Web Mining ed Analisi delle Reti Sociali

Proprietà delle Reti – Richiami di elementi di statistica

Dino Pedreschi
Dipartimento di Informatica
Università di Pisa
www.di.unipi.it/~pedre

"Natural" Networks and Universality

- Consider many kinds of networks:
  - social, technological, business, economic, content,…
- These networks tend to share certain *informal* properties:
  - large scale; continual growth
  - distributed, organic growth: vertices "decide" who to link to
  - interaction restricted to links
  - mixture of local and long-distance connections
  - abstract notions of distance: geographical, content, social,…
- Do natural networks share more *quantitative* universals?
- What would these "universals" be?
- How can we make them precise and measure them?
- How can we explain their universality?
- This is the domain of *social network theory*
- Sometimes also referred to as *link analysis*

Some Interesting Quantities

- *Connected components*:
  - how many, and how large?
- *Network diameter*:
  - maximum (worst-case) or average?
  - exclude infinite distances? (disconnected components)
  - the small-world phenomenon
- *Clustering*:
  - to what extent that links tend to cluster "locally"?
  - what is the balance between local and long-distance connections?
  - what roles do the two types of links play?
- *Degree distribution*:
  - what is the typical degree in the network?
  - what is the overall distribution?♪

The small-world effect

Consider an undirected network, and let us define ℓ to be the mean geodesic (i.e., shortest) distance between vertex pairs in a network:

$$\ell = \frac{1}{\frac{1}{2}n(n+1)} \sum_{i \geq j} d_{ij}, \qquad (1)$$

where d_{ij} is the geodesic distance from vertex i to vertex j.

$$\ell^{-1} = \frac{1}{\frac{1}{2}n(n+1)} \sum_{i \geq j} d_{ij}^{-1}.$$

Transitivity – the clustering coefficient

In the language of social networks, the friend of your friend is likely also to be your friend. In terms of network topology, transitivity means the presence of a heightened number of triangles in the network—sets of three vertices each of which is connected to each of the others. It can be quantified by defining a clustering coefficient C thus:

$$C = \frac{3 \times \text{ number of triangles in the network}}{\text{number of connected triples of vertices}}, \qquad (3)$$

where a "connected triple" means a single vertex with edges running to an unordered pair of others (see Fig. 5).

Transitivity – the clustering coefficient

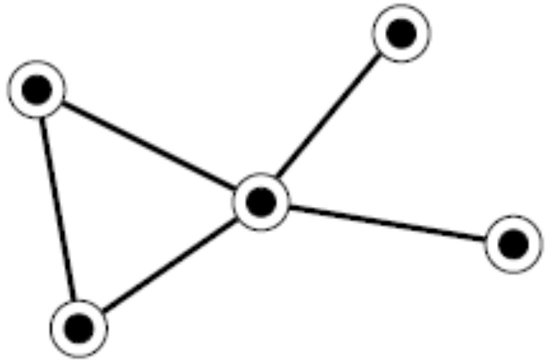


FIG. 5 Illustration of the definition of the clustering coefficient C, Eq. (3). This network has one triangle and eight connected triples, and therefore has a clustering coefficient of $3 \times 1/8 = \frac{3}{8}$. The individual vertices have local clustering coefficients, Eq. (5), of 1, 1, $\frac{1}{6}$, 0 and 0, for a mean value, Eq. (6), of $C = \frac{13}{30}$.

$$C = \frac{6\times \text{ number of triangles in the network}}{\text{number of paths of length two}}, \qquad (4)$$

Transitivity – the clustering coefficient

An alternative definition of the clustering coefficient, also widely used, has been given by Watts and Strogatz [416], who proposed defining a local value

$$C_i = \frac{\text{number of triangles connected to vertex } i}{\text{number of triples centered on vertex } i}. \qquad (5)$$

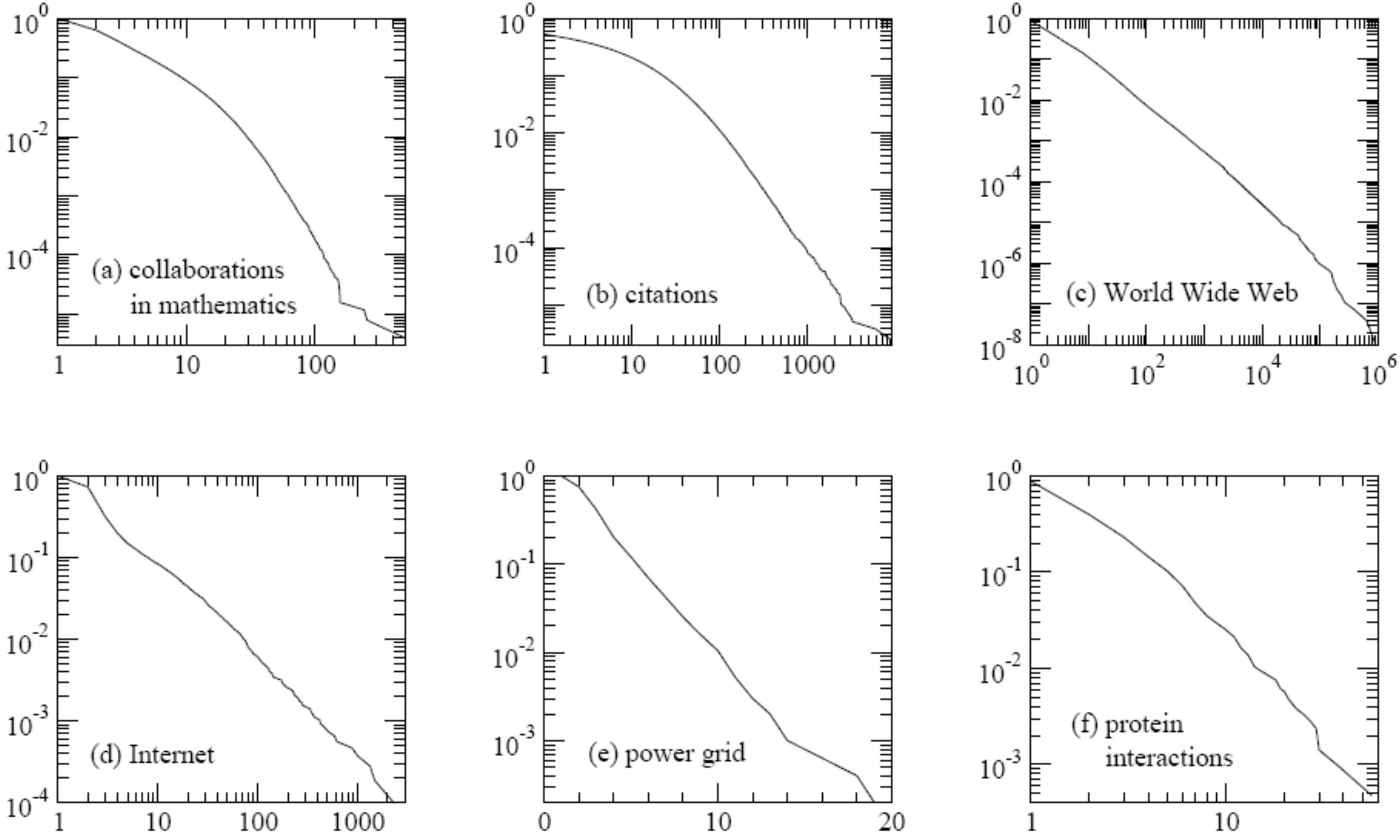
For vertices with degree 0 or 1, for which both numerator and denominator are zero, we put $C_i = 0$. Then the clustering coefficient for the whole network is the average

$$C = \frac{1}{n} \sum_i C_i. \qquad (6)$$

Degree distribution

- The **degree** of a vertex in a network is the number of edges incident on (i.e., connected to) that vertex.
- $\mathbf{p_k}$ = the fraction of vertices in the network that have degree k.
- Equivalently, $\mathbf{p_k}$ = the **probability** that a vertex chosen uniformly at random has **degree k**.
- A plot of $\mathbf{p_k}$ for any given network can be formed by a **histogram** of the degrees of vertices.
- This histogram is the **degree distribution** for the network

Degree distributions for six networks



Actor Connectivity (power law)

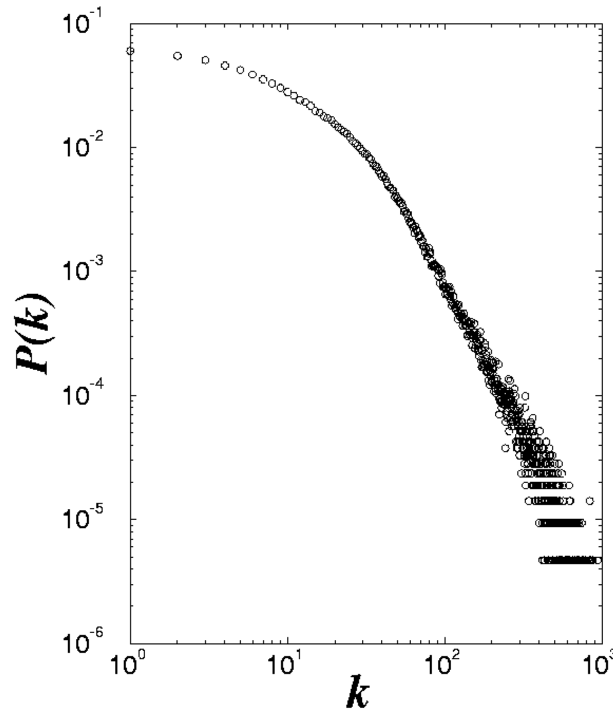
Days of Thunder (1990)
Far and Away (1992)
Eyes Wide Shut (1999)

N = 212,250 actors
$\langle k \rangle = 28.78$

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

$\gamma = 2.3$

Nodes: actors
Links: cast jointly



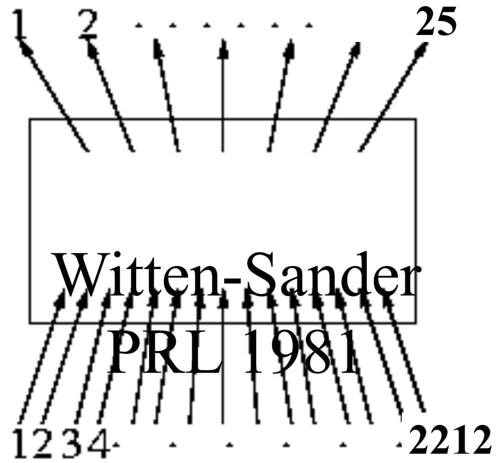
Science Citation Index (power law)

1,000 Most Cited Physicists
Out of over 500,000 E
(see http://www.ssl.n

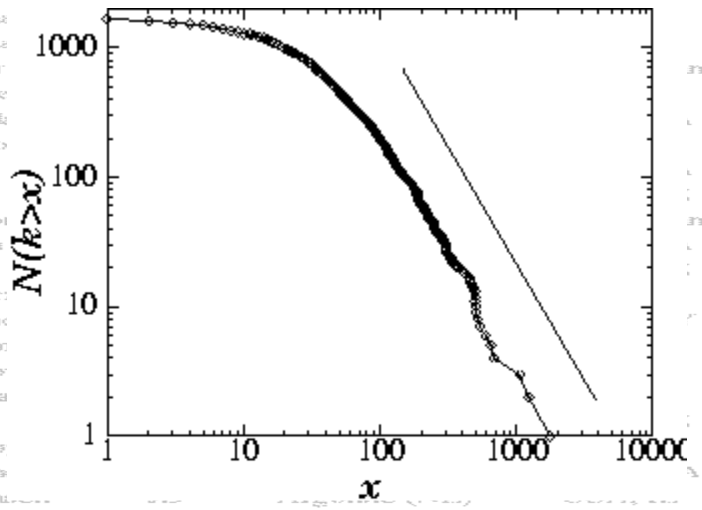
Author name			Country	Field				rank by total cit.
Witten	E		USA, NJ	High				1
Gossard	AC	UCSB (U)	USA, CA	Semi				2
Cava	R		USA, NJ	Supe				3
Batlogg	B		USA, NJ	Supe				4
Ploog	K	Max-Planck (NL)	Germany	Semi				5
Ellis	J	Euro Nuclear Cent.	Switzerland	Astro				6
Fisk	Z	Florida State (U)	USA, FL	Solid				7
Cardona	M	Max Planck (NL)	Germany	Semi				8
Nanopoulos	DV	Texas A&M (U)	USA, TX	High				9
Heeger	AJ	UCSB (U)	USA, CA	Poly				10
Lee*								11
Suzuki*								12
Anderson			NJ	Solid				13
Suzuki*	M				12	898	10417	14
Freem				Solid State (T)	27	389	10411	15
Tanaka					11	963	10404	16
Muller			nd	Superconductivity (E)	82	122	10049	17
Schnee				Superconductivity (E)	63	156	9768	18
Chem				Optics (E)	60	162	9668	19
Morko				Semiconductors (E)	20	477	9668	19
Miller				Semicond	67		9652	21
Chu				Supercond			9453	22
Bednor			nd	Supercond		85	9311	23
Cohen				Solid State (T)	33	284	9311	23
Meng				Superconductivity (E)	86	108	9300	25
Wasze				Superconductivity (E)	57	162	9170	26
Shiran				Supercond		269	8841	27
Wiegm				Semicondu		104	8822	28
Vando				Magnetism		129	8686	29
Uchida					28	301	8520	30
Hor				Superconductivity (E)	72	119	8512	31
Murph				Astronomy (E)	111	76	8439	32
Birgen			A	Superconductivity (E)				
Jorgen				Superconductivity (E)				
Hinks	DG	Argonne (NL)	USA, IL	Superconductivity (E)	37	223	8263	35

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

Nodes: papers
Links: citations



1736 PRL papers (1988)



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

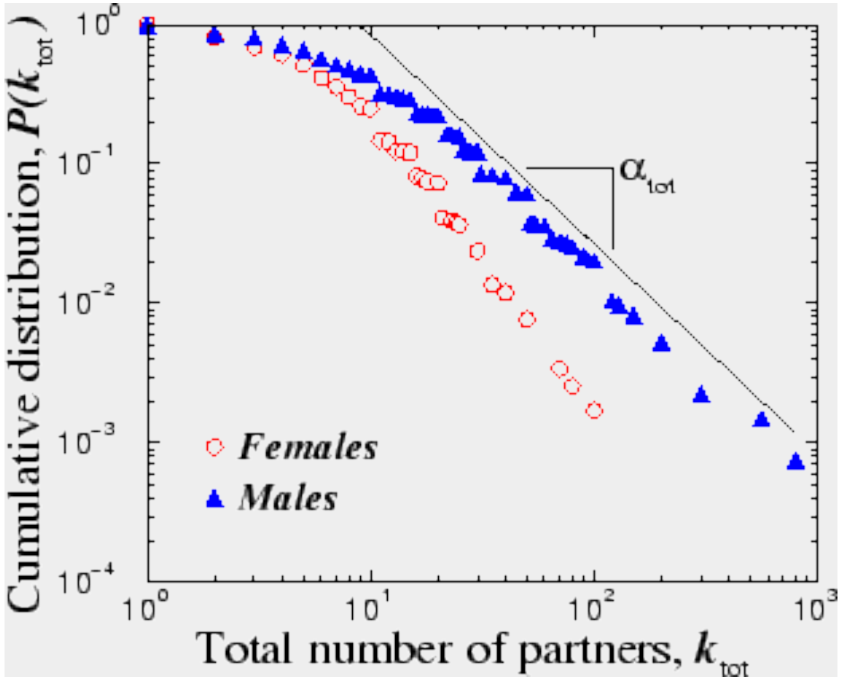
$(\gamma = 3)$

(S. Redner, 1998)

Sex-Web (power law)

Nodes: people (Females; Males)

Links: sexual relationships



4781 Swedes; 18-74;

59% response rate.

Liljeros et al. Nature 2001

Basic statisics for some published networks

	network	type	n	m	z	ℓ	α	$C^{(1)}$	$C^{(2)}$	r	Ref(s).
social	film actors	undirected	449 913	25 516 482	113.43	3.48	2.3	0.20	0.78	0.208	20, 416
	company directors	undirected	7 673	55 392	14.44	4.60	–	0.59	0.88	0.276	105, 323
	math coauthorship	undirected	253 339	496 489	3.92	7.57	–	0.15	0.34	0.120	107, 182
	physics coauthorship	undirected	52 909	245 300	9.27	6.19	–	0.45	0.56	0.363	311, 313
	biology coauthorship	undirected	1 520 251	11 803 064	15.53	4.92	–	0.088	0.60	0.127	311, 313
	telephone call graph	undirected	47 000 000	80 000 000	3.16		2.1				8, 9
	email messages	directed	59 912	86 300	1.44	4.95	1.5/2.0		0.16		136
	email address books	directed	16 881	57 029	3.38	5.22	–	0.17	0.13	0.092	321
	student relationships	undirected	573	477	1.66	16.01	–	0.005	0.001	−0.029	45
	sexual contacts	undirected	2 810				3.2				265, 266
information	WWW `nd.edu`	directed	269 504	1 497 135	5.55	11.27	2.1/2.4	0.11	0.29	−0.067	14, 34
	WWW Altavista	directed	203 549 046	2 130 000 000	10.46	16.18	2.1/2.7				74
	citation network	directed	783 339	6 716 198	8.57		3.0/–				351
	Roget's Thesaurus	directed	1 022	5 103	4.99	4.87	–	0.13	0.15	0.157	244
	word co-occurrence	undirected	460 902	17 000 000	70.13		2.7		0.44		119, 157
technological	Internet	undirected	10 697	31 992	5.98	3.31	2.5	0.035	0.39	−0.189	86, 148
	power grid	undirected	4 941	6 594	2.67	18.99	–	0.10	0.080	−0.003	416
	train routes	undirected	587	19 603	66.79	2.16	–		0.69	−0.033	366
	software packages	directed	1 439	1 723	1.20	2.42	1.6/1.4	0.070	0.082	−0.016	318
	software classes	directed	1 377	2 213	1.61	1.51	–	0.033	0.012	−0.119	395
	electronic circuits	undirected	24 097	53 248	4.34	11.05	3.0	0.010	0.030	−0.154	155
	peer-to-peer network	undirected	880	1 296	1.47	4.28	2.1	0.012	0.011	−0.366	6, 354
biological	metabolic network	undirected	765	3 686	9.64	2.56	2.2	0.090	0.67	−0.240	214
	protein interactions	undirected	2 115	2 240	2.12	6.80	2.4	0.072	0.071	−0.156	212
	marine food web	directed	135	598	4.43	2.05	–	0.16	0.23	−0.263	204
	freshwater food web	directed	92	997	10.84	1.90	–	0.20	0.087	−0.326	272
	neural network	directed	307	2 359	7.68	3.97	–	0.18	0.28	−0.226	416, 421

A "Canonical" Natural Network has…

- *Few* connected components:
  - often only 1 or a small number, indep. of network size
- *Small* diameter:
  - often a constant independent of network size (like 6)
  - or perhaps growing only logarithmically with network size or even shrink?
  - typically exclude infinite distances
- A *high* degree of clustering:
  - considerably more so than for a random network
  - in tension with small diameter
- A *heavy-tailed* degree distribution:
  - a small but reliable number of high-degree vertices
  - often of *power law* form♪

Probabilistic Models of Networks

- All of the network generation models we will study are *probabilistic* or *statistical* in nature
- They can generate networks of any size
- They often have various *parameters* that can be set:
  - size of network generated
  - average degree of a vertex
  - fraction of long-distance connections
- The models generate a *distribution* over networks
- Statements are always *statistical* in nature:
  - *with high probability*, diameter is small
  - on average, degree distribution has heavy tail
- Thus, we're going to need some basic statistics and probability theory♪

Social Network Analysis

- Social Network Introduction
- Statistics and Probability Theory
- Models of Social Network Generation
- Networks in Biological System
- Mining on Social Network
- Summary

Probability and Random Variables

- A *random variable* X is simply a variable that *probabilistically* assumes values in some set
  - set of possible values sometimes called the *sample space* S of X
  - sample space may be small and simple or large and complex
    - S = {Heads, Tails}, X is outcome of a coin flip
    - S = {0,1,…,U.S. population size}, X is number voting democratic
    - S = all networks of size N, X is generated by *preferential attachment*

- Behavior of X determined by its *distribution* (or *density*)
  - for each value x in S, specify Pr[X = x]
  - these probabilities sum to exactly 1 (mutually exclusive outcomes)
  - complex sample spaces (such as large networks):
    - distribution often defined *implicitly* by simpler components
    - might specify the probability that each *edge* appears independently
    - this *induces* a probability distribution over *networks*
    - may be difficult to *compute* induced distribution♪

Some Basic Notions and Laws

- *Independence*:
  - let X and Y be random variables
  - independence: for any x and y, Pr[X = x & Y = y] = Pr[X=x]Pr[Y=y]
  - intuition: value of X does not influence value of Y, vice-versa
  - dependence:
    - e.g. X, Y coin flips, but Y is always opposite of X
- *Expected (mean) value* of X:
  - only makes sense for *numeric* random variables
  - "average" value of X according to its distribution
  - formally, E[X] = Σ (Pr[X = x] X), sum is over all x in S
  - often denoted by μ
  - *always* true: E[X + Y] = E[X] + E[Y]
  - true only for *independent* random variables: E[XY] = E[X]E[Y]
- *Variance* of X:
  - Var(X) = E[(X – μ)^2]; often denoted by σ^2
  - *standard deviation* is sqrt(Var(X)) = σ
- *Union bound:*
  - for any X, Y, Pr[X=x & Y=y] <= Pr[X=x] + Pr[Y=y]

Convergence to Expectations

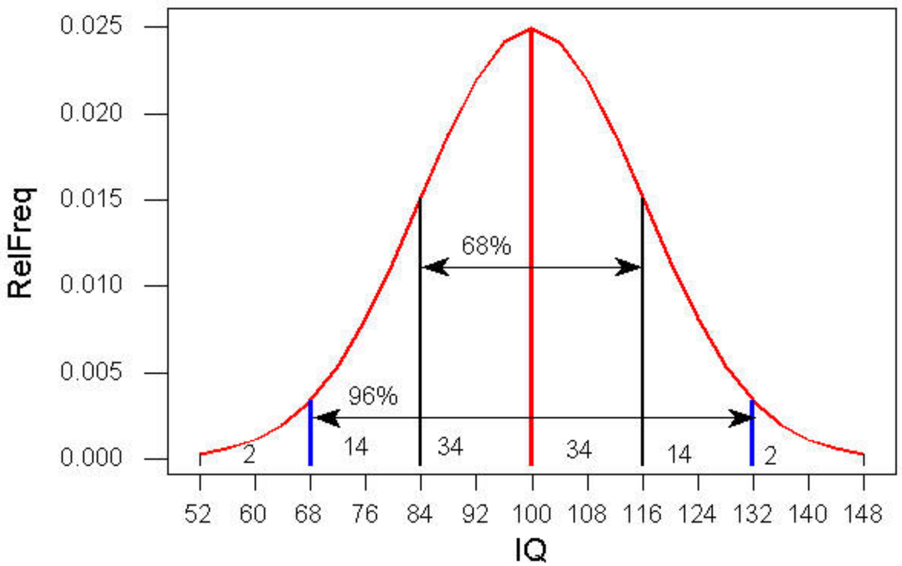
- Let $X_1, X_2, \ldots, X_n$ be:
  - *independent* random variables
  - with the *same* distribution Pr[X=x]
  - expectation μ = E[X] and variance σ^2
  - independent and identically distributed (i.i.d.)
  - essentially n repeated "trials" of the same experiment
  - natural to examine r.v. $Z = (1/n) \Sigma X_i$, where sum is over i=1,…,n
  - example: number of heads in a sequence of coin flips
  - example: degree of a vertex in the random graph model
  - E[Z] = E[X]; what can we say about the *distribution* of Z?

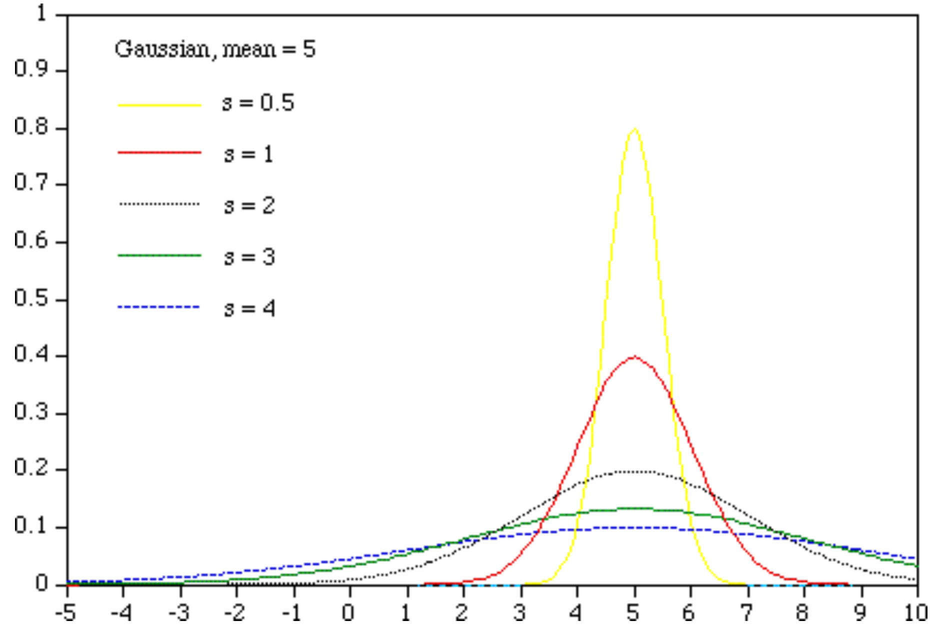
- *Central Limit Theorem:*
  - as n becomes large, Z becomes *normally distributed*
    - with expectation μ and variance σ^2/n

The Normal Distribution

- The *normal* or *Gaussian* density:
  - applies to continuous, real-valued random variables
  - characterized by mean (average) m and standard deviation s
  - *density* at x is defined as
    - $(1/(\sigma \text{ sqrt}(2\pi))) \exp(-(x-\mu)^2/2\sigma^2)$
    - special case $\mu = 0$, $\sigma = 1$: a $\exp(-x^2/b)$ for some constants a,b > 0
  - peaks at $x = \mu$, then dies off *exponentially* rapidly
  - the classic "bell-shaped curve"
    - exam scores, human body temperature,
  - remarks:
    - can control mean and standard deviation independently
    - can make as "broad" as we like, but always have *finite variance*♪

The Normal Distribution



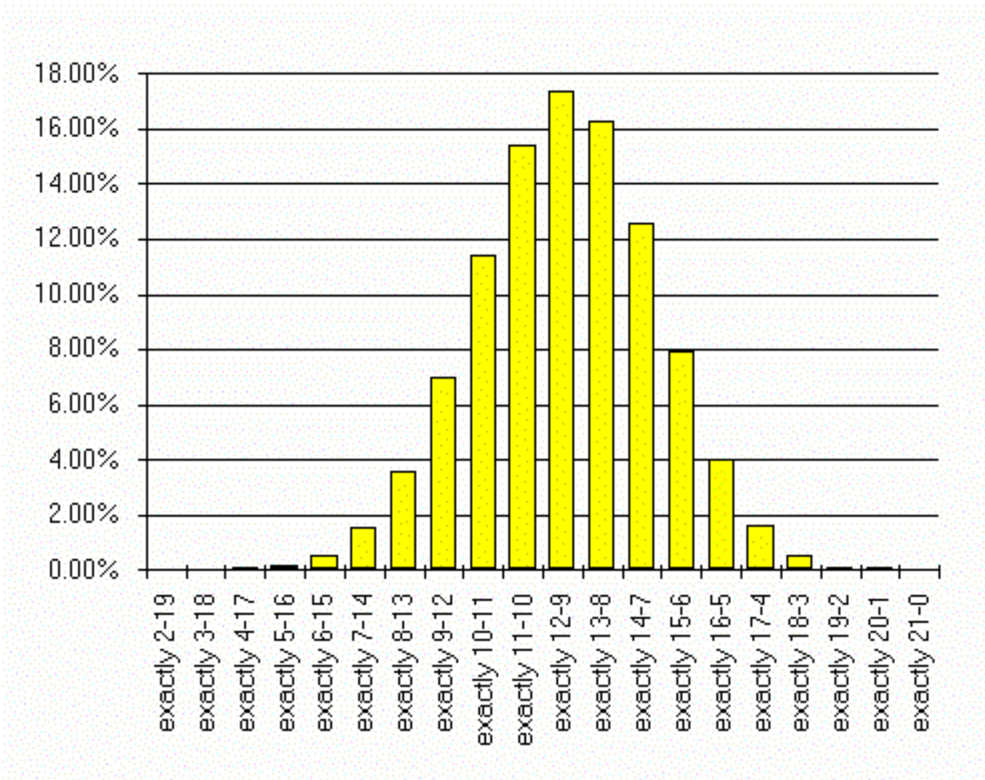


The Binomial Distribution

- coin with Pr[heads] = p, flip n times
- probability of getting exactly k heads:
  - choose(n,k) $p^k(1-p)^{n-k}$
- for large n and p *fixed:*
  - approximated well by a normal with

    $\mu = np$, $\sigma = \text{sqrt}(np(1-p))$
  - $\sigma/\mu \rightarrow 0$ as n grows
  - leads to strong large deviation bounds

The Binomial Distribution

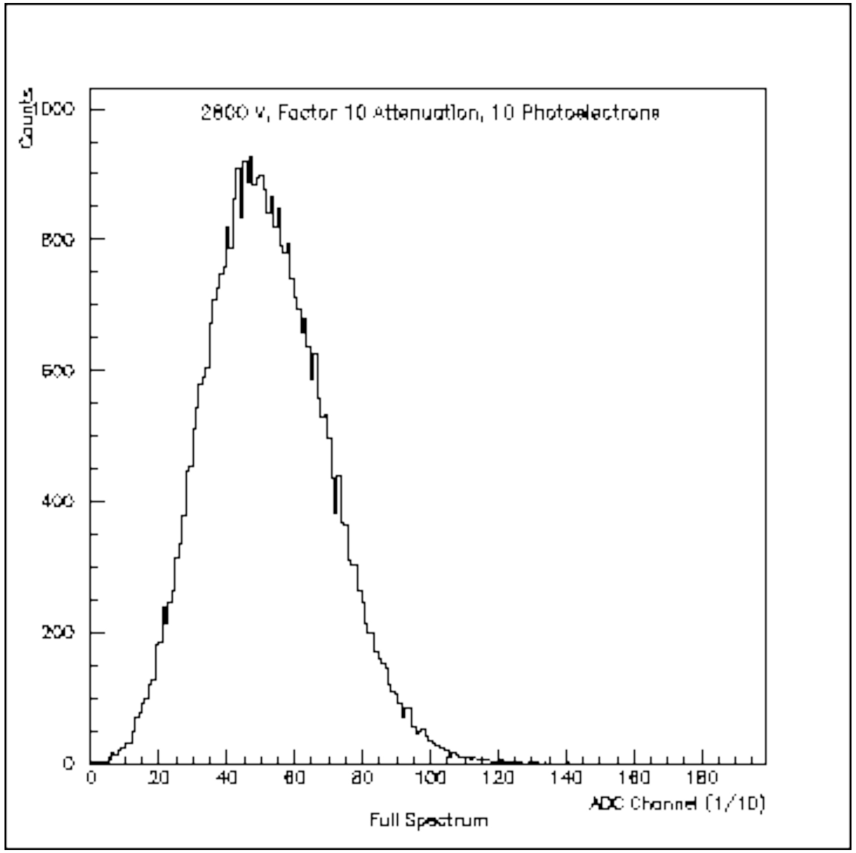


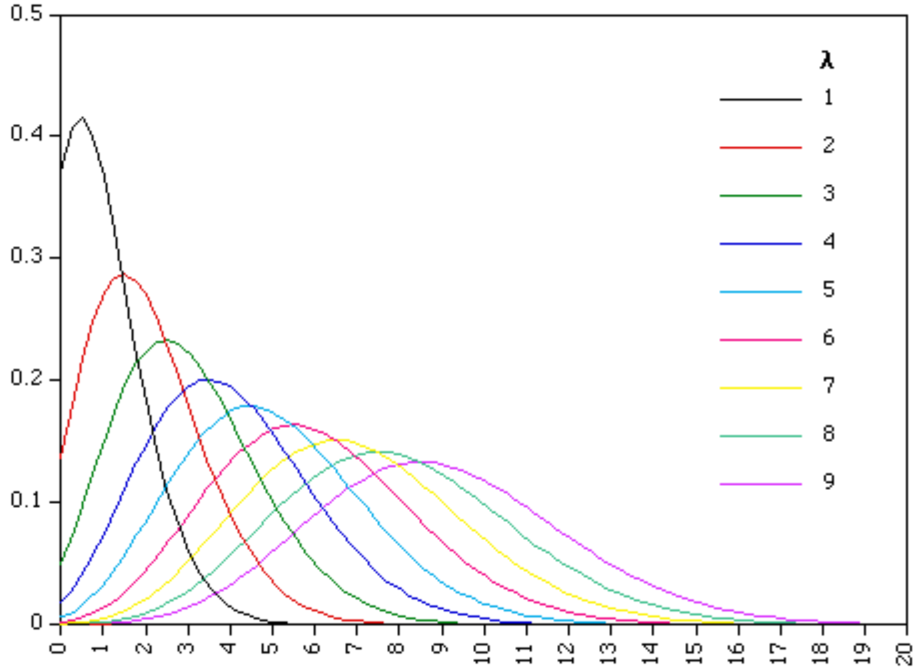
www.professionalgambler.com/ binomial.html ♪

The Poisson Distribution

- like binomial, applies to variables taken on integer values > 0
- often used to model *counts* of events
  - number of phone calls placed in a given time period
  - number of times a neuron fires in a given time period
- single free parameter λ
- probability of exactly x events:
  - $\exp(-\lambda)\, \lambda^x/x!$
  - mean and variance are both λ
- binomial distribution with n large, $p = \lambda/n$ (λ fixed)
  - converges to Poisson with mean λ

The Poisson Distribution



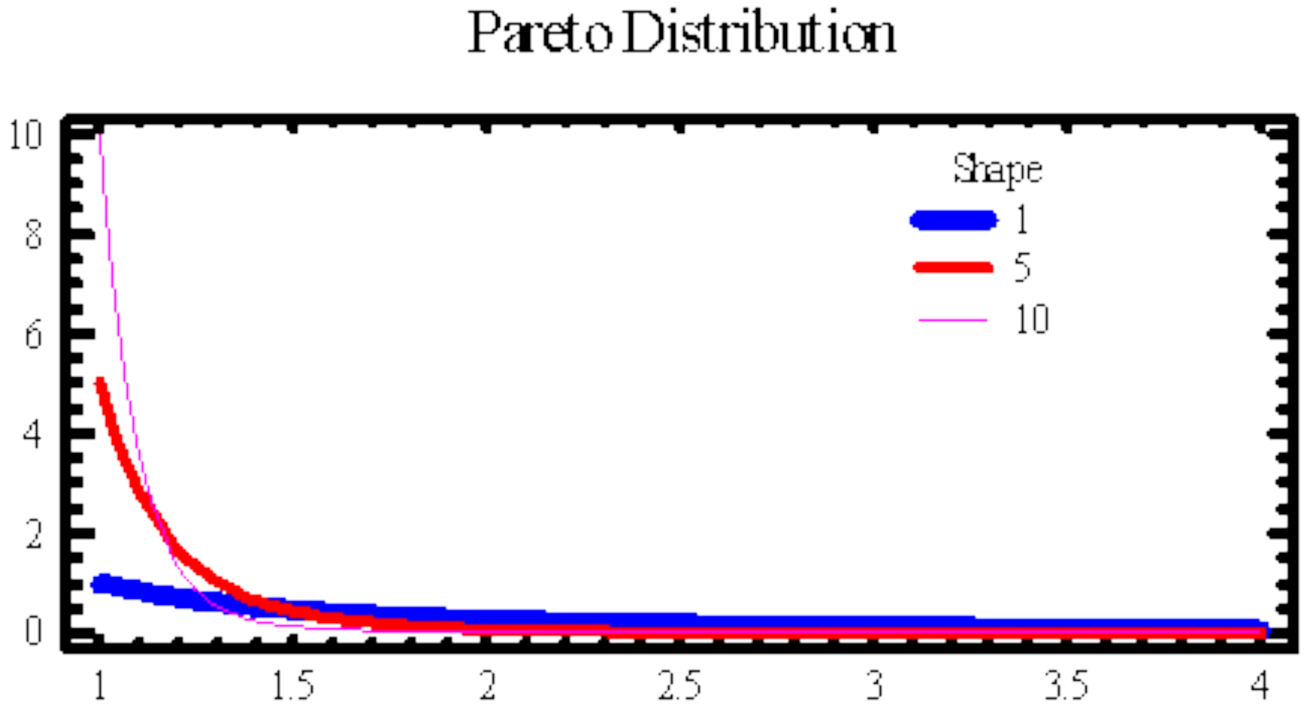


single photoelectron distribution ♪

Heavy-tailed Distributions

- *Pareto* or *power law* distributions:
  - for variables assuming integer values > 0
  - probability of value x ~ 1/x^a
  - typically 0 < a < 2; smaller a gives heavier tail
  - sometimes also referred to as being *scale-free*
- For binomial, normal, and Poisson distributions the tail probabilities approach 0 *exponentially* fast
- Inverse *polynomial* decay vs. *inverse* exponential decay
- What kind of phenomena does this distribution model?
- What kind of process would *generate* it?

Heavy-Tailed Distributions



Distributions vs. Data

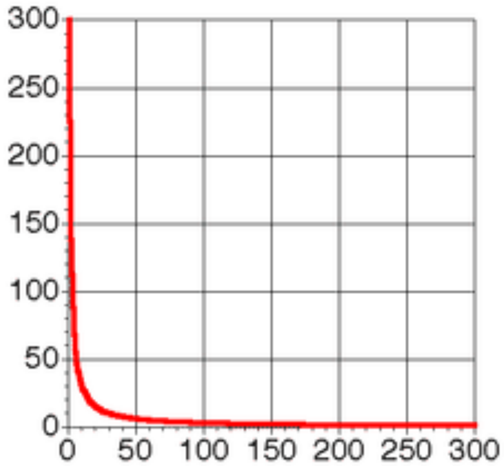
- All these distributions are *idealized models*
- In practice, we do not see distributions, but *data*
- Thus, there will be some *largest* value we observe
- Also, can be difficult to "eyeball" data and choose model
- So how do we distinguish between Poisson, power law, etc?
- Typical procedure:
  - might restrict our attention to a *range* of values of interest
  - accumulate *counts* of observed data into equal-sized bins
  - look at counts on a *log-log plot*
  - note that
    - power law:
      - $\log(\Pr[X = x]) = \log(1/x^{\alpha}) = -\alpha \log(x)$
      - linear, slope $-\alpha$
    - Normal:
      - $\log(\Pr[X = x]) = \log(a \exp(-x^2/b)) = \log(a) - x^2/b$
      - non-linear, concave near mean
    - Poisson:
      - $\log(\Pr[X = x]) = \log(\exp(-\lambda)\, \lambda^x/x!)$
      - also non-linear

Zipf's Law

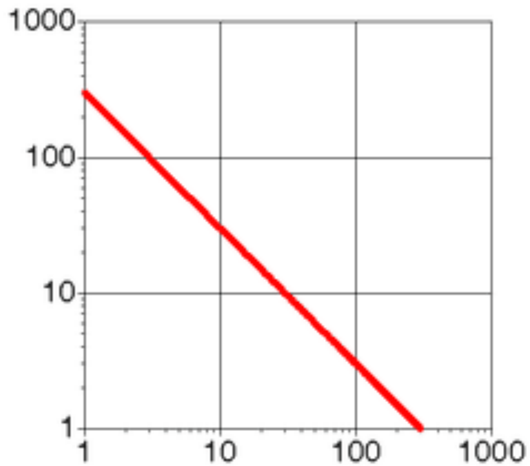
- Look at the frequency of English words:
  - "the" is the most common, followed by "of", "to", etc.
  - claim: frequency of the n-th most common ~ 1/n (power law, $\alpha = 1$)
- General theme:
  - *rank* events by their *frequency of occurrence*
  - resulting distribution often is a power law!
- Other examples:
  - North America city sizes
  - personal income
  - file sizes
  - genus sizes (number of species)
- People seem to dither over exact form of these distributions (e.g. value of α), but not heavy tails♪

Zipf's Law

The same data plotted on linear and logarithmic scales.
Both plots show a Zipf distribution with 300 datapoints



Linear scales on both axes



Logarithmic scales on both axes