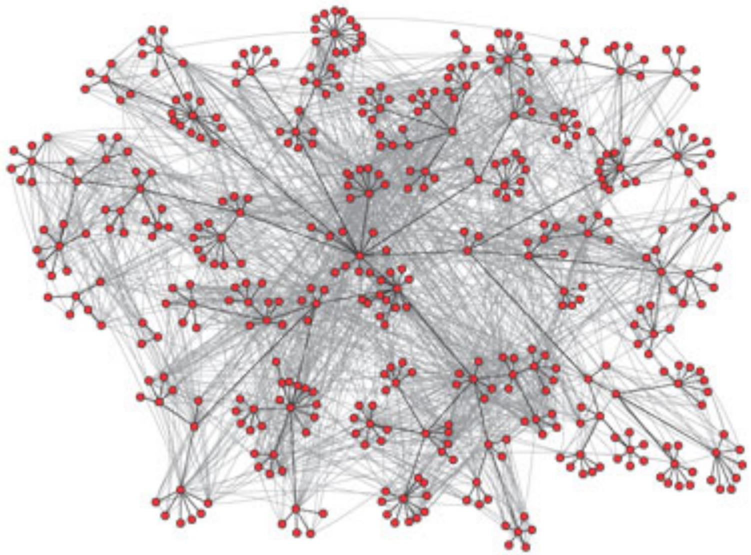
Network Formation Processes: Power-law degrees and Preferential Attachment

CS224W: Analysis of Networks
Jure Leskovec, Stanford University
http://cs224w.stanford.edu

Next Time: New Topics

Observations	Models	Algorithms
Small diameter, Edge clustering	Erdös-Renyi model, Small-world model	Decentralized search
Patterns of signed edge creation	Structural balance, Theory of status	Models for predicting edge signs
Viral Marketing, Blogosphere, Memetracking	Independent cascade model, Game theoretic model	Influence maximization, Outbreak detection, LIM
Scale-Free	Preferential attachment, Copying model	PageRank, Hubs and authorities
Densification power law, Shrinking diameters	Microscopic model of evolving networks	Link prediction, Supervised random walks
Strength of weak ties, Core-periphery	Kronecker Graphs	Community detection: Girvan-Newman, Modularity

Network Formation Processes

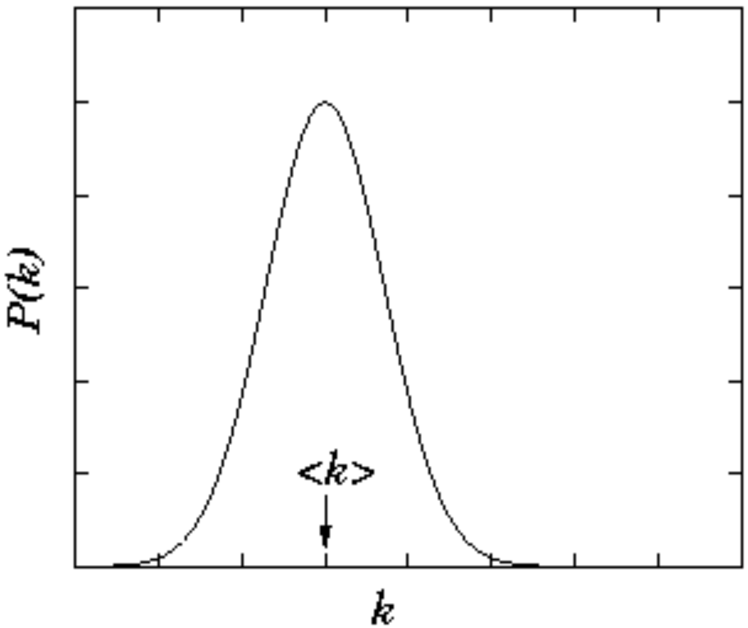


What do we observe that needs explaining?

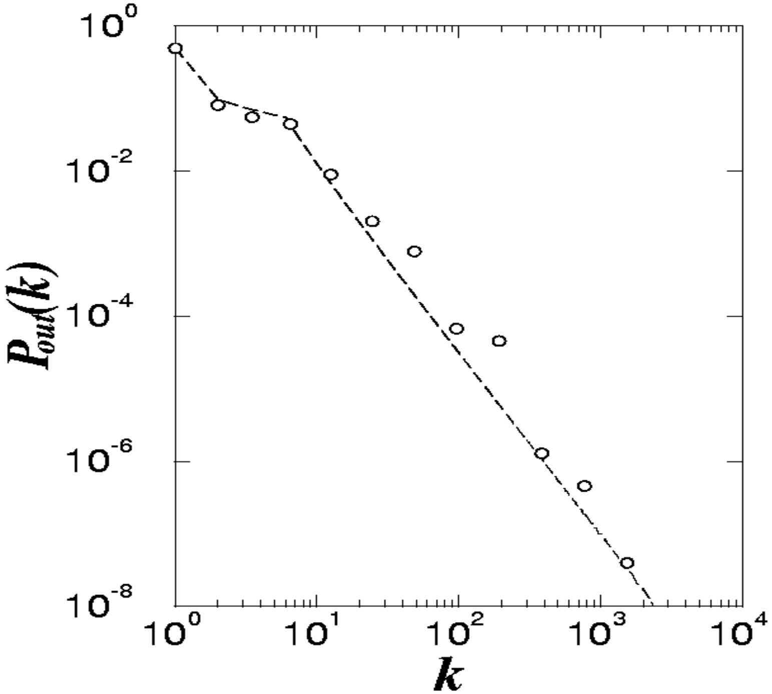
- **Small-world model:**
  - Diameter
  - Clustering coefficient
- **Preferential Attachment:**
  - **Node degree distribution**
    - What fraction of nodes has degree k (as a function of k)?
    - Prediction from simple random graph models: $p(k)$ = **exponential function of** k
    - **Observation: Often a power-law:** $p(k) \propto k^{-\alpha}$

Degree Distributions

Expected based on G_{np}



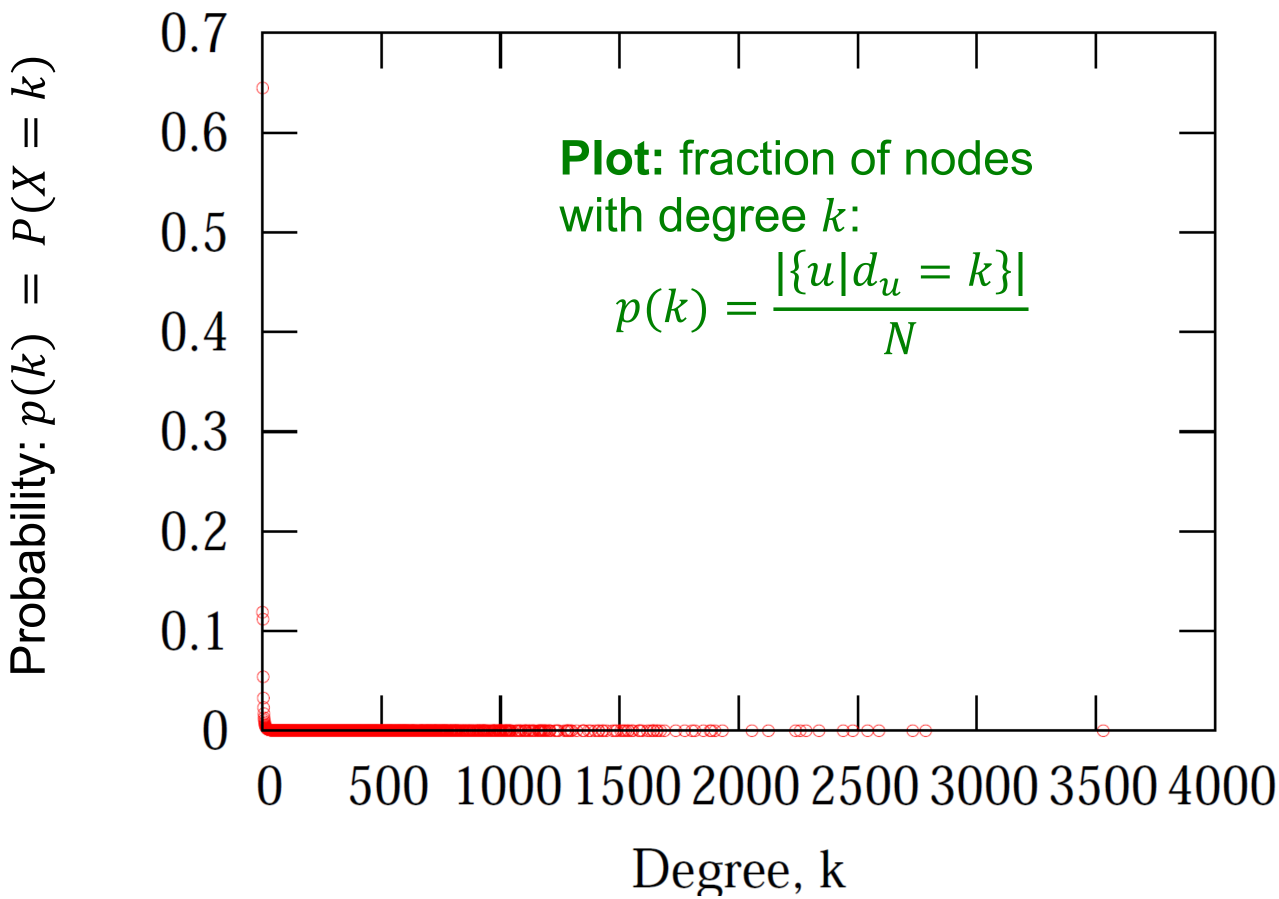
Found in data



$$P(k) \propto k^{-\alpha}$$

Node Degrees in Networks

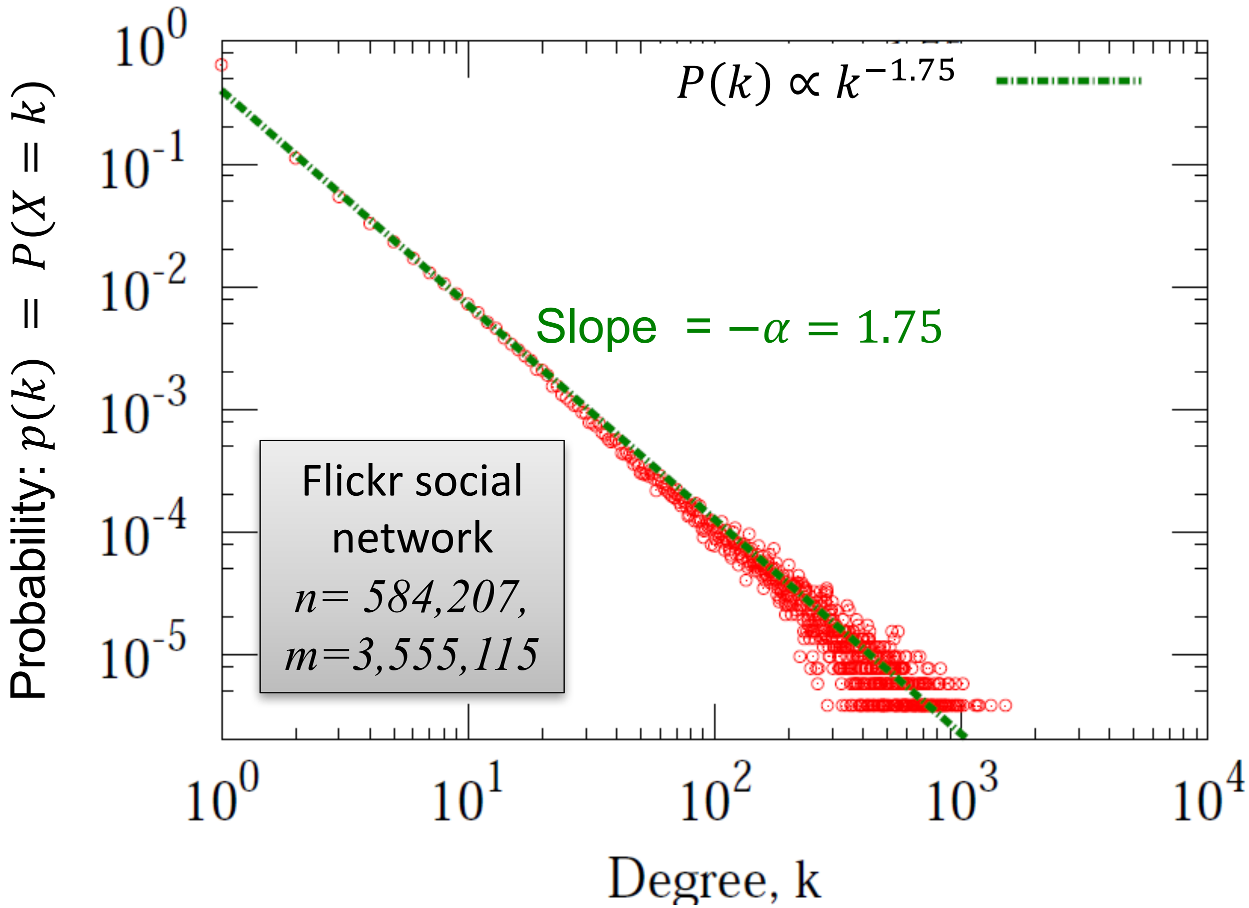
- **Take a network, plot a histogram of $P(k)$ vs. k**



Flickr social network
$n = 584{,}207$, $m = 3{,}555{,}115$

Node Degrees in Networks

- **Plot the same data on *log-log* scale:**

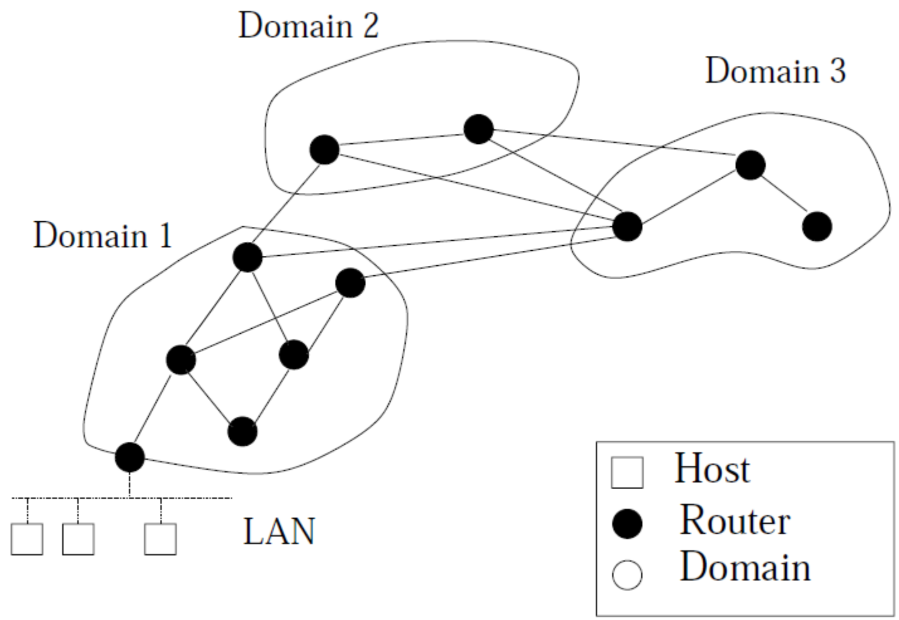


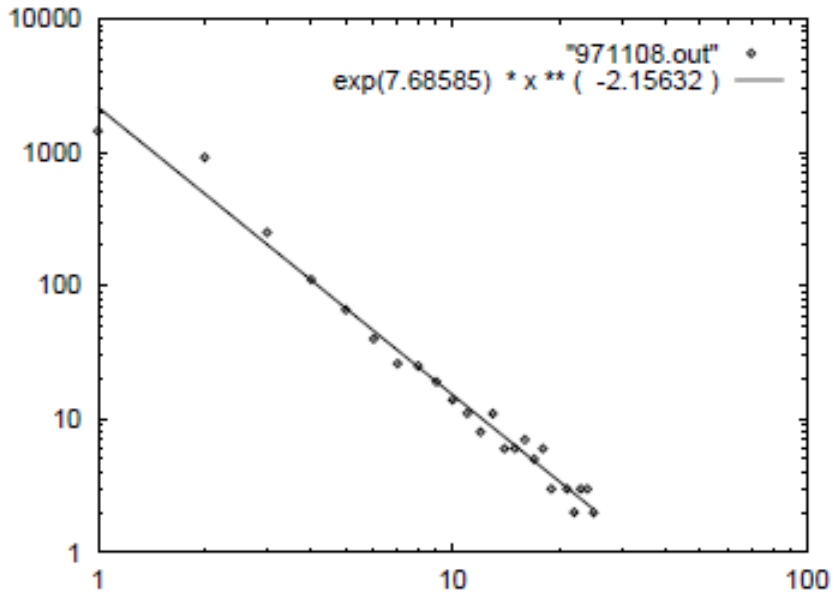
How to distinguish:
$P(k) \propto \exp(-k)$ **vs.**
$P(k) \propto k^{-\alpha}$ **?**

Take logarithms:
if $y = f(x) = e^{-x}$ then
$\log(y) = -x$
If $y = x^{-\alpha}$ then
$\log(y) = -\alpha \log(x)$
So on log-log axis power-law looks like a straight line of slope $-\alpha$!

Node Degrees: Faloutsos[3]

- **Internet Autonomous Systems**
  [Faloutsos, Faloutsos and Faloutsos, 1999]

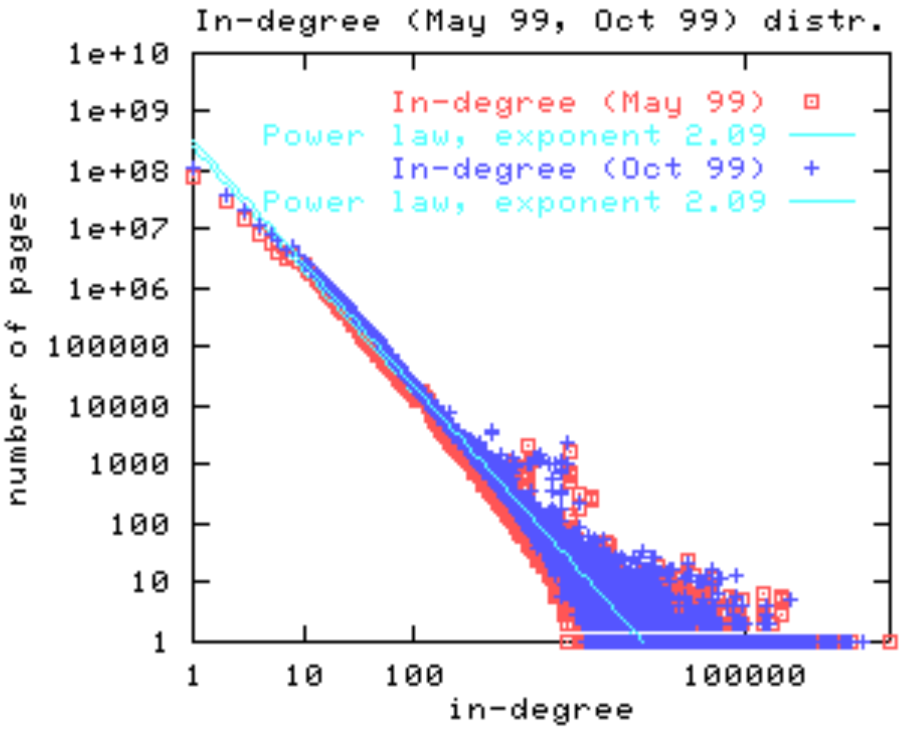


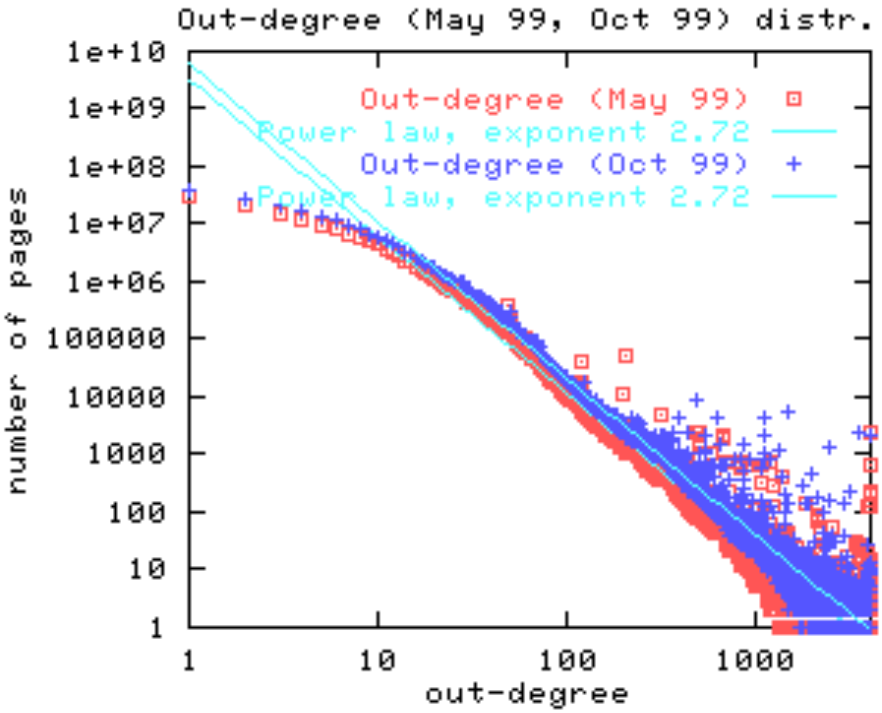


Internet domain topology

Node Degrees: Web

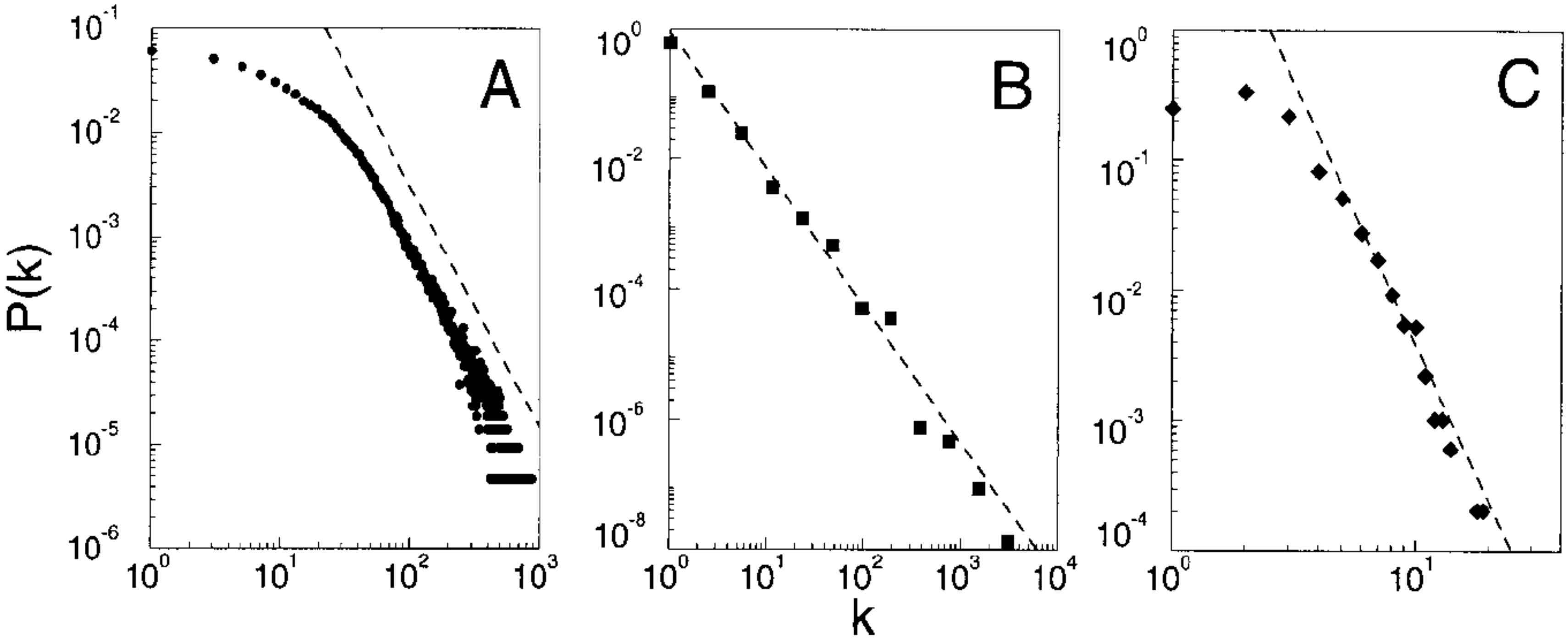
- **The World Wide Web** [Broder et al., 2000]





Node Degrees: Barabasi&Albert

- **Other Networks** [Barabasi-Albert, 1999]

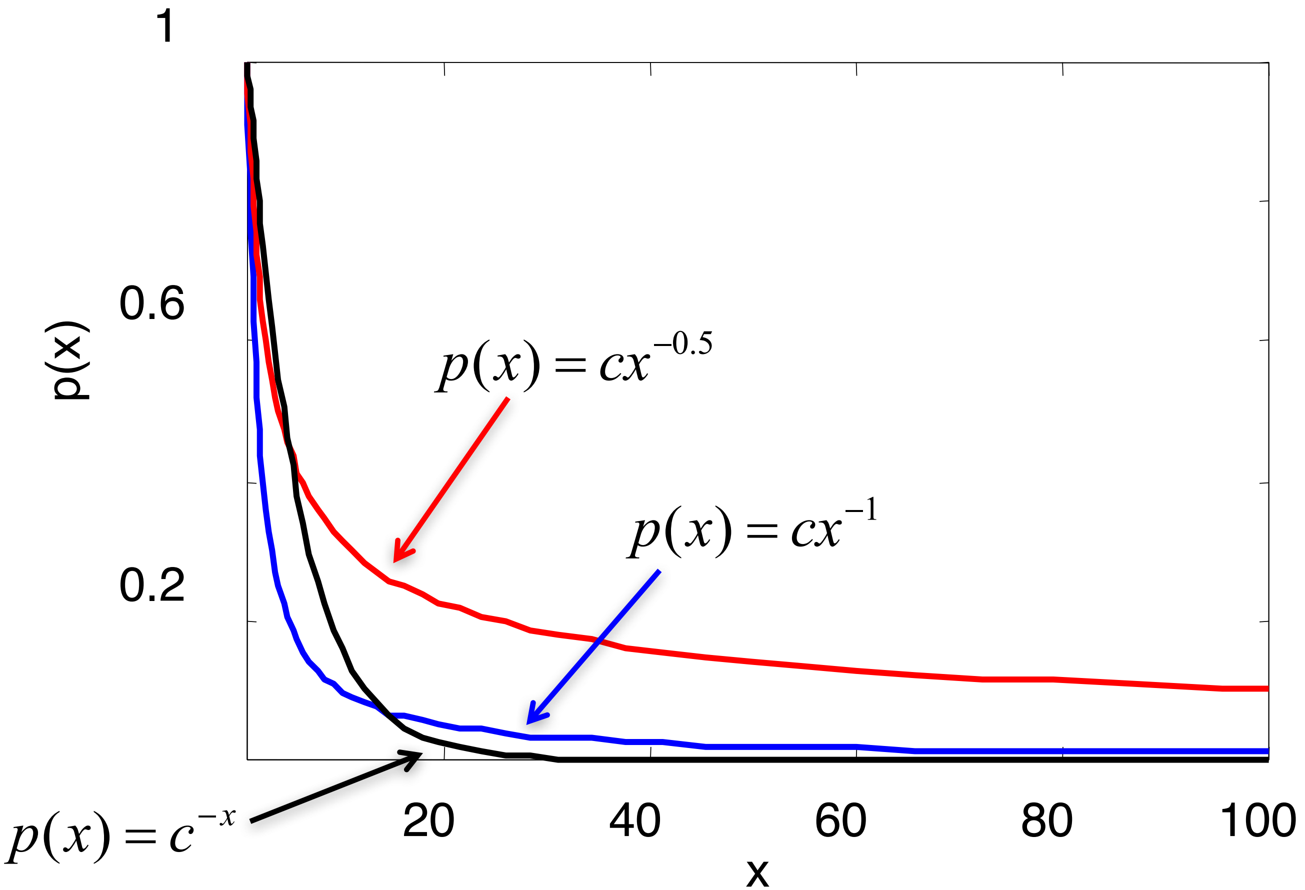


Actor collaborations

Web graph

Power-grid

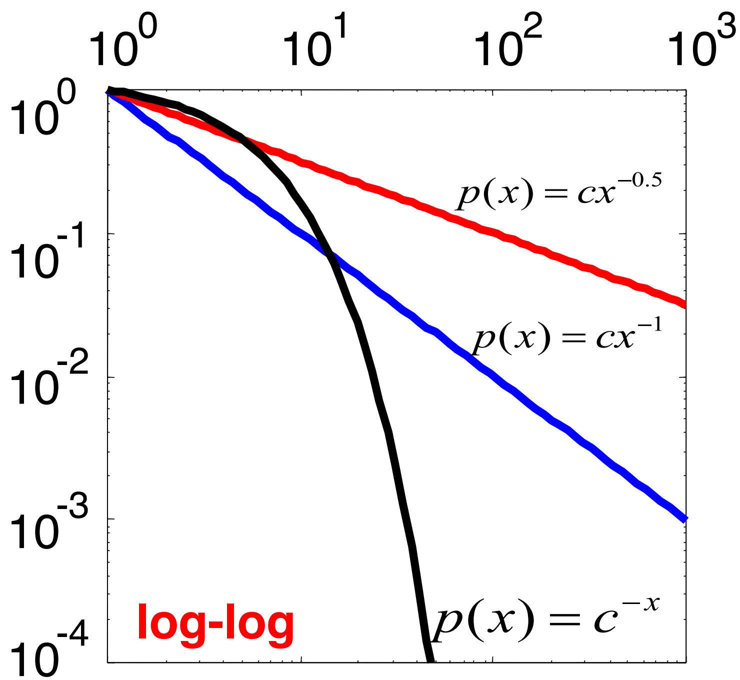
Exponential vs. Power-Law

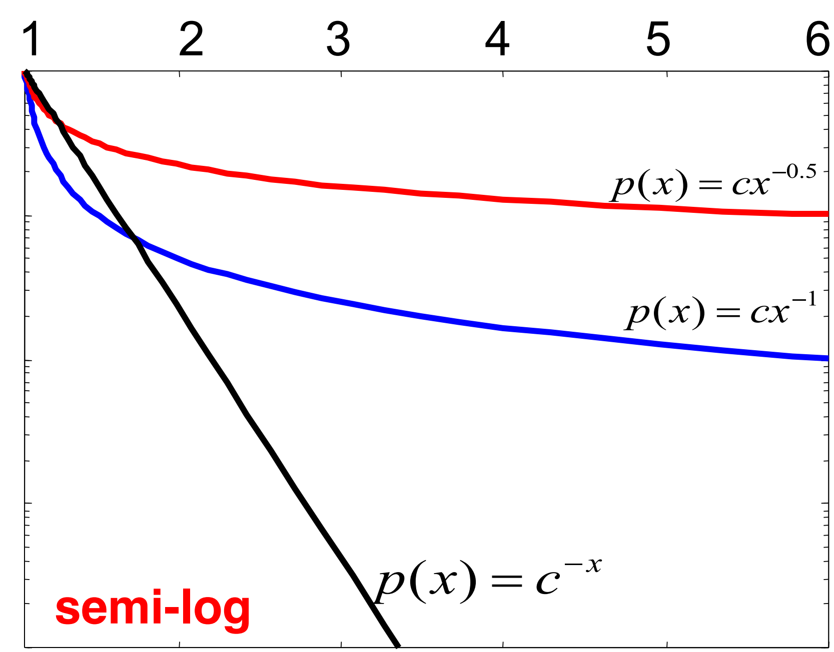


- Above a certain x value, the power law is always higher than the exponential!

Exponential vs. Power-Law

- **Power-law vs. Exponential on log-log and semi-log (log-lin) scales**

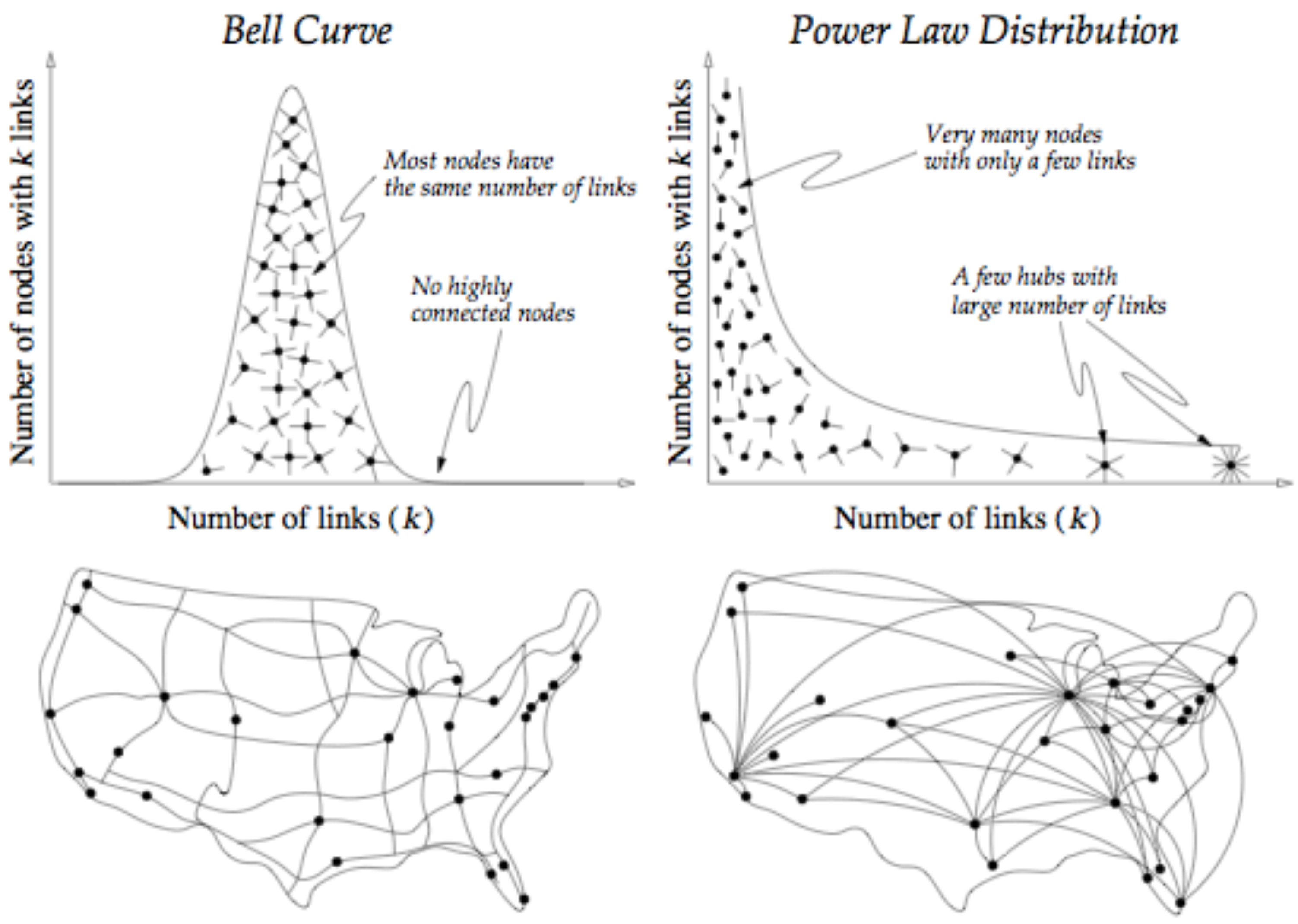




x … logarithmic axis
y … logarithmic axis

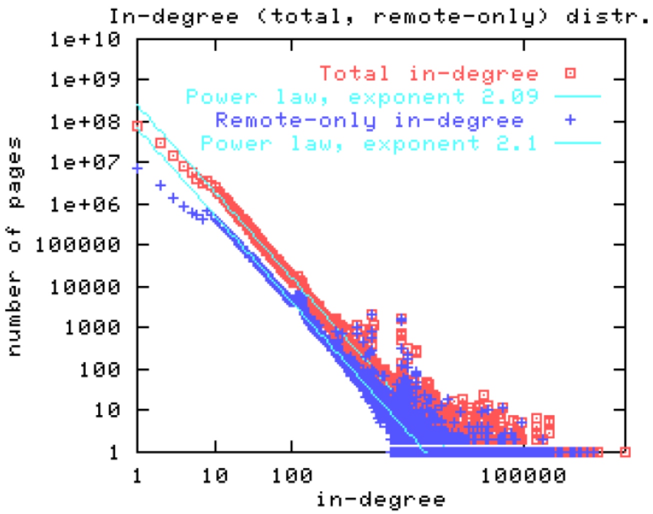
x … linear
y … logarithmic

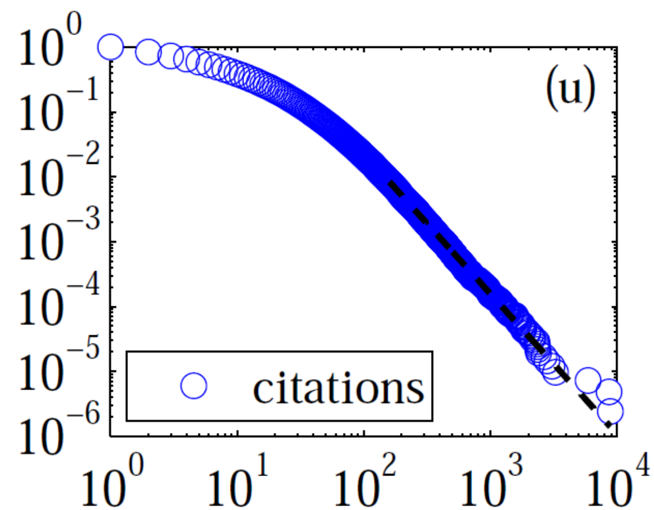
Exponential vs. Power-Law



Power-Law Degree Exponents

- **Power-law degree exponent is typically $2 < \alpha < 3$**
  - **Web graph:**
    - $\alpha_{in} = 2.1$, $\alpha_{out} = 2.4$ [Broder et al. 00]
  - **Autonomous systems:**
    - $\alpha = 2.4$ [Faloutsos^3, 99]
  - **Actor-collaborations:**
    - $\alpha = 2.3$ [Barabasi-Albert 00]
  - **Citations to papers:**
    - $\alpha \approx 3$ [Redner 98]
  - **Online social networks:**
    - $\alpha \approx 2$ [Leskovec et al. 07]





Scale-Free Networks

- **Definition:**
  Networks with a power-law tail in their degree distribution are called "scale-free networks"

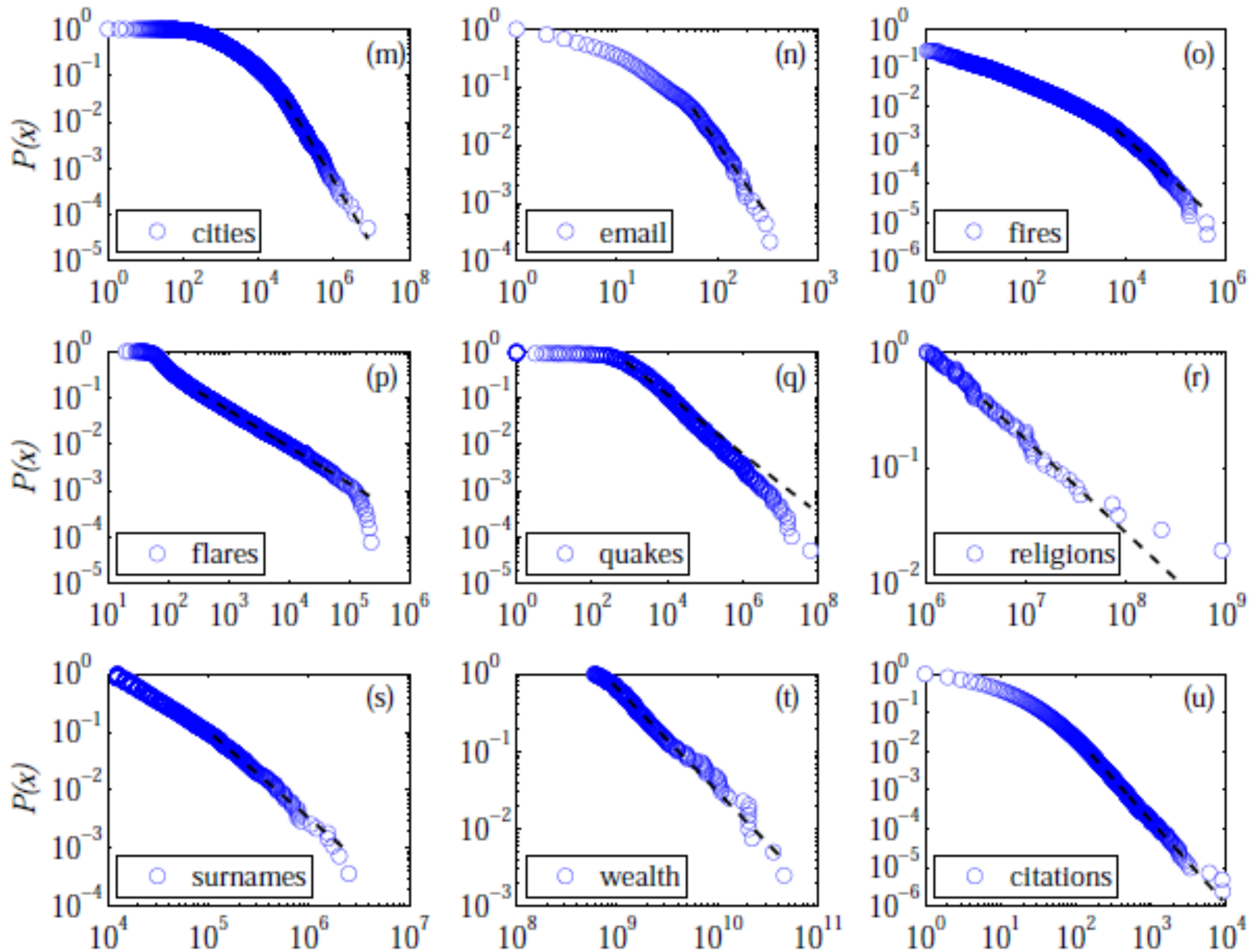
- **Where does the name come from?**
  - **Scale invariance:** There is no characteristic scale
    - **Scale invariance** is that laws do not change if scales of length, energy, or other variables, are multiplied by a common factor
  - **Scale-free function:** $\boldsymbol{f(ax) = a^{\lambda} f(x)}$
    - Power-law function: $\boldsymbol{f(ax) = a^{\lambda} x^{\lambda} = a^{\lambda} f(x)}$

Log() or Exp() are not scale free!

$f(ax) = \log(ax) = \log(a) + \log(x) = \log(a) + f(x)$

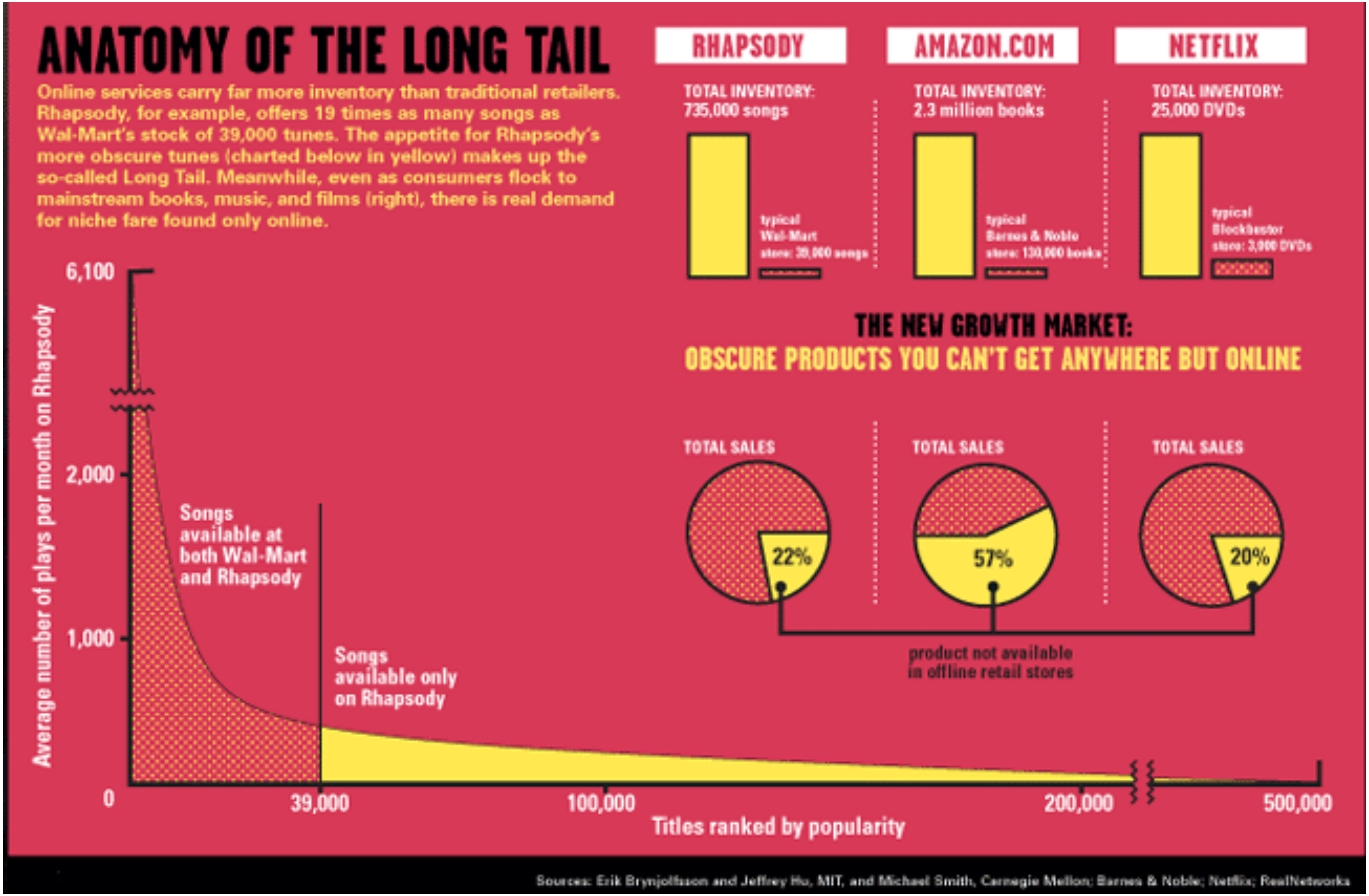
$f(ax) = \exp(ax) = \exp(x)^a = f(x)^a$

Power-Laws are Everywhere



Many other quantities follow heavy-tailed distributions

Anatomy of the Long Tail



Not Everyone Likes Power-Laws ☺

CMU grad-students at the G20 meeting in Pittsburgh in Sept 2009

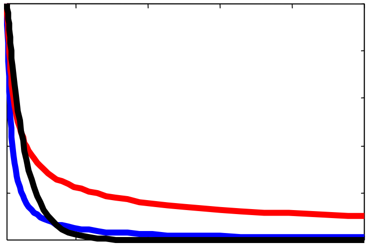
Mathematics of Power-Laws

Heavy Tailed Distributions

- **Degrees are heavily skewed:**

  Distribution $P(X > x)$ is **heavy tailed if:**

$$\lim_{x \to \infty} \frac{P(X > x)}{e^{-\lambda x}} = \infty$$



- **Note:**
  - **Normal PDF:** $p(x) = \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$
  - **Exponential PDF:** $p(x) = \lambda e^{-\lambda x}$
    - **then** $P(X > x) = 1 - P(X \le x) = e^{-\lambda x}$

are not heavy tailed!

[Clauset-Shalizi-Newman 2007]

Heavy Tailed Distributions

- **Various names, kinds and forms:**
  - Long tail, Heavy tail, Zipf's law, Pareto's law
- **Heavy tailed distributions:**
  - **P(x) is proportional to:**

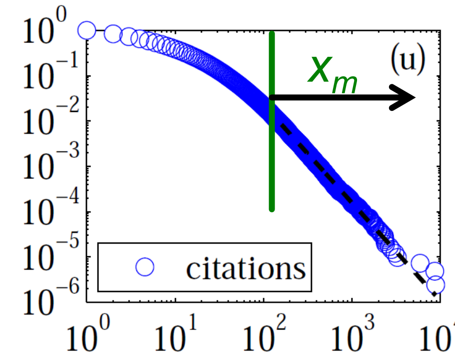
power law	$P(x) \propto \quad x^{-\alpha}$
power law with cutoff	$x^{-\alpha} \mathrm{e}^{-\lambda x}$
stretched exponential	$x^{\beta-1} \mathrm{e}^{-\lambda x^{\beta}}$
log-normal	$\frac{1}{x} \exp\left[-\frac{(\ln x-\mu)^2}{2\sigma^2}\right]$

Mathematics of Power-laws

- **What is the normalizing constant?**

  $p(x) = Z\, x^{-\alpha} \qquad Z = ?$

  - $\boldsymbol{p(x)}$ is a distribution: $\int \boldsymbol{p(x)\,dx = 1}$



p(x) diverges as x→0 so x_m is the minimum value of the power-law distribution x ∈ [x_m, ∞]

Continuous approximation

  - $1 = \int_{x_m}^{\infty} p(x)dx = Z \int_{x_m}^{\infty} x^{-\alpha} dx$
  - $= -\frac{Z}{\alpha - 1}\left[x^{-\alpha+1}\right]_{x_m}^{\infty} = -\frac{Z}{\alpha - 1}\left[\infty^{1-\alpha} - {x_m}^{1-\alpha}\right]$
  - $\Rightarrow Z = (\alpha - 1)x_m^{\alpha-1}$

Need: α > 1 !

$$\boldsymbol{p(x) = \frac{\alpha - 1}{x_m}\left(\frac{x}{x_m}\right)^{-\alpha}}$$

Integral:

$$\int (ax)^n = \frac{(ax)^{n+1}}{a(n+1)}$$

Mathematics of Power-laws

- **What's the expected value of a power-law random variable *X*?**
- $E[X] = \int_{x_m}^{\infty} x\, p(x) dx = Z \int_{x_m}^{\infty} x^{-\alpha+1} dx$
- $= \frac{Z}{2-\alpha}[x^{2-\alpha}]_{x_m}^{\infty} = \frac{(\alpha-1)x_m^{\alpha-1}}{-(\alpha-2)}[\infty^{2-\alpha} - x_m^{2-\alpha}]$

Need: **α > 2 !**

$$\Rightarrow \boldsymbol{E[X] = \frac{\alpha-1}{\alpha-2} x_m}$$

Power-law density:

$p(x) = \frac{\alpha-1}{x_m}\left(\frac{x}{x_m}\right)^{-\alpha}$

$Z = \frac{\alpha-1}{x_m^{1-\alpha}}$

Mathematics of Power-Laws

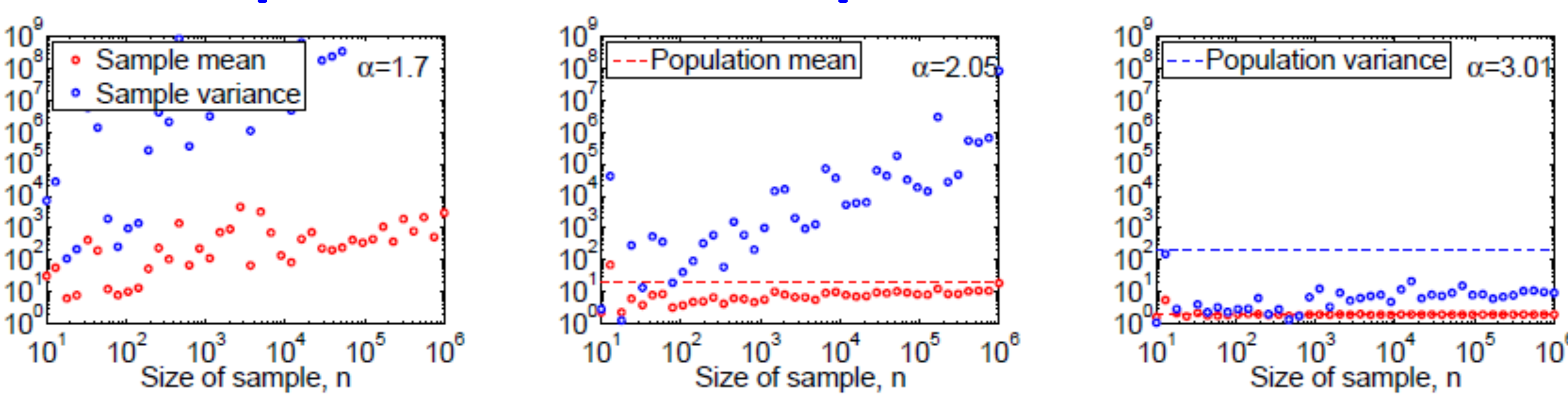
- **Power-laws have infinite moments!**

$$E[X] = \frac{\alpha - 1}{\alpha - 2} x_m$$

  - If $\alpha \leq 2 : E[X] = \infty$
  - If $\alpha \leq 3 : Var[X] = \infty$
    - **Average is meaningless, as the variance is too high!**

In real networks
2 < α < 3 so:
E[X] = const
Var[X] = ∞

- **Consequence: Sample average of *n* samples from a power-law with exponent α**

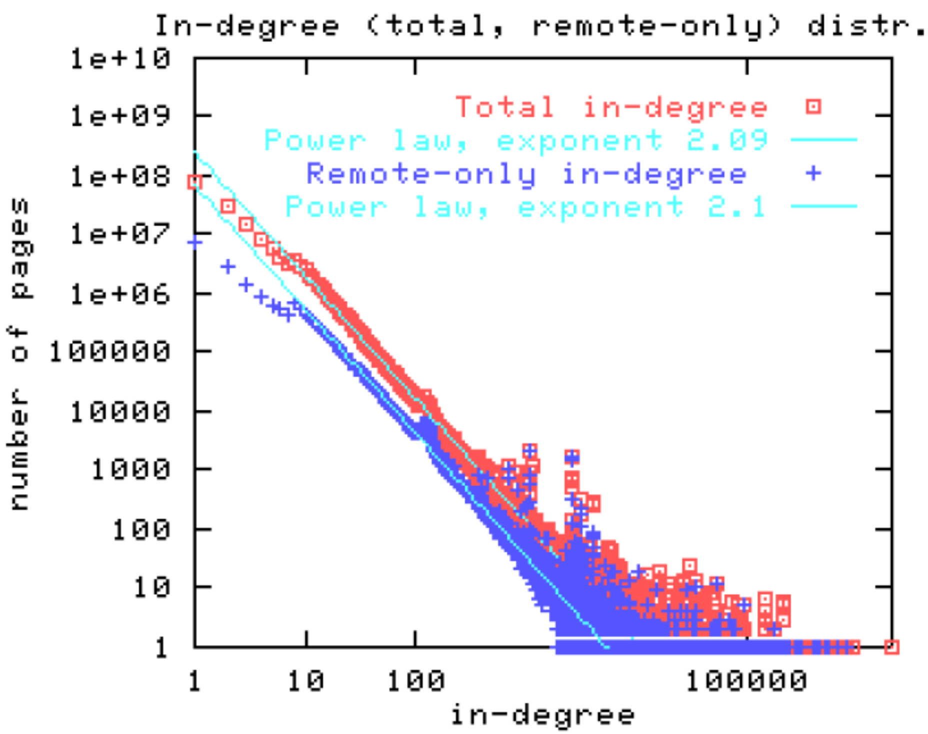


Estimating Power-law Exponent Alpha

Estimating Power-Law Exponent α

Estimating α from data:

- **(1)** Fit a line on log-log axis using least squares:
  - **Solve** $\arg\min_{\alpha} (\log(y) - \alpha \log(x) + b)^2$



BAD!

Estimating Power-Law Exponent α

OK!

Estimating α from data:

- Plot **Complementary CDF (CCDF)** $\boldsymbol{P(X \geq x)}$. Then the estimated $\boldsymbol{\alpha = 1 + \alpha'}$ where $\boldsymbol{\alpha'}$ is the slope of $\boldsymbol{P(X \geq x)}$.

- **Fact:** **If $\boldsymbol{p(x) = P(X = x) \propto x^{-\alpha}}$ then $\boldsymbol{P(X \geq x) \propto x^{-(\alpha-1)}}$**
  - $P(X \geq x) = \sum_{j=x}^{\infty} p(j) \approx \int_x^{\infty} Z\, y^{-\alpha} dy =$
  - $= \frac{Z}{1-\alpha}\left[y^{1-\alpha}\right]_x^{\infty} = \frac{Z}{1-\alpha} x^{-(\alpha-1)}$

Estimating Power-Law Exponent α

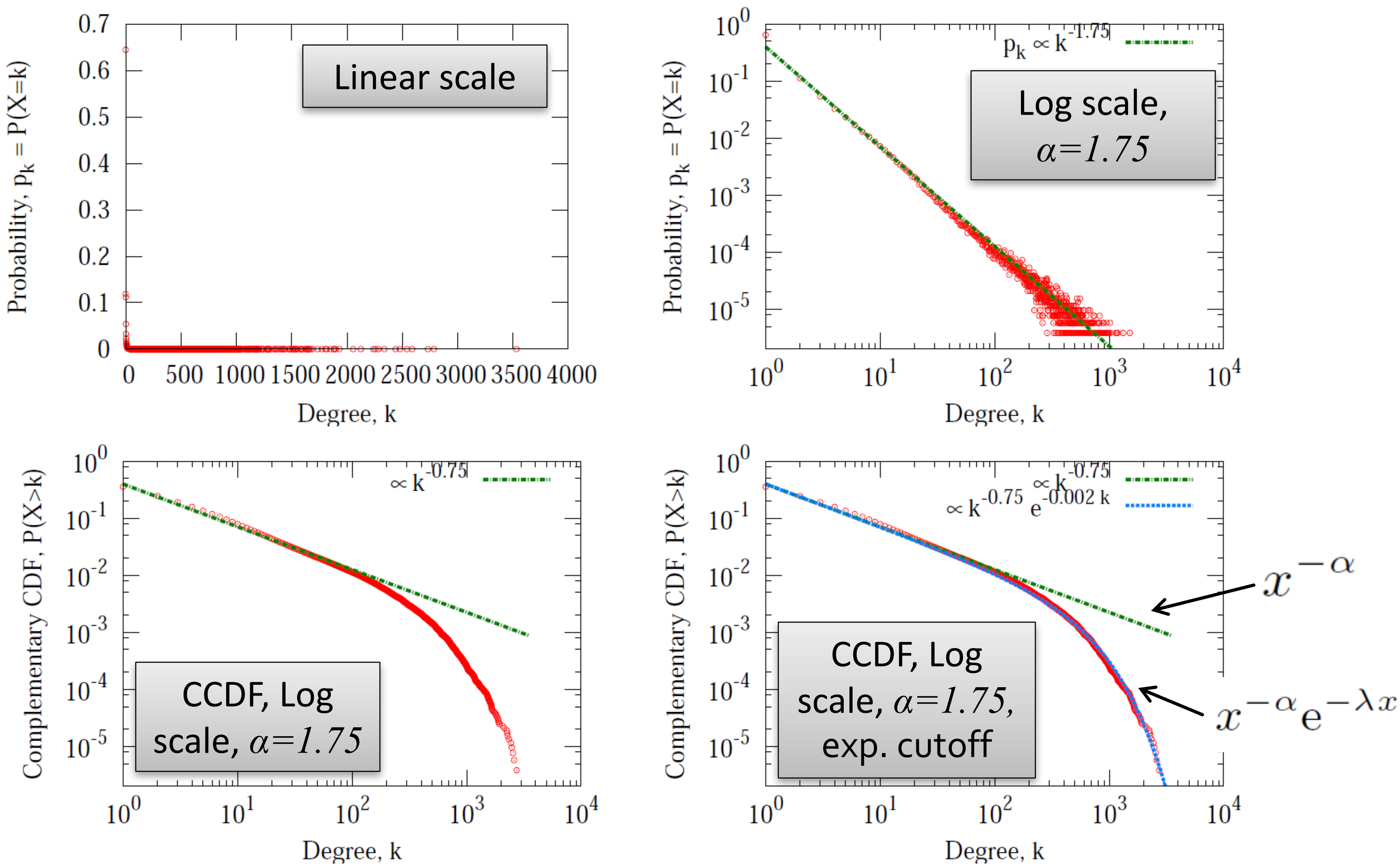
OK!

Estimating α from data:

- **Use maximum likelihood approach:**
  - **The log-likelihood of observed data d_i:**
    - $L(\alpha) = \ln(\prod_i^n p(d_i)) = \sum_i^n \ln p(d_i)$
    - $= \sum_i^n \left(\ln(\alpha - 1) - \ln(x_m) - \alpha \ln\left(\frac{d_i}{x_m}\right)\right)$
  - **Want to find α that max $L(\alpha)$: Set $\frac{\mathrm{dL}(\alpha)}{\mathrm{d}\alpha} = 0$**
  - $\frac{\mathrm{dL}(\alpha)}{\mathrm{d}\alpha} = 0 \;\Rightarrow\; \frac{n}{\alpha-1} - \sum_i^n \ln\left(\frac{d_i}{x_m}\right) = 0$
  - $\Rightarrow \hat{\alpha} = 1 + n\left[\sum_i^n \ln\left(\frac{d_i}{x_m}\right)\right]^{-1}$

Power-law density:
$p(x) = \frac{\alpha - 1}{x_m}\left(\frac{x}{x_m}\right)^{-\alpha}$

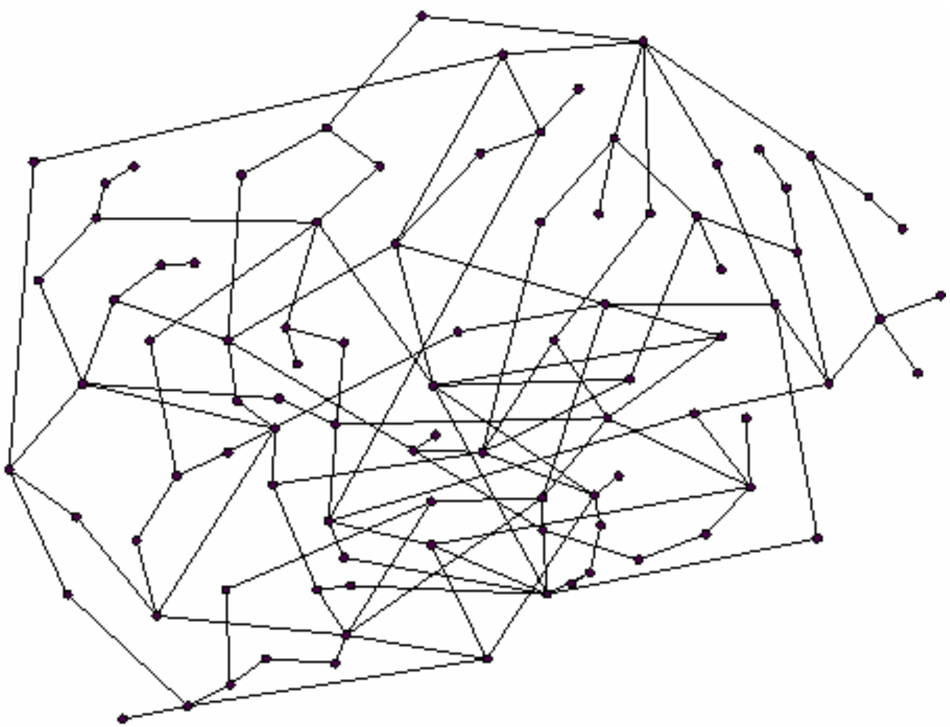
Flickr: Fitting Degree Exponent

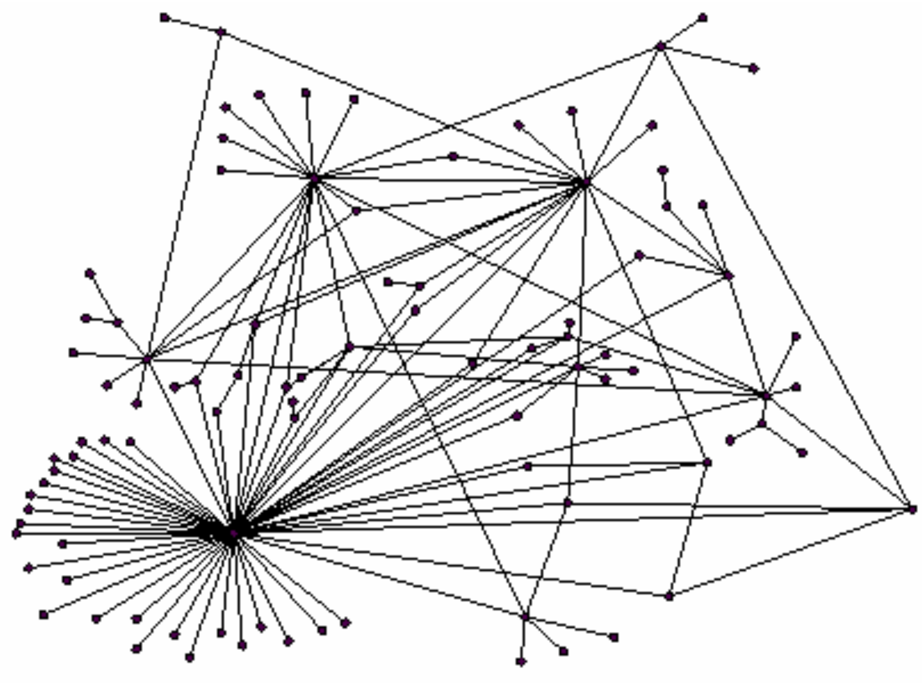


Why are Power-Laws Surprising

- **Can not arise from sums of independent events!**
  - **Recall:** in G_{np} each pair of nodes in connected independently with prob. p
    - X… degree of node v
    - X_w … event that w links to v
    - $X = \sum_w X_w$
    - $E[X] = \sum_w E[X_w] = (n-1)p$
  - **Now, what is $P(X = k)$? Central limit theorem!**
    - $X_1, \dots, X_n$: random vars with mean μ, variance σ^2
    - $S_n = \sum_i X_i$: $E[S_n] = n\mu$, $\mathrm{Var}[S_n] = n\sigma^2$, $\mathrm{SD}[S_n] = \sigma\sqrt{n}$
    - $P(S_n = E[S_n] + x \cdot \mathrm{SD}[S_n]) \sim \frac{1}{2\pi} e^{-\frac{x^2}{2}}$

Random vs. Scale-free network



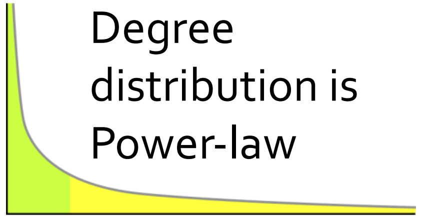


Random network

(Erdos-Renyi random graph)

Degree distribution is Binomial

Scale-free (power-law) network

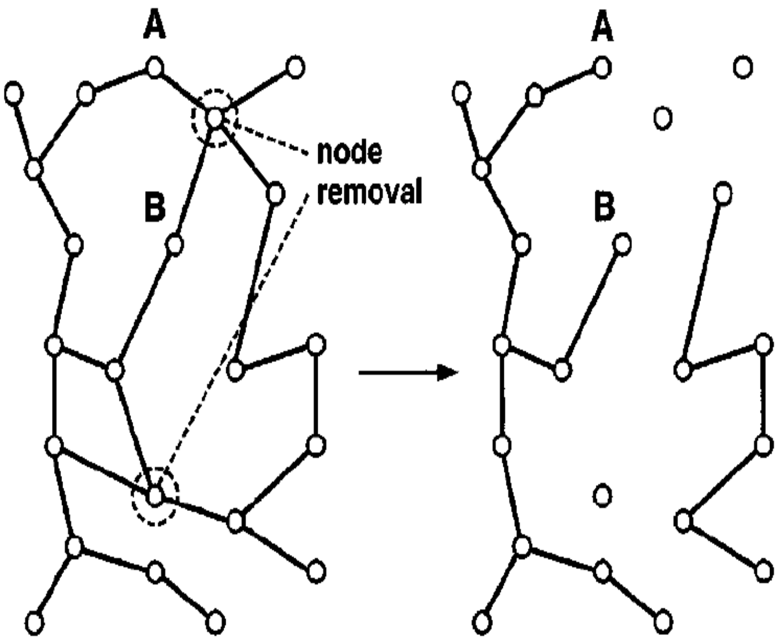


Degree distribution is Power-law

Consequence of Power-Law Degrees

Consequence: Network Resilience

- **How does network connectivity change as nodes get removed?**
  [Albert et al. 00; Palmer et al. 01]



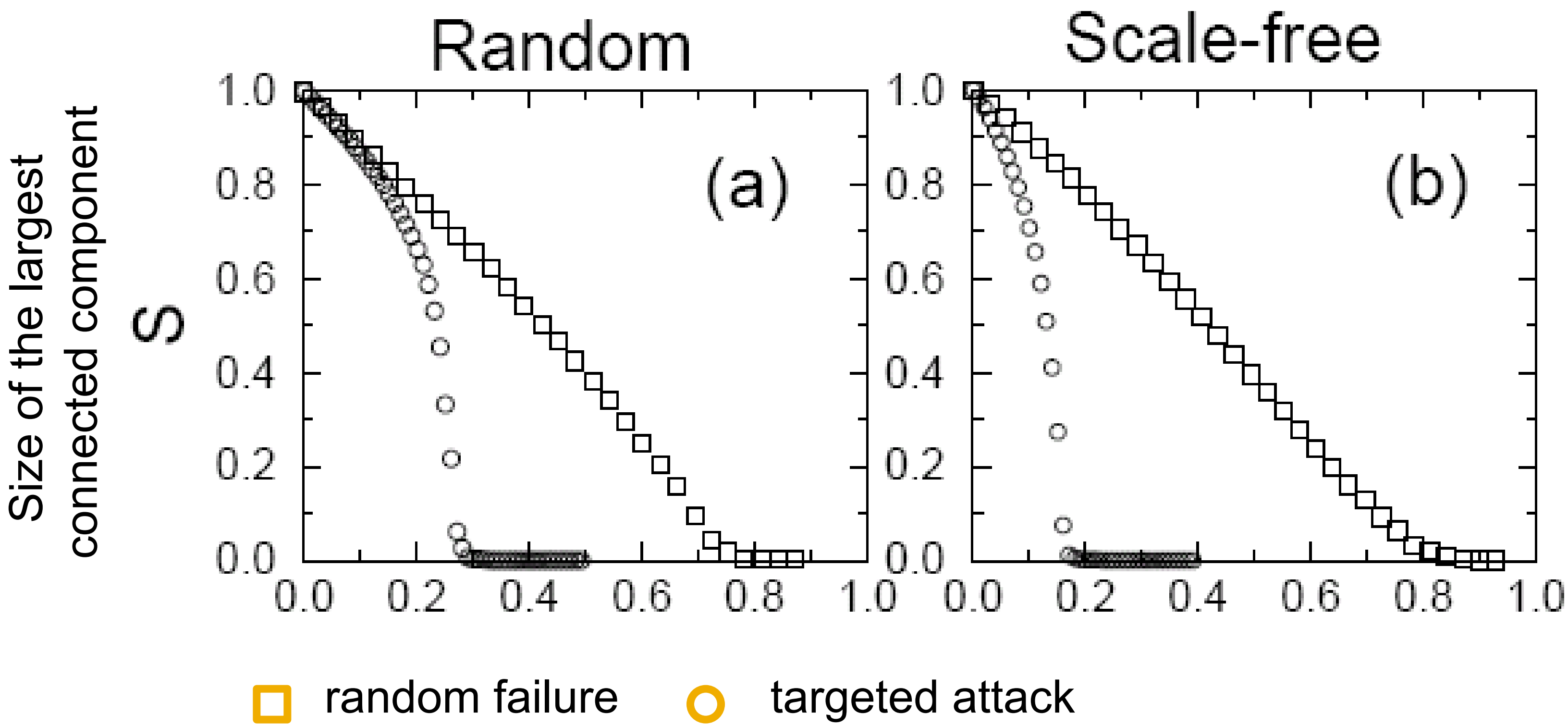
- **Nodes can be removed:**
  - **Random failure:**
    - Remove nodes uniformly at random
  - **Targeted attack:**
    - Remove nodes in order of decreasing degree
- This is important for **robustness of the internet** as well as **epidemiology**

Network Resilience

- Networks with equal number of nodes and edges:
  - ER random graph
  - Scale-free network
- Study the properties of the network as an increasing fraction of nodes are removed
  - Node selection:
    - **Random** (this corresponds to **random failures**)
    - **Nodes with largest degrees** (corresponds to **targeted attacks**)
- Measures:
  - Fraction of nodes in the **largest connected component**
  - **Average shortest path length** between nodes in the largest component

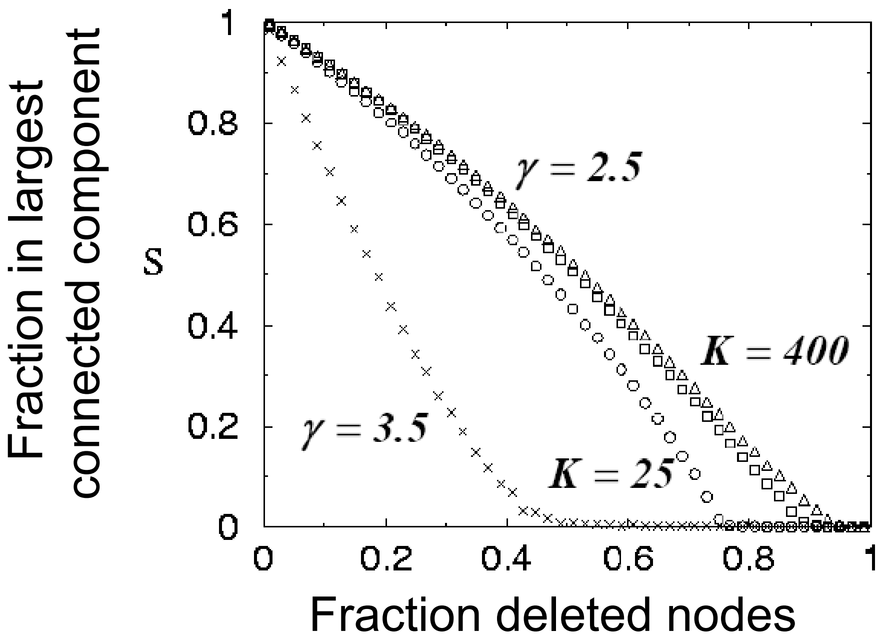
Network Resilience: Connectivity

- **Scale-free graphs are resilient to random attacks, but sensitive to targeted attacks:**



Network Resilience: Connectivity

What proportion of random nodes must be removed in order for the size (S) of the giant component to drop to 0?

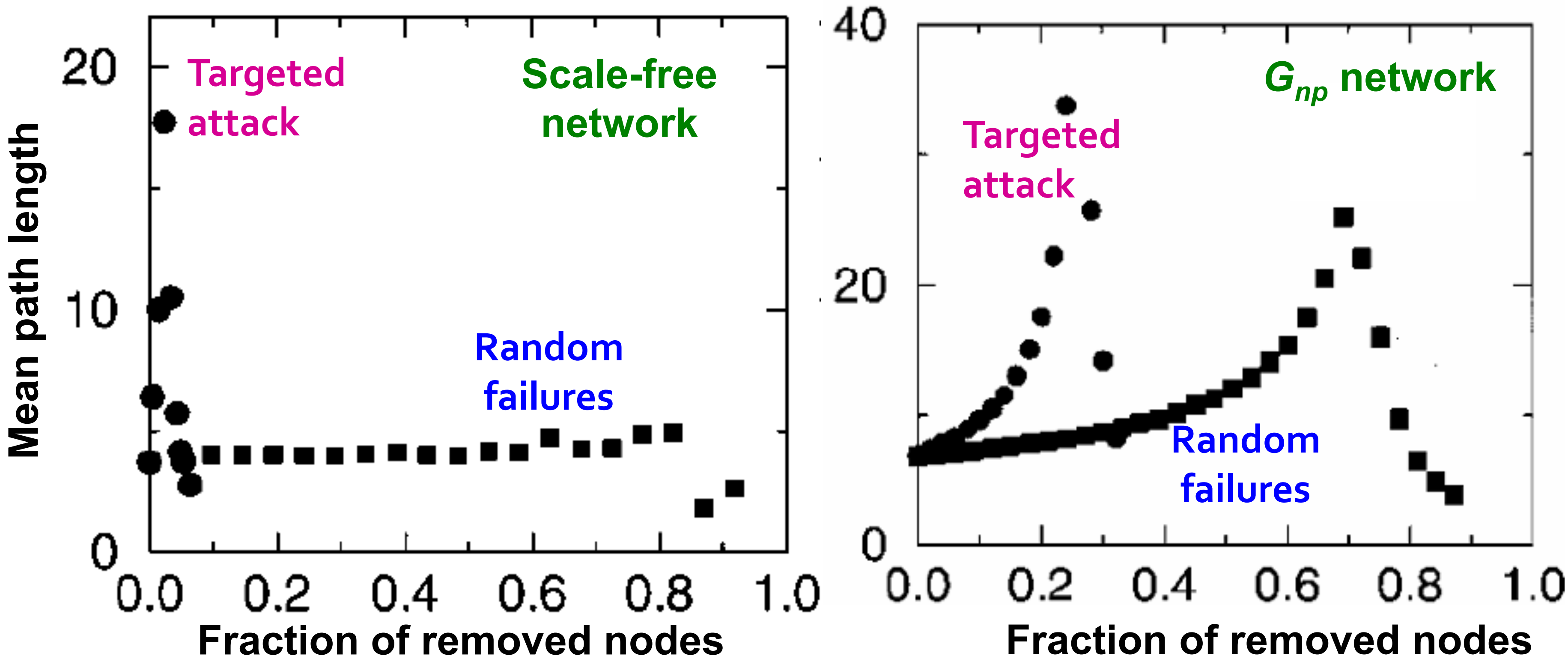


γ… degree exponent
K… maximum degree

- Infinite scale-free networks with $\gamma < 3$ never break down under random node failures

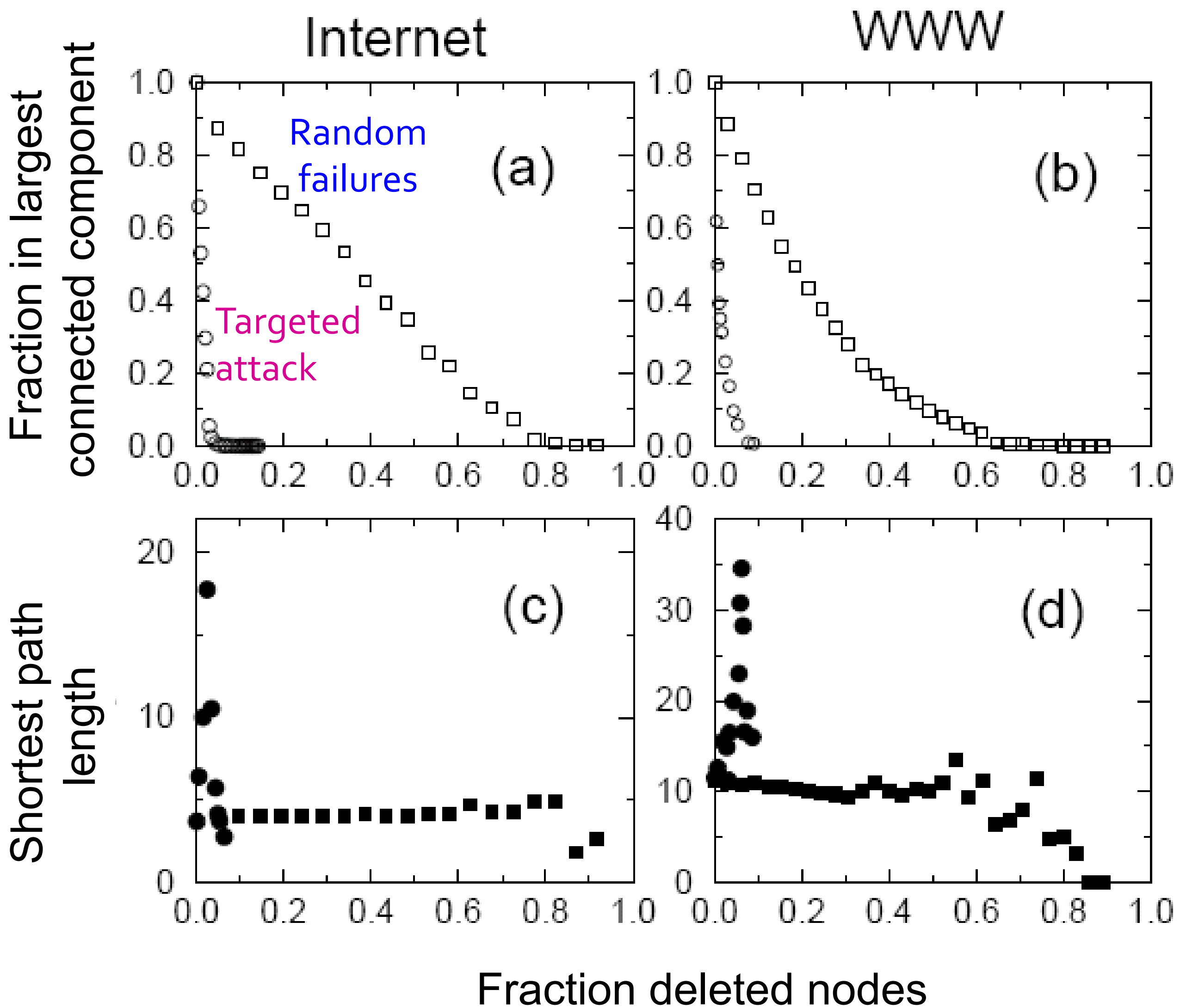
Source: Cohen et al., Resilience of the Internet to Random Breakdowns

Network Resilience: Path Length



- Real networks are resilient to <u>random failures</u>
- G_{np} has better resilience to <u>targeted attacks</u>
  - E.g., we need to remove all pages of degree **>5** to disconnect the Web. But this is a very small fraction of all web pages!

Resilience in Real Networks

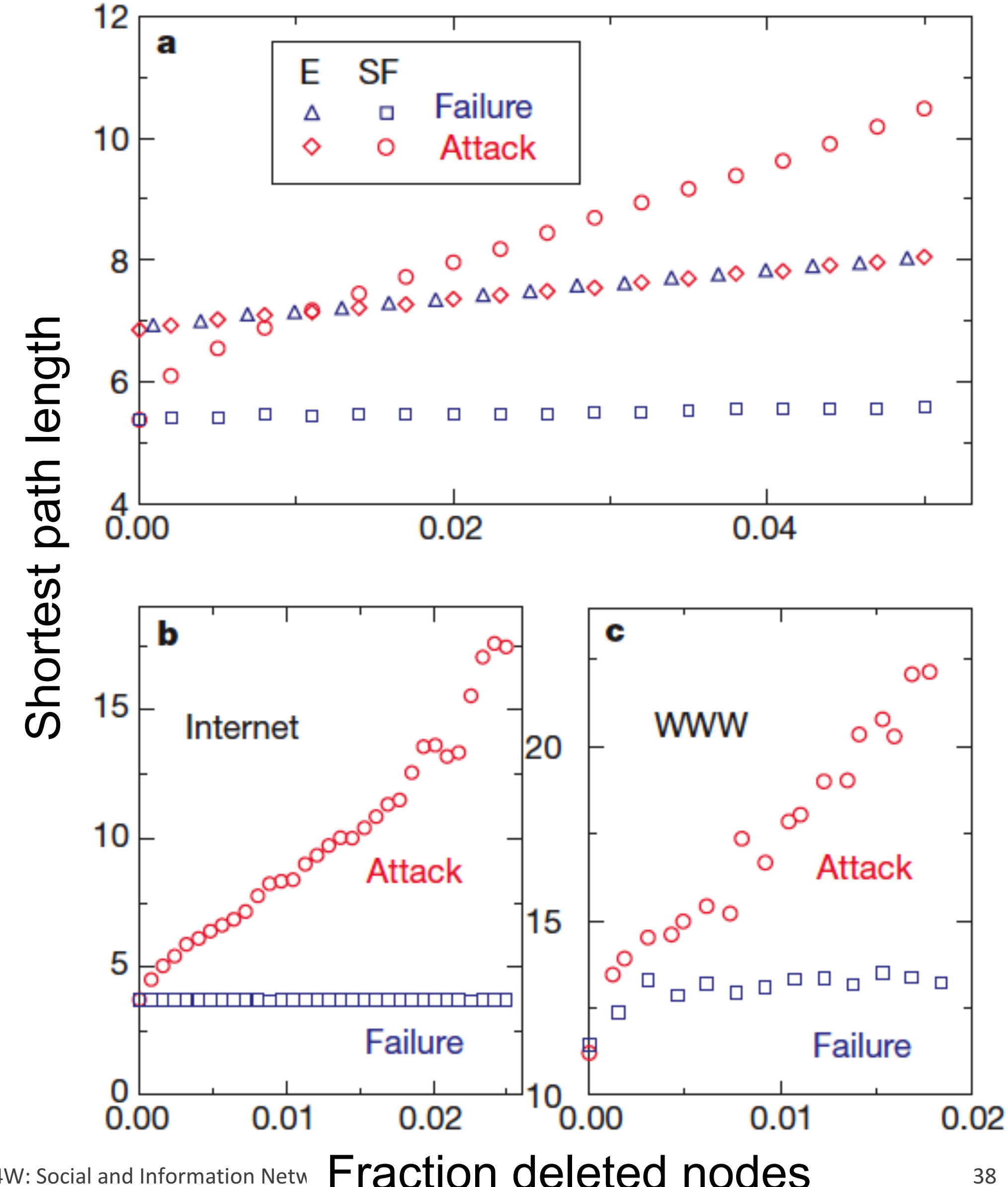


Source: Error and attack tolerance of complex networks. Réka Albert, Hawoong Jeong and Albert-László Barabási

Zoom-in

- **The first few % of nodes removed:**
  - **E: G_{np}**
  - **SF: Scale-free**

- Notice how targeted attacks very quickly disconnect the network

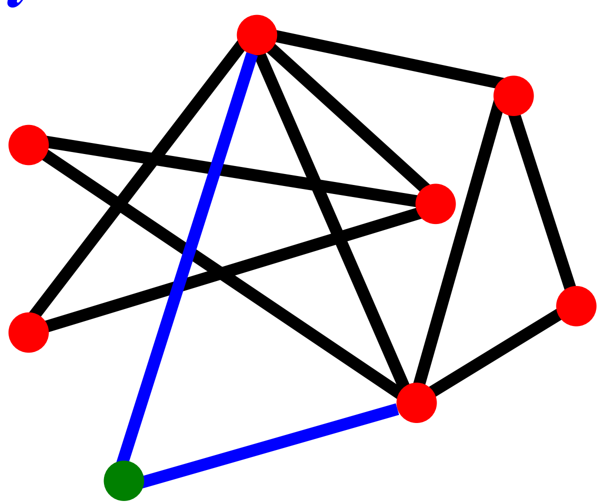


Preferential Attachment Model

Model: Preferential attachment

- **Preferential attachment**
  [Price '65, Albert-Barabasi '99, Mitzenmacher '03]
  - Nodes arrive in order **1,2,…,n**
  - At step j, let d_i be the degree of node $i < j$
  - A new node j arrives and creates m out-links
  - Prob. of j linking to a previous node i is **proportional to degree d_i of node i**

$$P(j \to i) = \frac{d_i}{\sum_k d_k}$$

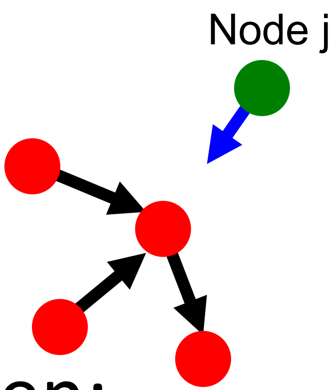


Rich Get Richer

- **New nodes are more likely to link to nodes that already have high degree**
- **Herbert Simon's result:**
  - Power-laws arise from "**Rich get richer**" (**cumulative advantage**)
- **Examples**
  - **Citations** [de Solla Price '65]: New citations to a paper are proportional to the number it already has
    - **Herding:** If a lot of people cite a paper, then it must be good, and therefore I should cite it too
  - **Sociology:** Matthew effect, http://en.wikipedia.org/wiki/Matthew_effect
    - "For whoever has will be given more, and they will have an abundance. Whoever does not have, even what they have will be taken from them."
    - Eminent scientists often get more credit than a comparatively unknown researcher, even if their work is similar

The Exact Model

[Mitzenmacher, '03]



We will analyze the following model:

- Nodes arrive in order $1,2,3, \ldots, n$
- When node j is created it makes a **single out-link** to an earlier node i chosen:
  - **1)** With prob. p, j links to i chosen **uniformly at random** (from among all earlier nodes)
  - **2)** With prob. $1-p$, node j chooses i uniformly at random & links **to a random node *l that i* points to**
    - **This is same as saying:** With prob. $1-p$, node j links to node l with prob. proportional to d_l (the in-degree of l)
  - **Our graph is directed:** Every node has out-degree 1

The Model Givens Power-Laws

- **Claim:** The described model generates networks where the fraction of nodes with in-degree $\boldsymbol{k}$ scales as:

$$P(d_i = k) \propto k^{-(1+\frac{1}{q})}$$

where q=1-p

So we get power-law degree distribution with exponent:

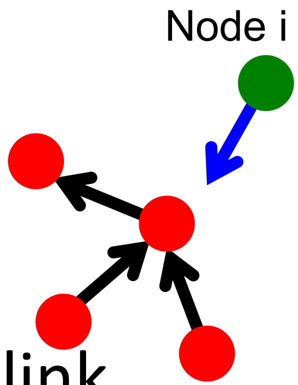
$$\alpha = 1 + \frac{1}{1-p}$$

Continuous Approximation

- **Consider deterministic and continuous approximation** to the degree of node $\boldsymbol{i}$ as a function of time $\boldsymbol{t}$
  - $\boldsymbol{t}$ is the number of nodes that have arrived so far
  - **In-Degree $d_i(t)$** of node $\boldsymbol{i}$ ($i = 1,2,\ldots,n$) is a **continuous quantity** and it **grows deterministically** as a function of time $\boldsymbol{t}$

- **Plan: Analyze $d_i(t)$ – continuous in-degree** of node $\boldsymbol{i}$ at time $\boldsymbol{t}$ $(\boldsymbol{t} > \boldsymbol{i})$
  - **Note: Node $\boldsymbol{i}$ arrives to the graph at time $\boldsymbol{i}$**

Continuous Degree: What We Know



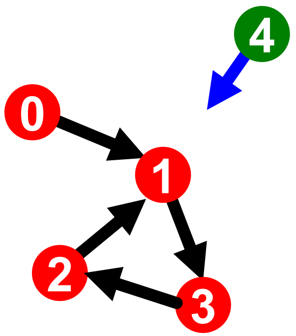
- **Initial condition:**
  - $d_i(t) = 0$, when $t = i$ **(node *i* just arrived)**
- **Expected <u>change</u> of $d_i(t)$ over time:**
  - Node i gains an in-link at step $t + 1$ only if a link from a newly created node $t + 1$ points to it
  - **What's the probability of this event?**
    - With prob. p node $t + 1$ links **randomly**:
      - Links to our node i with prob. $1/t$
    - With prob. $1 - p$ node $t + 1$ links **preferentially**:
      - Links to our node i with prob. $d_i(t)/t$
  - **Prob. node $t + 1$ links to i is:** $p\frac{1}{t} + (1-p)\frac{d_i(t)}{t}$

Note: each node creates exactly 1 edge. So after *t* nodes/steps there are *t* edges in total.

Continuous Degree

- **At $t = 4$ node $i = 4$ comes. It has out-degree of 1 to deterministically share with other nodes:**

Node i	$d_i(t)$	$d_i(t+1)$
0	0	$=0 + p\frac{1}{4} + (1-p)\frac{0}{4}$
1	2	$=2 + p\frac{1}{4} + (1-p)\frac{2}{4}$
2	0	$=0 + p\frac{1}{4} + (1-p)\frac{1}{4}$
3	1	$=1 + p\frac{1}{4} + (1-p)\frac{1}{4}$
4	/	0



- $d_i(t) - d_i(t-1) = \frac{\mathrm{d}d_i(t)}{\mathrm{d}t} = \mathbf{p}\frac{1}{t} + (1-p)\frac{d_i(t)}{t}$
- **How does $d_i(t)$ evolve as $t \to \infty$?**

What is the rate of growth of d_i?

- **Expected change of $d_i(t)$:**
  - $\underbrace{d_i(t+1) - d_i(t)} = p\frac{1}{t} + (1-p)\frac{d_i(t)}{t}$
  - $\frac{\mathrm{d}d_i(t)}{\mathrm{d}t} = p\frac{1}{t} + (1-p)\frac{d_i(t)}{t} = \frac{p+qd_i(t)}{t}$ — $q = (1-p)$
  - $\frac{1}{p+qd_i(t)}\mathrm{d}d_i(t) = \frac{1}{t}\mathrm{d}t$ — Divide by $p + q\, d_i(t)$
  - $\int \frac{1}{p+qd_i(t)}\mathrm{d}d_i(t) = \int \frac{1}{t}\mathrm{d}t$ — integrate
  - $\frac{1}{q}\ln\left(p + qd_i(t)\right) = \ln t + c$ — Exponentiate and let $A = e^c$
  - $p + qd_i(t) = e^{qc}\, t^q \;\Rightarrow\; d_i(t) = \frac{1}{q}\left((At)^q - p\right)$ **A=?**

What is the constant A?

$$\boldsymbol{d_i(t) = \frac{1}{q}(At^q - p)}$$

What is the value of constant A?

- **We know:** $d_i(i) = 0$
- **So:** $d_i(i) = \frac{1}{q}((Ai)^q - p) = 0$
- $\Rightarrow \boldsymbol{A = \frac{p}{i^q}}$
- **And so** $\Rightarrow \boldsymbol{d_i(t) = \frac{p}{q}\left(\left(\frac{t}{i}\right)^q - 1\right)}$

Observation: Old nodes (small i values) have higher in-degrees $d_i(t)$