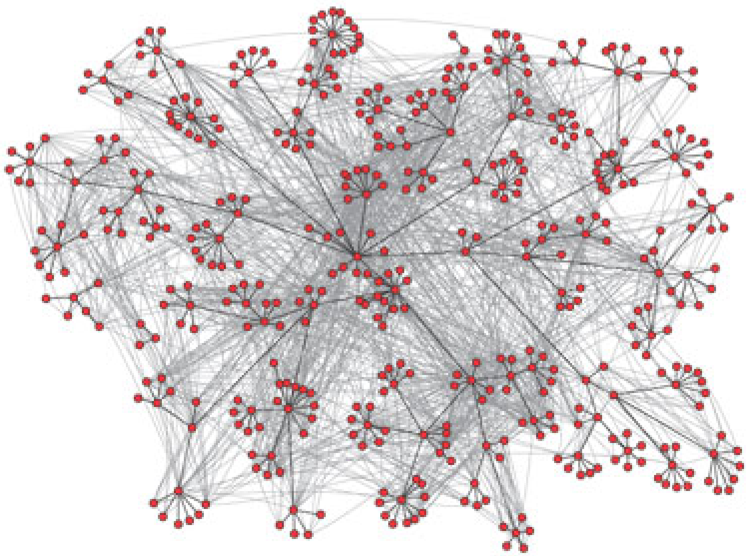
Network Formation Processes: Power-law degree distributions and Preferential Attachment

CS224W: Social and Information Network Analysis
Jure Leskovec, Stanford University
http://cs224w.stanford.edu

Network Formation Processes

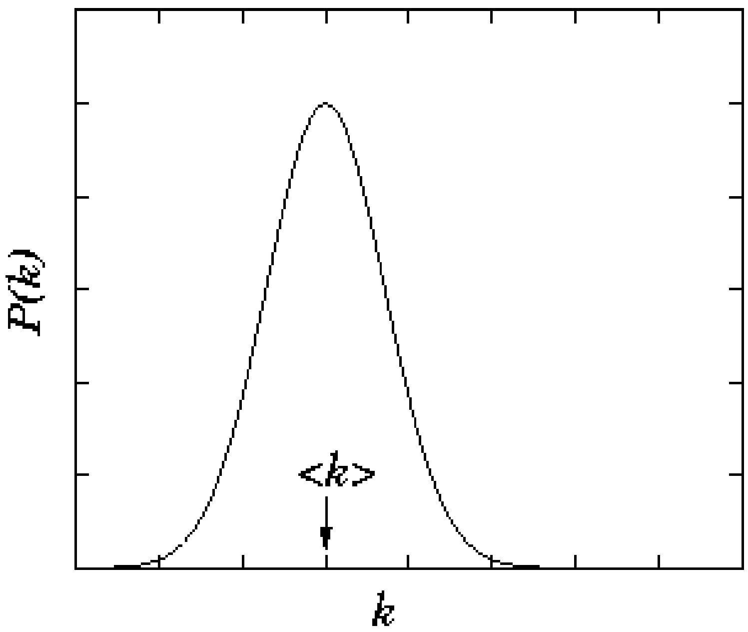
What do we observe that needs explaining

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

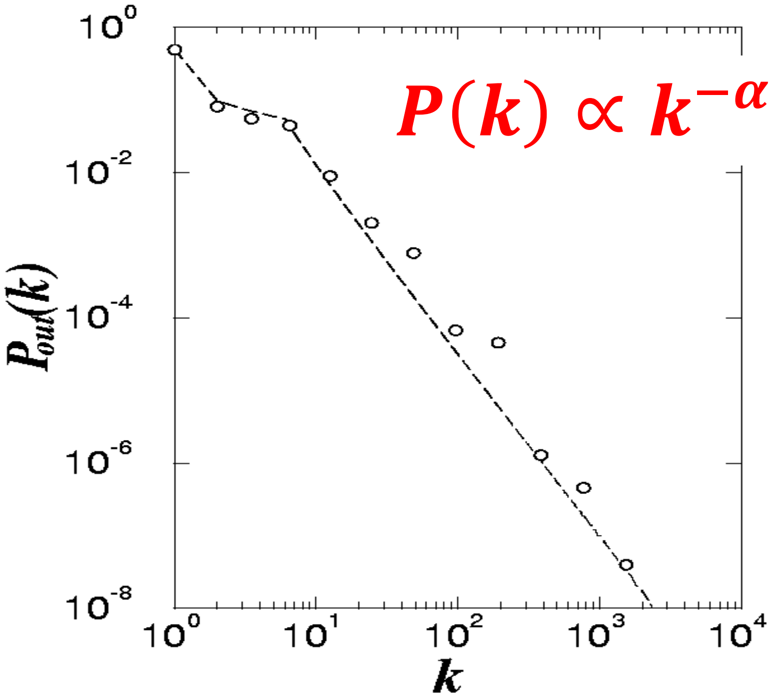


Degree Distributions

Expected based on G_{np}



Found in data

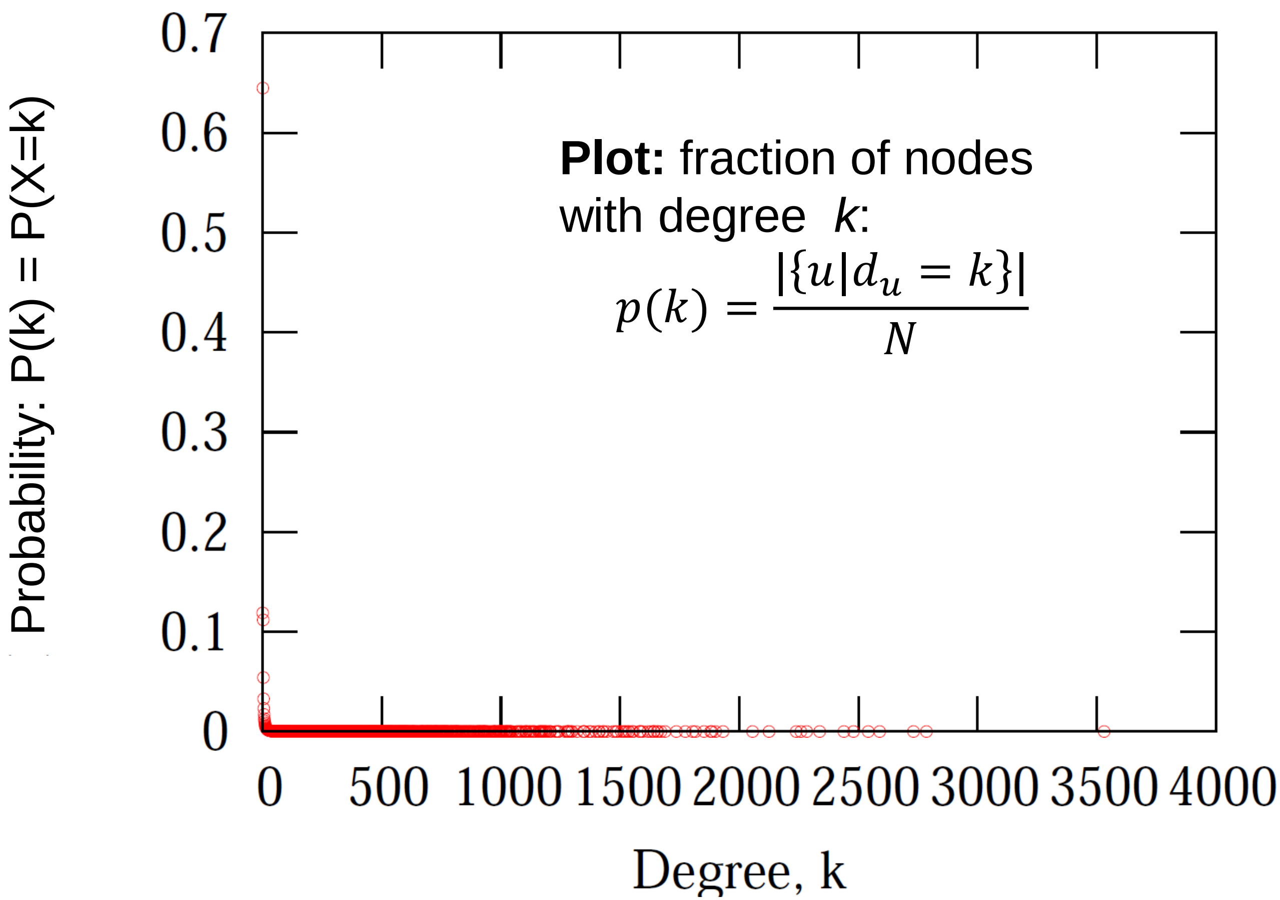


$$P(k) \propto k^{-\alpha}$$

[Leskovec et al. KDD '08]

Node Degrees in Networks

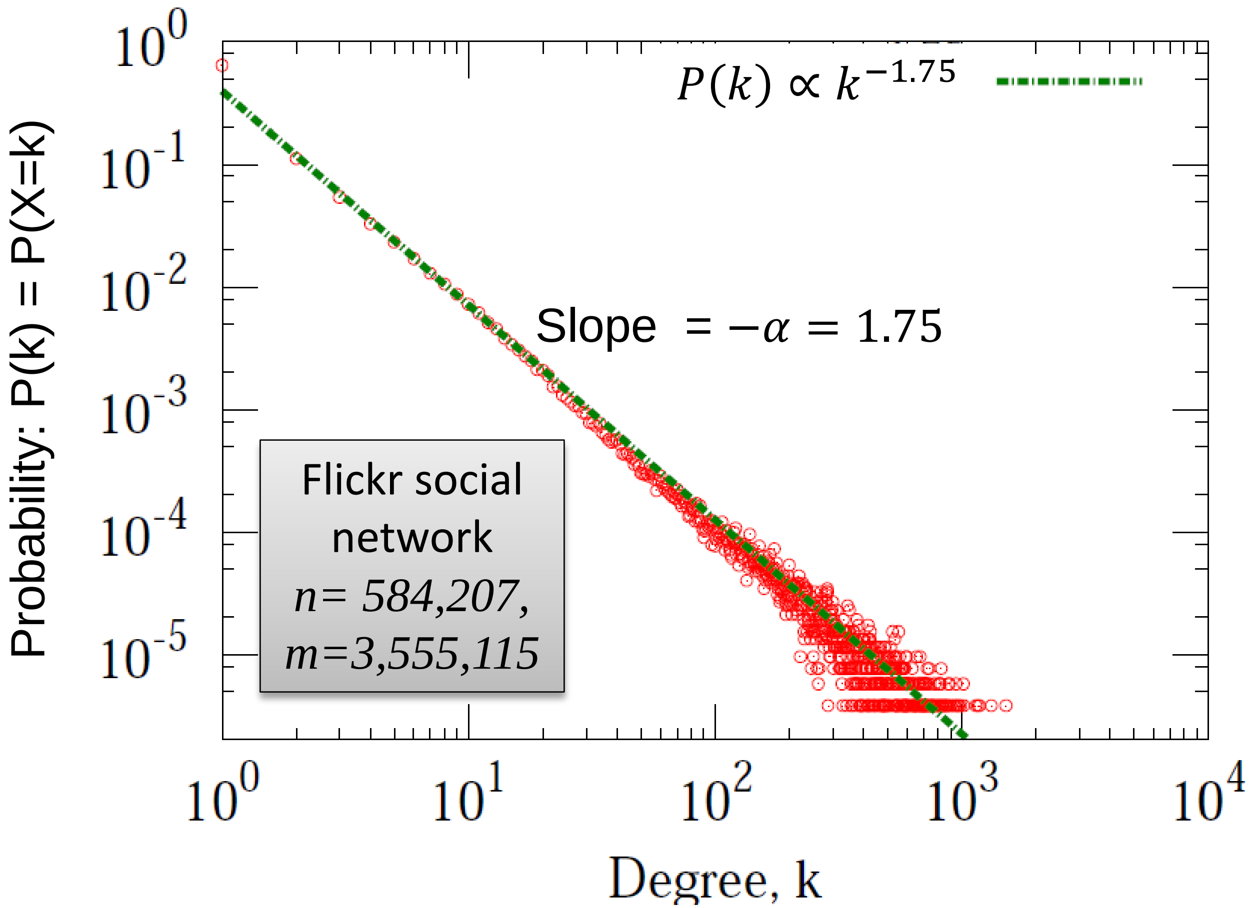
- **Take a network, plot a histogram of *P(k)* vs. *k***



Flickr social network
n= *584,207*,
m=3,555,115

Node Degrees in Networks

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

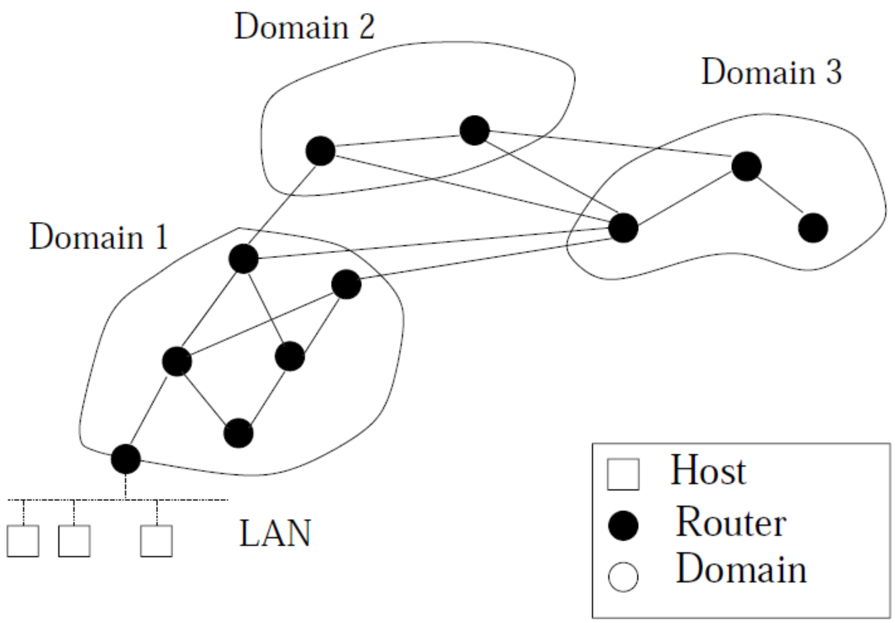


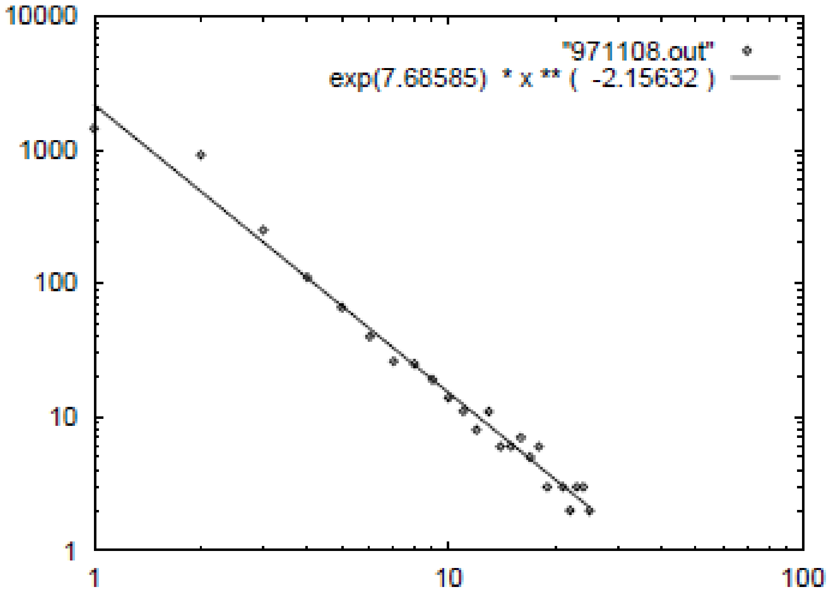
How to distinguish:
$P(k) \propto \exp(k)$ **vs.**
$P(k) \propto k^{-\alpha}$ **?**

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: Faloutsos³

- [Faloutsos, Faloutsos and Faloutsos, 1999]

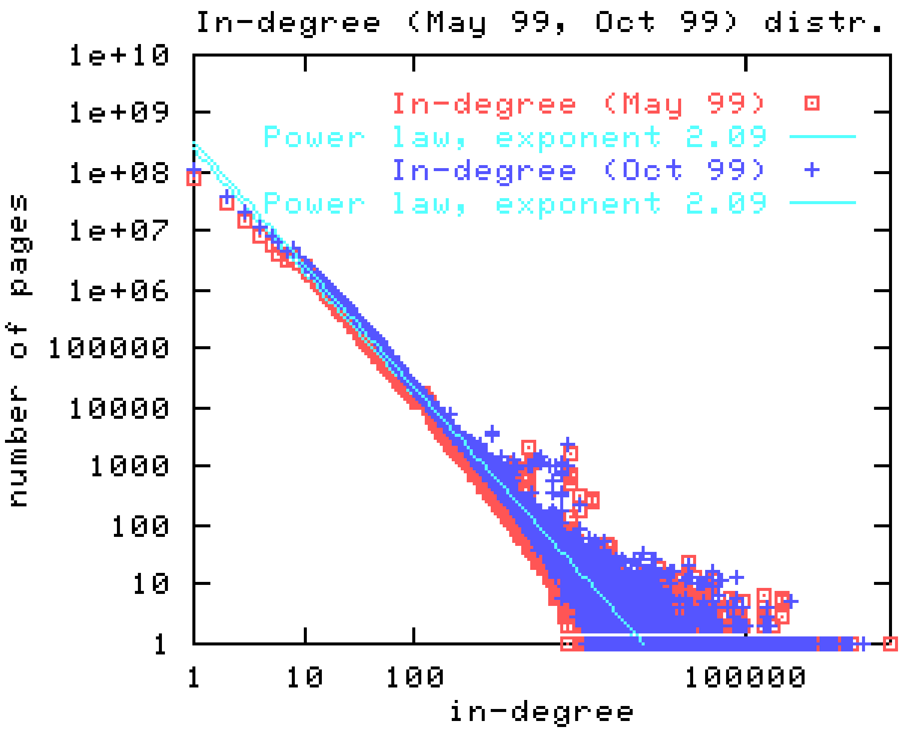


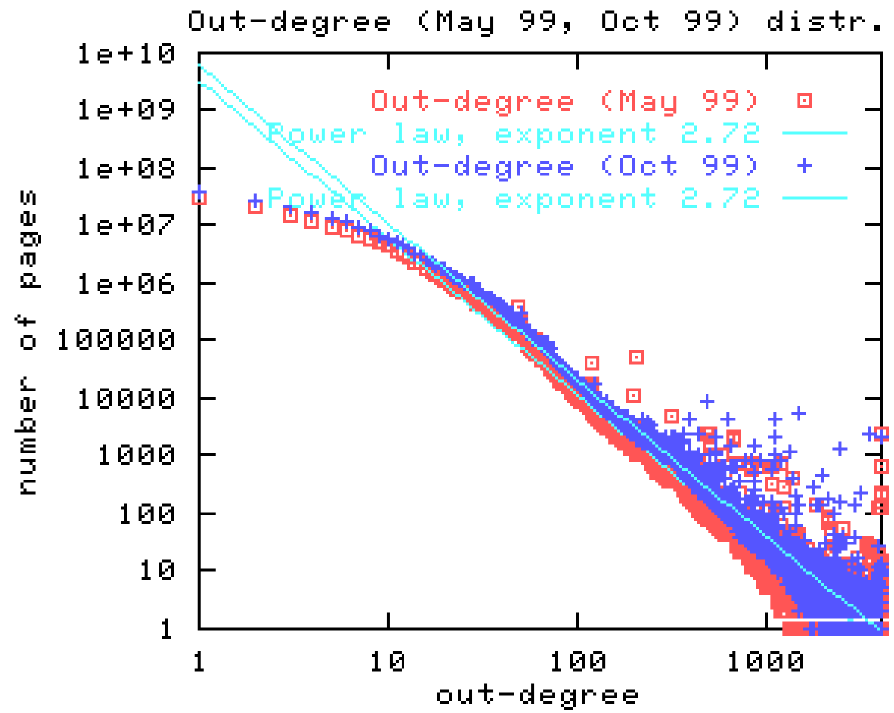


Internet domain topology

Node Degrees: Web

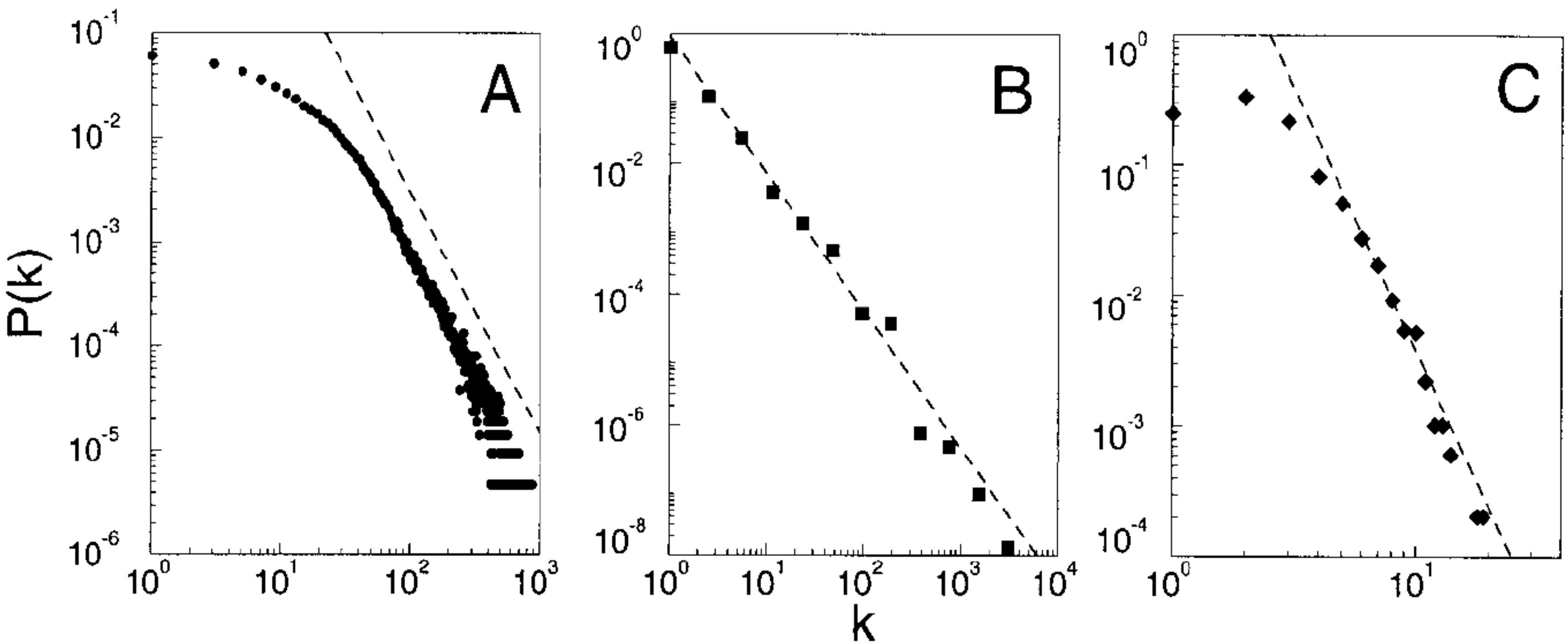
- [Broder et al., 2000]





Node Degrees: Barabasi&Albert

- [Barabasi-Albert, 1999]

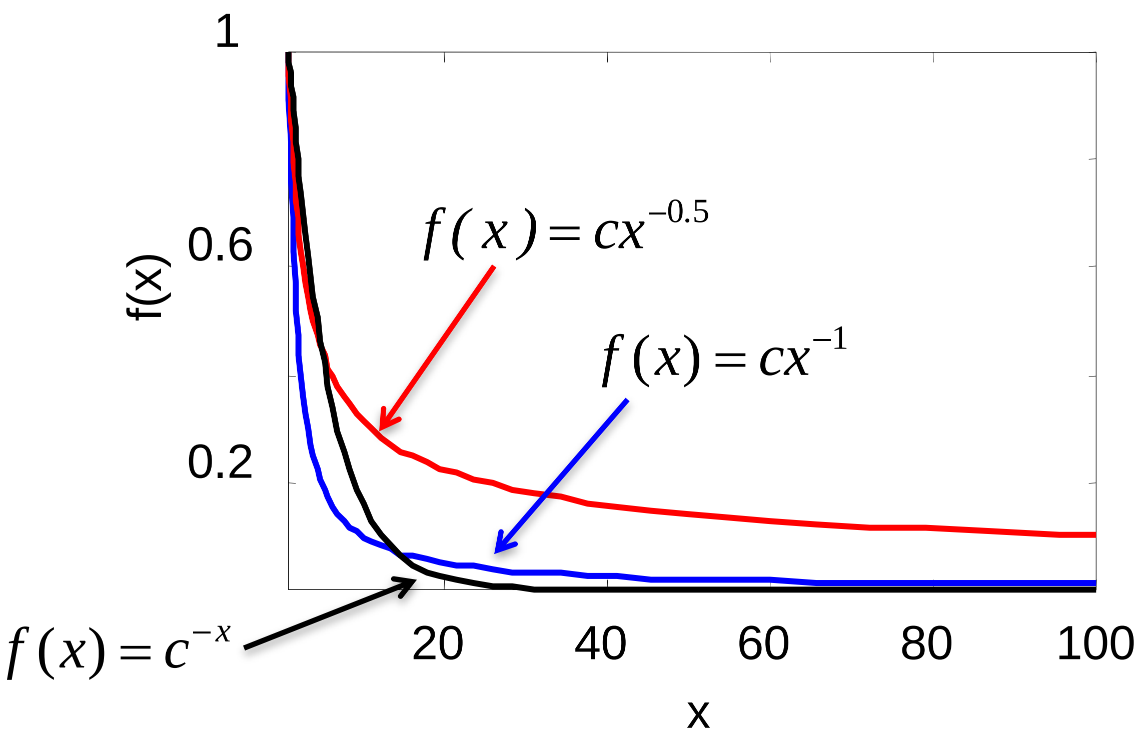


Actor collaborations

Web graph

Power-grid

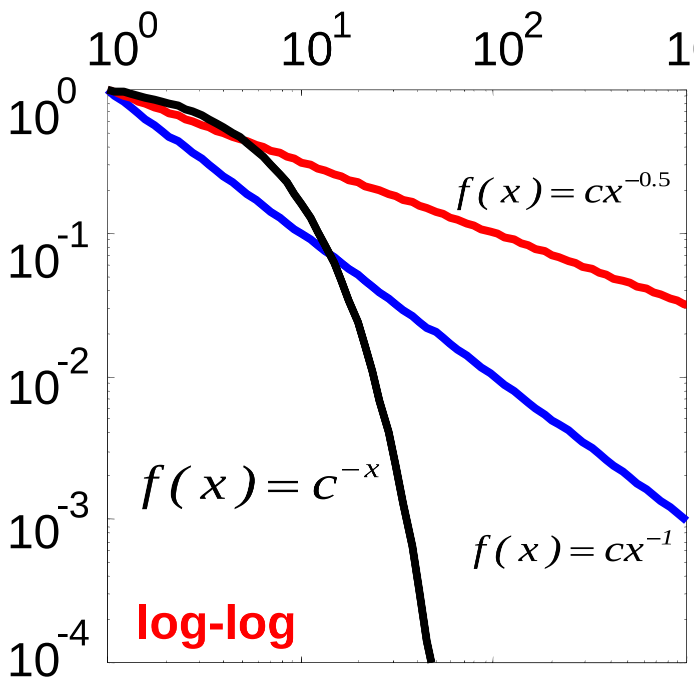
Exponential vs. Power-Law

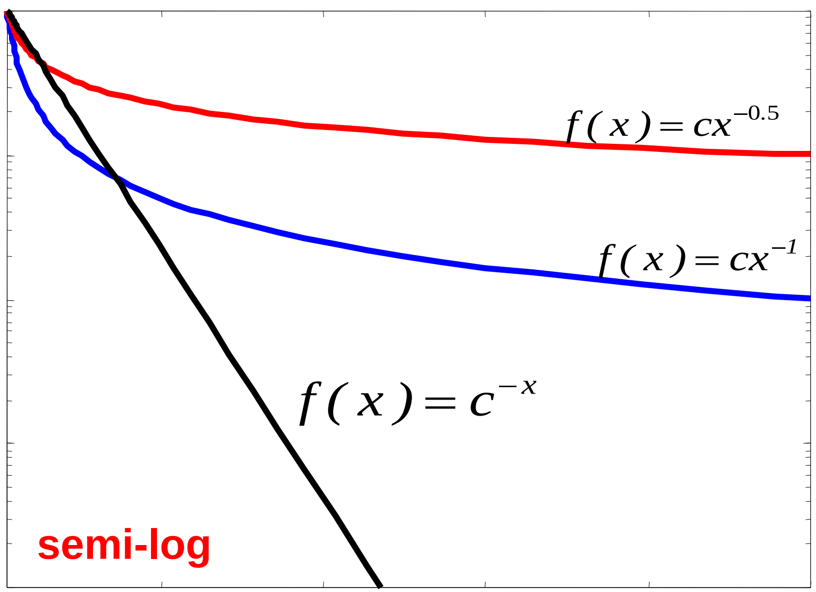


- Above a certain *x* value, the power law is always higher than the exponential.

Exponential vs. Power-Law

- Power-law vs. exponential on log-log and log-lin scales

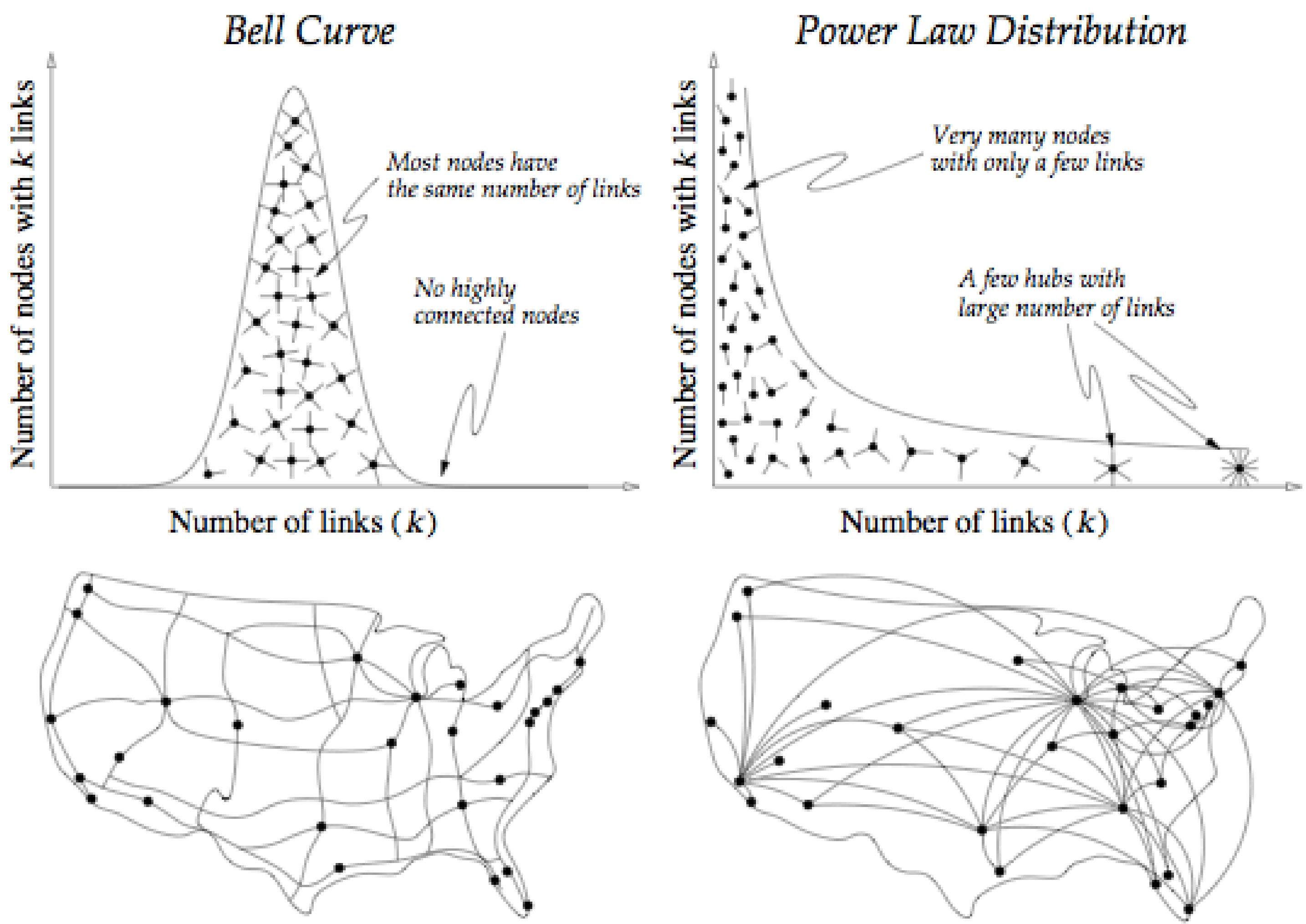




x … logarithmic
y … logarithmic

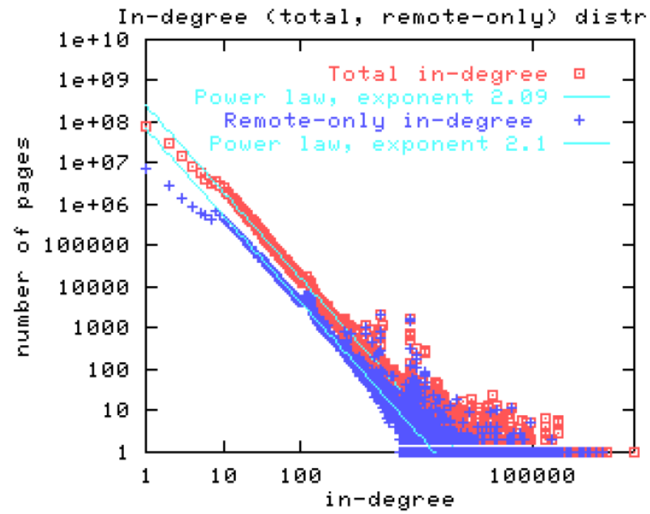
x … linear
y … logarithmic

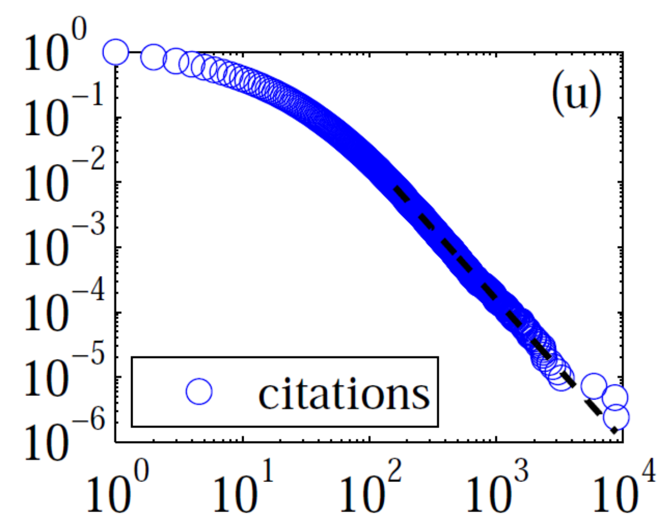
Exponential vs. Power-Law



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"

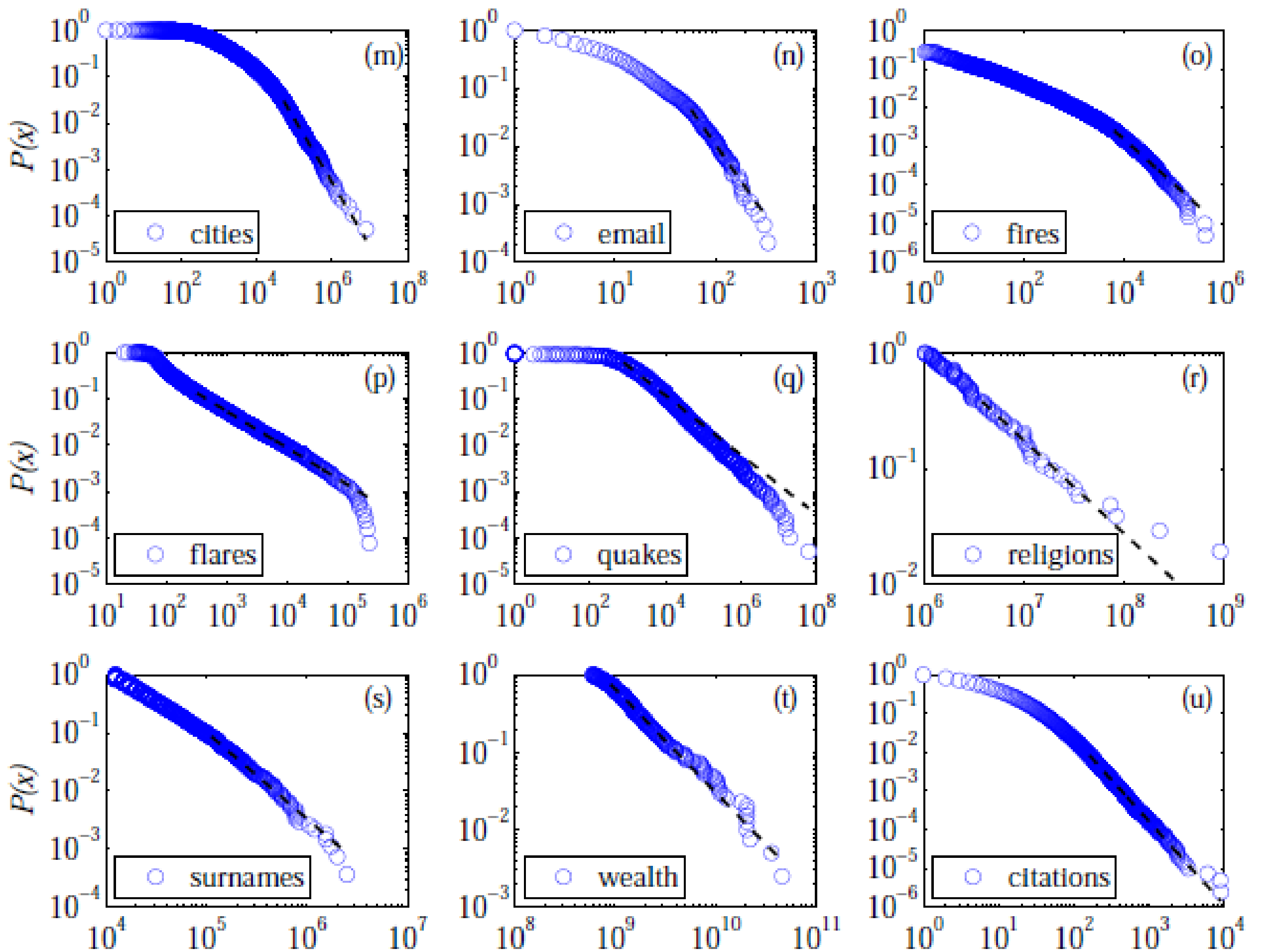
- Where does the name come from?
  - **Scale invariance**: there is no characteristic scale
  - **Scale-free function:** $\boldsymbol{f(ax) = a^{\lambda} f(x)}$
    - Power-law function: $f(ax) = a^{\lambda} x^{\lambda} = a^{\lambda} f(x)$

Power-laws are Everywhere

- **In social systems – lots of power-laws:**
  - Pareto, 1897 – Wealth distribution
  - Lotka 1926 – Scientific output
  - Yule 1920s – Biological taxa and subtaxa
  - Zipf 1940s – Word frequency
  - Simon 1950s – City populations

[Clauset-Shalizi-Newman 2007]

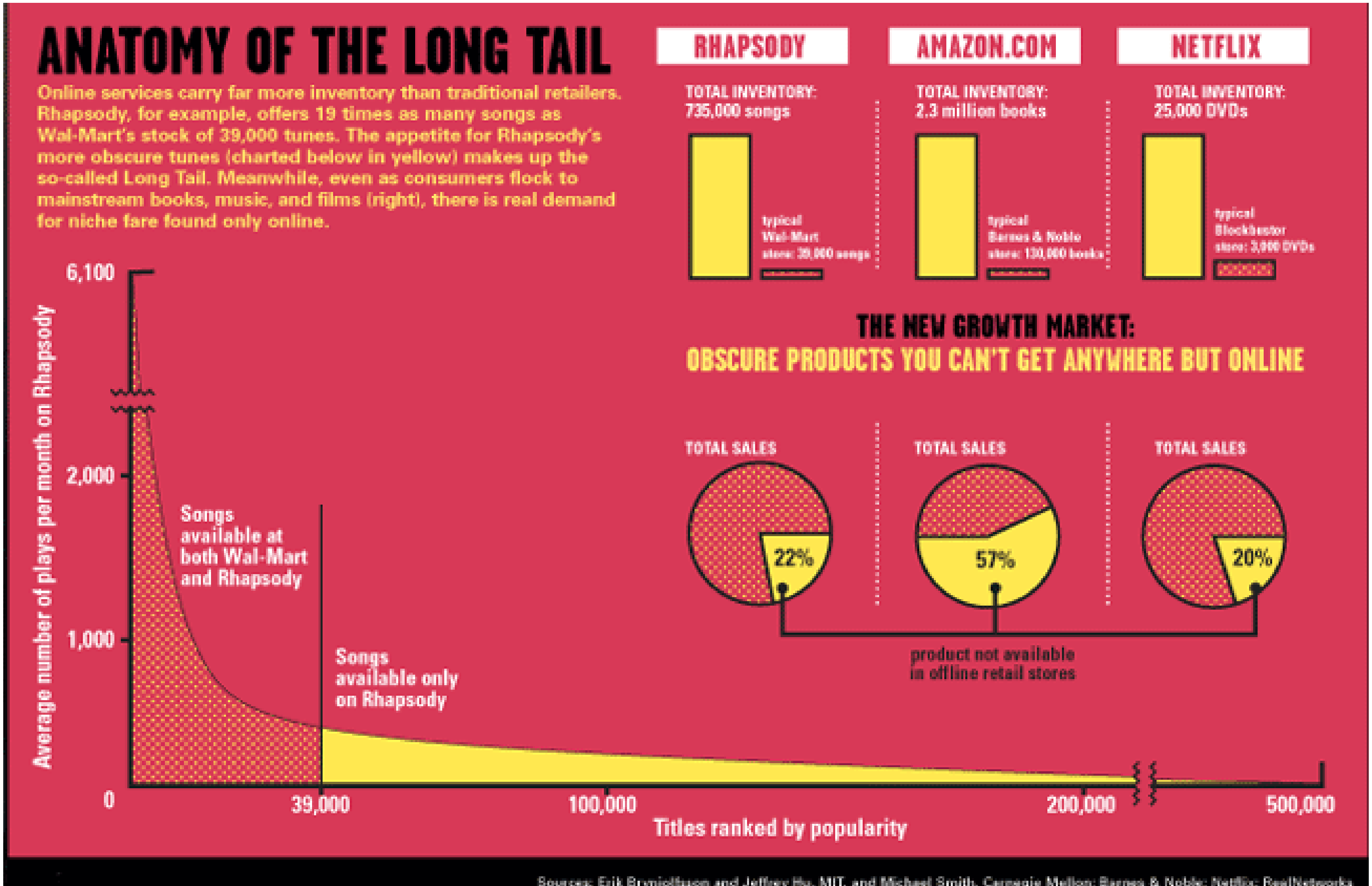
Power-laws are Everywhere



Many other quantities follow heavy-tailed distributions

[Chris Anderson, Wired, 2004]

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

are not heavy tailed!

[Clauset-Shalizi-Newman 2007]

Heavy Tails

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

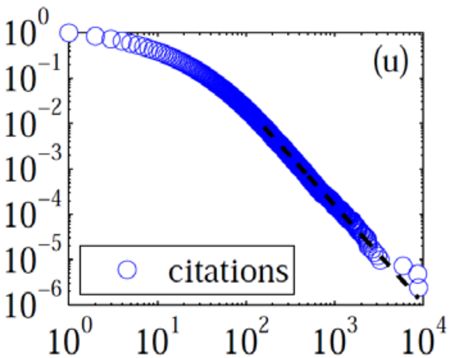
power law	$x^{-\alpha}$
power law with cutoff	$x^{-\alpha}\mathrm{e}^{-\lambda x}$
stretched exponential	$x^{\beta-1}\mathrm{e}^{-\lambda x^{\beta}}$
log-normal	$\frac{1}{x}\exp\left[-\frac{(\ln x-\mu)^2}{2\sigma^2}\right]$

Mathematics of Power-laws

- **What is the normalizing constant?**

  $P(x) = z\,x^{-\alpha}$ $\quad\quad z=?$

  - $P(x)$ is a distribution: $\int P(x)dx = 1$



Continuous approximation

P(x) diverges as x→0 so x_m is the minimum value of the power-law distribution $x \in [x_m, \infty]$

- $1 = \int_{x_{min}}^{\infty} P(x)dx = z \int_{x_m}^{\infty} x^{-\alpha} dx$
- $= \frac{z}{\alpha-1}[x^{-\alpha+1}]_{x_m}^{\infty}$
- $\Rightarrow z = (\alpha - 1)x_m^{\alpha-1}$

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

[Clauset-Shalizi-Newman 2007]

Mathematics of Power-laws

- **What's the expectation of a power-law random variable *x*?**
- $E[x] = \int_{x_m}^{\infty} x\, P(x) dx = z \int_{x_m}^{\infty} x^{-\alpha+1} dx$
- $= \frac{z}{2-\alpha}\left[x^{2-\alpha}\right]_{x_m}^{\infty} = \frac{(\alpha-1)x_m^{\alpha-1}}{2-\alpha}\left[\infty^{2-\alpha} - x_m^{2-\alpha}\right]$

Need: $\alpha > 2$

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

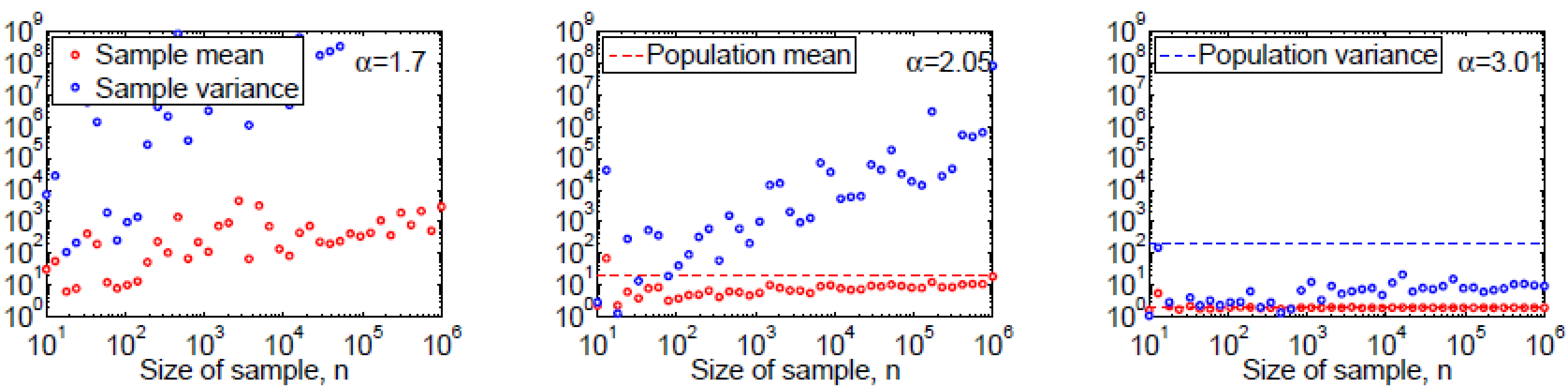
Mathematics of Power-Laws

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

- **Power-laws: Infinite moments!**
  - 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]=∞

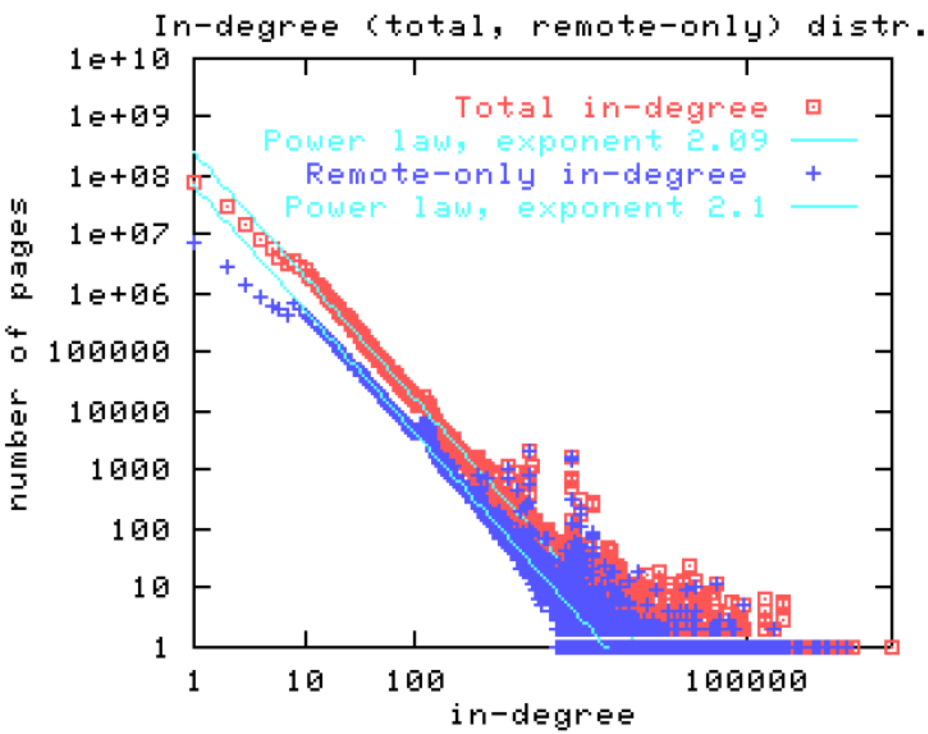
- **Sample average of *n* samples from a power-law with exponent α:**



Estimating Power-Law Exponent α

- **Estimating α from data:**
  1. Fit a line on log-log axis using least squares method: $\min_{\alpha}(\log(y) - \alpha\log(x))^2$

BAD!



Estimating Power-Law Exponent α

- **Estimating α from data:**
  2. Plot Complementary CDF $P(X > x)$. Then $\alpha = 1 + \alpha'$ where α is the slope of $P(X > x)$.

OK

If $\mathbf{P(X = x) \propto x^{-\alpha}}$ then $\mathbf{P(X = x) \propto x^{-(\alpha-1)}}$

- $P(X > x) = \sum_{j=x}^{\infty} P(j) \approx = \int_x^{\infty} z\, j^{-\alpha} dj =$
- $= \frac{z}{\alpha} \left[j^{1-\alpha} \right]_x^{\infty} = \frac{z}{\alpha} x^{-(\alpha-1)}$

Estimating Power-Law Exponent α

- **Estimating α from data:**

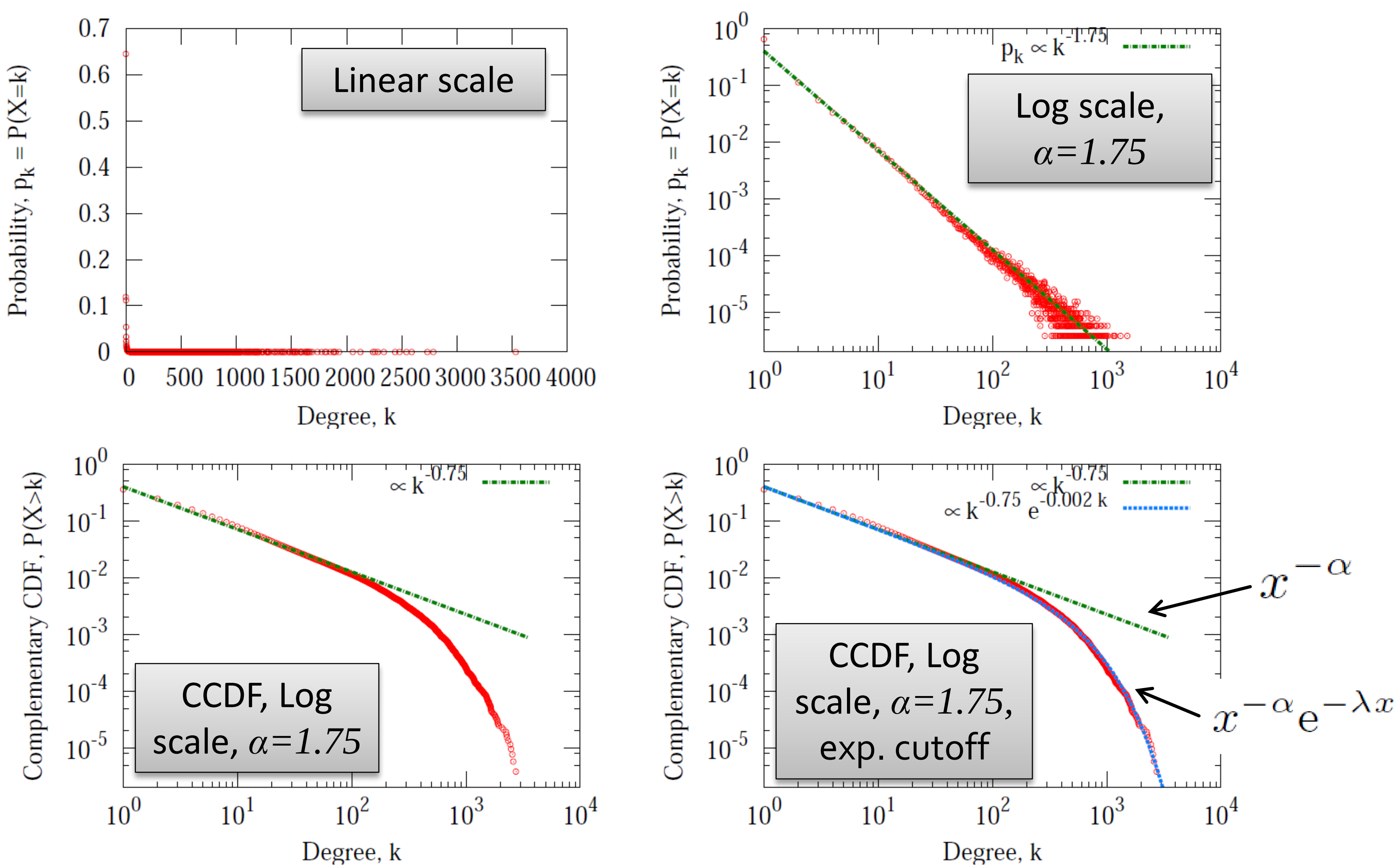
OK

3. Use MLE: $\hat{\alpha} = 1 + n\left[\sum_i^n \ln\left(\frac{d_i}{x_m}\right)\right]^{-1}$
   - $L(\alpha) = \ln(\prod_i^n p(d_i)) = \sum_i^n \ln p(d_i)$
   - $= \sum_i^n \ln(\alpha - 1) - \ln(x_m) - \alpha \ln\left(\frac{d_i}{x_m}\right)$
   - **Want to find α that max: set** $\frac{\mathrm{dL}(\alpha)}{\mathrm{d}\alpha} = 0$
   - $\frac{\mathrm{dL}(\alpha)}{\mathrm{d}\alpha} = 0 \Rightarrow \frac{n}{\alpha - 1} - \sum \ln\left(\frac{d_i}{x_m}\right) = 0$
   - $\Rightarrow \hat{\alpha} = 1 + n\left[\sum_i^n \ln\left(\frac{d_i}{x_m}\right)\right]^{-1}$

Power-law density:

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

Flickr: Fitting Degree Exponent



Maximum Degree

- **What is the expected maximum degree K in a scale-free network?**
  - The expected number of nodes with degree > K should be less than 1: $\int_K^\infty P(x)dx \approx \frac{1}{n}$
  - $= z\int_K^\infty x^{-\alpha}dx = \frac{z}{1-\alpha}[x^{1-\alpha}]_K^\infty =$
  - $= \frac{(\alpha-1)x_m^{\alpha-1}}{-\alpha+1}[0 - K^{1-\alpha}] = \frac{x_m^{\alpha-1}}{K^{\alpha-1}} \approx \frac{1}{n}$
  - $\Rightarrow K = x_m N^{\frac{1}{\alpha-1}}$

Power-law density:

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

Maximum Degree: Consequence

- **Why don't we see networks with exponents in the range of $\alpha = 4, 5, 6$?**
  - In order to reliably estimate α, we need 2-3 orders of magnitude of K. That is, $K \approx 10^3$
  - E.g., to measure an degree exponent $\alpha = 5$,we need to maximum degree of the order of:

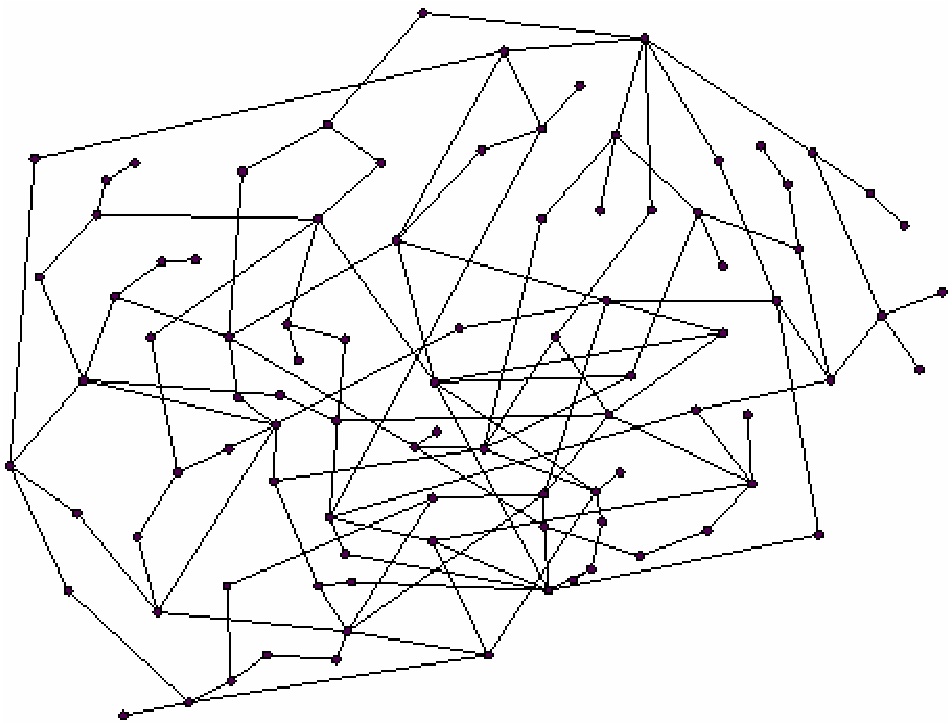
$$K = x_m N^{\frac{1}{\alpha - 1}}$$

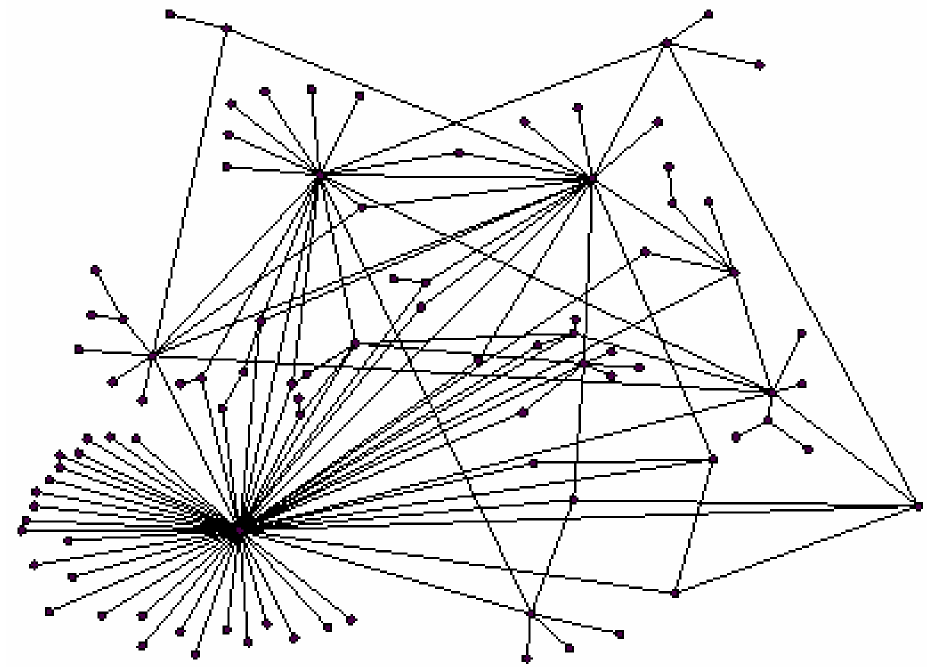
$$N = \left(\frac{K}{x_m}\right)^{\alpha - 1} \approx 10^{12}$$

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
  - $X \dots$ degree of node v,
  - $X_w \dots$ 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!**
  - $X, X, \dots, X_n$: rnd. vars with mean μ, variance σ^2
  - $S_n = \sum X_i \; : \; E[S_n] = n\mu \, , \mathrm{var}[S_n] = n\sigma^2, \mathrm{SD}[S_n] = \sigma\sqrt{n}$
  - $P(S_n = E[S_n] + x \cdot \mathrm{SD}[S_n]) \sim \frac{1}{2\pi} \mathrm{e}^{-\frac{x^2}{2}}$

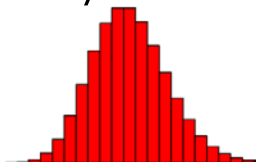
Random vs. Scale-free network





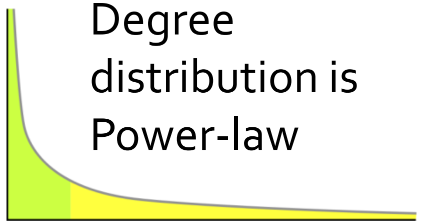
Random network

(Erdos-Renyi random graph)



Degree distribution is Binomial

Scale-free (power-law) network



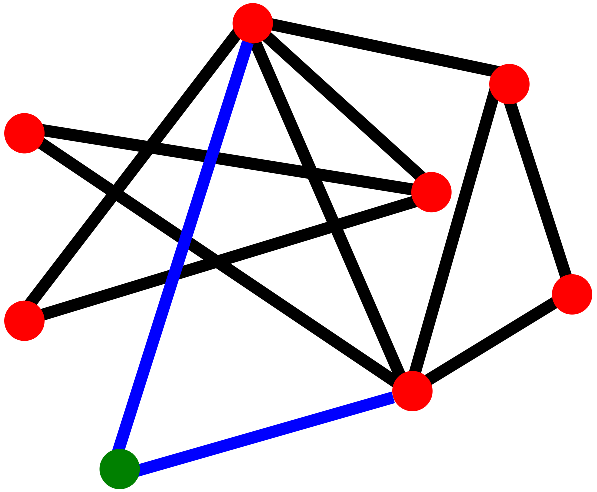
Degree distribution is Power-law

Model:
Preferential Attachment

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**

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



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** [Price 65]:
  - **Citations:** New citations to a paper are proportional to the number it already has

[Mitzenmacher, '03]

The Exact Model

- **We will analyze the following model:**
- Nodes arrive in order $1,2,3,\ldots,n$
- When node j is created it makes a single 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. $1\text{-}p$, node j chooses node i uniformly at random and links to a node i points to.
    - **Note this is same as saying:** With prob. $1\text{-}p$, node j links to node u with prob. proportional to d_u (the degree of u)

The Model Givens Power-Laws

- **Claim:** The described model generates networks where the fraction of nodes with degree k scales as:

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

where q=1-p

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

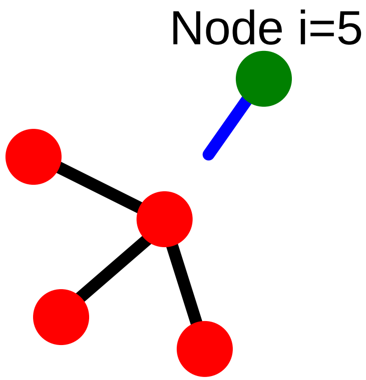
Continuous Approximation

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

Continuous Approximation

- **Plan:** Analyze continuous degree $d_i(t)$ of node i at time $t \geq i$
- **Node i=t=5 comes and has degree of 1 to share with other nodes:**

i	$d_i(t-1)$	$d_i(t)$
1	1	$=1 + p\frac{1}{4} + (1-p)\frac{1}{6}$
2	3	$=3 + p\frac{1}{4} + (1-p)\frac{3}{6}$
3	1	$=1 + p\frac{1}{4} + (1-p)\frac{1}{6}$
4	1	$=1 + p\frac{1}{4} + (1-p)\frac{1}{6}$
i=5	0	1



Continuous Degree: What We Know

- **Initial condition:**
  - $d_i(t)=0$, when $t=i$ (node i just arrived)
- **Expected change 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. $1\text{-}p$ node $t+1$ links **preferentially**:
      - Links to our node i with prob. $d_i(t)/t$
  - **So: Prob. node $t+1$ links to i is:** $p\frac{1}{t} + (1-p)\frac{d_i(t)}{t}$

What is the rate of growth of d_i?

- $\frac{\mathrm{d}d_i(t)}{\mathrm{d}t} = p\frac{1}{t} + (1-p)\frac{d_i(t)}{t} = \frac{p+qd_i}{t}$

Divide by p+q $d_i(t)$

- $\frac{1}{p+qd_i(t)}\mathrm{d}d_i(t) = \frac{1}{t}\mathrm{d}t$

integrate

- $\int \frac{1}{p+qd_i(t)}\mathrm{d}d_i(t) = \int \frac{1}{t}\mathrm{d}t$

Let $A=e^c$ and exponentiate

- $\frac{1}{q}\ln\left(p + qd_i\,(t)\right) = \ln t + c$
- $qd_i(t) + p = A\,t^q \;\Rightarrow d_i\,(t) = \frac{1}{q}(At^q - p)$

What is the constant A?

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

- **What is the constant A?**
- We know: $d_i(i) = 0$
- So: $d_i(i) = \frac{1}{q}(Ai^q - p) = 0$
- $\Rightarrow A = \frac{p}{i^q}$
- $\Rightarrow d_i(t) = \frac{p}{q}\left(\left(\frac{t}{i}\right)^q - 1\right)$

Degree Distribution

- What is F(d) the fraction of nodes that has degree at least d at time t?
  - **How many nodes i have degree > t?**
    - $d_i(t) = \frac{p}{q}\left(\left(\frac{t}{i}\right)^q - 1\right) > d$
    - then: $\mathrm{i} < \mathrm{t}\left(\frac{q}{p}d - 1\right)^{-\frac{1}{q}}$
- There are t nodes total at time t so F(d):

$$F(d) = \left[\frac{q}{p}d + 1\right]^{-\frac{1}{q}}$$

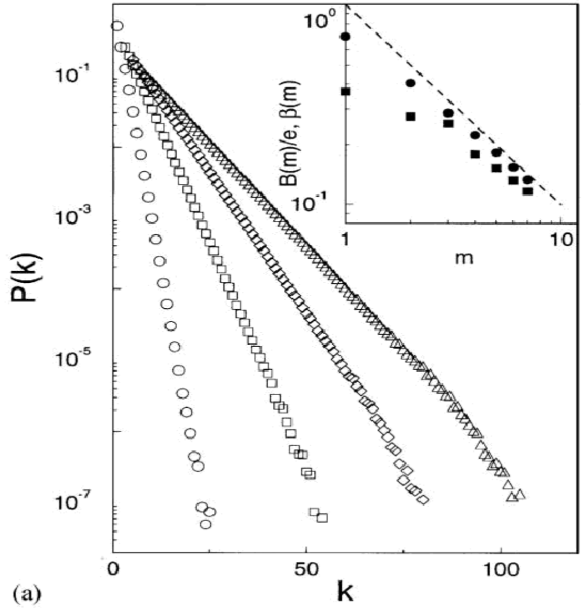
Degree Distribution

- **What is the fraction of nodes with degree exactly d?**
  - Take derivative of F(d):

$$F'(d) = \frac{1}{p}\left[\frac{q}{p}d + 1\right]^{-1-\frac{1}{q}} \Rightarrow \alpha = 1 + \frac{1}{q}$$

Preferential attachment: Reflections

- Two changes from the G_{np} model
  - The network grows
  - Preferential attachment
- Do we need both? **Yes!**
  - **Add growth to G_{np}** (assume $p=1$):
    - x_j = degree of node j at the end
    - $X_j(u) = 1$ if u links to j, else 0
    - $x_j = x_j(j+1)+x_j(j+2)+\ldots+x_j(n)$
    - $E[x_j(u)] = P[u \text{ links to } j] = 1/(u-1)$
    - $E[x_j] = \sum 1/(u-1) = 1/j + 1/(j+1)+\ldots+1/(n-1) = H_{n-1} - H_j$
    - $E[x_j] = \log(n-1) - \log(j) = \log((n-1)/j)$ **NOT** $(n/j)^{\alpha}$



(a)

H_n…n-th harmonic number: $= \sum_{k=1}^{n} \frac{1}{k}.$

Preferential attachment: Good news

- **Preferential attachment gives power-law degrees**
- Intuitively reasonable process
- Can tune p to get the observed exponent
  - On the web, $P[\textit{node has degree d}] \sim d^{-2.1}$
  - $2.1 = 1+1/(1-p) \rightarrow \underline{p \sim 0.1}$

There are also other network formation mechanisms that generate scale-free networks:

- Random surfer model
- Forest Fire model

PA-like Link Formation

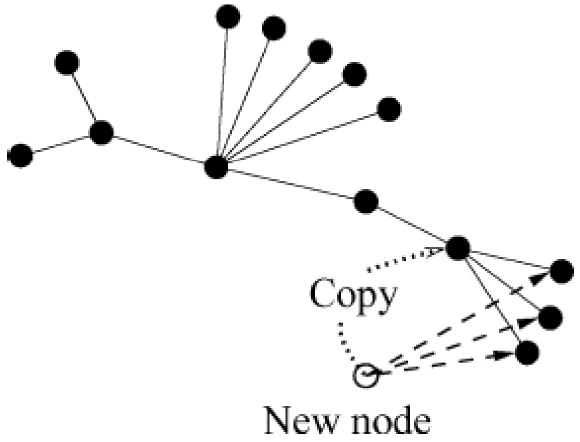
- **Copying mechanism** (directed network)
  - select a node and an edge of this node
  - attach to the endpoint of this edge
- **Walking on a network** (directed network)
  - the new node connects to a node, then to every
  - first, second, … neighbor of this node
- **Attaching to edges**
  - select an edge
  - attach to both endpoints of this edge
- **Node duplication**
  - duplicate a node with all its edges
  - randomly prune edges of new node

Preferential attachment: Bad news

- **Preferential attachment is not so good at predicting network structure**
  - Age-degree correlation
  - Links among high degree nodes
    - On the web nodes sometime avoid linking to each other
- **Further questions:**
  - What is a reasonable probabilistic model for how people sample through web-pages and link to them?
    - Short+Random walks
    - Effect of search engines – reaching pages based on number of links to them

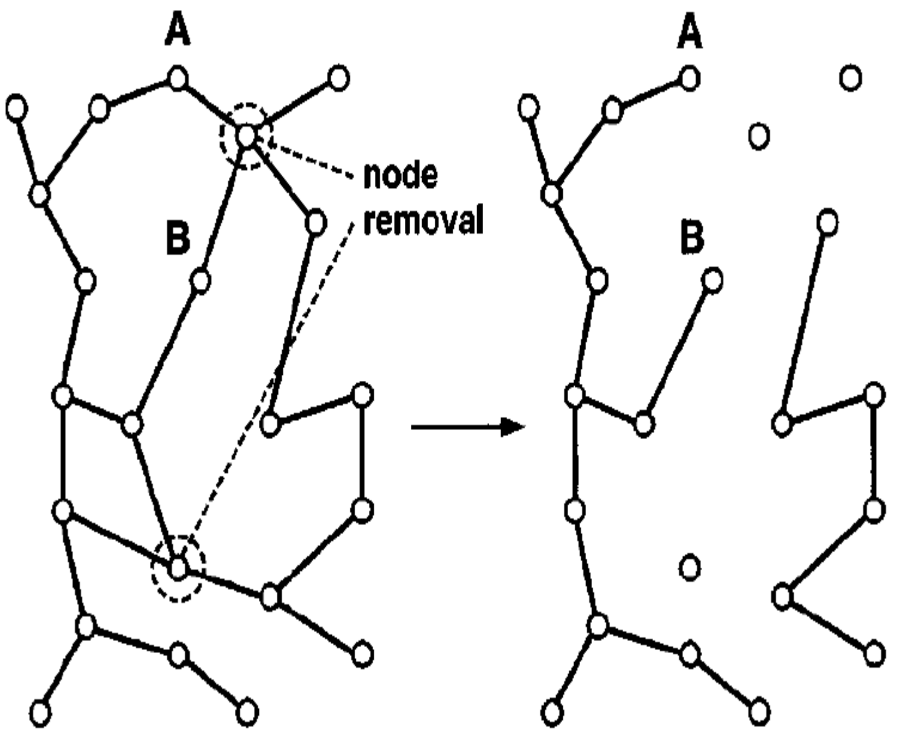
PA: Many Extensions & Variations

- Preferential attachment is a key ingredient
- **Extensions:**
  - Early nodes have advantage: node fitness
  - Geometric preferential attachment
- **Copying model:**
  - Picking a node proportional to the degree is same as picking an edge at random (pick node and then it’s neighbor)



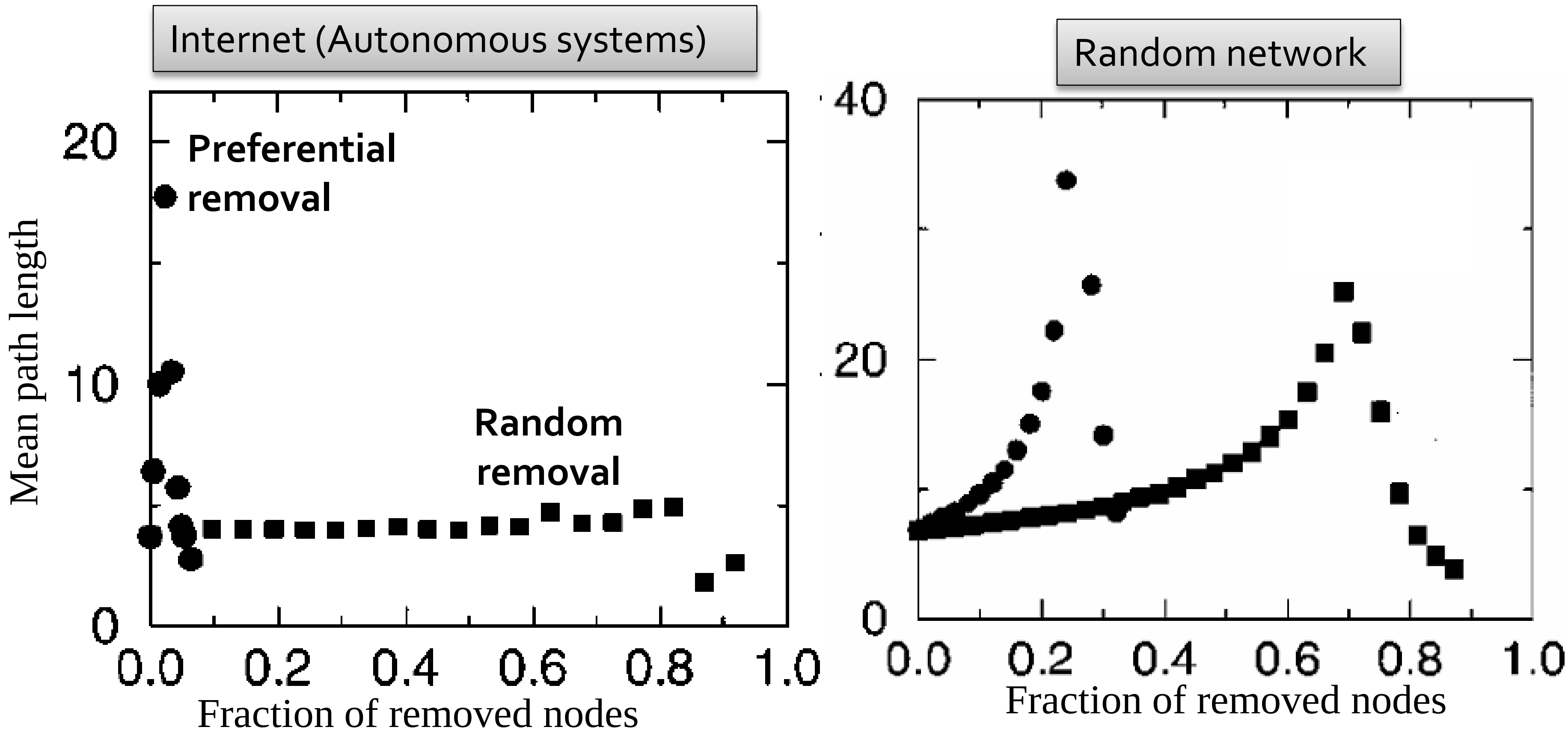
Network resilience (1)

- We observe how the connectivity (length of the paths) of the networl changes as the vertices get removed [Albert et al. 00; Palmer et al. 01]
- Vertices can be removed:
  - Uniformly at random
  - In order of decreasing degree
- It is important for epidemiology
  - Removal of vertices corresponds to vaccination



Network resilience (2)

- Real-world networks are resilient to random attacks
  - One has to remove all web-pages of degree > 5 to disconnect the web
  - But this is a very small percentage of web pages
- Random network has better resilience to targeted attacks



Network resilience (2)

- Real-world networks are resilient to random attacks
  - One has to remove all web-pages of degree > 5 to disconnect the web
  - But this is a very small percentage of web pages
- Random network has better resilience to targeted attacks