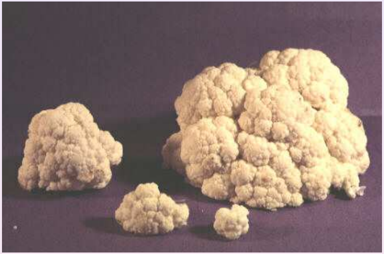
Power Laws in Networks

Why Wikipedia is like sex, trees are like computer files and lots of things are like cauliflower.



Richard G. Clegg (richard@richardclegg.org)— University of York, March 2006

(Pictures are mostly taken from wikipedia.)

Introduction

1. This talk is about current internet research and in particular the topic of power laws.

Introduction

1. This talk is about current internet research and in particular the topic of power laws.
2. Ongoing research — some questions asked have no answers yet, some answers given may be wrong.

Introduction

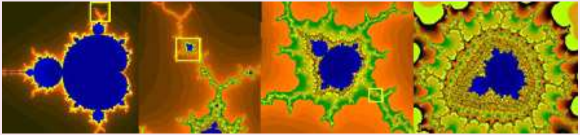
1. This talk is about current internet research and in particular the topic of power laws.
2. Ongoing research — some questions asked have no answers yet, some answers given may be wrong.
3. A taste of what is going on in mathematical internet research now.

Introduction

1. This talk is about current internet research and in particular the topic of power laws.
2. Ongoing research — some questions asked have no answers yet, some answers given may be wrong.
3. A taste of what is going on in mathematical internet research now.

The Main Topics

Topics considered: Statistical Self-Similarity, Heavy-Tails, Long-Range Dependence and Scale Free Networks.



What do we mean by a Power law?

A Power Law Relationship

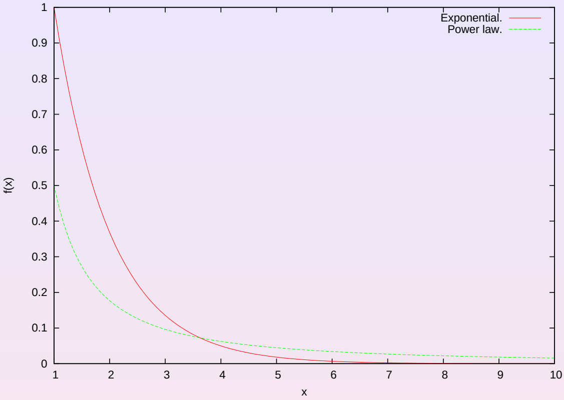
We will define a power law relationship for a function $f(x)$ for $x > 0$ as

$$f(x) \sim x^{-\alpha},$$

where $\alpha > 0$ is some constant and $\sim$ means asymptotically proportional to as $x \to \infty$.

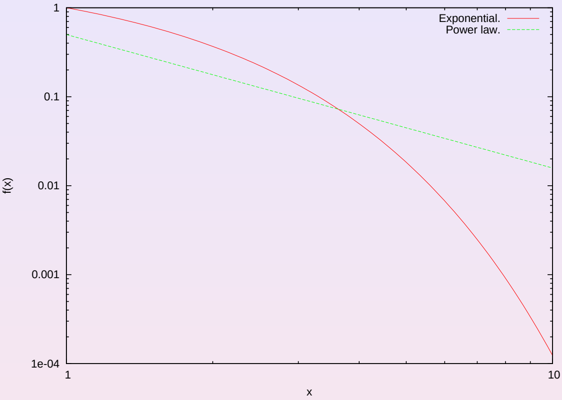
- This function falls off to zero quite slowly.
- This sort of function occurs in a surprising variety of contexts, particularly in the internet.
- Plotted on a log-log scale it will appear as a straight line gradient $-\alpha$.

Power law fall off



Heavy tailed $x^{-1.5}$ versus exponential e^{-x} decay. Both curves normalised so they have unit area in the interval $[1, \infty)$.

Power law fall off



The previous graph plotted on a logscale.

What is a Heavy-Tailed distribution?

Heavy-Tailed distribution

A variable X has a heavy-tailed distribution if

$$\mathbb{P}[X > x] \sim x^{-\beta},$$

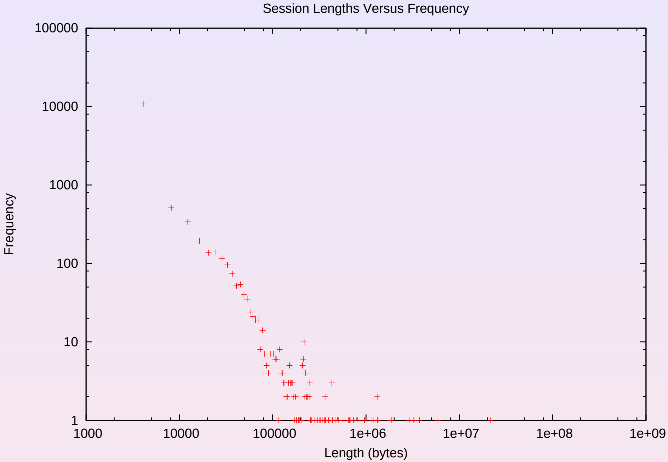
where $\beta \in (0, 2)$ and $\sim$ again means asymptotically proportional to as $x \longrightarrow \infty$.

- Obviously an example of a power law.
- A distribution where *extreme values* are still quite common.
- Examples: Heights of trees, frequency of words, populations of towns.
- Best known example, Pareto distribution $P(X > x) = (x/x_m)^{-\beta}$ where $x_m > 0$ is the smallest possible X.

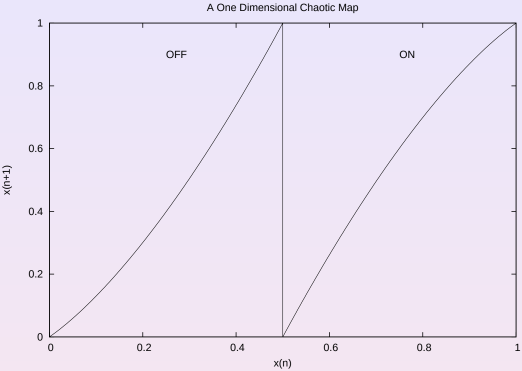
More about heavy tails

- The following internet distributions have heavy tails:
  1. Files on any particular computer.
  2. Files transferred via ftp.
  3. Bytes transferred by single TCP connections.
  4. Files downloaded by the WWW.
- This is more than just a statistical curiousity.
- Consider what this distribution would do to queuing performance (no longer Poisson).
- Non mathematicians are starting to take an interest in heavy tails.
- Some people refer to long-tails — this is a process which starts slowly but continues for a long time (e.g. a slow selling book).

TCP session lengths at York University



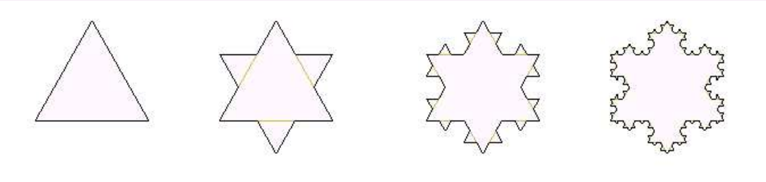
A model for heavy-tailed traffic



- LHS is: $f(x) = x + kx^{-\alpha}$, with $\alpha \in (0, 1)$.
- Generates series of zeros and ones which have heavy tails.
- Difficult analytically (Markov approximations easier).

Self Similarity

- Most of us are familiar with the concept of self similarity.
- A curve is thought of as self-similar if parts of the curve resemble other parts (perhaps after rescaling and rotation).
- A self-similar curve is what is popularly thought of as a fractal.
- (There is a mathematically rigorous definition of fractal, which will not be discussed here.)
- A typical example is the Koch snowflake shown below.



Fractals/Self-Similarity

Four "fractals" occurring in various areas.

Statistical Self-Similarity

Statistically self-similar

A stochastic process $Y_t : t \in \mathbb{R}_+$, is stastistically self-similar if it obeys

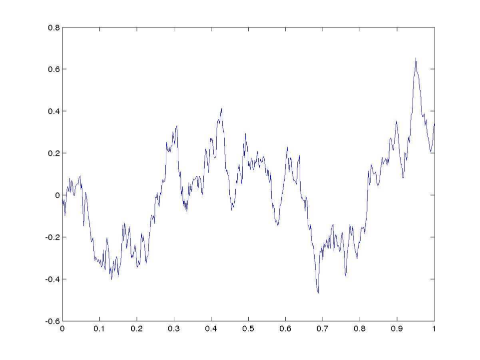
$$Y_t \stackrel{d}{=} c^{-H} Y_{ct},$$

for some constant $c > 0$ where $\stackrel{d}{=}$ means equal in distribution and H is a parameter known as the Hurst parameter.

- Crudely: when stretched by some factor c in the time dimension looks "the same" if stretched by c^{-H} in the y dimension.
- Most time series would look "flat" if stretched like this.

Statistical Self-Similarity

- Think of Statistical Self-Similarity in terms of mountainousness perhaps. As you zoom in, the small hills look "the same" as the larger mountains did.
- Some measurements of internet traffic exhibit SSS (see later).



Fractional Brownian Motion — a statistically self-similar process

Long-Range Dependence (LRD)

Let $\{X_1, X_2, X_3, \dots\}$ be a weakly stationary time series.

The Autocorrelation Function (ACF)

$$\rho(k) = \frac{\mathrm{E}\left[(X_t - \mu)(X_{t+k} - \mu)\right]}{\sigma^2},$$

where μ is the mean and σ^2 is the variance.

The ACF measures the correlation between X_t and X_{t+k} and is normalised so $\rho(k) \in [-1, 1]$. Note symmetry $\rho(k) = \rho(-k)$.
A process exhibits LRD if $\sum_{k=0}^{\infty} \rho(k)$ diverges (is not finite).

Definition of Hurst Parameter

The following functional form for the ACF is often assumed

$$\rho(k) \sim |k|^{-2(1-H)},$$

where $\sim$ means asymptotically proportional to and $H \in (1/2, 1)$ is the Hurst Parameter.

More about LRD

- Think of LRD as meaning that data from the distant past continue to effect the present.
- LRD was first spotted by a hydrologist (Hurst) looking at the flooding of the Nile river.
- For this reason Mandelbrot called it “the Joseph effect”.
- Stock prices (once normalised) also show LRD.
- LRD can also be seen in the temperature of the earth (once the trend is removed).
- Related to self-similarity. If Y_t is self-similar with Hurst $H \in (1/2, 1)$ and stationary increments $X_t = Y_{t+1} - Y_t$ then X_t is LRD with Hurst H.

LRD and the Internet

- In 1993 LRD (and self-similarity) was found in a time series of bytes/unit time measured on an Ethernet LAN [Leland et al '93].
- This finding has been repeated a number of times by a large number of authors (however recent evidence suggests this may not happen in the core).
- A higher Hurst parameter often increases delays in a network. Packet loss also suffers.
- If buffer provisioning is done using the assumption of Poisson traffic then the network will probably be underspecifed.
- The Hurst parameter is "a dominant characteristic for a number of packet traffic engineering problems".

The horrible properties of LRD

- Computationally, LRD is a nightmare to work with.
- Consider $\rho(k)$ — the effect we are looking for is at large k we only have many samples for small k. Standard estimators for $\rho(k)$ are biased for large k.
- The sample mean converges at a rate proportional to n^{2H-2} not n^{-1}.
- The sample variance S^2 is no longer an unbiased estimate of the variance σ^2.
- If we take standard techniques for confidence intervals then, as $n \to \infty$ a statistic will be outside a given confidence interval a.s. no matter how small that confidence interval.
- Only investigate LRD if you have a "large" data set (hundreds are good, thousands are better, millions are nice).

Where does LRD come from?

Where do we get LRD from? The research literature suggests four possibilities for the origin of LRD in the internet.

1. Data is LRD at Source
    - Claim arises from measurements on video traffic.
    - Pictures are updated by sending changes.
    - A still scene is few changes, a cut or pan is a lot of changes.
2. Data arise from aggregation of heavy tailed ON-OFF sources.
    - It can be shown that ON/OFF sources with heavy-tailed train lengths leads to self-similarity.
    - It has been observed that the sizes of files transferred on the internet follow a heavy-tailed distribution.

Where does LRD come from? (2)

3. LRD arises from feedback mechanisms in the TCP protocol.
   - This claim comes from Markov models of TCP timeout and retransmission.
   - A Markov model is used to show that the doubling of timeouts can cause correlations in timeseries of transmitted data.
   - Modelling shows that this can lead to LRD over certain timescales ("local" LRD).
4. LRD arises from network topology or routing.
   - Consider a simulation on a Manhattan network with randomly distributed sources and sinks.
   - The sources produce Poisson traffic.
   - Packets find their shortest route to the sink (accounting for the traffic on the next hop).
   - In this simple situation the aggregated traffic shows LRD.

Node Degree Distributions

Degree of a node

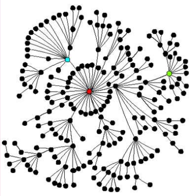
In an undirected graph, the degree of a node is the number of arcs incident to that node.

A scale free network

Let X be the degree of a node in a network. A network is said to be scale free if

$$\mathbb{P}[X = k] \sim k^{-\alpha},$$

where $\alpha \in (0, 2)$.



Scale-free networks

- A scale free network is a network where there are a significant number of highly connected nodes.
- What sort of networks (graphs) are scale free?
  1. The internet if a node is a computer/router and an arc is a connection between them.
  2. The web if a node is a web page and an arc is a hyperlink to (or from) that page.
  3. Citations if a node is an academic and an arc is when an author cites another.
  4. Wikipedia if a node is an article and an arc is when that article references another.
  5. Sexual relations if a node is a person and an arc is... um... a connection between them.
  6. Many many more networks.
- For some reason scale-free networks seem very common — but why?

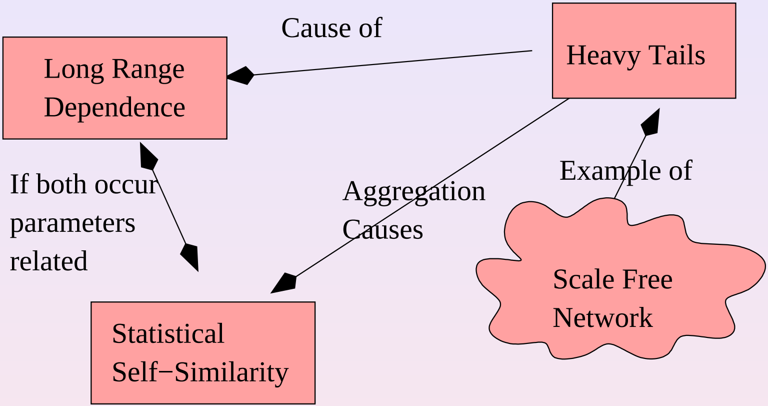
How might Scale Free Networks be formed?

- We might want to think about how a Scale free network forms (see [Albert and Barabassi 1999]).
- Consider a network where new arcs attach to nodes with a probability related to the number of links it has already.
- Let the probability a new node connects to a given existing node be proportional to k the degree of the node. In this model the network becomes scale free.
- But if the probability is proportional to $k^{1.0001}$ or $k^{0.9999}$ the network is not scale free.
- How such exact proportionality could be achieved is a mystery.

Random Walks and local formation of Scale Free Networks?

- Consider taking a random walk on a network.
- Leave any node down an arc completely at random.
- Consider the process as a Markov chain where the transition probability $p_{ij} = 1/k_i$ where k_i is the degree of node i.
- It can be easily shown that if the networks is ergodic, the equilibrium probabilities of the state $\pi_i = Kk_i$ for some constant K. (Left as an exercise for the student).
- This is exactly the probability required for our connection model.

Connections



Conclusions

- Power laws seem to appear in a huge number of places in the natural world.
- They are particularly common on the internet.
- There are some known interconnections but perhaps more remain to be discovered.
- Could it be that there is some underlying principle which explains why they are so ubiquitous?
- This area of research is rapidly developing with new discoveries every month.

References

1) Leland, W. E., Taqqu, M. S., Willinger, W., and Wilson, D. V. (1993). *On the self-similar nature of Ethernet traffic*. In Proc. ACM SIGCOMM, pages 183-193, San Francisco, California.
2) Barabsi, A.-L. and Albert, R. (1999), *Emergence of scaling in random networks*, Science 286, 509–512
3) Peitgen, H. O., Jrgens, H. and Saupe, D. (2005) *Chaos and Fractals: New Frontiers of Science*