A Brief History of Lognormal and Power Law Distributions and an Application to File Size Distributions

Michael Mitzenmacher

Harvard University

Motivation: General

- Power laws now everywhere in computer science.
  - See the popular texts *Linked* by Barabasi or *Six Degrees* by Watts.
  - File sizes, download times, Internet topology, Web graph, etc.
- Other sciences have known about power laws for a long time.
  - Economics, physics, ecology, linguistics, etc.
- We should know history before diving in.

Motivation: Specific

- Recent work on file size distributions
  - Downey (2001): file sizes have lognormal distribution (model and empirical results).
  - Barford et al. (1999): file sizes have lognormal body and Pareto (power law) tail. (empirical)
- Understanding file sizes important for
  - Simulation tools: SURGE
  - Explaining network phenomena: power law for file sizes may explain self-similarity of network traffic.
- Wanted to settle discrepancy.
- Found rich (and insufficiently cited) history.
- Helped lead to new file size model.

Power Law Distribution

- A power law distribution satisfies

$$\Pr[X \geq x] \sim cx^{-\alpha}$$

- Pareto distribution

$$\Pr[X \geq x] = \left({x}/{k} \right)^{-\alpha}$$

  - Log-complementary cumulative distribution function (ccdf) is exactly linear.

$$\ln \Pr[X \geq x] = -\alpha \ln x + \alpha \ln k$$

- Properties
  - Infinite mean/variance possible

Lognormal Distribution

- X is lognormally distributed if $Y = \ln X$ is normally distributed.
- Density function: $f(x) = \dfrac{1}{\sqrt{2\pi}\sigma x} e^{-(\ln x - \mu)^2 / 2\sigma^2}$
- Properties:
  - Finite mean/variance.
  - Skewed: mean > median > mode
  - Multiplicative: X_1 lognormal, X_2 lognormal implies $X_1 X_2$ lognormal.

Similarity

- Easily seen by looking at log-densities.
- Pareto has linear log-density.

$$\ln f(x) = -(\alpha - 1)\ln x + \alpha \ln k + \ln \alpha$$

- For large σ, lognormal has nearly linear log-density.

$$\ln f(x) = -\ln x - \ln \sqrt{2\pi}\sigma - \frac{(\ln x - \mu)^2}{2\sigma^2}$$

- Similarly, both have near linear log-ccdfs.
  - Log-ccdfs usually used for empirical, visual tests of power law behavior.
- Question: how to differentiate them empirically?

Lognormal vs. Power Law

- Question: Is this distribution lognormal or a power law?
  - Reasonable follow-up: Does it matter?
- Primarily in economics
  - Income distribution.
  - Stock prices. (Black-Scholes model.)
- But also papers in ecology, biology, astronomy, etc.

History

- Power laws
  - Pareto : income distribution, 1897
  - Zipf-Auerbach: city sizes, 1913/1940’s
  - Zipf-Estouf: word frequency, 1916/1940’s
  - Lotka: bibliometrics, 1926
  - Mandelbrot: economics/information theory, 1950’s+
- Lognormal
  - McAlister, Kapetyn: 1879, 1903.
  - Gibrat: multiplicative processes, 1930’s.

Generative Models: Power Law

- Preferential attachment
  - Dates back to Yule (1924), Simon (1955).
    - Yule: species and genera.
    - Simon: income distribution, city population distributions, word frequency distributions.
  - Web page degrees: more likely to link to page with many links.
- Optimization based
  - Mandelbrot (1953): optimize information per character.
  - HOT model for file sizes. Zhu et al. (2001)

Preferential Attachment

- Consider dynamic Web graph.
  - Pages join one at a time.
  - Each page has one outlink.
- Let $X_j(t)$ be the number of pages of degree j at time t.
- New page links:
  - With probability α, link to a random page.
  - With probability $(1- \alpha)$, a link to a page chosen proportionally to indegree. (Copy a link.)

Simple Analysis

$$\frac{dX_0}{dt} = 1 - \alpha \frac{X_0}{t}$$

$$\frac{dX_j}{dt} = \alpha \frac{X_{j-1}}{t} - \alpha \frac{X_j}{t} + (1-\alpha)(j-1)\frac{X_{j-1}}{t} - (1-\alpha) j \frac{X_j}{t}$$

- Assume limiting distribution where $X_j = c_j t$

$$\frac{c_j}{c_{j-1}} \sim 1 - \frac{2-\alpha}{1-\alpha}\frac{1}{j}$$

$$c_j \sim j^{-(2-\alpha)/(1-\alpha)}$$

Optimization Model: Power Law

- Mandelbrot experiment: design a language over a d-ary alphabet to optimize information per character.
  - Probability of jth most frequently used word is p_j.
  - Length of jth most frequently used word is c_j.
- Average information per word:

$$H = -\sum_j p_j \log_2 p_j$$

- Average characters per word:

$$C = \sum_j p_j c_j$$

Optimization Model: Power Law

- Optimize ratio $A = C/H$.

$$H = -\sum_j p_j \log_2 p_j \qquad\qquad C = \sum_j p_j c_j$$

$$dA \big/ dp_j = \left(c_j H + C \log_2(e p_j)\right) \big/ H^2$$

$$dA \big/ dp_j = 0 \text{ when } p_j = 2^{-Hc_j/C} / e$$

If $c_j \approx \log_d j$, power law results.

Monkeys Typing Randomly

- Miller (psychologist, 1957) suggests following: monkeys type randomly at a keyboard.
  - Hit each of n characters with probability p.
  - Hit space bar with probability $1 - np > 0$.
  - A word is sequence of characters separated by a space.
- Resulting distribution of word frequencies follows a power law.
- Conclusion: Mandelbrot's "optimization" not required for languages to have power law

Miller's Argument

- All words with k letters appear with prob.

$$p^{k}(1-pn)$$

- There are n^k words of length k.
  - Words of length k have frequency ranks

$$\left[1+\left(n^{k}-1\right)/\left(n-1\right),\left(n^{k+1}-1\right)/\left(n-1\right)\right]$$

- Manipulation yields power law behavior

$$p^{\log_N j+1}(1-np) \le p_j \le p^{\log_N j}(1-np)$$

- Recently extended by Conrad, Mitzenmacher to case of unequal letter probabilities.
  - Non-trivial: requires complex analysis.

Generative Models: Lognormal

- Start with an organism of size X_0.
- At each time step, size changes by a random multiplicative factor.

$$X_t = F_{t-1} X_{t-1}$$

- If F_t is taken from a lognormal distribution, each X_t is lognormal.
- If F_t are independent, identically distributed then (by CLT) X_t converges to lognormal distribution.

BUT!

- If there exists a lower bound:

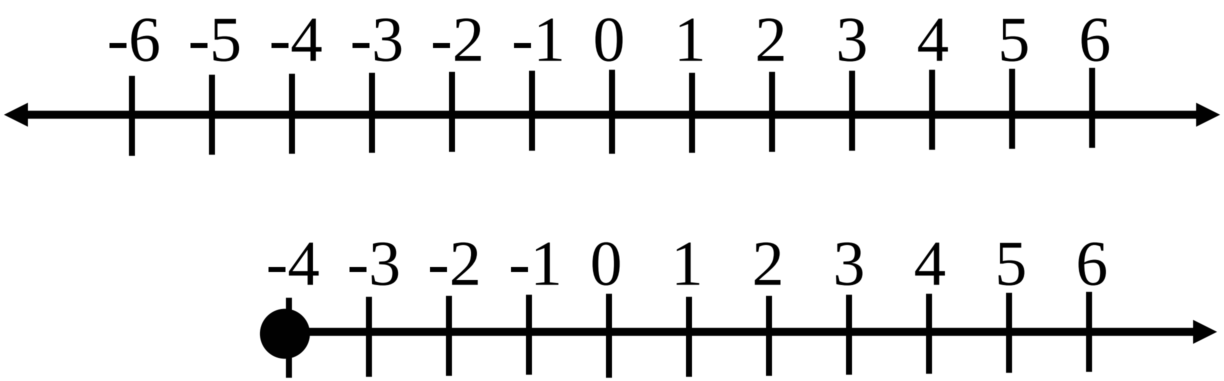
$$X_t = \max(\varepsilon, F_{t-1} X_{t-1})$$

then X_t converges to a power law distribution. (Champernowne, 1953)

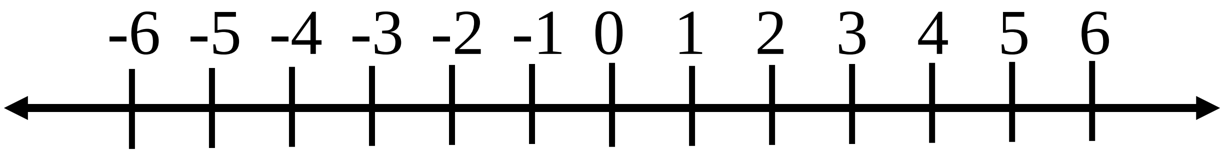
- Lognormal model easily pushed to a power law model.

Example

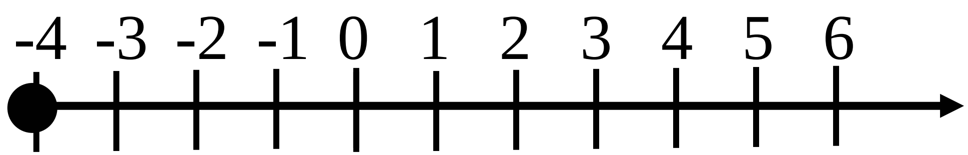
- At each time interval, suppose size either increases by a factor of 2 with probability 1/3, or decreases by a factor of 1/2 with probability 2/3.
  - Limiting distribution is lognormal.
  - But if size has a lower bound, power law.



Example continued



- After *n* steps distribution increases - decreases becomes normal (CLT).



- Limiting distribution:

$$\Pr[X \geq x] \sim 2^{-x} \Rightarrow \Pr[\text{size} \geq x] \sim 1/x$$

Double Pareto Distributions

- Consider continuous version of lognormal generative model.
  - At time t, log X_t is normal with mean μt and variance $\sigma^2 t$
- Suppose observation time is randomly distributed.
  - Income model: observation time depends on age, generations in the country, etc.

Double Pareto Distributions

- Reed (2000,2001) analyzes case where time distributed exponentially.

$$f(x) = \int_{t=0}^{\infty} \lambda e^{-\lambda t} \frac{1}{\sqrt{2\pi}\sigma x t} e^{-(\ln x - \mu t)^2 / 2\sigma^2 t} dt$$

  – Also Adamic, Huberman (1999).

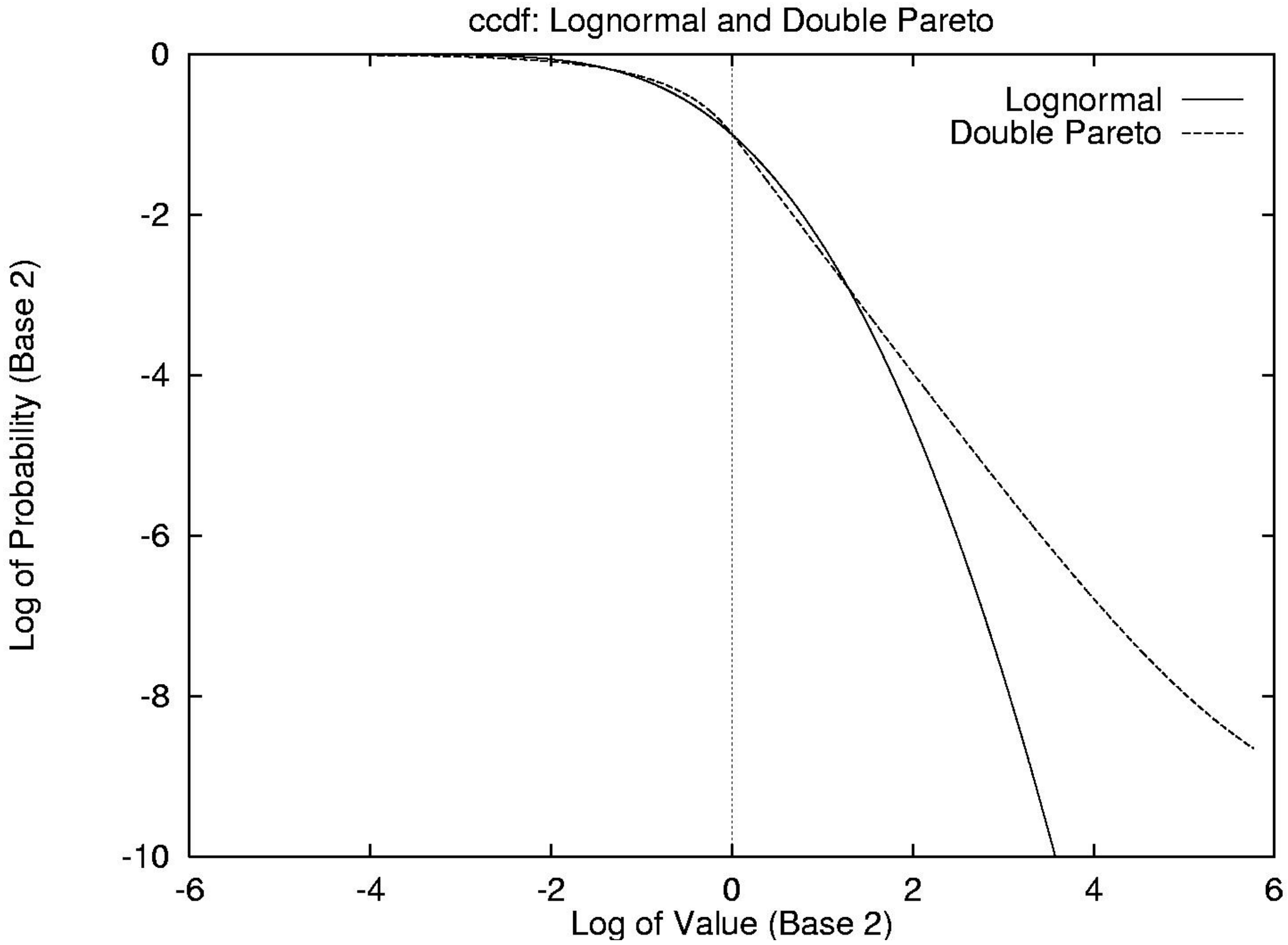
- Simplest case: $\mu = 0, \ \sigma = 1$

$$f(x) = \begin{cases} \sqrt{\lambda/2}\, x^{-1-\sqrt{2\lambda}} & \text{for } x \geq 1 \\ \sqrt{\lambda/2}\, x^{-1+\sqrt{2\lambda}} & \text{for } x \leq 1 \end{cases}$$

Double Pareto Behavior

- Double Pareto behavior, density
  - On log-log plot, density is two straight lines
  - Between lognormal (curved) and power law (one line)
- Can have lognormal shaped body, Pareto tail.
  - The ccdf has Pareto tail; linear on log-log plots.
  - But cdf is also linear on log-log plots.

Lognormal vs. Double Pareto



Double Pareto File Sizes

- Reed used Double Pareto to explain income distribution
  - Appears to have lognormal body, Pareto tail.
- Double Pareto shape closely matches empirical file size distribution.
  - Appears to have lognormal body, Pareto tail.
- Is there a reasonable model for file sizes that yields a Double Pareto Distribution?

Downey’s Ideas

- Most files derived from others by copying, editing, or filtering.
- Start with a single file.
- Each new file derived from old file.

$$\text{New file size} = F \times \text{Old file size}$$

- Like lognormal generative process.
  - Individual file sizes converge to lognormal.

Problems

- “Global” distribution not lognormal.
  - Mixture of lognormal distributions.
- Everything derived from single file.
  - Not realistic.
  - Large correlation: one big file near root affects everybody.
- Deletions not handled.

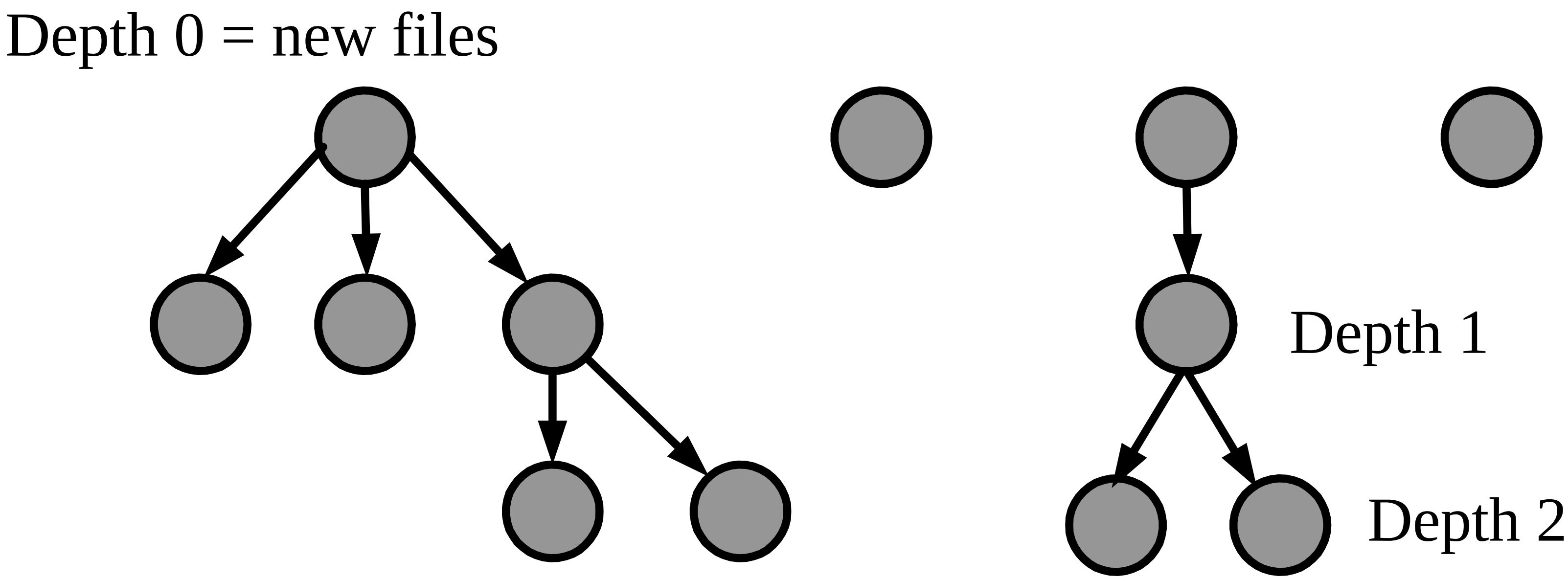
Recursive Forest File Size Model

- Keep Downey's basic process.
- At each time step, either
  - Completely new file generated (prob. p), with distribution F_1 or
  - New file is derived from old file (prob. $1 - p$):

$$\text{New file size} = F_2 \times \text{Old file size}$$

- Simplifying assumptions.
  - Distribution $F_1 = F_2 = F$ is lognormal.
  - Old file chosen uniformly at random.

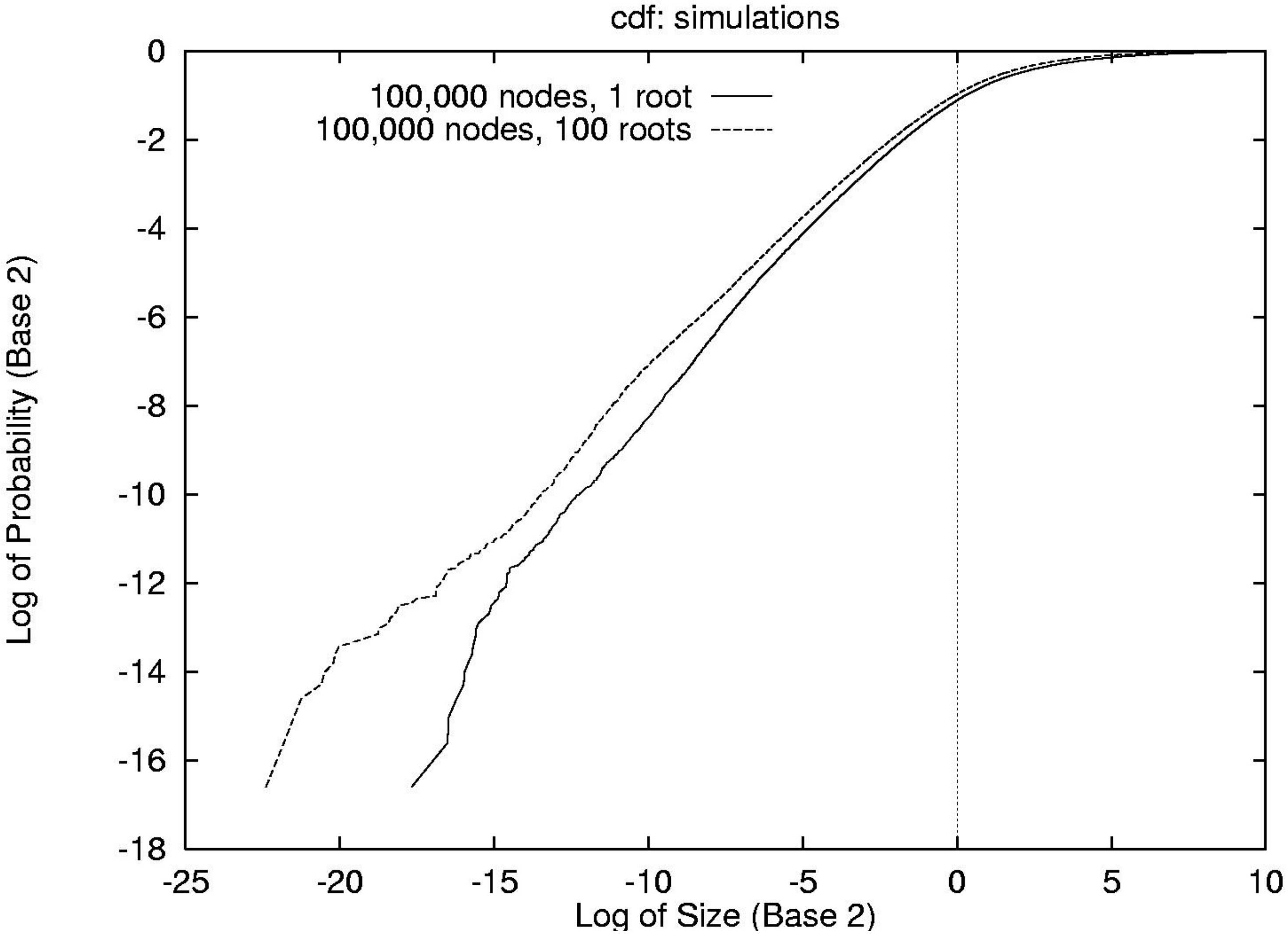
Recursive Forest



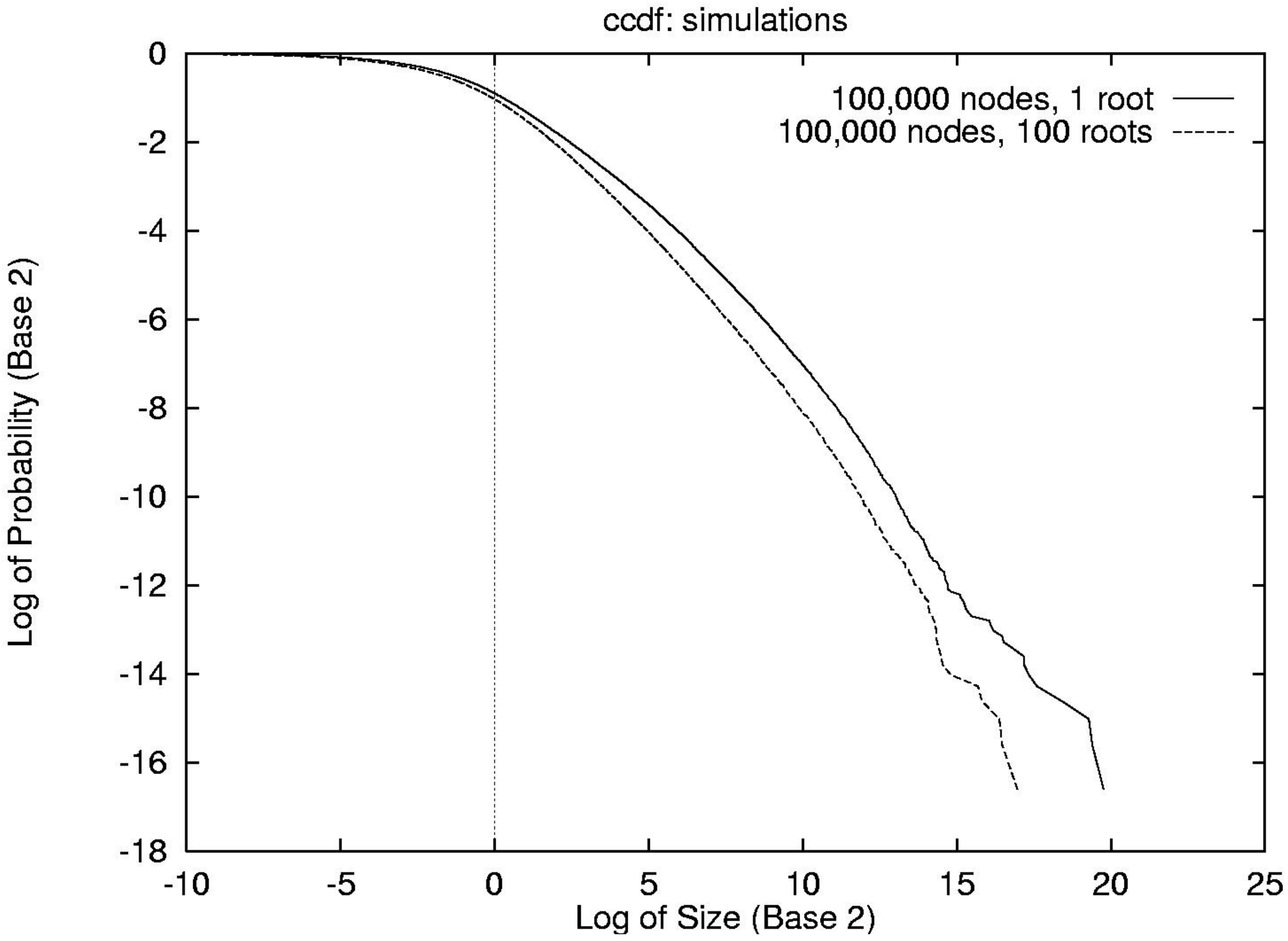
Depth Distribution

- Node depths have geometric distribution.
  - # Depth 0 nodes converge to pt; depth 1 nodes converge to $p(1-p)t$, etc.
  - So number of multiplicative steps is geometric.
  - Discrete analogue of exponential distribution of Reed’s model.
- Yields Double Pareto file size distribution.
  - File chosen uniformly at random has almost exponential number of time steps.
  - Lognormal body, heavy tail.
  - But no nice closed form.

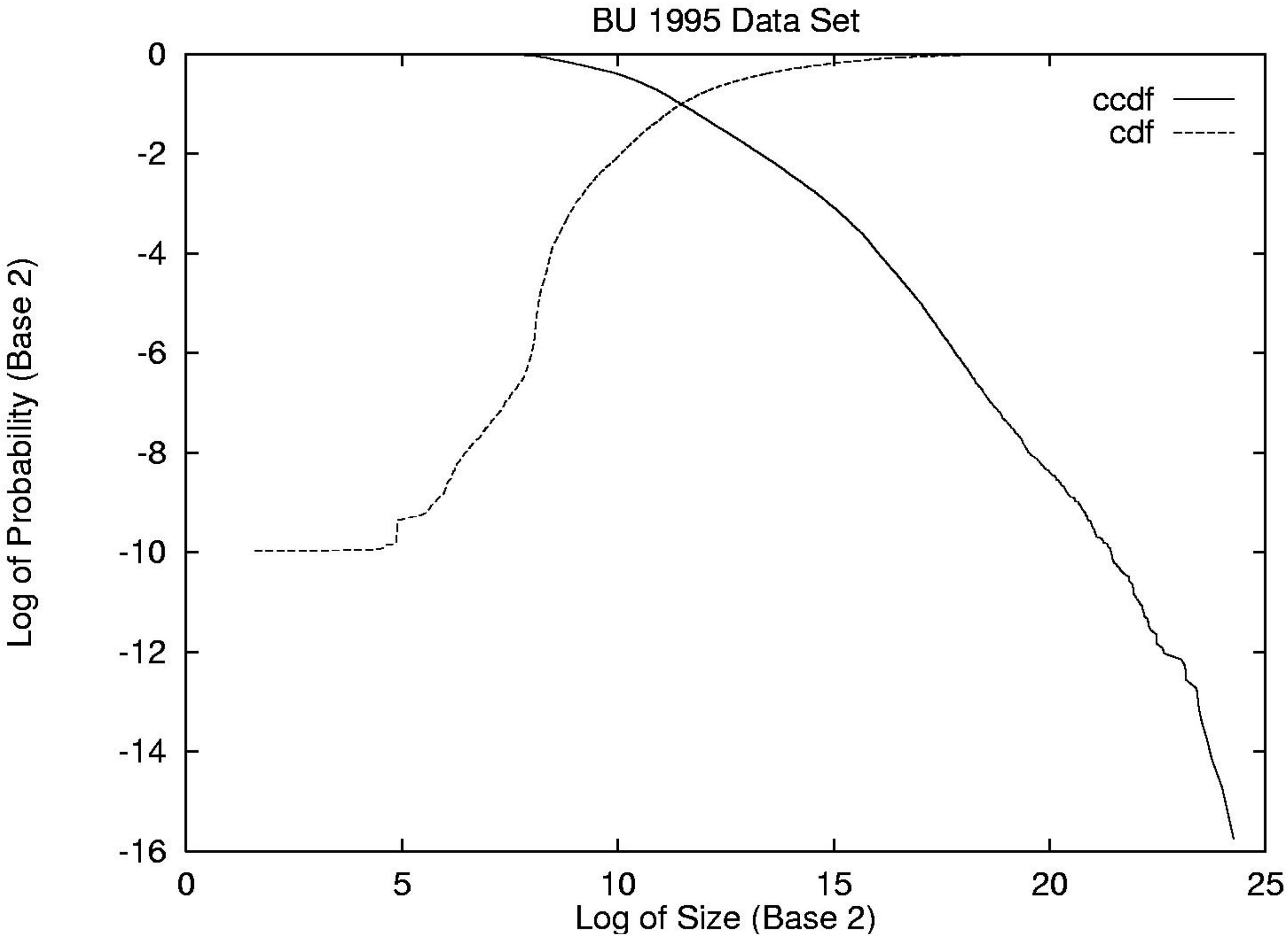
Simulations: CDF



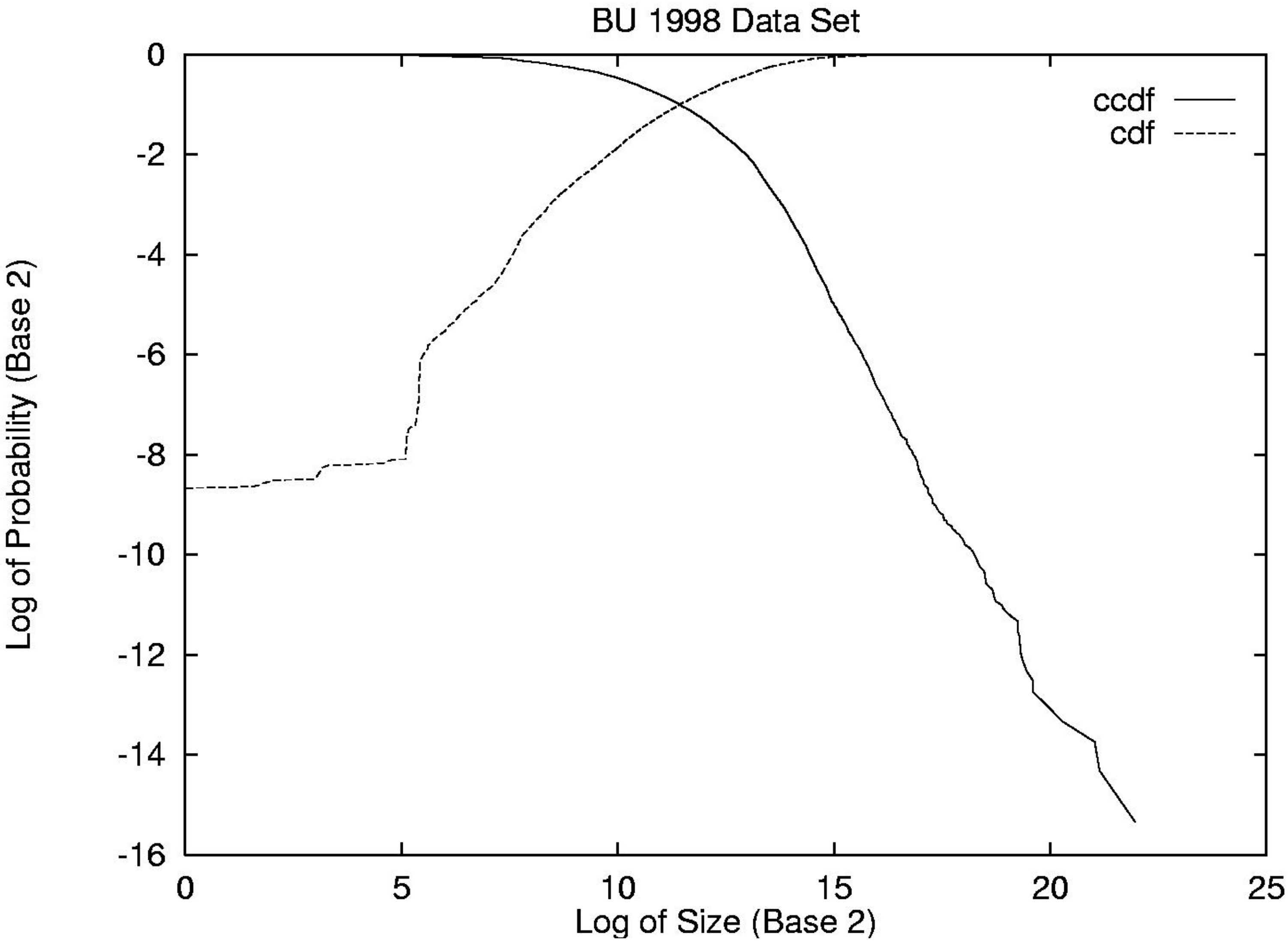
Simulation: CCDF



Boston Univ. 1995 Data Set



Boston Univ 1998 Data Set



Extension: Deletions

- Suppose files deleted uniformly at random with probability q.
  - New file generated with probability p.
  - New file derived with probability $1 - p - q$.
- File depths still geometrically distributed.
- So still a Double Pareto file size distribution.

Extensions: Preferential Attachment

- Suppose new file derived from old file with preferential attachment.
  - Old file chosen with weight proportional to $ax + b$, where x = #current children.
- File depths still geometrically distributed.
- So still get a double Pareto distribution.

Extensions: Correlation

- Each tree in the forest is small.
  - Any multiplicative edge affects few files.
- Martingale argument shows that small correlations do not affect distribution.
- Large systems converge to Double Pareto distribution.

Extensions: Distributions

- Choice of distribution F_1, F_2 matter.
- But not dramatically.
  - Central limit theorem still applies.
  - General closed forms very difficult.

Previous Models

- Downey
  - Introduced simple derivation model.
- HOT [Zhu, Yu, Doyle, 2001]
  - Information theoretic model.
  - File sizes chosen by Web system designers to maximize information/unit cost to user.
  - Similar to early heavy tail work by Mandelbrot.
  - More rigorous framework also studied by Fabrikant, Koutsoupias, Papadimitriou.
- Log-*t* distributions [Mitzenmacher,Tworetzky, 2003]

Summary of File Model

- Recursive Forest File Model
  - is simple, general.
  - combines multiplicative models and simple, well-studied random graph processes.
  - is robust to changes (deletions, preferential attachement, etc.)
  - explains lognormal body / heavy tail phenomenon.

Future Directions

- Tools for characterizing double-Pareto and double-Pareto lognormal parameters.
  - Fine tune matches to empirical results.
- Find evidence supporting/contradicting the model.
  - File system histories, etc.
- Applications in other fields.
  - Explains Double Pareto distributions in generational settings.

Conclusions

- Power law distributions are natural.
  - They are everywhere.
- Many simple models yield power laws.
  - New paper algorithm (to be avoided).
    - Find empirical power law with no model.
    - Apply some standard model to explain power law.
- Lognormal vs. power law argument natural.
  - Some generative models are extremely similar.
  - Power law appears more robust.
  - Double Pareto distributions may explain lognormal body / Pareto tail phenomenon.

New Directions for Power Law Research

Michael Mitzenmacher

Harvard University

My (Biased) View

- There are 5 stages of power law research.

1) Observe: Gather data to demonstrate power law behavior in a system.
2) Interpret: Explain the importance of this observation in the system context.
3) Model: Propose an underlying model for the observed behavior of the system.
4) Validate: Find data to validate (and if necessary specialize or modify) the model.
5) Control: Design ways to control and modify the underlying behavior of the system based on the model.

My (Biased) View

- In networks, we have spent a lot of time *observing* and *interpreting* power laws.
- We are currently in the *modeling* stage.
  - Many, many possible models.
  - I'll talk about some of my favorites later on.
- We need to now put much more focus on *validation* and *control*.
  - And these are specific areas where computer science has much to contribute!

Validation: The Current Stage

- We now have so many models.
- It may be important to know the *right* model, to extrapolate and control future behavior.
- Given a proposed underlying model, we need tools to help us validate it.
- We appear to be entering the validation stage of research…. BUT the first steps have focused on *invalidation* rather than *validation*.

Examples : Invalidation

- Lakhina, Byers, Crovella, Xie
  - Show that observed power-law of Internet topology might be because of biases in traceroute sampling.
- Chen, Chang, Govindan, Jamin, Shenker, Willinger
  - Show that Internet topology has characteristics that do not match preferential-attachment graphs.
  - Suggest an alternative mechanism.
    - But does this alternative match all characteristics, or are we still missing some?

My (Biased) View

- Invalidation is an important part of the process! BUT it is inherently different than validating a model.
- Validating seems much harder.
- Indeed, it is arguable what constitutes a validation.
- Question: what should it mean to say "This model is consistent with observed data."

To Control

- In many systems, intervention can impact the outcome.
  - Maybe not for earthquakes, but for computer networks!
  - Typical setting: individual agents acting in their own best interest, giving a global power law. Agents can be given incentives to change behavior.
- General problem: given a good model, determine how to change system behavior to optimize a global performance function.
  - Distributed algorithmic mechanism design.
  - Mix of economics/game theory and computer science.

Possible Control Approaches

- Adding constraints: local or global
  - Example: total space in a file system.
  - Example: preferential attachment but links limited by an underlying metric.
- Add incentives or costs
  - Example: charges for exceeding soft disk quotas.
  - Example: payments for certain AS level connections.
- Limiting information
  - Impact decisions by not letting everyone have true view of the system.

Conclusion : My (Biased) View

- There are 5 stages of power law research.
  1) Observe: Gather data to demonstrate power law behavior in a system.
  2) Interpret: Explain the import of this observation in the system context.
  3) Model: Propose an underlying model for the observed behavior of the system.
  4) Validate: Find data to validate (and if necessary specialize or modify) the model.
  5) Control: Design ways to control and modify the underlying behavior of the system based on the model.
- We need to focus on validation and control.
  - Lots of open research problems.

A Chance for Collaboration

- The observe/interpret stages of research are dominated by systems; modeling dominated by theory.
  - And need new insights, from statistics, control theory, economics!!!
- Validation and control require a strong theoretical foundation.
  - Need universal ideas and methods that span different types of systems.
  - Need understanding of underlying mathematical models.
- But also a large systems buy-in.
  - Getting/analyzing/understanding data.
  - Find avenues for real impact.
- Good area for future systems/theory/others collaboration and interaction.