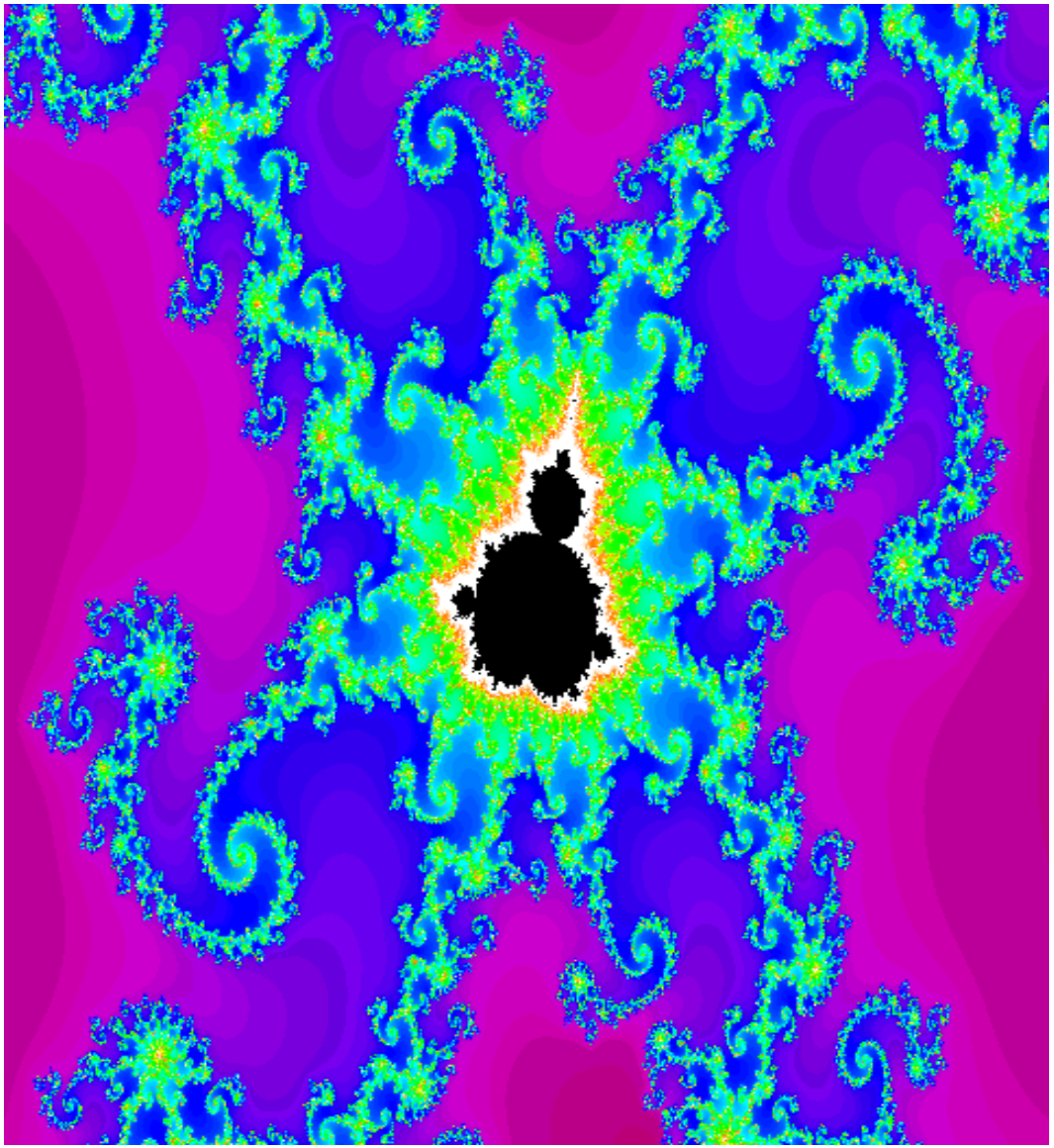
Why is Internet traffic self-similar?

Allen B. Downey
Wellesley College

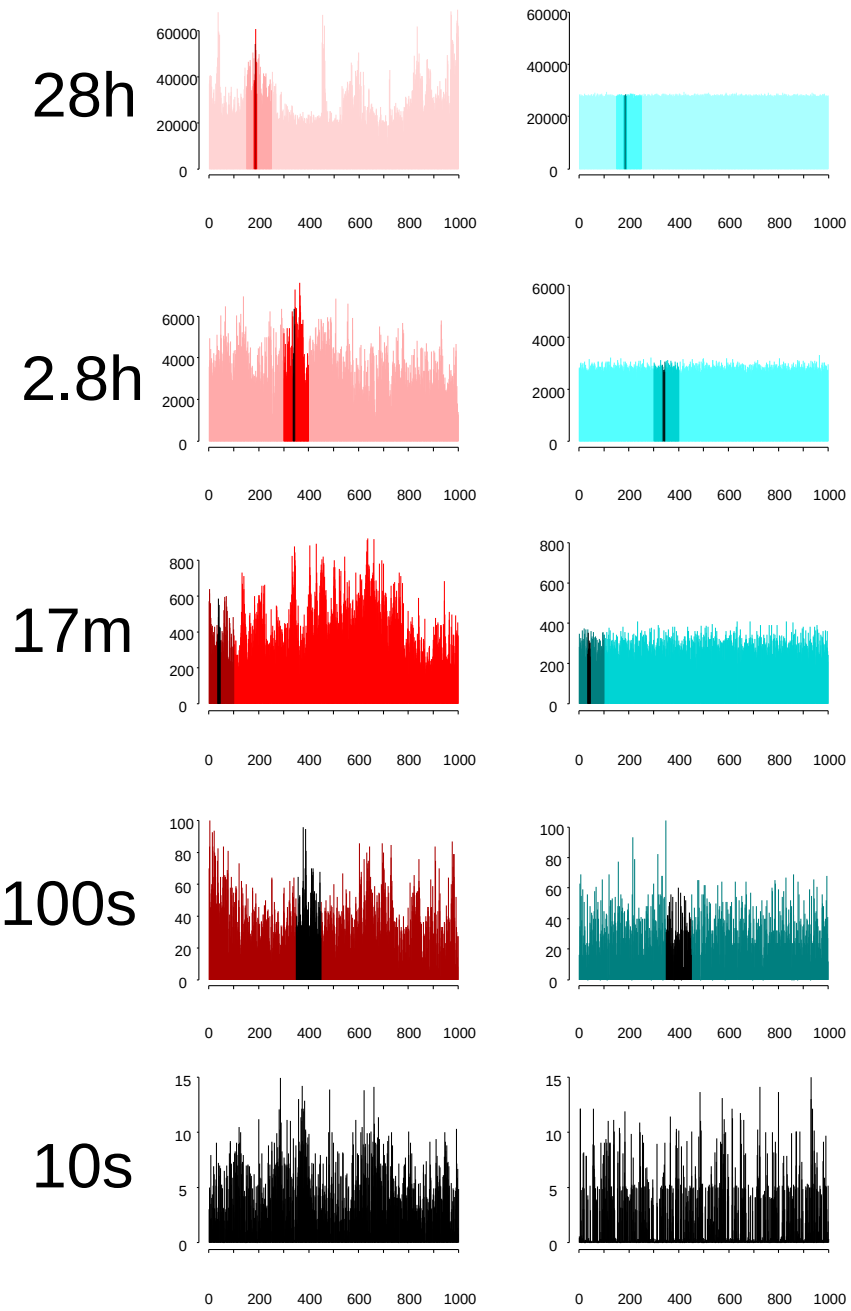
No Micro$oft products were used in the preparation of this talk.

What is self-similarity?



- Real-world: visually similar over range of spatial scales.
- Fractals: geometrically similar over all spatial scales.
- Time-series: statistically similar over range of time scales.

Network traffic



- Ethernet and WAN traffic appear self-similar.

  [WillingerEtAl95]

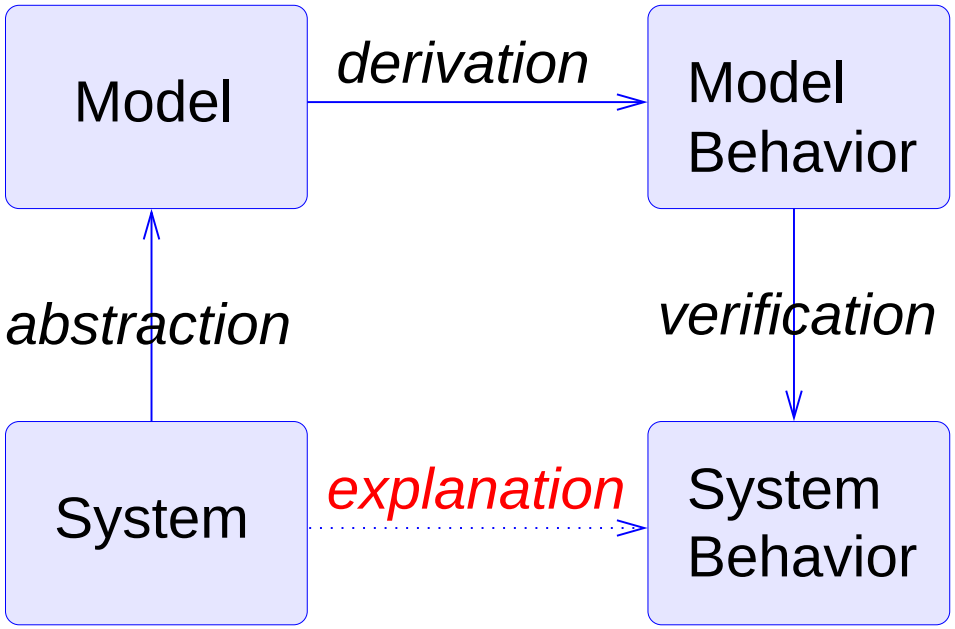
x = time in varying units

y = packets / unit time

- Visual self-similarity over 5 orders of magnitude!

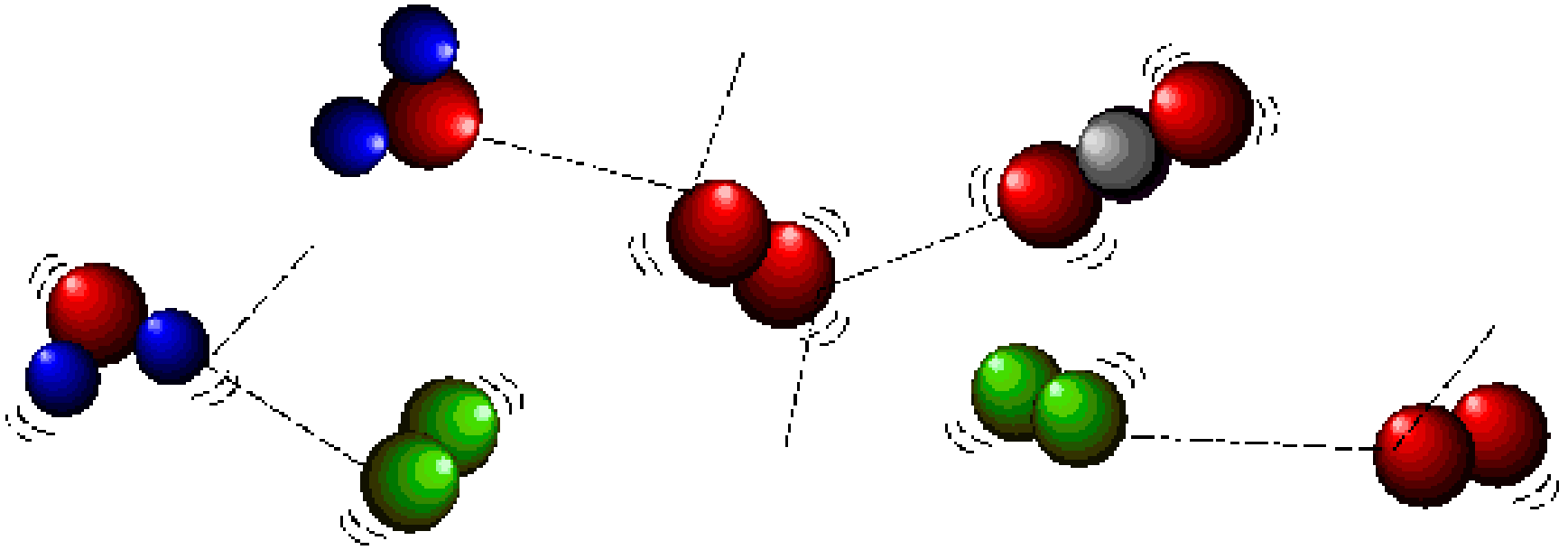
WHY?

Explanatory models



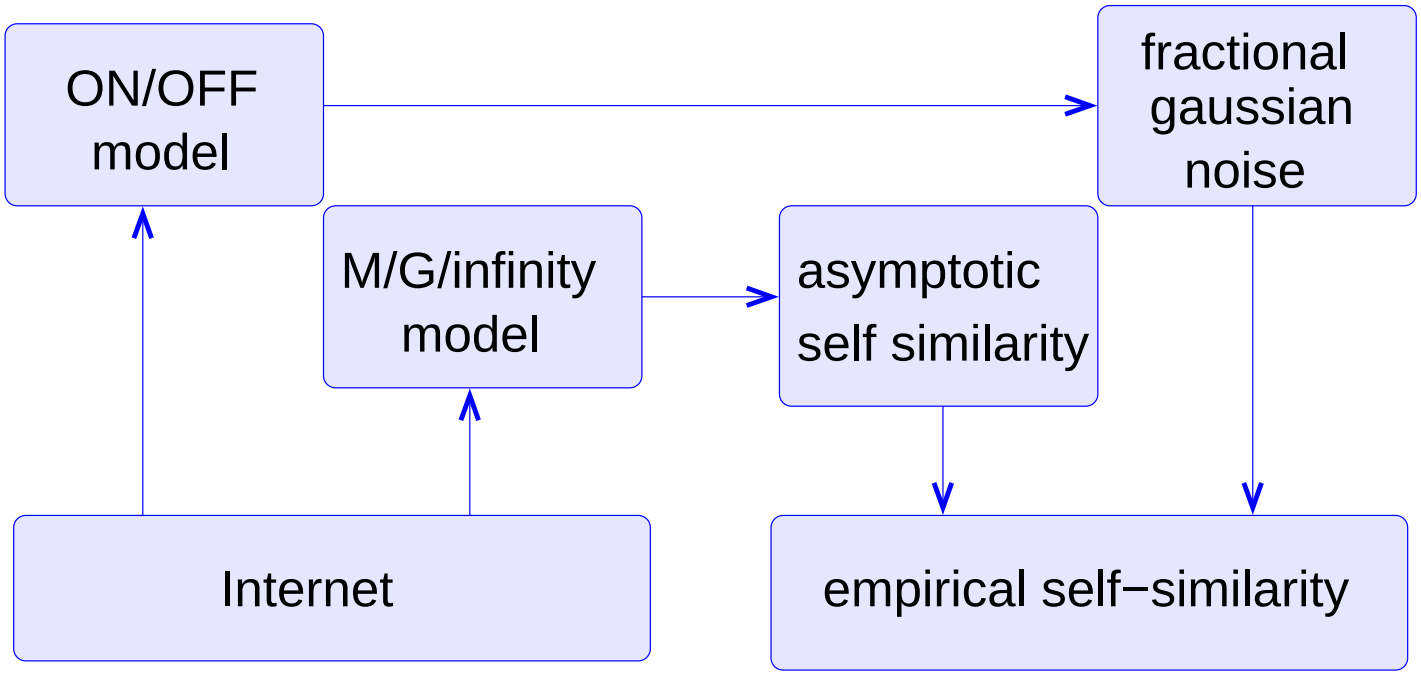
- Abstraction: is it realistic?
- Derivation: is it correct?
- Verification: is the behavior the same?
- Explanation: does this really explain?

Ideal gas law explained



- Abstraction: no interaction, elastic collision, etc.
- Derivation: you do the math (or simulation).
- Verification: most gas, most of the time.

Explanations of self-similarity

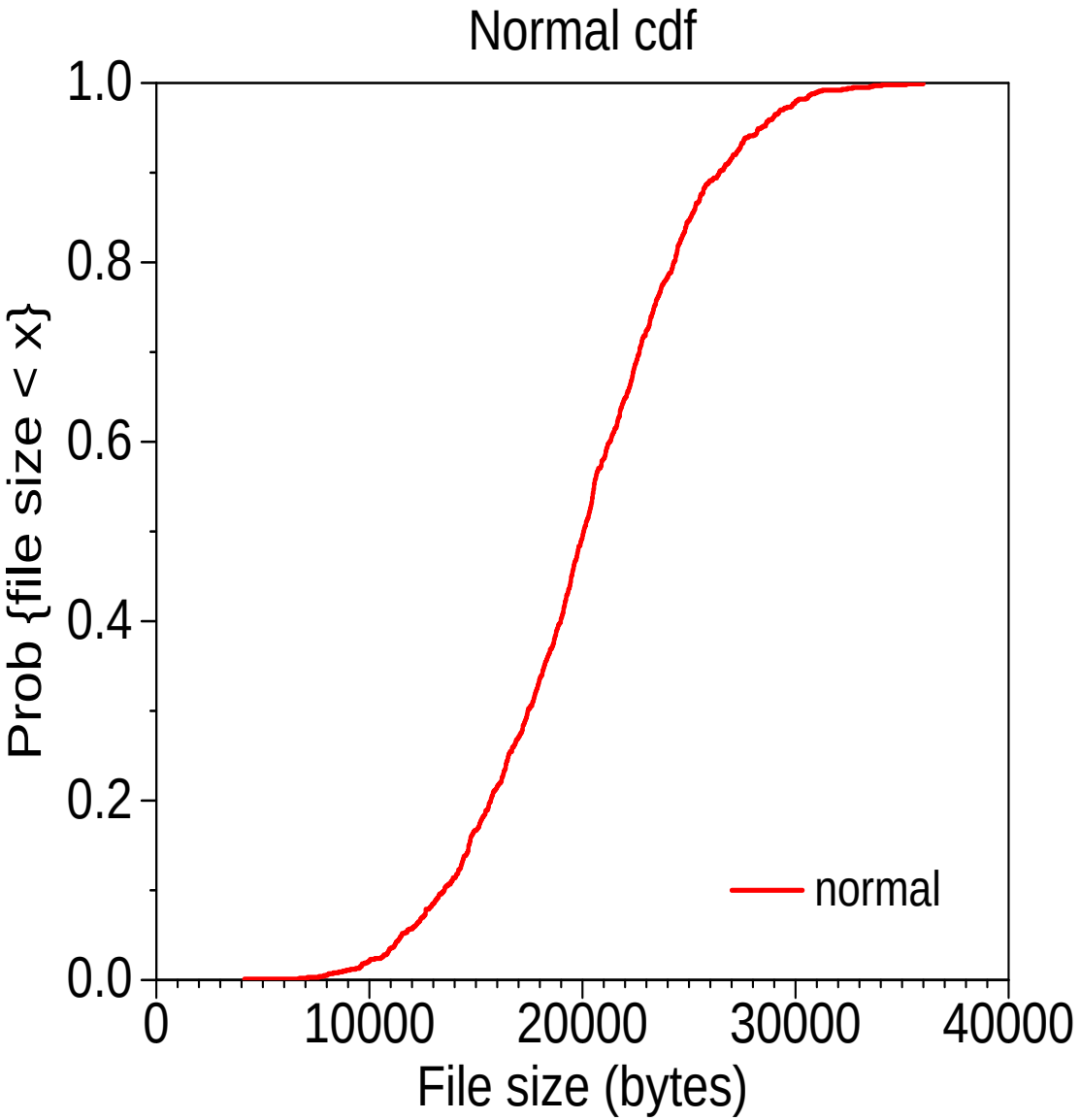


- Abstraction
  - Two aggregation models
  - Long-tailed distribution of file sizes
- Verification
  - FGN is self-similar.
  - ASY isn't, but it can pass.

Distribution of file sizes

- Is it long-tailed?
- If so, why?

Cumulative distributions

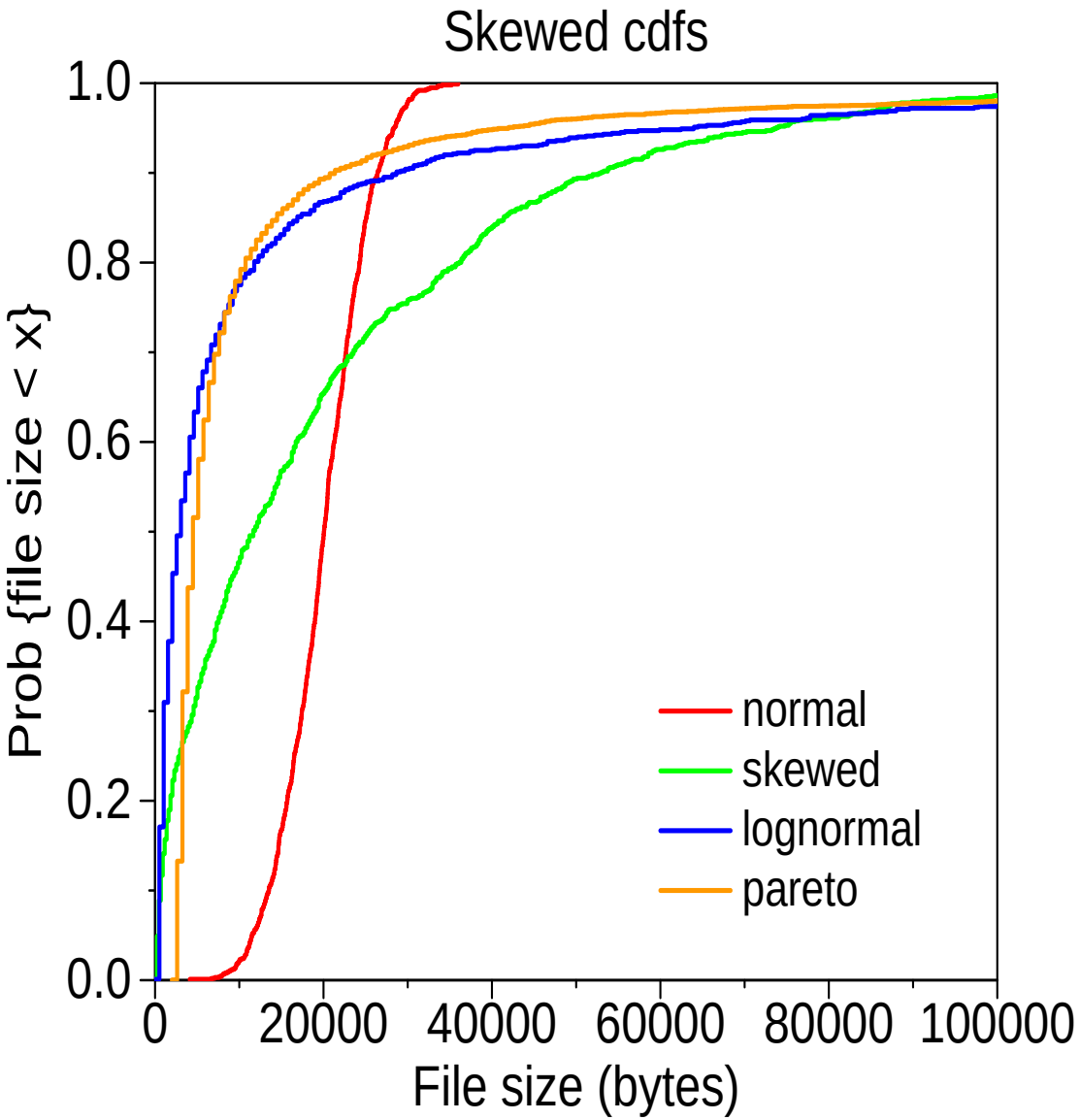


x = range of values

y = Prob {value < x}

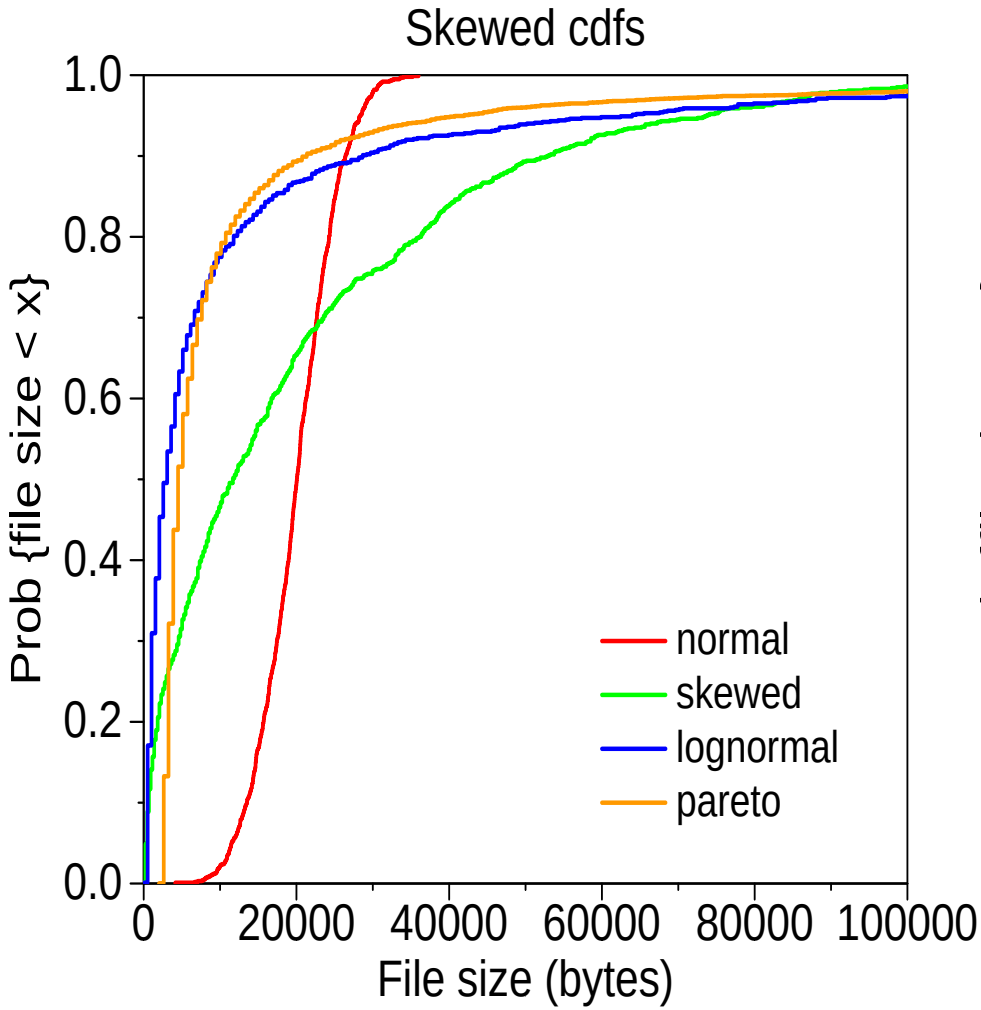
cdf maps values to percentiles

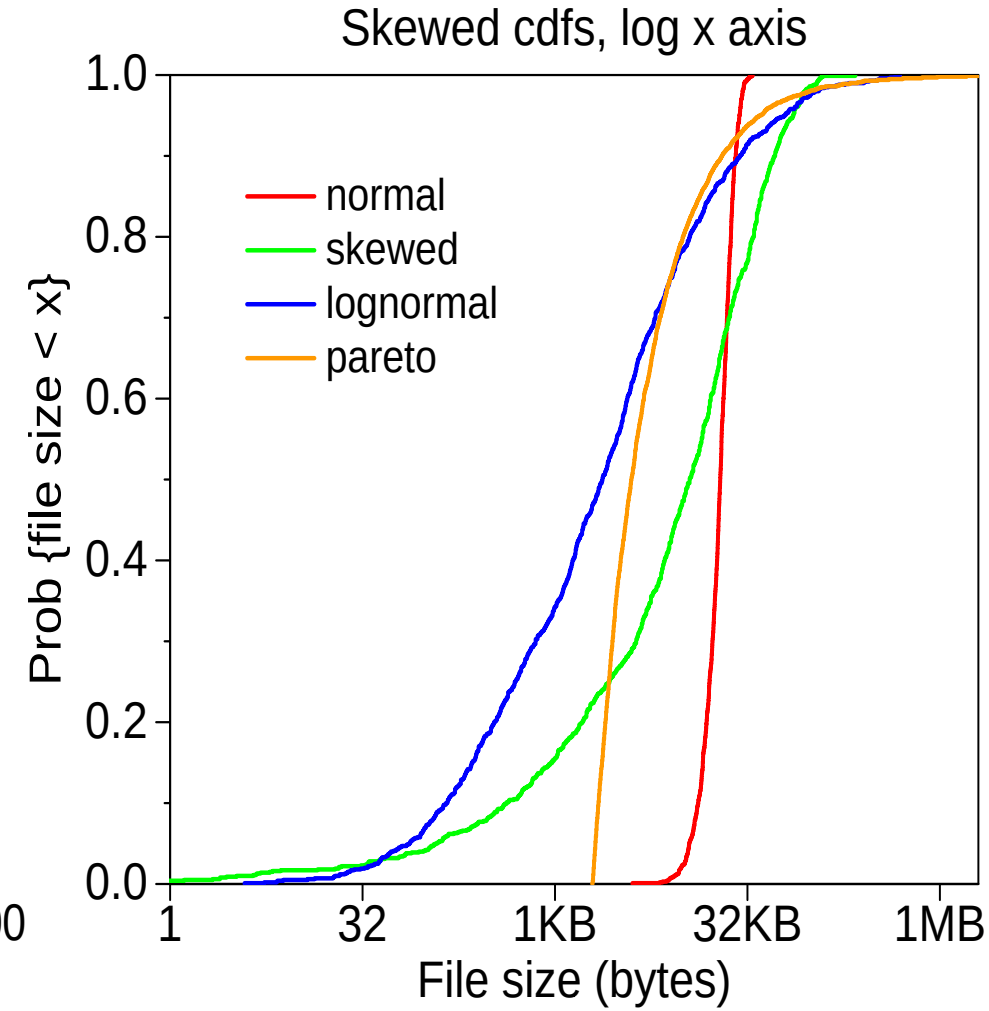
Skewed distributions

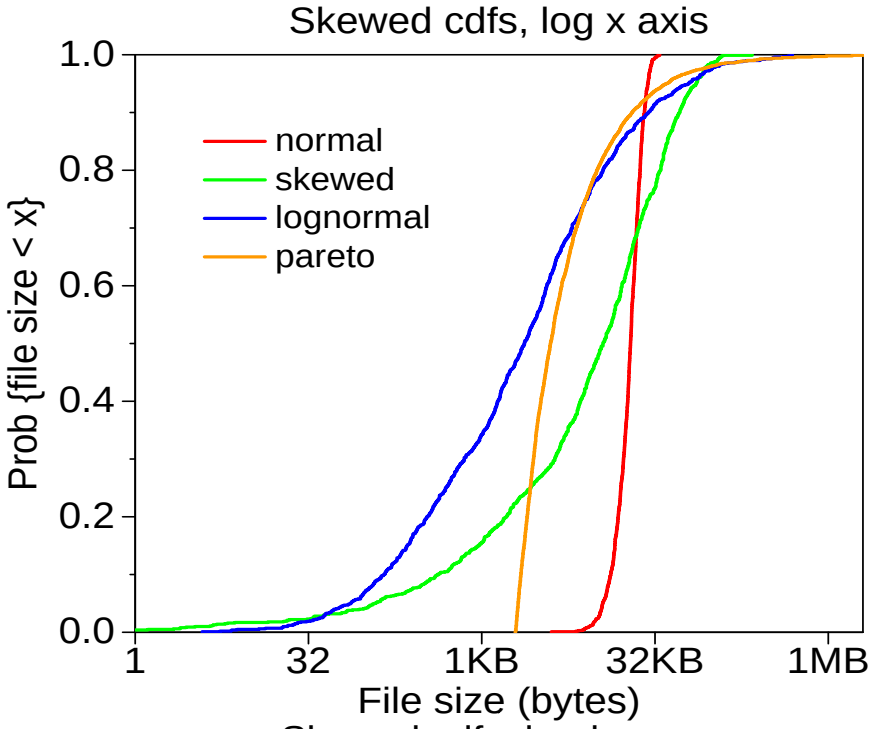


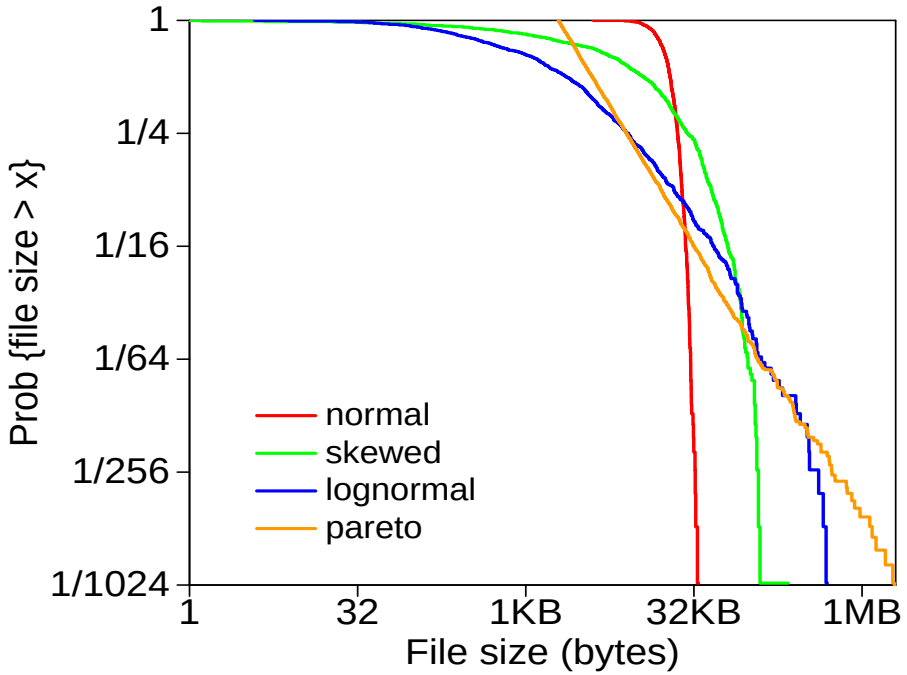
- normal distribution is symmetric.
- skewed has many small values and some large.
- lognormal even more skewed.
- pareto even more skewed.

Logarithmic x axis





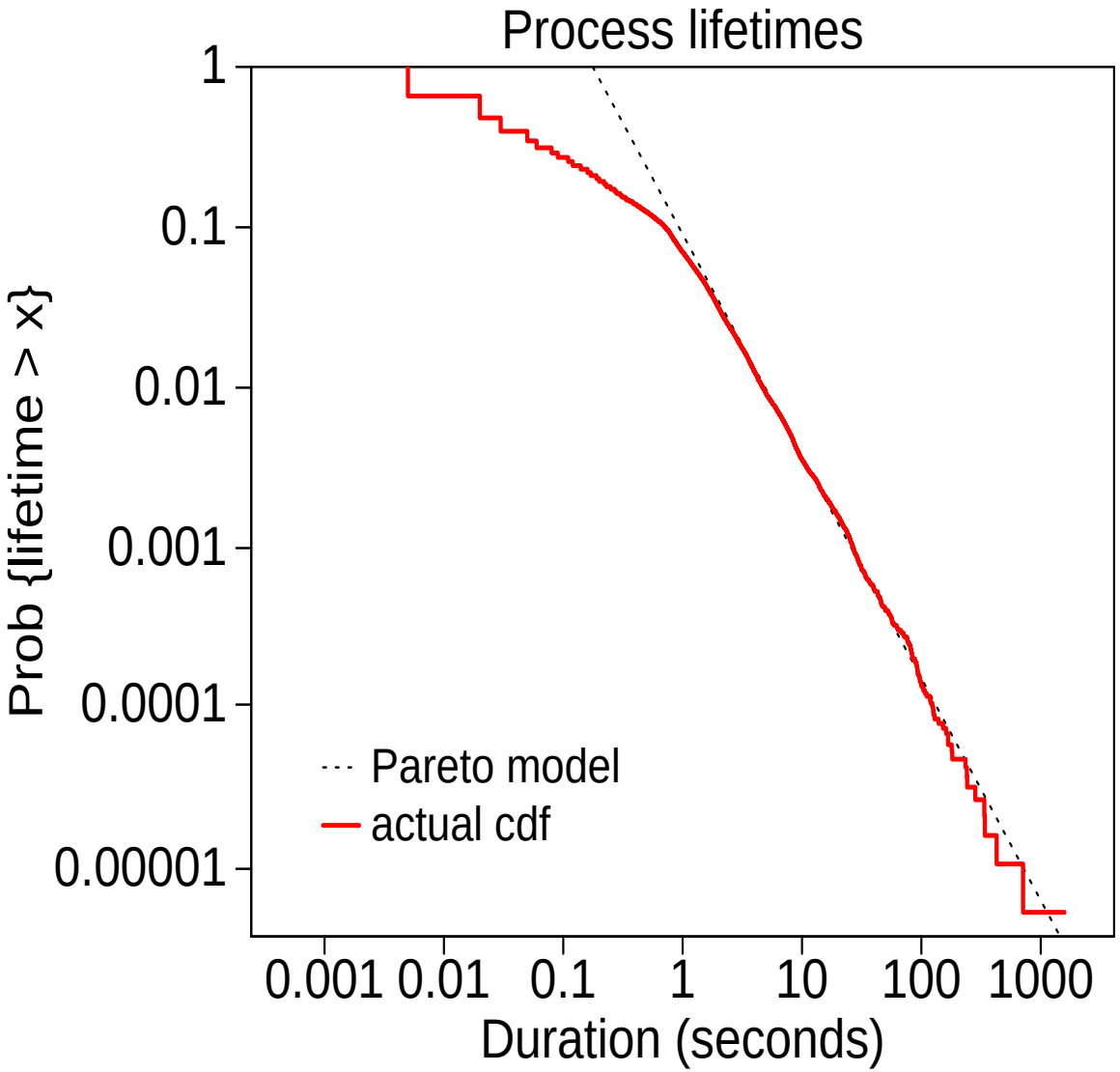




Log-log axes

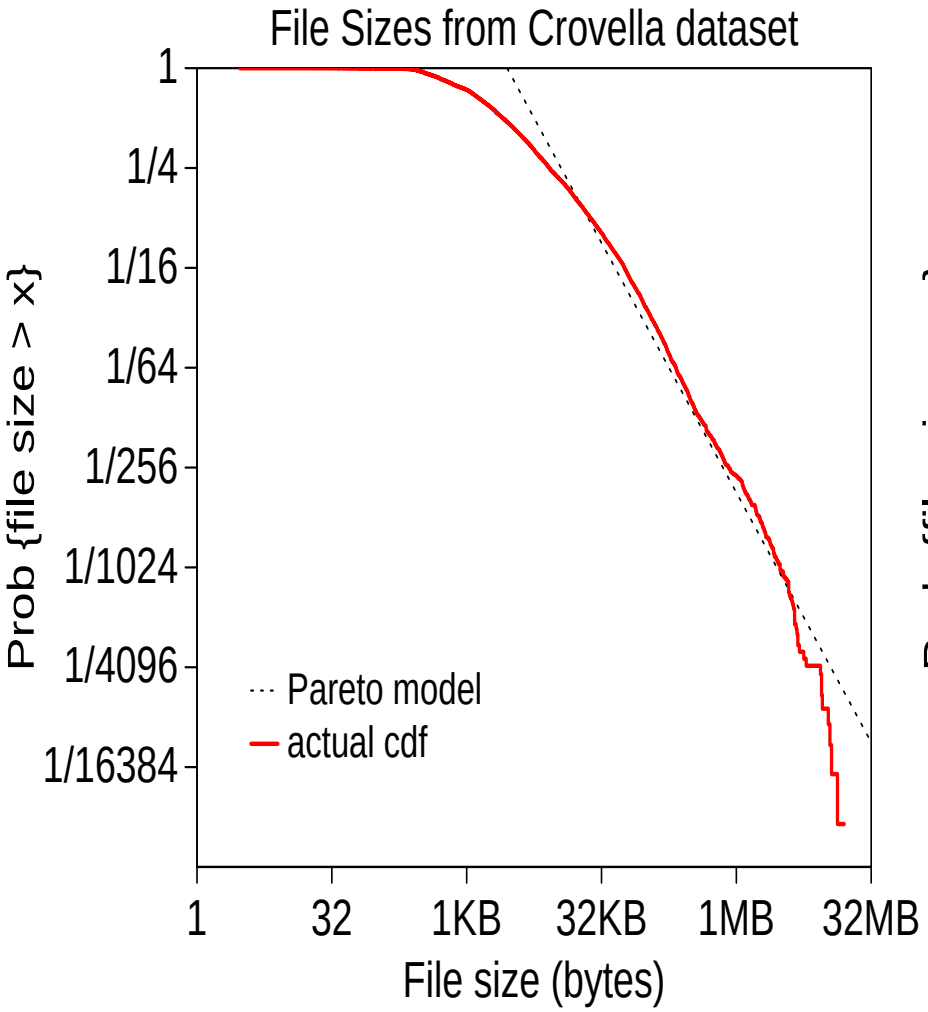
- Complementary cdf: Prob {value > x}
- Log y axis amplifies tail behavior.
- Pareto distribution is a straight line.

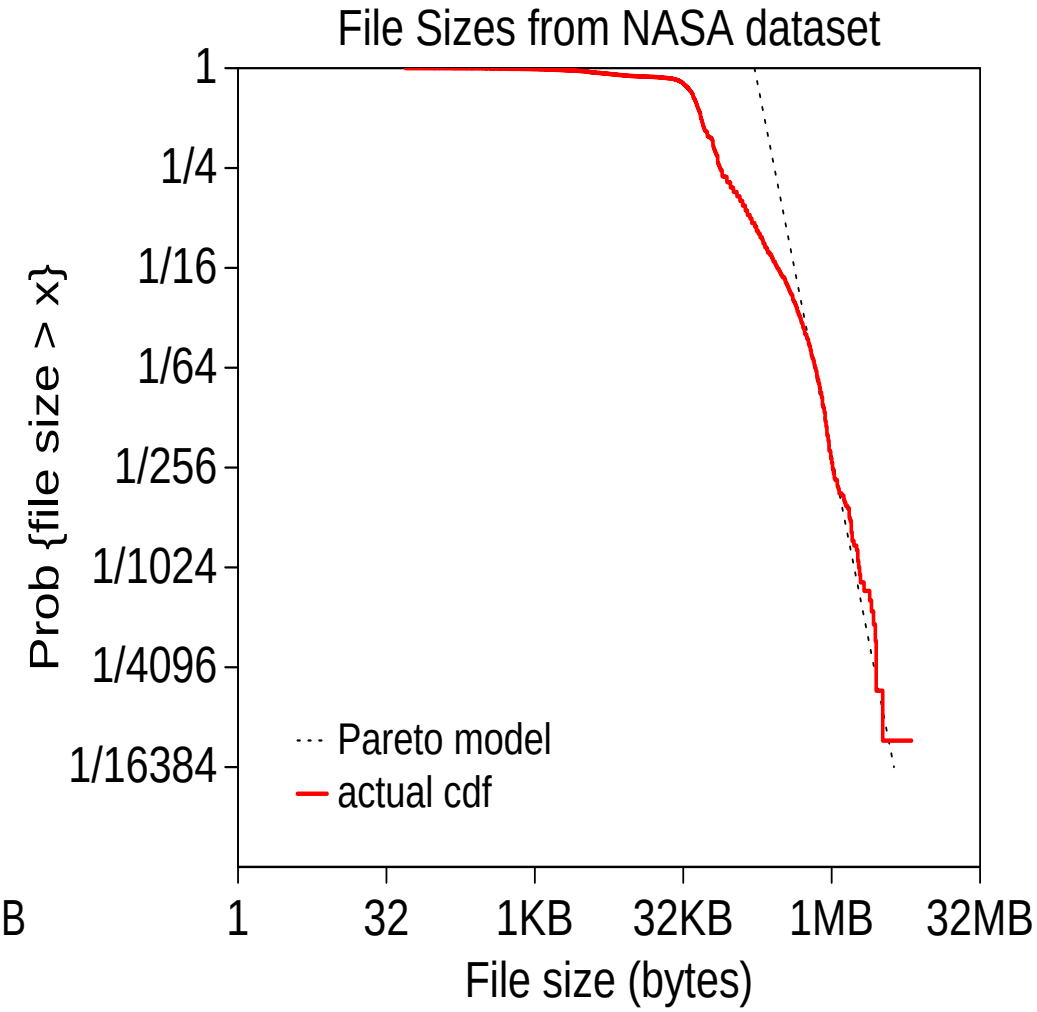
Evidence of long tails



- Is long-tailedness an empirical property?
- Long-tailed dist converges to Pareto.
- How do we know it keeps going?

File sizes in the WWW





Where we are

- Some empirical evidence of long tailed distributions.
- Explanatory model for WWW files.

  [CarlsonDoyle99]

- No explanation for other file systems.

Explanatory model

Goal:

- Model of user behavior that produces long-tailed distributions.

Hypothesis:

- Most new files are copies of old files.
- Many new files are translations of old files.
- New size is a small multiple of the old size.

User Model

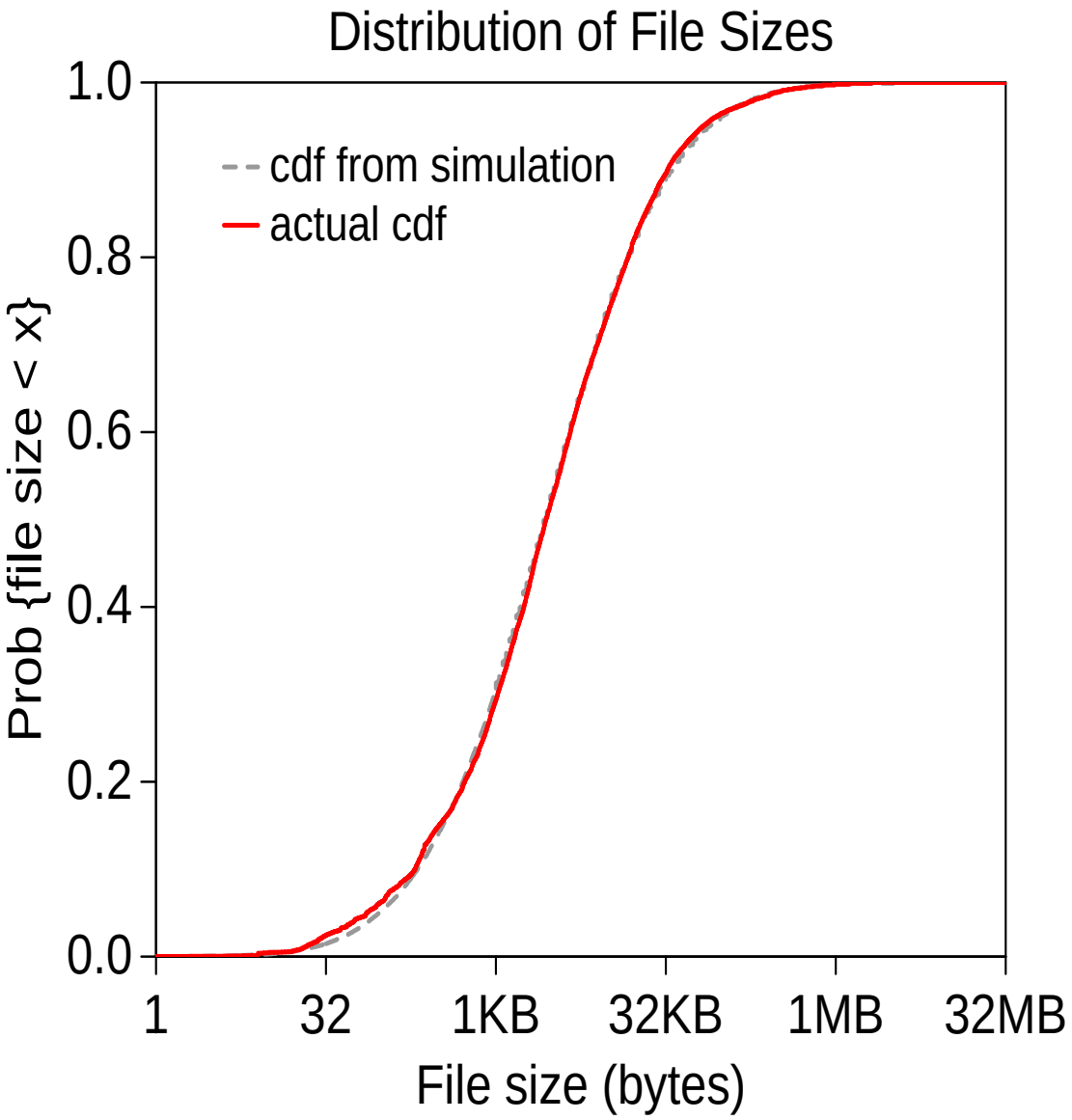
Model:

- Choose an existing file at random.
- Choose a small multiplier at random.
- new file size = old file size * multiplier
- Repeat.

Two parameters:

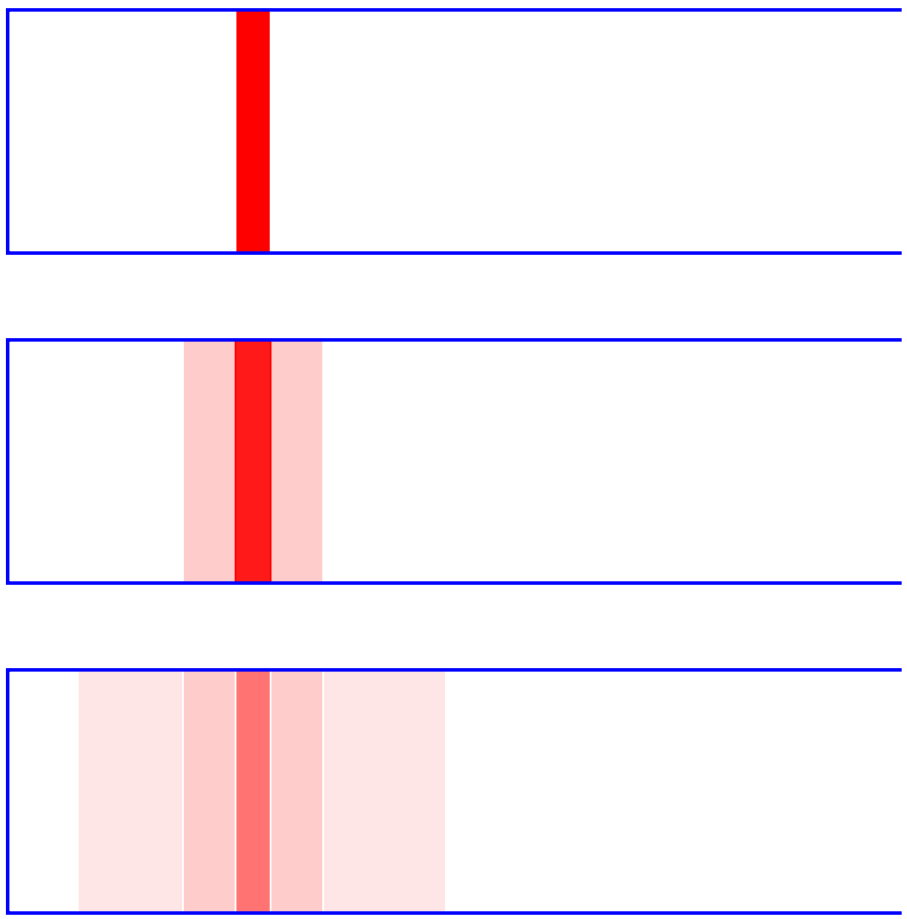
- Initial file size.
- Variability of multipliers.

Simulation of user model



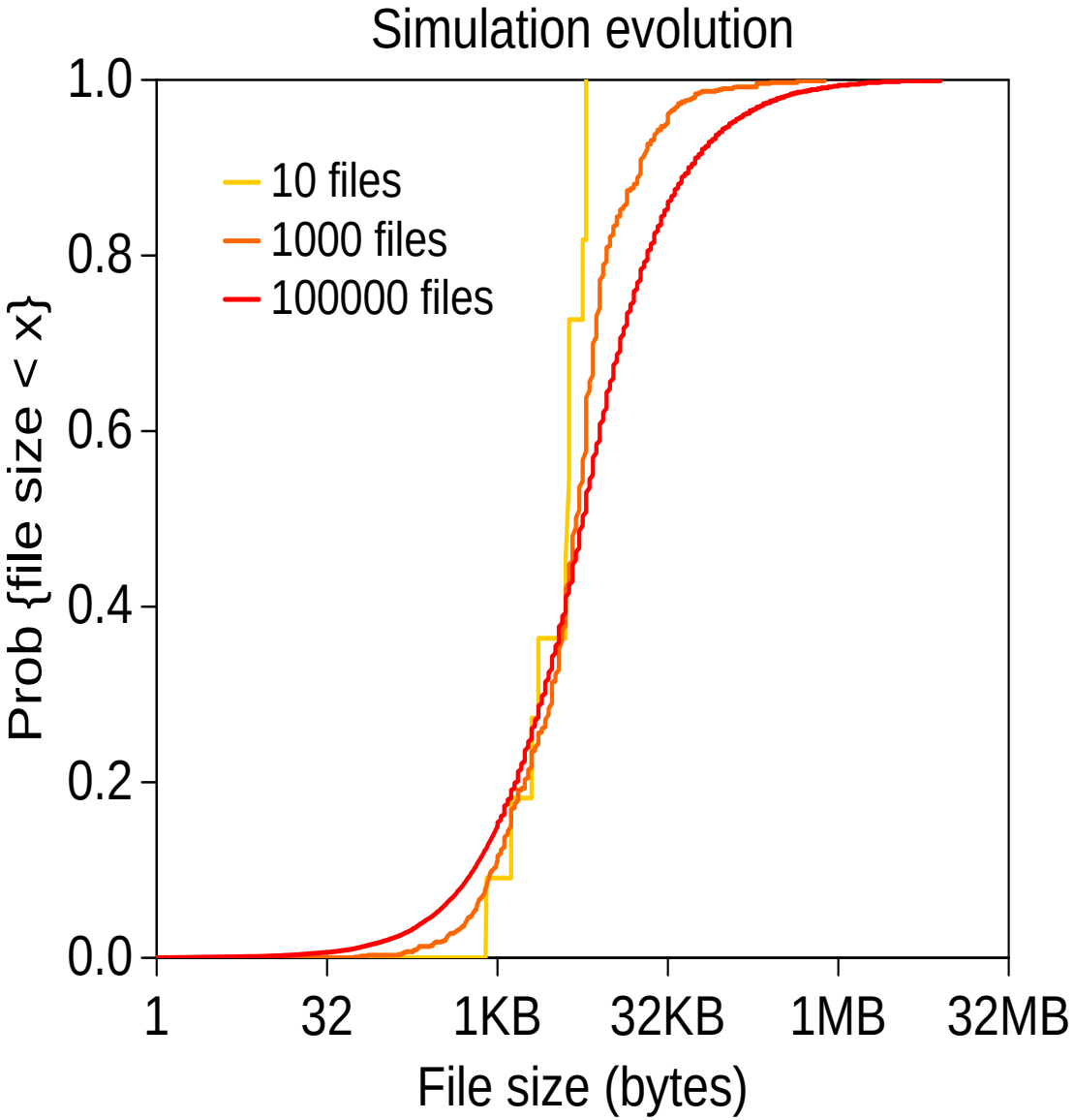
- 89,000 files on rocky.wellesley.edu
- Choose parameters to fit the distribution.
- Fits pretty good!
- Analytic form?

Continuous model



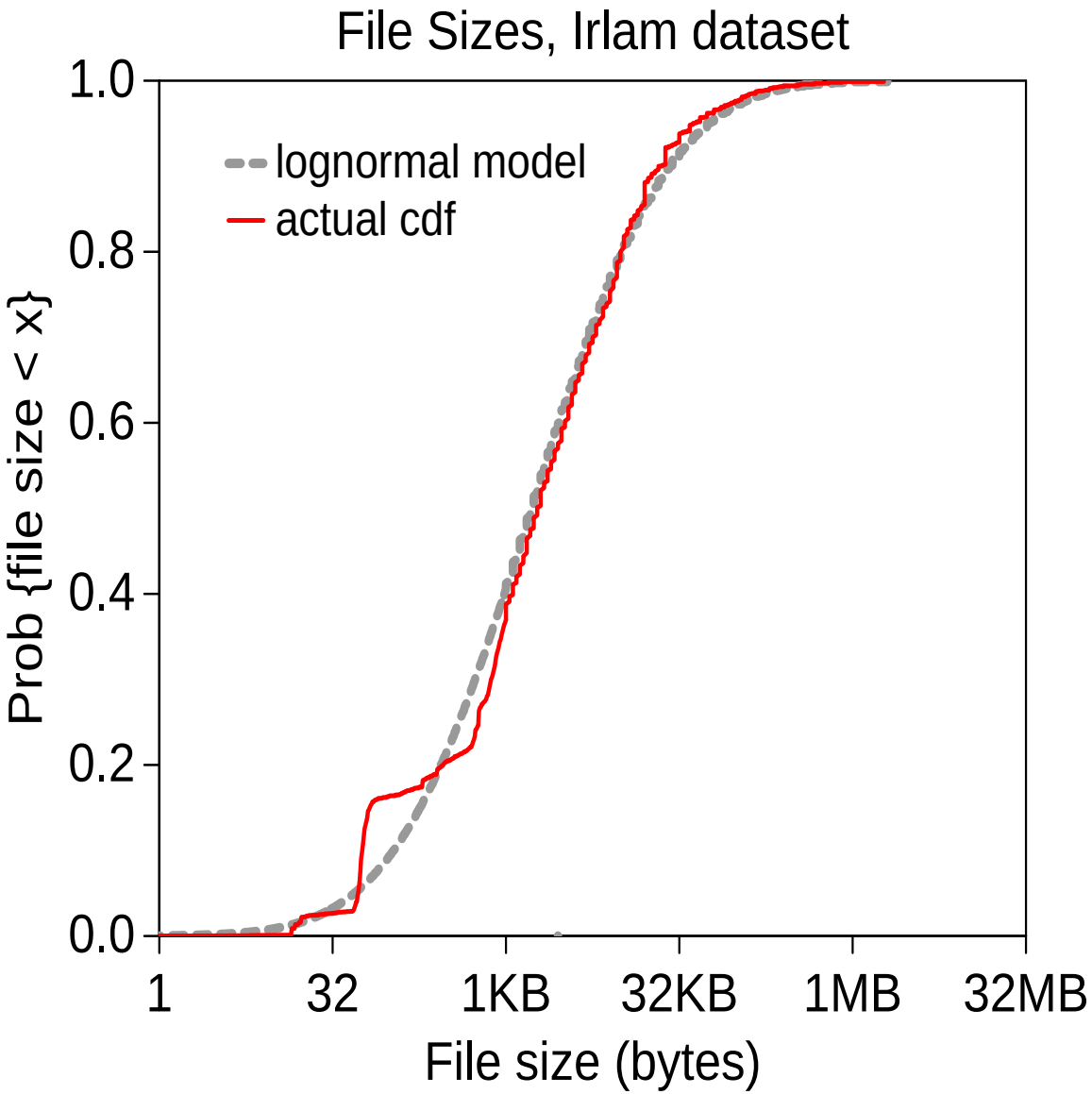
- Replace discrete file sizes with continuous.
- Simulation computes numerical solution of diffusion equation.
- Solution of PDE yields analytic model of the distribution.

Solve that PDE!



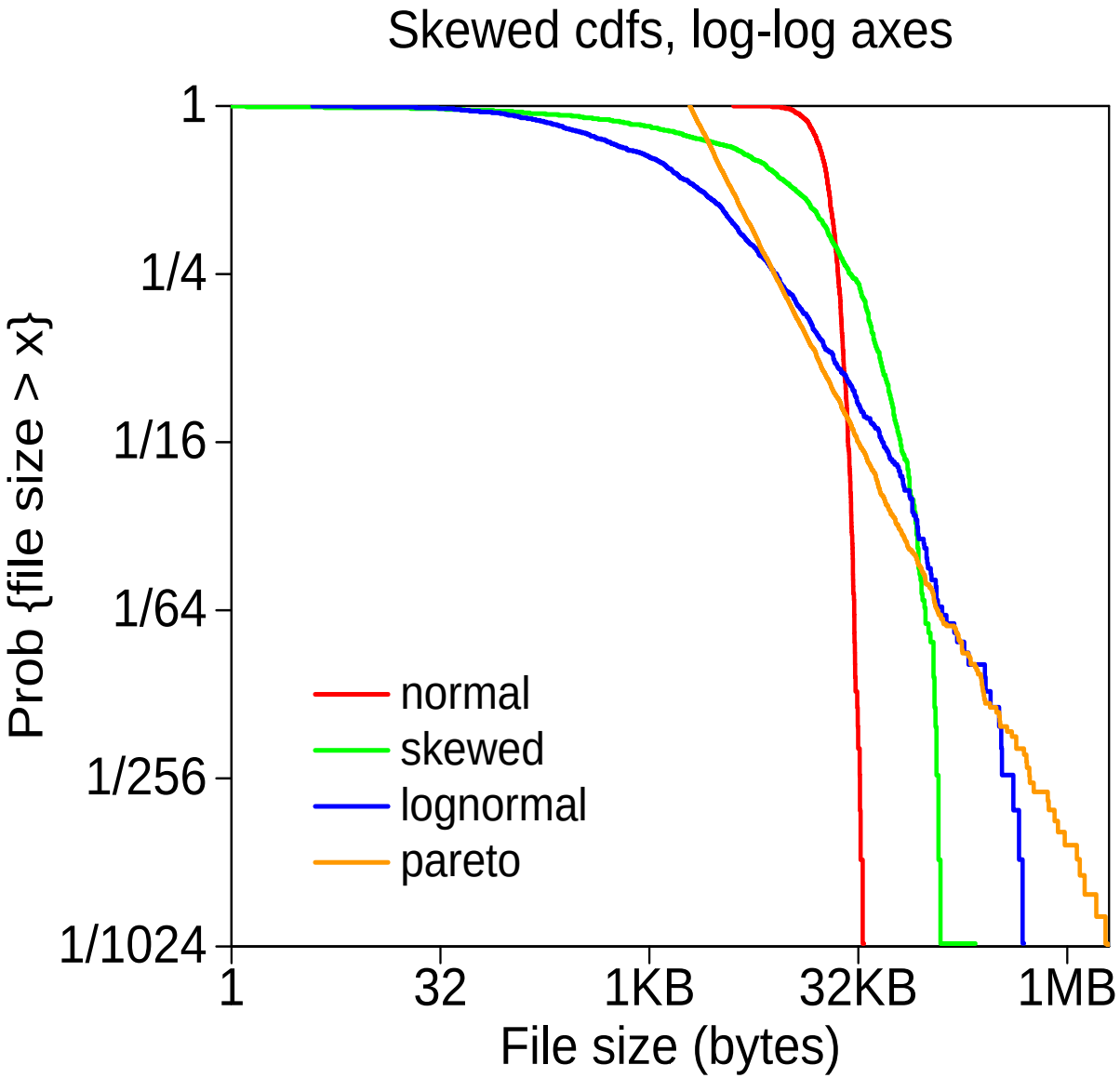
- Distribution of file sizes is normal on a log-x axis: LOGNORMAL.

Estimate those parameters!



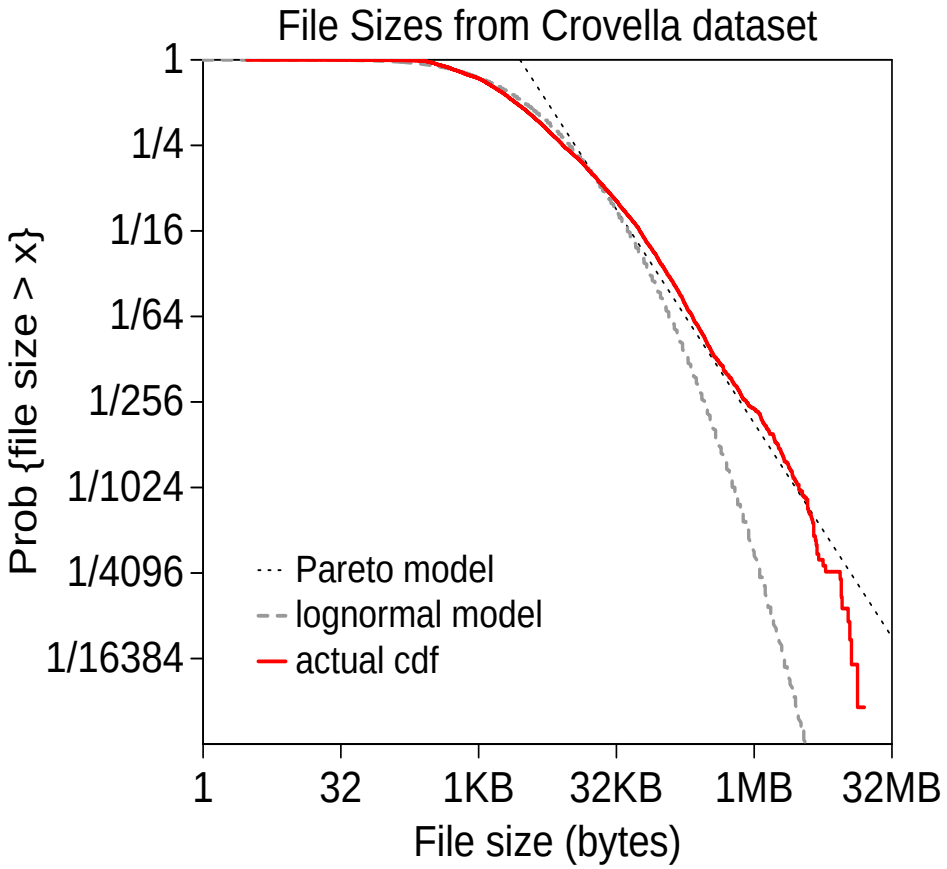
- Irlam collected file sizes from 500+ systems.
- Using the analytic model we can estimate parameters.
- Goodness of fit: Kolmogorov-Smirnov statistic.
- Range: 1.4 to 40
- Median: 8.0

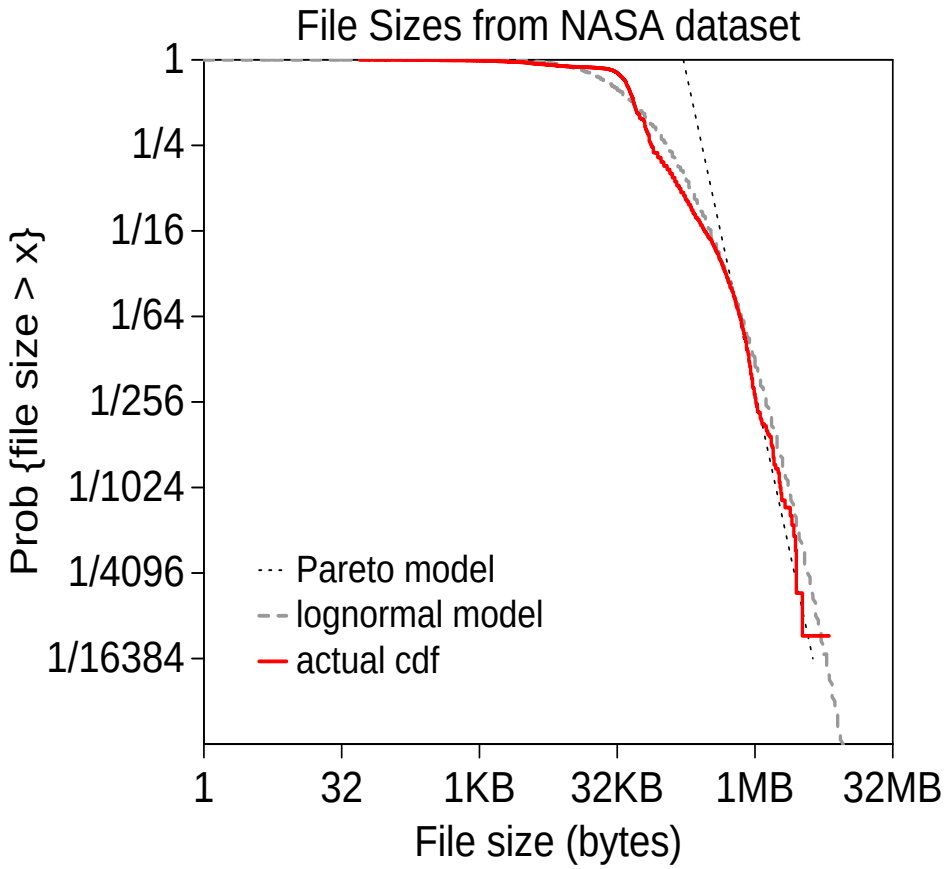
Oh, no!



- The lognormal distribution is not long-tailed.
- Under either aggregation model, lognormal file sizes yield self-similarity over a range of time scales, but not true self-similarity.

Tail behavior?





- To explain self-similarity, we only need a Pareto tail.
- Log-log ccdf amplifies tail.
- Which model is better?

Theory choice

- Accuracy
- Scope
- Consistency
- Simplicity
- Fruitfulness

Kuhn's criteria

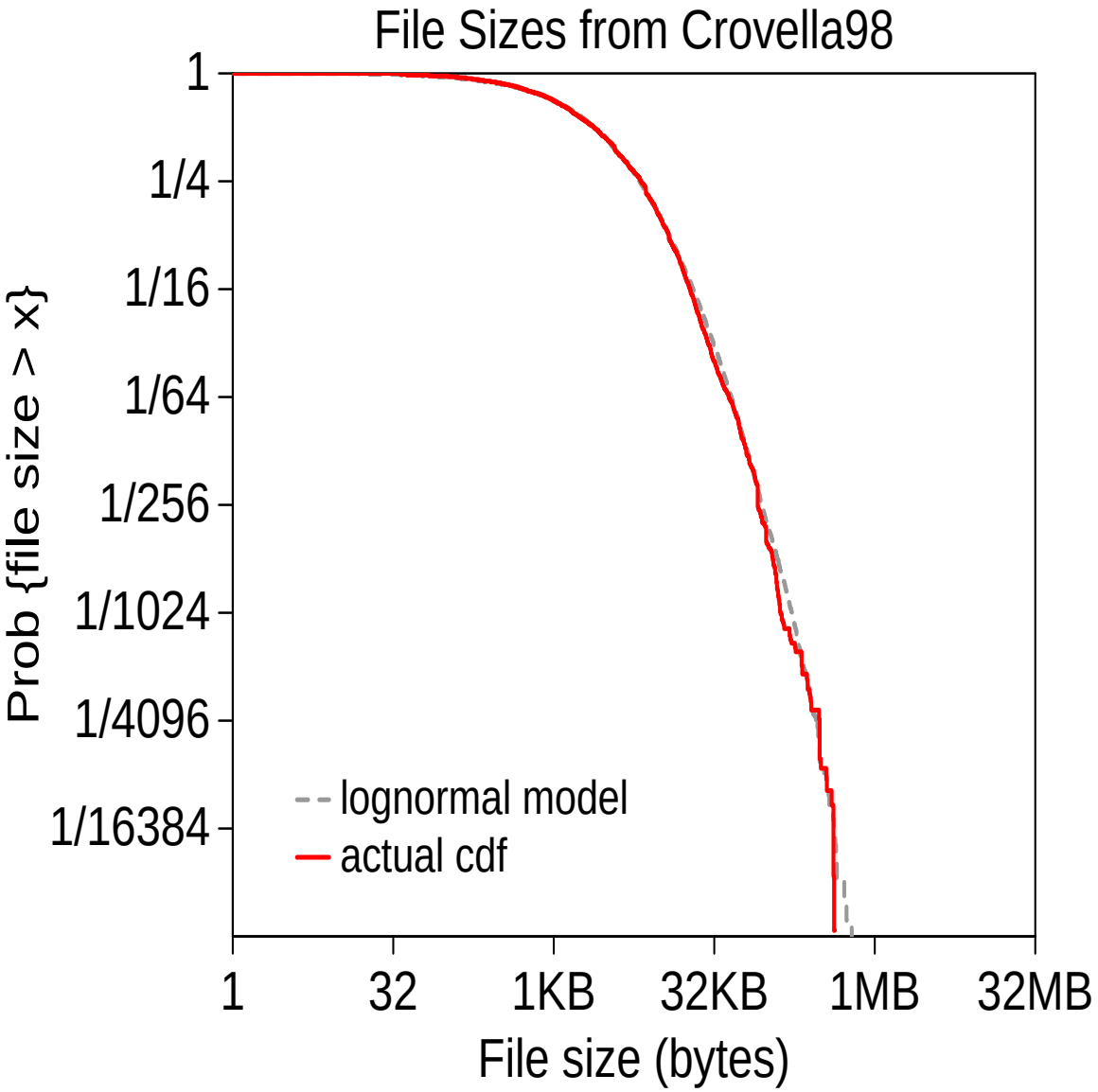
- Explanatory model

one more criterion

Lognormal vs. Pareto

- Accuracy and Scope
  - Diffusion model fits the bulk of the distribution.
  - Pareto model sometimes fits the tail better.
- Consistency
  - Diffusion model undermines self-sim explanation.
- Simplicity
  - Pick ’em.
- Fruitfulness
  - Long-tailed distributions are a nightmare for modelers.
- Explanatory model
  - Carlson and Doyle only explain Web files.
  - I think the diffusion model is more realistic.

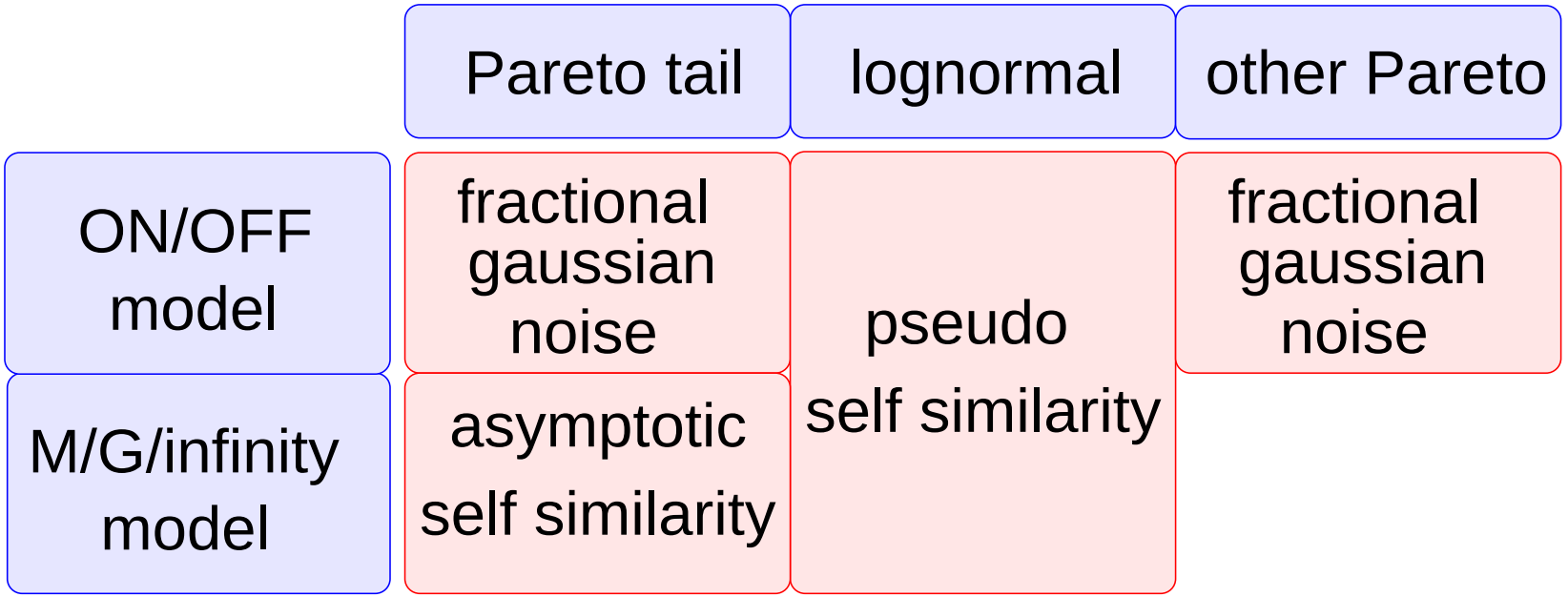
Trade simplicity for accuracy



- What if the primordial soup contained two files?
- Multimodal (5-parameter) lognormal model.
- Accuracy and complexity comparable to Crovella's hybrid model.

Is Internet traffic *really* self-similar?

- What seems to be an empirical question depends on theory choice.
- Theory choice is not determined (entirely) by evidence.



Where does that leave us?

- Realist:
  - There is a real world and we are capable of knowing about it.
  - Rational theory choice is capable of selecting the right theory.
  - The Internet either is or is not really self-similar.
- Instrumentalist:
  - Agnostic about the real world.
  - Our theories are tools that either work or not.
  - If it's useful to model the Internet as self-similar, go ahead.
- Other flavors of anti-realist.

Long-tailed marmot?