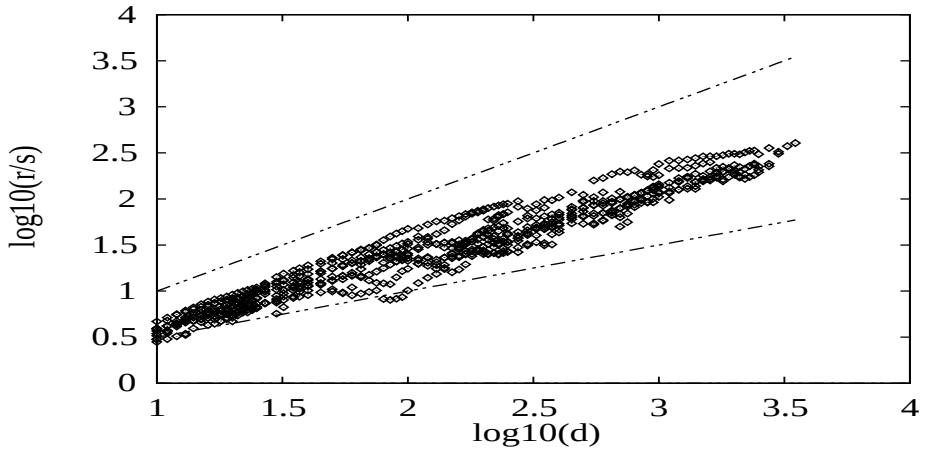
Self-similarity



- Throughput of an Ethernet router at Bellcore, 1989-92.
- Figure from Leland, Taqqu, Willinger, Wilson, 1993.

Long range dependence





- Internet: Paxson and Floyd, 1994.
- WWW: Figure from Crovella and Bestavros, 1996.
- Hurst parameter, $H \approx 0.7$

So what?

- Surprising for an engineered system.
- Contrary to engineering assumptions.

What does it mean?

"Internet traffic is self-similar."

S1: Network traffic has some characteristics of a self-similar process.

S2: A self-similar process is a useful model of network traffic.

S3: A self-similar process is the right model; other models are wrong.

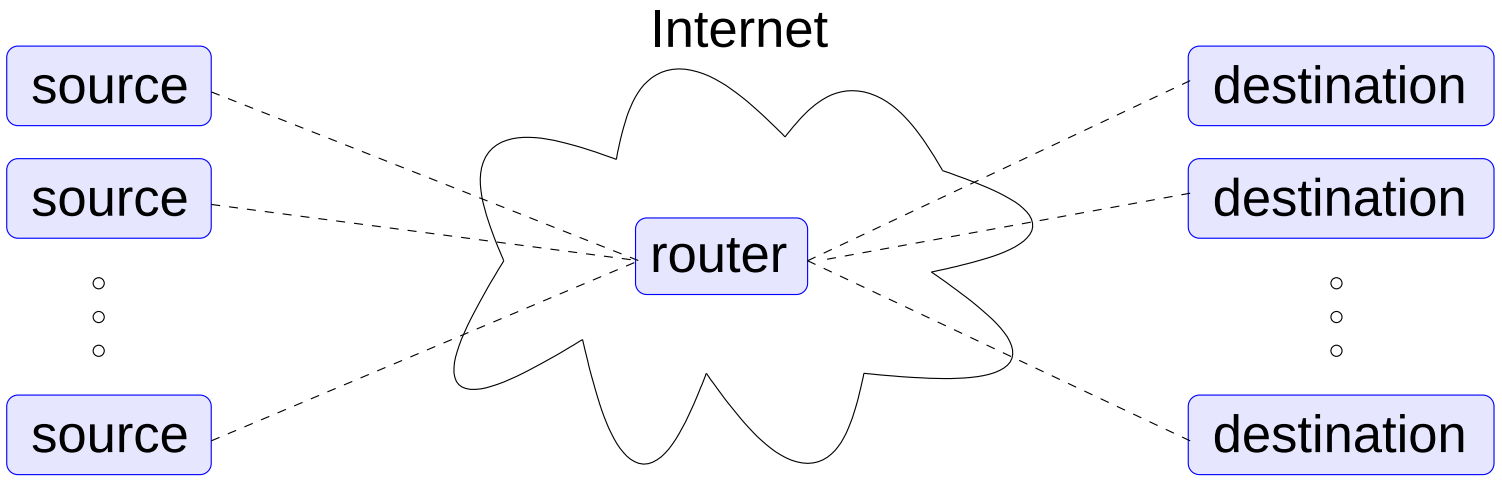
The self-sim hypothesis

What would support the strong version?

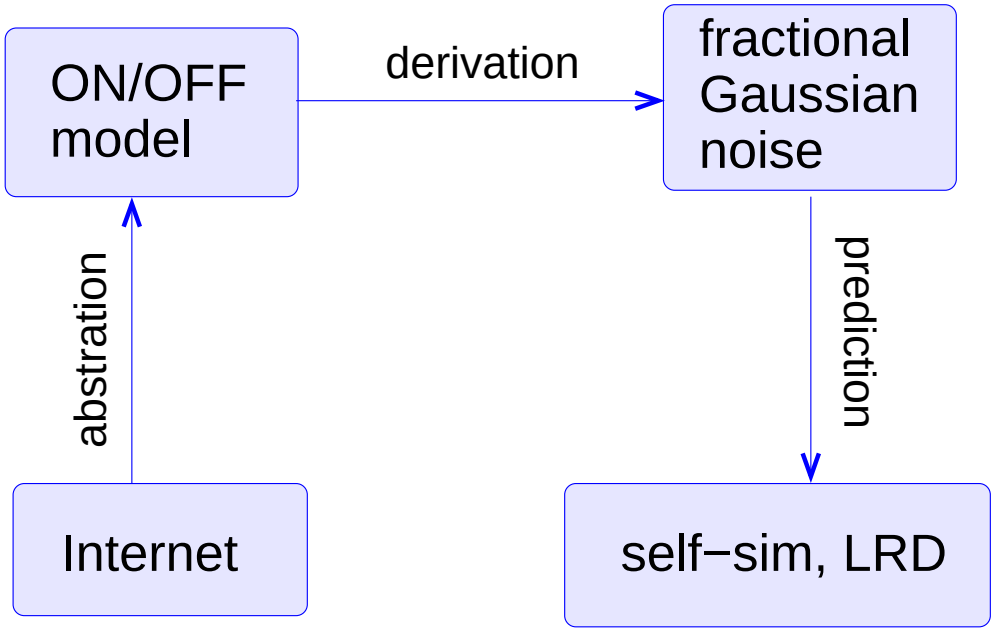
- Consistent and unambiguous evidence.
- Explanatory model(s).
- Lack of compelling alternatives.

Abstract models

- ON/OFF model: Taqqu, Willinger, Sherman, 1997.

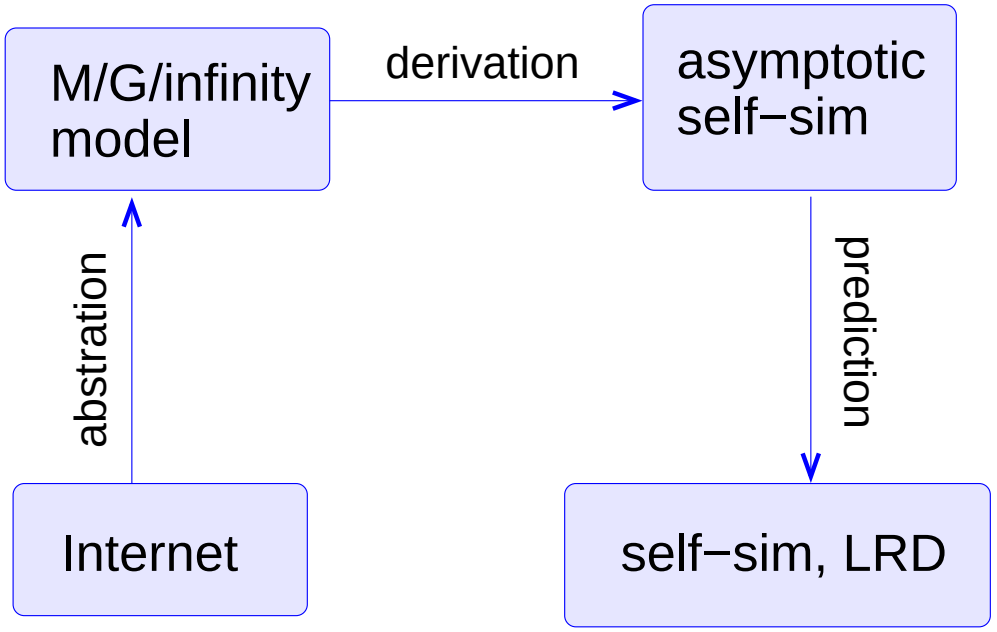


Abstract models



- Abstract model explains observed behavior.

Abstract models



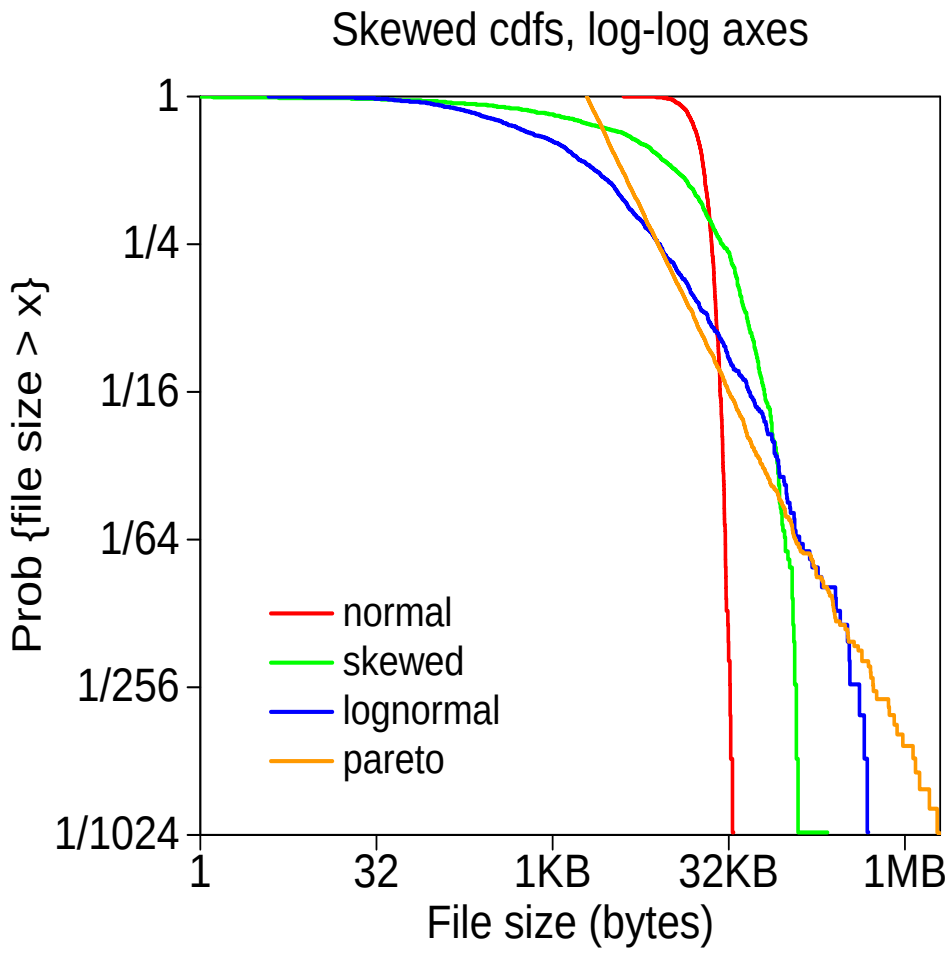
- M/G/∞ model: Parulekar and Makowski, 1997.

Long-tailed distributions

- Both models based on long-tailed distributions:
  - ON/OFF model: ON *or* OFF times.
  - M/G/∞ model: service times.
- In this context, long-tailed means:

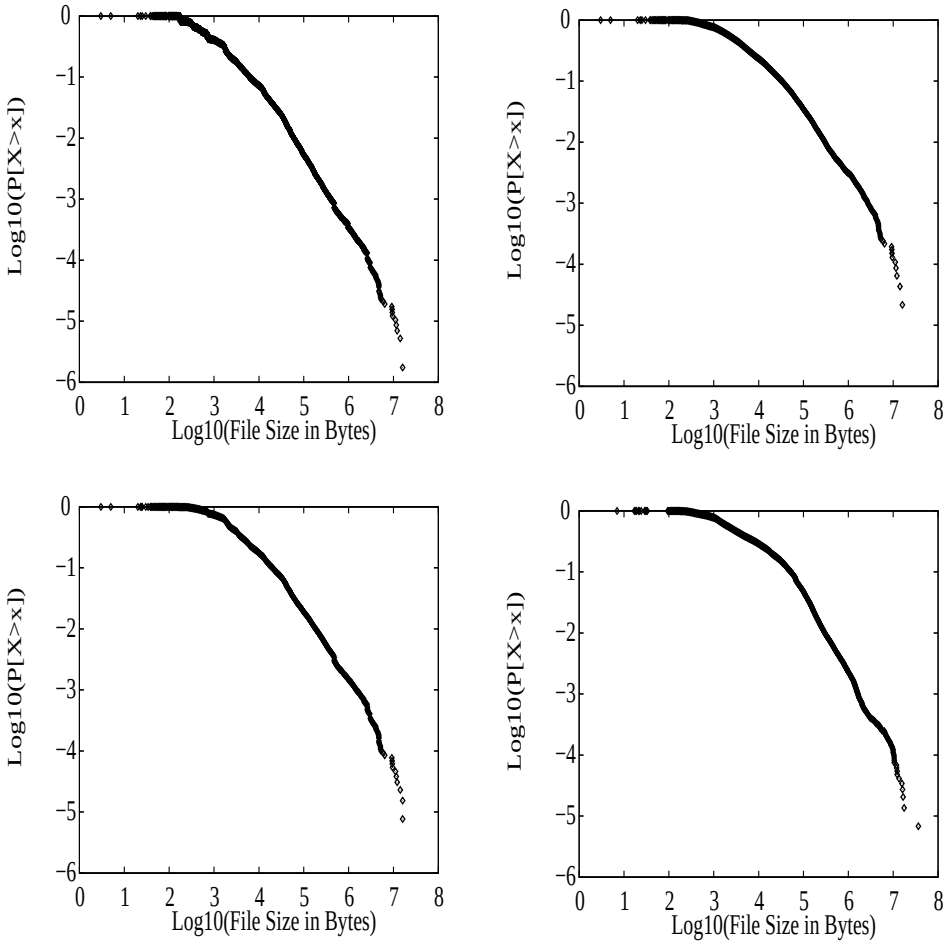
$$P\{X > x\} \sim cx^{-\alpha} \quad \text{as} \quad x \to \infty$$

Long tails



- Log-log plot of ccdf.
- Straight line indicates long tail.

Some evidence



- Figure from Crovella and Bestavros, 1996.
- WWW files requested, transferred, unique, available.

Explanatory model

- Highly optimized tolerance (HOT), Carlson and Doyle, 1999.
- Web content is a document.
- Designers optimize it so that high hit portions are in smaller files.
- Result is Pareto file sizes.

Explanatory model

- Realistic model?
- What about local file systems?
- Contradictory evidence.

Contradictory evidence

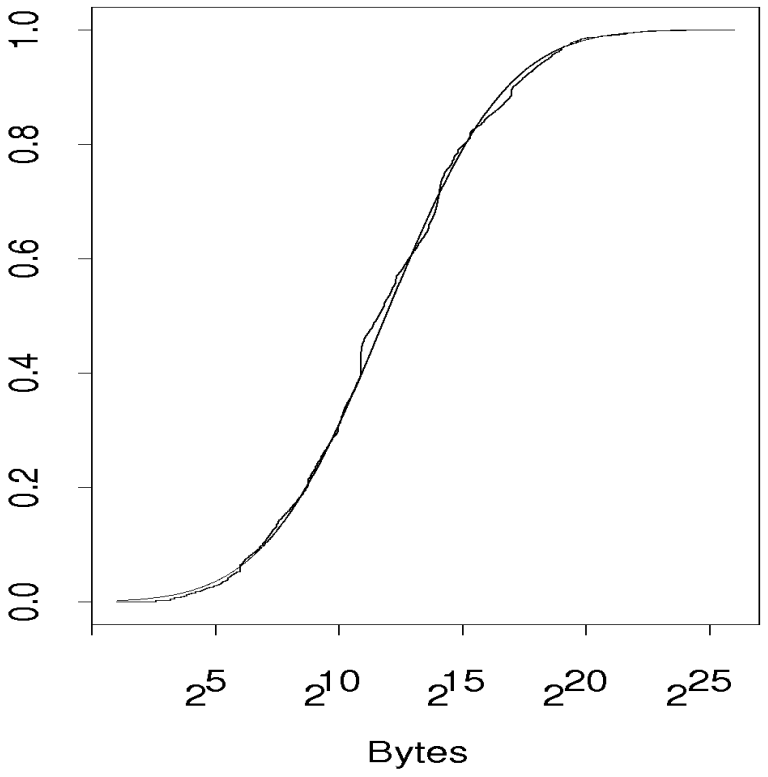
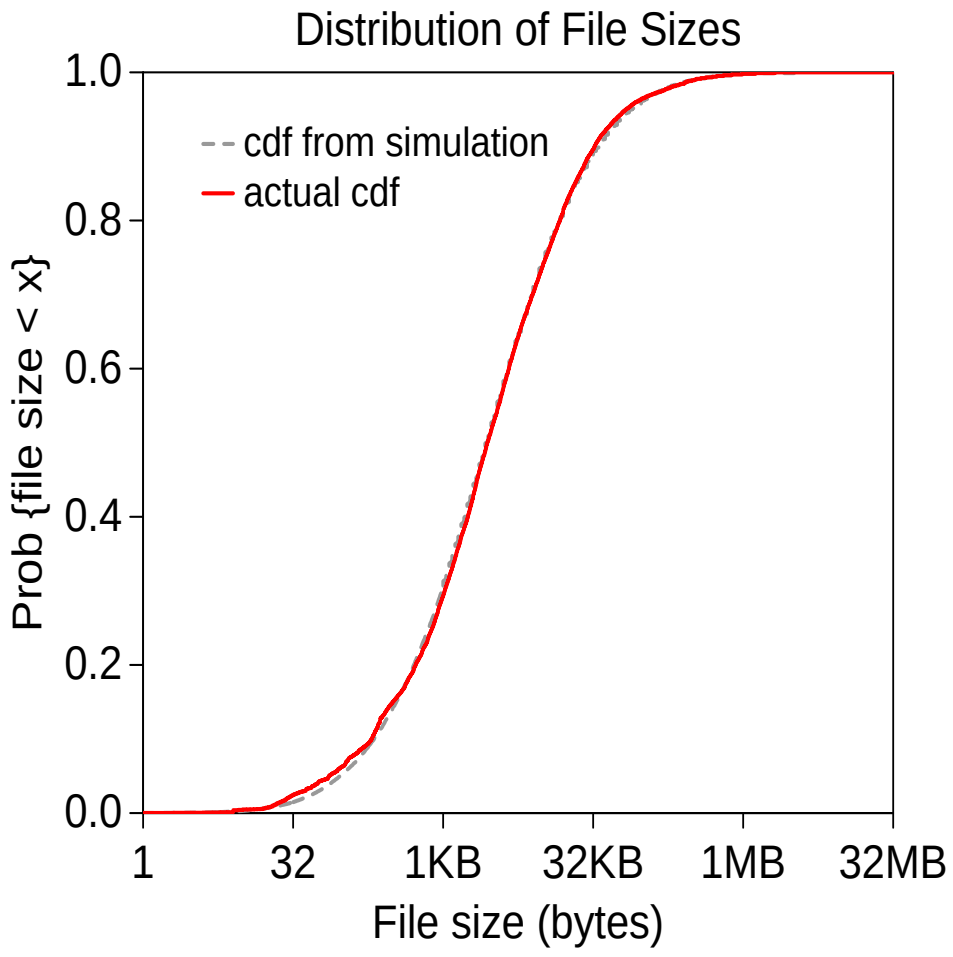


Figure 9: Log_2-Normal Fit to LBL-4 FTP Data Bytes

- Figure from Paxson 1994.
- Distribution of FTP transfer sizes.
- Lognormal model.

Distribution of file sizes



- My computer, 90,000 files.
- Lognormal model.

Pareto vs. lognormal

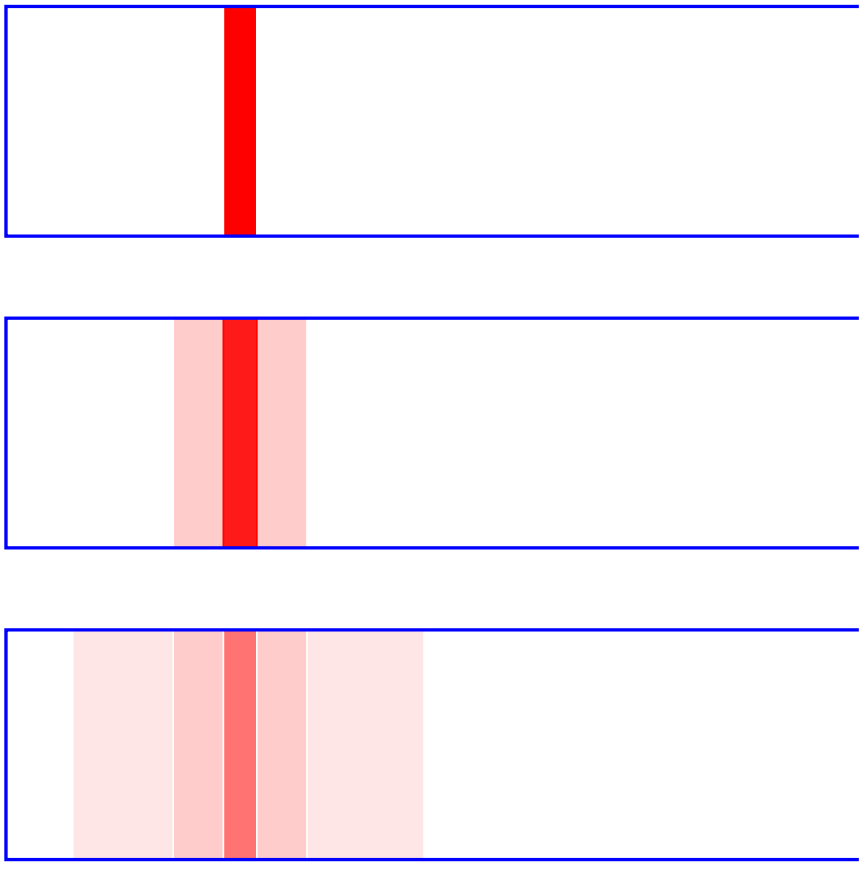
- It turns out that Pareto vs. lognormal is a running debate in many fields.
- But I didn’t know any better...

Explanatory model

"The structural cause of file size distributions," Downey, 2001.

- Most new files are based on existing files.
- New size proportional to old.
- User model:
  - Choose a base file, size s.
  - Choose a multiplicative factor, f.
  - Create a file, size fs.

Lognormal



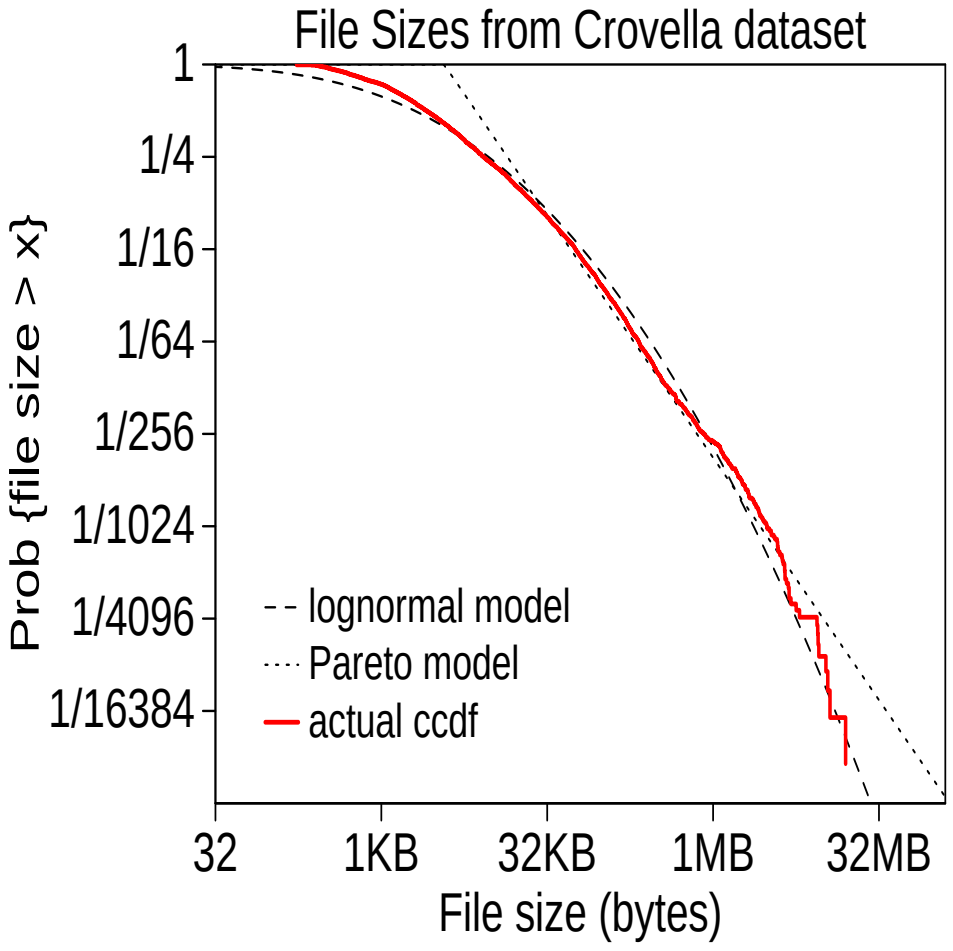
- Simulation: file sizes “diffuse” in log space.
- Analysis: product of random variables tends to lognormal.

Evidence

"Evidence for long-tailed distributions in the Internet," Downey, 2001.

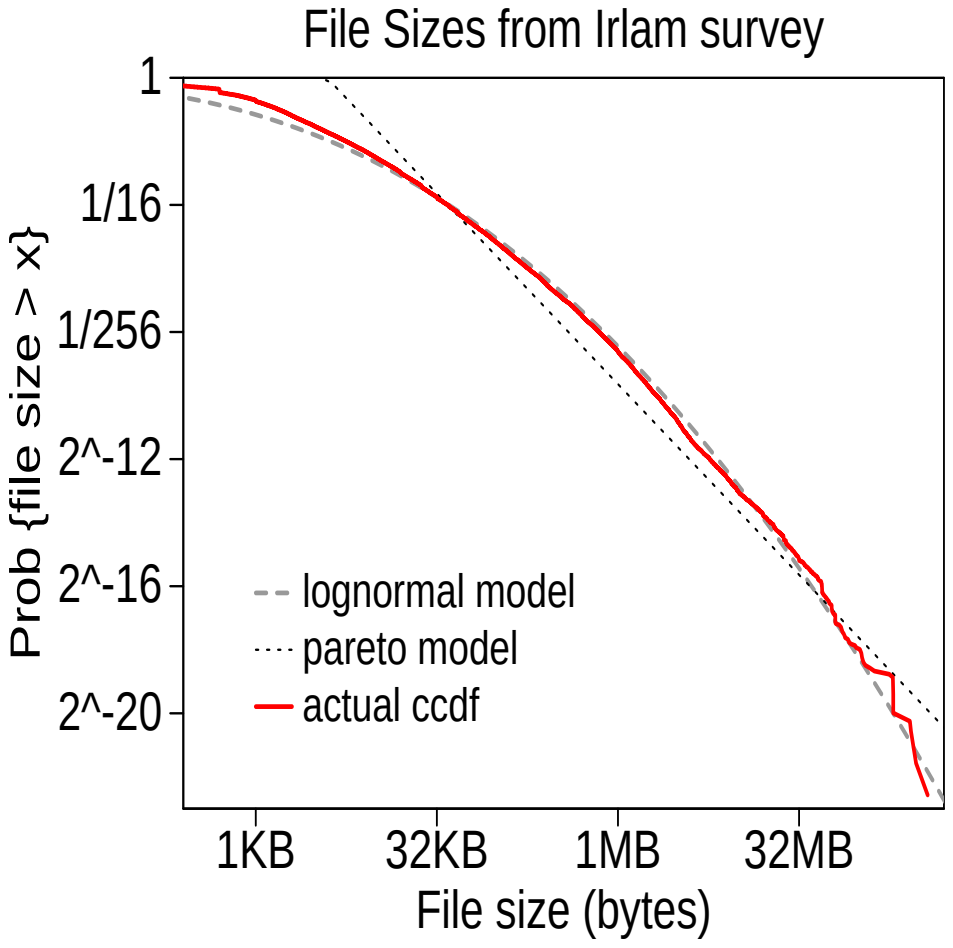
- Evaluated prior claims about long-tailed distributions.
- Most were based on ccdf test.
- Most showed some tail curvature.

Crovella dataset



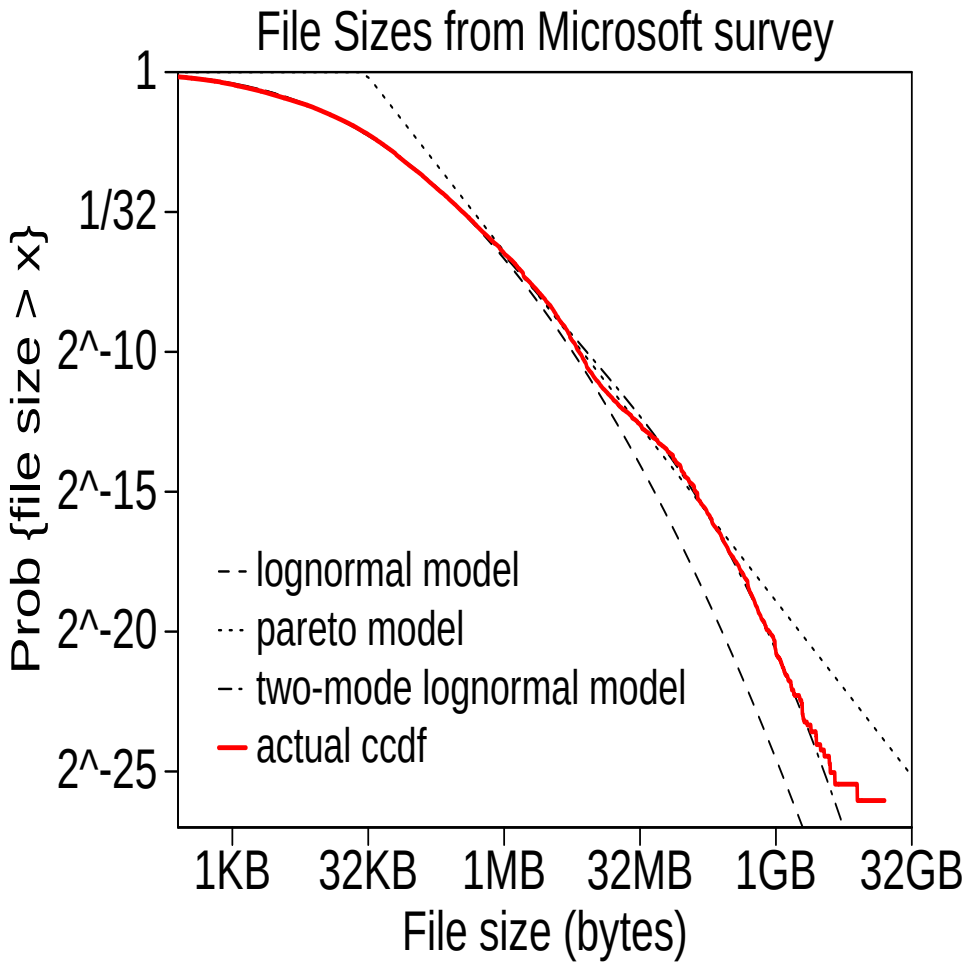
- Data from Crovella and Bestavros, 1996, my analysis.
- 36,000 files.
- Lognormal better fit for body and extreme tail.

Irlam dataset



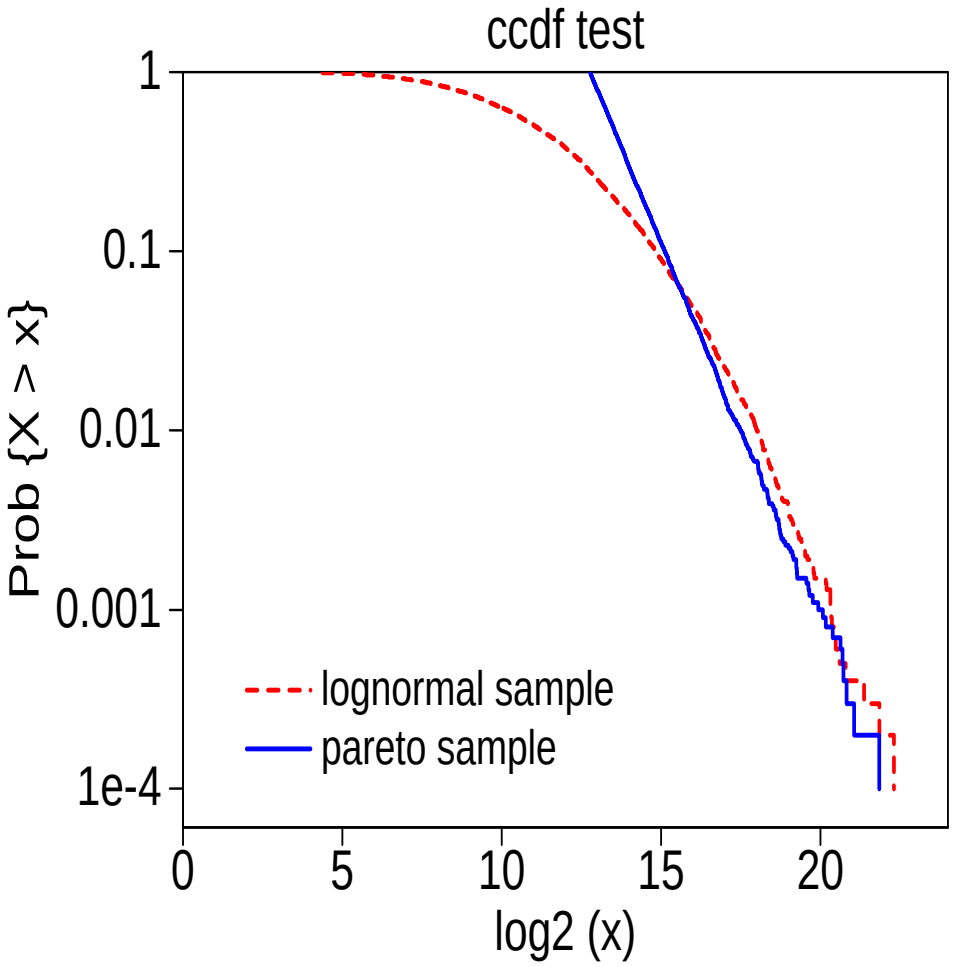
- 655 UNIX systems, 6 million files.
- Data from Irlam, 1993, my analysis.

Microsoft dataset



- 10,568 workstations, 140 million files.
- Data from Douceur and Bolosky, 1999, my analysis.

Limits of ccdf test



- Tail behavior of samples can be indistinguish-able.
- ccdf weights data inappropriately.

Tail curvature test

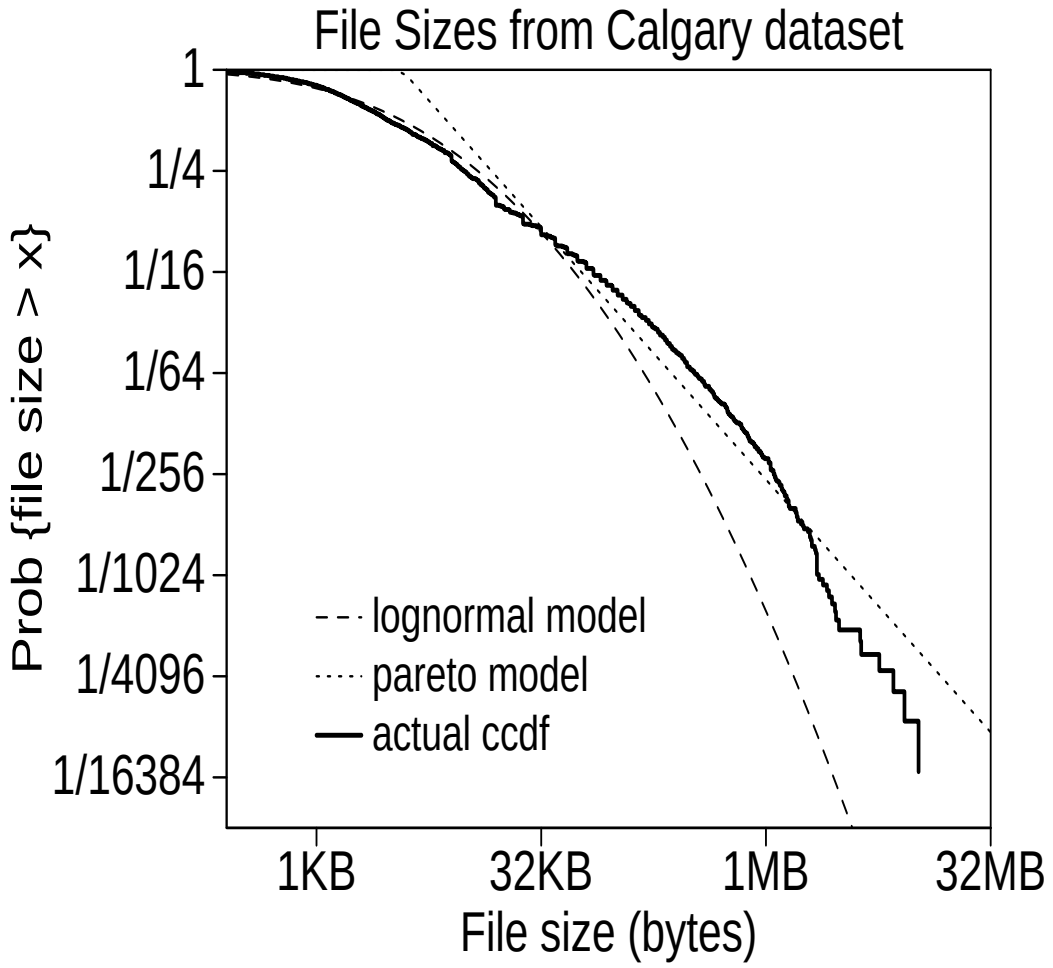
- Observation:
  - Sample from lognormal more likely to curve down than sample from Pareto with similar tail behavior.
  - Lack of curvature is definitive of long-tailed distribution.

$$P\{X > x\} \sim cx^{-\alpha} \quad \text{as} \quad x \to \infty$$

Tail curvature test

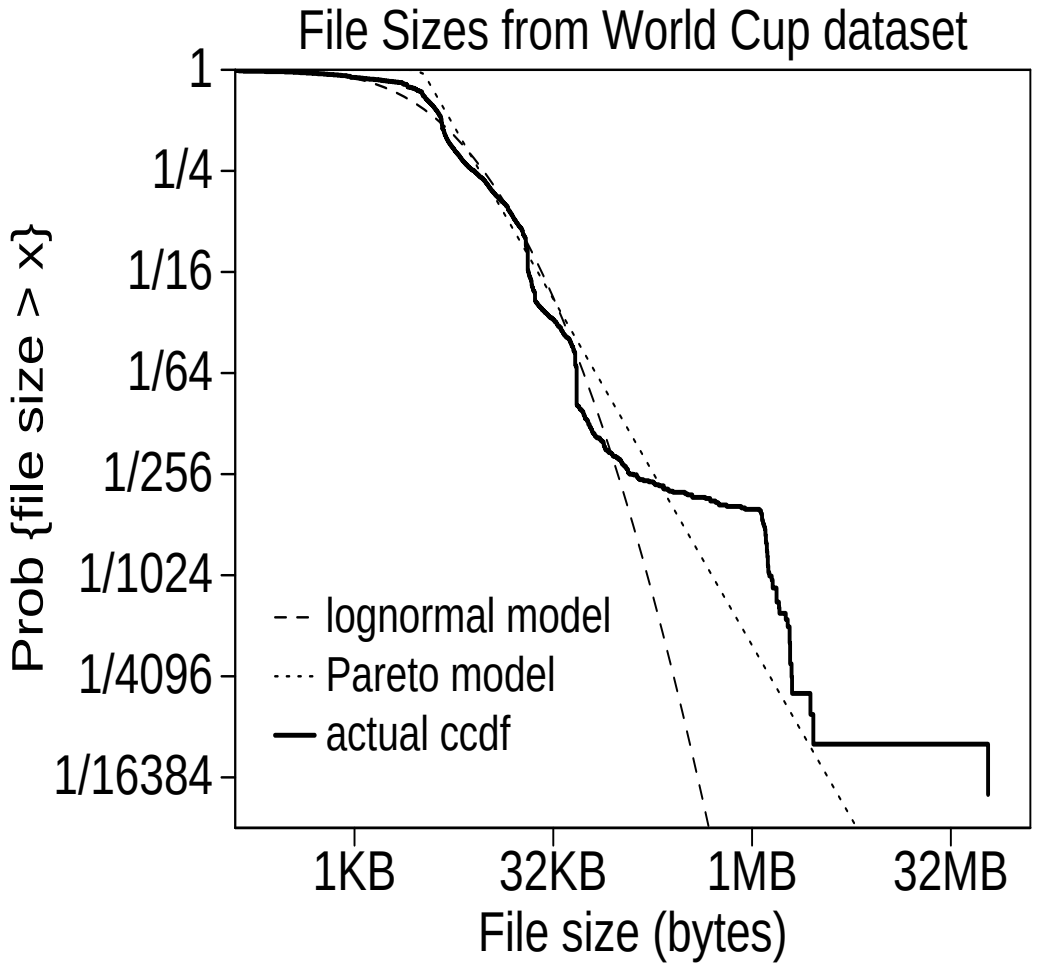
1. Estimate “tail curvature.”
2. Fit lognormal and Pareto models to data.
3. Generate samples from fitted distributions.
4. Compute p-values for measured curvature under fitted models.

Tail curvature test



- Quantifies ccdf test.
- Weights data appropriately.
- Useful discriminator.
- Data from Arlitt and Williamson, 1996, my analysis.

Tail curvature test



- Not much help with this.
- Data from Arlitt and Williamson, 1996, my analysis.

Summary so far

- Preponderance of evidence for lognormal model.
- Some examples hard to characterize.
- Good news for modelers.
- Bad news for explanations of self-similarity?
  - Lognormal is not long-tailed (enough).

Something else?

“Lognormal and Pareto distributions in the Internet”, Downey 2004.

- Even if file sizes are not long-tailed, maybe:
  - Interarrival times?
  - Mixture distribution?
  - Transfer times?
  - Burst durations?

Something else?

- Interarrival times? No.
- Mixture distribution? Probably not.
- Transfer times? Probably not.
- Burst durations? Best candidate.

Later work

“Dynamic models for file sizes and double Pareto distributions,” Mitzenmacher, 2004.

- Extension of my model with new “root” files and deletion.
- Result is double Pareto.
- Fits the body of the distribution better than Pareto.
- Still hard to discriminate tail.

Later work

"A pragmatic approach to dealing with high-variability in network measurements," Willinger, Alderson and Li, 2004.

- Check convergence of estimated σ as sample size increases.
- Estimates based on lognormal model converge to the "wrong value."
- "A number of technical and practical reasons for preferring scaling distributions over their Lognormal counterparts."

Later work

However:

- One dataset.
- One application, estimating σ.
- Not clear that the “wrong value” is wrong.

Later work

"Power-law distributions in empirical data," Clauset, Shalizi and Newman, 2007.

- Consider diverse "long-tailed" distributions.
- Propose a likelihood ratio test.
- HTTP dataset does not support long-tail hypothesis.
- Lognormal does well for many data sets, but not HTTP!
- Conclusion: neither and both.

Conclusions

File sizes:

- Some evidence that file sizes are lognormal, and some reasons to expect them to be.
- Some evidence that file sizes are Pareto, and some reasons to expect them to be.
- Both are justifiable choices.

Conclusions

S2: A self-similar process is a useful model.

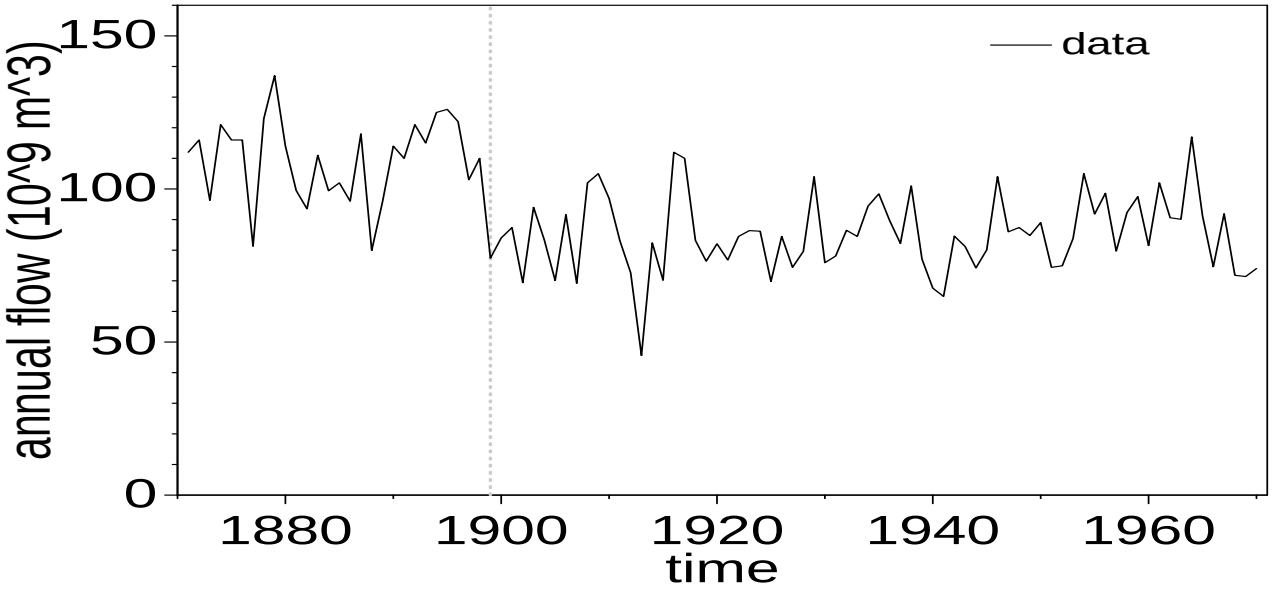
S3: A self-similar process is the right model.

What would support the strong version?

- Consistent and unambiguous evidence.
- Explanatory model(s).
- Lack of compelling alternatives.

Compelling alternative?

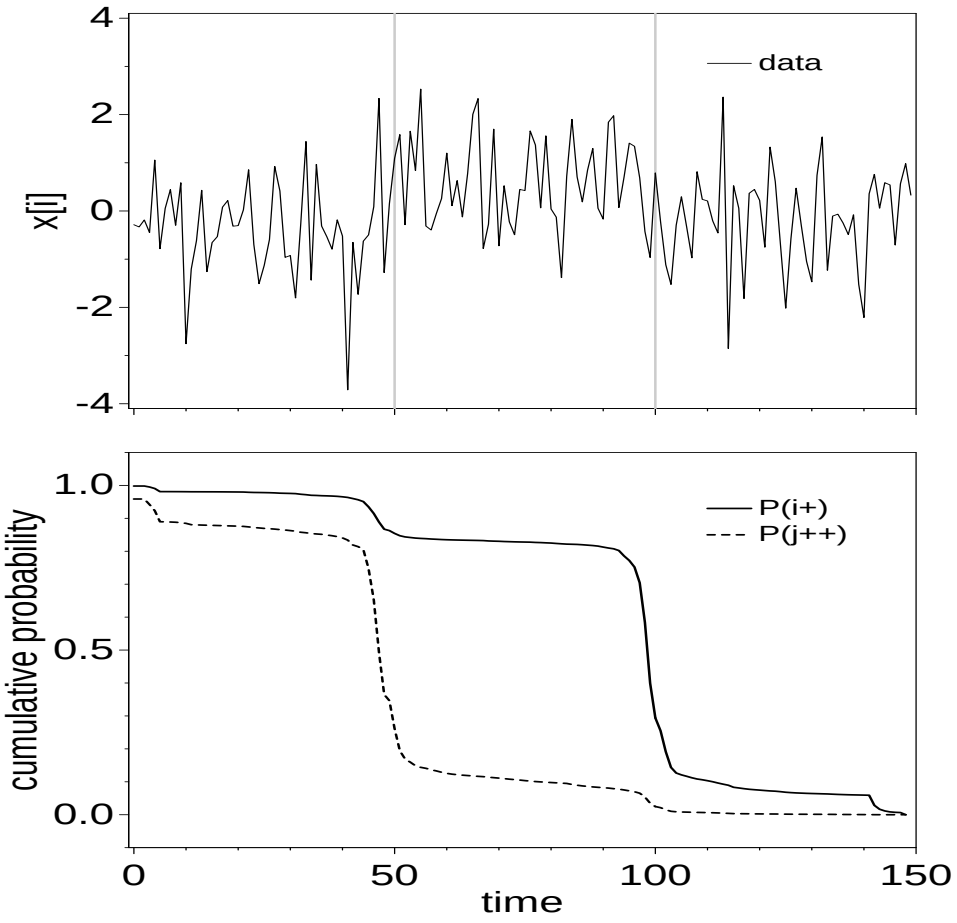
- Stationary model of non-stationary process yields long-range dependence.
- Most plausible explanation of Hurst's hydrological data (Koutsoyiannis 2002).



Ongoing projects

- Bayesian changepoint detection.
- Agent-based models of labor markets, meritocracy, social mobility, distribution of income.

Change detection



- Synthetic series with two changepoints.
- $P(i^+)$: probability that the last changepoint is at i.
- $P(j^{++})$: the second-to-last changepoint is at j.

Distribution of income

- The original Pareto vs. lognormal debate!
- Agent-based model of labor market to explore effects of meritocracy, return on talent, distribution of income.
- Intergenerational talent transfer, assortative mating and social mobility.

Questions

- Email: downey@allendowney.com
- Research: http://allendowney.com
- Textbooks: http://greenteapress.com