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]

  \[ x = \text{time in varying units} \]
  \[ y = \text{packets / unit time} \]

- 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

- Why is the distribution of file sizes long-tailed?
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.
- \text{new file size} = \text{old file size} \times \text{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 well!

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
Lognormal model of file sizes

- Lognormal model is
  - (reasonably) accurate,
  - well-behaved,
  - explainable.

- Only one problem:
  It’s not a long-tailed distribution!
Long-tailed distributions

- Definition depends on context
  - For self-similarity, tail behavior is definitive.
  - Tail must be asymptotic to Pareto distribution.

- Why did we think it was long-tailed?

- Review the evidence:
  - percentile-percentile plots
  - aest [CrovellaTaqqu99]
  - complementary cdf on log-log axes
CCDF test

- Complementary cdf: \( \text{Prob} \{ \text{value} > x \} \)
- Log y axis amplifies tail behavior.
- Pareto distribution is a straight line.
- Non-long-tailed falls away with increasing steepness.
File sizes on the WWW

- CrovellaBestavros96 instrumented browsers.
- 36208 unique file names.
- Fitted Pareto distribution to ccdf.
- Carlson and Doyle propose explanatory model (HOT).

![File Sizes from Crovella dataset](image)

- lognormal model
- Pareto model
- actual ccdf

File size (bytes)
Feldmann et al. collected session sizes from an ISP.

They "estimate the slope of the corresponding linear regions."
Server’s view

- Arlitt and Williamson collected unique files served by web servers:
  - University of Saskatchewan
  - NASA’s Kennedy Center
  - ClarkNet (an ISP)
  - NCSA
- Hard to characterize these datasets.
- This one looks lognormal...
Server’s view

File Sizes from Saskatchewan dataset

- lognormal model
- pareto model
- actual ccdf

- ...but this one looks Pareto (sort of).
- Increasing slope in extreme tail?
The Pareto model is a better fit.

But the shape matches the lognormal model.

Methodology?

- Estimate parameters, evaluate goodness of fit.
- How do we evaluate overall behavior?
Arlitt and Jin measured 20728 files on World Cup site.

Some site-specific features.

Hard to characterize.
Arlitt et al. measured 16 million unique HTML files from a proxy server.

- Top figure shows lognormal model (cdf on log-x axis).
- Bottom figure shows Pareto model (ccdf on log-log axes).
- Tail behavior characteristic of non-long-tailed dist.
Where are we?

- Some evidence for Pareto model.
- Preponderance for lognormal model.
- Good news for modelers.
- Not terribly satisfying as an explanation.
Arlitt et al. and Barford et al. proposed:

- Bulk of distribution is lognormal.
- Tail behavior is Pareto.
- Good match for the bulk and the tail.
- 4-5 parameters.
Multimodal model

- Extend lognormal model to two modes.
- 5 parameters (found by search to minimize K-S stat).
- Better fit for tail behavior.
Multimodal model

File Sizes from Saskatchewan dataset

- Multimodal lognormal handles problem cases.
- Long-tailed model is not necessary.
Theory choice

- Accuracy
- Scope
- Consistency
- Simplicity
- Fruitfulness

Explanatory model

Kuhn’s criteria

one more criterion
Lognormal vs. Pareto

• **Accuracy and Scope**
  • Lognormal model fits the bulk of the distribution.
  • Pareto model sometimes fits the tail better.

• **Consistency**
  • Lognormal 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.
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.

<table>
<thead>
<tr>
<th>Model</th>
<th>Pareto tail</th>
<th>lognormal</th>
</tr>
</thead>
<tbody>
<tr>
<td>ON/OFF model</td>
<td>fractional gaussian noise</td>
<td>pseudo self similarity</td>
</tr>
<tr>
<td>M/G/infinity model</td>
<td>asymptotic self similarity</td>
<td></td>
</tr>
</tbody>
</table>
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.

(recognizing differences in philosophic disposition can forestall fruitless argument)