Heavy Tails: The Origins and Implications for Large Scale Biological & Information Systems

Predrag R. Jelenković

Dept. of Electrical Engineering
Columbia University , NY 10027, USA
{predrag}@ee.columbia.edu

January 2007

Outline

1. General Course Information
2. Heavy Tails are Ubiquitous
   - 100+ years of repeated observations of power laws
   - Properties of heavy-tailed distributions
   - Formal definitions of heavy-tailed & subexponential distributions
3. Empirical Observations in Different Fields
   - Socioeconomic fields
   - Biological fields
   - Information technology fields
   - Scale free networks
4. Normal distributions and exponential distributions

Heavy Tails: The Origins and Implications for Large Scale Biological & Information Systems

Lecturer: Prof. Predrag Jelenkovic
Office hours: Wed. 4-5pm
Office: 812 Schapiro Research Bldg.
Phone: (212) 854-8174
Email: predrag@ee.columbia.edu
URL: http://www.comet.columbia.edu/ predrag

Day, time and place: Wed 12:10pm - 2:40pm,
Credits: 3

Required text: research papers will be primarily used as well as the lecture notes.

Project(s): small numerical or simulation problems might be periodically assigned.

Homework: Occasional assignments will be given.

Final: will consist of a project, in class presentation and a written paper.

Grading: Hwk (20%) + Final (80%)

Software requirements: Quantitative homework assignments may require the use of mathematical software packages MATHEMATICA or MATLAB.

Description

Since the early works of Pareto in 1897 and later of Zipf, heavy tails have been repeatedly observed for over a hundred years. Heavy-tailed distributions, in particular power laws, have been found in a vide variety of biological, technological and socioeconomic areas.

In this course, we will study general laws that explain the ubiquitous nature of heavy tails. Basically, the wide appearance of Gaussian/Normal distributions can be attributed to the generality of the central limit theorem. Similarly, we will present the existing and some very new laws that under very general conditions almost invariably result in heavy tails. We will study the implications that the heavy-tailed phenomena have on biological networks as well as on the design of future information networks and systems.

Outline

1. General Course Information
2. Heavy Tails are Ubiquitous
   - 100+ years of repeated observations of power laws
   - Properties of heavy-tailed distributions
   - Formal definitions of heavy-tailed & subexponential distributions
3. Empirical Observations in Different Fields
   - Socioeconomic fields
   - Biological fields
   - Information technology fields
   - Scale free networks
4. Normal distributions and exponential distributions

100+ years of repeated observations of power laws

Socioeconomic area

- Incomes, Pareto (1897)
- Population of cities Arrherbach (1913) & Zipf (1949)

Biological area

- Species-area relationship, Arrhenius (1921)
- Gene family sizes Huynen & Nimwegen (1998)

Technological area: the Internet

- Ethernet LAN traffic Leland, Willinger et al. (1993), Scenes in MPEG video streams Jelenković et al. (1997), WWW traffic Crovella & Bestavros (1997)
- Page requests Cunha et al. (1995), pages and visitors per Web site Adamic & Huberman (1999, 2000)

Are these observations merely a big coincidence?

Are there universal mathematical laws governing these phenomena?

Goal of this course

- Study the phenomena of heavy-tailed (power law) distributions
- Provide rigorously and robust models to explain the ubiquitous nature of heavy tails and, in particular, power laws
- Apply those models and their inferences to systems biology and information networks

Are these observations merely a big coincidence?

Are there universal mathematical laws governing these phenomena?

Goal of this course

- Study the phenomena of heavy-tailed (power law) distributions
- Provide rigorously and robust models to explain the ubiquitous nature of heavy tails and, in particular, power laws
- Apply those models and their inferences to systems biology and information networks

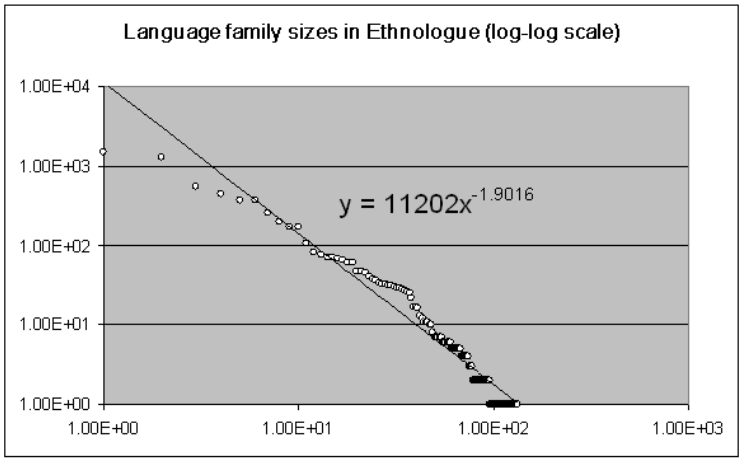
A first glance at power law distributions

Roughly speaking, a random variable X has a power law tail if there exists $\alpha > 0$, such that

$$\mathbb{P}[X > x] \sim \frac{H}{x^{\alpha}}$$

or, more generally,

$$\frac{\log \mathbb{P}[X > x]}{\log x} \to -\alpha$$



Therefore, in the log-log plot, a power law distribution is approximately a straight line with negative slope.

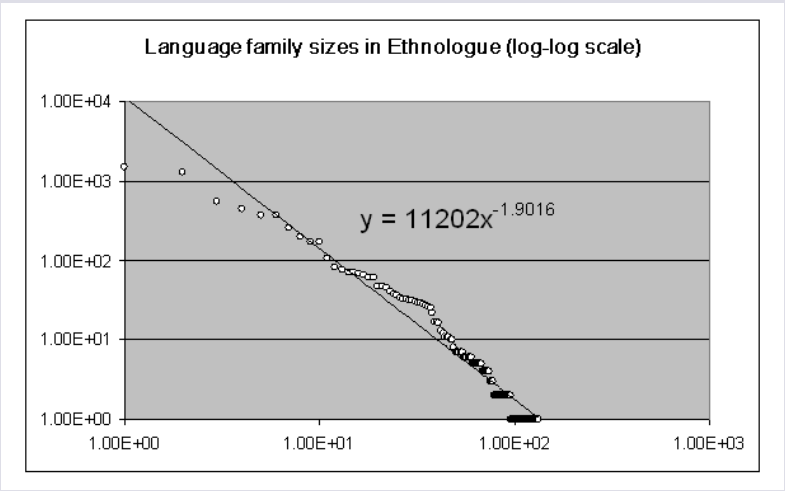
A first glance at power law distributions

Roughly speaking, a random variable X has a power law tail if there exists $\alpha > 0$, such that

$$\mathbb{P}[X > x] \sim \frac{H}{x^{\alpha}}$$

or, more generally,

$$\frac{\log \mathbb{P}[X > x]}{\log x} \to -\alpha$$



Therefore, in the log-log plot, a power law distribution is approximately a straight line with negative slope.

Outline

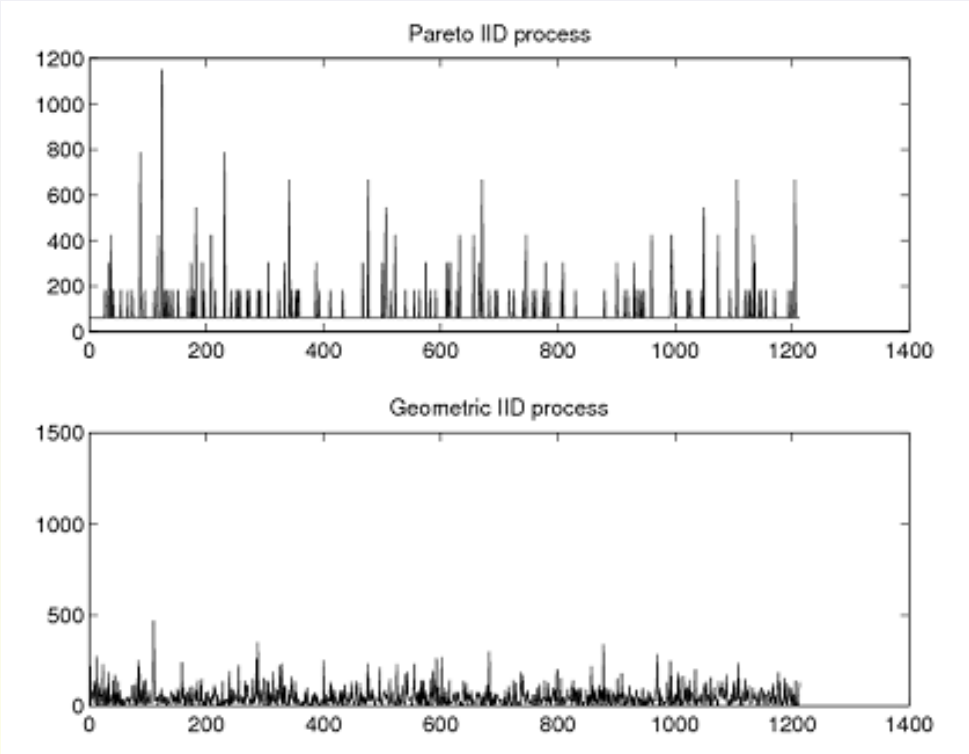
1. General Course Information
2. Heavy Tails are Ubiquitous
   - 100+ years of repeated observations of power laws
   - Properties of heavy-tailed distributions
   - Formal definitions of heavy-tailed & subexponential distributions
3. Empirical Observations in Different Fields
   - Socioeconomic fields
   - Biological fields
   - Information technology fields
   - Scale free networks
4. Normal distributions and exponential distributions

How are heavy tails different?

Properties

- Much heavier distribution tail than exponential distribution
- Large values strike the system often

Comparing the sample path of the power law with the geometric distribution of the same mean and variance.



System behavior is dominated by big excursions, not by averaging phenomena.

Figure: Accumulated strength

Figure: A big value

Examples that motivate the study of heavy tails

- Distribution of wealth, income of individuals
- City sizes vs. ranks - given the population, what is the city rank?
- The graphs of gene regulatory and protein-protein networks are scale free
- Long neuron inter-spike intervals in depressed mice
- Internet and WWW - scale free network (graph): fault tolerant, hubs are both the strength and Achilles' heels
- Scene lengths in VBR and MPEG video are heavy-tailed
- Computer files, Web documents, frequency of access are heavy-tailed
- Stock price fluctuations and company sizes
- Inter occurrence of catastrophic events, earthquakes - applications to reinsurance
- Frequency of words in natural languages (often called Zipf's law)

Outline

1. General Course Information
2. Heavy Tails are Ubiquitous
   - 100+ years of repeated observations of power laws
   - Properties of heavy-tailed distributions
   - Formal definitions of heavy-tailed & subexponential distributions
3. Empirical Observations in Different Fields
   - Socioeconomic fields
   - Biological fields
   - Information technology fields
   - Scale free networks
4. Normal distributions and exponential distributions

Heavy-tailed (long-tailed) distributions

A nonnegative random variable X is called *heavy-tailed* ($X \in \mathcal{L}$) if

$$\lim_{x\to\infty} \frac{\mathbb{P}[X > x + y]}{\mathbb{P}[X > x]} = 1, \quad y > 0$$

- Note that $\mathbb{P}[X > x + y]/\mathbb{P}[X > x]$ represents the conditional probability $\mathbb{P}[X > x + y | X > x]$.
- Hence, a random variable is heavy-tailed if the knowledge that X has exceeded a large value x implies that it will exceed an even larger value $x + y$ with a probability close to one.
- In other words, a heavy-tailed random variable exceeds a large value x by a substantial margin.
- If $X \in \mathcal{L}$, then heavier than exponential. Formally, if $X \in \mathcal{L}$ then $\mathbb{P}[X > x]e^{\alpha x} \to \infty$ as $x \to \infty$, for all $\alpha > 0$.

Subexponential distributions

We say that $X \geq 0$ is subexponetial ($X \in \mathcal{S}$) if for any $n \geq 1$ and $X_1, \ldots, X_n$ being n independent copies of X

$$\mathbb{P}\left[\sum_{i=1}^{n} X_i > x\right] \sim n\mathbb{P}[X > x] \quad \text{as} \quad x \to \infty$$

- Invented by Chistyakov in 1964
- Slightly smaller class than $\mathcal{L}$ ($\mathcal{S} \subset \mathcal{L}$)
- Sum of n i.i.d. subexponential random variables exceeds a large value x due to *exactly one* of them exceeding x
- Large Deviations - This remains true (under more restrictive assumptions) if both n and x are proportional and made large at the same time

Primary examples: Power (Pareto/Zipf) laws - regularly varying

- Best known class of subexponetial/heavy-tailed distributions
- *Regularly varying distributions* $\mathcal{R}_{-\alpha}$ (in particular Zipf/Pareto family); $F \in \mathcal{R}_{-\alpha}$ if it is given by

$$F(x) = 1 - \frac{l(x)}{x^{\alpha}} \quad \alpha \geq 0,$$

where $l(x) : \mathbb{R}_+ \to \mathbb{R}_+$ is a function of slow variation, i.e., $\lim_{x\to\infty} l(\delta x)/l(x) = 1, \delta > 1$, e.g., $l(x)$ can be *constant*, $\log x$, $\log\log x$, etc.

Other examples

- Lognormal distribution $F(x) = \Phi\left(\frac{\log x - \mu}{\sigma}\right), \quad \mu \in \mathbb{R}, \sigma > 0$, where Φ is the standard normal distribution.
- Weibull distribution $F(x) = 1 - e^{-x^{\beta}}$, for $0 < \beta < 1$.
- "Almost exponential" $F(x) = 1 - e^{-x(\log x)^{-a}}$, for $a > 0$.

Primary examples: Power (Pareto/Zipf) laws - regularly varying

- Best known class of subexponetial/heavy-tailed distributions
- *Regularly varying distributions* $\mathcal{R}_{-\alpha}$ (in particular Zipf/Pareto family); $F \in \mathcal{R}_{-\alpha}$ if it is given by

$$F(x) = 1 - \frac{l(x)}{x^{\alpha}} \quad \alpha \geq 0,$$

where $l(x) : \mathbb{R}_+ \to \mathbb{R}_+$ is a function of slow variation, i.e., $\lim_{x\to\infty} l(\delta x)/l(x) = 1, \delta > 1$, e.g., $l(x)$ can be *constant*, $\log x$, $\log\log x$, etc.

Other examples

- Lognormal distribution $F(x) = \Phi\left(\frac{\log x - \mu}{\sigma}\right), \quad \mu \in \mathbb{R}, \sigma > 0$, where Φ is the standard normal distribution.
- Weibull distribution $F(x) = 1 - e^{-x^{\beta}}$, for $0 < \beta < 1$.
- "Almost exponential" $F(x) = 1 - e^{-x(\log x)^{-a}}$, for $a > 0$.

Outline

1. General Course Information
2. Heavy Tails are Ubiquitous
   - 100+ years of repeated observations of power laws
   - Properties of heavy-tailed distributions
   - Formal definitions of heavy-tailed & subexponential distributions
3. Empirical Observations in Different Fields
   - Socioeconomic fields
   - Biological fields
   - Information technology fields
   - Scale free networks
4. Normal distributions and exponential distributions

City Sizes

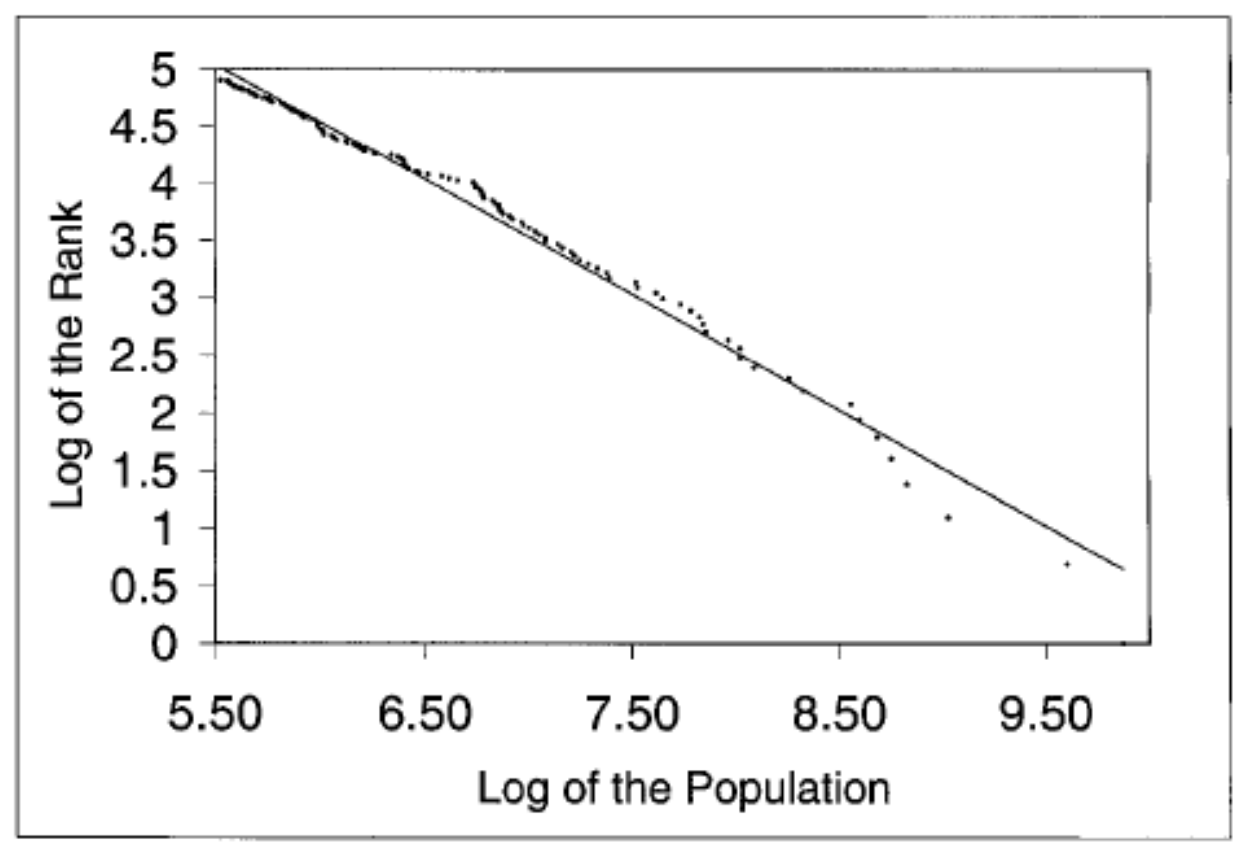
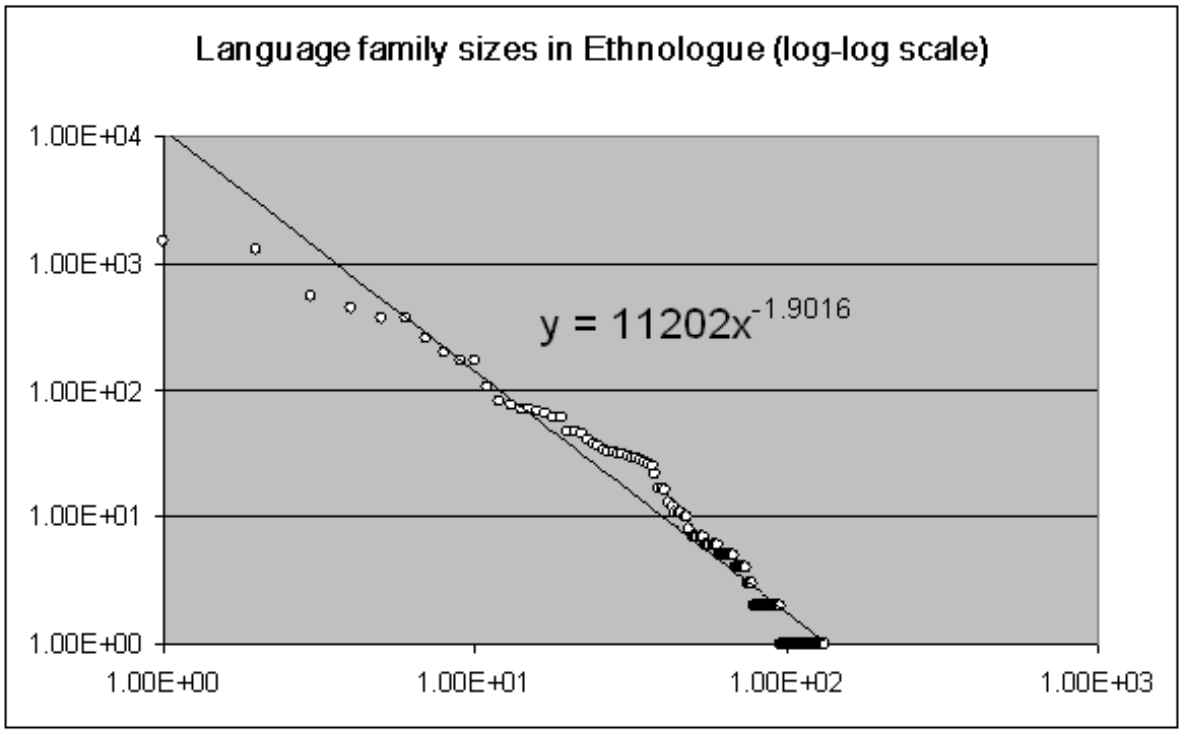


Figure: Log Size versus Log Rank of the 135 largest U.S. Metropolitan Areas in 1991 [cited from Gabaix (1999)].

Language Family Sizes



World Income Distribution

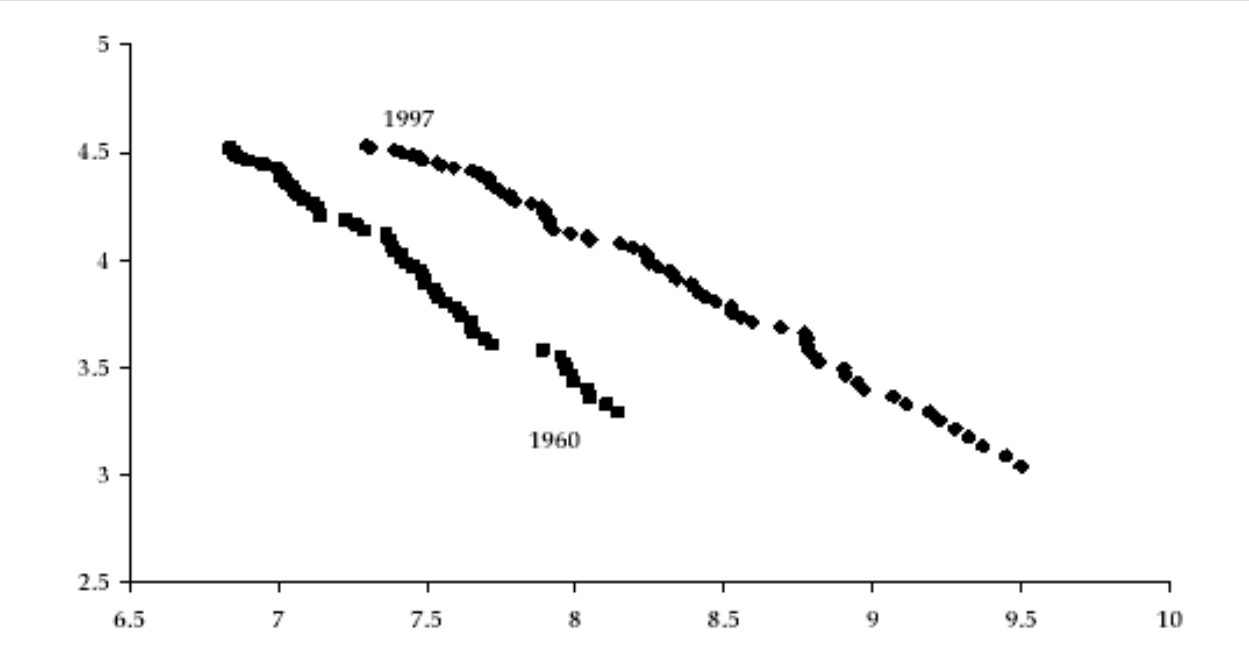


Figure: Zipf's plot of the 30th-85th percentiles of the world income distribution (GDP per capita) in 1960 and 1997.

Trading volume impacts stock price

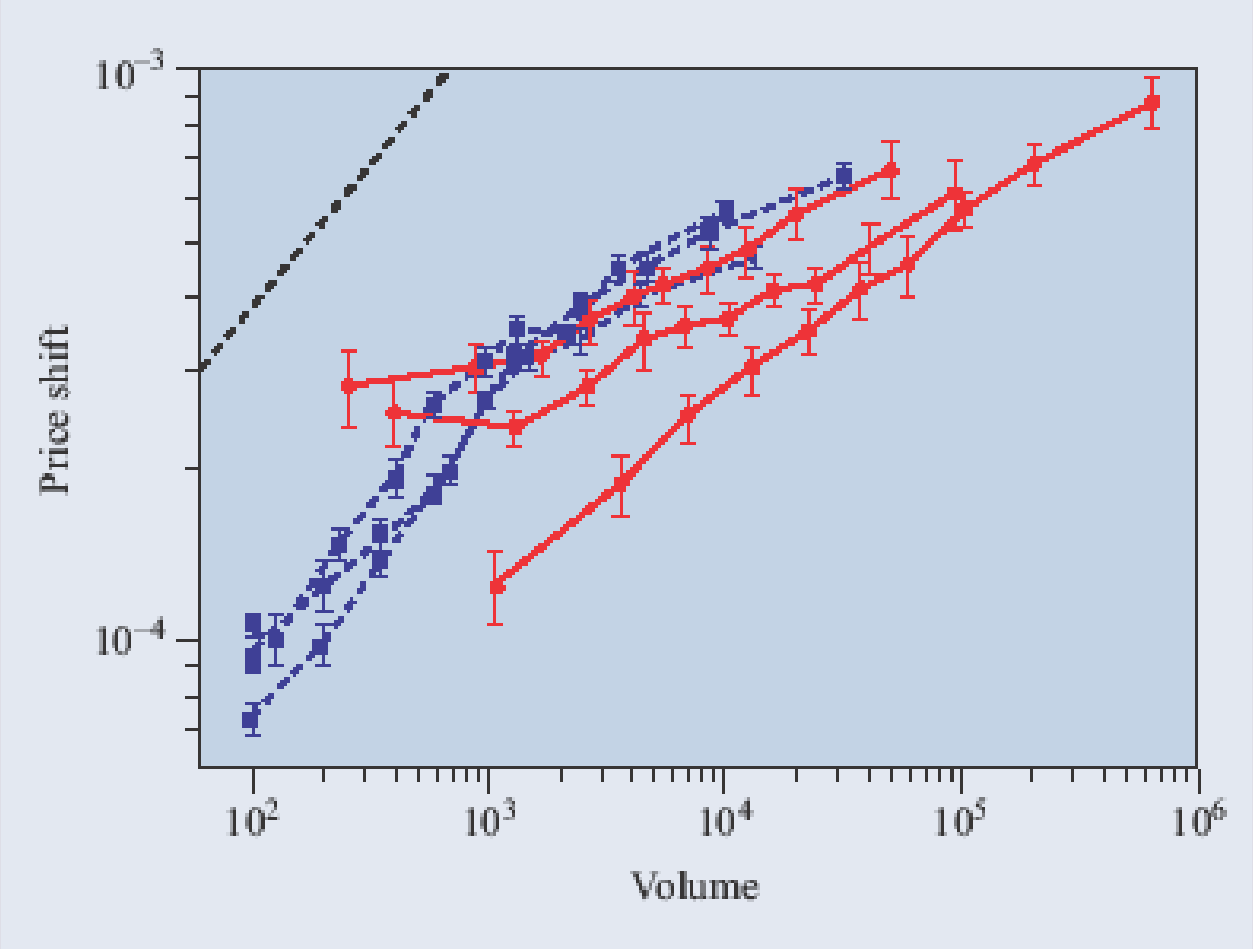


Figure: Market impact function for buy initiated trades of three stocks traded in the NYSE (dashed blue curve) and three stocks traded in the LSE (solid red curve).

Outline

1. General Course Information
2. Heavy Tails are Ubiquitous
   - 100+ years of repeated observations of power laws
   - Properties of heavy-tailed distributions
   - Formal definitions of heavy-tailed & subexponential distributions
3. Empirical Observations in Different Fields
   - Socioeconomic fields
   - Biological fields
   - Information technology fields
   - Scale free networks
4. Normal distributions and exponential distributions

Neuron spiking time-series

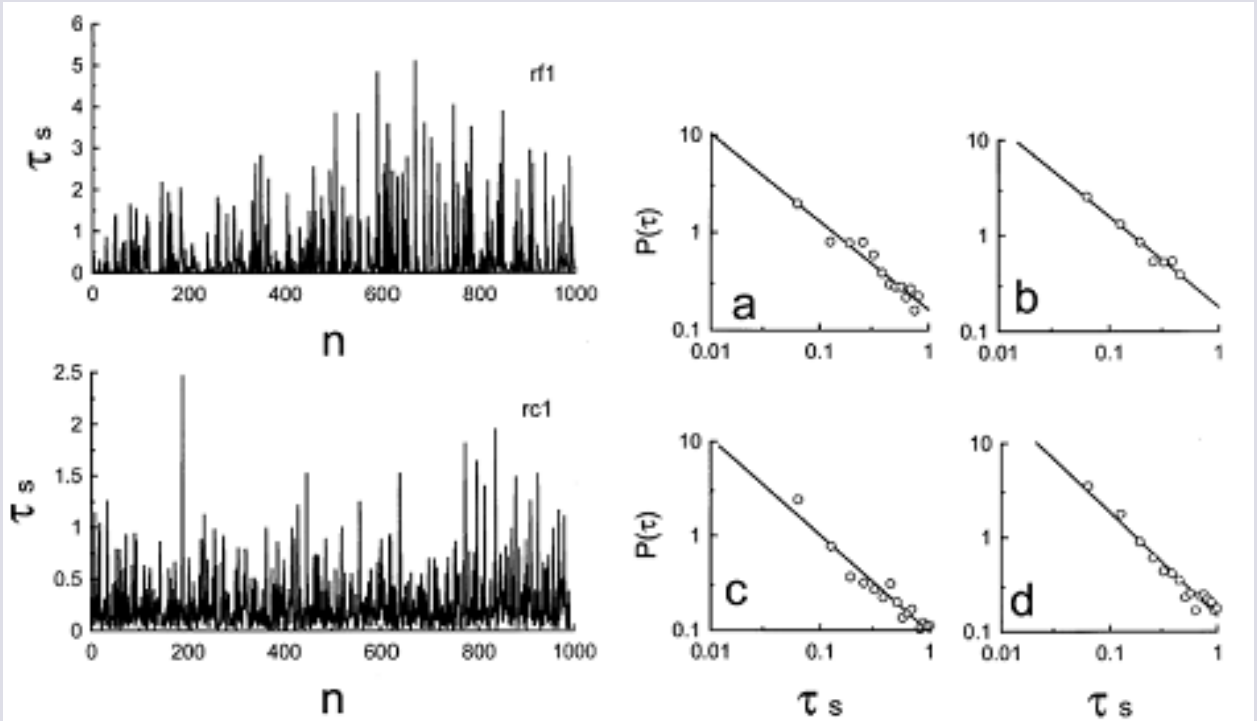


Figure: Power law inter-spike distribution for rat model of depression.

Fractal Characteristics of Neuronal Activity for Firing-code Patterns

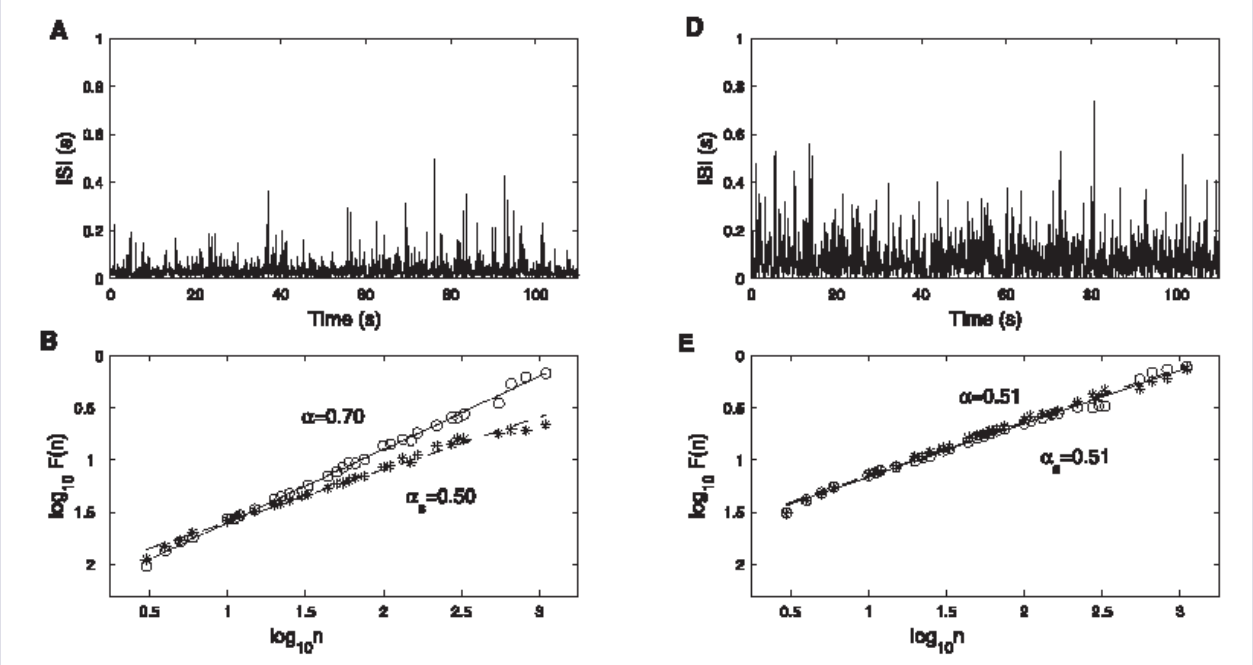


Figure: M. Rodriguez & E. Pereda & J. Gonzalez & P. Abdala & J. A. Obeso (2003)

Outline

1. General Course Information
2. Heavy Tails are Ubiquitous
   - 100+ years of repeated observations of power laws
   - Properties of heavy-tailed distributions
   - Formal definitions of heavy-tailed & subexponential distributions
3. Empirical Observations in Different Fields
   - Socioeconomic fields
   - Biological fields
   - Information technology fields
   - Scale free networks
4. Normal distributions and exponential distributions

File sizes

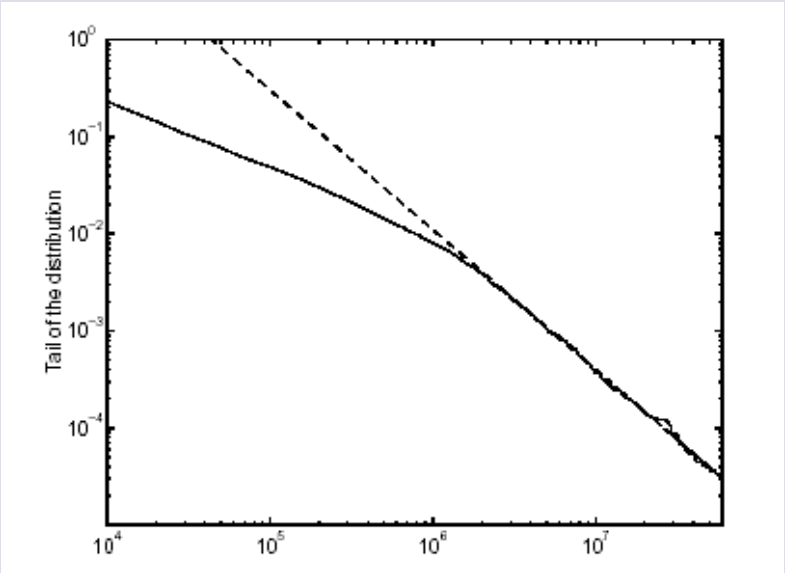


Figure: Log/log plot of the empirical distribution of the file sizes on five file servers in COMET Lab at Columbia University ($\alpha = 1.44$).

Visitors and pages per Web site

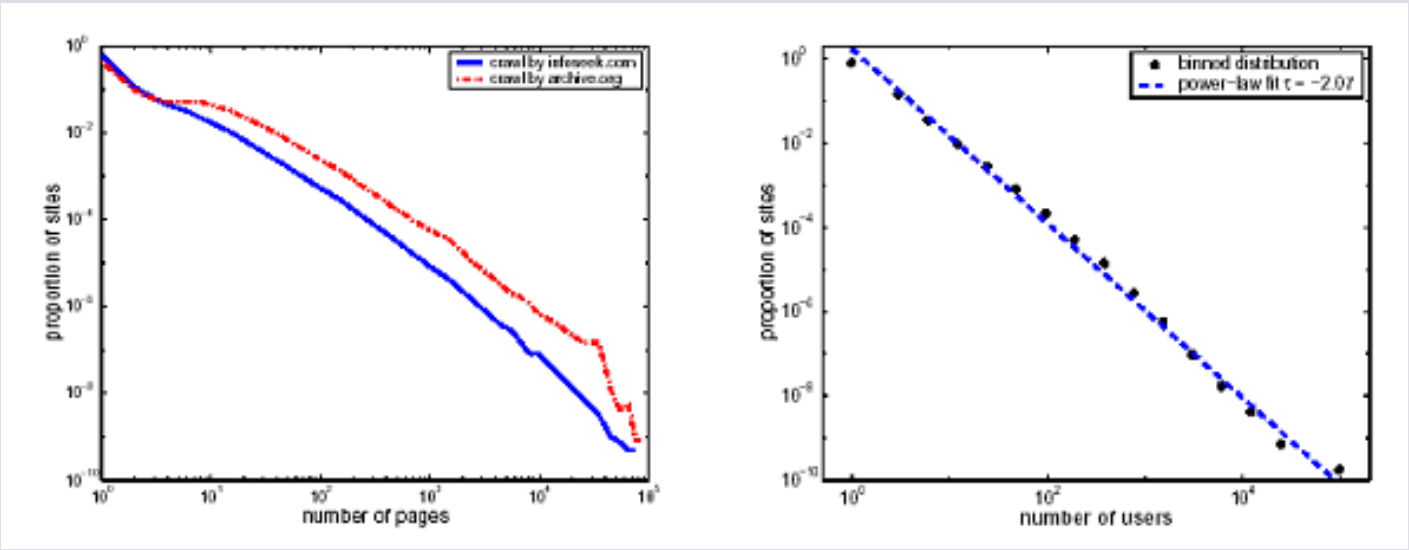


Figure: Fitted power law distributions of the number of pages and visitors per Web site.

Outline

1. General Course Information
2. Heavy Tails are Ubiquitous
   - 100+ years of repeated observations of power laws
   - Properties of heavy-tailed distributions
   - Formal definitions of heavy-tailed & subexponential distributions
3. Empirical Observations in Different Fields
   - Socioeconomic fields
   - Biological fields
   - Information technology fields
   - Scale free networks
4. Normal distributions and exponential distributions

Power Law Random Graph– Scale Free Network

The observations of power-law distributions in the connectivity of complex networks came as a surprise to researchers steeped in the tradition of random networks.

Traditional random graph - Erdos Renyi model VS Scale Free Network - Barabási model

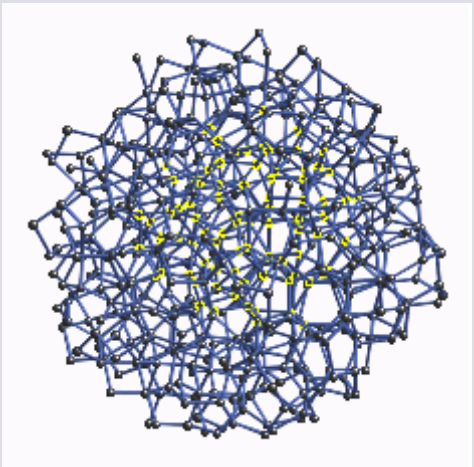


Figure: Concentrated Degree distribution: $\approx$ Poisson

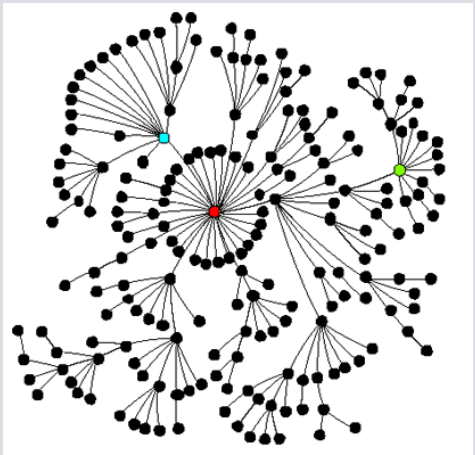


Figure: Power Law Degree distribution

Analogy of Internet topology and interacting proteins in yeast

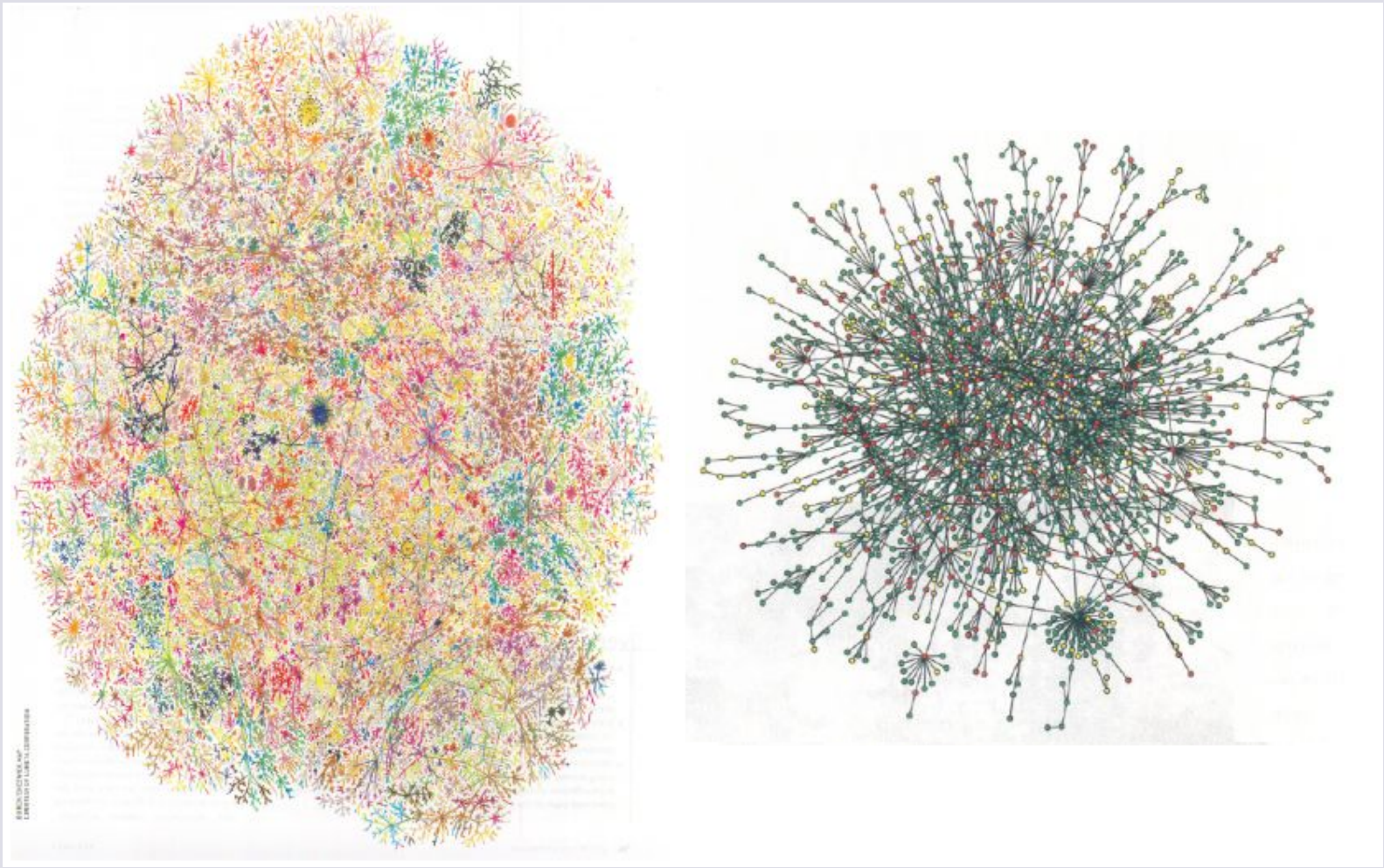


Figure: Image credit: Internet Mapping Project of Lumeta Corporation; Scientific American

Science citation index

1,000 Most Cited Physicists, 1981-June 1997

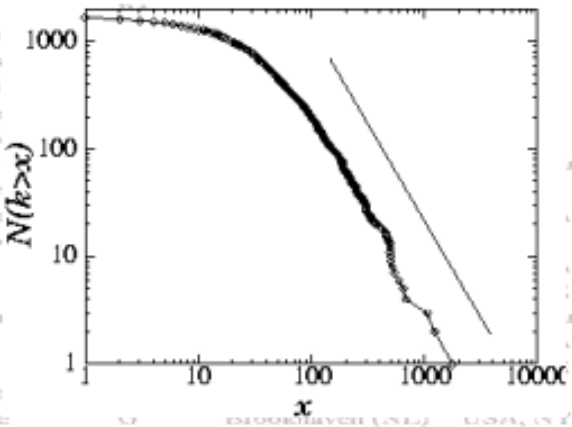
Out of over 500,000 Examined

(see http://www.sst.nrel.gov)

Author name		Institute	Country	Field	avg. cites	total art.	total cites	rank by total cit.
Witten	E	Princeton (U)	USA, NJ	High-energy (T)	168	138	23235	1
Gossard	AC	[illegible]	USA, [illegible]	[illegible]	[illegible]	[illegible]	16994	2
Cava	RJ	[illegible]	USA, [illegible]	[illegible]	[illegible]	[illegible]	14405	3
Batlogg	B	Bell Labs (I)	USA, NJ	Superconductors (E)	83	170	14164	4
Ploog	K	Max-Planck (NL)	Germany	Semiconductors (E)	19	712	13491	5
Ellis	J	Euro Nuclear Cent.	Switzerland	Astrophysics (E)	40	305	12255	6
Fisk	Z	Florida State (U)	USA, FL	Solid State (E)	23	520	12030	7
Cardona	M	Max Planck (NL)	Germany	Semiconductors (E)	20	571	11465	8
Nanopoulos	DV	Texas A&M (U)	USA, TX	High-energy (E)	39	293	11314	9
Heeger	AJ	UCSB (U)	USA, CA	Polymers (E)	34	320	10872	10
Lee*	[illegible]	[illegible]	[illegible]	[illegible]	73	146	10642	11
Suzuki	[illegible]	[illegible]	[illegible]	[illegible]	7.6	1401	10617	12
Anders	[illegible]	[illegible]	[illegible]	Solid State (T)	80	138	10439	13
Suzuki	[illegible]	[illegible]	[illegible]	[illegible]	12	898	10417	14
Freeman	[illegible]	[illegible]	[illegible]	Solid State (E)	21	889	10411	15
Tanaka	[illegible]	[illegible]	[illegible]	[illegible]	11	963	10404	16
Muller	[illegible]	[illegible]	[illegible]	nd Superconductivity (E)	82	122	10049	17
Schnee	[illegible]	[illegible]	[illegible]	Superconductivity (E)	63	156	9768	18
Chemla	[illegible]	[illegible]	[illegible]	Optics (E)	60	162	9668	19
Morko	[illegible]	[illegible]	[illegible]	Semiconductors (E)	20	477	9668	19
Miller	[illegible]	[illegible]	[illegible]	Semiconductors (E)	63	144	9652	21
Chu	[illegible]	[illegible]	[illegible]	Superconductivity (E)	42	213	9453	22
Bednor	[illegible]	[illegible]	[illegible]	nd Superconductivity (E)	110	85	9311	23
Cohen	[illegible]	[illegible]	[illegible]	Solid State (T)	33	284	9311	23
Meng	[illegible]	[illegible]	[illegible]	Superconductivity (E)	86	108	9300	25
Waszc	[illegible]	[illegible]	[illegible]	Superconductivity (E)	57	162	9170	26
Shirane	G	Brookhaven (NL)	USA, NY	Superconductivity (E)	33	269	8841	27
Wiegmann	W	Bell Labs (I)	USA, NJ	Semiconductors (E)	85	104	8822	28
Vandover	RB	Bell Labs (I)	USA, NJ	Magnetism (E)	67	129	8686	29
Uchida*	S	[illegible]	[illegible]	[illegible]	28	301	8520	30
Hor	PH	Houston Univ. (U)	USA, TX	Superconductivity (E)	72	119	8512	31
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	41	286	8375	33
Bergmann	JS	Argonne (NL)	USA, IL	[illegible]	50	167	8298	34
[illegible]	[illegible]	[illegible]	[illegible]	Superconductivity (E)	37	223	8263	35

* citation total may be skewed because of multiple authors with the same name

SCIENCE CITATION INDEX



$P(k) \sim k^{-\gamma}$

$(\gamma = 3)$

Albert László Barabási et al.: The Architecture of Complexity: From the Diameter of the WWW to the Structure of the Cell (power-point presentation) (http://www.nd.edu/~networks/papers.htm)

Scale free network caused by random walk

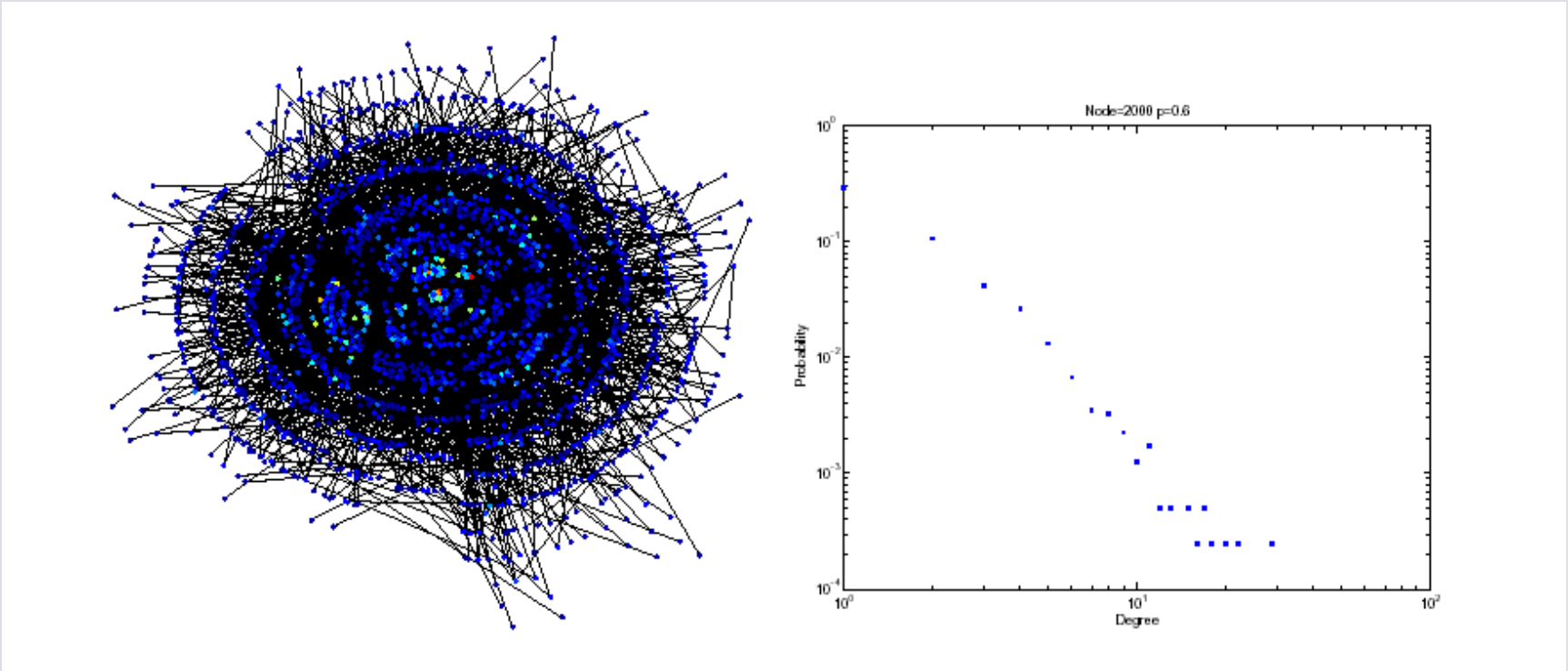


Figure: node=2000, random walk $p = 0.6$.

Why are normal and exponential distributions common?

Gaussian/normal: the central limit theorem

- X_i - identically and independently distributed (i.i.d.) random variables with mean zero and a finite variance σ^2.
- Then the probability density function $f_n(s)$ of the (normalized) sum $S_n = \left(\sum_{i=1}^{n} X_i\right) / (\sigma\sqrt{n})$ of X_i converges to a normal density with unit variance, as n becomes large,

$$f_n(s) \to \phi(s) = \frac{1}{2\sqrt{\pi}} \exp\left(-\frac{s^2}{2}\right).$$

Exponential: queueing theory

The superimum of an additive random walk with negative drift is exponential under quite general conditions.