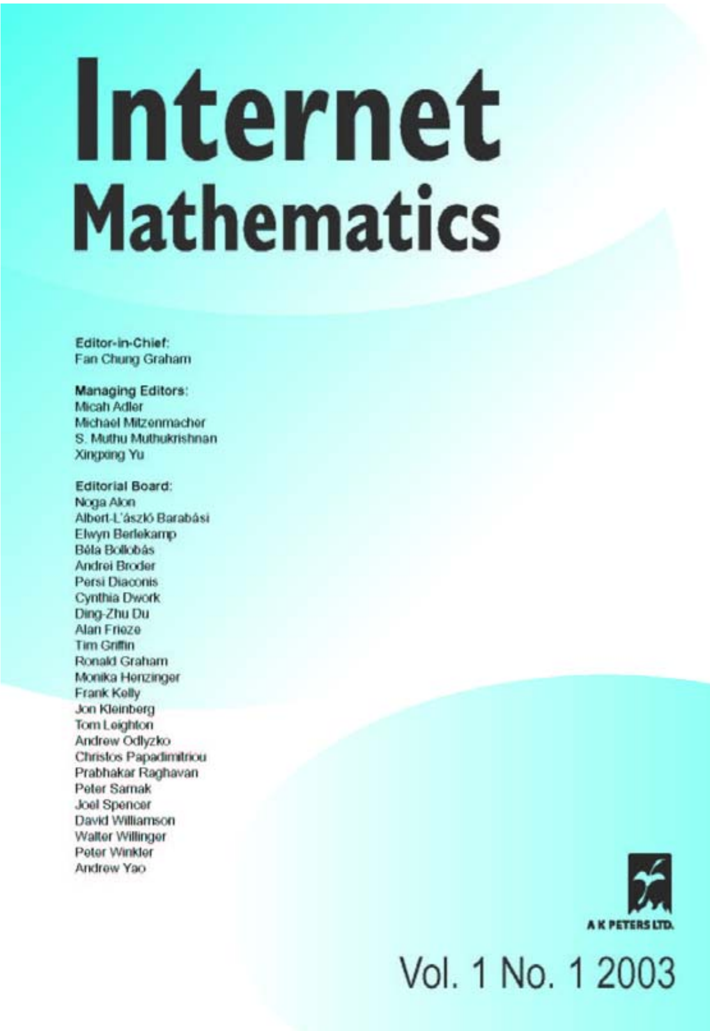
New Directions for Power Law Research

Michael Mitzenmacher
Harvard University

Internet Mathematics



Articles Related to This Talk

The Future of Power Law Research

Dynamic Models for File Sizes and Double Pareto Distributions

A Brief History of Generative Models for Power Law and Lognormal Distributions

Motivation: General

- Power laws (and/or scale-free networks) are now everywhere.
  - See the popular texts *Linked* by Barabasi or *Six Degrees* by Watts.
  - In computer science: file sizes, download times, Internet topology, Web graph, etc.
  - Other sciences: Economics, physics, ecology, linguistics, etc.
- What has been and what should be the research agenda?

My (Biased) View

- There are 5 stages of ~~power law~~ network 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!

Models

- After observation, the natural step is to explain/model the behavior.
- Outcome: lots of modeling papers.
  - And many models rediscovered.
- Lots of history…

History

- In 1990's, the abundance of observed power laws in networks surprised the community.
  - Perhaps they shouldn't have… power laws appear frequently throughout the sciences.
    - Pareto : income distribution, 1897
    - Zipf-Auerbach: city sizes, 1913/1940's
    - Zipf-Estouf: word frequency, 1916/1940's
    - Lotka: bibliometrics, 1926
    - Yule: species and genera, 1924.
    - Mandelbrot: economics/information theory, 1950's+
- Observation/interpretation were/are key to initial understanding.
- My claim: but now the mere existence of power laws should not be surprising, or necessarily even noteworthy.
- My (biased) opinion: The bar should now be very high for observation/interpretation.

Power Law Distribution

- A power law distribution satisfies

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

- Pareto distribution

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

  - 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.

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.)

Preferential Attachment History

- This model (without the graphs) was derived in the 1950’s by Herbert Simon.
  - … who won a Nobel Prize in economics for entirely different work.
  - His analysis was not for Web graphs, but for other preferential attachment problems.

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 leads to power law.

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

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:

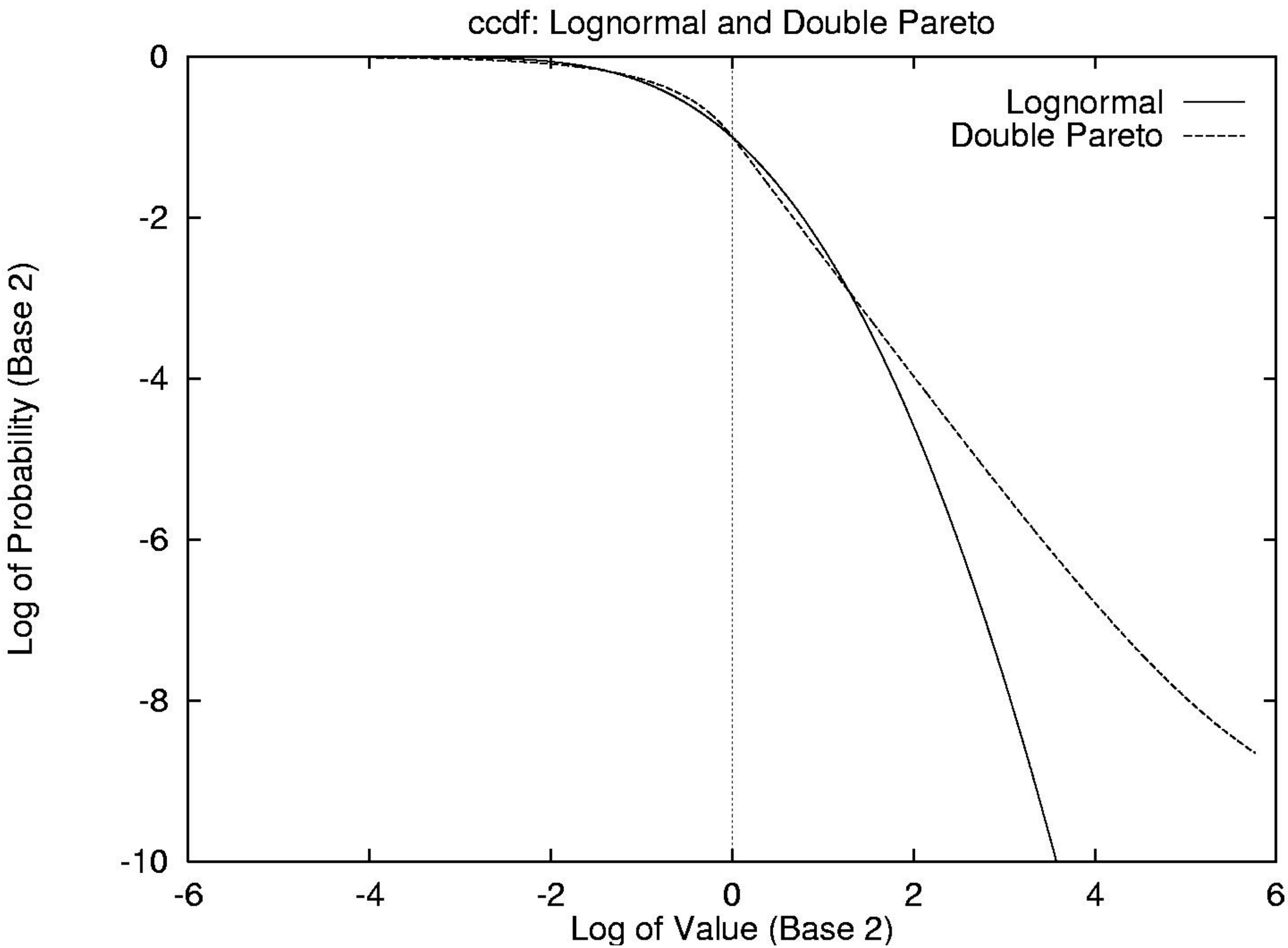
$$X_t = \max(\varepsilon, F_{t-1} X_{t-1})$$

  then X_t converges to a power law distribution. (Champernowne, 1953)
- Lognormal model easily pushed to a power law model.

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 distributed exponentially.
  - E.g., When Web size doubles every year.
- Resulting distribution is Double Pareto.
  - Between lognormal and Pareto.
  - Linear tail on a log-log chart, but a lognormal body.

Lognormal vs. Double Pareto



And So Many More…

- New variations coming up all of the time.
- Question : What makes a new power law model sufficiently interesting to merit attention and/or publication?
  - Strong connection to an observed process.
    - Many models claim this, but few demonstrate it convincingly.
  - Theory perspective: new mathematical insight or sophistication.
- My (biased) opinion: the bar should start being raised on model papers.

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."

Time-Series/Trace Analysis

- Many models posit some sort of actions.
  - New pages linking to pages in the Web.
  - New routers joining the network.
  - New files appearing in a file system.
- A validation approach: gather traces and see if the traces suitably match the model.
  - Trace gathering can be a challenging systems problem.
  - Check model match requires using appropriate statistical techniques and tests.
  - May lead to new, improved, better justified models.

Sampling and Trace Analysis

- Often, cannot record all actions.
  - Internet is too big!
- Sampling
  - Global: snapshots of entire system at various times.
  - Local: record actions of sample agents in a system.
- Examples:
  - Snapshots of file systems: full systems vs. actions of individual users.
  - Router topology: Internet maps vs. changes at subset of routers.
- Question: how much/what kind of sampling is sufficient to validate a model appropriately?
  - Does this differ among models?

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.