The structural cause of file size distributions

Allen B. Downey, Wellesley College

Summary

Background: Current explanations of self-similarity in the Internet are based on the assumption that the distribution of file sizes is long tailed.

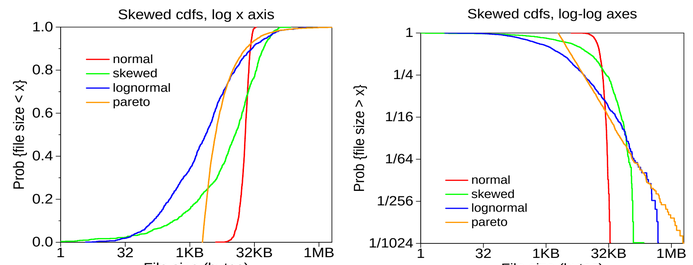
Initial Goal: Find a model of user behavior that explains the origin of long-tailed distributions.

Findings:

- A model of user behavior that yields a lognormal distribution of file sizes.
- Evidence that file sizes may not be long-tailed.

Problem

- The lognormal distribution is not long-tailed.
- For self-similarity, it has to be asymptotic to Pareto.
- CCDF test: plot complementary cdf on log-log axes.
  - Pareto distribution is a straight line.
  - Non-long-tailed falls away with increasing steepness.



User Model

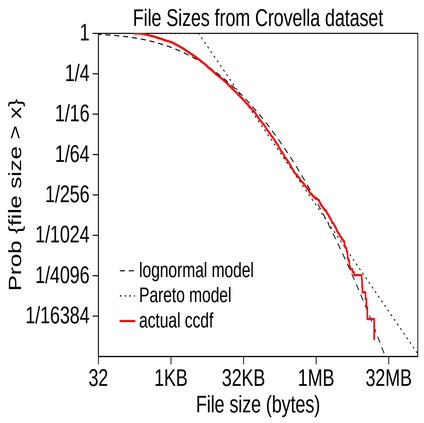
Model:

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

Two parameters:

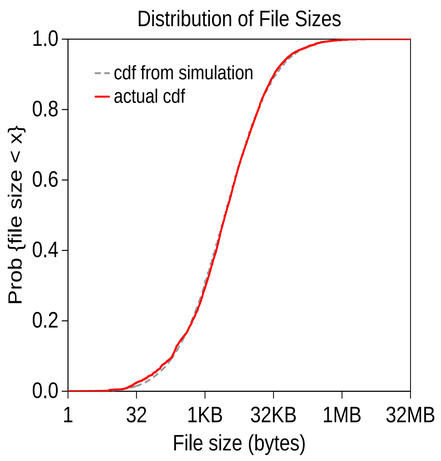
- Initial file size.
- Variability of multipliers.

File sizes on the WWW



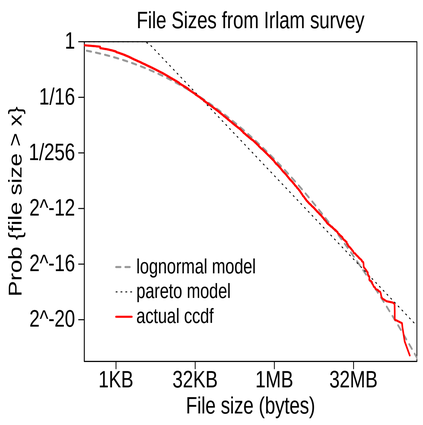
- CrovellaBestavros96 instrumented browsers to log Web downloads.
- Measured sizes of 36000 unique files.
- Fitted Pareto distribution to ccdf; good fit for most of the tail.
- Extreme tail looks more like lognormal.

Simulation of user model



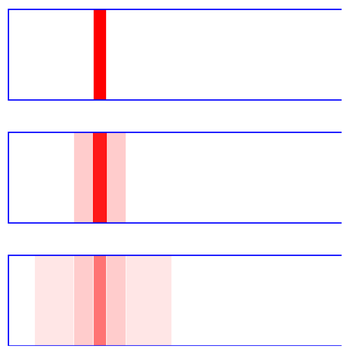
- Measured sizes of 89,000 files from rocky.wellesley.edu
- Chose simulation parameters to match the distribution.
- Simulation results agree with the actual distribution.
- Diffusion model yields analytic form.

File sizes on UNIX systems



- Irlam collected sizes of 2,092,227 files from 656 UNIX file systems.
- The lognormal model fits the ccdf well.
- The Pareto model doesn't.
- A mixture of lognormals with various means (and the same variance) is also lognormal.

Diffusion model



File sizes (on a log axis) diffuse over time like heat in a metal bar.

- The simulation is equivalent to a numerical technique for solving the diffusion equation.
- We guessed that the analytic solution would give the functional form of the distribution.
- The analytic solution is lognormal.

Read the paper

The full paper

- Reviews previous claims of long-tailed distributions, and finds that the preponderance of the evidence supports the lognormal model,
- Proposes a multimodal lognormal model and compares it to the hybrid lognormal-Pareto model,
- Discusses alternate explanations of self-similar Internet traffic, if the distribution of file sizes is not long-tailed.